МИНОБРНАУКИ РОССИИ

Федеральное государственное бюджетное образовательное учреждение высшего образования

«Российский технологический университет»

МИРЭА

Институт кибернетики

Кафедра информационной безопасности

ДОКЛАД

по дисциплине

«Криптографические протоколы»

На тему:

«Статистические тесты проверки качества случайных последовательностей»

Подготовил

студент группы ККСО–01–14 А.С. Першин

Руководитель работы

А.П. Никитин

Москва, 2019

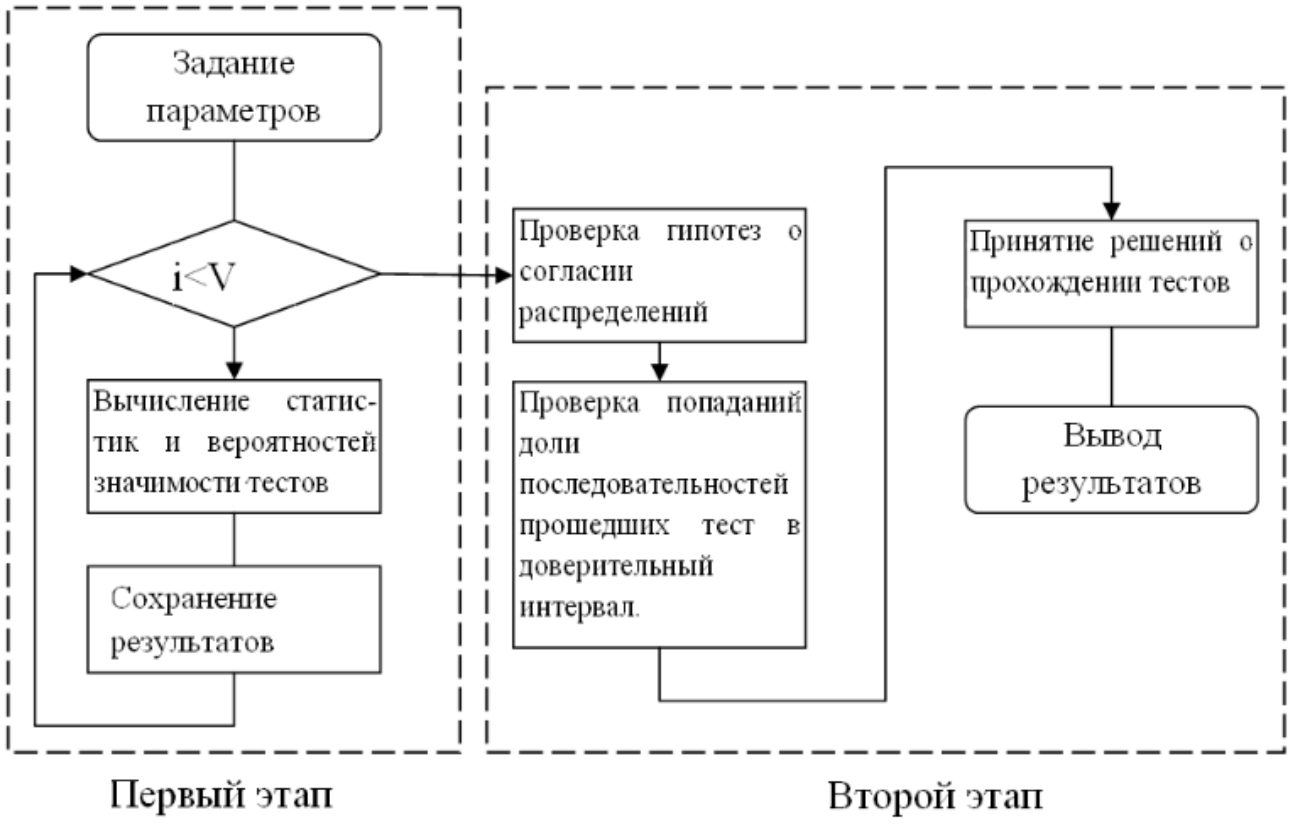
Оглавление

1. Общая схема

Статистический анализ последовательностей, как правило, проходит в два этапа.

