Повышение равномерности псевдослучайных чисел

А.С. Куцаев[1]

[1]ФГУ ФНЦ НИИСИ РАН, Москва, Россия, koutsaev@niisi.msk.ru

Аннотация. Качество выработки случайных тестов зависит от выбора генератора случайных чисел. Проверка популярных генераторов с помощью критерия Пирсона хи-квадрат показывает, что равномерное распределение с уровнем значимости 70% и выше наблюдается лишь у половины выборок. По этой причине, чтобы получить необходимое покрытие, нужно увеличивать объем тестов. Предлагается способ повышения равномерности распределения для генераторов случайных чисел, основанный на фильтрации выборки. Подбор параметров фильтра позволяет получить нужную равномерность распределения при умеренном числе пропусков.

Ключевые слова: генератор случайных чисел, равномерное распределение, хи-квадрат, фильтрация, случайные тесты.

1. Введение

В основе компьютерных датчиков псевдослучайных чисел (ПСЧ) обычно лежит моделирование равномерного распределения целых чисел на отрезке от нуля до наибольшего при данной разрядности. Уже первые опыты с датчиками показали, что сложные манипуляции с числами не гарантируют качество результата. Примером может служить метод "середины квадрата" Д. фон Неймана. Это рекурсивный алгоритм, в котором очередное число получается из предыдущего возведением его в квадрат и вырезанием отрезка битов из средней части результата. При этом количество единичных битов результата обычно снижается, и с какого-то момента датчик выдает только нули.

В настоящее время предложено много методов генерации ПСЧ, как универсальных, так и специализированных. Разрядность датчиков меняется от 32 до 128 битов. Генераторы большой разрядности используются в системах с повышенными требованиями к безопасности. Датчик ПСЧ должен моделировать равномерность распределения и случайность значений в выборках. Для компьютерного применения потребовались также минимум затрат, возможность точного воспроизведения уже полученной выборки, а затем и криптографическая стойкость, т.е. защита от попыток восстановить случайную последовательность по ее части.

Один из популярных способов получения ПСЧ предложен в 1948 году Д.Х. Лемером [1]. Рекурсивная схема основана на выражении $x_{n+1} = (a \cdot x_n + c)_{mod\ m}$, где a, c, m - константы. Полученная последовательность ПСЧ называется линейной конгруэнтной последовательностью. Она периодическая, и величина периода отчасти характеризует ее случайность. Разработаны правила для выбора констант a, c, m, но их соблюдение не гарантирует высокого качества выборок.

Другое семейство методов под общим названием XorShift предложено Д. Марсалья в 2003 году [2]. В его основе набор операций сдвига и исключающего ИЛИ над одним или несколькими предыдущими членами последовательности ПСЧ.

Из более сложных методов нужно отметить Mersenne Twister, разработанный М. Мацумото и Т. Нисимура в 1997 году и также имеющий несколько известных реализаций [3], [4]. Его особенностью среди прочего является "размазывание" результатов арифметических операций по достаточно длинному (до 624 слов) буферу, что позволяет получить огромный период датчика (для MT19937 это 2^{19937} - 1). Кроме того, после применения рекуррентного соотношения делается перемешивание битов результата, названное "закалкой" (tempering). Дальнейшее усложнение Mersenne Twister, названное WELL (Well Equidistributed Long-period Linear, [5]), позволило повысить недостаточную равномерность распределения.

При генерации случайных тестов часто используются небольшие выборки ПСЧ, равномерно распределенные на отрезке небольшой длины. Так, для случайного выбора аргументов инструкций требуется выбор 5-битового непосредственного значения или номера регистра, причем не все возможные значения допустимы. Неравномерность проявляется как частый выбор близких значений, и компенсировать ее можно только увеличением объема тестового материала. Равномерность здесь требуется в основном для наиболее значимых битов чисел.

В данной работе предлагается способ повышения равномерности распределения любого датчика ПСЧ с помощью фильтрации. В Главе 2

рассмотрена оценка равномерности для ряда популярных датчиков ПСЧ. В Главе 3 описан способ повышения равномерности распределения. В Главе 4 рассматриваются вероятностные оценки для получаемых выборок. Глава 5 содержит выводы. В Приложении приведены схемы использованных датчиков ПСЧ.

