Разработка и исследование генератора псевдослучайных чисел

Лабораторная работа

Ревизия: 0.1

Последняя модификация: Vlad Kovtun
Дата последней модификации: 13.11.2009 0:41:00

История изменений

20.09.2009 – Версия 0.1. Первичный документ. Владислав Ковтун

Содержание

Лабораторная работа 2. Разработка и исследование генератора псевдослучайных чисел

Вопросы

Общие положения.

Постановка задачи.

Методические рекомендации.

Постановка задачи

Цель

1. Разработать консольное приложение, которое формирует псевдослучайную последовательность.

2. Исследовать полученную псевдослучайную последовательность на случайность.

Задачи

1. Ознакомиться с требованиями, которые выдвигаются к псевдослучайным последовательностям.

2. Разработать консольное приложение для формирования псевдослучайной последовательности:

- линейным генератором;
- квадратичным генератором;
- кубическим генератором;
- встроенным генератором в язык программирования высокого уровня;
- встроенным криптографическим генератором в стандартный криптопровайдер Windows.

3. Исследовать сформированные последовательности.

Вход

Для управления приложением предлагается использовать ключи командной строки:

- /g:<название> – указывается тип генератора. Следует выделить следующие наборы:
- lc – линейный конгруэнтный генератор.
- qc – квадратичный конгруэнтный генератор.
- cc – кубический конгруэнтный генератор.
- csp – генератор встроенный в криптопровайдер ОС Windows.
- int – генератор встроенный в язык высокого уровня.
- /l:<длина> - длина формируемой последовательности в килобайтах. Если данный параметр не указан, то принимается число 12288 (что соответствует размеру в 12 Мб)
- /o:<filename> – полный путь к файлу, либо имени файла, который будет хранить сформированную последовательность. Если данный параметр не указывается, то вывод производится в файл rndseq.dat.
- /? - вывод информации о допустимых ключах командной строки.

Пример

Генерация последовательности, которая содержит 2 Мб посредством линейного конгруэнтного генератора, результат выводится в файл rnd.dat.

C:/>rndgen.exe /g:lc /o:rnd.dat /l:2048

Вывод

Во время работы приложения, рекомендуется выводить информацию о статусе приложения, а также о корректности его работы на консоль.

Допускается перенаправление вывода консоли в текстовый файл, например:

C:/>rndgen.exe /g:lc /o:rnd.dat /l:2048 >report.txt

Требования

Архитектура приложения строится по модульному принципу.

За основу принимается стандартная библиотека C++ (в случае разработки на C++) либо набор классов стандартной библиотеки C#.

Рекомендуется использовать защищенные ресурсы и указатели.

Алгоритм

Кратко опишем алгоритм данного приложения:

1. Чтение начальных значений для выбранного генератора из файла настроек приложения, проверка их корректности.
2. Открытие файлов с выходными данными. Проверка на корректность.
3. Инициализация модуля генерации псевдослучайных чисел.
4. Выполнение генерации последовательности указанной длины.
5. Закрытие файла с входными и выходными данными. Проверка на корректность.

Теоретические сведения

Наиболее распространенным на практике подходом к определению псевдослучайности является эвристический подход. При таком подходе псевдослучайный генератор рассматривается как программа (алгоритм), которая порождает битовую последовательность $s = s_0, s_1, .., s_{n-1}$ конечной длины n, которая проходит некоторые статистические тесты. Таким образом, свойство случайности либо псевдослучайной последовательности могут быть охарактеризованы и описаны в вероятностном виде.

Существует множество тестов, которые дают оценку, является ли последовательность случайной. Однако некоторый конечный набор тестов не считают достаточным. Кроме того, результаты статистического теста следует интерпретировать с некоторой осторожностью и предостережениями, что бы избежать неправильных выводов относительного исследуемого генератора.

Статистический тест формулируется для проверки определенной нулевой гипотезы H_0 о том, что последовательность - случайна. С этой нулевой гипотезой связана альтернативная гипотеза H_A, что последовательность не является случайной. Для каждого применяемого теста можно сделать вывод касательно принятия либо отклонения нулевой гипотезы, исходя из сформированной генератором последовательности.

Для каждого теста следует выбрать адекватную статистику случайности, на основании которой может быть принята либо отклонена нулевая гипотеза. Соответственно о допущении на случайность, такая статистика обладает некоторым распределением случайных значением. Теоретически для нулевой гипотезы распределение этой статистики определяется математическими методами. Из этого эталонного распределения определяется критическое значение. Во время проведения теста рассчитывается значение тестовой статистики. Это значение сравнивается с критическим значением. Если критическое значение тестовой статистики превышает критическое значение, нулевая гипотеза касательно случайности отклоняется. Иначе нулевая гипотеза принимается.

Проверка статистических гипотез работает благодаря тому, что эталонное распределение и критические значения зависят и генерируются в соответствии с допущением о случайности. Если допущение о случайности – истинно, то результат тестовой статистики для нее будет иметь очень низкую вероятность превышения критического значения (например: 0,01). Если расчетное значение тестовой статистики превышает критическое значение (тоесть возникает событие с очень низкой вероятностью), тогда с точки зрения проверки статистической гипотезы событие с низкой вероятностью не может встречаться по своей природе. В связи с этим, когда расчетное значение тестовой статистики превышает критическое значение, делается вывод, что первое допущение касательно случайности является подозрительным либо ошибочным. В таком случае делается вывод об отклонении H_0 (случайность), и принятии H_A (не случайность). Проверка статистической гипотезы является процедурой генерации выводов, во время выполнения которой можно либо принять H_0 (данные случайные) либо отклонить H_0 (данные не случайные). Таблица 1 связывает истинное (неизвестное) состояние данных с выводом, полученным посредством процедуры проверки.

Таблица 1. Виды ошибок

Ситуация	Вывод	
	Принять H_0	Принять H_A
Данные случайные (H_0 - истинна)	Нет ошибки	Ошибка 1-го рода
Данные неслучайны (H_A - истинна)	Ошибка 2-го рода	Нет ошибки

Если данные действительно случайные, то вывод об отклонении нулевой гипотезы принимается очень редко, этот вывод имеет название «Ошибка первого рода». Если данные не случайные, то вывод о принятии нулевой гипотезы (тоесть данные случайные) имеет название «Ошибка второго рода». Вывод о принятии H_0, когда данные действительно случайные, и отклонение H_0, когда данные не случайные является правильным.

Вероятность ошибка 2-го рода называется ***уровнем значимости теста***. Эта вероятность может быть установлена для тестов и обозначается как α. Для теста суть α состоит в том, что тест покажет неслучайность последовательности, тогда как в действительности она случайна, то есть последовательность имеет неслучайные свойства даже тогда, когда он сформирована «хорошим» генератором.

Вероятность ошибка 2-го рода обозначается как β. Для теста β - вероятность того, что тест покажет на случайность последовательности, когда она в действительности является неслучайной. То есть «плохой» генератор сформировал последовательность, которая, как будто обладает случайными свойствами. В отличие от α, ошибка 2-го рода β не является фиксированным значением. Она может принимать целый ряд различных значений, т.к. существует множество ситуаций, когда поток данных может быть неслучайным, и каждая из них может быть охарактеризована различными значениями β. Вычисление ошибки 2-го роду является более сложным заданием из-за существования большого количества различных типов неслучайности.

Для принятия решения о прохождении последовательностью случайных чисел статистического теста, используют **три основных подхода**.

Пусть дана двоичная последовательность $S = \{s_1, s_2, ..., s_n\}$, $s_i \in \{0,1\}$ длиною n бит. Необходимо принять решение, проходит ли данная последовательность статистический тест либо нет. Возможны такие подходы:

1. Критерий принятия решения на основе установки граничного уровня. Этот подход базируется на вычислении статистики теста $c(S)$, с дальнейшим сравнением с некоторым граничным значением $c_{bound}(S)$. Критерий принятия решения формируется таким образом: *предполагается, что двоичная последовательность S не проходит статистический тест каждый раз, когда статистика теста $c(S)$ принимает значение за граничный уровень $c_{bound}(S)$.*

Например, во время проверки сложности последовательности с использованием теста на основе алгоритма Лемпеля-Зива, для заданной двоичной последовательности S рассчитывается ее сложность $c(S)$. Для того, чтобы определить прошла ли данная последовательность тест или не, необходимо сравнить полученное значение $c(S)$ с граничным значением $n/\log_2 n$. Однако такой подход не является достаточно надежным. Как показали практические исследования, принятие решение на основе этого критерия, как единственного, часто приводит к ошибочным решениям.