1. Первый этап можно назвать подготовительным, он самый трудоемкий, здесь выполняется основная масса вычислений.
   1.1. При помощи исследуемого генератора формируются случайные последовательности.
   1.2. Для каждой последовательности вычисляется статистика теста. Если работает батарея тестов (проводится сразу несколько тестов), то статистика по последовательности вычисляется для каждого теста.
   1.3. Для каждой последовательности, вычисляется вероятность значимости.
   1.4. Полученные статистики и вероятности значимости сохраняются.
2. На втором этапе проводится обработка, полученных результатов.
   2.1. При помощи критериев согласия проверяются гипотезы о соответствии распределений статистик и вероятностей значимости гипотетическим распределениям.
   2.2. Определяется, число последовательностей, прошедших тест. Строится доверительный интервал для последней величины.
   2.3. Принимается решение о том, пройден ли тест.
   2.4. Окончательные выводы.

Схематично этот процесс можно представить примерно так:



2. Формирование статистики

Целью каждого теста является проверка гипотезы о том, что исследуемая последовательность имеет равномерное распределение. Если говорить более строго, то каждый знак случайной последовательности $X = \{\varepsilon_1, \varepsilon_2, ..., \varepsilon_n\}$ должен иметь равномерное распределение вероятностей: $P(\varepsilon = x) = \frac{1}{2}, x \in \{0,1\}$. Эта гипотеза обозначается через H_0 (говорят, нулевая гипотеза).

А от куда появляются формулы для статистик тестов? Давайте рассмотрим это на примере частотного теста из пакета NIST.

Каждый статистический тест проверяет некоторое предположение о том или ином свойстве, которым должна обладать случайная последовательность, удовлетворяющая гипотезе H_0.

Для частотного теста из набора тестов NIST таким предположением является утверждение: «Если последовательность удовлетворяет гипотезе H_0, то количество нулей и единиц в этой последовательности должно быть примерно одинаковым».

Если найти сумму знаков исследуемой последовательности, то итоговый результат будет представлять случайную величину, назовем ее ζ: $\zeta = \sum_{i=1}^{n} \varepsilon_i$. Её распределение будет выглядеть следующим образом: $\zeta = \begin{pmatrix} \frac{n}{2} & n \\ 1-\delta & \delta \end{pmatrix}$, где ζ имеет значение $x_1 = \frac{n}{2}$ с вероятностью $p_1 = 1-\delta$, и $x_2 = n$ с вероятностью $p_2 = \delta$.

Её математическое ожидание будет выглядеть следующим образом:

$$M_\zeta = \sum_{t=1}^{2} p_t x_t = \frac{n}{2}(1-\delta) + n\delta = \frac{n}{2} - \frac{n}{2}\delta + n\delta = \frac{n}{2} + \frac{n}{2}\delta$$

А дисперсия, при $\zeta^2 = \begin{pmatrix} \frac{n^2}{4} & n^2 \\ 1-\delta & \delta \end{pmatrix}$:

$$D_\zeta = \left(\sum_{t=1}^{2} p_t (x_t)^2\right) - \left(\sum_{t=1}^{2} p_t x_t\right)^2 = \left(\frac{n^2}{4}(1-\delta) + n^2\delta\right) - \left(\frac{n^2}{4} + \frac{n^2}{2}\delta + \frac{n^2\delta^2}{4}\right) = \frac{n^2}{4} - \frac{n^2\delta}{4} + n^2\delta - \frac{n^2}{4} -$$
$$- \frac{n^2\delta}{2} + \frac{n^2\delta^2}{4} = -\frac{3n^2\delta}{4} + n^2\delta + \frac{n^2\delta^2}{4} = \frac{n^2\delta}{4} + \frac{n^2\delta^2}{4}$$

Тогда её среднеквадратичным отклонение будет равно:

$$\sigma = \sqrt{D_\zeta} = \sqrt{\frac{n^2\delta}{4} + \frac{n^2\delta^2}{4}}.$$

Согласно, центральной предельной теореме случайная величина ζ в наших предположениях должна иметь распределение близкое к нормальному распределению с математическим ожиданием $M_\zeta = \lim\limits_{\substack{n\to+\infty \\ \delta\to+0}} \left(\frac{n}{2}+\frac{n}{2}\delta\right) = \frac{n}{2}$, дисперсией $D_\zeta = \lim\limits_{\substack{n\to+\infty \\ \delta\to+0}} \left(\frac{n^2\delta}{4}+\frac{n^2\delta^2}{4}\right) = \frac{n}{4}$ среднеквадратичным отклонением $\sigma = \sqrt{D_\zeta} = \frac{\sqrt{n}}{2}$.

