POLITECNICO DI TORINO
Repository ISTITUZIONALE

On statistical tests for randomness included in the NIST SP800-22 test suite and based on the binomial distribution

Original
On statistical tests for randomness included in the NIST SP800-22 test suite and based on the binomial distribution / Pareschi, Fabio; Rovatti, Riccardo; Setti, Gianluca. - In: IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY. - ISSN 1556-6013. - STAMPA. - 7:2(2012), pp. 491-505. [10.1109/TIFS.2012.2185227]

Availability:
This version is available at: 11583/2696608 since: 2022-03-28T18:32:02Z

Publisher:
Institute of Electrical and Electronics Engineers Inc.

Published
DOI:10.1109/TIFS.2012.2185227

Terms of use:

This article is made available under terms and conditions as specified in the corresponding bibliographic description in the repository

Publisher copyright
IEEE postprint/Author's Accepted Manuscript

©2012 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collecting works, for resale or lists, or reuse of any copyrighted component of this work in other works.

(Article begins on next page)

12 October 2024

On Statistical Tests for Randomness included in the NIST SP800-22 test suite and based on the Binomial Distribution

Fabio Pareschi*, *Member, IEEE,* Riccardo Rovatti, *Fellow, IEEE,* and Gianluca Setti, *Fellow, IEEE*

***Abstract*—In this paper we review some statistical tests included in the NIST SP 800-22 suite, which is a collection of tests for the evaluation of both true-random (*physical*) and pseudorandom (*algorithmic*) number generators for cryptographic applications. The output of these tests is the so-called p-value which is a random variable whose distribution converges to the uniform distribution in the interval $[0, 1]$ when testing an increasing number of samples from an ideal generator. Here, we compute the exact non-asymptotic distribution of p-values produced by few of the tests in the suite, and propose some computation-friendly approximations. This allows us to explain why intensive testing produces false-positives with a probability much higher than the expected one when considering asymptotic distribution instead of the true one. We also propose a new approximation for the *Spectral Test* reference distribution, which is more coherent with experimental results.**

***Index Terms*—CRY-OTHE**

I. Introduction

Statistical hypothesis testing is a classical approach used to evaluate whether an experimental set of data fits with a given hypothesis (the *Null Hypothesis*, usually indicated with $\mathcal{H}_0$).

In information technology, statistical tests are largely adopted in random number generators (RNGs) testing [1], [2], [3], where the tested experimental data is a long sequence of generated numbers, and $\mathcal{H}_0$ is the hypothesis that the generator under test is *ideal*, i.e., the generated symbols are *independent and identically distributed (iid)*, which means that they are independent of each other, and distributed according to the same desired probability distribution (usually uniform or normal).

One of the advantages of this approach is that it does not require any assumption on the generator under test, since it only looks for evidence of particular statistical recurrences in a generated (and allegedly random) stream. If present, these features may distinguish the generator under test from an ideal one. Note that this is coherent with the requirements of many security-related applications, especially cryptography, where a system is considered *secure* if discerning between a random stream and the encrypted signal is a computationally hard problem [4], [5], [6]. This also makes these tests suitable for both true-random generators (based on some *physical* phenomenon that is intrinsically random) and pseudorandom generators (based on a recursive algorithm on a finite state machine). Different approaches proposed in RNG testing require additional hypotheses, such as the knowledge of the physical or algorithmic nature of the generator [7], [8], [9].

In this paper we consider the statistical test suite SP 800-22 [2] first published in 2001 by the US National Institute of Standard and Technology (NIST) and recently revised (last known update is Apr. 2010). It consists of 15 p-value based tests[1] which take a n bit sequence as input and whose output p is a real number in $[0, 1]$. Under the assumption $\mathcal{H}_0$, p is modeled as a random variable uniformly distributed in its definition set.

In particular we take into account three tests of the suite:

- the *Frequency Test*
- the *Runs Test*
- the *Spectral Test* (also known as the *Discrete Fourier Transform Test*)

For these three tests[2], which are based on the *binomial distribution*, we provide a feasible approximation of the *exact* distribution of p assuming $\mathcal{H}_0$. In this way we can estimate when a statistical test is *reliable*, i.e., when there is a low probability of erroneously identify an ideal generator as not random. This is particularly important when considering the two-level approach, i.e., when testing N different generated sequences to check if the N obtained p-values are aligned with the expected distribution, since this test is known to always fail when considering very large values of N [12], [13].

Furthermore, for the Spectral Test, we propose a refinement of the semi-empirically found reference distribution [14], [15], which seems more compatible with simulated results.

The paper is organized as follows. In Section II we will introduce both standard and two-level testing approaches, providing also an example to understand why the two-level approach may be preferred, and why an error in the expected p-

Copyright (c) 2010 IEEE. Personal use of this material is permitted. However, permission to use this material for any other purposes must be obtained from the IEEE by sending a request to pubs-permissions@ieee.org.

F. Pareschi and G. Setti are with ENDIF - Engineering Department, University of Ferrara, Via Saragat 1, 44122 Ferrara, Italy. E-mail: {fabio.pareschi,gianluca.setti}@unife.it.

R. Rovatti is with DEIS - Department of Electronic, Computer science and Systems, University of Bologna, Via Risorgimento 2, 40136 Bologna, Italy. Email: riccardo.rovatti@unibo.it.

All authors are also with ARCES research center, University of Bologna, Via Toffano 2, 40125 Bologna, Italy.

[1] In the original document released in 2001 under the name of "A statistical test suite for random and pseudorandom number generator for cryptographic applications", 16 tests were present; however, the *Lempel-Ziv Compression Test* was removed afterward due to errors in the test reference distribution.

[2] Note that in the literature, different tests are known with the above names (see, for example, [1], [10], [11]); in this paper, with *Frequency*, *Runs* and *Spectral Tests* we always refer to the NIST version.

value distribution may lead to an erroneous test interpretation. In Section III we provide some mathematical details on statistical tests that are used in Section IV to suggest, for the three tests above, an approximated expression F'_p of the actual Cumulative Distribution Function (CDF) F_p of the p-value. With this expression, in Section V we are able to mathematically express a reliability condition for the two-level approach. Finally, we draw some conclusions. To improve the readability of the paper, almost all of the mathematical details of the computation of the F'_p are reported in the Appendices.

II. Standard/Two-level Statistical Testing Approach

In the following we present a brief introduction to the standard and level-two approaches for statistical tests for randomness in order to define two main properties: *significance* and *power*. As already outlined in the Introduction, $\mathcal{H}_0$ is that the sequence under test is composed by *iid* events (and distributed according to the desired distribution); the output p of the test is a random variable in $[0, 1]$ such that its CDF is $F_p^{(iid)}(x) = x$ when $\mathcal{H}_0$ is true, and $F_p^{(\not{iid})}(x) \neq x$ when $\mathcal{H}_0$ is false [2], [16], [17].

A. Standard (One-level) Testing Approach

Let us generate and test a sequence of n events. Then, let us compare the p-value p of this sequence with a significance level, usually indicated with α.

- If $p \leq \alpha$, then $\mathcal{H}_0$ is rejected and the generator under test is considered a non-ideal RNG.
- If $p > \alpha$ there is no evidence to reject $\mathcal{H}_0$, so the generator under test is accepted as ideal.

When $\mathcal{H}_0$ is true and $p \leq \alpha$ we have a *false positive* in the test interpretation. This is called Type I Error and its probability is

$$\Pr\left(p \leq \alpha \,|\, \mathcal{H}_0\right) = F_p^{(iid)}(\alpha) = \alpha$$

This is defined as the *statistical significance* of the test and should be set to a very small value (NIST suggests $\alpha = 0.01$).

A false negative (accepting the sequence as random when $\mathcal{H}_0$ is false) is called Type II Error, and its probability is

$$\Pr\left(p > \alpha \,|\, \not{\mathcal{H}_0}\right) = 1 - F_p^{(\not{iid})}(\alpha) = \beta \tag{1}$$

As long as both α and β are small, the approach is effective. Note in Equation (1) that the lower α, the higher β and vice-versa. The choice of α is a trade-off between the two possible errors. This approach is known as *one-side testing*[3] and is followed by NIST in [2].

The value $1 - \beta$ is also called *statistical power* of the test and is its main figure of merit. However, its exact computation is not possible and also not sensible, since it depends on the specific (non-ideal) generator under test. This is the main drawback of this approach (more generally, of the entire statistical testing approach) and the reason why many authors prefer different approaches in RNG testing, which however require many hypotheses on the generator under test [7], [8].

[3]Other authors prefer *two-side tests* [1], [3], i.e., they reject $\mathcal{H}_0$ when $p \leq \alpha/2$ or $1 - \alpha/2 \leq p$. The formal treatment is the same with anly a difference in the mathematical formulation of β in (1).

B. Two-level Testing Approach

In the two-level (also known as second-level) testing approach, a long sequence of events is partitioned into N sequences, each with n bits. A standard test is repeated for each sequence, and the distribution of the N obtained p-values is compared with $F_p^{(iid)}$. Note that, assuming $\mathcal{H}_0$, the N sequences (and so the N p-values) are independent of one another.

This approach has been known for a long time [1] and it may increase the testing power with respect to a standard approach [12]. This effect can be seen in the example of Table I, where we present testing results for two high-quality true-random number generators. The first one is a RNG designed by ourselves, which exploits chaotic dynamics [18]. The second one is based on a quantum effect (the reflection of a single photon on a semi-transparent mirror [19]). The example involves the two ways suggested by NIST [2, chap. 4] to perform a two-level test:

(a) given N sequences, let ζ be the fraction of them passing a basic test, i.e., with $p > \alpha$. If N is large enough, ζ can be approximated with a normal random variable, with mean $\mu = 1 - \alpha$ (for the suggested parameters, $\mu = 0.99$, so let us refer to this case as *99% test*) and standard deviation $\sigma = \sqrt{\alpha(1-\alpha)/N}$. $\mathcal{H}_0$ is rejected if ζ lies outside the significance interval $1 - \alpha \pm 3\sigma$.

(b) given N sequences, test the distribution of the N p-values against the uniform distribution with a *chi-square* goodness-of-fit test in k bins. This is again a statistical test, which yields a level-two p-value p_T. Given a significance α_T, $\mathcal{H}_0$ is rejected if $p_T \leq \alpha_T$. The NIST suggests to use $k = 10\,\text{bins}$ and $\alpha_T = 0.0001$.

In order to have the same significance in all tests, we have set $\alpha = \alpha_T = 0.01$ and restricted the significance interval in the 99% test to $1-\alpha \pm 2.575\sigma$. The higher power achieved by the two-level approaches in this particular example is clear.

C. Reliability of a Two-level Test

The choice of N in a two-level approach is usually a trade-off [20], [21]. For example, given $n \cdot N$ basic events, n may be limited by the memory size or by the computational power available. In other cases, one may prefer to intentionally limit n to test the short-term behavior.

It is also known [12], [13] that for extremely large values of N, the level-two approach always fails, i.e., it ends with $p_T \simeq 0$. In this case we can say that the test is *not reliable*, since its significance is sensibly different from the expected one. From this point of view, N is also *a trade-off between power and reliability* of a level-two test.