2. Оценки равномерности

Проверка закона распределения обычно начинается с оценок среднего, дисперсии и анализа частот с помощью критерия хи-квадрат [6]. В популярных датчиках среднее и дисперсия сходятся к теоретическим значениям, тогда как проверка равномерности не всегда дает желаемый результат.

При вычислении оценки критерия хи-квадрат для выборки из n чисел область изменения значений разбивается на k интервалов и находится сумма χ^2:

$$\chi^2 = n \cdot \sum_{i=1}^{k} \frac{(n_i/n - p_i)^2}{p_i}, \qquad (2.1)$$

где n_i - количество чисел выборки, попавших в интервал i, а p_i - теоретическая вероятность попадания этот же интервал. Для равномерного распределения, если все интервалы одной длины, $p_i = 1/k$ и выражение для оценки можно упростить.

$$\chi^2 = \sum_{i=1}^{k} \frac{(n_i - n/k)^2}{n/k} \qquad (2.2)$$

Полученная оценка сравнивается с распределением хи-квадрат Пирсона с k - 1 степенями свободы (так как сумма частот n_i равна n, независимых частот на единицу меньше). Если значение оценки меньше, чем квантиль функции распределения хи-квадрат для некоторой вероятности p, то с этой вероятностью выборка соответствует заданному закону распределения (в данном случае - равномерному). Иначе либо закон распределения принимается с меньшей вероятностью, либо отвергается.

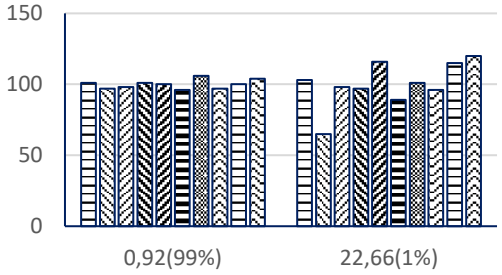

Рис. 1. Гистограммы выборок ПСЧ для крайних уровней значимости по хи-квадрат.

Для наглядности на Рис. 1 приведены гистограммы выборок ПСЧ для лучшего (χ^2 = 0.92) и худшего (χ^2 = 22.66) значений оценки на 10 интервалах при длине выборки 1000. Уровень значимости для равномерности составляет здесь 99% и 1%, соответственно.

Далее для сравнения были взяты линейный конгруэнтный генератор (LCG), две версии генератора XorShift и две версии Mersenne Twister. Алгоритмы и параметры генераторов подробно описаны в Приложении.

Датчики вырабатывают 32-битовые целые ПСЧ, равномерно распределенные на отрезке от 0 до 2^{31}-1. Далее они преобразуются к более удобному полуинтервалу чисел с плавающей точкой [0, 1), который разбивается на k одинаковых частей. При преобразовании использовались 23 наиболее значимых бита числа из 32.

Для оценки равномерности брались серии выборок, полученных при различных значениях "зерна". Значения χ^2 у выборок менялись в широких пределах, а соответствующий уровень значимости для равномерности распределения мог принимать значения от 1 до 99%. Для более детального анализа можно рассмотреть гистограммы получаемых значений χ^2 в серии выборок.

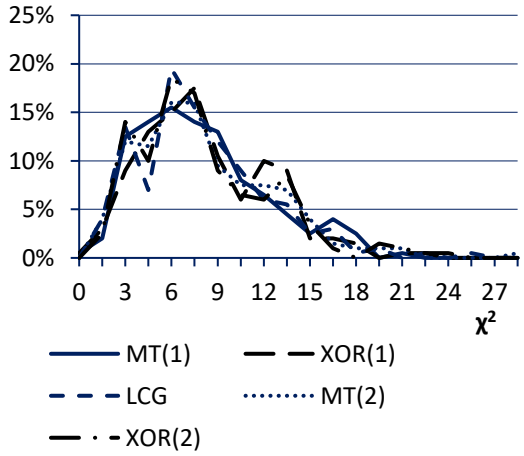

Рис. 2. Гистограммы значений χ^2 по 10 интервалам на 200 выборках для датчиков Mersenne Twister (MT 1 и 2), XorShift (XOR 1 и 2), LCG.

