More Powerful and Reliable Second-level Statistical Randomness Tests for NIST SP 800-22

Shuangyi Zhu[1,2,3], Yuan Ma[1,2⋆], Jingqiang Lin[1,2], Jia Zhuang[1,2], and Jiwu Jing[1,2]

[1] Data Assurance and Communication Security Research Center, Chinese Academy of Sciences, Beijing, China
[2] State Key Laboratory of Information Security, Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China
[3] University of Chinese Academy of Sciences, Beijing, China
{zhushuangyi,yma,linjq,jzhuang13,jing}@is.ac.cn

Abstract. Random number generators (RNGs) are essential for cryptographic systems, and statistical tests are usually employed to assess the randomness of their outputs. As the most commonly used statistical test suite, the NIST SP 800-22 suite includes 15 test items, each of which contains two-level tests. For the test items based on the binomial distribution, we find that their second-level tests are flawed due to the inconsistency between the assessed distribution and the assumed one. That is, the sequence that passes the test could still have statistical flaws in the assessed aspect. For this reason, we propose *Q-value* as the metric for these second-level tests to replace the original P-value without any extra modification, and the first-level tests are kept unchanged. We provide the correctness proof of the proposed Q-value based second-level tests. We perform the theoretical analysis to demonstrate that the modification improves not only the detectability, but also the reliability. That is, the tested sequence that dissatisfies the randomness hypothesis has a higher probability to be rejected by the improved test, and the sequence that satisfies the hypothesis has a higher probability to pass it. The experimental results on several deterministic RNGs indicate that, the Q-value based method is able to detect some statistical flaws that the original SP 800-22 suite cannot realize under the same test parameters.

Keywords: Statistical randomness test, NIST SP 800-22, random number generator, P-value

1 Introduction

As essential primitives, random number generators (RNGs) are important for cryptographic systems. The security of many cryptographic schemes and protocols is built on the perfect randomness of RNG outputs. RNGs are classified

⋆ Corresponding author

into two types: pseudo/deterministic and true/non-deterministic random number generators (PRNGs and TRNGs, respectively). In general, TRNGs based on some random physical phenomenons, may be used directly as random bit sources or generate seeds for PRNGs, and PRNGs extend the seeds to produce deterministic long sequences.

For any type of RNG, statistical hypothesis tests have been widely employed to assess the quality of the RNG, which evaluate whether the output sequences fit with the given hypothesis (i.e., the sequence has perfect randomness) or not. In addition, statistical randomness tests are also used to evaluate the outputs of other cryptographic primitives such as hash functions and block ciphers, to preliminarily validate the indistinguishability of their outputs from random mapping. The commonly used statistical test suites, each of which is composed of a serial of test items, include Diehard [7] proposed by Marsaglia and SP 800-22 [11] standardized by US National Institute of Standard and Technology (NIST).

The most commonly used NIST SP 800-22 test suite is composed of 15 test items, and provides comprehensive evaluation for different randomness aspects of assessed sequences. For example, the Frequency Test assesses the uniformity of the sequence, and the Runs Test assesses the transform frequency of 0's and 1's. In the beginning of the testing process, the whole bit sequence is divided into N blocks. In every test item, a test statistic value is computed for each data block. According to the assumed distribution of the test statistic value, 15 test items are divided into two types: *binomial distribution* based (binomial-based for short in this paper) and *chi-square distribution* based (chi-square based for short). Each test item uses its assumed distribution to compute the P-value, which roughly represents the probability that the block is random. A test item is considered to be passed when the computed P-value is larger than the *significance level*. Then, based on the computed N P-values for N blocks, each test item performs two-level tests: the first-level test and the second-level test, where passing the former is the premise to execute the latter. The second-level testing approach was found to increase the testing capability [9]. The first-level test focuses on the passing ratio of the N P-values, and the second-level test further focuses on the uniformity of the N P-values to assess whether the test statistic values follow the expected distribution, i.e., the standard normal distribution[1] or the chi-square distribution. In the remainder of this paper, *the test statistic* refers to the test statistic value that is assumed to follow the standard normal distribution in the binomial-based tests.

Related work. Several papers on the NIST SP 800-22 test suite have been presented in literature. Among the test items, Kim *et al.* [6] analyzed the correctness of the Spectral Test and the Lemple-Ziv Test, and Hamano [4, 5] adjusted the distribution parameters for the Spectral Test and corrected the Overlapping Test. Sulak *et al.* [14] found that the P-values for short sequences (less than 512 bits) follow a specific discrete distribution, rather than the assumed uniform distribution for long sequences. Pareschi *et al.* [9] investigated the reliability of the

[1] For a sufficiently large number of trials, the distribution of the binomial sum after normalizing, is closely approximated by a standard normal distribution [11].

second-level tests, and analyzed the sensitivity to the approximation errors introduced by the computation of P-values. Furthermore, as the sequence length is finite in practice and thus the set of possible statistic values is discrete, Pareschi *et al.* [10] provided the actual distributions of P-values for the Frequency Test, the Runs Test, and the Spectral Test, and evaluated the test errors for different testing methods based on P-values. In our preliminary work [15], we analyzed the correctness and the reliability of the second-level tests in the NIST SP 800-22 test suite.

Our contribution. In this paper, we find that the P-values derived from the binomial-based tests are unqualified for the second-level tests of the NIST SP 800-22 suite, though they are proper to be used for the first-level tests. The P-values in the binomial-based tests are computed, using the absolute values of the test statistics. Therefore, the second-level tests on P-values do not exactly tell whether the test statistics follow the standard normal distribution or not, because we cannot learn from the P-values that the test statistics are positive or negative. In particular, even if P-values follow the uniform distribution on $[0, 1]$, there still exists a non-ignorable probability that the test statistics are not aligned with the expected standard normal distribution; then, it fails to detect some imperfect random sequences.

We propose a new metric called *Q-value* in this paper,[2] for the second-level tests of the binomial-based tests (but not the chi-square based ones), to replace the original P-value without any extra modification. The Q-value is computed directly using the test statistics, rather than their absolute values. We prove that the uniformity of Q-values is equal to that the test statistics follow the standard normal distribution as expected. In the case that there exists some mean drifts of the assumed normal distribution, which is commonly caused by flawed generators, the Q-value based tests produce greater gaps than the P-value based ones under both the total variation distance (TVD) and the Kullback-Leibler divergence (KLD), i.e., Q-value is more sensitive to detect such drifts. Therefore, for the binomial-based tests, our Q-value based second-level tests have greater testing capability than the P-value based ones.

Furthermore, inspired by [10], we investigate the actual distributions of P-values and Q-values with a finite block length, for the binomial-based tests. The comparison in the Frequency Test shows that the distribution of Q-value is more smooth, i.e., it is closer to the uniform distribution on $[0, 1]$. Hence, the Q-value based second-level tests are more reliable, i.e., our improvement also decreases the probability of erroneously identifying an ideal generator as not random.

Finally, we perform the improved statistical tests on the outputs of several PRNGs. The experimental results demonstrate that the Q-value based second-level tests are able to detect some statistical flaws that the original SP 800-22 suite cannot detect under the same test parameters.

Organization. The rest of this paper is organized as follows. In Section 2, we introduce the two-level statistical tests included in the NIST SP 800-22 test

[2] The term of *q-value* is defined as a measure of significance in terms of the false discovery rate [12, 13], while in this paper we use Q-value as another definition.

suite. In Section 3, we state the problem in the second-level tests of binomial-based tests. In Section 4, we propose Q-value based second-level statistical tests, and investigate the detectability and the reliability. In Section 5, we apply the statistical tests on several popular PRNGs to validate the effectiveness. Section 6 concludes the paper.