As a further example, we have repeated the chi-square based two-level test on sequences obtained with the Blum-Blum-Shub (BBS) algorithm [5], which is known to asymptotically generate, at the cost of a very complex design, sequences almost indistinguishable from random ones [10], [6]. Using a pseudorandom generator is necessary to allow us to increase N at will. Results are shown in Table II, and confirm that for $N = 10^6$, almost no test is passed.

	single test		*99% test, 1000 seq.*		χ^2 *test, 1000 seq.*	
	Chaos-based	Quantis	Chaos-based	Quantis	Chaos-based	Quantis
Frequency	0.399902	0.514179	**0.948**	0.989	**0.000000**	0.402962
Block Frequency	0.437357	0.697836	0.989	0.991	0.342451	0.875539
Cumulative Sums	0.390490	0.411002	**0.940**	0.988	**0.000000**	0.612148
Runs	0.266725	0.585521	0.991	**0.981**	0.943242	**0.000001**
Longest Run of 1s	0.546938	0.536850	0.985	0.992	0.433590	0.203351
Matrix Rank	0.032352	0.715837	0.987	0.992	0.192724	0.083018
Spectral (DFT)	0.657982	0.797719	0.993	0.986	0.281232	0.048404
NOT Matching	0.656424	0.558425	0.983	0.990	0.212184	0.066882
OT Matching	0.856691	0.734679	**0.978**	0.987	**0.000006**	0.100109
Universal	0.855019	0.941268	0.985	0.992	0.340858	0.699313
Approx. Entropy	0.197721	0.373298	0.990	0.988	0.408275	0.440975
Random Excursion	0.346350	0.406225	0.989	0.983	0.374107	0.114584
Random Exc. Var.	0.526539	0.140243	0.987	0.995	0.055996	0.922084
Serial	0.722900	0.466486	0.988	0.993	0.230755	0.076658
Linear Complexity	0.236681	0.025910	0.993	0.989	0.073417	0.958485

TABLE I

RESULTS OF SP800-22 RANDOMNESS TEST FOR THE CHAOS-BASED RANDOM GENERATOR [18] AND THE QUANTIS GENERATOR [19], CONSIDERING BOTH THE STANDARD AND THE TWO-LEVEL NIST APPROACHES. TESTS WITH $p \leq 0.01$ (SINGLE TEST), WITH ζ OUTSIDE THE SIGNIFICANCE INTERVAL $1-\alpha \pm 2.575\sigma = 0.99 \pm 0.0081$ (99% TEST) OR WITH $p_T \leq 0.01$ (CHI-SQUARE TEST) ARE IN BOLD.

	χ^2 *test, N sequences*			
	$N = 10^3$	$N = 10^4$	$N = 10^5$	$N = 10^6$
Frequency	0.471146	0.106057	0.340080	**0.000000**
Block Frequency	0.848027	0.904981	0.295910	0.051232
Cumulative Sums	0.055010	0.612563	0.082745	0.026463
Runs	0.872425	0.857181	0.762307	0.024573
Longest Run of 1s	0.442831	0.197132	**0.000510**	**0.000000**
Matrix Rank	0.518106	0.134558	**0.000001**	**0.000000**
Spectral (DFT)	0.383827	0.305894	**0.000000**	**0.000000**
NOT Matching	0.112047	0.193871	0.606767	0.389343
OT Matching	0.131122	**0.000475**	**0.000000**	**0.000000**
Universal	0.239266	0.114846	**0.000000**	**0.000000**
Approx. Entropy	0.046870	0.018349	0.297963	**0.000000**
Random Excursion	0.730758	0.724267	0.205800	**0.000000**
Random Exc. Var.	0.294817	0.629095	**0.000020**	**0.000000**
Serial	0.467322	0.248395	0.935605	0.489325
Linear Complexity	0.948298	0.945195	0.368604	0.125629

TABLE II

RESULTS OF THE CHI-SQUARE BASED TWO-LEVEL RANDOMNESS TEST FOR THE BBS GENERATOR, WITH N RANGING FROM $N = 1,000$ TO $N = 1,000,000$. TESTS WHERE $p_T \leq 0.01$ ARE IN BOLD.

According to [12], this problem is due to approximation errors in the computation of the p-value of basic tests, which result in a deviation of the distribution of p from the expected uniform one. Thanks to Berry and Esséen inequality [22], in [12] we were able to compute an upper bound in the Frequency Test for the error between the approximated p-value p and the actual one p_0:

$$\max_{0 \leq p_0 \leq 1} |p - p_0| \leq \varepsilon_p = 2\,\frac{C\,\mathrm{E}\left[|X^{(i)}|^3\right]}{\sigma^3\sqrt{n}}$$

where n is the number of events $X^{(i)}$ in the basic test; σ and $\mathrm{E}\left[|X^{(i)}|^3\right]$ respectively the standard deviation and third order absolute moment of $X^{(i)}$; and $C \simeq 0.8$. Since $X^{(i)} \in \{-1, +1\}$, it is both $\sigma = 1$ and is $\mathrm{E}\left[|X^{(i)}|^3\right] = 1$; if $n = 10^6$, then $\varepsilon_p \simeq 1.6 \cdot 10^{-3}$.

From this, and assuming a chi-square test in k bins, the maximum error in the number of p-values in a bin is $\Delta_p = 2N\varepsilon_p$ independently of k. A very simple reliability condition is requiring that Δ_p is smaller than the variance of the distribution of N p-values in k bins, i.e., $\Delta_p < \sqrt{N(k-1)}/k$. In the standard case with $k = 10$ and $n = 10^6$, we get $N < 5722$.

In the following we make two steps further. First, we propose a feasible approximation of the exact CDF of the p-values in the three above mentioned tests. Then, this approximation is used to estimate the error in the computation of p_T. This approach is extremely useful when analyzing the reliability of the two-level testing approach.

III. MATHEMATICAL BACKGROUND ON STATISTICAL TESTS

Given n random events $X^{(0)}, X^{(1)}, \ldots, X^{(n-1)}$ with a continuous distribution and with $X^{(i)} \in \mathbb{X}$, let $\mathcal{H}_0$ be that the $X^{(i)}s$ are *iid*. A one-side statistical test is given by the function

$$T = T\left(X^{(0)}, X^{(1)}, \ldots, X^{(n-1)}\right), \quad T : \mathbb{X}^n \mapsto \Theta.$$

where T has to be defined over a normed space in order to define its expected value given $\mathcal{H}_0$, i.e., $T_0 = \mathrm{E}[T \,|\, \mathcal{H}_0]$ and the random variable $\xi = \|T - T_0\| \in \mathbb{R}$. The CDF of ξ is given by $F_{\|\cdot\|}(x) = \Pr(\xi \leq x \,|\, \mathcal{H}_0)$; to get a random variable uniformly distributed in $[0, 1]$ we can consider

$$p = 1 - F_{\|\cdot\|}(\xi) = 1 - F_{\|\cdot\|}(\|T_{obs} - T_0\|).$$

which is the definition of p-value.

Note however that if the $X^{(i)}s$ are discrete variables, also T and ξ are discrete (assuming n limited). In this case, the cardinality of the set of possible p-values $\mathcal{P} = \left(1 - F_{\|\cdot\|}\right) \circ T\left(\mathbb{X}^n\right)$ is $|\mathcal{P}| < \infty$. Assuming $\bar{p} \in \mathcal{P}$, with $\bar{p} = 1 - F_{\|\cdot\|}\left(\bar{\xi}\right)$, we can write $\Pr(p < \bar{p} \,|\, \mathcal{H}_0) = \Pr\left(\|T - T_0\| > \bar{\xi} \,|\, \mathcal{H}_0\right) = 1 - \Pr\left(\|T - T_0\| \leq \bar{\xi} \,|\, \mathcal{H}_0\right) = \bar{p}$ (note that $1 - F_{\|\cdot\|}$ is monotonically not increasing). This means that the CDF of p assuming $\mathcal{H}_0$ is:

$$F_p^{(iid)}(x) = \Pr(p \leq x \,|\, \mathcal{H}_0) = \inf_{p \in \mathcal{P}, p > x} p$$

which means $F_p^{(iid)}(x) \neq x$. Indeed, $\lim_{n\to\infty} F_p^{(iid)}(x) = x$ pointwise, i.e., p asymptotically converges [22] to a continuous random variable uniformly distributed in $[0, 1]$.

Furthermore, since $F_{\|\cdot\|}$ depends on n, it is common to compute p using $F^{\infty}_{\|\cdot\|} = \lim_{n\to\infty} F_{\|\cdot\|}$ which is usually a continuous function easier to implement with respect to the $F_{\|\cdot\|}$. This however generates an additional source of error in the CDF of p, which can now be written as

$$F_p(x) = x + \left(\inf_{p\in\mathcal{P},p>x} p - x\right) + \left(F^{\infty}_{\|\cdot\|}(\xi) - F_{\|\cdot\|}(\xi)\right) \quad (2)$$

where ξ is implicitly defined as any value for which $1 - F^{\infty}_{\|\cdot\|}(\xi) = \inf_{p\in\mathcal{P},p>x} p$, and where the two contributions discussed above are clearly identifiable.

In the general case, the most important cause of error is given by $|\mathcal{P}| < \infty$, i.e., by the *discrete* distribution of p. Intuitively, the higher $|\mathcal{P}|$, the smaller the distance between F_p and the continuous uniform CDF. This issue is already known in the literature. For example to minimize its effect in a two-side test, L'Ecuyer suggests in [1] to compute two different p-values, a right p-value p_R and a left p-value p_L.[4]

The main contribution of this paper is that, limiting ourselves to *binomial-based* statistical tests, i.e., tests where $F_{\|\cdot\|}$ is the combination of binomial coefficients, and where $F^{\infty}_{\|\cdot\|}$ is the CDF of a normal random variable, we are able to find a closed-form approximated expression F'_p for the CDF F_p in (2).

With this aim in mind, let us consider the following

Definition 1: A *generic binomial test* is a test where the basic event is the binomial variable:

$$X^{(i)} = \begin{cases} 1 & \text{with probability } u \\ 0 & \text{with probability } 1-u \end{cases}$$

and the test function is the sum of n basic events

$$T = \sum_{i=0}^{n-1} X^{(i)}$$

In this test, $\mathbb{X} = \{0,1\}$ and $\Theta = \{i \in \mathbb{N} \,|\, 0 \le i \le n\}$. However, according to the above considerations, we have to expand Θ to the set of reals $\mathbb{R}$ equipped with the usual absolute value function to have the normed space $(\mathbb{R}, |\cdot|)$. According to the central limit theorem, T has a normal limit distribution, with mean value $\mu = nu$ and variance σ^2 which depends on the $X^{(i)}$s correlation function.[5]

Proposition 1: In a generic binomial test, where the limit distribution of T is normal with mean value μ and variance σ^2

(a) the limit CDF $F^{\infty}_{\|\cdot\|}$ is given by

$$F^{\infty}_{\|\cdot\|}(\xi) = 2\,\Phi\left(\frac{\xi}{\sqrt{\sigma^2}}\right) - 1 = 1 - \operatorname{erfc}\left(\frac{\xi}{\sqrt{2\sigma^2}}\right) \quad (3)$$

where Φ is the standard normal CDF and erfc the complementary error function.

(b) the cardinality of $\mathcal{P}$ grows as n; more specifically

$$\frac{n+1}{2} \le |\mathcal{P}| \le n+1$$

(c) Indicating with ψ the fractional part of μ, i.e., $\psi = \mu_{(\bmod 1)}$, the CDF $F_p(x)$ can be approximated by

$$F'_p(x) = x + 2\,d(x)\; z\left(\sqrt{2\sigma^2}\operatorname{erfc}^{-1}(x)\right) \quad (4)$$

with

$$d(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(\operatorname{erfc}^{-1}(x))^2},$$

$$z(\xi) = \begin{cases} \xi_{(\bmod 1)} & \xi_{(\bmod 1)} < \min(\psi, 1-\psi) \\ \xi_{(\bmod 1)} - \frac{1}{2} & \min(\psi, 1-\psi) < \xi_{(\bmod 1)} < \max(\psi, 1-\psi) \\ \xi_{(\bmod 1)} - 1 & \xi_{(\bmod 1)} > \max(\psi, 1-\psi) \\ \lim_{x\to\xi^-} z(x) & \text{otherwise} \end{cases} \quad (5)$$

Where the last case in the definition of z ensures that z is left-continuous in its discontinuity points.

The proof can be found in Appendix A.

IV. Review of Binomial-based NIST Statistical Tests

On the basis of the results of Proposition 1, we review here the *Frequency Test*, the *Runs Test* and the *Spectral Test* included in the NIST suite. The mathematical notation is kept as close as possible to the one used in the NIST publication. These tests are applied to a sequence of random bits $X_i \in \{-1,+1\}$, and $\mathcal{H}_0$ is that the X_is are *iid uniform*, i.e., the desired distribution is $\Pr(X_i = -1) = \Pr(X_i = +1) = 1/2$.

