44,00
Рейтинг

Код Безопасности
Компания

igorek_seccode 12 ноября 2014 в 12:09

Статистическая проверка случайности двоичных последовательностей методами NIST

Криптография, Блог компании Код Безопасности

Любой, кто, так или иначе, сталкивался с криптографией, знает, что без генераторов случайных чисел в этом деле не обойтись. Одно из возможных применений таких генераторов, например, – генерация ключей. Но не каждый при этом задумывается, а насколько «хорош» тот или иной генератор. А если и задумывался, то сталкивался с тем фактом, что в мире не существует какого-либо единственного «официального» набора критериев, который бы оценивал, насколько данные случайные числа применимы именно для данной области криптографии. Если последовательность случайных чисел предсказуема, то даже самый стойкий алгоритм шифрования, в котором данная последовательность будет использоваться, оказывается, уязвим — например, резко уменьшается пространство возможных ключей, которые необходимо «перебрать» злоумышленнику для получения некоторой информации, с помощью которой он сможет «взломать» всю систему. К счастью, разные организации все же пытаются навести здесь порядок, в частности, американский институт по стандартам NIST разработал набор тестов для оценки случайности последовательности чисел. О них и пойдет речь в данной статье. Но сначала — немного теории (постараюсь изложить не нудно).

Случайные двоичные последовательности

Во-первых, под генерацией случайных чисел подразумевается получение последовательности из двоичных знаков 0 и 1, а не байтами, как бы ни хотелось программистам. Идеальным подобным генератором является подбрасывание «идеальной» монеты (ровная монета, у которой вероятности выпадения каждой из сторон одинаковы), которую бы подбрасывали столько раз, сколько нужно, но проблема в том, ничего идеального не сущ ествует, а производительность такого генератора оставляла бы желать лучшего (один подрос монеты = одному биту). Тем не менее, все тесты, описываемые ниже, оценивают, насколько исследуемый генератор случайных чисел «похож» или «не похож» на воображаемую идеальную монету (не по скорости получения «случайных» знаков, а их «качества»).

Во-вторых, все генераторы случайных чисел делятся на 2 типа —истинно случайные — физические генераторы/датчики случайных чисел (ДСЧ/ФДСЧ) и псевдослучайные – программные датчики/генераторы случайных чисел (ПДСЧ). Первые принимают на вход некий случайный бесконечный процесс, а на выходе дают бесконечную (зависит от времени наблюдения) последовательность 0 и 1. Вторые представляют собой заданную разработчиком детерминированную функцию, которая инициализируется т. н. зерном, после чего также на выходе выдает последовательность 0 и 1. Зная это зерно, можно предсказать всю последовательность. Хороший ПДСЧ — этот тот, для которого невозможно предсказать последующие значения, имея всю историю предыдущих значений, не имея зерна. Это свойство называется прямой непредсказуемостью. Есть еще обратная непредсказуемость — невозможность вычислить зерно, зная любое количество генерируемых значений.

Казалось бы, проще всего взять истинно случайные/физические ДСЧ и не думать ни о какой предсказуемости. Однако тут есть проблемы:

- Случайное явление/процесс, которое берется за основу, может быть не способно выдавать числа с нужной скоростью. Если вы вспоминаете, когда последний раз генерировали пару 2048битных ключей, то не обольщайтесь. Это происходит очень редко? Тогда вообразите себя сервером, принимающим сотни запросов на SSL-соединения в секунду (SSL handshake предполагает генерацию пары случайных чисел).
- С виду случайные явления могут быть не такими случайными, как казалось бы. Например, электромагнитный шум может быть суперпозицией нескольких более-менее однообразных периодических сигналов.

Каждый из тестов, предлагаемых NIST, получает на вход конечную последовательность. Далее вычисляется статистика, характеризующая некое свойство данной последовательности — это может быть и единичное значение, и множество значений. После чего эта статистика сравнивается с эталонной статистикой, которую даст идеально случайная последовательность. Эталонная статистика выводится математически, этому посвящено множество теорем и научных трудов. В конце статьи будут даны все ссылки на источники, где выводятся нужные формулы.

Нулевая и альтернативная гипотезы

В основе тестов лежит понятие *нулевой гипотезы*. Попробую объяснить, что это. Допустим, мы набрали некую статистическую информацию. Например, пусть это будет количество людей, заболевших раком легких в группе из 1000 человек. И пусть известно, что некоторые люди из этой группы являются курильщиками, а другие нет, причем известно, какие конкретно. Стоит следующая задача: понять, есть ли взаимосвязь между курением и заболеванием. Нулевая гипотеза — это предположение, что между двумя фактами отсутствует какая-либо взаимосвязь. В нашем примере это предположение, что курение не вызывает рак легких. Существует также *альтернативная гипотеза*, которая опровергает нулевую гипотезу: т.е. между явлениями взаимосвязь существует (курение вызывает рак легких). Если переходить к терминам случайных чисел, то за нулевую гипотезу принимается предположение, что последовательность является истинно случайной (знаки которой появляются равновероятно и независимо друг от друга). Следовательно, если нулевая гипотеза верна, то наш генератор производит достаточно «хорошие» случайные числа.