2 Two-level Statistical Tests in SP 800-22

2.1 Statistical Hypothesis Testing for Randomness

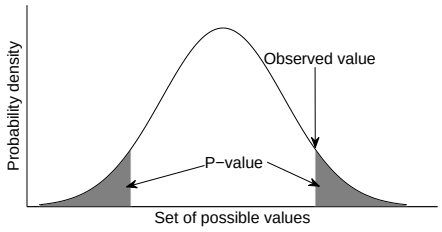
Hypothesis testing is a commonly used method to assess whether the tested data fit with the null hypothesis that is denoted as $\mathcal{H}_0$. In the statistical hypothesis testing, a statistic value is chosen and used to determine whether $\mathcal{H}_0$ should be accepted or rejected. Under the null hypothesis, the theoretical reference distribution of this statistic value is figured out by mathematical methods. From this reference distribution, a confidence interval is determined based on a preset *confidence level* γ (e.g., $\gamma = 0.99$), i.e., the probability that the statistic values are inside the confidence interval is γ.

The null hypothesis in statistical tests for randomness is that, the tested bit sequence is random. In the testing, the test statistic value is computed on the tested bit sequence, and then is compared to the bounds of the confidence interval. If the test statistic value lies outside the confidence interval, the null hypothesis that the sequence is random is rejected. Otherwise, $\mathcal{H}_0$ is accepted.

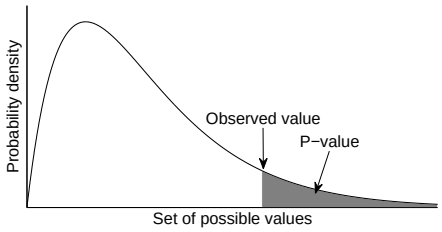
A randomness test suite may contain a serial of test items, which evaluate different aspects of randomness. These test items produce different confidence intervals based on the same confidence level. Then, *P-value* is employed as a unified metric for different test items, which is calculated using the test statistic. For a randomness test item, a P-value is the probability that a perfect random number generator would have produced a sequence less random than the tested sequence [11]. More specifically, the P-value is computed as the probability of obtaining a statistic value S equal to or "more extreme" than the observed value S_{obs} of the tested sequence. According to the definition of "more extreme" cases, the tests are generally divided into two categories: one-sided tests and two-sided tests.

In the NIST SP 800-22 test suite, a test is considered to be two-sided when S is assumed to follow a normal distribution, and the P-value is computed as $2\min\{\Pr(S > S_{obs}), \Pr(S < S_{obs})\}$. A test is considered to be one-sided when S is assumed to follow a chi-squared distribution, and the P-value is computed as $\Pr(S > S_{obs})$. Figure 1 shows the computations of P-values based on the observed values for the one-sided and two-sided tests included in the NIST SP 800-22 test suite, where the shaped areas are the P-values.

Then the test is performed by comparing P-value with a *significance level* denoted as α, and $\alpha = 1-\gamma$ where γ is the confidence level. If P-value $p < \alpha$,then $\mathcal{H}_0$ is rejected and the tested sequence is considered to be non-random. If $p \geq \alpha$, $\mathcal{H}_0$ is accepted and the sequence is considered to be random. When $\mathcal{H}_0$ is true and



(a) Two-sided test



(b) One-sided test

Fig. 1. P-value in one-sided and two-sided tests

$p < \alpha$, $\mathcal{H}_0$ is erroneously rejected, which is called Type I Error. The probability of Type I Error is α. On the contrary, the fact that, $p \geq \alpha$ when $\mathcal{H}_0$ is false, is called Type II Error. The significance level recommended by NIST is $\alpha = 0.01$.

2.2 Two-level Tests

The current version of the NIST SP 800-22 test suite [11] is composed of 15 test items. According to the assumed distribution of the test statistic values, these test items are divided into two categories: the binomial-based (i.e., the two-sided tests) and the chi-square based (i.e., the one-sided tests). The Frequency (Monobit) Test, the Runs Test, the Spectral Test, Maurer's "Universal Statistical" Test, and the Random Excursions Variant Test belong to the binomial-based tests, and the others are chi-square based.

In the testing process, according to the test parameters, the whole tested bit sequence is partitioned into N blocks, and each block contains n bits. For each test item, the hypothesis testing, where the null hypothesis is that the tested sequence is random, is executed for each data block, and then N P-values are obtained. Based on these P-values, the following two-level test is performed in each test item.

1. Count the number of the blocks whose P-values are equal or greater than α, and compute the passing ratio. If the ratio lies in the confidence interval defined as $1 - \alpha \pm 3\sqrt{\frac{(1-\alpha)\alpha}{N}}$, the first-level test is passed;

2. Divide the interval $[0, 1]$ into K equal sub-intervals, and count each number of the P-values in each sub-interval. Perform a chi-square goodness-of-fit test on these K numbers with the assumed uniform distribution, yielding another P-value p_T. If p_T is equal to or greater than another significance level α_T, the second-level test is considered to be passed. In the NIST SP 800-22 test suite, $K = 10$ and $\alpha_T = 0.0001$.

A test item is passed if the tested sequence passes the two-level test of this test item, and the SP 800-22 test suite is passed if all the 15 included test items are passed. The testing procedure is depicted in Figure 2, where we use $N = 1000$ as an example. Note that, some test items are further composed of a serial of sub-items (such as the Non-Overlapping Template Test), and each sub-item can be treated as a separate test item that has its own P-values and p_T. In addition, for the Random Excursions and Random Excursions Variant Tests, the P-values are computed only if the tested sequence block meets specific criteria, so the number of available P-values may be less than N for the N sequence blocks. These details are omitted in Figure 2 for simplicity.

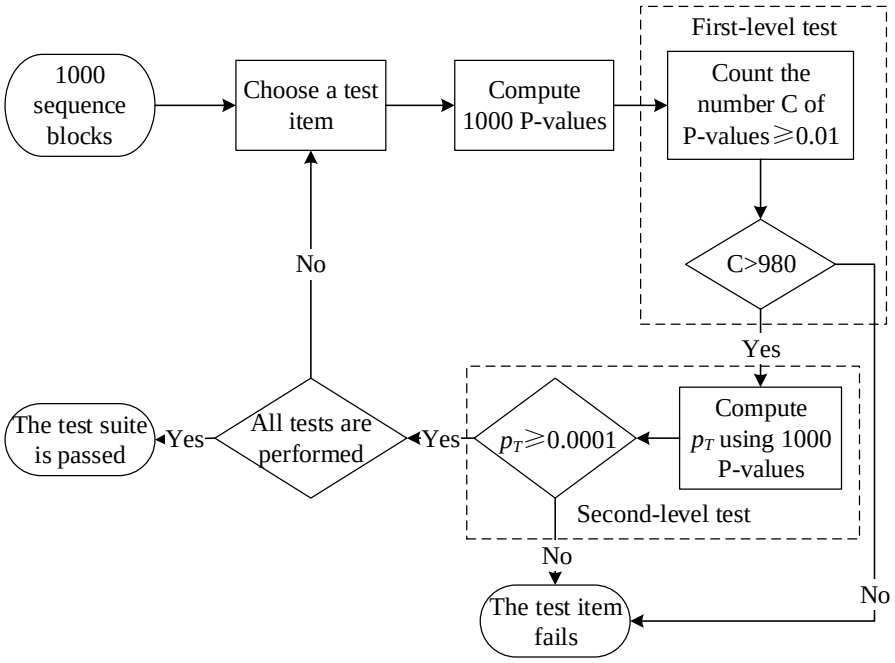


Fig. 2. The testing procedure of the NIST SP 800-22 test suite ($N = 1000$)

2.3 Frequency Test

We take the Frequency Test as an example to explain the P-value computation in the binomial-based tests. The bit block with length n is denoted as $\varepsilon = \{\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n\} \in \{0,1\}^n$. Then $S = \sum_{i=1}^{n}(2\varepsilon_i - 1)$ is computed. Under the null hypothesis, S is assumed to follow a binomial distribution. As n is always very large, the limiting binomial distribution is approximated as a normal distribution. Hence, S is assumed to follow the normal distribution $\mathcal{N}(u, \sigma^2)$, where $u = 0$ and $\sigma^2 = n$. The test statistic $d = (S - u)/\sigma$ follows $\mathcal{N}(0, 1)$. Then the P-value is computed using the cumulative distribution function (CDF) of the standard normal distribution $\Phi(\cdot)$ or the complementary error function erfc$(\cdot)$:

$$p = 2(1 - \Phi(|d|)) = \text{erfc}(\frac{|d|}{\sqrt{2}}),$$

where

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{\eta^2}{2}} d\eta,$$
$$\text{erfc}(x) = \frac{2}{\sqrt{\pi}} \int_{x}^{\infty} e^{-\eta^2} d\eta.$$

In all binomial-based tests, the same formula is used to compute P-values based on the test statistics, while each test item has a unique formula to compute the test statistic.