A. Frequency Test

Given the input sequence $X_i = \{+1,-1\}$, $i = 0 \ldots n-1$, the balance between symbols "+1" and "−1" is

$$S_n = \sum_{i=0}^{n-1} X_i$$

Assuming the two symbols have the same probability, S_n is a zero average random variable which can be approximated, for large values of n, with a normal distribution with variance $\sigma^2 = n$. The p-value is given by

$$p = 2\left(1 - \Phi\left(\frac{|S_n|}{\sqrt{\sigma^2}}\right)\right) = \operatorname{erfc}\left(\frac{|S_n|}{\sqrt{2n}}\right) \quad (6)$$

Accordingly to the notation used in Definition 1, this is a standard binomial based test with $u = 1/2$ and where $T = S_n/2 + n/2$ is binomial distributed. The limit distribution of T is normal with $\mu = n/2$ and $\sigma^2 = n/4$ due to symbols independence. Note that if n is an even number (which is the most common case), the number $|\mathcal{P}|$ of different p-values is minimum, i.e., the error between the CDF F_p of the p-values distribution and the uniform CDF is maximum.

The approximated expression F'_p obtained using (4) has been compared with the experimental CDF we found testing different true-random and pseudorandom generators. In all considered cases, the matching between the observed distribution and the one given by F'_p was better with respect to the

[4]The density of the p-values is not uniform and may present a large difference between the left and the right endpoint of $[0,1]$. In this way, one of the two p-values is always computed where the density is high enough.

[5]Under the assumption that the $X^{(i)}$s are independent, $\sigma^2 = nu(1-u)$

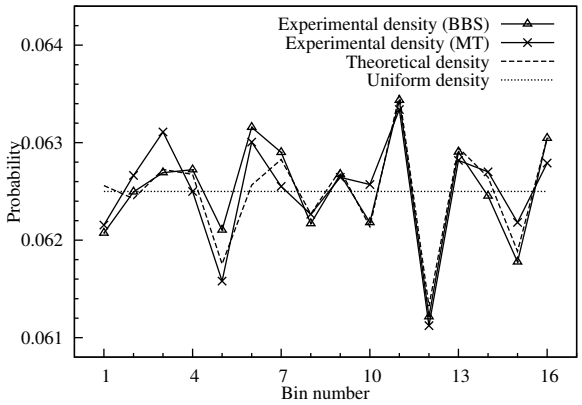


Fig. 1. Comparison between the uniform density, the theoretically expected one and the observed experimental one, discretized in $k = 16$ bins, for the p-values in the Frequency Test, with $n = 2^{20}$.

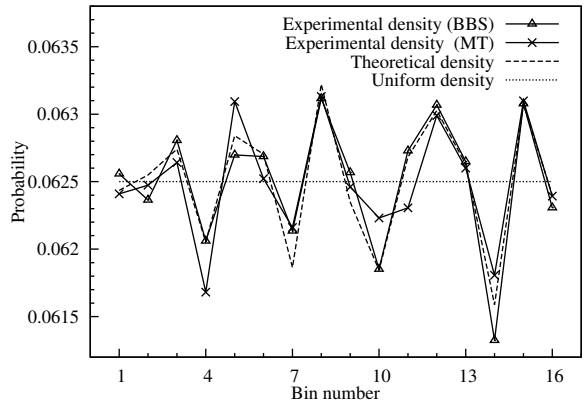


Fig. 2. Comparison between the uniform density, the theoretically expected one and the observed experimental one, discretized in $k = 16$ bins, for the p-values in the Runs Test, with $n = 100 \cdot 2^{10}$.

matching between the observed distribution and the uniform one. As an example, Figure 1 shows this comparison for the experimental distributions we get using the BBS algorithm [5] and the Mersenne-Twister algorithm [23] as pseudorandom sources, considering in both cases $N = 10^6$ sequences. The figure has been obtained dividing the interval $[0, 1]$ in $k = 16$ bins and computing the fraction of sequences giving a p-value in the considered bin. Results show that the theoretical expected density matches perfectly the experimental one.[6]

With the expression of F'_p, we can also compute the actual significance of this test. With the values suggested by NIST, i.e., $n = 10^6$ and $\alpha = 0.01$ the significance of this test under a standard approach is $F'_p(\alpha) = 1.0024 \cdot 10^{-2}$. If instead we use $n = 2^{20}$ and $\alpha = 0.01$ we get $F'_p(\alpha) = 1.00183 \cdot 10^{-2}$. Note that in both cases the error is small enough to consider the standard approach completely reliable.

B. Runs Test

Given the input sequence $X_i = \{+1, -1\}$, $i = 0 \ldots n-1$, let v_n be the total number of runs in the sequence. A run is an uninterrupted sequence of identical symbols, bound at both endpoints by the opposite symbol. Mathematically

$$v_n = 1 + \sum_{k=0}^{n-2} r_k, \quad \textit{where} \quad r_k = \begin{cases} 0 & \textit{if } X_k = X_{k+1} \\ 1 & \textit{otherwise} \end{cases}$$

Indicating with λ the fraction of symbols "+1" in the input sequence,[7] for n large, v_n is approximated with a normal distributed variable, with mean $\mu = 2n\lambda(1-\lambda)$ and variance $\sigma^2 = 4n\lambda^2(1-\lambda)^2$. The p-value is so given by

$$p = 2\left(1 - \Phi\left(\frac{|v_n - \mu|}{\sqrt{\sigma^2}}\right)\right) = \operatorname{erfc}\left(\frac{|v_n - 2n\lambda(1-\lambda)|}{2\sqrt{2n}\lambda(1-\lambda)}\right) \quad (7)$$

The Runs Test is slightly different from the above basic binomial test, since the $\mathcal{H}_0$ is that $n\lambda$ symbols "+1" and $n(1-\lambda)$ symbols "−1" are randomly distributed in a sequence of n symbols. This makes the Runs Test result independent of the probability u of the basic event (and so of the Frequency Test result) and is due to the fact that the NIST purpose was to collect a number of tests each one looking to a different statistical feature. However, the actual p-values distribution F_p depends on u.

We can prove that

Proposition 2: in the Runs Test:

(a) the cardinality of $\mathcal{P}$ increases as n^2; more specifically

$$|\mathcal{P}| \simeq \frac{n^2}{2}$$

(b) the actual CDF of the p-values F_p can be approximated with a sum of continuous functions

$$F'_p(x) = \sum_{v=0}^{n} F'^{(v)}_p(x; v) \quad (8)$$

The proof and the expression of $F'^{(v)}_p(x; v)$ in the case $u = 1/2$ can be found in Appendix B.

Note that the higher cardinality of $\mathcal{P}$ with respect to the Frequency Test case ensures a higher reliability. This could also be observed in the example of Table II.

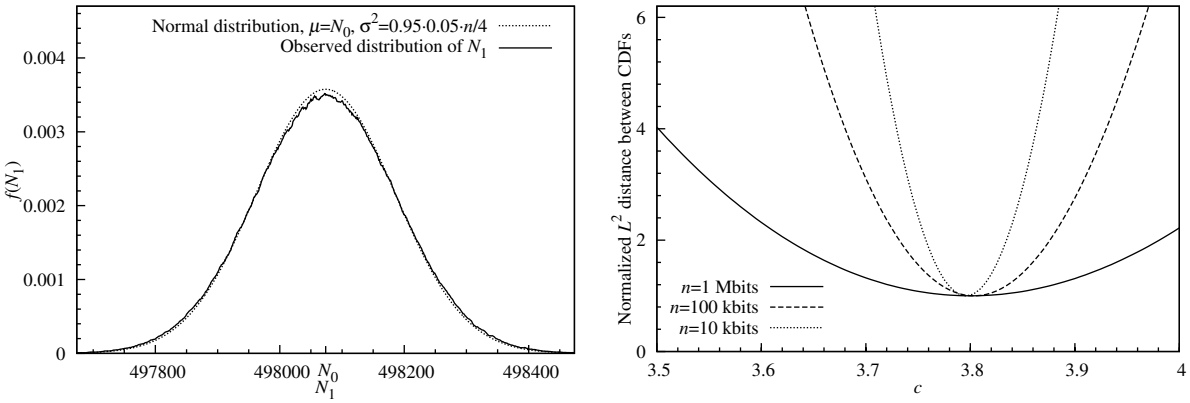
The distribution given by (8) has been compared in Figure 2 with the experimental ones found using the BBS and the Mersenne-Twister algorithms. As in the previous case, the matching is very good.[8]

The actual significance of the test with $n = 10^6$, and $\alpha = 0.01$ is $F'_p(\alpha) = 1.00059 \cdot 10^{-2}$, and considering $n = 2^{20}$ is $F'_p(\alpha) = 1.00057 \cdot 10^{-2}$.

[6]Note that, due to the limited number N of sequences considered, the experimental distribution is affected by a random variation, which can be estimated assuming a Gaussian approximation (the ratio of p-values in a bin can be considered a normal random variable with mean value approximately $\mu \simeq 1/k$ and variance approximately $\sigma^2 \simeq (k-1)/(Nk^2)$). In both considered cases $N = 10^6$, so $\sigma = 2.4 \cdot 10^{-4}$. Since the deviation from the uniform distribution in the experimental one is much higher than this value (the average distance is approximately $e_{\text{avg}} \simeq 4 \cdot 10^{-4}$) we can conclude that it is not due to random fluctuations.

[7]Note that in the NIST document the symbol π is used instead of λ; accordingly to [24] we prefer the latter symbol to avoid any misunderstanding with the classical constant $\pi \simeq 3.14$. Note also that, according to NIST, this test should be computed only if λ is close enough to $1/2$. We neglect this requirement for the sake of simplicity.

[8]Due to the smaller distance from the uniform distribution with respect to the Frequency Test case, ensuring a σ smaller than e_{avg} for $n = 2^{20}$ would require a N extremely high. For this reason, we set $n = 100 \cdot 2^{10}$; with these values and with $N = 10^6$, we have $e_{\text{avg}} = 3.7 \cdot 10^{-4}$ and $\sigma = 2.4 \cdot 10^{-4}$.



(a) (b)

Fig. 3. (a) Comparison between the theoretical probability density function (PDF) of N_1 according to [2], [14], [15] and the experimental one observed in $N = 10^6$ sequences generated with the BBS algorithm and with $n = 2^{20}$ bits. (b) Normalized L^2 distances between the observed PDF of N_1 in the Spectral Test and a normal PDF with $\sigma^2 = 0.95 \cdot 0.05 \cdot n/c$, with c ranging from 3.5 to 4, and with $n = 2^{20}$ bits, $n = 100 \cdot 2^{10}$ bits and $n = 10 \cdot 2^{10}$ bits. In all cases, the minimum distance is achieved for $c \simeq 3.8$.

C. Spectral Test

Given the input sequence $X_i = \{+1, -1\}$, $i = 0 \ldots n-1$, compute its Discrete Fourier Transform (DFT). Under the assumption that the sequence comes from an ideal generator, 95% of the bins in the unilateral *frequency spectrum should have an amplitude smaller than a threshold value $T_h = \sqrt{-\ln(0.05) \cdot n}$. The effective number of bins N_1 having an amplitude smaller than T_h converges to a normally distributed random variable, with mean value $\mu = 0.95 \cdot n/2 = N_0$, and variance $\sigma^2 = 0.95 \cdot 0.05 \cdot n/3.8$. So, the p-value is computed as*

$$p = 2\left(1 - \Phi\left(\frac{|N_1 - N_0|}{\sqrt{\sigma^2}}\right)\right) = \mathrm{erfc}\left(\frac{|N_1 - N_0|}{\sqrt{2\sigma^2}}\right) \quad (9)$$

The reference distribution for N_1 indicated by NIST in [2] is a normal distribution with $\mu = N_0 = 0.95 \cdot n/2$ and $\sigma^2 = 0.95 \cdot 0.05 \cdot n/4$. This has been experimentally found independently by many authors [14], [15].

We have found that:

(a) the distribution of N_1 is better approximated by a normal distribution with $\sigma^2 = 0.95 \cdot 0.05 \cdot n/3.8$. We suggest to refine the experimental value for the variance of the reference distribution to this value;

(b) the p-values in the Spectral Test are distributed as in standard binomial test, with a number of symbols $n' = n/2$ and a limit normal distribution with mean value and variance as above.

As far as *(a)* is concerned, in Figure 3-a we have plotted the observed distribution of N_1 coming from our simulations using the BBS algorithm and the theoretical distribution expected according to the NIST document. The variance of the experimental curve seems a bit larger than expected. The best fitting between the observed distribution and a normal distribution with variance $\sigma^2 = 0.95 \cdot 0.05 \cdot n/c$ (see Figure 3-b) is achieved for $c \simeq 3.8$ for many different values of n.

