On the Approximation Errors in the Frequency Test Included in the NIST SP800-22 Statistical Test Suite

This is the author's version of the article that has been presented at IEEE APCCAS2008
The editorial version of the paper is available at http://dx.doi.org/10.1109/APCCAS.2008.4746245

# On the Approximation Errors in the Frequency Test Included in the NIST SP800-22 Statistical Test Suite

Fabio Pareschi*‡, Riccardo Rovatti†‡, and Gianluca Setti*‡
*ENDIF - University of Ferrara, via Saragat 1, 44100 Ferrara - ITALY
†DEIS - University of Bologna, viale risorgimento 2, 40136 Bologna - ITALY
‡ARCES - University of Bologna, via Toffano 2/2, 40125 Bologna - ITALY
Email: {fabio.pareschi,gianluca.setti}@unife.it, rrovatti@arces.unibo.it

***Abstract*— In previous papers we have addressed the problem of testing Random Number Generators (RNGs) through statistical tests, with particular emphasis on the approach we called *second-level* testing. We have shown that this approach is capable of achieving much higher accuracy in exposing non-random generators, but may suffer from reliability issues due to approximations introduced in the test. Here we consider the NIST Frequency Test and present a mathematical expression of the error introduced by approximating the effective discrete distribution function with its continuous limit distribution. The matching against experimental data is almost perfect.**

## I. Introduction

Random Number Generators (RNGs) represent a fundamental component in many applications; for example they are critical for the security in some cryptographic primitives [1]. Several architectures of RNGs have been proposed in recent years, ranging from jitter measurements [2] to quantum effects observation [3], including generators based on chaotic dynamic [4]. In order to choose among this pletora of possible solutions, it is fundamental to assert the RNG quality.

For this reason the interest on tests for randomness has grown significantly. In this paper we focus on the class of tests known as *statistical tests for randomness* [5], [6]. Even if different methods have recently been proposed based on the direct estimation of the entropy a sequence of events [7], all these methods require some assumptions on the input sequence. On the contrary, statistical tests can effectively work as blackbox tests.

The main problem on statistical test is the interpretation of the results. Roughly speaking, while a failed test is a serious indicator for the weakness of a RNG, a passed test does not provide a direct positive proof for the quality of a RNG. Mathematically, a test can be schematically described as a function looking at a sequence of $n$ events (e.g. a sequence of $n$ bits) and giving as output a number in $[0,1]$, called a *p-value*. Intuitive speaking, the p-value is the quantification of how much the sequence under test "appears more random" than an effectively random generated sequence.

When we assume the input sequence is composed by random variables, also the p-value is a random variable depending on them. When the input sequence is *true random*, i.e. all events are independent and drawn according to the same known distribution, the p-value is uniform distributed in the interval $[0,1]$, and its cumulative distribution function (cdf) is $F^u(x) = x$. When we model the sequence generation as a process producing non-independent events (i.e. pattern or regularities are introduced in the sequence) or not distributed according to the expected distribution, the p-value is distributed according to a cdf $F^{nu}(x)$ (that depends on the test and on the generator model) where p-values around zero are much more probable than others.

Given a sequence and its p-value, the interpretation of the test is the following. Fixed a *level of significance* $\alpha$, we consider a sequence true random if its p-value $p > \alpha$. Immediately, one can found that this approach is not *exact* but it is possible to commit errors, hence the name "statistical":

- given a true-random sequence, the probability to fail the test is $\text{Prob}\{p \leq \alpha\} = F^u(\alpha) = \alpha$ *(Type I error)*
- given a dependent sequence, the probability to pass the test is $\text{Prob}\{p > \alpha\} = 1 - F^{nu}(\alpha) = \beta$ *(Type II error)*.

Usually this is not a problem, since both $\alpha$ and $\beta$ are typically small. The National Institute of Standard and Technology (NIST) in its suite SP800-22 [5] suggests $\alpha = 0.01$, while the value of $\beta$ depends on the test and on the statistics of the generator under test, and, of course, on $\alpha$. Its computation is not trivial; usually the better the test, the lower $\beta$; the more similar the generator to a true random process, the lower $\beta$. Note that this is the classical approach named *statistical hypothesis testing* [8].