2.4 Spectral Test

The Spectral Test is also known as the Discrete Fourier Transform (DFT) Test or the Fast Fourier Transform (FFT) Test. The purpose of this test is to detect periodic features (i.e., repetitive patterns that are near each other) in the tested sequence [11]. For tested sequence $\varepsilon = \{\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n\} \in \{0,1\}^n$, the observed value N_1 is assumed to follow $\mathcal{N}(u, \sigma^2)$, where u is the expected number of frequency components that are beyond the 95% threshold $T = \sqrt{(\ln \frac{1}{0.05})n}$. Then the test statistic d is computed as:

$$d = \frac{N_1 - u}{\sigma},$$

where $u = 0.95n/2$, $\sigma^2 = 0.95 \cdot 0.05 \cdot n/c$, and $c = 4$ in the NIST SP 800-22 test suite.

3 Incompleteness of P-value based Second-level Tests

In the binomial-based tests, the standard normal distribution should be used as the reference for the observed test statistics. However, we find that, when the computed P-values follow a uniform distribution, the test statistic values are aligned with the *half-normal distribution*[3], rather than the expected normal

[3] The half-normal distribution refers to the fold at the mean of the standard normal distribution in this paper.

distribution. We prove this observation using the following Lemma 1 [3] and Theorem 1. To ensure the continuity of the statistic values' CDF, we assume that the sequence block length n is large enough in this section.

Lemma 1 *Let F be a continuous CDF on $\mathbb{R}$ with inverse F^{-1} defined by*

$$F^{-1}(z) = \inf\{x : F(x) = z, 0 < z < 1\},$$

where inf *means the infimum. If Z is a uniform random variable on $[0, 1]$, then $F^{-1}(Z)$ has distribution function F. Also, if a random variable X has distribution function F, then $F(X)$ is uniformly distributed on $[0, 1]$.*

Proof. The first statement follows after noting that for all $x \in \mathbb{R}$,

$$\begin{aligned}\Pr(F^{-1}(Z) \leq x) &= \Pr(\inf\{y : F(y) = Z\} \leq x) \\ &= \Pr(Z \leq F(x)) = F(x).\end{aligned}$$

The second statement follows from the fact that for all $0 < z < 1$,

$$\begin{aligned}\Pr(F(X) \leq z) &= \Pr(X \leq F^{-1}(z)) \\ &= F(F^{-1}(z)) = z.\end{aligned}$$

□

Theorem 1 *Let d be the test statistic in a binomial-based test, and let p be the P-value computed in the test. The following two statements are equivalent: 1) $|d|$ follows the half-normal distribution, and 2) p is uniformly distributed on $[0, 1]$.*

Proof. Let Y be a random variable following the half-normal distribution, and let $F_Y(\cdot)$ be the CDF of Y. On one hand, if $|d|$ follows the half-normal distribution, $F_Y(|d|)$ is a uniformly distributed variable on $[0, 1]$ according to the second statement of Lemma 1. Since p is computed as $1-F_Y(|d|)$, p is also uniformly distributed on $[0, 1]$. On the other hand, if $p = 1 - F_Y(|d|)$ is uniformly distributed on $[0, 1]$, $F_Y(|d|)$ also follows the uniform distribution on $[0, 1]$. According to the first statement of Lemma 1 (by replacing Z with $F_Y(|d|)$), $F_Y^{-1}(F_Y(|d|)) = |d|$, has the same CDF with Y, thus $|d|$ follows the half-normal distribution. □

Obviously, the condition that $|d|$ follows the half-normal distribution is insufficient to deduce that d follows the normal distribution. Therefore, for the second-level tests of the binomial-based tests, checking the uniformity of P-values is unqualified to assess whether d satisfies the null hypothesis. Hence, the second-level tests in the binomial-based tests could fail to detect some imperfect random sequences or elaboratively constructed sequences.

Remark. As to the chi-square based tests in the NIST SP 800-22 test suite, we clarify that these tests do not have the mentioned problem. The chi-square based tests are one-sided, and their P-values are not computed from the absolute values of the test statistic values.

Biased "random" sequence construction. Below we will construct a biased sequence, yet it passes the NIST SP 800-22 test suite with given test parameters.

1. Generate a random bit sequence with an appropriate length that passes the test suite. For example, use the Blum-Blum-Shub generator (BBS) [2] which is acknowledged as a good PRNG.
2. Perform the Frequency Test according to the test parameters n and N: calculate the test statistic value d_i of the ith block ($i = 1, ..., N$). For each i, if d_i is less than zero (i.e., 0's are more than 1's), perform a bitwise NOT (negation) on the sequence block; otherwise, keep the block unchanged.

The processed sequence is significantly biased, as the number of 1's is larger than that of 0's for each block after processing. However, the processed sequence still has a very high probability to pass the test suite due to the following reasons.

- For the Frequency Test, the P-value for each block is unchanged since $|d|$ remains unchanged.
- For most test items, "0" and "1" have equal roles in the evaluation of randomness. For example, in the Block Frequency, Cumulative Sums, Runs, Spectral, Universal, Approximate Entropy, and Serial Tests, their P-values remain unchanged after processing.

The effectiveness of the construction is confirmed by the statistical testing for the original and processed BBS output sequences. The two test reports about the original and processed BBS outputs are presented in Appendix A. We emphasize that, the constructed sequence is elaborative, and changing the testing method (e.g., enlarging the block length adopted by the test) certainly can detect the bias. The goal of our construction is to demonstrate the incompleteness of the P-value based second-level test, rather than to construct a flawed sequence which can pass all the existing test methods. In practice, an undetectable flaw may occur in other manners more than the unbalance, or occur in the focused aspects of other binomial-based tests more than the Frequency Test.

4 Second-level Tests based on Q-value

4.1 Q-value

The bias in the constructed sequence above should be detected by the Frequency Test that assesses the balance of the tested sequence. In our construction experiment, as the P-value based second-level tests cannot assess the symmetry of the test statistics, the constructed sequence "bypasses" the Frequency Test, even the whole test suite. For this reason, we introduce *Q-value* to replace P-value in the second-level tests of the binomial-based tests, and Q-value is defined as

$$q = 1 - \Phi(d) = \frac{1}{2}\mathrm{erfc}(\frac{d}{\sqrt{2}}).$$

The relationship between p and q is

$$p = \begin{cases} 2q, & q \leq 0.5; \\ 2(1-q), & q > 0.5. \end{cases}$$

Referring to the proof of Theorem 1, we have Theorem 2 for Q-value.

Theorem 2 *Let d be the test statistic in a binomial-based test, and let q be the Q-value computed in the test. The following two statements are equivalent: 1) d follows the standard normal distribution, and 2) q is uniformly distributed on* $[0, 1]$.

Checking the uniformity of Q-value is equal to assessing the distribution of d rather than $|d|$. Therefore, we propose the Q-value based second-level tests to replace the original second-level tests for the binomial-based tests. In the testing process, the modification is only using N Q-values rather than P-values to perform the chi-square goodness-of-fit test.

Different from P-value, Q-value is computed directly using the test statistics, rather than their absolute values. Hence, Q-value based tests are able to assess the symmetry (to zero) of the test statistics, and have greater testing capability. The constructed sequence in Section 3 cannot pass the Q-value based second-level test of the Frequency Test, because all the derived Q-values are not greater than 0.5.