As far as *(b)* is concerned, Figure 4 shows the p-values CDFs experimentally found using the BBS and the Mersenne-Twister algorithms compared with the expected and the uniform one, all discretized in 16 bins in the case $n = 2^{20}$. When considering $\sigma^2 = 0.95 \cdot 0.05 \cdot n/4$ (the case *(a)*) the two distributions does not match; furthermore, the observed experimental density does not seem to converge to the uniform one. On the contrary, using $\sigma^2 = 0.95 \cdot 0.05 \cdot n/3.8$ (case *(b)*) the matching is very good.

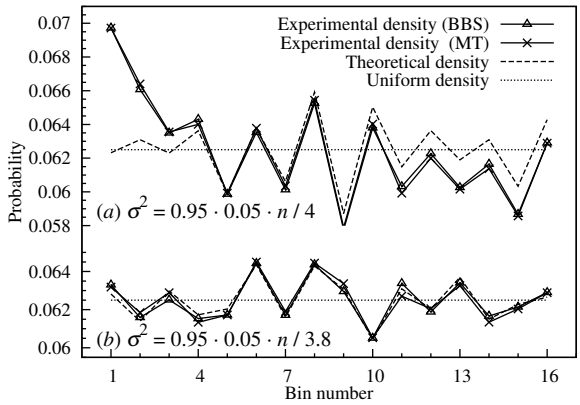


Fig. 4. Comparison between the uniform density, the theoretically expected one and the observed experimental ones achieved with the BBS and the Mersenne-Twister algorithms, all discretized in $k = 16$ bins, for the p-values in the Spectral Test, with $n = 2^{20}$, for the cases *(a)* the variance is computed as $\sigma^2 = 0.95 \cdot 0.05 \cdot n/4.0$ as indicated by the NIST document [2]; *(b)* the variance is computed as $\sigma^2 = 0.95 \cdot 0.05 \cdot n/3.8$. Note that the experimental densities in *(a)* and *(b)* are different since the p-values are computed with two different functions. The matching in the case *(b)* is almost perfect.

Using the refined proposed value for the variance in the distribution of N_1, the actual significance of the test under the standard approach is, considering $n = 10^6$ and $\alpha = 0.01$, $F'_p(\alpha) = 1.01258 \cdot 10^{-2}$, while when considering $n = 2^{20}$, the significance is $F'_p(\alpha) = 0.99742 \cdot 10^{-2}$.

V. Reliability of a Two-level Test

Assuming to know the exact p-value CDF F_p, when performing a two-level test, there are two possible options:

- compare the obtained distribution of the p-values with the theoretical one to get the correct level two p-value p_{T0};

- use the continuous uniform distribution as the reference one, accepting to compute an erroneous level-two p-value p_T.

The first solution is of course the optimal one. However, in many cases (for example, for the sake of simplicity) the second approach may be preferred. In the following we consider this second case, estimating the error between p_T and p_{T0}. In this way we are able to formulate a mathematical condition for the reliability of a two-level test, which is expressed by requiring that $|p_T - p_{T0}|$ is smaller than a given value.

In addition to the 99% and the chi-square based two-level tests proposed by NIST, we consider also a Kolmogorov-Smirnov based approach. For the last two tests we provide a basic introduction in order to understand the proposed condition. We also provide a few examples of designing a reliable level-two test based on the Frequency Test.

A. 99% based Two-level Test

The fraction ζ of sequences passing a basic test is expected to be in the range $1 - F_p(\alpha) \pm 3\sigma$, and *not* in the range $1 - \alpha \pm 3\sigma$.

In a two-level test, a first simple reliability condition is requiring that the above two intervals are almost superimposing. Since $F_p(\alpha) \simeq \alpha$, we can approximate $\sigma^2 = F_p(\alpha)(1 - F_p(\alpha))/N \simeq \alpha(1-\alpha)/N$. The reliability condition is

$$|F_p(\alpha) - \alpha| \ll \sqrt{\frac{\alpha(1-\alpha)}{N}}$$

i.e., $\sqrt{N} \ll \sqrt{\alpha(1-\alpha)}/\left|F_p(\alpha) - \alpha\right|$.

Example - Let us consider the Frequency Test, with $n = 2^{20}$, and $\alpha = 0.01$. For these parameters (see Section IV-A) $\left|F_p'(\alpha) - \alpha\right| \simeq 1.8 \cdot 10^{-5}$. The reliability condition is $\sqrt{N} \ll 5500$.

If a more precise condition is required, a p-value approach for this two-level test has to be considered. Counting the number of sequences $N\zeta$ which pass a basic test is in fact a standard binomial test (see Definition 1) with $\mu = N(1 - F_p(\alpha))$ and $\sigma^2 \simeq N\alpha(1-\alpha)$, where the p-value is given by

$$p_{T0} = 1 - F_{99\%}(\theta_0) = \operatorname{erfc}\left(\frac{|N\zeta - N\left(1 - F_p(\alpha)\right)|}{\sqrt{2N\alpha\left(1-\alpha\right)}}\right)$$

with $\theta_0 = \sqrt{N}\,|\zeta - (1 - F_p(\alpha))|$ and where the function $F_{99\%}(x)$ is implicitly defined.

When assuming $\mu = N(1-\alpha)$, the p-value is computed as $p_T = 1 - F_{99\%}\left(\theta\right)$ with

$$\theta = \sqrt{N}\,|\zeta - (1-\alpha)| =$$
$$= \sqrt{N}\,|\zeta - (1 - F_p(\alpha)) - (F_p(\alpha) - \alpha)| \leq \theta_0 + \sqrt{N}\,|F_p(\alpha) - \alpha|$$

Let us consider the worst case scenario (i.e., that $\hat{\theta} = \theta_0 + \sqrt{N}\,|F_p(\alpha) - \alpha|$) where the error between p_{T0} and $\hat{p}_T = 1 - F_{99\%}(\hat{\theta})$ is maximized. Indicating with $f_{99\%}(x) = dF_{99\%}(x)/dx$, and considering the two-term series expansion of $F_{99\%}(\theta)$ around θ_0, we get

$$\hat{p}_T = 1 - F_{99\%}\left(\theta_0 + \sqrt{N}\,|F_p(\alpha) - \alpha|\right) \simeq$$
$$\simeq 1 - F_{99\%}\left(\theta_0\right) - f_{99\%}\left(\theta_0\right)\sqrt{N}\,|F_p(\alpha) - \alpha|$$

which directly yields to

$$\hat{p}_T - p_{T0} \simeq -\sqrt{N} f_{99\%}\left(F_{99\%}^{-1}(1 - p_{T0})\right)|F_p(\alpha) - \alpha| \quad (10)$$

where $F_{99\%}^{-1}$ is the inverse function of $F_{99\%}$.

The error $|\hat{p}_T - p_{T0}|$:

(a) increases as $\sqrt{N}$;
(b) linearly depends on the function $f_{99\%}\left(F_{99\%}^{-1}(1 - p_{T0})\right)$ which does not depend on N.
(c) linearly depends on $|F_p(\alpha) - \alpha|$, i.e., on the error in the significance of the basic test approach.

Equation (10) ensures the reliability of a two-level test.

Example 1 - Let us consider a Frequency Test with $n = 2^{20}$ and $\alpha = 0.01$, and require a maximum error on the two-level p-value $|\hat{p}_T - p_{T0}| < 0.01$ on the whole range $0 \leq p_{T0} \leq 1$. In this case $\left|F_p'(\alpha) - \alpha\right| \simeq 1.8 \cdot 10^{-5}$. Furthermore

$$\sup_{0 \leq p_{T0} \leq 1} f_{99\%}\left(F_{99\%}^{-1}\left(1 - p_{T0}\right)\right) = 8.02$$

which means $N < 4798$.

Example 2 - Under the same assumptions as before, we are interested in a two-level test with $\alpha_T = 0.01$. In this case we need accuracy only for the p-values which lay around the value $p_{T0} = 0.01$, for example requiring that $|p_T - p_{T0}| < 0.001$. In this case

$$f_{99\%}\left(F_{99\%}^{-1}\left(1 - \alpha_T\right)\right) = 0.291$$

which results in $N < 36448$.

B. Chi-square Based Two-level Test

Let us distribute N samples $X^{(i)}$ in k subgroups, namely bins. *If the $X^{(i)}$s are continuous random variables, the bins are obtained as a partition of the definition set of the $X^{(i)}$s; let also π_j be the probability that a sample $X^{(i)}$ is in the j-th bin, with $j = 1, \ldots, k$. The observed number O_j of samples belonging to the j-th bin is compared with the expected number $E_j = N\pi_j$; the distance between O_j and E_j is given by:*

$$\chi^2 = \sum_{j=1}^{k} \frac{\left(E_j - O_j\right)^2}{E_j} \quad (11)$$

For a random input sequence, this is a random variable distributed according to a chi-square distribution with $k-1$ degree of freedom, so [25]

$$p_T = 1 - F_{\chi^2}\left(\chi^2\right) = 1 - \frac{\gamma\left(\left(k-1\right)/2; \chi^2/2\right)}{\Gamma\left(\left(k-1\right)/2\right)} \quad (12)$$

where $\gamma\left(k; x\right)$ and $\Gamma\left(k\right)$ are respectively the incomplete and the complete gamma functions.

In a chi-square based two-level test involving N p-values, when dividing the interval $[0, 1]$ uniformly in k bins $\left[\frac{j-1}{k}, \frac{j}{k}\right]$, with $j = 1, 2, \ldots, k$, the expected number of p-values in the j-th bin is

$$E_j = N\left(F_p\left(\frac{j}{k}\right) - F_p\left(\frac{j-1}{k}\right)\right) \quad (13)$$

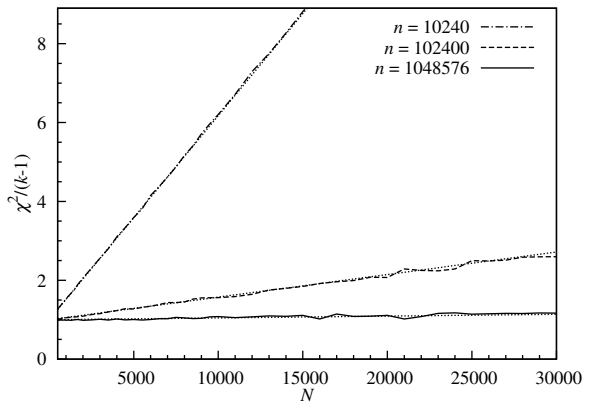


Fig. 5. Comparison between the average observed value of $\chi^2/(k-1)$ in a two-level Frequency Test for different values of N, with $k = 16$ and $n = 2^{20}$ (solid line), $n = 100 \cdot 2^{10}$ (dashed line), $n = 10 \cdot 2^{10}$ (dotted-dashed line), along with their values expected from (15) (dotted lines).

The random variable

$$\chi_0^2 = \sum_{j=1}^{k} \frac{\left(N\left(F_p\left(\frac{j}{k}\right) - F_p\left(\frac{j-1}{k}\right)\right) - O_j\right)^2}{N\left(F_p\left(\frac{j}{k}\right) - F_p\left(\frac{j-1}{k}\right)\right)}$$

has a chi-square distribution, and the two-level p-value is computed as $p_{T0} = 1 - F_{\chi^2}\left(\chi_0^2\right)$, where the CDF F_{χ^2} is implicitly defined in (12).

When assuming the continuous uniform distribution as the reference one, the expected number of p-values in each bin is $\tilde{E}_j = N/k$, and the variable

$$\begin{aligned}\chi^2 &= \sum_{j=1}^{k} \frac{\left(\tilde{E}_j - O_j\right)^2}{\tilde{E}_j} = \\ &= \sum_{j=1}^{k} \frac{\left(E_j - O_j - N\left(F_p\left(\frac{j}{k}\right) - F_p\left(\frac{j-1}{k}\right) - \frac{1}{k}\right)\right)^2}{\tilde{E}_j} = \\ &= \sum_{j=1}^{k} \frac{(E_j - O_j)^2}{\tilde{E}_j} + Nk \sum_{j=1}^{k} \left(F_p\left(\frac{j}{k}\right) - F_p\left(\frac{j-1}{k}\right) - \frac{1}{k}\right)^2 + \\ &\quad -2k \sum_{j=1}^{k} (E_j - O_j)\left(F_p\left(\frac{j}{k}\right) - F_p\left(\frac{j-1}{k}\right) - \frac{1}{k}\right)\end{aligned} \tag{14}$$

does not have a chi-square distribution.

