ОБОЗРЕНИЕ
ПРИКЛАДНОЙ И ПРОМЫШЛЕННОЙ
Том 21 МАТЕМАТИКИ Выпуск 5
2014

А. В. Паршин (Москва, ТВП). **О расширении набора статистических методов анализа физических датчиков случайных чисел.**

В настоящее время физические датчики случайных чисел (далее ФДСЧ) приобретают все большее значение применительно к защите информации ограниченного доступа (конфиденциальной). Невозможность повторения выходной случайной последовательности ФДСЧ позволяет обеспечить более качественную защиту информации по сравнению с программными датчиками случайных чисел.

Согласно нормативным документам, ФДСЧ не может быть использован для защиты конфиденциальной информации без разработки его математической модели и проверки соответствия указанной модели реальным физическим процессам в ФДСЧ.

Пояснить эти аспекты анализа ФДСЧ уместно на следующем примере.

Рассмотрим следующую математическую модель функционирования одного из экспериментальных образцов ФДСЧ:

- в ходе первичной оцифровки физического случайного процесса, на базе которого реализован ФДСЧ, вырабатывается элемент последовательности независимых одинаково распределенных случайных чисел $\{\mu_i\}$, имеющих биномиальное распределение со значением числа испытаний $m = 7$ и неизвестной вероятностью успеха p, близкой по значению к $1/2$;
- для формирования одного знака двоичной последовательности γ_i значение элемента последовательности $\{\mu_i\}$ приводится по модулю 2.

Проверка соответствия математической модели функционирования реальным физическим процессам в ФДСЧ проводилась с использованием выборок последовательности $\{\mu_i\}$ суммарным объемом 10^{12} значений и следующих статистических критериев.

1. Для проверки согласия выборочного распределения элементов случайной последовательности $\{\mu_i\}$ с биномиальным законом использовался критерий хи-квадрат проверки гипотезы о виде дискретного распределения.

2. Для проверки гипотезы о независимости элементов случайной последовательности $\{\mu_i\}$ использовались.

А) критерий хи-квадрат;

Б) вычисление выборочного коэффициента корреляции.

3. Дополнительно проверялось отсутствие в последовательности $\{\mu_i\}$ трендов с использованием критерия хи-квадрат для проверки гипотезы об однородности выборок.

В ходе проведенных исследований были получены следующие результаты:

1. При справедливости гипотезы H_0 : «элементы последовательности $\{\mu_i\}$ имеют распределение $Bi(7, 1/2)$» вычисленная статистика имеет распределение хи-квадрат с числом степеней свободы, равным 7. На суммарном объеме выборки 10^{12} значений указанная статистика приняла значение 9,518, которое соответствует уровню значимости (вероятности ошибки первого рода), равному 0,218.

Поскольку истинное значение p могло отличаться от $1/2$, дополнительно по выборке была вычислена точечная оценка параметра p по методу «минимума хи-квадрат». В результате проведенных вычислений была получена оценка $\widehat{p} = \frac{1}{2} - 4 \times 10^{-7}$, для которой значение статистики хи-квадрат приняло значение 4,525 при том же числе степеней свободы. Это значение позволяет говорить о согласии статистических данных с проверяемой гипотезой с уровнем значимости, большим 0,5.

2. Для удобства проведения вычислений по п. 2, связанных с необходимостью маркировки биграмм последовательности $\{\mu_i\}$, анализировались 100 выборок объемом 10^{10} значений каждая.

А) Статистика хи-квадрат критерия проверки гипотезы о независимости соседних значений последовательности $\{\mu_i\}$ на 100 выборках приняла 100 значений, соответствующих разным уровням значимости α_i, $i = 1, 2, \ldots, 100$. При справедливости проверяемой гипотезы элементы выборки $\alpha_1, \alpha_2, \ldots, \alpha_{100}$ должны иметь распределение, *близкое* к $R[0;1]$ (в точности распределение $R[0;1]$ невозможно получить в силу дискретности исходного вероятностного пространства). В результате анализа выборки $\alpha_1, \alpha_2, \ldots, \alpha_{100}$ установлено, что отсутствует преобладание, как значений, близких к 0,5, так и значений, близких к 0 и/или 1, а, следовательно, она близка к выборке из распределения $R[0;1]$.

