Public Comments on SP 800-22 Rev. 1a, A Statistical Test Suite for Random and Pseudorandom Number Generators for Cryptographic Applications

Comment period: August 6, 2021 -- October 1, 2021

On August 6, 2021, NIST's Crypto Publication Review Board initiated a review of SP 800-22 Rev. 1a, *A Statistical Test Suite for Random and Pseudorandom Number Generators for Cryptographic Applications* (April 2010). This document includes the public comments received during the comment period from August 6, 2021 to October 1, 2021.

More details about this review are available from NIST's Crypto Publication Review Project site.

LIST OF COMMENTS

1. Comments from Gary Woodcock, August 15, 2021

Anametric is developing an entropy source product, and we're heavily exercising the NIST STS v.2.1.2 software package (as referenced by SP 800-22 Rev. 1a) to help evaluate its quality. We consider the NIST STS tests to be vital tools in our overall development process, and we run them frequently. However, we've encountered multiple issues that impact our ability to most effectively apply the tests.

Consequently, we created a new framework to host the NIST STS tests that addresses these issues, which we would like to bring to your attention. The key issues we identified and amended are:

1) The NIST STS requires interactive console input. This precludes simple automation of tests runs via (for example) shell scripts, and also precludes scaling test runs in the cloud using dynamic resources.

2) No automatic validation of compiled builds. Though SP 800-22 Rev. 1a provides expected results in Appendix B for test runs with the five supplied test input data files included with the NIST STS package, there are no provided tools/scripts to actually execute these test configurations, nor to check the results against expected values. This must all be conducted manually.

3) No standardization of test parameters.

4) No standardization of test output/results. This makes parsing test results more difficult, as there is no agreed-upon structure for the output format.

5) Test outputs do not record any information regarding the test environment (for example, operating system, CPU architecture, versions, etc.), which makes it more difficult to directly compare test results between testers.

6) Some tests do not provide any useful feedback while running. For long-running tests, this makes it difficult to determine whether a test is executing or hung.

7) The NIST STS doesn't enforce any of the test restrictions documented in SP 800-22 Rev. 1a. For example, Section 4.2.2 states that in the computation of the distribution of p-values, "... to provide statistically meaningful results at least 55 sequences must be processed." This restriction is not enforced in the NIST STS source code.

8) Due to its software architecture, the NIST STS is unable to efficiently parallelize its test runs. All test runs are serialized by both bitstream and test, resulting in sub-optimal performance on multi-core/multi-processor compute platforms.

We believe that our new framework preserves the design and intent of the existing 15 tests in the NIST STS, while substantially improving flexibility, extensibility and performance. We're very excited and interested to open a discussion with the appropriate stakeholders at NIST to review our work – we believe it can have a positive impact for all users of the NIST STS going forward!

Thank you for your consideration,

Gary Woodcock

2. Comments from Canadian Centre for Cyber Security (CCCS), September 1, 2021

After review NIST SP 800-22 and the relevant literature, most of comments and recommendations relate to the behaviour of the tests with small sample sizes, the independence of the tests, the power of the tests to detect non-random sequences, and clarification of the choice of parameters. Comments and recommendations are as follows:

Editorial comments

In section 1.2, the definition for the complementary error function should say “See the definition for Erfc”.

In 2.2.3, χ^2 is not a measure of how well the observed proportions match for a given M-bit block; it considers all the M-bit blocks. This should be reworded.

Recommendations from reviewed papers

A General Method to Evaluate the Correlation of Randomness Tests

Link: https://link.springer.com/chapter/10.1007%2F978-3-319-05149-9_4

Description: This paper also considers the independence of the tests in the NIST suite and claims that their experiment does not agree with NIST’s claim that the tests included in the suite are independent.

Recommendation: This test should be replicated by NIST to ensure the tests in the suite are sufficiently independent as claimed in section 4.4. Out of all the recommendations given this one should be the highest priority as the results from this paper directly go against the claims made in SP 800-22.

Evaluation of Randomness Test Results for Short Sequences

Link: https://link.springer.com/chapter/10.1007/978-3-642-15874-2_27

Description: This paper notes that most of the NIST test suite applies only to relatively large sequences (see 4.2.1 of NIST SP 800-22). They then calculate exact distribution of the p-values (which is only asymptotically normal but assumed to be normal by the NIST test suite).

Recommendation: These calculated probabilities could be added to the test suite, or a reference could be added to this paper section 4.3 on sequence length. Also, note in section 4.2.2 that the p-values could not be uniform for short sequences.

