МИНИМАЛЬНАЯ ДЛИНА БИТОВЫХ ПОСЛЕДОВАТЕЛЬНОСТЕЙ ДЛЯ АСИМПТОТИЧЕСКИХ СТАТИСТИЧЕСКИХ ТЕСТОВ В КРИПТОГРАФИИ

Касьянов А.В.[1], Витчак И.Д.[2], Еремук В.В.[3]

Ключевые слова: *статистическое тестирование, случайные битовые последовательности, асимптотические приближения, теорема Берри — Эссеена, ошибка аппроксимации, размер выборки.*

Аннотация

Цель исследования: *определение минимальных размеров битовых случайных последовательностей, генерируемых (псевдо-)случайными генераторами, необходимых для корректного применения статистических тестов на случайность в криптографических приложениях, а также обоснование применения статистик, имеющих асимптотическую природу, в этих тестах.*

Методы исследования: *в исследовании использован метод теоретического анализа, основанный на применении теоремы Берри — Эссеена для оценки скорости сходимости распределения выборочного среднего к нормальному распределению. Также проведен сравнительный анализ источников литературы для отслеживания истории изменения границ коэффициента, используемого для оценки ошибки аппроксимации.*

Результаты исследования: *проведенное исследование установило, что для обеспечения точности аппроксимации распределения выборочного среднего исследуемой битовой случайной последовательности нормальным распределением на уровне 4—5 знаков после запятой, что необходимо для корректного применения статистических тестов, требуется размер выборки не менее n>550000 бит. Это означает, что для надежной оценки качества генераторов (псевдо-)случайных чисел, используемых в криптографии, длина генерируемых последовательностей должна быть не менее указанного значения, чтобы минимизировать ошибку, возникающую из-за асимптотической природы используемых статистик. Результаты исследования показали, что при меньших размерах последовательностей статистические тесты могут давать некорректные результаты, приводя к ложным выводам о случайности или неслучайности последовательности. Это критически важно для обеспечения надежности криптографических приложений, основанных на использовании генераторов (псевдо-)случайных чисел. Исследование также выявило историческую динамику изменения границ коэффициента C_0, используемого в теореме Берри — Эссеена для оценки ошибки аппроксимации. Анализ показал, что значения коэффициента C_0 уточнялись и корректировались с течением времени, что является важным аспектом для практического применения теоремы. Полученные результаты позволяют обосновать использование асимптотических статистик при тестировании на случайность, а также предоставляют практические рекомендации для определения минимальной длины последовательности при проведении такого тестирования. Это позволяет избежать ошибок и повысить надежность криптографических приложений.*

Научная новизна: *в работе впервые обоснованы границы размеров битовых случайных последовательностей, генерируемых (псевдо-)случайными генераторами, необходимые для корректного применения статистических тестов на случайность с использованием теоремы Берри — Эссеена. Полученные результаты предоставляют практические рекомендации для определения минимальной длины последовательности при проведении тестирования, что является значимым вкладом в область криптографического тестирования.*

DOI: 10.24412/1994-1404-2025-4-79-85

[1] ***Касьянов Александр Владимирович,*** *аспирант, Санкт-Петербургский государственный электротехнический университет «ЛЭТИ» им. В.И. Ульянова (Ленина), Санкт-Петербург, Россия. ORCID: 0009-0000-3319-8357, SPIN: 5334-0278.*
E mail: kasjanov@inbox.ru

[2] ***Витчак Иван Дмитриевич,*** *аспирант, ФГАОУ ВО «Национальный исследовательский университет «Высшая школа экономики», Санкт-Петербург, Россия. ORCID: 0009-0002-5411-7477.*
E mail: intdx@yandex.ru

[3] ***Еремук Владимир Вадимович,*** *аспирант, ФГАОУ ВО «Национальный исследовательский университет ИТМО», Санкт-Петербург, Россия. ORCID: 0000-0003-2337-0015.*
E mail: polar.vl@yandex.ru

Введение

Определение длины случайной битовой последовательности, достаточной для обоснованного применения статистических тестов на случайность, представляет собой нетривиальную задачу. Трудности возникают при использовании коротких последовательностей, поскольку многие статистические тесты, основанные на асимптотических приближениях предельных распределений, требуют существенно больших объёмов данных для получения достоверных результатов. Асимптотические приближения, используемые для вывода свойств распределения статистик, недостаточно точны при малых длинах последовательностей, что делает выводы о случайности ненадежными. Несмотря на значимость этого вопроса для криптографии, где генераторы (псевдо-) случайных последовательностей (далее Г(П)СЧ) играют важную роль, вопрос выбора оптимальной длины последовательности для проведения статистического тестирования случайных битовых последовательностей, используемых в криптографических приложениях, требует дальнейшего исследования. На данный момент отсутствуют достаточно обоснованные рекомендации по выбору длины исследуемых последовательностей.