Как проверяется гипотеза? С одной стороны, мы имеем статистику, подсчитанную на основе фактически собранных данных (т.е. по измеряемой последовательности). С другой стороны, есть эталонная статистика, получаемая математическими методами (теоретически вычисленная), которую бы имела истинно случайная последовательность. Очевидно, что собранная статистика не может сравняться с эталонной — насколько бы ни был хорошо наш генератор, он все равно не идеален. Поэтому вводят некую погрешность, например 5%. Она означает, что если, например, собранная статистика отклоняется от эталонной больше чем на 5%, то делается вывод о том, что нулевая гипотеза не верна с *большой надежностью*.

Так как мы имеем дело с гипотезами, то существует 4 варианта развития событий:

1. Сделан вывод о том, что последовательность случайна, и это верный вывод
2. Сделан вывод о том, что последовательность не случайна, хотя она была на самом деле случайна. Такие ошибки называют *ошибками первого рода*
3. Последовательность признана случайной, хотя на самом деле таковой не является. Такие ошибки называют *ошибками второго рода*
4. Последовательность справедливо отбракована

Вероятность ошибки первого рода называют *уровнем статистической значимости* и обозначают как α. Т.е. α — это вероятность отбраковать «хорошую» случайную последовательность. Это значение определяется областью применения. В криптографии принято α брать от 0.001 до 0.01.

В каждом тесте вычисляется т.н. *P-значение*: это вероятность того, что подопытный генератор произведет последовательность *не хуже*, чем гипотетический истинный. Если Pзначение = 1, то наша последовательность идеально случайна, а если оно = 0, то последовательность полностью предсказуема. В дальнейшем P-значение сравнивается с α, и если она больше α, то нулевая гипотеза принимается и последовательность признается случайной. В противном случае — отбраковывается.

В тестах берется α = 0.01. Из этого следует, что:

- Если P-значение ≥ 0.01, то последовательность признается случайной с уровнем доверия 99%
- Если P-значение < 0.01, то последовательность отбраковывается с уровнем доверия 99%

Итак, перейдем непосредственно к тестам.

Частотный побитовый тест

Очевидно, что чем более случайна последовательность, тем ближе это соотношение к 1. Данный тест оценивает, насколько это соотношение близко к 1.

Принимаем каждую «1» за +1, а каждый «0» за -1 и считаем сумму по всей последовательности. Это можно записать так:
$S_n = X_1 + X_2 + \ldots + X_n$, где $X_i = 2x_i — 1$.
Кстати, говорят, что распределение количества «успехов» в серии экспериментов, где в каждом эксперименте возможен *успех* или *неуспех* с

заданной вероятностью, имеет *биномиальное* распределение.

Возьмем такую последовательность: 1011010101

Тогда S = 1 + (-1) + 1 + 1 + (-1) + 1 + (-1) + 1 + (-1) + 1 = 2

Вычисляем статистику:

$$s_{obs} = \frac{|S|}{\sqrt{n}} = \frac{2}{\sqrt{10}} = 0.632455532$$

Вычисляем P-значение через *дополнительную функцию ошибок*:

$$P_{value} = erfc(\frac{s_{obs}}{\sqrt{2}}) = erfc(\frac{0.632455532}{\sqrt{2}}) = 0.527089$$

Дополнительная функция ошибок (complementary error function) определяется так:

$$erfc(x) = \frac{2}{\sqrt{\pi}} \int_x^\infty \mathrm{e}^{-t^2} \mathrm{d}t$$

Видим, что результат > 0.01, а значит наша последовательность прошла тест. Рекомендуется тестировать последовательности длиной не менее 100 бит.

Частотный блочный тест

Этот тест делается на основе предыдущего, только теперь значения пропорции «1»/«0» для каждого блока анализируются методом Хи-квадрат. Ясно, что это соотношение должно быть приблизительно равным 1.

Например, пусть дана последовательность 0110011010. Разобъем ее на блоки по 3 бита («бесхозный» 0 на конце отброшен):
011 001 101

Посчитаем пропорции π_i для каждого блока: $\pi_1 = 2/3$, $\pi_2 = 1/3$, $\pi_3 = 1/3$. Далее вычисляем статистику по методу Хи-квадрат с N степенями свободы (здесь N — количество блоков):

$$\chi^2_{obs} = 4 \cdot M \cdot \sum_{i=1}^{N} (\pi_i - 1/2)^2 = 4 \cdot 3 \cdot [(\frac{2}{3} - \frac{1}{2})^2 + (\frac{1}{3} - \frac{1}{2})^2 + (\frac{2}{3} - \frac{1}{2})^2] = 1$$

Вычислим P-значение через специальную функцию *Q*:

$$P_{value} = Q(\frac{N}{2}, \frac{\chi^2_{obs}}{2}) = Q(3/2, 1/2) = 0.801252$$

Q — это т.н. *неполная верхняя гамма-функция*, определяемая как:

$$Q(a, x) = \frac{1}{\Gamma(a)} \int_x^\infty \mathrm{e}^{-t} t^{a-1} dt$$

При этом функция Г — стандартная гамма-функция:

$$\Gamma(z) = \int_0^\infty \mathrm{t}^{z-1} e^{-t} dt$$

Последовательность считается случайной, если P-значение > 0.01. Рекомендуется анализировать последовательности длиной не менее 100 бит, а также должны выполняться соотношения M >= 20, M > 0.01n и N < 100.

Тест на одинаковые идущие подряд биты

В тесте ищутся все последовательности одинаковых битов, а затем анализируется, насколько количество и размеры этих последовательностей соответствуют количеству и размерам истинно случайной последовательности. Смысл в том, что если смена 0 на 1 (и обратно) происходит слишком редко, то такая последовательность «не тянет» на случайную.