On Statistical Tests for Randomness Included in the NIST SP800-22 Test Suite and Based on the Binomial Distribution

Link: https://ieeexplore.ieee.org/abstract/document/6135498

Description: This paper reviews the frequency, runs and spectral tests in NIST SP 800-22. It approximates the exact distribution of the p values of the frequency, number of runs, etc. on a small set, which are assumed to be binomial in the NIST version of the tests. It also suggests an improvement over the distribution used for the spectral test. This would help with testing small sample sizes.

Recommendation: The calculated distributions could be used in place of the binomial distribution in the NIST test code, particularly for small samples sizes.

A view on NIST randomness tests (In)Dependence

Link: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8166460

Description: This paper considers the independence of the tests provided in the NIST test suite, which is also considered in section 4.4 of NIST SP 800-22, the false acceptance rate of the NIST tests, and the minimum sample size required to obtain a certain rejection rate with a particular error.

Recommendation: NIST could repeat any or all of the three experiments, particularly testing the independence of the testing suite using the simple approach described in the paper may be worthwhile, rather than relying only on the factor analysis in section 4.4. An experiment like the one performed in this paper to estimate the false acceptance rate of the tests would also be useful to obtain estimates of the relative power of each of the tests.

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

Link: https://link.springer.com/chapter/10.1007/978-3-662-53887-6_11

Description: This paper claims that the testing method described in 4.2.2 testing the p-values against a uniform distribution is flawed, since the test statistic d is only guaranteed to be a half normal distribution. They provided an alternative q value to fix this problem.

Recommendation: If it is deemed necessary, NIST should replace the p value with the q value as suggested.

Other Comments/Suggestions

Besides the papers listed above, there have been other papers claiming to have new tests. Such new tests should not be added to the suite until/unless they have been shown to be useful and independent of the other tests in the suite.

For testing that the statistical test suite is operating properly (Appendix B), a test script could be added to run the tests on the sample files and compare with the expected results p-values given in Appendix B.

3. Comments from Yuan Ma, Chinese Academic of Sciences, September 14, 2021

Dear Sir/Madam,

We are a research team on randomness from Chinese Academic of Sciences. The following are our comments on NIST Special Publication (SP) 800-22 that still has huge influence in either academia or industry.

Our finding is that P-value is not qualified or suitable as the tested value in the second level tests of the five binomial distribution based test items included in SP 800-22.

The second level test means Goodness-of-Fit Distributional Test on the P-values, which is presented in Sect. 4.2.2 of SP 800-22. The five binomial distribution based test items are the Frequency Test, the Runs Test, the Spectral Test, the Universal Test, and the Random Excursions Variant Tests.

In particular, we found that there exists a flaw in the second-level testsbu of the binomial distribution based tests, yielding that some flawed sequences could pass the test suite. The conclusion is proved in both theory and practice, and these flawed sequences are real, which is confirmed by our experiments on the LCG outputs. These results are fully demonstrated in our AsiaCrypt2016 Paper titled "More Powerful and Reliable Second-Level

Statistical Randomness Tests for NIST SP 800-22", which is attached in this email.

We believe that the problem is important and has a **non-negligible impact** on the accuracy of the test suite. In order to fix this problem, we propose **a new metric Q-value (similar to P-value) in our paper**, which is dedicated for the second level tests, and the P-value is only fed to the first level tests (i.e., 4.2.1Proportion of Sequences Passing a Test in SP800-22). **We prove that Q-value is qualified as the tested value in the second-level tests.**

Furthermore, in addition to improve the testing accuracy (power), using Q-value has another benefit, i.e., **that is improving the realiability.** The problem of realiabiltiy for large block numbers is presented in an earlier IEEE TIFS 2012 paper titled "On Statistical Tests for Randomness Included in the NIST SP800-22 Test Suite and Based on the Binomial Distribution". The paper showed that, for large block numbers (10^5 or 10^6), the binomial distribution based tests returned the "fail" result for good random numbers. The improved results for the reliability are also demonstrated in the attached AsiaCrypt paper.

Thus, **one possible new test schedule of SP800-22 may be as following.** Firstly, when each test item generates P-value, Q value is also generated. For the binomial distribution based tests, Q-value is constructed as in our paper; for other tests, Q-value equals to P-value. Then, the P-values are used for the first-level tests as the same, and the Q-values are used for the second-level tests instead of the former P-values. The others (including the judgment rules) remain the same as the original test suite.