2. Критерий принятия решений на основе установки фиксированного доверительного интервала. При таком подходе критерий принятия решения формируется следующим образом: предполагается, что двоичная последовательность S *не проходит статистический тест каждый раз, когда статистика теста $c(S)$ принимает значение за пределами граничного интервала, которые вычислены для заданного уровня значимости α*. Например, пусть к двоичной последовательности S длиной $n = 800$ бит применяется частотный тест. Значениям статистики тесту S является количество единиц в последовательности S, причем ожидается, что последовательность будет приблизительно 400 единиц и 400 нулей. Если зафиксировать уровень значимости на уровне 5% (α =0,05), то последовательность S не пройдет частотный тест, если количество единиц будет находится за пределами доверительного интервала $400 \pm 1{,}96/2 \times \sqrt{800} = [373{,}427]$.

Данный критерий, в сравнении с первым, является более надежным. Необходимо только учитывать, что разным уровням значимости соответствуют различные доверительные интервалы.

3. Третий подход построения критерия принятия решения опирается на вычисление, для статистики теста $c(S)$ соответствующего значения вероятности $\mathbf{P}-value.$ Тут статистика теста рассматривается как реализация случайной величины, которая подчиняется известному закону распределения. Статистика теста строится таким образом, чтобы ее меньшие значения указывали на любой дефект случайности последовательности. Значения вероятности $\mathbf{P}-value$ является вероятность события, что статистика теста примет значение большее за значения, что наблюдается при испытании последовательности, в предсказании ее случайности. Таким образом небольшие значения $\mathbf{P}-value$ ($\mathbf{P}-value$ < 0,05 либо $\mathbf{P}-value$ < 0,01) интерпретируются как доказательство того, что последовательность не случайна. Результирующее правило формируется следующим образом: *для фиксированного значения уровня значимости* α, двоичная последовательность S *не проходит статистический тест, если значение вероятности* $\mathbf{P}-value < \alpha$. Значение α рекомендуется выбирать из интервала [0.001, 0.01].

Значение α =0.001, говорит о том, что из 1000 случайных последовательностей, которые тестируются, не прошла бы тест лишь одна. При $\mathbf{P}-value \geq 0.001$ последовательность рассматривается как случайная с уровнем доверия 99,9 %. При $\mathbf{P}-value < 0.001$ последовательность рассматривается, как неслучайная с уровнем доверия 99,9%.

Значение α =0.01, говорит о том, что из 100 случайных последовательностей не прошла бы тест лишь одна. При $\mathbf{P}-value \geq 0.01$ последовательность рассматривается как случайная с уровнем доверия 99 %. При $\mathbf{P}-value < 0.01$ последовательность рассматривается, как неслучайная с уровнем доверия 99%.

Одной из основных целей тестов, которые строятся в соответствии с третьим подходом, является минимизация вероятности ошибки второго рода. Иначе говоря, минимизация вероятности принятия последовательности, сформированной «плохим» генератором за последовательность, сформированную «хорошим» генератором. Вероятности α, β связаны между собой и с длиной последовательности n, что проверяется: если два из этих значений определены, третье определяется автоматически. На практике, обычно, выбирают размер n и значение α (вероятность ошибки второго рода). Тогда критическая точка выбирается таким образом, чтобы получить наименьшее значение β (вероятность ошибки второго рода).

Таким образом, на сегодняшний день, основным подходом, который можно использовать для исследования свойств генераторов (сформированных ими последовательностей), является эвристический подход. Именно такой подход в дальнейшем используется в лабораторных исследованиях. Одной из важнейших задач в использовании эвристического подхода на практике является обоснование набора статистических тестов. Состав тестов зависит от назначения генераторов и способов использования последовательностей. С философской точки зрения не существует возможности, посредством эвристического подхода, доказать случайность последовательности. В этом случае конкретной последовательности следует пройти целый ряд тестов. Однако и здесь нельзя утверждать, что такая послдедовательность случайна, потому что возможно создание нового теста, по которому она может не пройти. В связи с этим невозможно также построение и универсального теста, например универсальный тест Маурэра. Как показали исследования, возможны варианты, когда такие тесты проходят и неслучайные последовательности.

На сегодняшний день существует несколько признанных методик статистического тестирования, как иностранных, так и отечественных специалистов. В таблице 2 приведена информация об известных системах статистических тестов. В лабораторных исследованиях используются методика статистического тестирования генераторов, которая предложена Национальным институтом стандартизации и технологий США.

Таблица 2. Системы статистических тестов

Авторы	Источник
Д. Кнут (Стенфордский университет, США)	Искусство программирования. Т.2. Получисленные алгоритмы
Дж. Марсалья (Флоридский Госдуратсвенный университет, США)	Система статистического тестирования DIEHARD

Х. Густавсон и др. (Куиндслэндский технологический университет, Австралия)	Система CRYPT-S
Методика тестирования FIPS 140-2	NIST PUB FIPS 140-2
І. Горбенко, О. Потій (Харьковский военный университет, Украина)	Методика статистического тестирования генераторов псевдослучайных последовательностей
Методика NIST (США)	NIST Statistical test Suite
Методика тестирования RIPE	www.nessie.org

Коротко остановимся на наиболее популярной.

Методика тестирования NIST STS

Общие положения

Набор тестов NIST STS был предложен в ходе проведения конкурса на новый национальный стандарт США блокового шифрования. Этот набор использовался для исследования статистических свойств кандидатов на новый блоковый шифр. Сегодня методика тестирования, предложенная NIST. является наиболее распространенной у разработчиков криптографических средств защиты информации.

Порядок тестирования отдельной двоичной последовательности S имеет такой вид:

- Выдвигается нулевая гипотеза H_0 - допущение того, что данная двоичная последовательность S случайна.
- Далее рассчитывается статистика теста $c(S)$.
- Используя специальную функцию статистики теста, вычисляется значение вероятности $\mathbf{P} = f(c(S)),\ P \in [0,1]$.
- Значение вероятности $\mathbf{P}$ сравнивается с уровнем значимости $\alpha \in [0.001, 0.01]$. Если $\mathbf{P} \geq \alpha$, то гипотеза $\mathbf{H}_0$ принимается. В противном случае принимается альтернативная гипотеза.

Пакет содержит в себе 16 статистически тестов. Однако фактически, в зависимости от входных параметров вычисляется 189 значений вероятности $\mathbf{P}$, которые можно рассматривать как результат работы отдельных тестов. В таблице 3 перечисляются собранные данные по всем тестам с указанием количества значений вероятности $\mathbf{P}$, которые вычисляются, физического смысла статистики и дефекта на определение которого направлен тест.

Таким образом, в результате тестирования двоичной последовательности формируется вектор значений вероятности $\mathbf{P} = \{\mathbf{P}_1, \mathbf{P}_2, \ldots, \mathbf{P}_{189}\}$. Анализ значений $\mathbf{P}_i$ данного вектора позволяет указать конкретные дефекты случайности тестируемой последовательности.

Таблица 3. Статистические тесты NIST STS

Номер	Статистический тест	Статистика теста $c(S)$	Выявляемый дефект
1	Частотный (монобитный тест)	Нормализованная абсолютная сума значений элементов последовательности	Слишком много нулей либо единиц в последовательности
2	Частотный тест (в средине блока)	Мера согласования количества наблюдаемых единиц, с тем, что ожидается теоретически	Локализованные отклонения частоты появления единиц в блоке от идеального значения ½
3	Проверка накопленных сумм	Максимальное отклонение значений накопленной суммы элементов последовательности от начальной точки отсчета (точка 0)	Большое количество единиц либо нулей в начале либо в конце двоичной последовательности
4	Проверка серий	Общее количество серий на все длине последовательности	Слишком быстрое либо слишком медленное изменение знака в ходе генерации последовательности
5	Проверка	Мера согласованности значений	Отклонение от

	максимальной длины серии в блоке	максимальной длины, которые наблюдаются, с теми значениями, что ожидаются теоретически	теоретического закона распределения максимальной длины серий единиц
6	Проверка ранга двоичной матрицы	Мера согласованности значений рангов различного порядка, которые наблюдаются, с значениями, которые ожидаются теоретически	Отклонение эмпирического закона распределения значений рангов матриц от теоретического, что указывает на зависимость элементов последовательности
7	Спектральный анализ на основе дискретного преобразования Фурьє	Нормализованная разница количества частотных компонент, что наблюдаются с той, что ожидается теоретически, если превышает пороговый уровень 95%	Выявление периодических слагаемых (трендов) в двоичной последовательности
8	Проверка перекрывающихся шаблонов	Мера согласованности количества наблюдаемых перекрывающихся шаблонов в последовательности с теоретическими значениями	Большое количество m-битных серий единиц в последовательности.
9	Универсальный тест Маурера	Сумма логарифма расстояний между l–битными шаблонами	Возможность сжатия последовательности
10	Энтропийный тест	Мера согласованности наблюдаемого значения энтропии источника, с тем, что теоретически ожидается из случайного источника	Неравномерность распределения m-битных слов в последовательности (регулярность свойств источника)
11	Проверка случайных отклонений	Мера согласованности наблюдаемого количества визитов при случайном блуждании с заданное состояние в середине цикла, в сравнении с тем, что ожидается теоретически	Отклонение от теоретического закона распределения визитов в конкретное состояние при случайном блуждании
12	Проверка случайных отклонений (вариант)	Общее количество визитов при случайном блуждании	Отклонение от теоретически ожидаемого количества общего количества визитов при случайном блуждании в заданное состояние
13	Последовательный тест	Мера согласованности количества наблюдаемых всех вариантов m- битных шаблонов с тем количеством, которое ожидается теоретически	Неравномерность распределения m- битных слов в последовательности
14	Проверка сжатия согласно алгоритму Лемпеля-Зива	Количество разных слов в последовательности	Большая степень сжатия тестируемой последовательности и степенью сжатия ожидаемой случайной последовательности
15	Проверка шаблонов, что не перекрываются	Мера согласованности ожидаемого количества непериодических шаблонов в последовательности с теоретическим значением	Большое количество заданных непериодических шаблонов в последовательности
16	Проверка линейной сложности	Мера согласованности ожидаемого количества событий, что лежат в основе появления фиксированной длины эквивалентного ЛРР для	Отклонение эмпиричного распределения длины эквивалентных ЛРР для последовательностей фиксированной длины от