We showed in [9] that using a *second-level* approach it is possible to get more accurate results, i.e. fixed the same Type I error probability, it is more difficult for a non-true random sequence to pass a test. This approach is already discussed by NIST in its document [5, chap. 4], and consists in repeating the statistical test over many different sequences, and checking if the distribution of the obtained p-values matches an uniform distribution.

With this aim, NIST suggests a chi-square goodness-of-fit test. Note that this test is again a statistical test and gives another (a second level) p-value; for this reason we call this approach *second level testing*.

We showed that regrettably this approach may produce unreliable results. In fact, in every statistical test some approximations are adopted, introducing errors in the p-value computation and so in the p-value distribution. As a result, matching a distribution of p-values coming from true random sequences against an uniform distribution may result in failing the test.

Following [9], in this paper we focus on the Frequency Test included in the NIST suite. This tests uses a binomial distribution approximated with a normal distribution. In [9] we analyzed this test, finding a limit on the sensitivity of a chi-square test used as a second-level test to ensure that the normal approximation does not generate unreliable results. Here we consider this approximation under a different point of view,

This is the author's version of the article that has been presented at IEEE APCCAS2008
The editorial version of the paper is available at http://dx.doi.org/10.1109/APCCAS.2008.4746245

i.e. that we are approximating a discrete distribution with a continuous one. With this approach we will be able not only to say if the second-level test is reliable or not, but also to compute the error in the second-level p-value.

The paper is organized as follows. Section II will provide a mathematical background on the Frequency Test and on the chi-square goodness-of-fit test. In section III we will find an expression for the effective distribution of the p-values in a Frequency Test, and how deviations from uniform reflect in error in the second-level test. Finally, section IV provides some examples of how to ensure a second-level reliable test.

## II. Mathematical Background

In the following we provide a very short mathematical background on a statistical test considering two cases, the NIST Frequency Test and the Pearson's chi-square test.

Given a sequence of $n$ events $X^{(1)}, \ldots, X^{(n)}$, a statistical test can be defined as a test function $T = T\left(X^{(1)}, \ldots, X^{(n)}\right)$. If we assume a true random input sequence, $T$ is a random variable whose mean value $T_0$ and probability distribution can be computed from the $X^{(i)}$ statistics. The p-value of an observed sequence $X^{(1)}_{obs}, \ldots, X^{(n)}_{obs}$ is by definition the probability that a random sequence has a $T$ more distant from $T_0$ with respect to $T_{obs} = T\left(X^{(1)}_{obs}, \ldots, X^{(n)}_{obs}\right)$. So, given a distance $\|T - T_0\| \to R^+$ and the cumulative probability distribution function $F_{\|\cdot\|}(x) = \text{Prob}\{\|T - T_0\| < \|x - T_0\|\}$, the p-value is expressed as $p = 1 - F_{\|\cdot\|}(T_{obs})$. In this way:

- $p = 1$, if $T = T_0$;
- $p \to 0$, if $\|T - T_0\| \to \infty$;
- for a true random sequence, $p$ is a random variable uniformly distributed in $[0, 1]$.

### A. NIST Frequency Test

The sequence $X^{(i)}$ is a sequence of bits, with $X^{(i)} = \{-1, +1\}$. The test function is the sum of all $X^{(i)}$:

$$S = \sum_{i=1,\ldots,n} X^{(i)}$$

$S$ is distributed according to a binomial distribution, with mean value $S_0 = 0$. The distance from $S_0$ can be simply computed as $|S|$. Since for large $n$, the binomial distribution can be approximated with a normal distribution, the distribution of $|S|$ can be approximated with a half-normal distribution, so

$$p = \text{erfc}\left(\frac{|S_{obs}|}{\sqrt{2n}}\right)$$

where $\text{erfc}(x)$ is the complementary error function.

### B. chi-square goodness-of-fit test

The test function is the distribution of the $n$ samples $X^{(i)}$ in $k$ subgroups, called *bins*. If the $X^{(i)}$ have a continuous distribution, the bins are obtained as a partition of the definition set of the $X^{(i)}$; let also $\pi_j$ 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:

$$\chi^2 = \sum_{j=1,\ldots,k} \frac{(E_j - O_j)^2}{E_j}$$

For a random input sequence, this is a random variable distributed according to a chi-square distribution with $k - 1$ degree of freedom, so [8]

$$p = 1 - \frac{\gamma\left((k-1)/2; \chi^2_{obs}/2\right)}{\Gamma\left((k-1)/2\right)}$$

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

## III. Error on a second-level NIST Frequency Test

In the NIST Frequency test, $S$ can only assume values in a subset of all integer numbers between $-n$ and $n$, more precisely *even* numbers if $n$ is even, and *odd* numbers it $n$ is odd. The probability that $S$ is equal to $r$ is given by

$$f_r = f_{-r} = 2^{-n}\binom{n}{\frac{r+n}{2}} \simeq \sqrt{\frac{2}{\pi n}} e^{-\frac{r^2}{2n}} \tag{1}$$

This is a *standard Gaussian approximation*, since (1) is the probability density function (pdf) of a normal random variable.