На Рис. 2 показаны гистограммы χ^2 для пяти датчиков ПСЧ. Длина выборок 2000, значение χ^2 вычислялось по 10 интервалам на 200 выборках. Отрезок изменения χ^2, общий для всех выборок, разбит на 20 частей, показан процент значений, попадающих в каждую часть.

У всех гистограмм значения медиан (от 6.4 до 7.1) отвечают уровню значимости около 70%. С ростом значений χ^2 соответствующая доля выборок убывает, оставаясь ненулевой и для очень малых уровней значимости. Это значит, что мак-

симум значений χ^2 не подходит в качестве критерия для оценки и сравнения качества датчиков. Также ни один датчик не дает гарантии равномерного распределения. Для оценки качества можно брать долю выборок, для которых уровень значимости по хи-квадрат превышает определенный порог, например, 80%. Для сравнения датчиков можно также использовать среднее значение χ^2 по серии выборок.

3. Фильтрация

Испытанным приемом при генерации случайных чисел является пропуск части выработанных значений согласно заданному условию. Такие пропуски используются, например, при генерации случайных чисел с нормальным распределением либо для ускорения работы. Этот же подход можно использовать для повышения равномерности любого датчика ПСЧ.

Для фильтрации разобьем область значений ПСЧ на F одинаковых интервалов. С каждым интервалом связан счетчик попавших в него чисел. Введем порог для разности максимума и минимума текущих значений счетчиков. Если очередное ПСЧ приводит к превышению порога, оно пропускается. Таким образом, значения всех счетчиков отличаются друг от друга на величину, не превышающую порог. Порог выбирается так, чтобы обеспечить нужную равномерность, а доля пропусков была бы не слишком велика.

Такая фильтрация выравнивает гистограмму выборки в случае, когда число интервалов гистограммы k равно числу интервалов фильтрации F. Для других k влияние фильтрации сложнее, но и в этом случае происходит определенное выравнивание гистограмм.

При выборе числа интервалов фильтрации F и порога d нужно учитывать, что фильтрация делает оценку хи-квадрат зависимой от размера выборки n. При $k=F$ величины n_i в (2.1) совпадают со значениями счетчиков фильтрации. Можно показать, что при пороге фильтрации $|n_i - n_j| \le d$ и четном $k=F$ максимум суммы в χ^2 достигается, если половина счетчиков имеет значение $n/k+d/2$, и половина $n/k-d/2$.

$$\chi^2 = \sum_{i=1}^{k} \frac{(n_i - n/k)^2}{n/k} \le \frac{k^2 d^2}{4n} \qquad (3.1)$$

В этом случае размер выборки существенно влияет на значение χ^2. Например, при $k=F=16$, $d=10$ и $n=1000$ оценка (3.1) дает значение 6.4, что ниже квантиля распределения хи-квадрат с $r=15$ степенями свободы для уровня значимости 95%. Иначе говоря, при большом размере выборки оценка хи-квадрат покажет сколь угодно высокую равномерность выборки после фильтрации независимо от качества использованного датчика ПСЧ. Вообще неравенство (3.1) с заменой k на F выполняется при $Ak = F$ для любых целых A. Для других значений k неравенство (3.1), вообще говоря, не выполняется, но ограничение фильтрации сказывается и здесь: с ростом размера выборки падает чувствительность оценки хи-квадрат.

Пример зависимости среднего значения χ^2 от числа интервалов k показан на Рис. 3. Использован датчик MT19937, фильтрация с порогом 10 при числе интервалов F=16 и F=32. Осреднение делалось по 200 выборкам, длина выборки 2000. Здесь же приведены квантили распределения хи-квадрат для уровня значимости P = 90%, а также изменение χ^2 в отсутствие фильтрации (F = 0).