Получили нормальное распределение N_{a,δ^2}.

Для того, чтобы воспользоваться теоретическими вероятностями, масштаб реальных данных нужно «подогнать» под эталон стандартного нормального распределения $N_{0,1}$. Делается это довольно просто с помощью процедуры нормирования:

$$S = \frac{\left|\zeta - M_\zeta\right|}{\sigma} = \frac{\left|\sum\limits_{i=1}^{n}\varepsilon_i - \frac{n}{2}\right|}{\frac{\sqrt{n}}{2}} = \frac{\left|2\sum\limits_{i=1}^{n}\varepsilon_i - n\right|}{\sqrt{n}}$$

Математическое ожидание и дисперсия новой переменной *S* теперь также равны 0 и 1 соответственно. Такую *оценку* можно напрямую сравнивать с теоретическими вероятностями, т.к. ее масштаб совпадает с эталоном.

Получилась случайная величина *S*, зависящая от исследуемой последовательности. Функция распределения этой случайной величины в предположении гипотезы H_0 равна:

$$F(x) = \begin{cases} P(S \le x) = 2\Phi(x) - 1 & \text{при } x > 0 \\ 0 & \text{при остальных } x \end{cases}$$

Где $\Phi(x)$ - функция стандартного нормального распределения. Значения которой можно получить из таблицы, а $2\Phi(x)-1$ описывает вероятность отклонения в обе стороны от нуля (число битов нулей (или единиц) больше эталона или наоборот меньше).

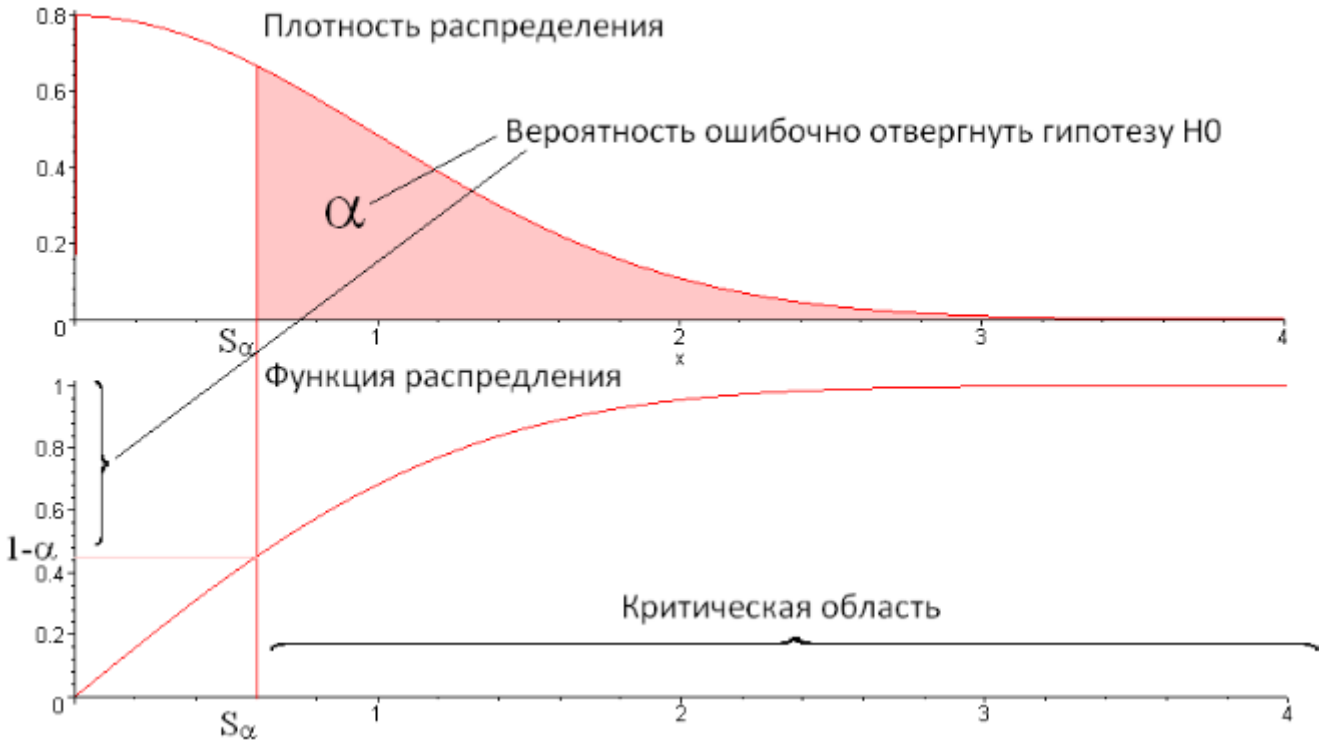
Величину *S*, вычисленную по конкретной последовательности *X* называют «статистикой». Если заглянуть в описание тестов NIST, то можно найти именно эту формулу для вычисления статистики теста.