В данной работе исследуются границы длин битовых случайных последовательностей, вырабатываемых генераторами, с целью обоснования применения статистических тестов на случайность с асимптотическим поведением. Анализ имеющейся литературы [1-7] выявил недостаточную проработку этого вопроса в контексте криптографических приложений. Например, авторы работ [1, 7] предлагают достаточным для проведения статистических тестов размер выборки более 100 бит, [3] — более 1000 бит, рассматривая примеры на длине 100 бит. Главным критерием для определения размеров выборки для данных работ является уровень значимости α; анализ ошибок аппроксимации между распределениями в данных работах не проводится.

Целью работы является определение критериев выбора длины последовательности, позволяющих надёжно применять статистические тесты, основанные на асимптотических приближениях, для оценки качества генераторов, используемых в криптографических протоколах и приложениях с уровнем значимости $\alpha = 0{,}01$. Это исследование позволит установить практические рекомендации по оценке качества и надёжности Г(П)СЧ, что будет способствовать повышению безопасности криптографических систем.

Исследование границ

Рассмотрим бесконечную последовательность $X_1, X_2, \ldots$ независимых и одинаково распределенных случайных величин с конечным математическим ожиданием $M(X_1) = M(X_1) = \cdots = \mu$, и рассмотрим $\bar{X}_n = \frac{1}{n}(X_1 + X_2 + \ldots + X_n), n \in \mathbb{N}$.

Пусть $\forall i, D(X_i) = \sigma^2 < +\infty$, независимость случайных величин не предполагает корреляции между ними, следовательно мы имеем:

$$D(\bar{X}_n) = D\left(\frac{1}{n}(X_1 + X_2 + \ldots + X_n)\right) = \frac{1}{n^2} D(X_1 + X_2 + \ldots + X_n) = \frac{1}{n^2}\sum_{i=1}^{n} D(X_i) = \frac{n \cdot \sigma^2}{n^2} = \frac{\sigma^2}{n}. \quad (1)$$

Математическое ожидание случайной величины $\bar{X}_n$ представляет собой среднее значение выборочного среднего:

$$M(\bar{X}_n) = \mu. \quad (2)$$

Используя неравенство Чебышева для $\bar{X}_n$, получаем, что вероятность:

$$P(|\bar{X}_n - \mu| \geq \varepsilon_1) \leq \frac{\sigma^2}{n \cdot \varepsilon_1^2}, \quad (3)$$

где ε_1 — ошибка аппроксимации.

Это неравенство используем для получения следующего:

$$P(|\bar{X}_n - \mu| < \varepsilon_1) = 1 - P(|\bar{X}_n - \mu| \geq \varepsilon_1) \geq 1 - \frac{\sigma^2}{n \cdot \varepsilon_1^2}. \quad (4)$$

Так как $\forall i, D(X_i) = \sigma^2 < +\infty$ и $\varepsilon_1 = \text{const}$ тогда:

$$1 \geq P(|\bar{X}_n - \mu| < \varepsilon_1) \geq 1 - \frac{\sigma^2}{n \cdot \varepsilon_1^2} \xrightarrow[n\to\infty]{} 1. \quad (5)$$

Чтобы понять с какого n мы можем использовать закон больших чисел, рассмотрим различные значения ε_1, также для определения границ определим, что нас будет удовлетворять условие, что вероятность $P(|\bar{X}_n - \mu| < \varepsilon_1) = 0{,}99$.

Но перед этим рассмотрим этот вопрос с точки зрения центральной предельной теоремы (далее по тексту ЦПТ).

Пусть $X = \{X_1, X_2, \ldots, X_n\}$ — есть последовательность независимых одинаково распределённых случайных величин, имеющих конечные математическое ожидание μ и дисперсию σ^2. Пусть также:

$$S_n = \sum_{i=1}^{n} X_i,$$

тогда по ЦПТ:

$$\frac{S_n - \mu n}{\sigma\sqrt{n}} \xrightarrow[n\to\infty]{} \mathcal{N}(0,1), \quad (6)$$

где $\mathcal{N}(0,1)$ — нормальное распределение с нулевым математическим ожиданием и стандартным отклонением, равным единице. Пусть случайная величина $\bar{X}_n$ имеет распределение Бернулли, следовательно:

$$\mu = p, \sigma^2 = pq, q = 1 - p, \quad (7)$$

Определим выборочное среднее первых n величин как:

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i. \quad (8)$$

Перепишем (6) в другой форме:

$$S_n \xrightarrow[n\to\infty]{} \mathcal{N}(\mu n, n\sigma^2),$$

Поделим обе части на n и применим (7):

$$\bar{X}_n = \frac{S_n}{n} \xrightarrow[n\to\infty]{} \frac{\sigma\sqrt{n}\mathcal{N}(0,1) + \mu n}{n} =$$

$$\frac{\sigma}{\sqrt{n}}\mathcal{N}(0,1) + \mu = \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$$

т. е.

$$\bar{X}_n \xrightarrow[n\to\infty]{} \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right). \quad (9)$$

Перепишем (9) в другой форме:

$$\bar{X}_n = \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right) + o\left(\frac{1}{\sqrt{n}}\right), \quad (10)$$

из (7), (10) следует, что:

$$\bar{X}_n = \mathcal{N}\left(0, \frac{pq}{n}\right) + \mu + o\left(\frac{1}{\sqrt{n}}\right). \quad (11)$$

Вычтем из обоих частей уравнения (11) μ и используя (10) получим:

$$\bar{X}_n - \mu = \mathcal{N}\left(0, \frac{pq}{n}\right) + \mu + o\left(\frac{1}{\sqrt{n}}\right) - \mu =$$

$$= \mathcal{N}\left(0, \frac{p(1-p)}{n}\right) + o\left(\frac{1}{\sqrt{n}}\right), \quad (12)$$

положим,

$$p = \frac{1}{2}. \quad (13)$$

Из (12), (13) получаем, что:

$$\bar{X}_n - \mu = \mathcal{N}\left(0, \frac{1}{4n}\right) + o\left(\frac{1}{\sqrt{n}}\right).$$

Таким образом, для нахождения n при различных значениях ε_1, с граничным условием, что $P(|\bar{X}_n - \mu| < \varepsilon_1) = 0{,}99$ для $\alpha = 0.01$, значение σ^2 следует выбирать:

$$\sigma^2 = \frac{0{,}25}{n}. \quad (14)$$

Предположим, что в (3) $P(|\bar{X}_n - \mu| \geq \varepsilon_1) = 0{,}01$ $P(|\bar{X}_n - \mu| \geq \varepsilon_1) = 0{,}01$, тогда найдем при каких n n формула (3) выполняется. Произведем эту замену в формуле (3), получим:

$$0{,}01 \leq \frac{\sigma^2}{n \cdot \varepsilon_1^2}. \quad (15)$$

Из (14), (15) следует:

$$0{,}01 \leq \frac{0{,}25}{n^2 \cdot \varepsilon_1^2} \Rightarrow n \leq \frac{5}{\varepsilon_1}. \quad (16)$$

Из (16) следует, что при $n \leq \frac{5}{\varepsilon_1}$ выполняется формула (3). И, следовательно, условие того, что:

$$P(|\bar{X}_n - \mu| < \varepsilon_1) = 0{,}99$$

выполняется при условии:

$$n > \frac{5}{\varepsilon_1}. \quad (17)$$

Для примера: при $\varepsilon_1 = 0{,}01$ необходимо, чтобы случайная битовая последовательность содержала $n > 500$ бит; см. рис. 1.