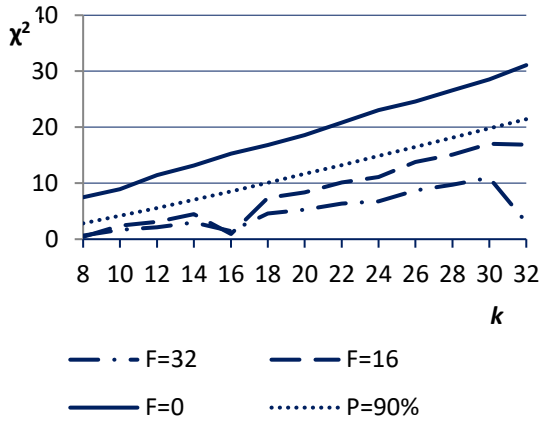

Рис. 3. Средние значения χ^2 на 200 выборках для датчика MT19937 при фильтрации с F = 32 и F = 16 интервалами, а также без фильтрации (F = 0), в зависимости от числа интервалов оценки хи-квадрат k. Для справки показаны квантили распределения хи-квадрат для уровня значимости P = 90%.

На графиках χ^2 видны провалы при $k = F$. Также есть провалы при $2k = F$ и $4k = F$, согласно замечанию для $Ak = F$ выше. Средние значения χ^2 при наличии фильтрации лежат ниже графика квантилей для уровня значимости 90%, но максимумы в серии заметно выше этого графика. Поведение кривых показывает, что случаи, когда F кратно k, не подходят для подбора параметров фильтрации. Для этого можно использовать осреднение по набору используемых на практике значений k, исключив из него случаи кратности.

При оценках равномерности нужно учитывать размер выборки n. Для больших размеров (n = 2000 и выше) равномерность при фильтрации может завышаться, поскольку при вычислении оценки ограниченные величины делятся на n.

Это видно по провалам графиков на Рис. 3. Размеры n = 100 и ниже не подходят, так как частоты попадания в интервалы оказываются меньше 5, что делает оценку хи-квадрат слишком грубой. Далее при оценках используются размеры выборки 200 и 400.

Подбор параметров фильтрации удобнее начать с анализа доли пропущенных значений, так как она зависит в основном от порога d и числа интервалов фильтрации F. На Рис. 4 показана зависимость доли пропусков от порога при n = 400. Для каждого числа интервалов F (16 и 32) кривые от разных датчиков близки и сливаются, поэтому они заменены одной кривой, проведенной по середине полосы, содержащей исходные кривые. Отклонение не превышает 1.6% (в единицах оси ординат).

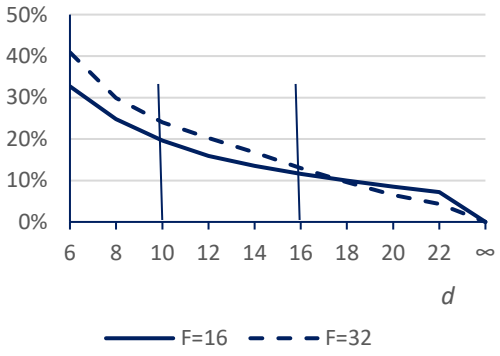

Рис. 4. Зависимость доли пропусков от порога при F=16 и 32.

Здесь видно, что приемлемые величины порога d от 10 до 16: выше этих значений фильтрация мало влияет, а ниже слишком много отбрасывает. На этом отрезке сравнение кривых показывает, что при F=16 доля пропусков меньше.

Далее нужно выбрать тип датчика и число интервалов фильтрации. Оценки делались для всех типов датчиков по каждому числу интервалов хи-квадрат k из набора {10, 12, 14, 18, 20, 22} и затем осреднялись. Основной интерес представляли значения порога d от 10 до 16, выбранные выше. На этом отрезке датчик MT19937 дает лучшие результаты для размеров выборки 200 и 400 при F=16 и F=32, причем при F=16 равномерность выше.

Рис. 5 показывает повышение равномерности распределения за счет фильтрации с выбранными параметрами F=16 и d=10. Здесь показана зависимость доли приемлемых выборок от числа интервалов критерия хи-квадрат k. Приемлемыми считаются выборки, у которых оценка χ^2 отвечает равномерному распределению с уровнем значимости 90%. Результаты получены на 500 выборках, длина выборки 200 и 400.