Note that we are approximating a discrete distribution with a continuous distribution; let us assume that in all the points where the discrete pdf is defined (i.e. in all possible value of $S$) the introduced error is negligible. In other word, let us try to use the approximated expression instead of the binomial coefficients, and look for deviations from the uniform in the p-values distribution due only to the discrete pdf.

Let us also suppose for simplicity that $n$ is even; so $|S|$ can only assume the $n/2+1$ even values between 0 and $n$. This means that also the p-value can assume only $n/2+1$ values; since we know the probability of every $S$, we can compute the pdf of the generated p-values as the discrete distribution

$$f^{(erfc)}(x) = \sum_{r=2,4,\ldots,n} 2\sqrt{\frac{2}{\pi n}} e^{-\frac{r^2}{2n}} \delta\left(x - \text{erfc}\left(\frac{r}{\sqrt{2n}}\right)\right) + \sqrt{\frac{2}{\pi n}}\delta(x-1) \tag{2}$$

where $\delta(x)$ is the classical Dirac delta function.

Note that for the case $p = 1$ the general rule does not apply. In fact the p-value $p = 1$ is generated only by $S = 0$, while for any other p-value the probability is doubled since it can be generated both by $S$ and by $-S$.

The integral of (2) gives the cdf, that is

$$F^{(erfc)}(x) = \sum_{r=2,4,\ldots,n} 2\sqrt{\frac{2}{\pi n}} e^{-\frac{r^2}{2n}} \text{u}\left(x - \text{erfc}\left(\frac{r}{\sqrt{2n}}\right)\right) + \sqrt{\frac{2}{\pi n}}\text{u}(x-1) \tag{3}$$

where $\text{u}(x)$ is the unit step function and again the case $p = 1$ has to be considered separately.

The cdf (3) is a step-wise function approximating the continuous uniform cdf; an example for $n = 100$ is shown in Figure 1. Note that this distribution effectively converges to the continuous uniform distribution, i.e. (2) converges *weakly* to the uniform pdf, and (3) converges *punctually* to the uniform cdf [10].

Note that the presence of the sum in (3) may lead to computational problems; for this reason we have found an approximation

$$F^{(erfc)}(x) \simeq x + d(x)\, z(x) = x + \varepsilon(x) \tag{4}$$

with

$$d(x) = \sqrt{\frac{2}{\pi n}} e^{-\left(\text{erfc}^{-1}(x)\right)^2}$$

$$z(x) = \sqrt{2n}\ \text{erfc}^{-1}(x) \pmod 2 - 1$$

This is the author's version of the article that has been presented at IEEE APCCAS2008
The editorial version of the paper is available at http://dx.doi.org/10.1109/APCCAS.2008.4746245

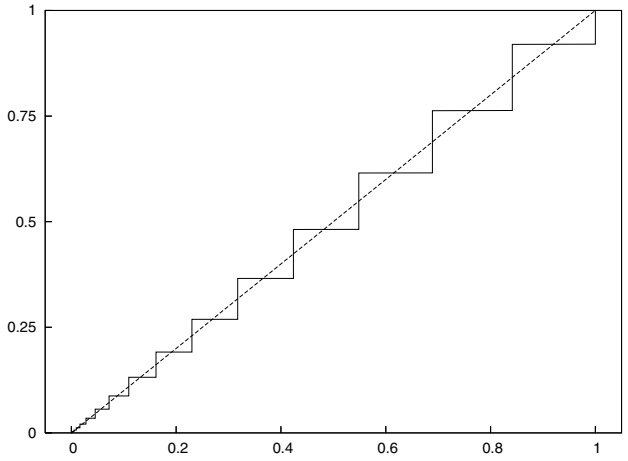


Fig. 1. Comparison between discrete cumulative distribution function for p-values generated by Frequency Test for $n = 100$ and continuous uniform cumulative distribution function.