Note however that the first term of (14) can be approximated with χ_0^2; the second term is a constant; and the third one is a random variable which depends on the O_j.

Let us consider the expected value $\bar{\chi}^2$ of χ^2 given χ_0^2, i.e., let us consider all O_j sequences giving χ_0^2 in (13). The O_j are not independent, since there are two constraints, the first given by χ_0^2 and the second by $\sum_j O_j = N$. However, since every O_j has E_j as expected value, the average contribution of the third term in (14) vanishes:

$$\bar{\chi}^2 \simeq \chi_0^2 + Nk \sum_{j=1}^{k} \left(F_p\left(\frac{j}{k}\right) - F_p\left(\frac{j-1}{k}\right) - \frac{1}{k}\right)^2 = \chi_0^2 + NC_{\chi^2} \tag{15}$$

where C_{χ^2} is implicitly defined and depends only on n and k.

Figure 5 shows a comparison between the expected value of the normalized[9] value $\bar{\chi}^2/(k-1)$ computed through (15) and the average value of $\chi^2/(k-1)$ from experimental results for different values of n and N. The matching between the curves confirms the validity of the above approximations.

From the knowledge of $\bar{\chi}^2$ it is possible to compute the error between p_{T0} and $\bar{p}_T = 1 - F_{\chi^2}\left(\bar{\chi}^2\right)$. By proceeding with a two-term series expansion

$$\bar{p}_T = 1 - F_{\chi^2}\left(\chi_0^2 + NC_{\chi^2}\right) \simeq 1 - F_{\chi^2}\left(\chi_0^2\right) - f_{\chi^2}\left(\chi_0^2\right) NC_{\chi^2}$$

i.e.,

$$\bar{p}_T - p_{T0} \simeq -f_{\chi^2}\left(F_{\chi^2}^{-1}\left(1 - p_{T0}\right)\right) NC_{\chi^2} \tag{16}$$

where $f_{\chi^2}(x)$ is the PDF of a chi-square distribution with $k-1$ degrees of freedom. The error $\bar{p}_T - p_{T0}$ is always negative ($f_{\chi^2}(x)$ is a PDF), and linearly depends on N.

Equation (16) can be used to verify or ensure the reliability of a chi-square based two-level test.

Example 1 - Let us consider the Frequency Test with $n = 2^{20}$ and suppose that we require an average error on the level-two p-value $|\bar{p}_T - p_{T0}| < 0.01$ over the whole range $0 \leq p_{T0} \leq 1$ using $k = 10$ bins. In this case

$$C_{\chi^2} \sup_{0 \leq p_{T0} \leq 1} f_{\chi^2}\left(F_{\chi^2}^{-1}\left(1 - p_{T0}\right)\right) = 1.67361 \cdot 10^{-6}$$

which means $N < 5975$. Note that this value is coherent with the value found in Section II-C and coming from [12].

Example 2 - Under the same assumptions as before and with $\alpha_T = 0.01$, we are interested in checking if $p_T \leq \alpha_T$. In this case we require that $|\bar{p}_T - p_{T0}| < 0.001$ for $p_T \simeq \alpha_T$. Since

$$C_{\chi^2} f_{\chi^2}\left(F_{\chi^2}^{-1}\left(1 - \alpha_T\right)\right) = 5.70634 \cdot 10^{-8}$$

we get $N < 17524$.

Example 3 - Let us consider the Frequency Test with $n = 2^{20}$ and $N = 10^4$. We want to know for which values of k the average error $|\bar{p}_T - p_{T0}| < 0.01$, i.e., for which k we get

$$C_{\chi^2} \sup_{0 \leq p_{T0} \leq 1} f_{\chi^2}\left(F_{\chi^2}^{-1}\left(1 - p_{T0}\right)\right) < 10^{-6}$$

Note that, despite the general trend, the sequence $C_{\chi^2} \sup_{0 \leq p_{T0} \leq 1} f_{\chi^2}\left(F_{\chi^2}^{-1}\left(1 - p_{T0}\right)\right)$ may not be strictly monotonically increasing in k. For this reason we can expect a set of solutions that is not a range of integers between two bounds. In fact, in the case of the example we have $k = \{3, 4, 5, 6, 9\}$.

C. Kolmogorov-Smirnov Based Two-level Test

Let us consider N samples $X^{(i)}$ with assumed continuous *CDF $F(x)$. Let us also define the empirical CDF $F_N(x)$ as*

$$F_N(x) = \frac{1}{N} \sum_{j=0}^{N-1} H\left(x - X^{(j)}\right)$$

[9]The expected value for a chi-square distributed random variable with $k-1$ degree of freedom is $k-1$.

where $H(x)$ is the unity-step function, defined as $H(x) = 1$ if $x \geq 0$ and 0 elsewhere. The Kolmogorov-Smirnov statistic for a given CDF $F(x)$ is

$$D_N = \sup_x |F_N(x) - F(x)|$$

Indicating with Q the CDF of D_N, the p-value is given by $p = 1 - Q(D_N)$.

The distribution of D_N have been intensively studied. It is known that $\sqrt{N} D_N$ converges, for large N, to a Kolmogorov distribution [26], which does not, however, fit well for small N. A comprehensive review of all the proposed approximations and their validity regions can be found in [27], from where the code used to compute Q in this paper is taken.

Being an extremely complex test from a mathematical point of view, we have to introduce strong hypotheses and approximations to build a theoretical framework as in the two previous cases. The first assumption we make is that

$$D_{N0} = \sup_x |F_N(x) - F_p(x)|$$

is distributed according to Q despite the fact that F_p is the CDF of a discrete random variable. Under this assumption, the correct p-value for this test is given by $p_{T0} = Q(D_{N0})$. Note that both F_N and F_p are piece-wise constant functions, and so $F_N - F_p$, where all discontinuities are in $x \in \mathcal{P}$; the maximum is achieved in an interval (p_i, p_j). Let us indicate with $\hat{x}$ a point of this interval.

When using the continuous uniform CDF as reference one, the computed p-value is $p_T = Q(D_N)$, with

$$D_N = \sup_x |F_N(x) - x| = \sup_x |F_N(x) - F_p(x) + F_p(x) - x| \quad (17)$$

A relation between D_{N0} and D_N can be found by analyzing separately the two contributions $F_N(x) - F_p(x)$ and $F_p(x) - x$ of (17). As observed above, $F_N - F_p$ is step-wise; $F_p(x) - x$ is a piece-wise linear, usually zero average, oscillating function (see, for example, Proposition 1). Let us indicate with $\tilde{\varepsilon}(x)$ the *envelope* of $|F_p(x) - x|$, i.e., a function which approximates the maximum value of $|F_p(x) - x|$ in any interval in which F_p is constant.

With the additional hypothesis that $\tilde{\varepsilon}$ is "smooth", i.e., that the steps of $F_n - F_p$ are higher than $\tilde{\varepsilon}$,[10] we can assume that the maximum for D_N is reached in the same interval in which D_{N0} has its maximum. In this case

$$D_N \simeq D_{N0} + \tilde{\varepsilon}(\hat{x})$$

Note that, given D_{N0}, the value of $\hat{x}$ (and so D_N) is not uniquely determined, i.e., $\hat{x}$ is still a random variable. Let us introduce $\bar{D}_N$ as the average value of D_N given D_{N0}, whose exact computation would require the knowledge of the statistic of $\hat{x}$. Yet, $F_N(x) - F_p(x)$ is a *Brownian bridge*, i.e., a Brownian motion pinned at both ends $x = 0$ and $x = 1$ ($F_N(x)$ and $F_p(x)$ are both CDF), so it is unlikely that $\hat{x}$ would lie in the area around 0 or 1.

[10]In a standard binomial test, it is $\tilde{\varepsilon}(x) = e^{-(\mathrm{erfc}^{-1}(x))}/\sqrt{2\pi\sigma^2} \propto 1/\sqrt{n}$, while the steps of $F_N - F_p$ has two contributions, one of which is equal to $1/N$ and the other one is proportional to $\tilde{\varepsilon}$. According to this, for values of N much smaller than n, this hypothesis is verified.

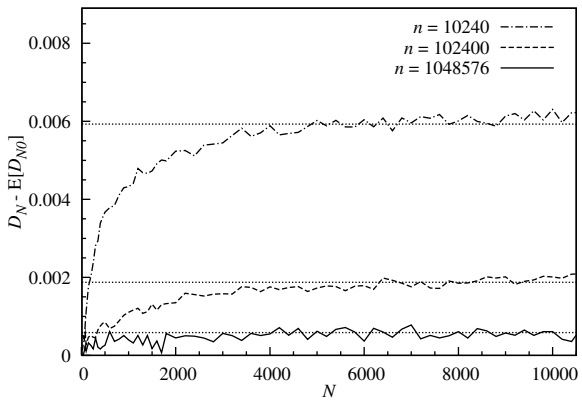


Fig. 6. Comparison between the average observed value of $D_N - \mathrm{E}[D_{N0}]$ in a two-level Frequency Test for different values of N, with $n = 2^{20}$ (solid line), $n = 100 \cdot 2^{10}$ (dashed line), $n = 10 \cdot 2^{10}$ (dotted-dashed line), along with the approximated theoretically expected values (dotted horizontal lines).

Limiting ourselves to the Frequency Test, we can numerically compute the expected value of $\tilde{\varepsilon}(\hat{x})$ assuming different distributions for $\hat{x}$. In all cases we get values very similar to $0.6/\sqrt{n}$, which can be assumed as the actual expected value, independently of D_{N0}

$$\bar{D}_N = D_{N0} + \mathrm{E}[\tilde{\varepsilon}(\hat{x}) | D_{N0}] \simeq D_{N0} \, \mathrm{E}[\tilde{\varepsilon}(\hat{x})] \simeq D_{N0} + \frac{0.6}{\sqrt{n}}$$

To get an experimental verification, we have computed the average D_N and compared with $\mathrm{E}[D_{N0}]$ in many simulations of a two-level Frequency Test for different values of n and N. The results, which can be seen in Figure 6, show (especially for n and N large) a good matching between the expected and the observed curves, proving that the approximations introduced in this section are acceptable.

Proceeding as in the previous cases, if $\bar{p}_T = 1 - Q(\bar{D}_N)$, the average error on the p-value is given by

$$\bar{p}_T - p_{T0} \simeq -q\left(Q^{-1}(1 - p_{T0})\right) \mathrm{E}[\tilde{\varepsilon}(\hat{x})] \quad (18)$$

with q the PDF associated to Q. Interestingly enough, the value of $q\left(Q^{-1}(1 - p_{T0})\right)/\sqrt{N}$ has a very weak dependence on N, and this makes easy to use Equation (18) to verify the reliability of a Kolmogorov-Smirnov based two-level statistical test.

Example 1 - Let us consider a Frequency Test with $n = 2^{20}$ and suppose we require an average error on the two-level p-value $|\bar{p}_T - p_{T0}| < 0.01$ on the whole range $0 \leq p_{T0} \leq 1$. Assuming $\mathrm{E}[\tilde{\varepsilon}(\hat{x})] \simeq 0.6/\sqrt{n}$, and since for N in the range of a few hundreds

$$\sup_{0 \leq p_{T0} \leq 1} q\left(Q^{-1}(1 - p_{T0})\right)/\sqrt{N} \simeq 1.69$$

we have $N < 102$.

Example 2 - Under the same assumptions as before, we require that $|\bar{p}_T - p_{T0}| < 0.001$ only for p-values neighboring the significance level $\alpha_T = 0.01$.

$$q\left(Q^{-1}(1 - \alpha_T)\right)/\sqrt{N} \simeq 0.065$$

with N in the range of a few hundreds. The reliability condition is ensured for $N < 685$.

Note that, according to these examples, a Kolmogorov-Smirnov based two-level test has a lower reliability compared to a chi-square based one.

VI. Conclusion

In this paper we have considered the Frequency Test, the Runs Test and the Spectral Test included in the NIST SP800-22 statistical test suite. For these tests we have proposed a *computationally feasible approximation* of the actual p-values distribution when testing an ideal random generator. This is a discrete distribution which may be considerably different from the uniform one.