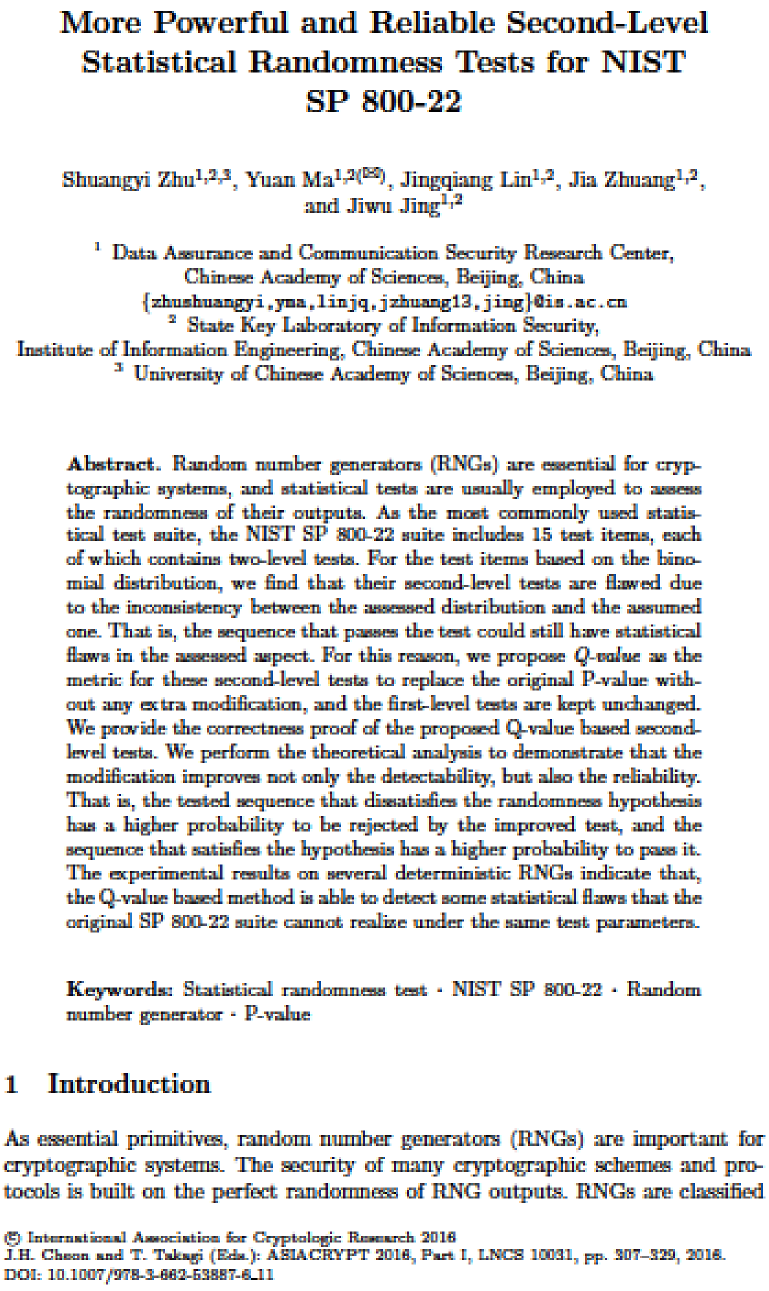
We hope that our work could be a substantial contribution to improve SP 800-22. Any feedback or discussion will be more than appreciated.

Thanks for your work!

____

Best regards,
Yuan Ma
Chinese Academic of Sciences

[The commenter attached a copy of the paper below, which is available from the publisher at https://doi.org/10.1007/978-3-662-53887-6_11.]



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
{zhushuangyi,yma,linjq,jzhuang13,jing}@is.ac.cn
[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

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

J.H. Cheon and T. Takagi (Eds.): ASIACRYPT 2016, Part I, LNCS 10031, pp. 307–329, 2016.
DOI: 10.1007/978-3-662-53887-6_11

4. Comments from Dan Brown, September 20, 2021

Dear NIST Crypto Publication Review Board,

Thank you for your excellent work, and the opportunity for public comments.

Please find attached BlackBerry's comments about NIST Special Publication 800-22.

Best regards,

Dan Brown

Senior Manager, Standards

BlackBerry

September 14, 2021

National Institute of Standards and Technology
100 Bureau Drive, Gaithersburg, MD 20899

Re: 800-22, A Statistical Test Suite for Random and Pseudorandom Number Generators for Cryptographic Applications

Dear NIST,

BlackBerry commends NIST's leadership in developing, maintaining, and reviewing cryptography-related publications that help to better secure user information. We appreciate the opportunity for public comments on FIPS 198-1 and Special Publications on Hash Functions, Statistical Randomness Tests, and Block Cipher Modes of Operation.

BlackBerry encourages the **withdrawal** of NIST Special Publication 800-22, *A Statistical Test Suite for Random and Pseudorandom Number Generators for Cryptographic Applications*.