4.2 Testing Capability on the Drift of Test Statistics

The second-level tests in the binomial-based tests are designed to assess the difference between the theoretical reference distribution (i.e., the standard normal distribution) and the observed distribution. In the practical testing on the output sequences of RNGs (rather than the elaboratively constructed sequences), we emphasize that both the P-value based and Q-value based tests can detect the statistical flaws when the observed distribution is quite different from the standard normal distribution. Hence, we focus on the case that the observed distribution is (or is similar to) a normal distribution, but the distribution parameters (such as the mean or the variance) drift from the ideal ones. Next, we compare the sensitivity to the drifts between the Q-value based test and the P-value based test.

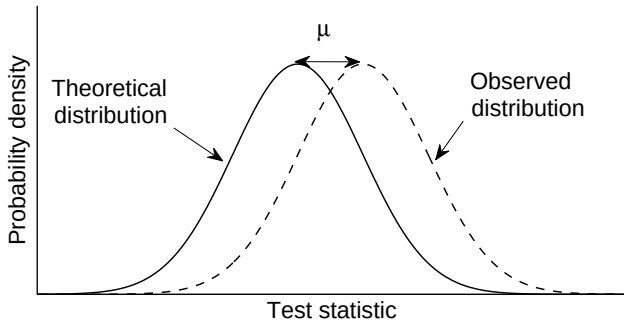


Fig. 3. The mean drift between the theoretical distribution and the observed one

Mean drift. The mean drift is defined as the distance between the mean values of the theoretical distribution and the observed one. We assume that the test statistic d, which is computed by a formula on the tested data, follows the standard normal distribution. The mean drift with μ for the test statistic is depicted in Figure 3.

Either an error in the computation formula of d or the flawed data can cause a drift. For example, Kim *el al.* [6] improved the formula in the Spectral Test, which makes the the distribution of the calculated test statistics from good RNGs show better consistency with the theoretical reference distribution. The other case that, the tested data are flawed, is more common in the testing. Below we show the consequence if one uses a biased generator of a noticeable mean drift.

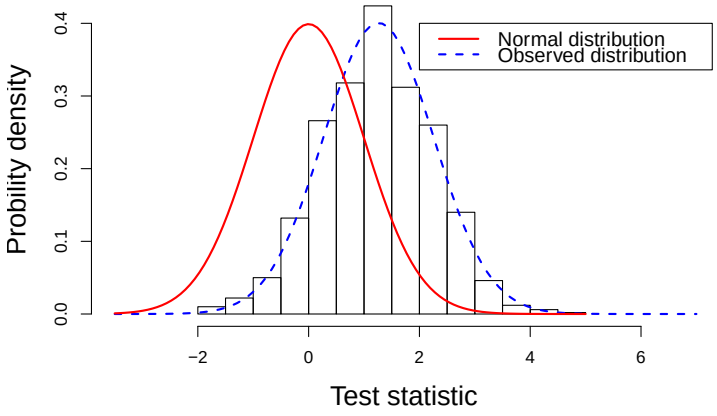


Fig. 4. The mean drift caused by the simulated biased sequence

The statistical flaw on the frequency (i.e., the tested sequence is biased) is common for imperfect RNGs, especially for those TRNGs where the physical phenomenons are not ideal. We assume that the flawed generator outputs a biased sequence with 50.2% 1's. The generator is simulated by the R software [1], and the observed test statistics are computed with the parameters $n = 10^5$ and $N = 1000$. Due to the existence of the bias, the distribution of the observed test statistics has a mean drift from the expected standard normal distribution, as shown in Figure 4. The histogram in Figure 4 is plotted using the probability density values computed on the 1000 test statistics. The mean of the observed distribution drifts to 1.265. Knowing the inherent bias of the generator output sequence, one can optimize brute-force attacks to reduce the breaking complexity

for cryptographic systems. Hence, it is important to detect the mean drift for the testing of RNGs.

KLD and TVD. We denote the probability distribution function (PDF) of the ideal test statistic d as $f(x)$, and the PDF of $|d|$ as $g(x)$. When the mean of the observed test statistics d_μ has a drift μ ($\mu \neq 0$), the PDFs of d_μ and $|d_\mu|$ are represented as $f_\mu(x)$ and $g_\mu(x)$, respectively. If the deviation caused by μ between $f(x)$ and $f_\mu(x)$ is larger than that between $g(x)$ and $g_\mu(x)$, we say that $f(x)$ is more sensitive to the drift, and the statistical test based on $f(x)$ has greater testing capability on detecting the drift.

We choose Kullback-Leibler divergence (KLD) and total variation distance (TVD) as the measurements of the sensitivity. For PDFs $h_A(x)$ and $h_B(x)$ of two continuous random variables A and B, the KLD between them is defined as

$$D_{\mathrm{KL}}(h_A(x)\|h_B(x)) = \int_{-\infty}^{\infty} h_A(x)\log\frac{h_A(x)}{h_B(x)}\mathrm{d}x, \tag{1}$$

and the TVD between them is defined as

$$\delta(h_A(x), h_B(x)) = \frac{1}{2}\int_{-\infty}^{\infty} |h_A(x) - h_B(x)|\mathrm{d}x. \tag{2}$$

Roughly speaking, KLD represents the amount of information lost when $h_B(x)$ is used to approximate $h_A(x)$, and TVD represents the largest possible difference between the probabilities that the two variables A and B have the same value.

When d is assumed to follow the standard normal distribution, we get

$$f(x) = \frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}, f_\mu(x) = \frac{1}{\sqrt{2\pi}}e^{-\frac{(x-\mu)^2}{2}}, x \in (-\infty, +\infty),$$

and

$$g(x) = \frac{2}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}, g_\mu(x) = \frac{1}{\sqrt{2\pi}}(e^{-\frac{(x-\mu)^2}{2}} + e^{-\frac{(x+\mu)^2}{2}}), x \in [0, +\infty).$$

Then, substituting $f(x)$ and $f_\mu(x)$ into Equation (1), we get the KLD between $f(x)$ and $f_\mu(x)$, as shown in Equation (3).

$$\begin{aligned} D_{\mathrm{KL}}(f(x)\|f_\mu(x)) &= \int_{-\infty}^{\infty} f(x)\log\frac{f(x)}{f_\mu(x)}\mathrm{d}x \\ &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}\log\frac{e^{-\frac{x^2}{2}}}{e^{-\frac{(x-\mu)^2}{2}}}\mathrm{d}x = \frac{\mu^2}{2} \end{aligned} \tag{3}$$

By noting that $e^{-\frac{(x-\mu)^2}{2}} + e^{-\frac{(x+\mu)^2}{2}} \geq 2e^{-\frac{x^2+\mu^2}{2}}$, we get the KLD between $g(x)$ and $g_\mu(x)$, which is strictly smaller than $\mu^2/2$, as shown in Equation (4).

$$D_{\mathrm{KL}}(g(x)\|g_\mu(x)) = \int_{-\infty}^{\infty} g(x)\log\frac{g(x)}{g_\mu(x)}\mathrm{d}x \tag{4}$$
$$= \int_0^{\infty} \frac{2}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \log\frac{2e^{-\frac{x^2}{2}}}{(e^{-\frac{(x-\mu)^2}{2}} + e^{-\frac{(x+\mu)^2}{2}})}\mathrm{d}x < \frac{\mu^2}{2}$$

By observing that $e^{-\frac{x^2}{2}}$ and $e^{-\frac{(x-\mu)^2}{2}}$ are symmetrical to $x = 0$ and $x = \mu$, respectively, we compare the result between $\delta(f(x), f_\mu(x))$ and $\delta(g(x), g_\mu(x))$, as shown in Equation (5).