		заданного блока с теоретическим	теоретического закона распределения для случайной последовательности, что указывает на недостаточную сложность тестируемой последовательности

Методика тестирования NIST STS

Методика тестирования имеет следующий вид:

1. Для каждого ГСП необходимо оценить и принять решение о том, что он формирует случайные двоичные последовательности. Генератору следует формировать двоичную последовательность $S = \{s_1, s_2, \ldots, s_n\}$, $s_i \in \{0,1\}$ произвольной длины n.

2. Для фиксированного значения n формируют множество из m двоичных последовательностей:

$$\begin{aligned} &S_1 = s_{11}, s_{12}, \ldots, s_{1n} \\ &S_2 = s_{21}, s_{22}, \ldots, s_{2n} \\ &\ldots \\ &S_m = s_{m1}, s_{m2}, \ldots, s_{mn} \end{aligned} \qquad (1.1)$$

Таким образом, для тестирования необходимо сформировать выборку объемом $N = m \times n$.

3. Каждую последовательность проверяют с использованием пакета NIST STS. В результате формируется ***статистический портрет*** генератора, таблица 4.

Таблица 4. Статистический портрет генератора

№ теста j / № последовательности i	1	2	…	q
S_1	P_{11}	P_{12}		P_{1q}
S_2	P_{21}	P_{22}		P_{2q}
	:			
S_m	P_{m1}	P_{m2}		P_{mq}

либо в матричном виде

$$\begin{pmatrix} P_{11} & P_{12} & \cdots & P_{1q} \\ P_{21} & P_{22} & \cdots & P_{1q} \\ \vdots & \vdots & \ddots & \vdots \\ P_{m1} & P_{m2} & \cdots & P_{mq} \end{pmatrix} \qquad (1.2)$$

Статистическим портретом генератора является матрица размерностью $m \times q$, де m – количество проверяемых двоичных последовательностей, а q – количество статистических тестов, используемых для тестирования каждой последовательности. Элементы матрицы $P_{ij} \in [0,1]$, где $i = \overline{1,m}$ и $j = \overline{1,q}$ представляют собой значения вероятности, которая получена в результате тестирования i-ой последовательности j-ым тестом.

4. Согласно полученному статистическому портрету определяют долю последовательностей, которые прошли каждый статистический тест. Для этого задают уровень значимости $\alpha \in [0.001,\ 0.01]$ и выполняют подсчет значений вероятности, что превышают установленный уровень значимости α для каждого из q тестов, то есть определяют коэффициент:

$$r_j = \frac{\#\{P_{ij} \geq \alpha | i = \overline{1,m}\}}{m}. \qquad (1.3)$$

В результате формируется вектор коэффициентов $\mathbf{R} = \{r_1, r_2, r_q\}$, элементы которого характеризуют, в процентах, прохождения последовательности S_i всех статистических тестов.

Правило 1. Предполагается, что генератор G прошел тестирование по j-ому тесту, если значение коэффициента r_j находится в пределах доверительного интервала $[r_{\max}, r_{\min}]$. Границы доверительного интервала определяются в соответствии с выражением:

$$r_{\max(\min)} = \hat{p} \pm 3\sqrt{\frac{\hat{p}(1-\hat{p})}{m}}, \text{ где } \hat{p} = 1-\alpha. \qquad (1.4)$$

5. Производится статистический анализ статистического портрета. Полученные значения вероятностей P_{ij} подчиняются равновероятному закону распределения на интервале $[0, 1]$. Для вектора-столбца статистического портрета строится гистограмма частот F_k попадания значений P_{ij} в каждый из $k = \overline{1,10}$ подинтервалов, на которые разбит интервал $[0, 1]$. Равновероятность распределения значений вероятностей P_{ij}, проверяется с использованием критерия χ^2. Для этого рассчитывается статистика вида:

$$\chi_j^2 = \sum_{k=1}^{10} \frac{\left(F_k - m/10\right)^2}{m/10}, \qquad (1.5)$$

которая подчиняется распределению χ^2 с девятью степенями свободы.

Правило 2. Предполагается, что генератор G прошел тестирование по j-му тесту, если выполняется условие $P(\chi_j^2) > 0.0001$.

6. Предпоследнее решение принимают в соответствии с правилом: предполагается, что генератор G прошел статистическое тестирование пакетом NIST STS, если значение коэффициентов r_j для всех $j = \overline{1,q}$ находятся в пределах доверительного интервала $[r_{\min}, r_{\max}]$ и выполняется условие $P(\chi_j^2) > 0.0001$ для всех $j = \overline{1,q}$.

Задание на лабораторную работу

Провести статистические испытания указанных генераторов случайных последовательностей в соответствии с методикой NIST STS. Провести анализ полученных результатов и сформировать акт испытания.

1. С помощью разработанного программного обеспечения сформировать псевдослучайную последовательность.

2. Провести тестирование сформированных последовательностей с использованием всех тестов пакета NIST STS.

3. Провести анализ результатов путем использования статистического критерия χ^2. Для этого:

Построить гистограмму распределения частот, исходя из полученных результатов, и сравнить их с теоретическим распределением, используя критерий χ^2;

Построить эмпирическую функцию распределения, сравнить ее с теоретическим распределением, пользуясь критерием Колмогорова.

Сравнить таблицу результатов статистических исследований, в которой сравниваются результаты с эталонной выборкой BBS.

4. Подготовить и оформить отчет о проведенной лабораторной работе.

Порядок выполнения работы

1. Запустить приложение NIST_STS.exe для статистического тестирования.

2. Выполнить статистическое тестирование заданных ГВП. Для этого необходимо воспользоваться руководством Приложение А.

3. Используя обобщенные результаты тестирования, которые находятся в файле finalAnalysisReport, выполнить интерпретацию результатов тестирования (См. пример в Приложение Б).

4. Используя результаты тестирования из файла finalAnalysisReport, для указанных в работе статистических тестов построить теоретическую и эмпирическую гистограмму частостей F_k попадания значений P_{ij} в каждый из интервалов $k=\overline{1,10}$ подинтервалов, на которые разбит интервал $[0,1]$, рис. 1.

5. Согласно выражения:

$$\chi_j^2 = \sum_{k=1}^{10} \frac{\left(F_k - m/10\right)^2}{m/10}.$$

Рассчитать значение χ^2 и проверить согласованность теоретического и эмпирического закона распределения значений P_{ij} при $\alpha = 0{,}01$ и $\alpha = 0{,}05$. Критические значения для χ^2 представлены в Приложении В.

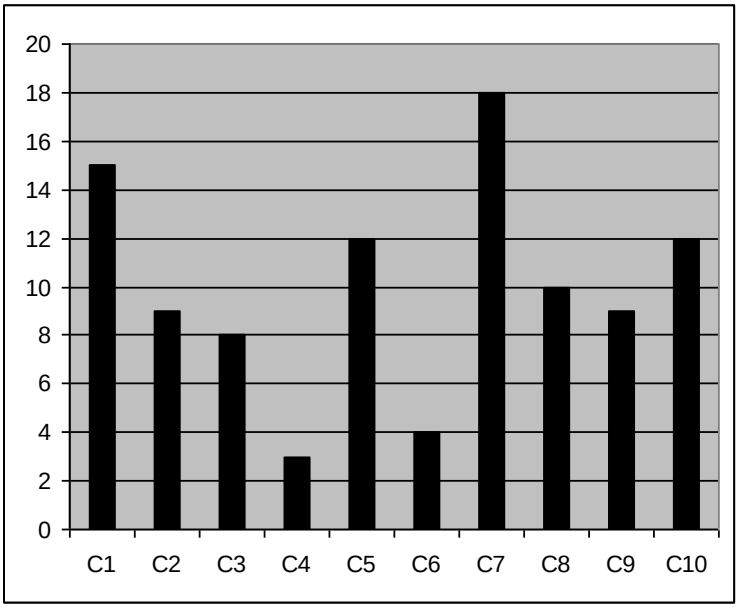

Рис. 1. Образец построения гистограммы

6. Проверить согласованность эмпирического и теоретического закона распределения значений P_{ij} для указанных в работе статистических тестов посредством критерия Колмогорова-Смирнова. Для этого, посредством результатов тестирования, что находятся в файле result, соответствующего теста, построить эмпирическую функцию распределения и теоретическую функцию распределения для равновероятного закона. Вычислить статистику Колмогорова-Смирнова, см. Приложение В. Как критерий оценки разницы оценки теоретической $F_t(P_{ij})$ и эмпиричной $F_e(P_{ij})$ функции распределения рассматривается максимальное значение модуля разницы:

$$D = \max\left|F_T(P_{ij}) - F_e(P_{ij})\right|.$$

Определяется значение:

$$\beta = D\sqrt{m},$$

где m – количество точек, по которым строится эмпирическая кривая.