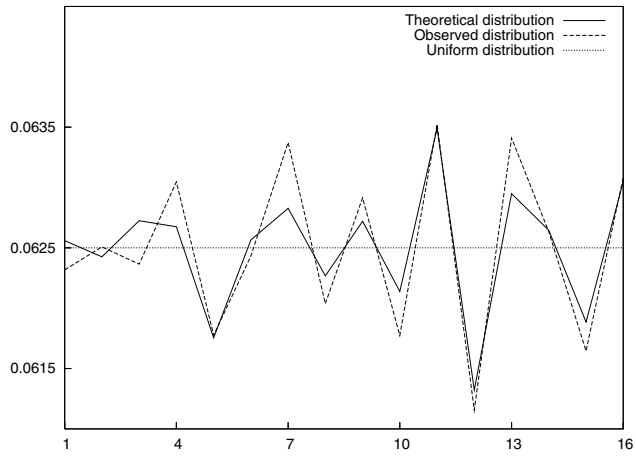


Fig. 2. Comparison between the theoretical distribution (6), the observed distribution and the uniform distribution (5) for $n = 2^{20}$ and $k = 16$.

The computation of (4) is not reported here; intuitively $d(x)$ expresses the height of the steps, while $z(x)$ the shape. Supposing (1) holds, the introduced error on both edges of all steps is zero; the only approximation error (that can be shown to be very small) is given by the fact that (4) is actually not stepwise constant.

Having an expression for $F^{(erfc)}(x)$, we can look at the deviation from uniform in the p-value distribution. Let us assume to have $N$ p-values, and perform a chi-square test dividing the interval $[0,1]$ in $k$ subintervals $\left[\frac{j-1}{k}, \frac{j}{k}\right]$. If we look for "classical" uniformity, i.e. if we assume $F(x) = x$, we have

$$\pi'_j = F\left(\frac{j}{k}\right) - F\left(\frac{j-1}{k}\right) = \frac{1}{k} \tag{5}$$

however since we have an analytic form for $F^{(erfc)}(x)$, we can compute

$$\pi''_j = \frac{1}{k} + \varepsilon\left(\frac{j}{k}\right) - \varepsilon\left(\frac{j-1}{k}\right) \tag{6}$$

Distribution (6) has (5) as limit distribution when $n$ grows to infinity. For a finite value of $n$ a small difference exists: Figure 2 shows the comparison in the case $n = 2^{20}$ between (5), (6) and the experimental distribution we have found averaging results from some different RNGs including [2], [3], [4] and the BBS generator [11]. The empirical distribution matches (6).

Even if the difference between distributions (5) and (6) is relatively small (below 2% in the above example), when increasing $N$ the chi-square test may become sensitive enough to distinguish between the two distributions. In this case, testing observed frequencies of the p-values against (5) or (6) may lead to completely different results.

More precisely, we know that the expected distribution is well approximated by (6); however, for sake of simplicity, we want to test the observed distribution against (5); this results in an error in the computation of the distance $\chi^2$. We anticipate here that we are able to evaluate this error as an average error. In fact suppose to know that testing an observed distribution $O_j$ against (6) gives a distance $\chi^2_0$. Testing the same $O_j$ against (5) gives a distance $\chi^2$ that is generally different from $\chi^2_0$:

$$\chi^2 = \sum_{j=1,\ldots,k} \frac{\left(\frac{N}{k} - O_j\right)^2}{\frac{N}{k}} = \sum_{j=1,\ldots,k} \frac{\left(E_j - N\varepsilon\left(\frac{j}{k}\right) + N\varepsilon\left(\frac{j-1}{k}\right) - O_j\right)^2}{\frac{N}{k}} =$$

$$= \sum_{j=1,\ldots,k} \frac{(E_j - O_j)^2}{\frac{N}{k}} + Nk \sum_{j=1,\ldots,k} \left(\varepsilon\left(\frac{j}{k}\right) - \varepsilon\left(\frac{j-1}{k}\right)\right)^2$$

$$+ 2k \sum_{j=1,\ldots,k} (E_j - O_j)\left(\varepsilon\left(\frac{j}{k}\right) - \varepsilon\left(\frac{j-1}{k}\right)\right) \tag{7}$$

where we indicated with $E_j$ the expected values of the $O_j$ assuming a distribution (6), while $N/k$ is the expected value of the $O_j$ assuming distribution (5).