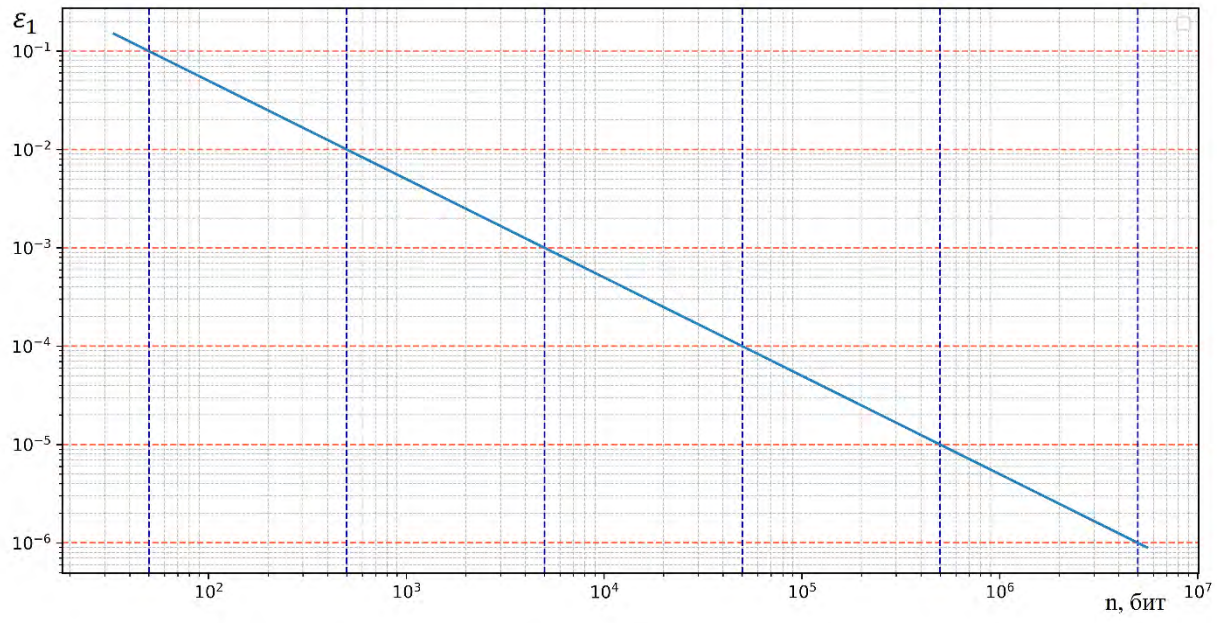

Рис. 1. Логарифмический график зависимости n от ε_1

Теперь рассмотрим данный вопрос с точки зрения теоремы Берри — Эссеена. Формулировки теоремы различаются, поскольку она была независимо открыта двумя математиками, Эндрю К. Берри (в 1941 году) и Карлом-Густавом Эссееном (в 1942 году) [8]. Неравенство Берри—Эссеена дает количественный результат, определяя скорость сходимости и давая границу максимальной ошибки аппроксимации между нормальным распределением и распределением масштабированного выборочного среднего независимых случайных величин. Ошибка аппроксимации определяется расстоянием Колмогорова [9].

Пусть $X_1, X_2, \ldots$ последовательность независимых и одинаково распределенных случайных величин с математическим ожиданием $M(X_i) = 0$, $\forall i, D(X_i) = \sigma^2 > 0, < +\infty, \quad M(|X_i^3|) = \beta_3 < \infty$; для натурального $n \geq 1$ положим:

$$\bar{S}_n = \frac{\sum_{i=1}^{n} X_i}{\sigma\sqrt{n}}, \tag{18}$$

функцию распределения случайной величины $\bar{S}_n$ и стандартную нормальную функцию распределения соответственно обозначим $F_n(x)$ и $\Phi(x)$.

Тогда [3] существует константа C_0 такая, что при всех $n \geq 1$:

$$\sup_{x\in\mathbb{R}} |F_n(x) - \Phi(x)| \leq \frac{C_0\beta_3}{\sigma^3\sqrt{n}}. \tag{19}$$

Рассмотрим, как меняется ошибка аппроксимации ε_2 между нормальным распределением и распределением масштабированного выборочного среднего независимых одинаково распределенных случайных величин в зависимости от размера выборки n:

$$\sup_{x\in\mathbb{R}} |F_n(x) - \Phi(x)| \leq \frac{C_0\widetilde{\beta_3}}{\sigma^3\sqrt{n}} < \varepsilon_2 \tag{20}$$

$$\widetilde{\beta_3} = \begin{cases} \beta_3, & \text{если } \beta_3 \geq \dfrac{3}{\sqrt{2}}, \\ \dfrac{1}{0{,}7804 - 0{,}1457 \cdot \beta_3}, & \text{если } \beta_3 < \dfrac{3}{\sqrt{2}}. \end{cases} \tag{21}$$

В нашем случае $\beta_3 = 1, \sigma = 1$, следовательно:

$$\widetilde{\beta_3} = \frac{1}{0{,}7804 - 0{,}1457} \approx 1{,}5755475, \tag{22}$$

$$C_0 = 0{,}651$$

Данные границы C_0 были получены С. Цалем в [3] (1963 г.).

$$\varepsilon_2 > \frac{1{,}5755475 \cdot 0{,}651}{\sqrt{n}} \approx \frac{1{,}02568142}{\sqrt{n}}.$$

Выразим n:

$$n > \left(\frac{1{,}02568142}{\varepsilon_2}\right)^2. \tag{23}$$

Для примера: при $\varepsilon_2 = 0{,}01$ необходимо, что бы случайная битовая последовательность содержала $n > 10\,520$ бит.

В работе [10] были получены уточненные границы, и доказано, что для всех $n \geq 1$ и всех распределений с нулевым средним, единичной дисперсией и конечным третьим абсолютным моментом β_3 справедливо неравенство:

$$\sup_{x\in\mathbb{R}} |F_n(x) - \Phi(x)| \leq \frac{0{,}34445 \cdot \beta_3 + 0{,}16844}{\sqrt{n}}. \tag{24}$$

Следовательно, при условии $\beta_3 = 1$ верно следующее:

$$\frac{0{,}51289}{\sqrt{n}} < \varepsilon_2,$$

Выразим n:

$$n > \left(\frac{0{,}51289}{\varepsilon_2}\right)^2. \tag{25}$$

Для примера: при $\varepsilon_2 = 0{,}01$ необходимо, что бы случайная битовая последовательность содержала $n > 2630$ бит.