7. Зная β, представлено в таблице 2 Приложения В вычислить вероятность $P(\beta)$. Если эта вероятность небольшая, то гипотеза о согласованности законов отклоняется, м наоборот, если вероятность большая, то гипотеза принимается.

Проверить согласованность результатов, которые были получены с помощью критерия χ^2 и критерия Колмогорова-Смирнова, сделать выводы.

Замечание: С целью упрощения заполнения отчета, следует пользоваться электронными таблицами Microsoft Excel.

Практические задания

1. Произведите тестирование произвольного файла, потом выполните его архивацию (RAR, ZIP), повторите тестирование. Как изменились результаты? Поясните причину изменения.

2. Выполните тестирование последовательности посредством частотного теста (в средине блока). Измените длину блока, сравните результаты, сделайте выводы.

3. На основе полученных в ходе проведения лабораторной работы данных рассчитайте среднюю частоту прохождения всех тестов одной последовательностью.

4. Для заданной длины, количества последовательностей и длины шаблона, вычислить количество элементарных операций, которые необходимо провести для выполнения теста шаблонов, что не перекрываются.

5. Для заданной длины, количества последовательностей и длины вычислить количество элементарных операций, которые необходимо провести для выполнения теста перекрываемых шаблонов. Сравните полученные результаты с результатом пункта 4.

Литература

1. Microsoft Developer Network. Доступно по адресу: www.msdn.com

2. Теоретические материалы и программное обеспечение доступны по адресу: www.nrjetix.com/r-and-d/lectures

3. Руководство пользователя к пакету NIST STS. Доступно: NIST STS Guide

4. Средство для статистического анализа случайных последовательностей. Доступно: NIST STS Software

5. Потій О.В., Лєншин А.В., Горбенко Ю.І. Методичні вказівки до лабораторних робіт за дисципліною «стандартизація та сертифікація в галузі інформаційної безпеки». ХНУРЕ. -2005. –С. 105.

Приложение А. Руководство исследователя к тесту NIST STS

Средством тестирования NIST STS следует пользоваться из командной строки. Для этого необходимо произвести запуск исполняемого файла NIST_STS.exe, и длины исследуемой последовательности. Например необходимо исследовать файл размером 12 Мб (12 582 912 байт), что составит 100 663 296 бит (такая длина позволяет говорить о 100 последовательностях длиной 1 000 000 бит, допускаются вариации на тему разложения числа 100 000 000 на два множителя: количество двоичных последовательностей и их двоичной длины), см. рис. 2.

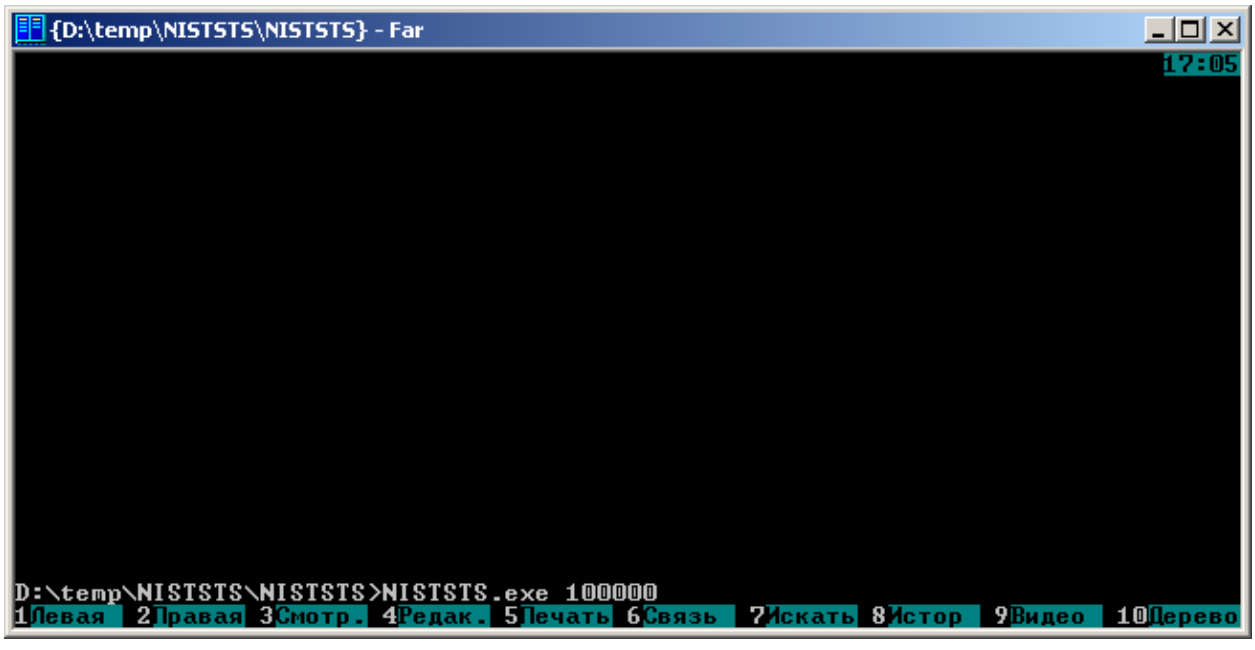

Рис. 2. Запуск приложения NISTSTS.exe

После запуска приложения необходимо указать, какие данные необходимо тестировать, «OPTION---->»:

- Тестовые данные будут сформированы генератором, реализованным в тесте NISTSTS.exe. Для этого необходимо ввести значение от 1 до 9. Например вводим 9 для выбора генератора основанного на функции хеширования SHA-1 («G Using SHA-1»). После чего будет сформирована последовательность длиной 12 Мб.
- Тестовые данные находятся в файле, который содержит последовательность, которую необходимо протестировать. Для этого необходимо указать значение 0. В строке «User Prescribed Input File: » вводится имя файла, например «output.bin», Рис. 3.

После чего будет предоставлен перечень статистических тестов, которые могут быть применены к тестируемой последовательности. Если исследователь не желает применить существующие тесты к исследуемой последовательности, ему следует ввести в строке «Enter Choise: » цифру 0, в противном случае 1, рис. 4 и нажать клавишу «Enter».

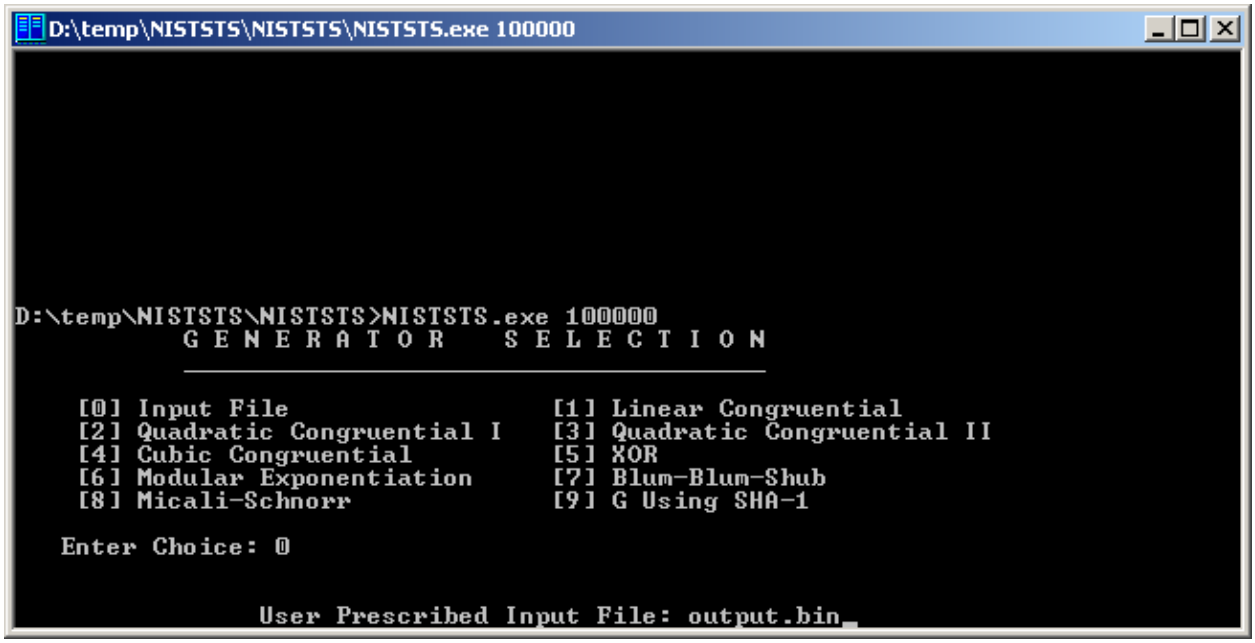

Рис. 3. Выбор тестируемой последовательности

Далее программа предложит выбрать, какие именно тесты следует применить к последовательности. Все присутствующие тесты пронумерованы от 01 до 15. Исследователю следует ввести в нижней строке под номером соответствующего теста цифру 1, если тест следует применить, в противном случае – цифру 0, рис. 4.