The first term of (7) can be approximated with a small error with $\chi^2_0$; the second term is a constant; the third term is a random variable, depending on the $O_j$

We can consider the average error on all $O_j$ sequences giving $\chi^2_0$. Note that in this case the $O_j$ are random variables that are not independent (there are two constraints, the first given by $\chi^2_0$ and the second by $\sum O_j = N$); however every $O_j$ has $E_j$ as expected value. This means that the average contribute of this third term vanishes:

$$\chi^2 \simeq \chi^2_0 + Nk \sum_{j=1,\ldots,k} \left(\varepsilon\left(\frac{j}{k}\right) - \varepsilon\left(\frac{j-1}{k}\right)\right)^2 = \chi^2_0 + NC_{\chi^2}$$

where $C_{\chi^2}$ is implicitly defined.

Figure 3 shows a comparison between the expected value of the normalized value $\chi^2/(k-1)$ theoretically computed and experimental results for different values of $n$ and $N$. There is an almost perfect matching between the curves. Note that we used this normalization since $\chi^2_0$ is a chi-square distributed random variable whose expected value is $k - 1$. The trend of figure has been observed for all RNGs mentioned above. Figure 3 refers to the BBS, since this was the only one able to to generate the amount of bits necessary to compute the $\approx 10^6$ p-values used to plot the figure.

From the knowledge of the error on the distance $\chi^2$ we can compute the error on the p-value $p$ of the chi-square test. If

This is the author's version of the article that has been presented at IEEE APCCAS2008
The editorial version of the paper is available at http://dx.doi.org/10.1109/APCCAS.2008.4746245

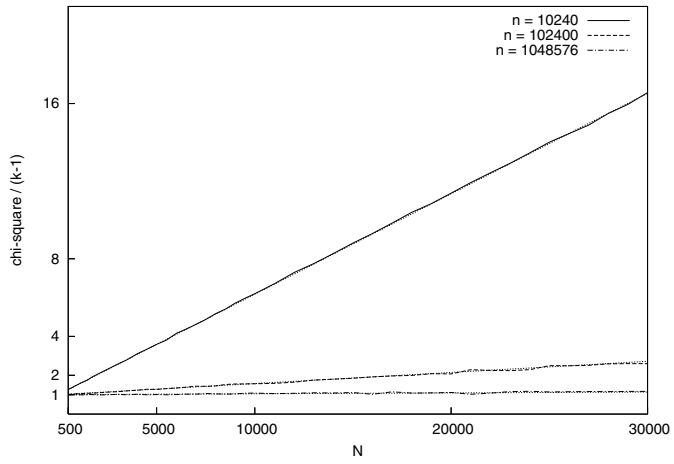


Fig. 3. Comparison between the observed value of $\chi^2/(k-1)$ for different values of $N$, with $k = 16$ and $n = 10 \cdot 2^{10}$ (solid line), $n = 100 \cdot 2^{10}$ (dashed line), $n = 2^{20}$ (dotted-dashed line), along with their theoretically expected values (dotted lines).

$p_0 = 1 - F_{\chi^2}\left(\chi_0^2\right)$ is the p-value we get using (6), and we suppose that the error is small, we can write:

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

i.e. the average error on the chi-square test p-value is

$$p - p_0 \simeq -f_{\chi^2}\left(F_{\chi^2}^{-1}\left(1-p_0\right)\right) NC_{\chi^2} \tag{8}$$

where $F_{\chi^2}(x)$ and $f_{\chi^2}(x)$ are respectively the cdf and the pdf of a chi-square distribution with $k-1$ degrees of freedom.

We can notice that the error $p - p_0$ is always negative ($f_{\chi^2}(x)$ is a pdf), and depends linearly on N. Furthermore, for $k \geq 4$, $f_{\chi^2}\left(F_{\chi^2}^{-1}(x)\right)$ is a concave function, with a maximum in $x = \gamma\left((k-1)/2, (k-3)/2\right)$.