На основе исследований, полученных в работе [11], определена наиболее точная граница для константы C_0:

$$C_0 \leq 0{,}409954, \tag{26}$$

и доказано, что для всех $1 \leq n \leq 500000$ (ограничение связано с тем, что для больших значений в работе [11] проверка не проводилась) и всех распределений с нулевым средним, единичной дисперсией и конечным третьим абсолютным моментом β_3 справедливо неравенство:

$$\sup_{x\in\mathbb{R}} |F_n(x) - \Phi(x)| \leq \frac{C_0 \cdot \beta_3}{\sigma^3 \cdot \sqrt{n}}.$$

Следовательно: $\varepsilon_2 > \frac{C_0 \cdot \beta_3}{\sigma^3 \cdot \sqrt{n}}, \beta_3 = 1, \sigma = 1,$ откуда:

$$n > \left(\frac{0{,}409954}{\varepsilon_2}\right)^2. \tag{27}$$

Для примера: при $\varepsilon_2 = 0.01$ необходимо, что бы случайная битовая последовательность содержала $n > 1680$ бит.

В 1966 году [12] Золотарев выдвинул гипотезу о том, что константа:

$$C_0 = C_E, \text{где } C_E = \frac{3 + \sqrt{10}}{6\sqrt{2\pi}} \approx 0{,}4097321837, \tag{28}$$

откуда:

$$n > \left(\frac{0{,}4097321837}{\varepsilon_2}\right)^2. \tag{29}$$

Для примера: при $\varepsilon_2 = 0{,}01$ необходимо, что бы случайная битовая последовательность содержала $n > 1678$ бит.

В 2005 году в [13] Хипп и Маттнер доказали, что для функций распределений сумм случайных чисел, имеющих симметричное распределение Бернулли, коэффициент:

$$C_0 \leq \frac{1}{\sqrt{2 \cdot \pi}} \approx 0{,}39894228, \tag{30}$$

откуда:

$$n > \left(\frac{0{,}39894228}{\varepsilon_2}\right)^2. \quad (31)$$

Для примера: при $\varepsilon_2 = 0.01$ необходимо, что бы случайная битовая последовательность содержала $n > 1591$ бит, а при $\varepsilon_2 = 0{,}0005$ необходимо, что бы случайная битовая последовательность содержала $n > 636619$ бит; см. рис. 2.