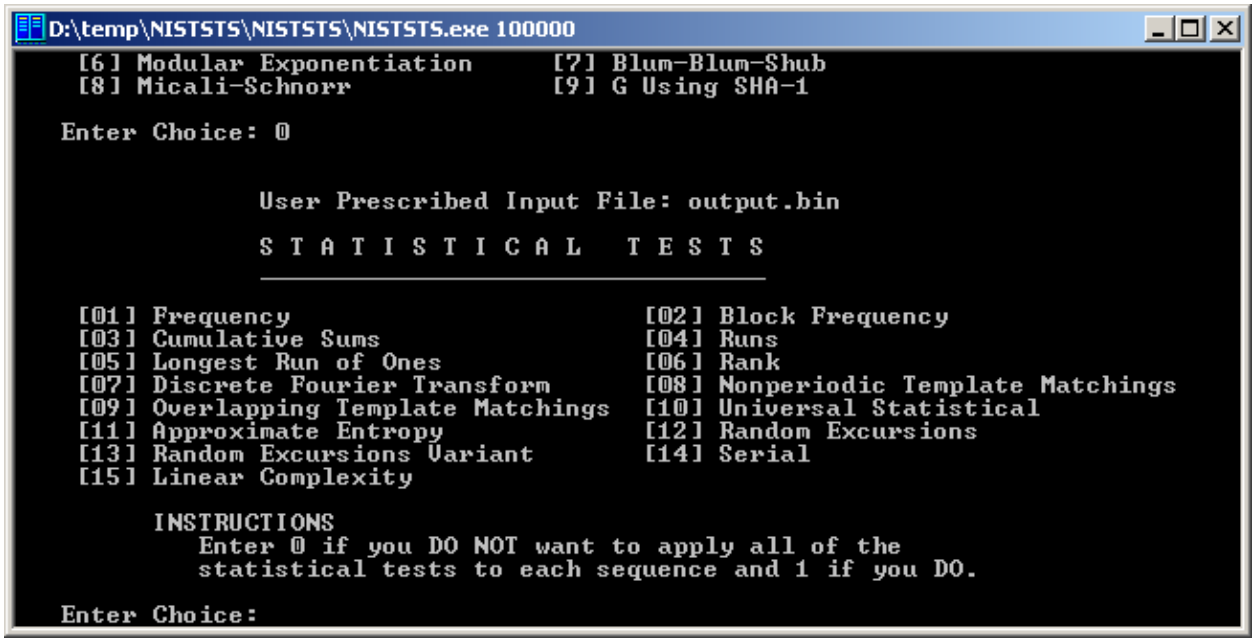

Рис. 4. Выбор единственного статистического теста

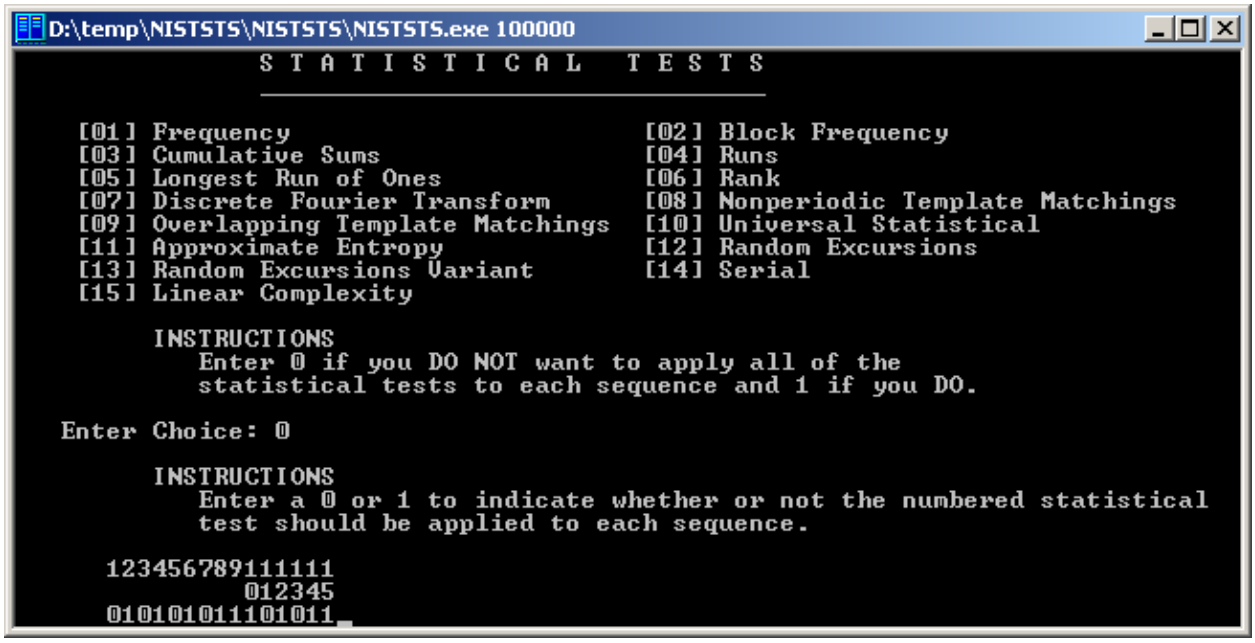

Рис. 5. Выбор множества статистических тестов, которые будут последовательно применяться к исследуемой последовательности

После выбора необходимых тестов, приложение предложит ввести параметры, согласно которым будет производиться тестирование последовательности.

Обратим внимание, что существуют параметризированные и непараметризированные тесты. К непараметризируемым тестам следует отнести:

- Cumulative Sums
- Runs
- Longest Runs of Ones
- Rank
- Spectral DFT
- Random Excursions Variant
- Lempel Ziv Complexity
- Frequency

Для указанных тестов необходимо указать лишь длину последовательности и их количество. К параметризованным тестам следует отнести (укажем также их параметры):

- Block Frequency – длина блока, по умолчанию – 128 бит.
- NonOverlapping Template – длина блока, по умолчанию – 9 бит.
- Overlapping Template – длина блока, по умолчанию – 9 бит.
- Approximate entropy – длина блока, по умолчанию – 10 бит.
- Serial – длина блока, по умолчанию – 16 бит.
- Linear Complexity – длина блока, по умолчанию – 500 бит.

Кроме того, следует указать в каком формате будут представлены данные в файле:

- «текстовый» – один символ – один бит;
- «двоичный» - один байт содержит 8 бит последовательности.

В строке «How many bitstreams should be generated?» исследователю следует ввести количество тестируемых последовательностей. NIST STS рекомендует число 100 в качестве количества подпоследовательностей.