3. Определение критерия принятия решения

Нужно определить значимо ли отклоняется статистика *S* от нуля или нет.

Отклонение не должно быть «слишком» большим. Слово «слишком», не дает конкретных указаний к действию. Поэтому выберем некоторый критический уровень α - вероятность ошибочно отвергнуть гипотезу H_0.

Данному α соответствует критическое значение S_α, которое определяет критическую область. Продемонстрируем это на картинках:



Иными словами S_α представляет собой границу для принятия решения, прошла ли последовательность тест или нет. Если значение $S < S_\alpha$, то считаем, что последовательность «хорошая» — прошла тест. Если S попало в критическую область $S \geq S_\alpha$, то считаем последовательность плохая — не прошла тест.

В пакетах статистических исследований чаще применяется другой эквивалентный метод принятия решения. По статистике S вычисляют вероятность значимости p, которая равна: $p = 1 - F(S)$. Данное значение будет меньше или равно α ($p \leq \alpha$), только в случае попадания S в критическую область.

Обычно α выбирают в диапазоне от 0.01 до 0.001. Чем больше α, тем больше последовательностей будут отвергаться.

4. Дальнейшие выводы

Напрашивается вопрос: «Как отбраковать генератор (псевдо)случайных чисел, если часть последовательностей плохая, а часть хорошая?»

На самом деле любой генератор случайных последовательностей всегда будет генерировать часть последовательностей, которые не пройдут тест, если все последовательности пройдут тест, то это очень подозрительно.

На первом этапе генерируется V последовательностей длины n – $X_1, X_2, ..., X_V$. Для каждой последовательности X_i вычисляется статистика S_i и вероятность

значимости p_i, т.е. получаются два набора: набор статистик $S_1, S_2, ..., S_V$ и набор вероятностей значимости $p_1, p_2, ..., p_V$.

На втором этапе производится обработка полученных выборок $S_1, S_2, ..., S_V$ и $p_1, p_2, ..., p_V$.

Сначала к этим последовательностям применяются критерии согласия. Выборка $S_1, S_2, ..., S_V$ должна иметь распределение, описанное выше, а выборка $p_1, p_2, ..., p_V$ должна иметь равномерное распределение на отрезке $p_i \in [0,1]$.

Для первой последовательности $S_1, S_2, ..., S_V$ например, применяется критерий согласия Колмогорова-Смиронова, а для второй $p_1, p_2, ..., p_V$ критерий согласия Пирсона.

5. Применение критерия Пирсона

Для $P = \{p_1, p_2, ..., p_V\}$, где $|P| = V$ критерий Пирсона применяется так:

Область возможных значений *P* (отрезок [0,1]) делится на *T* одинаковых отрезков. Значение *T* обычно выбирается не очень большим, например, 10 или 20. В предположении, что *P* имеет равномерное распределение, в среднем в каждый интервал должно попадать V/T значений (кстати, T надо выбрать так, чтобы V/T>5).

По выборке *P* формируется гистограмма $\{F_0, F_1, ..., F_{T-1}\}$ – количество значений, попавших в каждый интервал.

Далее вычисляется статистика Пирсона:

$$\chi^2 = \sum_{i=0}^{T-1} \frac{\left(F_i - \frac{V}{T}\right)^2}{\frac{V}{T}}$$

Указанная статистика в предположении, что исходная последовательность распределена равномерно на отрезке [0,1], должна иметь распределение *Хи-квадрат* с *T-1* степенью свободы (пусть $F_{T-1}(x)$ — функция распределения *Хи-квадрат* с *T-1* степенью свободы). Значение функции $F_{T-1}(x)$ вычисляют программно или по специальным таблицам (при ручном применении критерия).