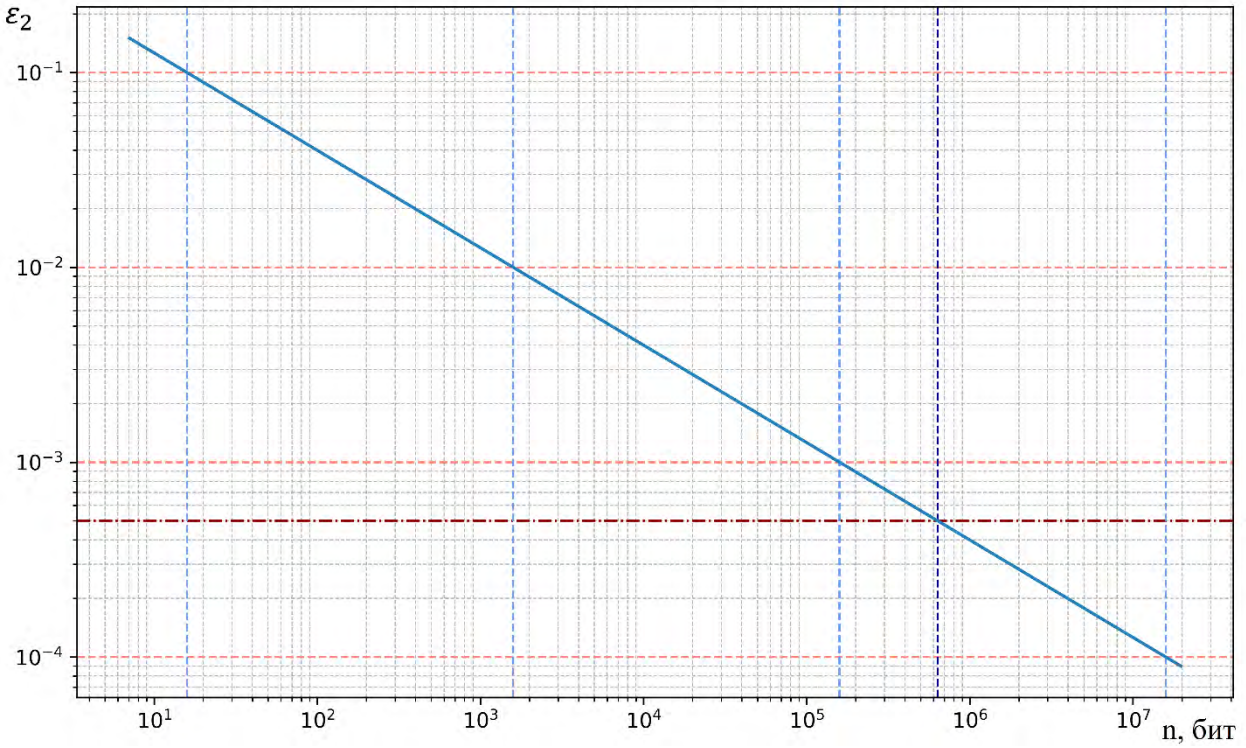

Рис. 2. Логарифмический график зависимости n от ε_2 с использованием границ для коэффициента C_0, полученных Хилл и Маттнером

Заключение

При исследовании генераторов случайных чисел на случайность, как правило, нужна точность 4—5 знаков после запятой, таким образом, чтобы в проводимом статистическом тесте пренебречь ошибкой аппроксимации между нормальным распределением и распределением случайной величины, представляющей собой сумму бит случайной битовой последовательности, размер выборки должен быть не меньше $n > 550000$ бит (если говорить о точности для функций распределений сумм случайных чисел, имеющих симметричное распределение Бернулли, то ошибка ε_2 должна быть $\varepsilon_2 \approx 0{,}0005$, а соответственно размер выборки должен быть не меньше $n > 636619$ бит).

В случае если мы хотим пренебречь разницей между выборочным средним и математическим ожиданием (с точки зрения применимости центральной предельной теоремы, в данном случае мы говорим, что пренебрегаем при $\varepsilon_1 = 0{,}0005$) размер выборки должен быть не меньше $n > 10000$ бит.

Подводя итог данной работы, можно сказать, что исследования на случайность битовых последовательностей необходимо проводить на длинах более $n \geq 5{,}5 \cdot 10^5$ бит.

Полученные результаты предоставляют практические, научно обоснованные рекомендации по корректному применению статистических тестов к генераторам случайных чисел, что способствует повышению надежности и безопасности криптографических приложений. Фактически данное исследование устанавливает четкие количественные границы для размеров выборок (длин битовых случайных последовательностей), которые необходимо использовать при оценке качества генераторов случайных чисел.

Литература

1. NIST/SEMATECH e-Handbook of Statistical Methods. 7.2.4.2. Sample sizes required // U.S. Department of Commerce, 2012. URL: http://www.itl.nist.gov/div898/handbook (дата обращения: 02.12.2024). Текст: электронный.