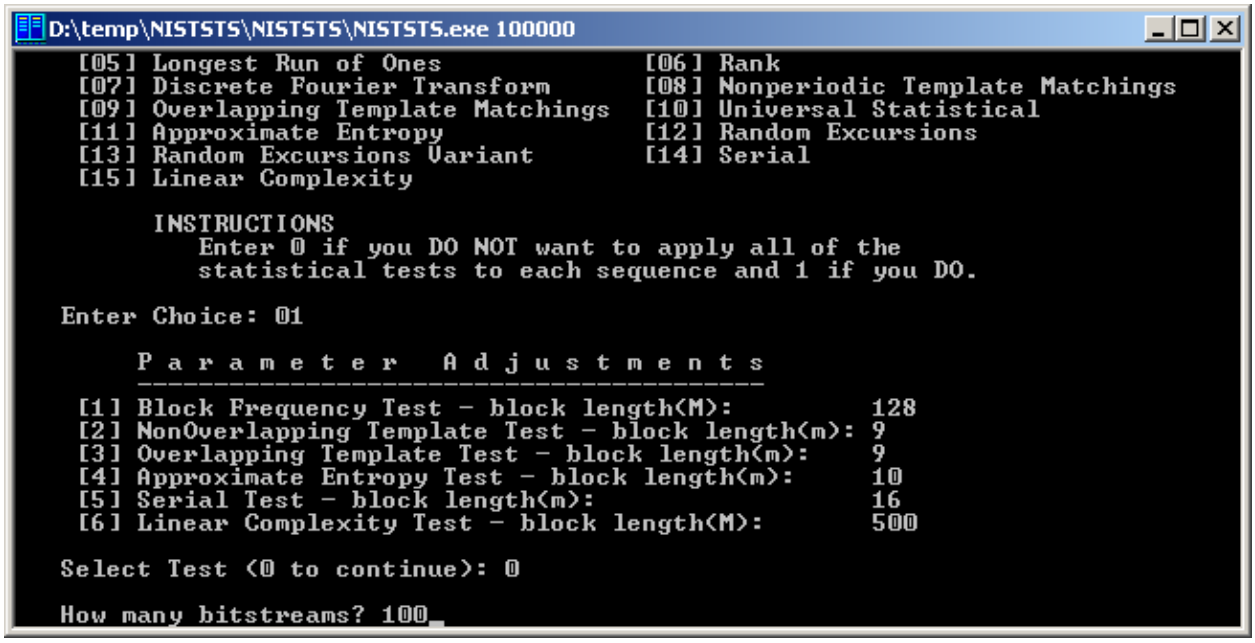

Рис. 6. Настройка параметров Частотного теста, например

В строке «Select input mode:» исследователь указывает тип тестируемой подпоследовательности: в случае если последовательность представлена в ASCII формате, следует ввести 0, если в двоичном, то 1, рис. 7.

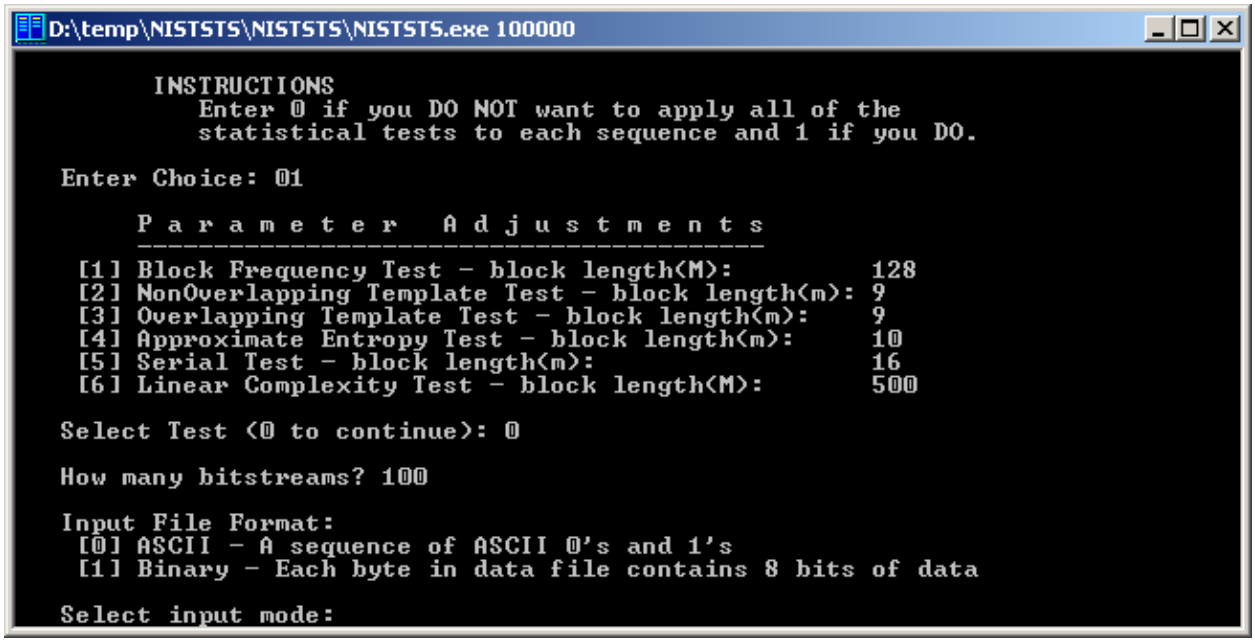

Рис. 7. Выбор формата представления последовательности

После ввода типу последовательности приложение выведет сообщение «Statistical Testing In Progress…….», рис. 8.

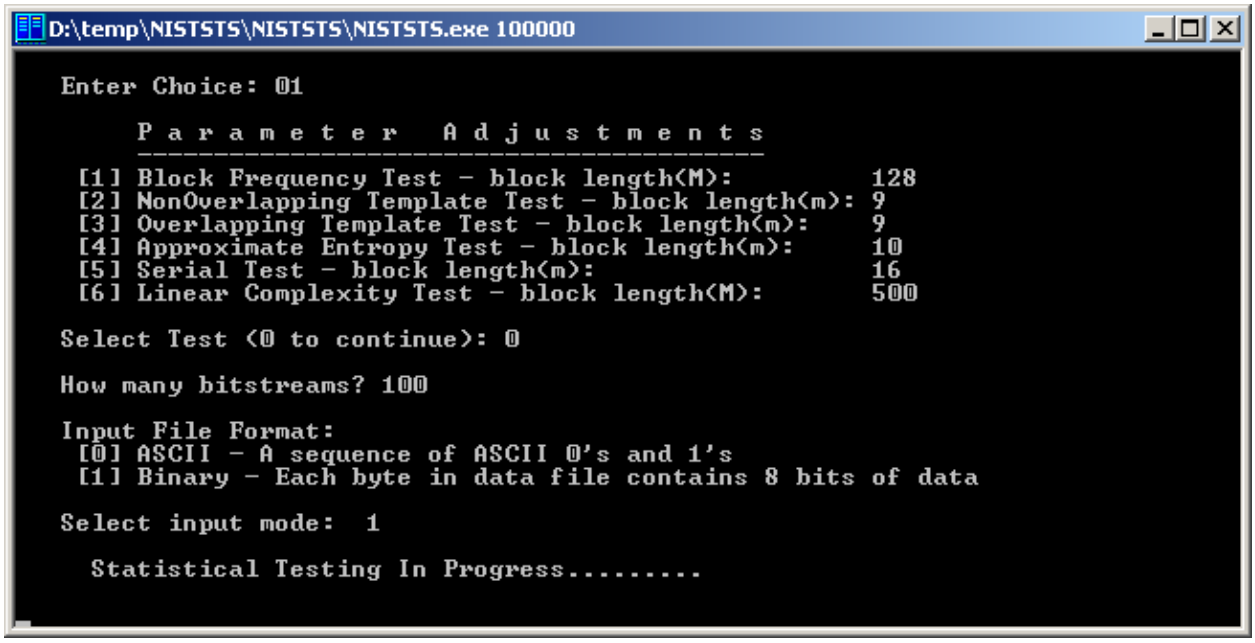

Рис. 8. Отображение статуса теста

После завершения работы приложения, все суммарные расчетные данные размещаются в том же каталоге, где находится само приложение, в файле finalAnalysisReport, рис. 9 и рис. 10.

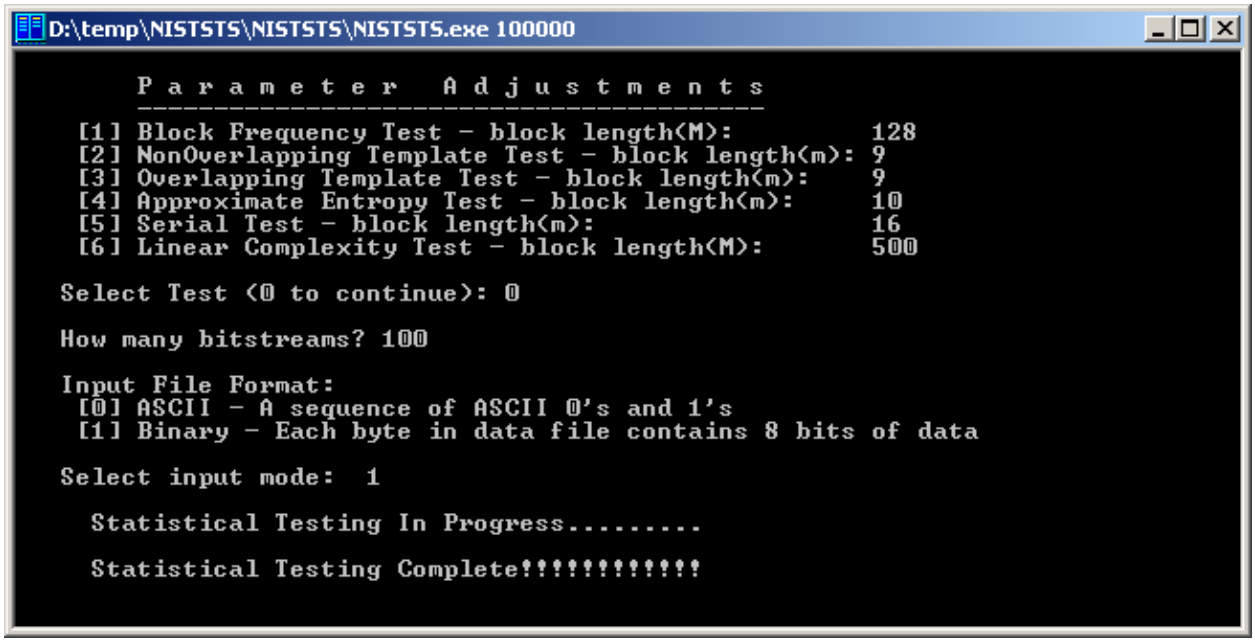

Рис. 9. Сообщение о завершении теста

За более детальной информацией (промежуточной) следует обращаться в папку experiments, в которой перечислены папки (с названиями соответствующих тестов), будут находиться 2 файла stats и results.