Для статистики χ^2 вычисляется вероятность значимости $p = 1 - F_{T-1}(\chi^2)$. Если $p > \alpha$, то гипотеза о равномерном распределении не отвергается.

Далее исследуется количество последовательностей, прошедших тест.

Так как вероятностей значимости распределены равномерно на отрезке [0,1], то в среднем тест должно пройти $V(1-\alpha)$ последовательностей. Но сравнивать

полученное число с $V(1-\alpha)$ нет смысла. Вместо этого строится доверительный интервал.

Последовательности $\{p_1, p_2, ..., p_V\}$ соответствует последовательность V независимых испытаний Бернулли $\{\xi_1, \xi_2, ..., \xi_V\}$:

$$\xi_i = 1, p_i \geq \alpha$$
$$\xi_i = 0, p_i < \alpha$$

$$\xi = \sum_{i=0}^{V} \xi_i$$

Тогда случайная величина $\xi = \begin{pmatrix} V & 0 \\ 1-\alpha & \alpha \end{pmatrix}$ распределенная по закону Бернулли будет иметь:

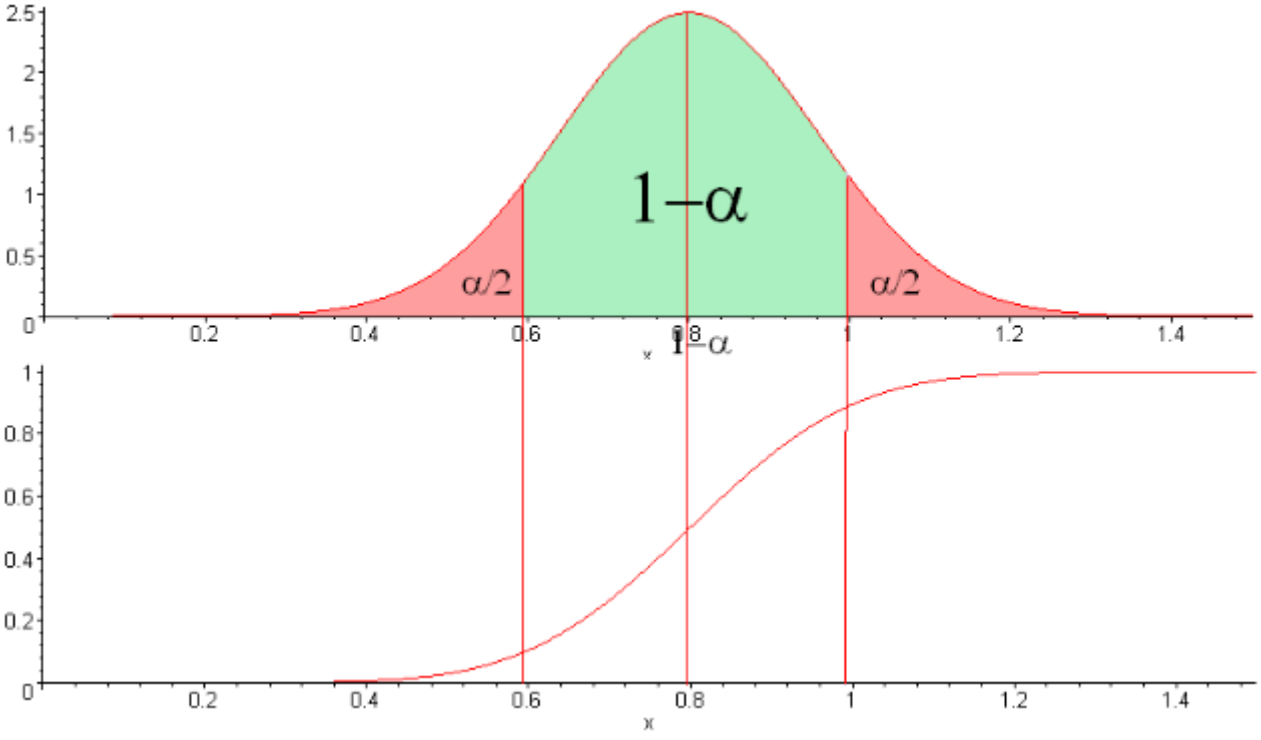
Математическое ожидание:

$$M_\xi = V(1-\alpha)$$

Дисперсию:

$$D_\xi = V(1-\alpha)\alpha$$

Тогда, согласно центральной предельной теореме, распределение числа успехов в последовательности испытаний Бернулли, совпадающее с числом последовательностей, прошедших тест, можно считать нормальным с математическим ожиданием $V(1-\alpha)$ и дисперсией $V(1-\alpha)\alpha$