The knowledge of this true distribution allows us to design a level-two test which is *reliable*, i.e., a test in which the probability of a false positive is aligned with the expected one. We have considered the level-two approaches proposed by NIST, i.e., counting the sequences passing a basic test and checking the p-value distribution with a chi-square test, as well as an additional approach consisting of checking the p-value distribution with a Kolmogorov-Smirnov test.

Furthermore, for the Spectral Test, we have proposed a refinement of the reference distribution for the p-value computation. With this refinement, the issues of these tests that are known and addressed in many other papers [14], [15] are strongly reduced.

Appendix A
Proof of proposition 1

Proof of (a) - Assuming $\mathcal{H}_0$, the PDF of T is a discrete function with binomial coefficients which can be approximated with the Central Limit Theorem:

$$f_r = \Pr(T = r) \simeq f_T(r) = \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{\frac{(r-\mu)^2}{2\sigma^2}} \tag{A.1}$$

where the f_r are defined for $r \in \{0, 1, \ldots, n\}$, and the function f_T is the extension of the f_r in $\mathbb{R}$. The CDF is

$$F_T(\theta) = \Pr(T \le \theta) = \sum_{r \le \theta} f_r \tag{A.2}$$

which can be approximated with the integral of (A.1)

$$F_T^{(lim)}(\theta) = \int_{-\infty}^{\theta} f_T(r)\, dr = 1 - \frac{1}{2}\operatorname{erfc}\left(\frac{\theta - nu}{\sqrt{2nu(1-u)}}\right) \tag{A.3}$$

Now, consider the distance between T and its mean value μ

$$\begin{aligned} F_{\|\cdot\|}^{\infty}(\xi) &= \Pr(|T-\mu| \le \xi) = \Pr(T \le \mu+\xi) - \Pr(T \le \mu-\xi) \\ &= F_T^{(lim)}(\mu+\xi) - F_T^{(lim)}(\mu-\xi) \end{aligned} \tag{A.4}$$

which leads directly to (3). □

Note that, despite the fact that $F_T^{(lim)}$ is globally a good approximation of F_T, the error in all points $\theta \in \{0, 1, \ldots, n\}$, i.e., in all the possible values of T, is quite large. To cope with this, we adopt in the following the *continuity correction*

$$F_T(\theta) \simeq F_T^{(lim)}\left(\theta + \frac{1}{2}\right), \quad \forall\, \theta \in \{0, 1, \ldots, n\} \tag{A.5}$$

Proof of (b) - Let us enumerate and order all different p-values. Since $p = 1 - F_{\|\cdot\|}^{\infty}(|r-\mu|)$ with $r = 0, 1, \ldots, n$, and $F_{\|\cdot\|}^{\infty}$ is strictly increasing, the number of different p-values is $|\mathcal{P}| = n+1$ except if couples (r_1, r_2) exist, for which

$$|r_1 - \mu| = |r_2 - \mu|, \quad r_1 \ne r_2 \tag{A.6}$$

It is easy to see that, indicating with $\psi = \mu_{(\bmod 1)}$, this is possible only for $\psi = 0$ or $\psi = 1/2$.

As an example, let us assume that $\psi = 0$ (i.e., μ is integer), and indicate with $\xi = |r-\mu|$; $|\mathcal{P}|$ is also given by the number of different values of ξ. Let us also enumerate $\xi_i,\ i = 0, 1, 2, \ldots$ all different values that ξ can assume, with $\xi_i < \xi_{i+1}$. In this case $\xi_i = i$, with i ranging from 0 to $n/2$ (in the case $\mu = n/2$) or to $n+1$ in the degenerate cases when $\mu \to 0$ or $\mu \to n$.

It is interesting in this case also the computation of $\tilde{f}_i = \Pr(p = p_i) = \Pr(\xi = \xi_i)$.

$$\tilde{f}_i = \begin{cases} f_\mu & \text{if } i = 0 \\ f_{\mu-i} + f_{\mu+i} & \text{otherwise} \end{cases} \tag{A.7}$$

since $\xi_0 = 0$ corresponds only to $r = \mu$, while all other values of ξ correspond to two values of r.

In a very similar way, it is possible to prove that in the case $\psi = 1/2$, $|\mathcal{P}|$ ranges from $(n+1)/2$ when $\mu = n/2$, to $n+1$ in the two degenerate cases. For all other values of ψ, we have $|\mathcal{P}| = n+1$. □

Proof of (c): According to the notation introduced above, the exact PDF f_p and CDF F_p can be respectively written as

$$f_p(x) = \sum_i \tilde{f}_i \delta(x - p_i) \tag{A.8}$$

$$F_p(x) = \sum_i \tilde{f}_i H(x - p_i) \tag{A.9}$$

where $\delta(\cdot)$ is the Dirac distribution and $H(\cdot)$ is the unity-step function. An example of them for $n = 250$, $\mu = n/2$ and $\sigma^2 = n/4$ is depicted in Figure 7. This case corresponds to $\psi = 0$, and the probability $\tilde{f}_0$ is halved with respect to the general trend. Note that when $n \to \infty$, f_p converges in the *weak* sense [22] to the continuous uniform PDF, while F_p converges puctually to $F(x) = x$.

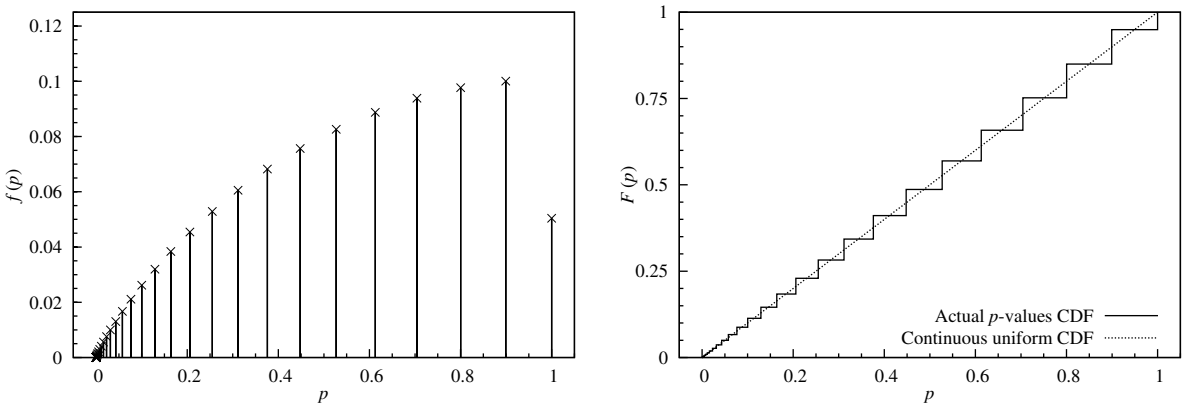
This CDF is a piece-wise constant right-continuous function, with a discontinuity in all p_i; in these points (A.9) can be simplified in

$$F_p(p_i) = \sum_{j \ge i} \tilde{f}_j = 1 - \sum_{j=0}^{i-1} \tilde{f}_j \tag{A.10}$$

with

$$\lim_{x \to p_i^+} F_p(x) = F_p(p_i), \quad \lim_{x \to p_i^-} F_p(x) = F_p(p_{i+1}) \tag{A.11}$$

We propose the function F_p' in 4 as a low computational complexity approximation for the CDF F_p for a twofold reason. First, we can show that the two limits of (A.11) hold also for the F_p' under the assumption of (A.1) and (A.5). Second, despite the fact that F_p' is not piecewise constant (and furthermore not montonically not decreasing), its oscillations



(a) (b)

Fig. 7. Example of p-values PDF (a) and CDF (b) in a generic binomial test, with $n = 250$, $\mu = n/2$ and $\sigma^2 = n/4$ ($\psi = 0$).

between two discontinuity points (i.e., where F_p is constant) are very small compared to the steps into the discontinuities, i.e., the relative error we commit in considering F'_p stepwise is extremely small.

In the following, for the sake of simplicity, we show that the first limit in (A.11) holds also for the F'_p in the case $\psi = 0$; the other cases are very similar to show. To this purpose, let us consider the identity

$$\begin{aligned}\sum_{j=a}^{b} f_j = \sum_{j=0}^{b} f_j - \sum_{j=0}^{a-1} f_j = F_T(b) - F_T(a-1) \simeq \\ \simeq F_T^{(lim)}\left(b+\frac{1}{2}\right) - F_T^{(lim)}\left(a-\frac{1}{2}\right)\end{aligned} \tag{A.12}$$

which allow us to recast, in the case $\psi = 0$

$$\begin{aligned}F_p(p_i) = 1-\sum_{j=0}^{i-1}\tilde{f}_j = 1-\sum_{j=\mu-i+1}^{\mu+i-1} f_j \simeq \\ \simeq 1 - F_T^{(lim)}\left(\mu+i-\frac{1}{2}\right) + F_T^{(lim)}\left(\mu-i+\frac{1}{2}\right)\end{aligned} \tag{A.13}$$

Note that when $\psi = 0$, then $\xi_i = i = \sqrt{2\sigma^2}\,\mathrm{erfc}^{-1}(p_i)$; by using the series expansion of $F_T^{(lim)}$ around $\mu \pm \xi_i$, we get

$$\begin{aligned}\lim_{x\to p_i^+} F_p(x) \simeq 1-F_T^{(lim)}(\mu+\xi_i)+\frac{1}{2}f_T(\mu+\xi_i)+ \\ +F_T^{(lim)}(\mu-\xi_i)+\frac{1}{2}f_T(\mu-\xi_i) = \\ = p_i+\frac{1}{2}\Big(f_T(\mu+\xi_i)+f_T(\mu-\xi_i)\Big) = p_i+d(p_i)\end{aligned} \tag{A.14}$$

where (A.1) and (A.4) have been used.

When considering F'_p, we have

$$\lim_{x\to p_i^+} F'_p(x) = p_i + 2\,d(p_i)\lim_{\xi\to i^-} z(\xi) = p_i + d(p_i) \tag{A.15}$$

□

To assert the quality of the proposed approximation, we show in Figure 8 the comparison between the actual CDF F_p and its approximation F'_p. Actually, to make the Figure more readable we prefer to show the comparison between $F_p(p)-p$ and $F'_p(p) - p$; in this way any difference among the two functions is more evident. We also considered different values of ψ; in all cases the matching is almost perfect, as the two curves can be hardly distinguished.

APPENDIX B
PROOF OF PROPOSITION 2

Proof of (a) - Let us consider a generator of independent basic events X_i, with $\Pr(X_i = +1) = u$. Given λ the fraction of events $+1$ in a n events sequence, we get $\mathrm{E}[\lambda] = u$; however, the actual value of λ could range from 0 to 1.

Let us enumerate with λ_j all possible values of λ, i.e., $\lambda_j = j/n$, $j = 0, 1, \ldots, n$, and let us indicate with $p_{v,j}$ the p-value

$$p_{v,j} = \mathrm{erfc}\left(\frac{|v - 2n\lambda_j(1-\lambda_j)|}{2\sqrt{2n}\lambda_j(1-\lambda_j)}\right) \tag{B.1}$$

Given λ_j, there are $n\lambda_j$ events $+1$ and $n(1-\lambda_j)$ events -1. In this case the number of runs v_n lies in the range $2 \le v_n \le \min(2n\lambda_j+1, 2n(1-\lambda_j)+1, n)$ except in the two degenerate cases $\lambda_0 = 0$ and $\lambda_n = 1$, where $v_n = 1$. Similarly to the case of the generic binomial test, different values of v_n can give rise to the same p-value only if $\psi = (2n\lambda_j(1-\lambda_j))_{(\mathrm{mod}\,1)} = (-2j^2/n)_{(\mathrm{mod}\,1)}$ has particular values. Since for almost all values of j it is $\psi \neq 0$ and $\psi \neq 1/2$, we can approximate the number of different p-values given λ_j with the number of different values of v_n. With this approximation the cardinality of $\mathcal{P}$ in a Runs Test is the different number of couples (j, v)

$$|\mathcal{P}| \simeq 2 + \sum_{j=1}^{n-1}\min\left(n\lambda_j, n(1-\lambda_j), n\right) = \begin{cases}\dfrac{n^2}{2}+1 & n \text{ even}\\ \dfrac{n^2}{2}+\dfrac{3}{2} & n \text{ odd}\end{cases} \tag{B.2}$$

i.e., $|\mathcal{P}| \simeq n^2/2$. □

Proof of (b) - Let us consider a generator of independent basic events X_i, with $\Pr(X_i = +1) = u$, with λ the fraction of events $+1$ in a n events sequence.