$$\delta(f(x), f_\mu(x)) = \frac{1}{2}\cdot\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} |e^{-\frac{x^2}{2}} - e^{-\frac{(x-\mu)^2}{2}}|\mathrm{d}x \tag{5}$$
$$= \frac{1}{2\sqrt{2\pi}}\int_{-\infty}^{\frac{-\mu}{2}} e^{-\frac{x^2}{2}} - e^{-\frac{(x-\mu)^2}{2}}\mathrm{d}x + \frac{1}{2\sqrt{2\pi}}\int_{-\frac{\mu}{2}}^{\frac{\mu}{2}} e^{-\frac{x^2}{2}} - e^{-\frac{(x-\mu)^2}{2}}\mathrm{d}x$$
$$+ \frac{1}{2\sqrt{2\pi}}\int_{\frac{\mu}{2}}^{\infty} e^{-\frac{(x-\mu)^2}{2}} - e^{-\frac{x^2}{2}}\mathrm{d}x$$
$$= \frac{1}{2\sqrt{2\pi}}\int_{\frac{\mu}{2}}^{\infty} |(e^{-\frac{(x-\mu)^2}{2}} - e^{-\frac{x^2}{2}})| + |(e^{-\frac{x^2}{2}} - e^{-\frac{(x+\mu)^2}{2}})|\mathrm{d}x$$
$$+ \frac{1}{2\sqrt{2\pi}}\int_0^{\frac{\mu}{2}} |2e^{-\frac{x^2}{2}} - e^{-\frac{(x-\mu)^2}{2}} - e^{-\frac{(x+\mu)^2}{2}}|\mathrm{d}x$$
$$> \frac{1}{2\sqrt{2\pi}}\int_0^{\infty} |2e^{-\frac{x^2}{2}} - e^{-\frac{(x-\mu)^2}{2}} - e^{-\frac{(x+\mu)^2}{2}}|\mathrm{d}x = \delta(g(x), g_\mu(x))$$

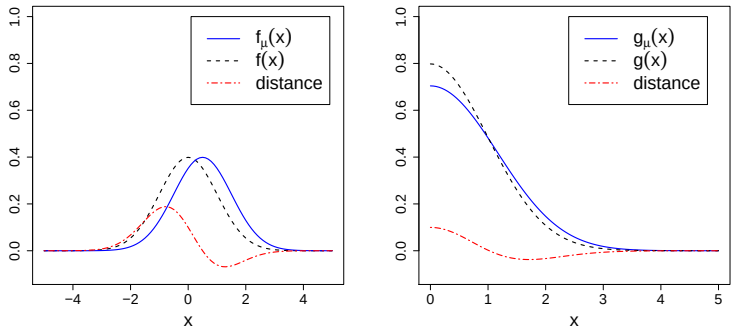
From Equations (3)–(5), we deduce $D_{\mathrm{KL}}(f(x)\|f_\mu(x)) > D_{\mathrm{KL}}(g(x)\|g_\mu(x))$ and $\delta(f(x), f_\mu(x)) > \delta(g(x), g_\mu(x))$.

The KLD and TVD results with $\mu = 0.5$ on the normal distribution $f(x)$ and the half-normal distribution $g(x)$ are also depicted in Figure 5, where the distances represent the integral parts in Equations (1) and (2). We can see that the change caused by μ in the normal distribution is larger than that in the half-normal distribution, which means that the test based on the normal distribution is more sensitive to the drift. Thus, we conclude that Q-value based second-level tests are more powerful than the P-value based ones to detect the mean drift of the test statistics.

Regarding to the drift of the variance, we note that the testing capability of the two testing methods are identical, as their KLDs (or TVDs) are equal.

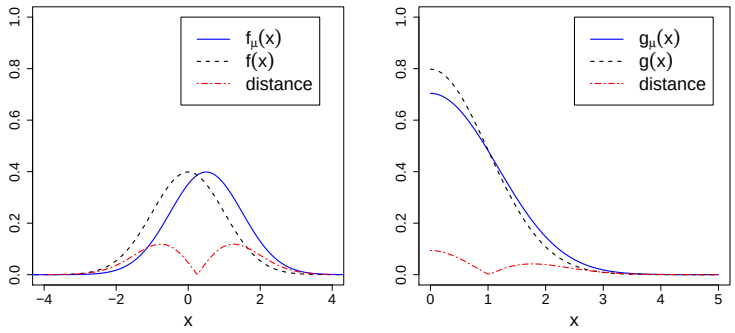
4.3 Testing Reliability Analysis based on Actual Distribution

An asymptotic distribution refers to the limiting distribution when n approaches infinity. The asymptotic distribution of P-value is the uniform distribution on $[0, 1]$. However, in the practical cases that n is finite, the number of possible P-values is limited, i.e., the set of P-values is discrete. This fact makes the actual



(a) The integral part in KLD between $f(x)$ and $f_\mu(x)$

(b) The integral part in KLD between $g(x)$ and $g_\mu(x)$



(c) The integral part in TVD between $f(x)$ and $f_\mu(x)$

(d) The integral part in TVD between $g(x)$ and $g_\mu(x)$

Fig. 5. The comparison of TVDs and KLDs with the drift $\mu = 0.5$

distribution of P-value is not a perfect uniform distribution on $[0, 1]$. When the number of blocks N is very large, the inconsistency is revealed and the observed P-values do not follow the assumed uniform distribution, which makes these P-values fail the chi-square test in the second-level test. This decreases the reliability of the statistical tests, i.e., increases the probability of erroneously identifying an ideal generator as not random.

In order to investigate the reliability of the Q-value based second level tests, we deduce the actual distributions of Q-values for the binomial-based tests, and compare them with those of P-values. The actual distributions of P-values for the binomial-based tests have been analyzed in [10]. The actual distribution of Q-value is closer to the assumed uniform distribution, meaning that the Q-value based test has a lower probability that a sequence with perfect randomness fails the test, i.e., higher reliability.

Actual distribution. As we mentioned in Section 2.3, each binomial-based test computes its normally distributed value $S \sim \mathcal{N}(u, \sigma^2)$. For an n-bit sequence block, the number of possible values of S is denoted as m. The possible values are increasingly ordered as $\mathcal{S} = \{s_1, s_2, ..., s_m\}$, i.e., $s_{i-1} < s_i$ for $i = 2, \ldots, m$. Note that the variables u, σ, m depend on the specific test item, such as the Frequency, Runs, Spectral, Universal, and Random Excursions Variant Tests. As our goal is to provide a general conclusion for the binomial-based tests, we do not consider the specific values of these variables.

For simplicity, we consider a common situation that m is odd and $\mathcal{S}$ is symmetrical with respect to u. For each s_i, P-value $p_i = \text{erfc}(\frac{|s_i - u|}{\sqrt{2}\sigma})$, and Q-value $q_i = \frac{1}{2}\text{erfc}(\frac{s_i - u}{\sqrt{2}\sigma})$. The sets of possible P-values and possible Q-values are denoted as $\mathcal{P}$ and $\mathcal{Q}$, respectively. According to the symmetry of S, it is observed that $p_i = p_{m+1-i}$ and $q_i + q_{m+1-i} = 1$, thus the cardinality $|\mathcal{P}| = m/2 + 1$ and $|\mathcal{Q}| = m$.

The actual CDFs of P-value and Q-value are represented as:

$$F'_p(x) = \sum_{i=1}^{m} \Pr\{S = s_i\} U(x - p_i), \tag{6}$$

$$F'_q(x) = \sum_{j=1}^{m} \Pr\{S = s_j\} U(x - q_j), \tag{7}$$

where

$$U(x) = \begin{cases} 1, & x \geq 0; \\ 0, & x < 0. \end{cases}$$

Using the property $p_i = p_{m+1-i}$, $F'_p(x)$ is rewritten as:

$$F'_p(x) = 2 \sum_{i=1}^{(m-1)/2} \Pr\{S = s_i\} U(x - p_i) + \Pr\{S = s_{\frac{m+1}{2}}\} U(x - p_{\frac{m+1}{2}}). \tag{8}$$

In fact, these two CDFs are both stepladder-like functions. We compare the number, height, and width of the steps between $F'_p(x)$ and $F'_q(x)$. Note that $|\mathcal{Q}|$ is almost as twice as $|\mathcal{P}|$, so the number of steps in $F'_q(x)$ is approximately as twice as that in $F'_p(x)$. The coefficient of the step function in $F'_p(x)$ is as twice as that in $F'_q(x)$, so the maximum width and height of the step in $F'_p(x)$ are also as twice as those in $F'_q(x)$. Therefore, the actual distribution of Q-values is more smooth, and is closer to the uniform distribution than that of P-values.