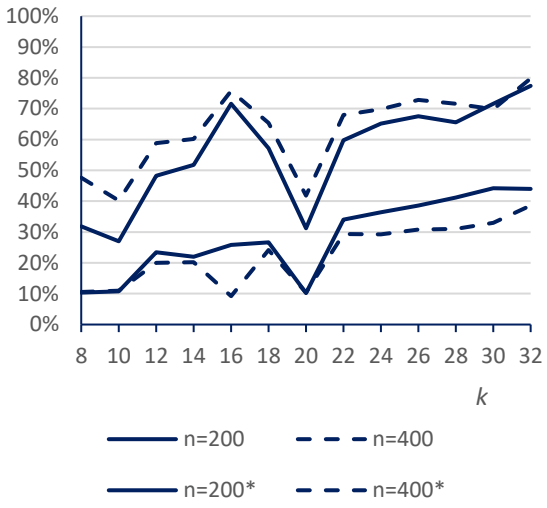

Рис. 5. Зависимость доли приемлемых выборок от числа интервалов в оценке хи-квадрат для датчика MT19937 на 500 выборках, длина выборки 200 и 400. Две верхние кривые с фильтрацией при F=16 и d=10. Две нижние кривые (отмечены *) без фильтрации.

На Рис. 5 видно, что фильтрация повышает долю выборок с равномерным распределением, оцениваемую по критерию хи-квадрат, более чем вдвое. Результат обычно зависит от исходной степени равномерности. Кроме того, влияет специфика цифровой генерации ПСЧ: все датчики во всех случаях дают ухудшение оценок при k=20. Это может быть связано с особенностями перемешивания битов в датчиках.

4. Вероятностные оценки

Оценки влияния фильтрации позволяют убедиться, что фильтрация не ухудшает свойства выборок (помимо равномерности), существенные при генерации тестов. В примерах использовался датчик MT19937.

Пусть датчик ПСЧ моделирует непрерывную случайную величину X, равномерно распределенную на полуинтервале [-1/2, 1/2). Разобьем полуинтервал на k одинаковых частей. Для оценки влияния фильтрации выберем k равным числу интервалов фильтрации F. Значения X представим в виде $x = y + z$, где

$$y = \left[\left(x + \frac{1}{2}\right) \cdot k\right] / k - \frac{1}{2}, \quad z = x - y \quad (4.1)$$

Здесь квадратные скобки обозначают целую часть. Будем рассматривать y и z как значения случайных величин Y и Z, соответственно. Дискретная случайная величина Y принимает значения Y_j, j = 1, 2,... k. В разбиении полуинтервала

[-1/2, 1/2) на равные части Y_j являются координатами середин частей. Значения непрерывной случайной величины Z лежат на полуинтервале [-1/2k, 1/2k). Можно показать, что значения Y равновероятны, а Z распределена равномерно. Случайные величины Y и Z независимы, поэтому математическое ожидание и дисперсия X могут быть получены сложением соответствующих величин для Y и Z. M[Y] = M[Z] = 0, D[X] = 1/12, D[Z] = 1/12k^2. Дисперсию D[Y] можно вычислить как разность D[X] - D[Z] либо непосредственно.

Выборку на выходе датчика ПСЧ обозначим $\{x_i\}$, i=1, 2,... n. Величины x_i моделируют значения случайной величины X. Из выборки $\{x_i\}$ с помощью (4.1) получим выборки $\{y_i\}$и $\{z_i\}$ для случайных величин Y и Z, соответственно. В отсутствие фильтрации можно принять, что выборки содержат независимые значения случайных величин X, Y и Z. Также нет зависимости между Y и Z. Это обеспечивается качеством датчика и может быть проверено эмпирически.