- the probability to generate a sequence with a given λ is $u^{n\lambda}(1-u)^{n(1-\lambda)}$;

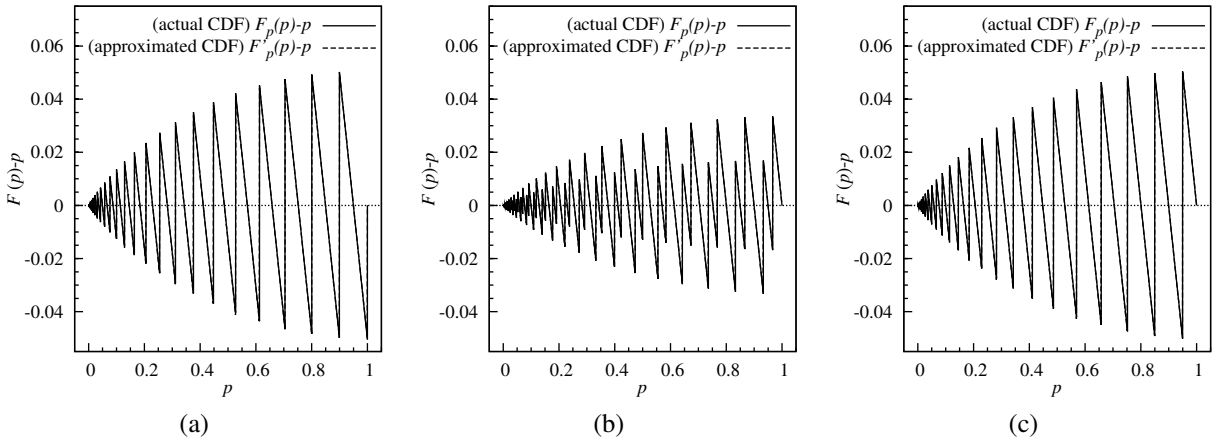


Fig. 8. Comparison between $F_p(p) - p$ (difference between the actual CDF and the continuous uniform one, solid lines) and $F'_p(p) - p$ (difference between the approximated CDF and the continuous uniform one, dashed lines) for $n = 250$, $\mu = n/2 + \psi$, $\sigma^2 = n/4$, in the cases: (a) $\psi = 0$; (b) $\psi = 1/3$; and (c) $\psi = 1/2$.

- the number of different sequences with v runs is

$$N(v,\lambda) = \begin{cases} 2\binom{n\lambda-1}{\frac{v}{2}-1}\binom{n-n\lambda-1}{\frac{v}{2}-1} & \text{if } v \text{ is even} \\ \binom{n\lambda-1}{\frac{v-1}{2}}\binom{n-n\lambda-1}{\frac{v-3}{2}}+ & \\ +\binom{n\lambda-1}{\frac{v-3}{2}}\binom{n-n\lambda-1}{\frac{v-1}{2}} & \text{if } v \text{ is odd} \end{cases} \tag{B.3}$$

Equation (B.3) can be explained by observing that the number of different ways in which we can distribute k objects in n groups with no empty groups is $\binom{n-1}{k-1}$. When v is even, we have to consider all sequences with $v/2$ continuous groups of symbols "$+1$" interleaved with $n/2$ continuous groups of "-1", considering both sequences starting with "$+1$" and ending with "-1", and sequences starting with "-1" and ending with "$+1$". When instead we consider r odd, we have to count all sequences starting and ending with "$+1$", with $(r+1)/2$ groups of "$+1$" and $(r-1)/2$ groups of "-1", as well as sequences starting and ending with "-1", with $(r-1)/2$ groups of "$+1$" and $(r+1)/2$ groups of "-1".

On the basis of the notation introduced, the actual CDF for the Runs Test can be written as

$$\begin{aligned} F_p(x) &= \sum_{j=0}^{n}\sum_{v=0}^{n} u^{n\lambda_j}(1-u)^{n-n\lambda_j} N(v,\lambda_j)\cdot \\ &\quad \cdot H\left(x - \operatorname{erfc}\left(\frac{|v-2n\lambda_j(1-\lambda_j)|}{2\sqrt{2n}\lambda_j(1-\lambda_j)}\right)\right) = \\ &= \sum_{j,v} u^{n\lambda_j}(1-u)^{n-n\lambda_j} N(v,\lambda_j), \\ &\qquad \forall j,v \,:\, x \geq \operatorname{erfc}\left(\frac{|v-2n\lambda_j(1-\lambda_j)|}{2\sqrt{2n}\lambda_j(1-\lambda_j)}\right) \end{aligned} \tag{B.4}$$

where we have considered $N(v,\lambda_j) = 0$ for all values of v outside the range $2 \leq v \leq \min(2n\lambda_j + 1, 2n(1-\lambda_j)+1, n)$. Note that this distribution is dependent on u.

An example of F_p for $u = 1/2$ and $n = 250$ is depicted in Figure 9. Note that the distance from the uniform distribution is much smaller with respect to the generic binomial test with the same n, as expected by the higher cardinality of $\mathcal{P}$.

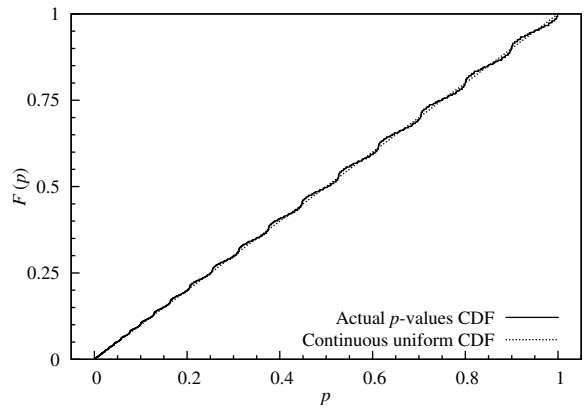


Fig. 9. Example of the p-values CDF for the Runs Test, with $n = 250$ and $u = 1/2$.

In order to get our result, let us recast (B.4) as

$$F_p(x) = \sum_{v=0}^{n} F_p^{(v)}(x;v) \tag{B.5}$$

$$\begin{aligned} F_p^{(v)}(x;v) &= \sum_j u^{n\lambda_j}(1-u)^{n-n\lambda_j} N(v,\lambda_j), \\ &\quad \forall j \,:\, x \geq \operatorname{erfc}\left(\frac{|v-2n\lambda_j(1-\lambda_j)|}{2\sqrt{2n}\lambda_j(1-\lambda_j)}\right) \end{aligned} \tag{B.6}$$

$F_p^{(v)}$ is a family of stepwise functions, where the height of the steps is inversely proportional to their distance, i.e., the higher the steps, the nearest they are. This suggest that $F_p^{(v)}$ can be approximated with a continuous function.

The approximation we propose is the following one

$$\begin{aligned} F_p'^{(v)}(x;v) &= \int_\Lambda e^{-C_0 - C_1(\lambda-u) - \frac{1}{2}C_2(\lambda-u)^2}\, d\lambda, \\ \Lambda &= \left\{\lambda \,:\, x \geq \operatorname{erfc}\left(\frac{|v-2n\lambda_j(1-\lambda_j)|}{2\sqrt{2n}\lambda_j(1-\lambda_j)}\right)\right\} \end{aligned} \tag{B.7}$$

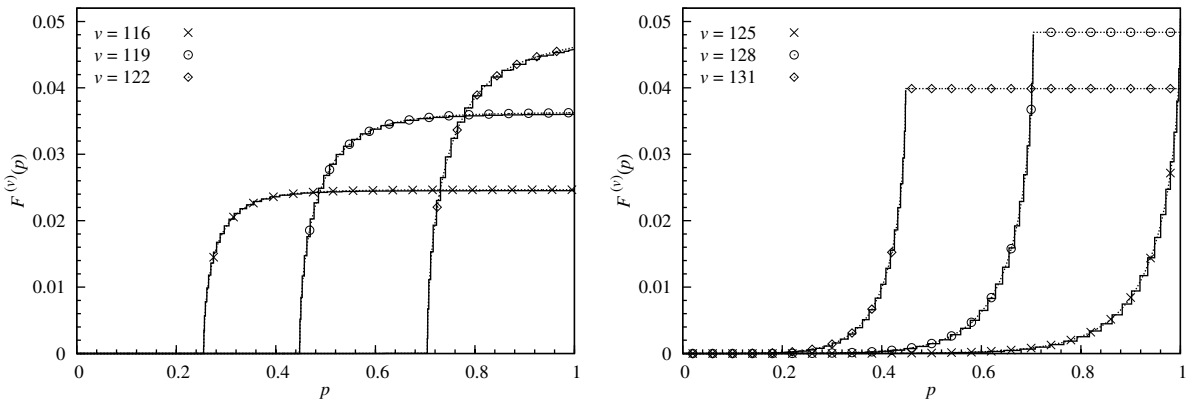


Fig. 10. Comparison between the actual $F_p^{(v)}$ (piece-wise constant, dotted line) and its approximation $F_p'^{(v)}$ (continuous function, solid line) for the Runs Test, with $n = 250$, $u = 1/2$ and different values of v.

where the explicit definition of Λ can be expressed as

$$\Lambda = \begin{cases} [0, \lambda^{(1)}] \cup [\lambda^{(4)}, 1] & \text{if } x \leq \operatorname{erfc} \dfrac{|n-2v|}{\sqrt{2n}} \\ [0, \lambda^{(1)}] \cup [\lambda^{(2)}, \lambda^{(3)}] \cup [\lambda^{(4)}, 1] & \text{if } x > \operatorname{erfc} \dfrac{|n-2v|}{\sqrt{2n}},\ v < \dfrac{n}{2} \\ [0,1] & \text{if } x > \operatorname{erfc} \dfrac{|n-2v|}{\sqrt{2n}},\ v > \dfrac{n}{2} \end{cases} \tag{B.8}$$

with $\lambda^{(i)}$ given by the four combinations of the following expression, with $\lambda^{(i)} < \lambda^{(i+1)}$

$$\lambda^{(i)} = \frac{1}{2} \pm \frac{1}{2}\sqrt{1 - \frac{2v}{n \pm \sqrt{2n}\operatorname{erfc}^{-1}(x)}} \tag{B.9}$$

and where C_0, C_1 and C_2 come from the series expansion of $u^{n\lambda}(1-u)^{n-n\lambda}N(v,\lambda)$ where the binomial coefficients in $N(v,\lambda)$ can be substituted with a standard Gaussian approximation.