Пусть дана последовательность 1001101011. Сначала вычисляем долю единиц в общей массе:

$$\pi = \frac{\sum_j X_j}{n} = \frac{6}{10} = \frac{3}{5}$$

Дальше проверяется условие:

$$\left|\pi - \frac{1}{2}\right| < \frac{2}{\sqrt{n}}$$

Если оно не удовлетворяется, то весь тест считается неуспешным и на этом все заканчивается. В нашем случае 0.63246 > 0.1, а значит идем

дальше.

Вычисляем суммарное число знакоперемен V:

$$V_n = \sum_{k=1}^{n-1} r(k) + 1$$

где $r(k) = 0$ если $X_i = X_{i+1}$, или $r(k) = 1$ в противном случае.

$$V_{10} = (1 + 0 + 1 + 0 + 1 + 1 + 1 + 1 + 0) + 1 = 7$$

Вычисляем P-значение через функцию ошибок:

$$P_{value} = erfc\left(\frac{|V_n - 2n\pi(1-\pi)|}{2\sqrt{2n}\pi(1-\pi)}\right) = erfc\left(\frac{7 - (2 \cdot 10 \cdot \frac{3}{5} \cdot (1 - \frac{3}{5}))}{2 \cdot \sqrt{2 \cdot 10} \cdot \frac{3}{5} \cdot (1 - \frac{3}{5})}\right) = 0.147232$$

Если результат >= 0.01 (как в нашем примере), то последовательность признается случайной.

Тест на самую длинную последовательность из единиц в блоке

Исходная последовательность из n битов разбивается на N блоков, каждый по M бит, после чего в каждом блоке ищется самая длинная последовательность единиц, а затем оценивается, насколько показатель близок к такому же показателю для истинно случайной последовательности. Очевидно, что аналогичного теста на нули не требуется, так как если единицы распределены *хорошо*, то нули также будут распределены *хорошо*.

Какую взять длину блока? NIST рекомендует несколько опорных значений, как разбивать на блоки:

Общая длина, n	Длина блока, M
128	8
6272	128
750000	10000

Пусть дана последовательность:
11001100 00010101 01101100 01001100 11100000 00000010
01001101 01010001 00010011 11010110 10000000 11010111
11001100 11100110 11011000 10110010

Разобьем ее на блоки по 8 бит (M=8), после чего посчитаем максимальную последовательность из единиц для каждого блока:

Блок	Длина единиц
11001100	2
00010101	1
01101100	2
01001100	2
11100000	3
00000010	1
01001101	2
01010001	1
00010011	2
11010110	2
10000000	1
11010111	3
11001100	2
11100110	3

11011000	2
10110010	2

Далее считаем статистику по разным длинам на основе следующей таблицы:

v_i	M = 8	M = 128	M = 10000
v_0	≤1	&le4	&le10
v_1	2	5	11
v_2	3	6	12
v_3	≥4	7	13
v_4		8	14
v_5		≥9	15
v_6			≥16

Как пользоваться этой таблицей: у нас M = 8, поэтому смотрим только один соответствующий столбец. Считаем v_i:
v_0 = { *кол-во блоков с макс. длиной ≤ 1* } = 4
v_1 = { *кол-во блоков с макс. длиной = 2* } = 9
v_2 = { *кол-во блоков с макс. длиной = 3* } = 3
v_3 = { *кол-во блоков с макс. длиной ≥ 4* } = 0

Вычисляем Хи-квадрат:

$$\chi^2 = \sum_{i=0}^{K} \frac{(v_i - R\pi_i)^2}{R\pi_i}$$

Где значения K и R берутся исходя из такой таблицы:

M	K	R
8	3	16
128	5	49
10000	6	75

Теоретические вероятности π_i задаются константами. Например, для K=3 и M=8 рекомендуется взять π_0 = 0.2148, π_1 = 0.3672, π_2 = 0.2305, π_3 = 0.1875. (Значения для других K и M приведены в [2]).

$$\chi^2_{(obs)} = \frac{(4-16\cdot 0.2148)^2}{16\cdot 0.2148} + \frac{(9-16\cdot 0.3672)^2}{16\cdot 0.3672} + \frac{(3-16\cdot 0.2305)^2}{16\cdot 0.2305} + \frac{(0-16\cdot 0.1875)^2}{16\cdot 0.1875} = 4.882605$$

Далее вычисляем P-значение:

$$P_{value} = igamc(\frac{K}{2}, \frac{\chi^2_{(obs)}}{2}) = igamc(\frac{3}{2}, \frac{4.882605}{2}) = 0.180598$$

Если оно > 0.01, как в нашем примере, то последовательность считается достаточно случайной.