It should be noted that we assume $\mathcal{S}$ is symmetrical with respect to u, the mean of the asymptotic distribution. The assumption is appropriate for the Frequency Test; however, in other binomial-based tests, there may be a little deviation between u and the mean of S. We leave the study on this case as our future work.

Actual distribution in the Frequency Test. We take the Frequency Test as an example to demonstrate the difference between the distributions of P-value and Q-value. Without loss of generality, the length n of the sequence block is

assumed to be even. It is easy to figure out that $|\mathcal{S}| = n+1$, $|\mathcal{P}| = \frac{n}{2}+1$, $|\mathcal{Q}| = n+1$, and $u = 0$, $\sigma^2 = n$. Then, from Equation (8) we get the actual CDF of P-value:

$$F'_p(x) = 2\sum_i \Pr\{S = s_i\}U(x - p_i) + \frac{2}{\sqrt{2\pi n}}U(x-1),$$

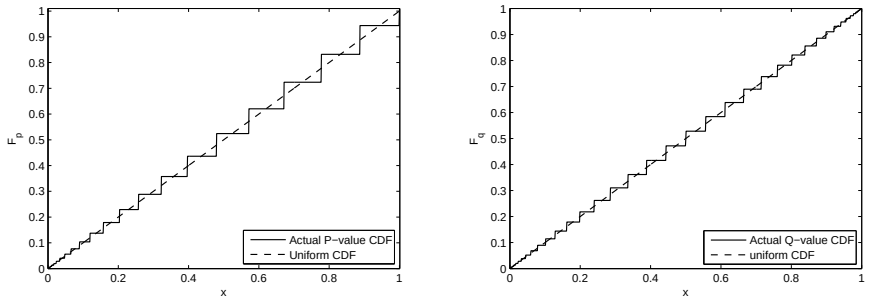
where $\Pr\{S = s_i\} = 2^{-n}\binom{n}{i-1} \approx \frac{2}{\sqrt{2\pi n}}e^{\frac{-(2i-n-2)^2}{2n}}$, $p_i = \text{erfc}(\frac{|2i-n-2|}{\sqrt{2n}})$, and $i \in \{1, 2, ..., \frac{n}{2}\}$.

From Equation (7), we get the actual CDF of Q-value:

$$F'_q(x) = \sum_j \Pr\{S = s_j\}U(x - q_j),$$

where $\Pr\{S = s_j\} = 2^{-n}\binom{n}{j-1} \approx \frac{2}{\sqrt{2\pi n}}e^{\frac{-(2j-n-2)^2}{2n}}$, $q_j = \frac{1}{2}\text{erfc}(\frac{2j-n-2}{\sqrt{2n}})$ and $j \in \{1, 2, ..., n+1\}$.

For the parameter $n = 200$, we plot the actual CDFs of P-value and Q-value, as shown in Figure 6. It is observed that Q-value's actual CDF is closer to the uniform distribution than P-value's.



(a) The actual CDF of P-value (b) The actual CDF of Q-value

Fig. 6. Actual CDF comparison between P-value and Q-value for the Frequency Test ($n = 200$)

Then, we compare the uniformity between actual P-values and Q-values through the chi-square goodness-of-fit test. Here we choose $n = 2^{20}$ and $K = 16$ to better express the difference between Q-values and P-values in the chi-square test. As shown in Equation (9), the statistic value χ^2 is computed using O_i which is the number of P-values or Q-values in the ith sub-interval.

$$\chi^2 = \sum_{i=1}^{K} \frac{(O_i - N/K)^2}{N/K} \tag{9}$$

Using the Q-value and P-value CDFs with $n = 2^{20}$, we calculate two sets of O_i based on P-values and Q-values, respectively. As expected, the set of O_i based on Q-values shows better consistency with the uniform distribution than that based on P-values, as shown in Figure 7. Therefore, we conclude that, under the same test parameters, the Q-value based second-level test has higher reliability than the P-value based one.

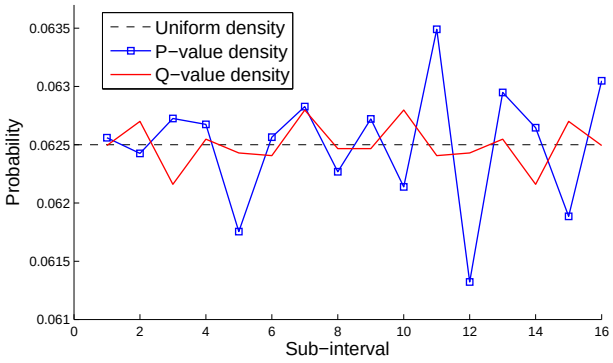


Fig. 7. The probability comparison between P-values and Q-values in each sub-interval ($K = 16,\ n = 2^{20}$)

To verify the correctness of the derived actual CDF of Q-value, we test the BBS output sequence with test parameters $n = 2^{10}$ and $N = 100000$, and count the number of Q-values in each sub-interval. The experimental and theoretical counting results in each sub-interval are shown in Figure 8, which shows good consistency between the theory and the experiment.

5 Statistical Tests on PRNGs

In this section, our experiments confirm that the Q-value based second-level tests have lower probabilities to erroneously identify good RNGs as not random, and also demonstrate that they have greater testing capability.

5.1 Experiment Setup

We choose several popular PRNGs including BBS, Linear Congruential Generator (LCG), Modular Exponentiation Generator (MODEXPG), and Micali-Schnorr Generator (MSG), and test their original output sequences using the NIST Statistical Test Suite (sts v2.1) [8] and our version using Q-values.

The test parameters adopted by each test item are the default values specified in the sts v2.1 toolkit. Also, we run the PRNG functions included in the toolkit

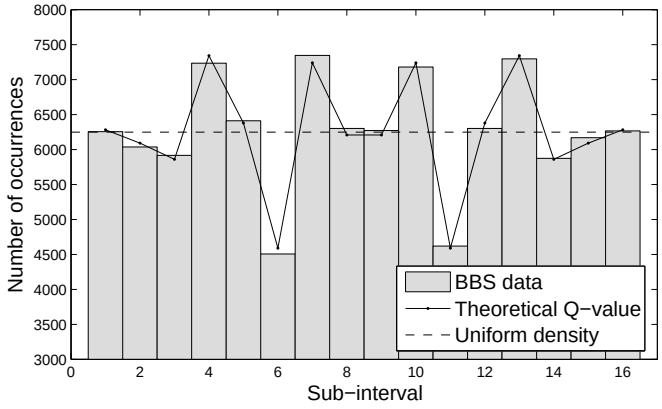


Fig. 8. Q-value comparison between the experimental and theoretical results in each sub-interval ($n = 2^{10},\ N = 100000,\ K = 16$)

to generate the output sequences, and the input parameters for these PRNGs are the default values fixed in the source code of sts v2.1, where the default seed of LCG is 23482349.

5.2 Statistical Testing

Using the recommended test parameters $n = 10^6$ and $N = 1000$, we perform statistical tests on the output sequences of these PRNGs. We only list the second-level test results (i.e., p_t's) for the Frequency Test, the Runs Test, the Spectral Test, and the Universal Test, as shown in Table 1. We omit the results of the Random Excursions Variant Test, for 18 different subitems are included in this item and all these subitems are passed.