Фильтрация состоит в том, что из выборки $\{y_i\}$ удаляется очередной элемент, нарушающий ограничение $|n_i - n_j| \le d$, где n_j - накопленное количество значений Y_j в выборке, j = 1, 2, ... k, d - порог фильтрации. Соответственные элементы удаляются также из выборок $\{x_i\}$и $\{z_i\}$. В результате значения y_i. становятся зависимыми от предыдущих. Как показано далее, зависимость возникает и усиливается по мере роста размера выборки n. Поскольку фильтрация не зависит от значений z_i, для них пропуск значений является случайным, а оставшиеся значения по-прежнему независимы.

По свойствам датчика ПСЧ, оценки математического ожидания и дисперсии для выборок $\{x_i\}$, $\{y_i\}$ и $\{z_i\}$ должны сходиться к соответствующим значениям для случайных величин X, Y и Z, соответственно. При включении фильтрации происходит выравнивание частот n_j, что усиливает сходимость оценок для Y к теоретическим значениям. Фильтрация не влияет на статистические свойства выборок Z.

Влияние взаимозависимости значений y_i при фильтрации можно проследить на примере оценки математического ожидания X, $\tilde{m}(\mathrm{X})$. Эта случайная величина равна среднему арифметическому n одинаково распределенных случайных величин X. В отсутствие фильтрации принимается, что указанные случайные величины независимы. В этом случае дисперсия $\mathrm{D}[\tilde{m}(\mathrm{X})]$ равна дисперсии D[X] (т.е. 1/12), деленной на n. На Рис. 6 показано отношение оценки дисперсии $\mathrm{D}[\tilde{m}(\mathrm{X})]$ для серии из 200 выборок к теоретическому значению 1/12n (кривая d=∞). Это отношение близко к единице в широком диапазоне изменения n.

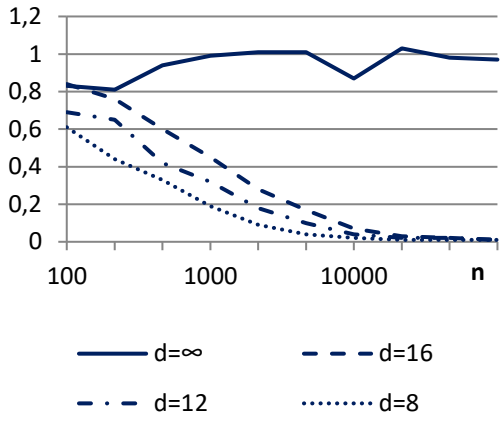

Рис. 6. Отношение оценки дисперсии $\mathrm{D}[\tilde{m}(\mathrm{X})]$ для серии выборок к значению 1/12n при разных порогах фильтрации.

При фильтрации пропуск значений зависит от совокупности значений, выданных датчиком ПСЧ ранее. Тем самым, значения в выборке не будут независимыми. Как показано на Рис. 6, при этом оценка дисперсии $\mathrm{D}[\tilde{m}(\mathrm{X})]$ убывает с ростом n быстрее, чем при независимости значений. Характер этого убывания можно оценить, используя представление X в виде суммы Y + Z (см. выше). Так как значения Z независимы, дисперсию $\mathrm{D}[\tilde{m}(\mathrm{Z})]$ можно считать близкой к теоретическому значению $1/12nk^2$. Вычислим также оценку $\tilde{m}(\mathrm{Y})$, дисперсию которой нужно найти:

$$\tilde{m}(\mathrm{Y}) = \frac{1}{n}\sum_{j=1}^{k} n_j Y_j =$$

$$= \frac{1}{n}\sum_{j=1}^{k}(n_j - n_{min})Y_j \qquad (4.2)$$

Здесь, как и выше, n_j - число значений Y_j в выборке, n_{min} –минимум n_j для j=1, 2,,...k. Сумма $n_{min}Y_j$ нулевая в силу четности набора Y_j. Из-за ограничения $0 \le n_j - n_{min} < d$ сумма в (4.2) ограничена, и оценка $\tilde{m}(\mathrm{Y})$ убывает как 1/n. Так как Y и Z взаимно независимы, дисперсия $\mathrm{D}[\tilde{m}(\mathrm{X})]$ равна сумме $\mathrm{D}[\tilde{m}(\mathrm{Y})]$ и $\mathrm{D}[\tilde{m}(\mathrm{Z})]$, откуда получим:

$$\mathrm{D}[\tilde{m}(\mathrm{X})] = \mathrm{M}\left[\left(\tilde{m}(\mathrm{Y})\right)^2\right] + \mathrm{D}[\tilde{m}(\mathrm{Z})] =$$

$$= \frac{C}{n^2} + \frac{1}{12nk^2}, \qquad (4.3)$$

где $C = C\,(n, d)$ – ограниченная функция, равная дисперсии сумм n_jY_j. Значения C можно получить непосредственно из (4.3), подставляя оценки $\mathrm{D}[\tilde{m}(\mathrm{X})]$ из численных экспериментов. Эти оценки для датчика MT19937 показаны на Рис.7.

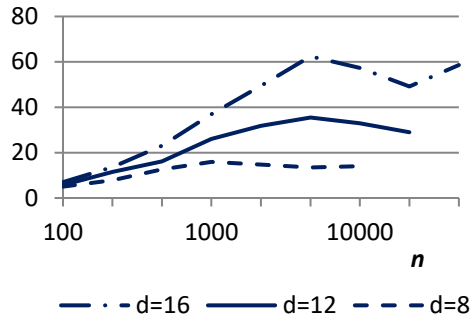

Рис. 7. $C(n, d)$, дисперсия оценки сумм дискретной случайной величины Y для разных значений порога фильтрации d.

На Рис. 7 видно, что $C(n, d)$ сначала растет с n, затем выходит на некоторый уровень, приближенно пропорциональный d^2. Чем больше порог фильтрации d, тем меньше ограничение. В отсутствие фильтрации ограничения нет, и дисперсия оценки сумм дискретной составляющей оказывается пропорциональной $1/n$, а не $1/n^2$. На этом примере можно видеть, как влияет взаимозависимость элементов выборки.

Рассмотрим также частоту выдачи пар одинаковых значений для дискретного равномерного распределения на целочисленном отрезке небольшой длины. При независимых значениях в выборке вероятность совпадения с предыдущим значением равна $1/k$, где k - число возможных значений. Будем анализировать произведение полученной частоты таких пар p на k, $T = p \cdot k$, осредненное по серии выборок. В экспериментах число выборок 200, длина выборки 4000. Результаты показывают, что в отсутствие фильтрации T отличается от единицы не более чем на 1% и не зависит от длины отрезка. При F=16 отрезках фильтрации и для порога фильтрации d=12 T меняется от 0.97 до 0.95 при изменении длины отрезка n от 8 до 32. С ростом порога (и с ослаблением фильтрации) T приближается к единице. Таким образом, при фильтрации частота пар одинаковых значений снижается на несколько процентов. Это объясняется устройством механизма фильтрации: у пары одинаковых значений вероятность превысить порог выше, чем у одиночного значения. При этом второе значение пропускается, и пары больше нет.

5. Заключение

Наиболее известные датчики случайных чисел могут давать недостаточно равномерное распределение для генерации случайных тестов. Повысить равномерность для любого датчика может фильтрация на основе выравнивания гистограммы выдаваемых чисел. Тем самым снижается их независимость, однако это не ухудшает статистические свойства выборок. Подбор параметров фильтрации позволяет получить нужную равномерность распределения при относительно небольшой доле пропусков.

Публикация выполнена в рамках государственного задания ФГУ ФНЦ НИИСИ РАН по теме № FNEF-2024-0003 "Методы разработки аппаратно-программных платформ на основе защищенных и устойчивых к сбоям систем на кристалле и сопроцессоров искусственного интеллекта и обработки сигналов".