Тест рангов бинарных матриц

Тест анализирует матрицы, которые составлены из исходной последовательности, а именно — рассчитывает ранги непересекающихся подматриц, построенных из исходной двоичной последовательности. В основе тест лежат исследования Коваленко [6], где ученый исследовал случайные матрицы, состоящие из 0 и 1. Он показал, что можно спрогнозировать вероятности того, что матрица M x Q будем иметь ранг R, где R = 0,1,2,...min(M,Q). Эти вероятности равны:

$$P_R = 2^{R(Q+M-R)-MQ} \prod_{i=0}^{R-1} \frac{(1-2^{i-Q})(1-2^{i-m})}{1-2^{i-R}}$$

NIST рекомендует брать M = Q = 32, а также, чтобы длина последовательности n = M^2 * N. Но мы для примера возьмем M = Q = 3. Далее нужны

вероятности P_M, P_{M-1} и P_{M-2}. С небольшой долей погрешности формулу можно упростить, и тогда эти вероятности равны:

$$P_M \approx \prod_{j=1}^{\infty}[1 - \frac{1}{2^j}] \approx 0.2888$$

$$P_{M-1} \approx 2P_M \approx 0.5776$$

$$P_{M-2} \approx \frac{4P_M}{9} \approx 0.1284$$

$$\Rightarrow 1 - 0.2888 - 0.5776 = 0.1336$$

Итак, пусть дана последовательность 01011001001010101101. «Раскладываем» ее по матрицам — хватило на 2 матрицы:

$$A_1 = \begin{vmatrix} 0 & 1 & 0 \\ 1 & 1 & 0 \\ 0 & 1 & 0 \end{vmatrix} A_2 = \begin{vmatrix} 0 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 1 \end{vmatrix}$$

Определяем ранг матриц: получается $R_1 = 2$, $R_2 = 3$. Для теста нужно 3 числа:

- F_M = {*кол-во матриц с рангом M*} = {*кол-во матриц с рангом 3*} = 1
- F_{M-1} = 1 (аналогично)
- N — F_M — F_{M-1} = 2 — 1 — 1 = 0

Вычисляем Хи-квадрат:

$$\chi^2 = \sum \frac{(F_i - NP_i)^2}{NP_i}$$

$$\chi^2_{(obs)} = \frac{(F_M - 0.2888N)^2}{0.2888N} + \frac{(F_{M-1} - 0.5776N)^2}{0.5776N} + \frac{(N - F_M - F_{M-1} - 0.1336N)^2}{0.1336N} = 0.596953$$

Вычисляем P-значение:

$$P_{value} = igamc(1, \frac{\chi^2_{(obs)}}{2}) = e^{\frac{-\chi^2_{(obs)}}{2}} = e^{-0.596953/2} = 0.741948$$

Если результат > 0.01, то последовательность признается случайной. NIST рекомендует, чтобы общая длина последовательности была >= 38MQ.

Спектральный тест

Подопытная последовательность рассматривается как дискретный сигнал, для которого делается спектральное разложение с целью выявить частотные пики. Очевидно, что такие пики будут свидетельствовать о наличии периодических составляющих, что не есть гут. Если вкратце, то тест выявляет пики, превышающие 95%-й барьер, после чего проверяет, не превышает ли доля этих пиков 5%.

Как нетрудно догадаться, для представления последовательности в виде суммы периодических составляющих будем использовать дискретное преобразование Фурье. Оно выглядит так:

$$X_j = \sum_{k=0}^{n-1} x_k e^{\frac{2\pi i k j}{n}}$$

Здесь x_k — исходная последовательность, в которой единице соответствует +1, а нулю -1, X_j — полученные значения комплексных амплитуд (комплексные означает, что в них содержится как вещественное значение амплитуды, так и фаза).

Вы спросите, а где же здесь периодичности? Ответ — экспоненту можно выразить через тригонометрические функции:

$$e^{\frac{2\pi i k j}{n}} = cos(\frac{2\pi k j}{n}) + i \cdot sin(\frac{2\pi i k j}{n})(j = 0, ..., n-1, i \equiv \sqrt{-1})$$

Для нашего теста интересны не фазы, а абсолютные значения амплитуд. И если мы вычислим эти абсолютные значения, то окажется, что они симметричны (это общеизвестный факт при переходе от комплексных значений к вещественным), поэтому для дальнейшего рассмотрения мы возьмём только половину этих значений (от 0 до n/2) — остальные не несут дополнительной информации.

Покажем все это на примере. Пусть задана последовательность 1001010011.
Тогда x = { 1, -1, -1, 1, -1, 1, -1, -1, 1, 1 }.

Вот как разложение Фурье можно сделать, например, в программе GNU Octave:

```
octave:1> x = [1, -1, -1, 1, -1, 1, -1, -1, 1, 1]

x =
   1  -1  -1   1  -1   1  -1  -1   1   1

octave:2> abs(fft(x))

ans =

   0.0000  2.0000  4.4721  2.0000  4.4721  2.0000  4.4721  2.0000  4.4721  2.0000
```

Видим, что наблюдается симметрия. Поэтому нам хватит и пять значений: 0, 2, 4.4721, 2, 4.4721.

Далее вычисляем граничное значение по формуле

$$T = \sqrt{log(\frac{1}{0.05}) \cdot n} \approx 5.47$$

Оно означает, что если последовательность истинно случайная, то 95% пиков не должны превышать эту границу.

Вычислим предельное число пиков, которых должно быть меньше, чем T:

$$N_0 = \frac{0.95n}{2} = 4.75$$

Далее смотрим на результат разложения и видим, что все наши 4 пика меньше граничного значения. Далее оцениваем эту разницу:

$$d = \frac{N_1 - N_0}{\sqrt{n \cdot 0.95 \cdot 0.05/4}} = \frac{4 - 4.75}{\sqrt{10 \cdot 0.95 \cdot 0.05/4}} = -2.176429$$

Вычисляем P-значение:

$$P_{value} = erfc(\frac{|d|}{\sqrt{2}}) = erfc(\frac{2.176429}{\sqrt{2}}) = 0.029523$$

Оно получилось >0.01, поэтому гипотеза о случайности принимается. И да, для теста рекомендуется брать не менее 1000 бит.