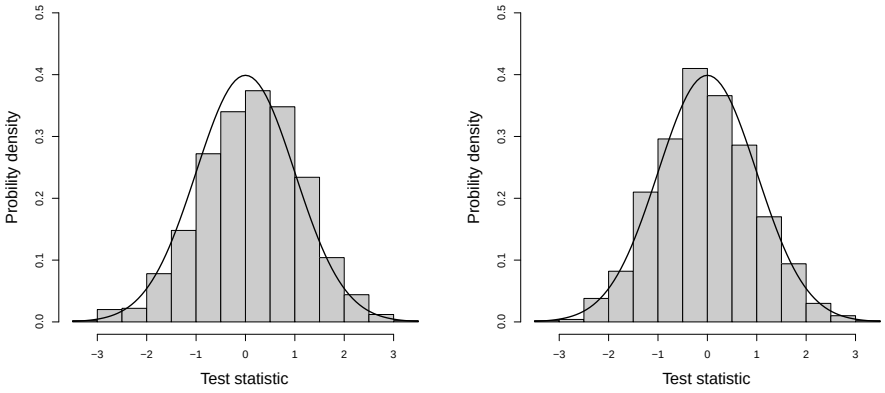
Table 1. Second-level test results for PRNGs ($n = 10^6$, $N = 1000$)

PRNG	Second-level Test	Frequency	Runs	Spectral ($c = 4$)	Spectral ($c = 3.8$)	Universal
BBS	P-value	0.6641	0.6350	0.5281	0.2480	0.4299
	Q-value	0.4817	0.9379	0.0218	0.0113	0.4263
MSG	P-value	0.3899	0.1746	0.6642	0.9619	0.7734
	Q-value	0.8055	0.1786	0.1825	0.6350	0.9996
LCG	P-value	0.8596	0.7075	0.4788	0.6392	0.8111
	Q-value	0.4769	0.8905	**0.0007**	**0.0026**	0.4447
MODEXPG	P-value	0.0	*	0.4541	0.2636	0.2676
	Q-value	0.0	*	0.1538	0.1107	0.7578

* The first-level test fails.

All the tested sequences of these PRNGs pass the whole original SP 800-22 test suite, except for MODEXPG. For BBS, MSG, and LCG, the p_t's of the three binomial-based tests are all greater than the preset threshold $\alpha_T = 0.0001$, thus these PRNGs pass both P-value based and Q-value based second-level tests. For MODEXPG, the Frequency Test fails either for P-value or Q-value. However, the Q-value's p_t of LCG in the Spectral Test becomes very small (0.0007), which indicates that the test statistics are not well consistent with the standard normal distribution, though p_t is still greater than α_T.

In order to confirm the discovery in the Spectral Test, we plot the histograms using the probability density values computed on the 1000 test statistics from LCG or MSG, and compare them with the PDF of the standard normal distribution, as depicted in Figure 9. The distribution of the test statistics of MSG has better consistency with the standard normal distribution, while the distribution of the test statistics of LCG drifts to the right. As we analyzed in Section 4.2, the Q-value based test is more sensitive to the mean drift, thus detects the drift better than the P-value based test.



(a) The test statistics of LCG (b) The test statistics of MSG

Fig. 9. Comparison between the standard normal distribution and the distribution of test statistics for LCG and MSG

For the Spectral Test, Pareschi *et al.* [10] pointed out that the variance $\sigma^2 = 0.95 \cdot 0.05 \cdot n/c$ with $c = 3.8$, is closer to the ideal distribution than the original value ($c = 4$) in the NIST SP 800-22 test suite [6, 4, 11]. Here we emphasize that the modification only adjusts the variance of the test statistic value, rather than the mean. The reason why the tested sequence almost fails the Spectral Test is the asymmetry of the statistic values. Therefore, the mean drift (or the asymmetry) still exists after modifying the variance, thus the Q-value based

test can still detect the drift. This is confirmed by the experiment, and the experimental results for $c = 3.8$ are also shown in Table 1.

5.3 Further Analysis on LCG

We repeat the Spectral Test on the output sequences of LCG with different seeds, and the p_t results of the P-value and Q-valued based second-level tests are presented in Table 2. From Table 2, we confirm that the conflict in Table 1 is not a coincidence or individual example, as similar results are also obtained for other seeds. It is noted that the choice of the LCG parameters has an impact on the quality of the output, thus the output sequences derived from some seeds are possible to show better statistical properties, as shown in the latter rows of Table 2.

Table 2. The second-level test results of the Spectral Test on the outputs of LCG with different seeds ($n = 10^6$, $N = 1000$, $c = 4$)

Seed	P-value based test	Q-value based test
73724612	0.3635	0.00006
12876498	0.2882	0.00030
52731971	0.0329	0.00096
92134122	0.0142	0.00106
82345342	0.1478	0.01581
59823781	0.6890	0.02959
23646172	0.2167	0.03732

Although we get small p_t's in the Q-value based second-level tests for LCG outputs, the sequence is still considered to pass the test ($p_t \geq \alpha_T = 0.0001$). Therefore, we further test the LCG outputs using a longer block length $n = 10^7$ to improve the testing capability, and the tested sequence is the same with that in Table 1. We find that, out of $N = 100$ blocks only 2 blocks pass the Spectral Test, i.e., the first-level test fails. For comparison, we also perform the test with $n = 10^7$ and $N = 100$ on the same BBS output sequence in Table 1, and the test is still passed. The detailed test reports are presented in Appendix B.

It is reasonable to conclude that the Q-value based second-level tests improve the detectability under the same test parameters. In the process of increasing the block length to improve the testing capability, the Q-value based second-level tests discover statistical flaws sooner.

6 Conclusion

We investigate the testing capability of the second-level tests of the binomial-based tests in the NIST SP 800-22 test suite, and find that, the sequence that

passes the tests could still have statistical flaws in the assessed aspect. Hence, we propose *Q-value* as the metric for the second-level tests to replace the original P-value without any extra modification. The Q-value based second-level test is applicable for all the five binomial-based tests, including the Frequency, Runs, Spectral, Universal, and Random Excursions Variant Tests. We provide the correctness proof of the proposed Q-value based second-level tests, and the distance analyses show that the modification improves the testing capability. Surprisingly, the comparison between the P-value's and Q-value's actual distributions indicates that the testing reliability is also improved. The experiments on several popular PRNGs demonstrate that the Q-value based second-level tests improve the detectability under the same test parameters. In the future, we will study the effectiveness of our method on TRNGs, and further analyze the properties of the Q-value based second-level tests.

Acknowledgments

We thank the anonymous reviewers of CHES 2016 and ASIACRYPT 2016, for their invaluable suggestions and comments to improve the quality and fairness of this paper. This work was partially supported by National Basic Research Program of China (973 Program No. 2013CB338001) and National Natural Science Foundation of China (No. 61602476).

References

1. The R project for statistical computing. `www.r-project.org`
2. Blum, L., Blum, M., Shub, M.: A simple unpredictable pseudo-random number generator. SIAM J. Comput. 15(2), 364–383 (1986)
3. Devroye, L.: Non-uniform random variate generation, chap. 2, p. 28. Springer-Verlag, New York (1986)
4. Hamano, K.: The distribution of the spectrum for the discrete fourier transform test included in SP800-22. IEICE Transactions 88-A(1), 67–73 (2005)
5. Hamano, K., Kaneko, T.: Correction of overlapping template matching test included in NIST randomness test suite. IEICE Transactions 90-A(9), 1788–1792 (2007)
6. Kim, S., Umeno, K., Hasegawa, A.: Corrections of the NIST statistical test suite for randomness. IACR Cryptology ePrint Archive 2004, 18 (2004), `http://eprint.iacr.org/2004/018`
7. Marsaglia, G.: Diehard Battery of Tests of Randomness. `http://www.stat.fsu.edu/pub/diehard/`
8. NIST: Statistical test suite (sts 2.1). `csrc.nist.gov/groups/ST/toolkit/rng/documents/sts-2.1.2.zip`
9. Pareschi, F., Rovatti, R., Setti, G.: Second-level NIST randomness tests for improving test reliability. In: International Symposium on Circuits and Systems (ISCAS 2007). pp. 1437–1440 (2007)
10. Pareschi, F., Rovatti, R., Setti, G.: On statistical tests for randomness included in the NIST SP800-22 test suite and based on the binomial distribution. IEEE Transactions on Information Forensics and Security 7(2), 491–505 (2012)