Б) Вообще говоря, распределение выборочного коэффициента корреляции доподлинно известно только для исходной выборки, имеющей нормальное распределение, а для анализируемых выборок из последовательности $\{\mu_i\}$ сделать вывод о значимом отличии любых полученных значений от нуля невозможно. В результате вычисления выборочного коэффициента корреляции между соседними значениями последовательности $\{\mu_i\}$ на 100 выборках объемом 10^{10} значений каждая установлено, что все значения принадлежат интервалу $[-3,6 \times 10^{-5}; 3,0 \times 10^{-5}]$. Полученный разброс значений имеет тот же порядок, что и среднеквадратичное отклонение выборочного коэффициента корреляции между нормальными случайными величинами для объема выборки 10^{10} значений.

3. Для проверки гипотезы об однородности статистических данных каждая из 100 выборок объемом 10^{10} значений дополнительно разбивалась на 100 выборок объемом 10^8 значений каждая. Было вычислено 100 значений статистики хи-квадрат критерия проверки гипотезы об однородности. Дальнейший анализ полученный выборки из 100 значений проводился по аналогии с п. 2 А, путем сведения полученной выборки к выборке из значений уровня значимости $\alpha_1, \alpha_2, \ldots, \alpha_{100}$, и также показал близость распределения элементов выборки $\alpha_1, \alpha_2, \ldots, \alpha_{100}$ к распределению $R[0;1]$, а, следовательно, «хорошее» согласие статистических данных с проверяемой гипотезой.

На этом предполагалось окончание исследований с подготовкой положительного заключения о соответствии ФДСЧ его математической модели.

Проведение дополнительных исследований последовательности $\{\mu_i\}$, связанных с использованием статистик типа «выборочная энтропия» было обусловлено проведенными на указанном экспериментальном образце исследованиями, направленными на проверку возможности использования всей случайности, содержащейся в элементах последовательности $\{\mu_i\}$. Здесь и далее наблюдаемая последовательность $\{\mu_i\}$ трактовалась как выборка значений некоторой случайной величины μ.

Как известно, количественной мерой случайности, содержащейся в распределении случайной величины μ, является энтропия $H(\mu)$, значение которой может быть оценено по выборочной энтропии:

$$\widehat{H}(\mu) = -\sum_{j=j_{\min}}^{j_{\max}} \frac{\nu_j}{N} \log_2 \frac{\nu_j}{N}, \tag{1}$$

где N — объем выборки, $j_{\min}$, $j_{\max}$ — минимальное и максимальное значения, принимаемые случайной величиной μ, ν_j — частота встречаемости значения j в выборке.

Коротко о проведенных исследованиях и их результатах.

В ходе исследований попытки использования всей случайности, содержащейся в элементах последовательности $\{\mu_i\}$, осуществлялись путем применения к элементам данной последовательности преобразования, ставящего в соответствие каждому значению последовательности $\{\mu_i\}$ отрезок двоичной последовательности некоторой длины (кодирования). Для проведения исследований использовалась выборка полного объема 10^{12} значений случайных величин $\{\mu_i\}$. Для данной выборки значение выборочной энтропии $\widehat{H}(\mu)$ составило 2,44664064.

Преобразование указанной выборки проводилось путем обработки файла, содержащего элементы выборки, программой, реализующей кодирование по Хаффману, характеризующееся минимальной избыточностью. В результате применения кодирования:

- из 10^{12} значений последовательности $\{\mu_i\}$ было получено примерно $2,5\times 10^{12}$ значений двоичной последовательности (избыточность кода примерно равна 0,05336);
- для полученной двоичной последовательности отклонение от $1/2$ выборочной вероятности появления знака «0» составило $3,1 \times 10^{-3}$.