Файл stats содержит статистическую информацию по каждому тесту, а также формализованный результат: «Прошел» либо «Не прошел».

Файл results содержит лишь значения $\mathbf{P}$-вероятности, которая также указывается в файле stats.

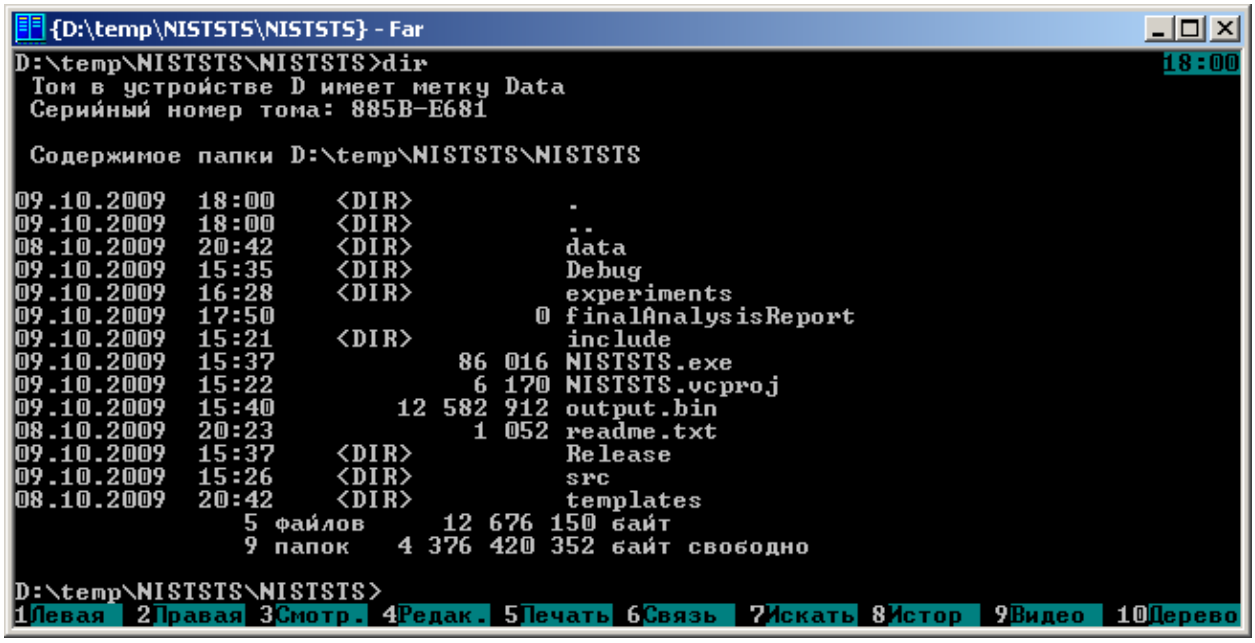

Рис. 10. Информация о результатах тестирования

Приложение Б. Пример интерпретации результатов теста генератора

Випробування здійснювалися у спеціальній лабораторії ЗАТ «Інститут інформаційних технологій» відповідно до методики тестування NIST SP 800-22. Випробування здійснювалися представниками на випробувальному стенді.

Інтерпретація результатів перевірки програмної реалізації алгоритму блокового шифрування FIPS 197 у режимі лічильника

Програмна реалізації алгоритму блокового шифрування FIPS 197 у режимі лічильника була піддана статистичному тестуванню з використанням методики NIST STS, що рекомендована Національним інститутом зі стандартизації і технологій США - NIST SP 800-22.

З використанням методики NIST STS було здійснене тестування програмної реалізації алгоритму блокового шифрування FIPS 197 у режимі лічильника, а також, з метою порівняльного аналізу, генератора псевдовипадкових чисел BBS (тестова вибірка, рекомендована NIST).

Для здійснення тестувань були обрані такі параметри:

- довжина послідовності, що тестується $n = 10^6$ біт;
- кількість послідовностей, що тестується $m = 100$;
- рівень значущості $\alpha = 0{,}01$.

Таким чином,

- обсяг вибірки, що тестується, склав $N = 10^6 \times 100 = 10^8$ біт;
- кількість тестів (q) для різних довжин $q = 189$, таким чином, статистичний портрет генератора містить 18900 значень імовірності P.

В ідеальному випадку при $m = 100$ і $\alpha = 0{,}01$ у ході тестування може бути відкинута тільки одна послідовність зі ста, тобто коефіцієнт проходження кожного тесту має складати 99%. Але це занадто жорстке правило. Тому застосовується правило на основі довірчого інтервалу для r_j. Нижня межа в цьому випадку складе значення $r_{min} = 0{,}96015$. З цих позицій аналізуються результати тестування генераторів.

У таблиці Б.1 наведено результати проходження тестування програмної реалізації алгоритму блокового шифрування FIPS 197 у режимі лічильника за Правилом 1.

Всі тести пройшли (або вказати який тест не пройшов, та у відповідності до керівництва NIST STS надати інтерпретацію недоліку, що виявляється цім тестом)

У таблиці Б.2 наведено результати проходження тестування програмної реалізації алгоритму блокового шифрування FIPS 197 у режимі лічильника за Правилом 2.

Таблиця Б.1– Результати тестування алгоритму FIPS 197 за правилом 1

Генератор	Кількість тестів, у яких тестування	Кількість тестів, у яких

	пройшли більш 99% послідовностей	тестування пройшли більш 96% послідовностей
BBS	134 (71%)	189 (100%)
FIPS 197	126 (67%)	189 (100%)

Таблиця Б.2 – Результати тестування алгоритму FIPS 197 за правилом 2

Генератор	Кількість тестів, у яких значення *P < 0,001*	Кількість тестів, у яких значення *P < 0,01*
BBS	134 (71%)	189 (100%)
FIPS 197	126 (67%)	189 (100%)

(Якщо необхідно, вказати за якими тестами значення **P** було нижче визначеної межі, вказати, які недоліки виявляють ці тести, та порівняти ці результати з результатами тестування за Правилом 1).

Таким чином, можна зробити висновок, що програмна реалізація алгоритму блокового шифрування FIPS 197 у режимі лічильника пройшла комплексний контроль за методикою NIST STS.

Результати тестування

Таблиця 5. Результати тестування

Гістограма значень імовірності **P**											Значення імовірності P	Доля проходження тесту	Статистичний тест
Номер тесту	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..1,0			
1	10	10	15	12	11	6	6	9	14	7	0,455937	1,0000	Frequency
2	9	6	11	13	16	8	8	11	10	8	0,574903	0,9800	Block-Frequency
3	10	9	11	10	12	13	16	13	3	3	0,071177	0,9900	Cusum
4	10	13	11	11	12	9	9	13	8	4	0,678686	1,0000	Cusum
5	11	10	19	10	9	6	10	11	8	6	0,213309	0,9900	Runs
6	7	9	9	8	8	12	16	11	7	13	0,55442	1,0000	Long-Run
7	8	16	14	8	12	6	7	10	10	9	0,437274	0,9800	Rank
8	10	11	11	8	7	14	8	12	12	7	0,816537	0,9900	FFT
9	13	10	9	10	11	8	11	6	12	10	0,935716	0,9800	Periodic-Template
10	13	12	11	13	10	13	7	9	3	9	0,419021	0,9700	Universal
11	5	8	8	9	8	16	19	11	9	7	0,055361	1,0000	Apen
12	6	3	6	9	10	5	7	0	9	4	0,028817	0,9661	Random-Excursion
13	6	5	6	4	7	6	8	7	1	9	0,304126	0,9831	Random-Excursion
14	4	8	5	5	7	3	7	7	5	8	0,637119	0,9831	Random-Excursion
15	4	3	5	4	8	8	4	5	10	8	0,224821	1,0000	Random-Excursion
16	10	7	3	6	4	3	6	8	7	5	0,304126	1,0000	Random-Excursion
17	8	4	4	6	6	6	5	6	6	8	0,834308	0,9661	Random-Excursion
18	6	6	5	6	3	6	3	6	8	10	0,401199	0,9831	Random-Excursion
19	5	6	4	6	9	8	9	5	4	3	0,366918	0,9661	Random-Excursion
20	4	2	8	8	6	8	3	7	7	6	0,334538	1,0000	Random-Excursion-V
…	…	…	…	…	…	…	…	…	…	…	…	…	…
37	5	9	4	2	4	5	5	7	8	10	0,162606	1,0000	Random-Excursion-V
38	9	9	11	11	8	7	15	13	13	4	0,383827	1,0000	Serial
39	13	10	8	5	13	11	9	10	16	5	0,275709	1,0000	Serial
40	8	11	13	9	9	7	12	7	14	10	0,798139	1,0000	Lempel-Ziv
41	6	10	7	9	6	15	18	8	10	11	0,137282	1,0000	Aperiodic-Template
…	…	…	…	…	…	…	…	…	…	…	…	…	…
188	5	3	14	10	10	12	11	10	14	11	0,262249	1,0000	Aperiodic-Template
189	12	12	5	3	9	12	7	9	19	12	0,032923	0,9800	Linear-Complexity