11. Rukhin, A., et al.: A statistical test suite for random and pseudorandom number generators for cryptographic applications. NIST Special Publication 800–22. http://csrc.nist.gov/publications/nistpubs/800-22-rev1a/SP800-22rev1a.pdf
12. Storey, J.D.: The positive false discovery rate: A bayesian interpretation and the q-value. Annals of Statistics 31(6), 2013–2035 (2003)
13. Storey, J.D., Tibshirani, R.: Statistical significance for genomewide studies. Proceedings of the National Academy of Sciences 100(16), 9440–9445 (2003)
14. Sulak, F., Doganaksoy, A., Ege, B., Koçak, O.: Evaluation of randomness test results for short sequences. In: Sequences and Their Applications (SETA 2010). pp. 309–319 (2010)
15. Zhuang, J., Ma, Y., Zhu, S., Lin, J., Jiwu, J.: Q_value test: a new method on randomness statistical test. Journal of Cryptologic Research 3(2), 192–201 (2016), (in Chinese)

A Statistical test results on the original and processed BBS output sequences

Table 3. Statistical test report of the original BBS outputs ($n = 10^6$, $N = 10^3$)

C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	P-VALUE	PROPO	STATISTICAL TEST
102	88	101	108	101	89	93	95	111	112	0.664168	995/1000	Frequency
86	109	99	91	89	106	106	94	116	104	0.474986	987/1000	BlockFrequency
88	102	85	106	112	98	88	109	103	109	0.463512	994/1000	CumulativeSums
97	96	83	104	103	94	107	109	103	104	0.807412	996/1000	CumulativeSums
89	115	89	103	102	95	100	102	93	112	0.635037	990/1000	Runs
98	90	97	101	116	102	99	93	104	100	0.883171	994/1000	LongestRun
103	98	80	92	102	96	116	94	108	111	0.371941	989/1000	Rank
107	108	83	99	109	101	90	94	96	113	0.528111	983/1000	FFT
97	104	101	118	84	86	112	94	97	107	0.319084	993/1000	NonOverlappingTemplate
103	101	104	106	112	94	90	95	90	105	0.841226	992/1000	OverlappingTemplate
114	118	104	101	98	93	93	97	84	98	0.429923	987/1000	Universal
107	98	97	89	95	99	101	101	106	107	0.965860	995/1000	ApproximateEntropy
62	58	67	60	68	61	53	63	60	52	0.906970	598/604	RandomExcursions
59	53	55	56	66	59	73	51	58	74	0.380976	600/604	RandomExcursionsVariant
90	96	81	105	109	96	104	117	109	93	0.323668	994/1000	Serial
81	97	91	104	112	100	105	103	113	94	0.484646	988/1000	Serial
111	95	100	107	113	88	97	97	97	95	0.779188	993/1000	LinearComplexity

Table 4. Statistical test report of the processed BBS outputs ($n = 10^6$, $N = 10^3$)

C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	P-VALUE	PROPO	STATISTICAL TEST
102	88	101	108	101	89	93	95	111	112	0.664168	995/1000	Frequency
86	109	99	91	89	106	106	94	116	104	0.474986	987/1000	BlockFrequency
88	102	85	106	112	98	88	109	103	109	0.463512	994/1000	CumulativeSums
97	96	83	104	103	94	107	109	103	104	0.807412	996/1000	CumulativeSums
89	115	89	103	102	95	100	102	93	112	0.635037	990/1000	Runs
101	88	88	111	118	100	104	99	92	99	0.518106	993/1000	LongestRun
99	100	81	85	107	98	111	98	110	111	0.361938	990/1000	Rank
107	108	83	99	109	101	90	94	96	113	0.528111	983/1000	FFT
98	102	88	105	91	105	104	97	102	108	0.926487	994/1000	NonOverlappingTemplate
122	89	90	98	112	96	109	108	84	92	0.147815	991/1000	OverlappingTemplate
114	118	104	101	98	93	93	97	84	98	0.429923	987/1000	Universal
107	98	97	89	95	99	101	101	106	107	0.965860	995/1000	ApproximateEntropy
61	56	57	53	74	57	64	65	65	52	0.654467	597/604	RandomExcursions
56	60	58	55	76	54	69	51	54	71	0.280306	601/604	RandomExcursionsVariant
90	96	81	105	109	96	104	117	109	93	0.323668	994/1000	Serial
81	97	91	104	112	100	105	103	113	94	0.484646	988/1000	Serial
97	92	110	99	101	105	98	97	108	93	0.953089	992/1000	LinearComplexity

B Statistical test results with the longer block length on the LCG and BBS output sequences

Table 5. Statistical test report of the LCG outputs ($n = 10^7$, $N = 10^2$)

C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	P-VALUE	PROPO	STATISTICAL TEST
12	16	4	12	12	8	10	8	11	7	0.334538	100/100	Frequency
12	10	8	11	11	8	11	10	6	13	0.911413	100/100	BlockFrequency
12	13	14	6	6	11	8	7	13	10	0.494392	99/100	CumulativeSums
9	11	10	11	11	17	10	6	9	6	0.474986	100/100	CumulativeSums
9	10	7	12	13	8	13	12	9	7	0.834308	99/100	Runs
9	10	9	10	13	16	7	6	12	8	0.534146	100/100	LongestRun
14	10	12	7	11	7	9	12	9	9	0.867692	98/100	Rank
100	0	0	0	0	0	0	0	0	0	**0.000000**	**2/100**	FFT
6	11	10	10	6	15	12	11	14	5	0.319084	100/100	NonOverlappingTemplate
17	13	7	10	13	7	8	6	9	10	0.304126	97/100	OverlappingTemplate
6	8	7	8	12	10	7	17	12	13	0.289667	98/100	Universal
10	3	8	10	8	12	13	6	21	9	0.013569	99/100	ApproximateEntropy
8	14	7	13	3	9	6	13	9	8	0.213309	88/90	RandomExcursions
9	8	13	11	7	9	11	11	6	5	0.694743	89/90	RandomExcursionsVariant
5	9	10	9	4	12	8	18	10	15	0.066882	100/100	Serial
8	10	11	7	13	6	12	13	10	10	0.816537	100/100	Serial
7	11	6	8	13	11	14	7	9	14	0.514124	100/100	LinearComplexity

Table 6. Statistical test report of the BBS outputs ($n = 10^7$, $N = 10^2$)

C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	P-VALUE	PROPO	STATISTICAL TEST
12	10	7	6	11	6	15	7	8	18	0.096578	99/100	Frequency
11	8	7	11	8	7	6	12	14	16	0.350485	100/100	BlockFrequency
12	10	4	15	8	10	8	8	13	12	0.437274	99/100	CumulativeSums
12	5	10	12	9	4	11	13	11	13	0.437274	99/100	CumulativeSums
11	12	10	12	8	6	8	8	13	12	0.834308	100/100	Runs
7	16	14	11	5	7	13	9	5	13	0.122325	100/100	LongestRun
11	11	12	6	8	7	10	11	10	14	0.816537	99/100	Rank
16	13	13	6	14	10	7	7	5	9	0.162606	97/100	FFT
13	11	12	7	6	7	6	10	11	17	0.249284	97/100	NonOverlappingTemplate
18	12	10	11	8	9	11	7	7	7	0.334538	96/100	OverlappingTemplate
15	10	9	8	8	8	12	11	7	12	0.779188	100/100	Universal
9	11	11	6	7	8	10	7	8	23	0.010988	99/100	ApproximateEntropy
11	8	9	7	7	9	9	7	7	14	0.689019	88/88	RandomExcursions
9	8	8	6	10	7	4	11	17	8	0.105618	87/88	RandomExcursionsVariant
9	12	10	9	13	10	7	6	7	17	0.366918	100/100	Serial
7	20	8	8	9	9	8	8	9	14	0.108791	99/100	Serial
11	7	11	10	14	12	6	8	12	9	0.779188	100/100	LinearComplexity