2. David A. Kenn. Statistics for the social and behavioral sciences // Little, Brown and Company, Boston, 1987. P. 203-223. URL: https://archive.org/details/statisticsforsoc0000kenn (дата обращения: 28.11.2024). Текст: электронный.
3. NIST SP 800-22. Revision 1a. A Statistical Test Suite for Random and Pseudorandom Generators for Cryptographic Applications, April 2010. URL: https://dl.acm.org/doi/pdf/10.5555/2206233. (дата обращения: 05.11.2024). Текст: электронный.
4. Молдовян Н.А. Криптография: от примитивов к синтезу алгоритмов / Н.А. Молдовян, А.А. Молдовян, М.А. Еремеев. СПб. : БХВ-Петербург, 2004. 448 с. ISBN 5-94157-524-6. Текст: непосредственный.
5. Кнут Д.Э. Искусство программирования, том 2. Получисленные алгоритмы / Д.Э. Кнут, 3-е изд.; пер. с англ. М. : Издательский дом «Вильямс», 2001. 832 с. ISBN 5-8459-0081-6 (рус.). Текст: непосредственный.
6. Тихова Г.П. Планируем клиническое исследование. Вопрос № 1: как определить необходимый объем выборки? / Г.П. Тихова. Текст: электронный // Регионарная анестезия и лечение острой боли. 2014. № 3. URL: https://cyberleninka.ru/article/n/planiruem-klinicheskoe-issledovanie-vopros-1-kak-opredelit-neobhodimyy-obem-vyborki (дата обращения: 15.11.2024).
7. Оспанов Р. Статистическое тестирование криптографической хеш-функции, построенной на основе модифицированной схемы sponge / Р. Оспанов, Е. Сейткулов, Н. Сисенов, Б. Ергалиева, Н. Ташатов. Текст: электронный // Вестник КазАТК, 121, 2022, С. 591-599. URL: https://doi.org/10.52167/1609-1817-2022-121-2-591-599 (дата обращения: 02.12.2024).
8. Kolmogoroff A.L. Sulla determinazione empirica di una legge di distribuzione // Giornale dell' Instituto Italiano degli Attuari. 1933. V.4. № 1. P.83-91. URL: https://www.researchgate.net/post/Kolmogorovs_original_paper_Sulla_determinazione_empirica_di_una_legge_di_distribuzione_from_1933 (дата обращения: 02.12.2024). Текст: электронный.
9. Орлов А.И. Непараметрические критерии согласия Колмогорова, Смирнова, омега-квадрат и ошибки при их применении / А.И. Орлов. Текст: электронный // Научный журнал КубГАУ. 2014. № 97. URL: https://cyberleninka.ru/article/n/neparametricheskie-kriterii-soglasiya-kolmogorova-smirnova-omega-kvadrat-i-oshibki-pri-ih-primenenii (дата обращения: 22.10.2024).
10. Королев В.Ю. О верхней оценке абсолютной постоянной в неравенстве Берри — Эссеена. Теория вероятн. и ее примен. / В.Ю. Королев, И.Г. Шевцова. Текст: электронный // Theory Probab. Appl., 54:4 (2010), P. 638—658 — URL: https://www.mathnet.ru/php/getFT.phtml?jrnid=tvp&paperid=3534&what=fullt&option_lang=rus (дата обращения: 22.10.2024).
11. Zolotukhin A. On a bound of the absolute constant in the Berry—Esseen inequality for i.i.d. Bernoulli random variables / A. Zolotukhin, S. Nagaev, V. Chebotarev. Текст: электронный // Modern Stoch. Theory Appl. 5(2018), no. 3, P. 385-410. URL: https://vmsta.org/journal/VMSTA/article/129/text (дата обращения: 22.10.2024).
12. Zolotarev V.M. A sharpening of the inequality of Berry — Esseen. Z. Wahrscheinlichkeitstheor. verw. Geb. 8, 332—342 (1967). MR0221570. https://doi.org/10.1007/BF00531598 (дата обращения: 25.10.2024). Текст: электронный.
13. Hipp C. On the normal approximation to symmetric binomial distributions / C. Hipp, L. Mattner. Текст: электронный // Theory of Probability and its applications. Vol 52(2008), iss. 3, 516-523. https://www.math.uni-trier.de/~mattner/Papers/NormBin20050401.pdf (дата обращения: 25.10.2024).

SECTION:
INFORMATION PROTECTION METHODS AND SYSTEMS

MINIMUM LENGTH OF BIT SEQUENCES FOR ASYMPTOTIC STATISTICAL TESTS IN CRYPTOGRAPHY

Kasyanov A.V.[4], Vitchak I.D.[5], Eremuk V.V.[6]

Keywords: *statistical testing, random bit sequences, asymptotic approximations, Berry-Essen theorem, approximation error, sample size.*

[4] ***Alexander V. Kasyanov,*** *Postgraduate Student, V.I. Ulyanov (Lenin) St. Petersburg Electrotechnical University "LETI", St. Petersburg, Russia. ORCID: 0009-0000-3319-8357, SPIN: 5334-0278.*
E-mail: kasjanov@inbox.ru

[5] ***Ivan D. Vitchak,*** *Ph.D. student, National Research University Higher School of Economics, St. Petersburg, Russia. ORCID: 0009-0002-5411-7477.*
E-mail: intdx@yandex.ru

[6] ***Vladimir V. Eremuk,*** *Ph.D. student, National Research University ITMO, St. Petersburg, Russia. ORCID: 0000-0003-2337-0015.*
E-mail: polar.vl@yandex.ru

Abstract

Purpose of the study: *is to determine the minimum sizes of bit random sequences generated by (pseudo-)random generators necessary for the correct application of statistical tests for randomness in cryptographic applications, and to justify the use of statistics that are asymptotic in nature in these tests.*

Research methods: *the study used a theoretical analysis method based on the application of the Berry-Essen theorem to estimate the rate of convergence of the distribution of the sample mean to the normal distribution. A comparative literature analysis was also carried out to trace the history of variation in the bounds of the coefficient used to estimate the error of approximation.*

Results of the study: *the study conducted established that a sample size of at least n>550000 bits is required to ensure that the distribution of the sample mean of the bit random sequence under study is accurately approximated by a normal distribution at the level of 4—5 decimal places, which is necessary for the correct application of statistical tests. This means that in order to reliably assess the quality of (pseudo-)random number generators used in cryptography, the length of the generated sequences must be at least the specified value in order to minimise the error arising from the asymptotic nature of the statistics used. The results of the study showed that with smaller sequence sizes, statistical tests can produce incorrect results, leading to false conclusions about the randomness or non-randomness of the sequence. This is critical for ensuring the robustness of cryptographic applications based on (pseudo-)random number generators. The study also revealed the historical evolution of the bounds of the C_0 coefficient used in the Berry-Essen theorem to estimate the approximation error. The analysis showed that the values of the C_0 coefficient have been refined and adjusted over time, which is an important aspect for the practical application of the theorem. The results obtained allow us to justify the use of asymptotic statistics in randomness testing and provide practical guidelines for determining the minimum sequence length when performing such testing. This makes it possible to avoid errors and increase the reliability of cryptographic applications.*