При использовании для получения двоичной последовательности приведения элементов последовательности $\{\mu_i\}$ по модулю 2 отклонение выборочной вероятности появления знака «0» от $1/2$ на объеме 10^{12} значений составило $6,5\times 10^{-7}$, что означает статистическую неотличимость (по крайней мере, по указанной статистике) выработанной двоичной последовательности от равновероятной и независимой.

В свете проведенных дополнительных исследований возникла идея о проверке соответствия ФДСЧ его математической модели по пп. 1–3 с помощью статистик, представляющих собой выборочную энтропию. В работе Г. П. Башарина «О статистической оценке энтропии последовательности независимых случайных величин» показано, что выборочная энтропия $\widehat{H}$ имеет асимптотически нормальное распределение с параметрами:

$$E\widehat{H} = H - \frac{s-1}{SN}\log_2 e + O\Big(\frac{1}{N^2}\Big),$$

$$D\widehat{H} = \frac{1}{N}\Big(\sum_{j=1}^{s} p_j(\log_2 p_j)^2 - H^2\Big) + O\Big(\frac{1}{N^2}\Big) \qquad (2)$$

где N — объем выборки, s — число значений, принимаемых случайной величиной.

Непосредственное использование формулы (2) затруднительно, поскольку неизвестны точные значения ни самой энтропии H, ни вероятностей p_j. Для построения предварительной оценки в формуле для дисперсии $D\widehat{H}$ вместо H использовалась $\widehat{H}$ из формулы (1), а вместо p_j их оценка ν_j/N. В результате вычислений получено:

$$D\widehat{H} \approx \frac{4,39}{N} + O\Big(\frac{1}{N^2}\Big),$$

3. Однородность выборок проверялась опосредованно, путем анализа значений выборочной энтропии по ста выборкам объема 10^{10} значений. Все значения попали в отрезок [2,4466188; 2,4466705]. Разброс значений, равный $5,2 \times 10^{-5}$, меньше трех среднеквадратичных отклонений, а, следовательно, серьезные изменения в выборочных распределениях отсутствуют.

2. Так же опосредованно проверялась независимость соседних значений последовательности $\{\mu_i\}$. Одномерная выборочная энтропия умножалась на 2 и сравнивалась с выборочной энтропией биграмм последовательности $\{\mu_i\}$, замаркированных без зацепления. Существенное различие свидетельствовало бы о наличии зависимости, но максимальная разность по 100 выборкам составила $1,2 \times 10^{-8}$.

1. Наиболее интересный результат был получен при попытке построить точечную оценку неизвестного параметра биномиального распределения $Bi(7,p)$, ис-

ходя из значения выборочной энтропии $\widehat{H}(\mu)$. Нетрудно установить (например, путем непосредственного вычисления), что максимальное значение энтропии $H(\mu)$ для $\mu \sim Bi(7,p)$ достигается на значении параметра $p = 1/2$ и равно 2,44663975, что меньше 2,44664064 — значения, вычисленного на общем объеме 10^{12} значений выборки. Вообще говоря, для биномиального распределения не существует таких значений вероятности успеха p, для которых энтропия случайной величины, имеющей распределение $Bi(7,p)$, составляла бы 2,44664064. Однако разница значений составляет всего лишь $8,9 \times 10^{-7}$, что для объема выборки $N = 10^{12}$ значений меньше одного среднеквадратичного отклонения случайной величины $\widehat{H}(\mu)$, а значит, допустимо.

Вывод. Статистические данные, полученные с образца ФДСЧ, согласуются с математической моделью ФДСЧ, что подтверждается как априори выбранными классическими статистическими критериями, так и приведенными рассуждениями, основанными на результатах вычисления выборочной энтропии.