Тест на встречающиеся непересекающиеся шаблоны

Подопытная последовательность разбивается на блоки одинаковой длины. Например:

```
1010010010 1110010110
```

В каждом блоке будем искать какой-нибудь шаблон, например «001». Слово *непересекающиеся* означает, что в случае нахождения шаблона внутри последовательности, следующее сравнение не будет захватывать ни одного бита найденного шаблона. В результате поиска для каждого i-го блока будет найдено число W_i, равное кол-ву найденных случаев.

Итак, для наших блоков $W_1 = 2$ и $W_2 = 1$:
101 **001 001** 0
111 **001** 0110

Вычислим математические ожидание и дисперсию, как если бы наша последовательность была подлинно случайна. Ниже приведены формулы. Здесь N = 2 (кол-во блоков), M = 10 (длина блока), m = 3 (длина образца).

$$\mu = \frac{M - m + 1}{2^m} = \frac{10 - 3 + 1}{2^3} = 1$$

$$\sigma^2 = M \cdot (\frac{1}{2^m} - \frac{2m - 1}{2^{2m}}) = 10 \cdot (\frac{1}{2^3} - \frac{2 \cdot 3 - 1}{2^{2 \cdot 3}})$$

Вычислим Хи-квадрат:

$$\chi^2_{obs} = \sum_{j=1}^{N} \frac{(W_j - \sigma)^2}{\sigma^2} = \frac{(2-1)^2 + (1-1)^2}{0.46875} = \frac{1+0}{0.46785} = 2.133333$$

Вычислим итоговое P-значение через неполную гамма-функцию:

$$Q(a, x) = Q(\frac{N}{2}, \frac{\chi^2_{obs}}{2}) = Q(\frac{2}{2}, \frac{2.13333}{2}) = 0.344154$$

Видим, что P-значение > 0.1, а значит, последовательность достаточно случайна.

Мы оценили только один шаблон. На самом деле нужно проверить все комбинации шаблонов, да и ещё к тому же для разной длины этих шаблонов. Сколько того и другого нужно, определяется исходя из конкретных требований, но обычно *m* берут 9 или 10. Чтобы получить осмысленные результаты, следует брать N < 100 и M > 0.01 * n.

Тест на встречающиеся пересекающиеся шаблоны

Этот тест отличается от предыдущего тем, что при нахождении шаблона «окно» поиска сдвигается не на длину шаблона, а только на 1 бит. Чтобы не загромождать статью, мы не станем приводить пример расчета по этому методу. Он полностью аналогичен.

Универсальный тест Мауэра

Тест оценивает, насколько «далеко» друг от друга отстоят шаблоны внутри последовательности. Смысл теста в том, чтобы понять, насколько последовательность сжимаема (конечно, имеется в виду сжатие без потерь). Чем более сжимаема последовательность, тем она менее случайна. Алгоритм этого теста весьма громоздкий для Хабра-формата, поэтому опустим его.

Тест на линейную сложность

В основе теста лежит предположение, что подопытная последовательность была получена через *регистр сдвига с линейной обратной связью* (или LFSR, Linear feedback shift register). Это общеизвестный метод получения бесконечной последовательности: тут каждый следующий бит получается как некая функция бит, «сидящих» в регистре. Минус LFSR в том, что он всегда имеет конечный период, т.е. последовательность обязательно будет рано или поздно повторяться. Чем больше длина LFSR, тем *лучше* случайная последовательность.

Исходная последовательность разбивается на равные блоки длиной M. Далее для каждого блока с помощью алгоритма Берлекэмпа — Мэсси [10] находится его линейная сложность (L_i), т.е. длина LFSR. Затем для всех найденных L_i оценивается распределение Хи-квадрат со 6 степенями свободы. Покажем на примере.

Пусть дан блок 1101011110001 (M=13), для которого алгоритм Берлекэмпа — Мэсси выдал L = 4. Убедимся, что это так. Действительно, нетрудно догадаться, что для этого блока каждый следующий бит получается как сумма (по модулю 2) 1-го и 2-го бита (нумерация с 1):
$x_5 = x_1 + x_2 = 1 + 1 = 0$
$x_6 = x_2 + x_3 = 1 + 0 = 1$
$x_7 = x_3 + x_4 = 1 + 0 = 1$
и.т.д.

Вычисляем математическое ожидание по формуле
$$\mu = \frac{M}{2} + \frac{9+(-1)^{M+1}}{36} - \frac{\frac{M}{3}+\frac{2}{9}}{2^M} = \frac{13}{2} + \frac{9+(-1)^{13+1}}{36} - \frac{\frac{13}{3}+\frac{2}{9}}{2^{13}} = 6.777222$$

Для каждого блока вычисляем значение T_i:
$$T_i = (-1)^M \cdot (L_i - \mu) + \frac{2}{9} = (-1)^{13} \cdot (4 - 6.777222) + \frac{2}{9} = 2.999444$$

Далее на основе множества T вычисляем набор $v_0,...,v_6$ таким образом:

- если $T_i <= -2.5$, то v_0++
- если $-2.5 < T_i <= -1.5$, то v_1++
- если $-1.5 < T_i <= -0.5$, то v_2++
- если $-0.5 < T_i <= 0.5$, то v_3++
- если $0.5 < T_i <= 1.5$, то v_4++
- если $1.5 < T_i <= 2.5$, то v_5++
- если $T_i > 2.5$, то v_6++

Имеем 7 возможных исходов, а значит вычисляем Хи-квадрат с числом степеней свободы 7 — 1 = 6:
$$\chi^2 = \sum_{i=0}^{K} \frac{(v_i - N\pi_i)^2}{N\pi_i}$$
Вероятности π_i в тесте жестко заданы и равны соответственно: 0.010417, 0.03125, 0.125, 0.5, 0.25, 0.0625, 0.020833. (π_i для большего числа степеней свободы можно вычислить по формулам, данным в [2]).