## IV. Examples

Equation (8) can be easily used to verify or ensure the reliability of a chi-square test. For example, we can consider a Frequency test with $n = 2^{20}$ and suppose to require an error on the chi-square p-value $|p - p_0| < 0.01$ on the whole range $0 \leq p_0 \leq 1$ using $k = 10$ bins. In this case

$$C_{\chi^2} \sup_{0 \leq p_0 \leq 1} f_{\chi^2}\left(F_{\chi^2}^{-1}\left(1-p_0\right)\right) = 1.67361 \cdot 10^{-6}$$

that means

$$N < 5975$$

Under the same assumptions as before, we could be interested in accuracy not in the whole range $0 \leq p_0 \leq 1$ but in a smaller range. For example we have a level of significance $\alpha = 0.01$ and we are interested only in the probability that $p < \alpha$. In this case we need accuracy only around $p = 0.01$:

$$C_{\chi^2} f_{\chi^2}\left(F_{\chi^2}^{-1}\left(1-0.01\right)\right) = 5.70634 \cdot 10^{-8}$$

of course in this case we must require a much smaller error, for example $|p - p_0| < 0.001$; with this bound we get

$$N < 17524$$

As last example, consider to have $n = 2^{20}$ and $N = 10000$. We want to know if it is possible to have a maximum average error $|p - p_0| < 0.01$. We need to find all $k$ for which

$$C_{\chi^2} \sup_{0 \leq p_0 \leq 1} f_{\chi^2}\left(F_{\chi^2}^{-1}\left(1-p_0\right)\right) < 10^{-6}$$

Note that despite the trend, $C_{\chi^2} \sup_{0 \leq p_0 \leq 1} f_{\chi^2}\left(F_{\chi^2}^{-1}\left(1-p_0\right)\right)$ is not strictly increasing. For this reason we can expect a non-compact set of solution. In the example we have

$$k = \{3, 4, 5, 6, 9\}$$

## V. Conclusion

In this paper we have analyzed the Frequency Test included in the well known NIST SP 800-22 test suite looking for deviation from uniform in the p-value distribution. In particular we have supposed to ignore the errors introduced by the normal approximation used in the test but still consider a discrete distribution. We then focused on the error given by approximating a discrete distribution with a continuous distribution. We were able not only to express an upper bound on the reliability of a second-level test based on the Frequency test, but also to give an expression for the average error on the second-level p-value. This is supported by experimental results.

The condition we have introduced here can be used to choose the parameters of a second-level test; in this paper we have presented few examples of how to perform a reliable second-level test given different targets and conditions.

## Acknowledgment

This work has been supported by MIUR under the FIRB framework.

## References

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

[2] Cryptography Research, "Evaluation of VIA C3 Nehemiah Random Number Generator", white paper prepared by Cryptography Research, Inc., San Farncisco (USA). February 27, 2003. Available at http://www.cryptography.com/resources/whitepapers/VIA_rng.pdf

[3] idQuantique, "Random Numbers Generation using Quantum Physics" white paper, 2004. Available at http://www.idquantique.com/products/files/quantis-whitepaper.pdf

[4] F. Pareschi, G. Setti and R. Rovatti, "A Fast Chaos-based True Random Number Generator for Cryptographic Applications," in *Proceedings ESSCIRC2006*, pp 130-133. Montreux, Switzerland, 19-21 September 2006.

[5] National Institute of Standard and Technology, "A statistical test suite for random and pseudorandom number generators for cryptographic applications", Special Publication 800-22, May 15, 2001. Available at http://csrc.nist.gov/rng/SP800-22b.pdf

[6] G. Marsaglia, "The diehard test suite", 2003. Available at http://www.csis.hku.hk/~diehard/

[7] W. Killmann and W. Schindler, AIS 31: Functionality classes and evaluation methodology for true (physical) random number generators, version 3.1, Bundesamt fur Sicherheit in der Informationstechnik (BSI), Bonn, 2001.

[8] H. Cramer, "Mathematical Methods of Statistics", Princeton University Press, September 1946.

[9] F. Pareschi, R. Rovatti, and G. Setti, "Second Level NIST Randomness Test for Improving Test Reliability", in *Proceedings of ISCAS2007*, pp. 1437–1440. New Orleans (USA), May 27–30, 2007.

[10] A. N. Shiryaev, and R. P. Boas "Probability" (Graduate Texts in Mathematics), Springer-Verlag, 1995.

[11] L. Blum, M. Blum, and M. Shub, "A Simple Unpredictable Pseudo-Random Number Generator", in *SIAM Journal on Computing*, vol. 15, pp. 364–383, May 1986.