More recent NIST Special Publications 800-90A and 800-90B describe much better approaches to random number generation in cryptography: 800-90B Section 4.1 already says that 800-22 is not useful for health tests of noise sources. Web pages providing withdrawn versions of 800-22 should include a prominent pointer to the newer 800-90A and 800-90B.

As an alternative to withdrawal, 800-22 could be amended to address only *non-cryptographic* applications of random number generators.

Respectfully submitted,

Dan Brown
Senior Standards Manager

5. Comments from Fatih Sulak, Atilim University, October 1, 2021

Dear Madam/Sir,

I send our comments on NIST Special Publication (SP) 800-22 Rev. 1a in the attachment.

Kind regards.

Assoc. Prof. Dr. Fatih SULAK

Atılım University

Mathematics Department

Comments on NIST Special Publication (SP) 800-22 Rev. 1a

Fatih Sulak, Ali Doğanaksoy, Muhiddin Uguz

Department of Mathematics, Atılım University, Ankara, Turkey
Department of Mathematics, Middle East Technical University, Ankara, Turkey
fatih.sulak@atilim.edu.tr, aldoks@metu.edu.tr, muhid@metu.edu.tr,

In the NIST Special Publication (SP) 800-22 Rev. 1a (in this document it will be called as NIST Test Suite), there are two randomness tests considering the number of occurrences of a pre-defined template in a sequence, namely the overlapping template matching test and the non-overlapping template matching test.

Definition 1. *Period of a template $\tau = \tau_1 \cdots \tau_\rho \cdots \tau_{\lambda_1}$ is defined as the smallest index ρ such that there exists a template $\eta = \eta_1 \cdots\cdots\cdots \eta_{\lambda_2}$ satisfying*

$$\begin{aligned} \tau\eta &= \tau_1 \cdots \tau_\rho \cdots \tau_{\lambda_1} \eta_1 \cdots\cdots\cdots \eta_{\lambda_2} \\ &= \tau_1 \cdots \tau_\rho \tau_1 \cdots \tau_\rho \ \cdots\cdots\cdots \ \eta_{\lambda_2} \end{aligned}$$

For example, consider the template $\tau = \mathbf{01}0101$. This template has period 2, since for $\eta = 01\cdots$, $\tau\eta = 01010101\cdots$. Similarly, the template $\tau = \mathbf{01101}011$ has period 5, as for $\eta = 01\cdots$, $\tau\eta = 0110101101\cdots$.

If the period of the template is maximum then the template is called non-overlapping, otherwise it is called as overlapping template. Although, the NIST Test Suite contains a test for the case $n = 1032$, namely the overlapping template matching test, the probabilities stated in the NIST Test Suite are valid only for $\lambda = 9$ ($B = 111111111$). The probabilities are not stated for the other templates.

In a recent paper, it is observed that the probabilities are the same for any two templates having the same period, and all probabilities for all possible templates are given [1]. Notice that only the first column of the Table 1 is given in the NIST Test Suite.

Table 1: Bin probabilities for Overlapping Template Matching Test, $n = 1032$, $\lambda = 9$

	$\rho = 1$	$\rho = 2$	$\rho = 3$	$\rho = 4$	$\rho = 5$	$\rho = 6$	$\rho = 7$	$\rho = 8$	$\rho = 9$
π_0	0.364091	0.218724	0.168833	0.148856	0.139322	0.135077	0.132952	0.131889	0.130823
π_1	0.185659	0.252422	0.266954	0.269936	0.270569	0.270658	0.270658	0.270647	0.270632
π_2	0.139381	0.206593	0.240861	0.257798	0.266743	0.270929	0.273075	0.274162	0.275260
π_3	0.100571	0.140936	0.161571	0.172140	0.177889	0.180632	0.182049	0.182768	0.183493
π_4	0.0704323	0.0855144	0.0892548	0.0900851	0.0902333	0.0902316	0.0902106	0.090194	0.0901712
π_5	0.139865	0.095811	0.0725254	0.0611858	0.0552441	0.0524724	0.0510559	0.0503402	0.0496206

In our opinion, a revision in the NIST Test Suite is needed for the other templates. Moreover, overlapping template matching test can be applied to the sequences of minimum length 10^6 as stated in NIST Test Suite. However, using the setting $n = 128$, $\lambda = 4$, and $\rho = 4$, the periodic template test can be applied to any sequence whose length is greater than 4864. The probabilities can be easily evaluated for any block length and period using the theorems given in [1].

References

1. Sulak F, Doganaksoy A, Uguz M, Kocak O, Periodic Template Tests: A Family of Statistical Randomness Tests for a Collection of Binary Sequences, Discrete Applied Mathematics, 2019.