Вычислить P-значение:

$$P_{value} = igamc(\frac{K}{2}, \frac{\chi^2}{2})$$

Если результат получился > 0.01, то последовательность признается случайной. Для реальных тестов рекомендуется брать n >= 10^6 и M в пределах от 500 до 5000.

Тест на подпоследовательности

Анализируется частота нахождения всевозможных последовательностей длиной «m» бит внутри исходной последовательности. При этом каждый образец ищется независимо, т.е. возможно как бы «наложение» одного найденного образца на другой. Очевидно, что количество всевозможных образцов будет 2^m. Если последовательность достаточно велика и случайна, то вероятности нахождения каждого из этих образцов одинакова. (Кстати, если m = 1, то этот тест «вырождается» в уже описанный ранее тест на соотношение «0» или «1»).

В основе теста лежат работы [8] и [11]. Там описываются 2 показателя ($\nabla\psi^2_m$ и $\nabla^2\psi^2_m$), которые характеризуют, насколько частоты появления образцов соответствуют этим же частотам для истинно случайной последовательности. Покажем алгоритм на примере.

Пусть дана последовательность 0011011101 длиной n = 10, а также m = 3.

Сначала формируется 3 новых последовательности, каждая из которых получается добавлением m-1 первых битов последовательности к её концу. Получается:

- Для m = 3: 0011011101 00 (добавили 2 бита к концу)
- Для m-1 = 2: 0011011101 0 (добавили 1 бит к концу)
- Для m-2 = 1: 0011011101 (исходная последовательность)

Далее найдем частоты появления всех блоков длиной m, m-1 и m-2 соответственно:

- $v_{000} = 0$, $v_{001} = 1$, $v_{010} = 1$, $v_{011} = 2$, $v_{100} = 1$, $v_{101} = 2$, $v_{110} = 2$, $v_{111} = 0$
- $v_{00} = 1$, $v_{01} = 3$, $v_{10} = 3$, $v_{11} = 3$
- $v_0 = 4$, $v_1 = 6$

Вычисляем нужные статистики по формулам:

$$\psi^2_m = \frac{2^m}{n} \cdot \sum_{i_1 \dots i_m} (v^2_{i_1 \dots i_m}) - n$$
$$\psi^2_{m-1} = \frac{2^{m-1}}{n} \cdot \sum_{i_1 \dots i_{m-1}} (v^2_{i_1 \dots i_{m-1}}) - n$$
$$\psi^2_{m-2} = \frac{2^{m-2}}{n} \cdot \sum_{i_1 \dots i_{m-2}} (v^2_{i_1 \dots i_{m-2}}) - n$$

Подставляем:

$$\psi^2_1 = \frac{2}{10}(16 + 36) - 10 = 10.4 - 10 = 0.4$$
$$\psi^2_2 = \frac{2^2}{10}(1 + 9 + 9 + 9) - 10 = 11.2 - 10 = 1.2$$
$$\psi^2_3 = \frac{2^3}{10}(0 + 1 + 1 + 4 + 1 + 4 + 4 + 0) - 10 = 12 - 10 = 2$$

Тогда:

$$\nabla\psi^2_m = \psi^2_m - \psi^2_{m-1} = \psi^2_3 - \psi^2_2 = 2 - 1.2 = 0.8$$
$$\nabla^2\psi^2_m = \psi^2_m - 2\psi^2_{m-1} + \psi^2_{m-2} = 2 - 2 \cdot 1.2 + 0.4 = 0$$

Итоговые значения:

$$P_{value1} = igamc(2^{m-2}, \nabla\psi^2_m) = igamc(2, \frac{0.8}{2}) = 0.93845$$
$$P_{value2} = igamc(2^{m-3}, \nabla^2\psi^2_m) = igamc(1, \frac{0}{2}) = 1$$

Итак, оба P-значения > 0.01, а значит последовательность признается случайной.

Приблизительная энтропия

Метод приблизительной энтропии (Approximate Entropy) изначально проявил себя в медицине, особенно в кардиологии. Вообще, согласно

классическому определению, энтропия является мерой хаоса: чем она выше, тем более непредсказуемые явления. Хорошо это или плохо, зависит от контекста. Для случайных последовательностей, используемых в криптографии, важно иметь высокую энтропию — это значит, что будет сложно предсказать последующие случайные биты на основе того, что уже имеем. А вот, например, если за случайную величину взять сердечный ритм, измеряемый с заданным периодом, то ситуация иная: есть множество исследований (например, [12]), доказывающих, что чем ниже вариабельность сердечных ритмов, тем реже вероятность инфарктов и прочих неприятных явлений. Очевидно, что сердце человека не может биться с постоянной частотой. Однако одни умирают от инфарктов, а другие нет. Поэтому метод приблизительной энтропии позволяет оценить, насколько с виду случайные явления *действительно случайны*.