Доверительным называется интервал, который с заданным уровнем доверия $1-\alpha$ покрывает оцениваемый параметр (в нашем случаем математическое ожидание доверительного интервала $M_{1-\alpha}$).

Если выбрать уровень доверия $(1-\alpha)$, то доверительный интервал его математического ожидания равен:

$$P\left(M_\xi + \Phi^{-1}\left(\frac{\alpha}{2}\right)\times\frac{\sigma}{\sqrt{n}} \le M_{1-\alpha} \le M_\xi - \Phi^{-1}\left(\frac{\alpha}{2}\right)\times\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha$$

или для нашего случая:

$$\left((1-\alpha)+\Phi^{-1}\left(\frac{\alpha}{2}\right)\sqrt{\frac{(1-\alpha)\alpha}{V}};(1-\alpha)-\Phi^{-1}\left(\frac{\alpha}{2}\right)\sqrt{\frac{(1-\alpha)\alpha}{V}}\right)$$

Где: $\Phi^{-1}\left(\frac{\alpha}{2}\right)=t \Leftrightarrow \Phi(t)=\frac{\alpha}{2}$, которое можно найти по таблице стандартного нормального распределения.

На рисунке доверительный интервал – значения *x* под зеленой областью.

В предположении, что исходные последовательности имеют равномерное распределение, доля последовательностей прошедших тест с вероятностью $(1-\alpha)$ должна попасть в этот интервал.

6. Принцип работы статистических тестов

Выше рассмотрено описание тестирования по одному тесту. Если батарея содержит несколько тестов, то описанные исследования проводятся для каждого теста.

На выходе выдается табличка, какие тесты пройдены, и процент пройденных тестов.

7. Базовые тесты в пакете Dieharder

Тесты diehard — это набор статистических тестов для измерения качества набора случайных чисел. Они были разработаны Джорджем Марсальей в течение нескольких лет и впервые опубликованы на CD-ROM, посвящённом случайным числам. Вместе они рассматриваются как один из наиболее строгих существующих наборов тестов (отсюда и название — англ. «die-hard» в качестве прилагательного означает приблизительно «трудноубиваемый» и обычно переводится на русский фразеологизмом «крепкий орешек»).

- **Дни рождения** (Birthday Spacings) — выбираются случайные точки на большом интервале. Расстояния между точками должны быть

асимптотически распределены по Пуассону. Название этот тест получил на основе парадокса дней рождения.

- **Пересекающиеся перестановки** (Overlapping Permutations) — анализируются последовательности пяти последовательных случайных чисел. 120 возможных перестановок должны получаться со статистически эквивалентной вероятностью.

- **Ранги матриц** (Ranks of matrices) — выбираются некоторое количество бит из некоторого количества случайных чисел для формирования матрицы над {0,1}, затем определяется ранг матрицы. Считаются ранги.

- **Обезьяньи тесты** (Monkey Tests) — последовательности некоторого количества бит интерпретируются как слова. Считаются пересекающиеся слова в потоке. Количество «слов», которые не появляются, должны удовлетворять известному распределению. Название этот тест получил на основе теоремы о бесконечном количестве обезьян.

- **Подсчёт единичек** (Count the 1's) — считаются единичные биты в каждом из последующих или выбранных байт. Эти счётчики преобразуется в «буквы», и считаются случаи пятибуквенных «слов».

- **Тест на парковку** (Parking Lot Test) — единичные окружности случайно размещаются в квадрате 100×100. Если окружность пересекает уже существующую, попытаться ещё. После 12 000 попыток, количество успешно «припаркованных» окружностей должно быть нормально распределено.

- **Тест на минимальное расстояние** (Minimum Distance Test) — 8000 точек случайно размещаются в квадрате 10 000×10 000, затем находится минимальное расстояние между любыми парами. Квадрат этого расстояния должен быть экспоненциально распределён с некоторой медианой.