Scientific novelty: *in this paper, we first justify bounds on the size of bit random sequences generated by (pseudo-)random generators, necessary for the correct application of statistical randomness tests using the Berry-Essen theorem. The results provide practical guidelines for determining the minimum sequence length for testing, which is a meaningful contribution to the field of cryptographic testing.*

References

1. NIST/SEMATECH e-Handbook of Statistical Methods. 7.2.4.2. Sample sizes required // U.S. Department of Commerce, 2012. URL: http://www.itl.nist.gov/div898/handbook.
2. David A. Kenn. Statistics for the social and behavioral sciences // Little, Brown and Company, Boston, 1987. P.203-223. URL: https://archive.org/details/statisticsforsoc0000kenn.
3. NIST SP 800-22. Revision 1a. A Statistical Test Suite for Random and Pseudorandom Generators for Cryptographic Applications, April 2010. URL: https://dl.acm.org/doi/pdf/10.5555/2206233.
4. Moldovyan N.A. Kriptografiya: ot primitivov k sintezu algoritmov / Moldovyan N.A., Moldovyan A.A., Eremeev M.A. // SPb. : BXV-Peterburg, 2004. 448 s. ISBN 5-94157-524-6
5. D. Knut. Iskusstvo programmirovaniya dlya E`VM. Poluchislenny`e algoritmy`. T. 2. M. : Mir, 1977. 700 s.
6. Tixova G.P. Planiruem klinicheskoe issledovanie. Vopros № 1: kak opredelit` neobxodimy`j ob``em vy`borki? // Regionarnaya anesteziya i lechenie ostroj boli. 2014. № 3. URL: https://cyberleninka.ru/article/n/planiruem-klinicheskoe-issledovanie-vopros-1-kak-opredelit-neobhodimyy-obem-vyborki.
7. Ospanov R. Statisticheskoe testirovanie kriptograficheskoj xesh-funkcii, postroennoj na osnove modificirovannoj sxemy` sponge / Sejtkulov E., Sisenov N., Ergalieva B., Tashatov N. // Vestnik KazATK, 121, 2022, 591-599. URL: https://doi.org/10.52167/1609-1817-2022-121-2-591-599.
8. Kolmogoroff A. Sulla determinazione empirica di una legge di distribuzione // Giornale dell' Instituto Italiano degli Attuari. 1933. V.4. № 1. P.83-91. URL: https://www.researchgate.net/post/Kolmogorovs_original_paper_Sulla_determinazione_empirica_di_una_legge_di_distribuzione_from_1933.
9. Orlov A.I. Neparametricheskie kriterыii soglasiya Kolmogorova, Smirnova, omega-kvadrat i oshibki pri ix primenenii // Nauchny`j zhurnal KubGAU. 2014. № 97. URL: https://cyberleninka.ru/article/n/neparametricheskie-kriterii-soglasiya-kolmogorova-smirnova-omega-kvadrat-i-oshibki-pri-ih-primenenii.
10. Korolev V.Yu., Shevczova I.G. O verxnej ocenke absolyutnoj postoyannoj v neravenstve Berri—E`sseena. Teoriya veroyatn. i ee primen., 54:4 (2009), 671—695; Theory Probab. Appl., 54:4 (2010), 638—658. https://www.mathnet.ru/php/getFT.phtml?jrnid=tvp&paperid=3534&what=fullt&option_lang=rus.
11. Zolotukhin A., Nagaev S., Chebotarev V. On a bound of the absolute constant in the Berry—Esseen inequality for i.i.d. Bernoulli random variables. Modern Stoch. Theory Appl. 5(2018), no. 3, 385-410, https://vmsta.org/journal/VMSTA/article/129/text.
12. Zolotarev V.M. A sharpening of the inequality of Berry — Esseen. Z.Wahrscheinlichkeitstheor. verw. Geb. 8, 332—342 (1967). MR0221570. https://doi.org/10.1007/BF00531598.
13. Hipp C., Mattner L. On the normal approximation to symmetric binomial distributions. Theory of Probability and its applications. Vol 52(2008), iss. 3, 516-523. https://www.math.uni-trier.de/~mattner/Papers/NormBin20050401.pdf