Конкретно, тест вычисляет частоты появления всевозможных образцов заданной длины (m), а затем аналогичные частоты, но уже для образцов длиной m+1. Затем распределение частот сравнивается с эталонным распределением Хи-квадрат. Как и в предыдущем тесте, образцы могут перекрываться.

Покажем на примере. Пусть дана последовательность 0100110101 (длина n = 10), и возьмём m = 3.

Для начала дополним последовательность первыми m-1 битами. Получится 0100110101 01.

Посчитаем встречаемость каждого из 8 всевозможных блоков. Получится:
$k_{000} = 0$, $k_{001} = 1$, $k_{010} = 3$, $k_{011} = 1$, $k_{100} = 1$, $k_{101} = 3$, $k_{110} = 1$, $k_{111} = 0$.

Посчитаем соответствующие частоты по формуле $C_i^m = k_i / n$:
$C_{000}^3 = 0$, $C_{001}^3 = 0.1$, $C_{010}^3 = 0.3$, $C_{011}^3 = 0.1$, $C_{100}^3 = 0.1$, $C_{101}^3 = 0.3$, $C_{110}^3 = 0.1$, $C_{111}^3 = 0$.

Аналогичным образом считаем частоты появления подблоков длиной m+1=4. Их уже 2^4=16:
$C_{0011}^4 = C_{0100}^4 = C_{0110}^4 = C_{1001}^4 = C_{1101}^4 = 0.1$, $C_{0101}^4 = 0.2$, $C_{1010}^4 = 0.3$. Остальные частоты = 0.

Вычисляем φ^3 и φ^4 (заметьте, что здесь натуральный логарифм):

$$\phi^m = \sum_{i=0}^{2^m-1} \pi_i \ln \pi_i = \sum_{i=0}^{2^m-1} C_i^3 \ln C_i^3 = 0 + 0.1(ln0.1) + 0.3(ln0.3) + 0.1(ln0.1) + 0.1(ln0.1) + 0.3(ln0.3) + 0.1(ln0.1) = -1.64341772$$

$$\phi^{m+1} = 0 + 0 + 0 + 0.1(ln0.01) + 0.1(ln0.01) + 0.2(ln0.02) + 0.1(ln0.01) + 0 + 0 + .1(ln0.01) + 0.3(ln0.03) + 0 + 0 + 0.1(ln0.01) + 0 + 0 = -1.83437197$$

Вычисляем Хи-квадрат:

$$\chi^2 = 2n(ln2 - (\phi^3 - \phi^4) = 2 \cdot 10 \cdot (0.693147 + 1.64341772 - 1.83437197) = 10.044$$

P-значение:

$$P_{value} = igamc(2^{m-1}, \frac{\chi^2}{2}) = igamc(4, \frac{10.044}{2}) = 0.261961$$

Получившееся значение > 0.01, а значит последовательность признается случайной.

Тест кумулятивных сумм

Примем каждый нулевой бит исходной последовательности за -1, а каждый единичный — за +1, после чего посчитаем сумму. Интуитивно понятно, что чем более случайна последовательность, тем быстрее эта сумма будет стремиться к нулю. С другой стороны, представим, что дана последовательность, состоящая из 100 нулей и 100 единиц, идущих подряд: 00000...001111...11. Здесь сумма получится равной 0, однако очевидно, что назвать *такую* последовательность случайной «рука не поднимется». Следовательно, нужен более глубокий критерий. И этим критерием являются частичные суммы. Будем постепенно считать суммы, начиная от первого элемента:
$S_1 = x_1$
$S_2 = x_1 + x_2$ $S_3 = x_1 + x_2 + x_3$…
$S_n = x_1 + x_2 + x_3 + \ldots + x_n$

Далее находится число z = максимум среди этих сумм.

Наконец, считается P-значение по следующей формуле (её вывод см. в [9]):

$$P_{value} = 1 - \Sigma_1 + \Sigma_2$$

Где:

$$\Sigma_1 = \sum_{k=(\frac{-n}{z}+1)\cdot 4}^{(\frac{n}{z}-1)\cdot 4} [\Phi(\frac{(4k+1)z}{\sqrt{n}}) - \Phi(\frac{(4k-1)z}{\sqrt{n}})]$$

$$\Sigma_2 = \sum_{k=(\frac{-n}{z}-3)\cdot 4}^{(\frac{n}{z}-1)\cdot 4} \left[\Phi\left(\frac{(4k+3)z}{\sqrt{n}}\right) - \Phi\left(\frac{(4k+1)z}{\sqrt{n}}\right)\right]$$

Здесь Ф — функция распределения стандартной нормальной случайной величины. Напоминаем, что стандартное нормальное распределение — это всем известное гауссово распределение (в форме колокола), у которого математическое ожидание 0 и дисперсия 1. Выглядит так:

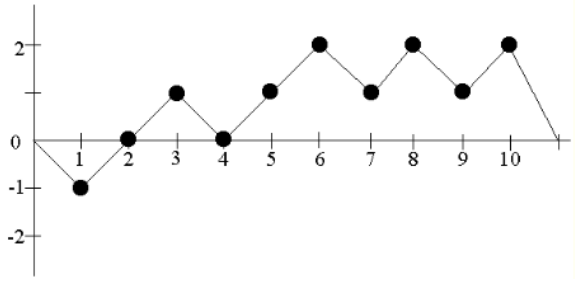
$$\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} \mathrm{e}^{\frac{-u^2}{2}} du$$

Если получившееся P-значение > 0.01, то последовательность признается случайной.