Despite the general case is still difficult to handle with, limiting ourselves to $u = 1/2$, we get

$$C_0 = \begin{cases} \dfrac{(n-2v+2)^2}{2(n-2)} + \log \dfrac{\pi(n-2)}{2} & n \text{ even} \\ \dfrac{(n-2v+2)^2+4}{2(n-2)} + \log \dfrac{\pi(n-2)}{2} & n \text{ odd} \end{cases}$$

$$C_1 = 0$$

$$C_2 = \begin{cases} \dfrac{4n^2(4(v-2)^2 - n + 2)}{(n-2)^3} & n \text{ even} \\ \dfrac{4n^2(4(n-6)(v-2)^2 - (n-4)^2 + 4)}{(n-2)^4} & n \text{ odd} \end{cases} \tag{B.10}$$

Thanks to $C_1 = 0$, the integral in (B.7) can be solved into a closed form. By setting

$$F_1'^{(v)}(x;v) = n\int_0^{\lambda^{(1)}} e^{-C_0 - \frac{1}{2}C_2(\lambda-\frac{1}{2})^2} d\lambda + n\int_{\lambda^{(4)}}^{1} e^{-C_0 - \frac{1}{2}C_2(\lambda-\frac{1}{2})^2} d\lambda = \\ = \frac{n\sqrt{2\pi}e^{-C_0}}{\sqrt{C_2}}\left(\operatorname{erfc}\left(\frac{\sqrt{C_2}}{2\sqrt{2}}\sqrt{1 - \frac{2v}{n+\sqrt{2n}\operatorname{erfc}^{-1}(x)}}\right) - \operatorname{erfc}\frac{\sqrt{C_2}}{2\sqrt{2}}\right) \tag{B.11}$$

$$F_2'^{(v)}(x;v) = n\int_{\lambda^{(2)}}^{\lambda^{(3)}} e^{-C_0 - \frac{1}{2}C_2(\lambda-\frac{1}{2})^2} d\lambda = \\ = \frac{n\sqrt{2\pi}e^{-C_0}}{\sqrt{C_2}}\left(1 - \operatorname{erfc}\left(\frac{\sqrt{C_2}}{2\sqrt{2}}\sqrt{1 - \frac{2v}{n-\sqrt{2n}\operatorname{erfc}^{-1}(x)}}\right)\right) \tag{B.12}$$

$$F_3'^{(v)}(v) = n\int_0^1 e^{-C_0 - \frac{1}{2}C_2(\lambda-\frac{1}{2})^2} d\lambda = \frac{n\sqrt{2\pi}e^{-C_0}}{\sqrt{C_2}}\left(1 - \operatorname{erfc}\frac{\sqrt{C_2}}{2\sqrt{2}}\right) \tag{B.13}$$

as well as by selecting the correct expression for C_0 and C_2 depending if v is even or odd, we come into

$$F_p'^{(v)}(x;v) = \begin{cases} F_1'^{(v)}(x;v) & \text{if } p \leq \operatorname{erfc} \dfrac{|n-2v|}{\sqrt{2n}} \\ F_1'^{(v)}(x;v) + \\ \quad + F_2'^{(v)}(x;v) & \text{if } p > \operatorname{erfc} \dfrac{|n-2v|}{\sqrt{2n}},\ v < \dfrac{n}{2} \\ F_3'^{(v)}(v) & \text{if } p > \operatorname{erfc} \dfrac{|n-2v|}{\sqrt{2n}},\ v > \dfrac{n}{2} \end{cases} \tag{B.14}$$

A comparison between the actual $F_p^{(v)}$ and its proposed approximation $F_p'^{(v)}$ for $n = 250$ and different values of v can be seen in figure 10. In all cases the matching is very good.

Using (B.14), the p-value CDF F_p in the Runs Test can be approximated with

$$F_p'(x) = \sum_{v=0}^{n} F_p'^{(v)}(x;v) \tag{B.15}$$

Note that (B.15) presents a computational complexity which can be quantified in $O(n)$, instead of $O(n^2)$ as (B.4). With a small computational effort, F_p' can be computed even for quite large values of n. □

REFERENCES

[1] P. L'Ecuyer and R. Simard, "TestU01: A C Library for Empirical Testing of Random Number Generators," *AMC Transaction on Mathematical Software*, vol. 33, no. 4, article 22, Aug. 2007.

[2] *A statistical test suite for random and pseudorandom number generators for cryptographic applications*, National Institute of Standards and Technology (NIST) Special Publication 800-22, Rev. 1a, Apr. 2010.

[3] G. Marsaglia, "The Marsaglia Random Number CD-ROM including the DieHard Battery of test of randomness," 1995. [Online]. Available: http://stat.fsu.edu/pub/diehard/

[4] A. J. Menezes, P. C. van Oorschot, and S. A. Vanstone, *Handbook of Applied Cryptography*. CRC Press, 1996.

[5] L. Blum, M. Blum, and M. Shub, "A Simple Unpredictable Pseudo-Random Number Generator," *SIAM Journal on Computing*, vol. 15, pp. 364–383, May 1986.
[6] P. L'Ecuyer and R. Proulx, "About polynomial-time "unpredictable" generators," in *Proc. of the IEEE 1989 Winter Simulation Conference*, Dec. 1989, pp. 467–476.
[7] W. Schindler, *AIS 20: Functionality Classes and Evaluation Methodology for Deterministic Random Number Generators*, Bundesamt fur Sicherheit in der Informationstechnik (BSI) version 2.0, Dec. 1999.
[8] W. Killmann and W. Schindler, *AIS 31: Functionality Classes and Evaluation Methodology for True (Physical) Random Number Generators*, Bundesamt fur Sicherheit in der Informationstechnik (BSI) version 3.1, Sep. 2001.
[9] J. S. Coron, "On the Security of Random Sources," in *Public Key Cryptography. Second International Workshop on Practice and Theory in Public Key Cryptography*, Lecture Notes in Computer Science 1560. Springer, 1999, pp. 29–42.
[10] D. E. Knuth, *The Art of Computer Programming.* Volume 2: *Seminumerical Algorithms*, 3rd ed. Addison-Wesley Professional, 1997.
[11] A. M. Law and W. D. Kelton, *Simulation Modeling and Analysis*, 3rd ed. New York: McGraw-Hill, 2000.
[12] F. Pareschi, R. Rovatti, and G. Setti, "Second-level NIST Randomness Tests for Improving Test Reliability," in *Proc. IEEE International Symposium on Circuits and Systems (ISCAS2007)*, May 2007, pp. 1437–1440.
[13] H. Sackrowitz and E. Samuel-Cahn, "P values as random variables—expected P values," *The American Statistician*, vol. 53, no. 4, pp. 326–331, Nov. 1999.
[14] S. J. Kim, K. Umeno, and A. Hasegawa, "Corrections of the NIST Statistical Test Suite for Randomness," Cryptology ePrint Archive, Tech. Rep. 2004/018, 2004.
[15] K. Hamano, "The Distribution of the Spectrum for the Discrete Fourier Transform Test included in SP800-22," *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, vol. E88, no. 1, pp. 67–73, Jan. 2005.
[16] J. D. Gibbons and J. W. Pratt, "P-Values: Interpretation and Methodology," *The American Statistician*, vol. 29, no. 1, pp. 20–25, Feb. 1975.
[17] M. J. Schervish, "P Values: What They Are and What They Are Not," *The American Statistician*, vol. 50, no. 3, pp. 203–206, Aug. 1996.
[18] F. Pareschi, G. Setti, and R. Rovatti, "Implementation and Testing of High-Speed CMOS True Random Number Generators Based on Chaotic Systems," *IEEE Transactions on Circuits and Systems—Part I: Regular Papers*, vol. 57, no. 12, pp. 3124–3137, Dec. 2010.
[19] "Random Numbers Generation using Quantum Physics," White Paper, idQuantique, 2004. [Online]. Available: http://www.idquantique.com/products/files/quantis-whitepaper.pdf
[20] P. L'Ecuyer, R. Simard, and S. Wegenkittl, "Sparse serial tests of uniformity for random number generators," *SIAM Journal on Scientific Computing*, vol. 24, no. 2, pp. 652–668, 2002.
[21] P. L'Ecuyer, J. F. Cordeau, and R. Simard, "Close-point spatial tests and their application to random number generators," *Operations Research*, vol. 48, no. 2, pp. 308–317, 2000.
[22] A. N. Shiryaev, *Probability*, 2nd ed. Springer-Verlag New York, 1995.
[23] M. Matsumoto, "Mersenne Twister: A 623-Dimensionally Equidistributed Uniform Pseudo-Random Number Generator," *ACM Transactions on Modeling and Computer Simulation*, vol. 8, no. 1, pp. 3–30, Jan. 1998.
[24] J. D. Gibbons and S. Chakraborti, *Nonparametric Statistical Inference*, 3rd ed. Marcel Dekker Ed., 1992.
[25] H. Cramer, *Mathematical Methods of Statistics*. Princeton University Press, Sep. 1946.
[26] W. Feller, "On the Kolmogorov-Smirnov limit theorems for empirical distributions," *Annals of Mathematical Statistics*, vol. 21, no. 2, pp. 301–302, 1948.
[27] R. Simard and P. L'Ecuyer, "Computing the two-sided Kolmogorov-Smirnov distribution," *Journal of Statistical Software*, vol. 39, no. 11, Mar. 2011, Available: http://www.jstatsoft.org/v39/i11.

Fabio Pareschi (S'05-M'08) received the Dr.Eng. degree (with honors) in electronic engineering from University of Ferrara, Italy, in 2001, and the Ph.D. in Information Technology under the European Doctorate Project (EDITH) from University of Bologna, Italy, in 2007.

He is currently with Department of Engineering (ENDIF), University of Ferrara, and also with Advanced Research Center on Electronic Systems, University of Bologna. In 2006, he spent six months as a Visiting Scholar at the Department of Electrical Engineering of the Catholic University of Leuven, Belgium. He is currently Associated Editor for the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMSPART II.

His research activity is mainly focused on Analog and Mixed Mode electronic circuit design (in particular on non-linear and chaotic circuit implementation), statistical signal processing, random number generation and testing, and electromagnetic compatibility. He was co-recipient of the best paper award at ECCTD2005 and the best student paper award at EMCZurich2005.

Riccardo Rovatti (M'99-SM'02-F'12) received the M.S. degree (summa cum laude) in electronic engineering and the Ph.D. degree in electronics, computer science, and telecommunications from the University of Bologna, Italy, in 1992 and 1996, respectively.

Since 2001 he is Associate Professor of Electronics with the University of Bologna. He is the author of more than 230 technical contributions to international conferences and journals and of two volumes. He is co-editor of the book Chaotic Electronics in Telecommunications (CRC, Boca Raton) as well as one of the guest editors of the May 2002 special issue of the PROCEEDINGS OF THE IEEE on Applications of Non-linear Dynamics to Electronic and Information Engineering. His research focuses on mathematical and applicative aspects of statistical signal processing especially those concerned with nonlinear dynamical systems. Prof. Rovatti was an Associated Editor of the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS–PART I. In 2004 he received the Darlington Award of the Circuits and Systems Society. He was the Technical Program Cochair of NDES 2000 (Catania) and the Special Sessions Cochair of NOLTA 2006 (Bologna).

Gianluca Setti (S'89-M'91-SM'02-F'06) received the Dr.Eng. degree (with honors) in electronic engineering and the Ph.D. degree in electronic engineering and computer science from the University of Bologna, Italy, in 1992 and 1997, respectively, for his contribution to the study of neural networks and chaotic systems. From May 1994 to July 1995, he was with the Laboratory of Nonlinear Systems (LANOS) of the Swiss Federal Institute of Technology in Lausanne (EPFL) as Visiting Researcher. Since 1997 he has been with the School of Engineering at the University of Ferrara, Italy, where he is currently a Professor of Circuit Theory and Analog Electronics. He held several visiting position at Visiting Professor/Scientist at EPFL (2002, 2005), UCSD (2004), IBM T. J. Watson Laboratories (2004, 2007) and at the University of Washington in Seattle (2008), and is also a permanent faculty member of ARCES, University of Bologna. He is co-editor of the book Chaotic Electronics in Telecommunications (CRC Press, Boca Raton, 2000) and Circuits and Systems for Future Generation of Wireless Communications (Springer, 2009) as well as one of the guest editors of the May 2002 special issue of the PROCEEDINGS OF THE IEEE on Applications of Non-linear Dynamics to Electronic and Information Engineering. His research interests include nonlinear circuits, recurrent neural networks, implementation and application of chaotic circuits and systems, statistical signal processing, electromagnetic compatibility, wireless communications, and sensor networks.

Dr. Setti received the 1998 Caianiello prize for the best Italian Ph.D. thesis on Neural Networks and he is corecipient of the 2004 IEEE CAS Society Darlington Award, as well as of the best paper award at ECCTD 2005 and the best student paper award at EMCZurich 2005 and ISCAS 2011. He served as an Associate Editor for the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS–PART I (1999-2002 and 2002-2004) and for the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS–PART II (2004-2007), as the Deputy-Editor-in-Chief, for the IEEE Circuits and Systems Magazine (2004-2007) and as the Editor-in-Chief for the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS–PART II (2006-2007) and for the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS–PART I (2008-2009).

He was the 2004 Chair of the Technical Committee on Nonlinear Circuits and Systems of the of the IEEE CAS Society, a Distinguished Lecturer (2004-2005), a member of the Board of Governors (2005-2008), and President (2010) of the same society. He was also the Technical Program Co-Chair of NDES2000 (Catania) the Track Chair for Nonlinear Circuits and Systems of ISCAS 2004 (Vancouver), the Special Sessions Co-Chair of ISCAS 2005 (Kobe) and ISCAS 2006 (Kos), the Technical Program Cochair of ISCAS2007 (New Orleans), and ISCAS 2008 (Seattle), as well as the General Co-Chair of NOLTA 2006 (Bologna).