6. Приложение. Схемы использованных датчиков ПСЧ

Линейный конгруэнтный генератор (LCG): рекурсивный, вырабатывает 32-битовое целое с помощью 64-битовых операций:

```
t = 214013 * x[i] + 2531011;
x[i+1] = (t ^ (t >> 15)) & 0xFFFFFFFF
```

Генератор XorShift: также рекурсивный, вырабатывает 32-битовое целое, операции 32-битовые. Первая версия:

```
s = x[i] ^ (x[i] << 13);
t = s ^ (s >> 17);
x[i+1] = t ^ (t << 5)
```

Во второй версии алгоритм использует первое и последнее из четверки ранее выработанных чисел x_i, x_{i-1}, x_{i-2}, x_{i-3}.

```
s = x[i-3] ^ (x[i-3] << 11);
t = s ^ (s >> 8);
x[i+1] = (t ^ x[i]) ^ (x[i] >> 19)
```

В этой версии для инициализации нужны четыре числа. Одно из них задавалось "зерном", три других получались умножением "зерна" на простые числа 8179, 8191, 8209.

Генератор Mersenne Twister: первая версия взята из www.agner.org/random в 2003 году. Она использует циклический буфер z размером K=17 пар 32-битовых чисел. На каждом шаге текущая позиция i в буфере обновляется и затем циклически сдвигается назад.

```
struct {int x, y;} z [K];
t = _lrotl (z[i].x, 19) + z[i+10].x;
z[i].x = _lrotl (z[i].y, 17) +
         z[i+10].y;
z[i].y = t;
i = i-1 при i>0, иначе i = K-1.
```

Функция _lrotl() выполняет побитовый цик-

лический сдвиг влево. Таким образом, на каждом шаге генерируется пара ПСЧ, но для выработки 32-битового целого используется только второе (т.е. *t*). Пара ПСЧ используется при выработке числа с плавающей точкой двойной точности.

Вторая версия генератора Mersenne имеет обозначение MT19937 и использует циклический буфер *x*, содержащий *K*=624 32-битовых числа. На каждом шаге текущая позиция *i* в буфере обновляется и затем циклически сдвигается вперед. Перед выдачей числа делается дополнительное перемешивание его битов, т.н. "закалка".

Обновление текущей позиции:

```
j=(i+1)mod K, m=(i+397)mod K

t = (x[i] & 0x80000000) |
    (x[j] & 0x7FFFFFFF);
Если t&1 == 0, t =  t >> 1;
Иначе t = (t >> 1) ^ 0x9908B0DF;
x[i] = t = t ^ x[m];
```

Сдвиг текущей позиции:

```
i = j;
```

"Закалка":

```
t = t ^ (t >> 11);
t = t ^ ((t << 7)  & 0x9D2C5680);
t = t ^ ((t << 15) & 0xEFC60000);
t = t ^ (t >> 18);
```

Improving the Uniformity of Pseudorandom Numbers

A.S. Koutsaev

Abstract. The quality of random tests generation depends on the choice of random number generator. Testing popular generators using the Pearson chi-square test shows that a uniform distribution with a significance level of 70% or higher is observed for only half of the samples. For this reason, to obtain the necessary coverage, you need to increase the volume of tests. A method is proposed to increase the uniformity of distribution for random number generators, based on filtering the sample. Selection of filter parameters allows you to obtain the desired uniformity of distribution with a moderate number of rejects.

Keywords: random number generator, uniform distribution, chi-square, filtering, random tests.

Литература

1. D. H. Lehmer, Mathematical methods in large-scale computing units, Proceedings of a Second Symposium on Large-Scale Digital Calculating Machinery, 1949, Harvard University Press, Cambridge, Mass., 1951, P. 141-146. MR 0044899 (13,495f)

2. G. Marsaglia. Xorshift RNGs. Journal of Statistical Software, 2003, Vol. 8, P. 1-6.

3. www.agner.org/random.

4. M. Matsumoto, T. Nishimura. Mersenne twister: a 623-dimensionally equidistributed uniform pseudo-random number generator. ACM Transactions on Modeling and Computer Simulation, 1998, Vol. 8 (1), P. 3-30.

5. F.O. Panneton, P. l'Ecuyer, M. Matsumoto. Improved long-period generators based on linear recurrences modulo 2. ACM Transactions on Mathematical Software, 2006, 32 (1), P. 1-16.

6. Е.С. Вентцель, Л.А. Овчаров. Теория вероятностей и ее инженерные приложения. М., "Наука", 1988.