Кстати, у этого теста есть 2 режима: первый мы только что рассмотрели, а во втором суммы считаются начиная с последнего элемента.

Тест на произвольные отклонения

Этот тест похож на предыдущий: аналогичным образом считаются частичные суммы нормализованной последовательности (т.е. состоящей из -1 и 1). Пусть дана последовательность 0110110101 и пусть S(i) — это частичная сумма с 1 по i-й элемент. Нанесем эти точки на график, предварительно прибавив «0» к началу и концу последовательности S(i) — это нужно для целостности дальнейших рассчетов:



Отметим точки, где график пересекает горизонтальную ось — эти точки будут делить последовательность на т.н. *циклы*. Здесь мы имеем 3 цикла: {0, -1, 0}, {0, 1, 0} и {0, 1, 2, 1, 2, 1, 2, 0}. Далее, говорят, что каждый из этих циклов последовательно принимает различные *состояния*. Например, первый цикл 2 раза принимает состояние «0» и 1 раз состояние "-1". Для данного теста интересуют состояния от -4 до 4. Занесем все нахождения в этих состояниях в такую таблицу:

Состояние (x)	Цикл №1	Цикл №2	Цикл №3
-4	0	0	0
-3	0	0	0
-2	0	0	0
-1	1	0	0
1	0	1	3
2	0	0	3
3	0	0	0
4	0	0	0

На основе этой таблицы формируем другую таблицу: в ней по горизонтали пойдут количества циклов, принимающих заданное состояние:

Состояние (x)	Ни разу	1 раз	2 раза	3 раза	4 раза	5 раз
-4	3	0	0	0	0	0
-3	3	0	0	0	0	0
-2	3	0	0	0	0	0
-1	2	1	0	0	0	0
1	1	1	0	1	0	0
2	2	0	0	1	0	0
3	3	0	0	0	0	0
4	3	0	0	0	0	0

Далее для каждого из восьми состояний вычисляется Хи-квадрат статистики по формуле

$$\chi^2_{obs} = \sum_{k=0}^{5} \frac{(v_k(x) - J\pi_k(x))^2}{J\pi_k(x)}$$

Где $v_k(x)$ — значения в таблице для данного состояния, J — количество циклов (у нас 3), $\pi_k(x)$ — вероятности того, что состояние «x» возникнет k раз в подлинно случайном распределении (они известны).

Например, для x=1 получается:

$$\chi^2 = \frac{(1-3\cdot0.5)^2}{3\cdot0.5} + \frac{(1-3\cdot0.25)^2}{3\cdot0.25} + \frac{(0-3\cdot0.125)^2}{3\cdot0.125} + \frac{(1-3\cdot0.0625)^2}{3\cdot0.0625} + \frac{(0-3\cdot0.0312)^2}{3\cdot0.0312} + \frac{(0-3\cdot0.0312)^2}{3\cdot0.0312} = 4.333$$

Значения π для остальных *x* смотрите в [2].

Вычисляем P-значение:

$$P_{value} = igamc(\frac{k}{2}, \frac{\chi^2_{obs}}{2}) = igamc(5/2, 4.333/2) = 0.502529$$

Если оно > 0.01, то делается вывод о случайности. В итоге необходимо вычислить 8 P-значений. Какие-то могут оказаться больше 0.01, какие-то — меньше. В таком случае финальное решение о последовательности делается на основе других тестов.

Разновидность теста на произвольные отклонения

Практически похож на предыдущий тест, но берется более широкий набор состояний: -9, -8, -7, -6, -5, -4, -3, -2, -1, +1, +2, +3, +4, +5, +6, +7, +8, +9. Но главное отличие в том, что здесь P-значение вычисляется не через гамма-функцию (igamc) и Хи-квадрат, а через функцию ошибок (erfc). За точными формулами читатель может обратиться к исходному документу.

Ниже привожу список источников, которые можно посмотреть, если хочется углубиться в тему:

1. csrc.nist.gov/groups/ST/toolkit/rng/stats_tests.html
2. csrc.nist.gov/groups/ST/toolkit/rng/documents/SP800-22rev1a.pdf
3. Центральная предельная теорема
4. Anant P. Godbole and Stavros G. Papastavridis, (ed), Runs and patterns in probability: Selected papers. Dordrecht: Kluwer Academic, 1994
5. Pal Revesz, Random Walk in Random and Non-Random Environments. Singapore: World Scientific, 1990
6. И. Н. Коваленко, Теория вероятностей и её применения, 1972
7. O. Chrysaphinou, S. Papastavridis, “A Limit Theorem on the Number of Overlapping Appearances of a Pattern in a Sequence of Independent Trials.” Probability Theory and Related Fields, Vol. 79, 1988
8. I. J. Good, “The serial test for sampling numbers and other tests for randomness,” Cambridge, 1953
9. A. Rukhin, “Approximate entropy for testing randomness,” Journal of Applied Probability, 2000
10. Алгоритм Берлекэмпа — Мэсси
11. D. E. Knuth, The Art of Computer Programming. Vol. 2 & 3, 1998
12. www.ncbi.nlm.nih.gov/pubmed/8466069

Теги: случайные числа, криптография, математическая статистика