Припустиме значення долі проходження тесту для вибірки розміром 100 двійкових послідовностей дорівнює	0,96015
Припустиме значення долі проходження тесту для вибірки розміром 71 двійкових послідовностей для тесту Random-Excursion дорівнює	0,954575
Кількість тестів, в яких тестування пройшло 99% послідовностей	126
Кількість тестів, в яких тестування пройшло 96% послідовностей	189
Кількість тестів, в яких значення імовірності P <=0,01	0
Кількість тестів, в яких значення імовірності P <=0,001	0
Кількість тестів, в яких значення імовірності P <=0,05	8

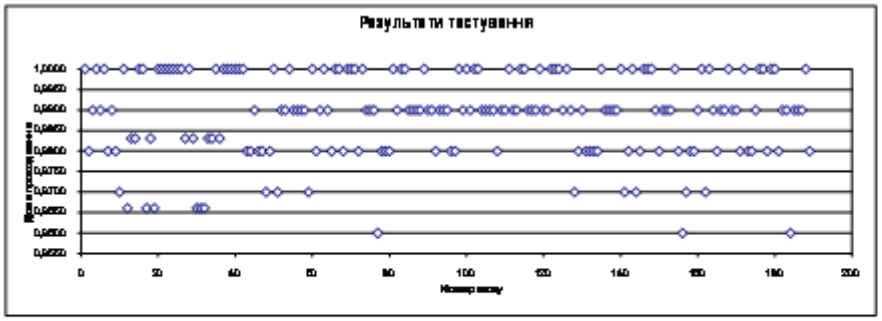

Рис. 11. Статистичний портрет програмної реалізації алгоритму блокового шифрування FIPS 197 у режимі лічильника

Приложение В. Критерии согласованности

При побудові ключових систем одним з основних завдань є одержання випадкових і псевдовипадкових послідовностей, які не можуть бути відрізнені від випадкових і мають великий період. Після генерації послідовності чисел $X = \{x_1, x_2, \ldots, x_n\}$ необхідно впевнитись у тому, що випадкова величина X має рівномірний закон розподілу, її реалізації випадкові та незалежні. Методи математичної статистики дають нам можливість побудувати статистичні тести для перевірки гіпотез про рівномірність, випадковість і незалежність випадкових величин.

Для перевірки гіпотези про закон розподілу скористаємося критеріями χ^2 Пірсона та критерієм Колмогорова-Смірнова.

Критерій Пірсона χ^2

Критерій Пірсона χ^2. перевіряє узгодженість гіпотетичних ймовірностей $P_k = P(x_k)$ випадкових величин $x_1, x_2, \ldots, x_n$ з їхніми відносними частотами $h_k = v_k / n$ у вибірці з n незалежними спостереженнями. Статистика критерію має вигляд:

$$\chi^2 = n\sum_{k=1}^{m}\frac{(h_k - P_k)^2}{P_k},$$

де m – кількість інтервалів розбивки. Граничне значення статистики для рівня значущості визначається за формулою:

$$\chi_\alpha^2 \approx l\left(1 - \frac{2}{9l} + z_\alpha\sqrt{\frac{2}{9l}}\right)^3,$$

де l – кількість ступенів волі; z_α - граничне значення стандартного нормального розподілу.

Гіпотеза про узгодженість емпіричного закону розподілу спостережуваної випадкової величини з теоретичним відкидається, якщо спостережене значення $\chi^2 > \chi_\alpha^2$.

Критерій Колмогорова-Смірнова

Критерій Колмогорова-Смірнова заснований на розподілі величини

$D_n = \max|F_n(x) - F(x)|$,

де $F_n(x)$ – емпіричний закон розподілу; $F(x)$ – гіпотетичний закон розподілу. Відомо, що яка б не була безперервна функція розподілу $F(x)$, імовірність

$$P\left\{D_n < \frac{\lambda}{\sqrt{n}}\right\}$$

при $n \to \infty$ прагне до межі: $K(\lambda) = 1 - 2\sum_{k=1}^{\infty}(-1)^{k-1} e^{-2\lambda^2 k^2}$.

Таблиці В.1– Деякі дані для розподілення χ^2

	p=99%	p=95%	p=75%	p=50%	p=25%	p=5%	p=1%
ν =1	0,00016	0,00393	0,1015	0,4549	1,323	3,841	6,635
ν =2	0,00201	0,1026	0,5753	1,386	2,773	5,991	9,210
ν =3	0,1148	0,3518	1,213	2,366	4,108	7,815	11,34
ν =4	0,2971	0,7107	1,923	3,357	5,385	9,488	13,28
ν =5	0,5543	1,1455	2,675	4,351	6,626	11,07	15,09
ν =6	0,8720	1,635	3,455	5,348	7,841	12,59	16,81
ν =7	1,239	2,167	4,225	6,346	9,037	14,07	18,48
ν =8	1,646	2,733	5,071	7,344	10,22	15,51	20,09
ν =9	2,088	3,325	5,889	8,343	11,39	16,92	21,67
ν =10	2,558	3,940	6,737	9,342	12,55	18,31	23,21
ν =11	3,053	4,575	7,584	10,34	13,70	19,68	24,73
ν =12	3,571	5,226	8,438	11,34	14,84	21,03	26,22
ν =15	5,229	7,261	11,04	14,34	18,25	25,00	30,58
ν =20	8,260	10,85	15,45	19,34	23,85	31,41	37,57
ν =30	14,95	18,49	24,48	29,34	34,80	43,77	50,89
ν =50	29,71	34,76	42,94	49,33	56,33	67,50	76,15
ν >30				Приблизно $\nu + 2\sqrt{\nu \cdot x}_p + \frac{4}{3}x_p^2 - \frac{2}{3}$			
x_p	−2,33	−1,64	−0,675	0,00	0,675	1,64	2,33

Для практичних обчислень слід застосовувати такі формули:

$$D_n^+ = \max_{1 \le m \le n}\left(\frac{m}{n} - F(\xi_m)\right),$$

$$D_n^- = \max_{1 \le m \le n}\left(F(\xi_m) - \frac{m-1}{n}\right),$$

$$D_n = \max\left(D_n^+, D_n^-\right),$$

де $\xi_1 \le \xi_2 \le \ldots \le \xi_n$ – упорядковані значення випадкової величини.

Статистика D_n має певний розподіл (розподіл Колмогорова), що має табульований для деяких значень n. При $n \ge 10$ для визначення граничного значення $D_n(\alpha)$ на відрізку $0{,}01 \le \alpha \le 0{,}2$ потрібно користуватися формулою

$$D_n(\alpha) = \sqrt{\frac{1}{2n}(y) - \frac{2y^2 - 4y - 1}{18n}} - \frac{1}{6n} \approx \sqrt{\frac{y}{2n}} - \frac{1}{6n},\ y = -\ln(0{,}5\alpha).$$

При $n \ge 100$ зазначена формула правильна для всіх $0{,}0001 \le \alpha \le 0{,}5$.

Якщо в результаті досвіду виявиться, що $D_n \ge D_n(\alpha)$ то гіпотезу про узгодженість емпіричного та гіпотетичного законів розподілу слід відкинути з рівнем значущості α .

Таблиця В.2– Деякі дані для обчислення критерію Колмогорова-Смірнова

β	$P(\beta)$	β	$P(\beta)$	β	$P(\beta)$	β	$P(\beta)$	β	$P(\beta)$	β	$P(\beta)$
0	1,0	0,3	1,0	0,6	0,654	0,9	0,393	1,2	0,112	1,5	0,022
0,1	1,0	0,4	0,997	0,7	0,711	1,0	0,270	1,3	0,068	1,6	0,012
0,2	1,0	0,5	0,994	0,8	0,544	1,1	0,179	1,4	0,040	1,7	0,006