- **Тест случайных сфер** (Random Spheres Test) — случайно выбираются 4000 точек в кубе с ребром 1000. В каждой точке помещается сфера, чей радиус является минимальным расстоянием до другой точки. Минимальный объём сферы должен быть экспоненциально распределён с некоторой медианой.

- **Тест сжатия** (The Squeeze Test) — 2^{31} умножается на случайные вещественные числа в диапазоне [0,1) до тех пор, пока не получится 1. Повторяется 100 000 раз. Количество вещественных чисел необходимых для достижения 1 должно быть распределено определённым образом.

- **Тест пересекающихся сумм** (Overlapping Sums Test) — генерируется длинная последовательность вещественных чисел из интервала [0,1). В ней суммируются каждые 100 последовательных чисел. Суммы должны быть нормально распределены с характерными средним и дисперсией.

- **Тест последовательностей** (Runs Test) — генерируется длинная последовательность на [0,1). Подсчитываются восходящие и нисходящие последовательности. Числа должны удовлетворять некоторому распределению.

- **Тест игры в кости** (The Craps Test) — играется 200 000 игр в кости, подсчитываются победы и количество бросков в каждой игре. Каждое число должно удовлетворять некоторому распределению.

- **Статистические тесты NIST** — пакет статистических тестов, разработанный Лабораторией информационных технологий (англ. *Information Technology Laboratory*), являющейся главной исследовательской организацией Национального института стандартов и технологий (NIST). В его состав входят 15 статистических тестов, целью которых является определение меры случайности двоичных последовательностей, порождённых либо аппаратными, либо программными генераторами случайных чисел. Эти тесты основаны на различных статистических свойствах, присущих только случайным последовательностям.

Установка: на сайте http://webhome.phy.duke.edu/~rgb/General/rand_rate.php описана установка для ОС семейства Unix.

Для установки под ОС Windows необходимо установить Cygwin с пакетами:

- Gsl;
- Auoconf;
- Automake;
- Binutils;
- Cygport;
- Gcc-core;
- Libtools;
- Make.

Далее слинковать компилятор сборки с именем gcc:

```
$ ln -sf x86_64-pc-cygwin-gcc.exe gcc
```

Далее нужно собрать библиотеку:

```
$ ./configure
$ make
```

Библиотека готова к работе.

Запуск проверки сгенерированной последовательности осуществляется командой:

```
dieharder -g 201 -f testrands.txt -a
```

Где:

- -g 201 (формат тестируемых данных – полученный на выходе ГПСЧ файл ASCII с ПСП);

- -f (указывает путь к файлу testrands.txt);
- -a (выполнить проверку по всем тестам, которые есть в сборке библиотеки, посмотреть конкретные тесты можно флагом -l, запуск через флаг -d [номер теста])

Литература

1. Статья «Как устроены пакеты для проверки качества случайных последовательностей» Хабрахабр [Интернет ресурс], ссылка https://habr.com/ru/post/276535/
2. Ubuntu.com «Manuals of dieharder» [Интернет ресурс], ссылка u.com/manpages/bionic/man1/dieharder.1.html
3. Новосибирский государственный университет «Интервальное оценивание» [Интернет ресурс], ссылка https://nsu.ru/mmf/tvims/chernova/ms/lec/node31.html
4. Новосибирский государственный университет «Математические ожидания и дисперсии стандартных распределений» [Интернет ресурс], ссылка https://nsu.ru/mmf/tvims/chernova/tv/lec/node46.html
5. Статический анализ данных в MS Excel и R «Таблица нормального распределения» [Интернет ресурс], ссылка https://statanaliz.info/statistica/teoriya-veroyatnostej/tablitsa-normalnogo-raspredeleniya/
6. Статический анализ данных в MS Excel и R «Стандартное нормальное распределение» [Интернет ресурс], ссылка https://statanaliz.info/statistica/teoriya-veroyatnostej/tablitsa-normalnogo-raspredeleniya/
7. «Доверительный интервал» [Интернет ресурс], ссылка http://cito-web.yspu.org/link1/metod/theory/node40.html
8. Химфак МГУ им. М.В. Ломоносова «Лекция 2. Распределения и доверительные интервалы» [Интернет ресурс], ссылка http://td.chem.msu.ru/uploads/files/courses/general/statexp/Lecture02.pdf