*Article Submitted to *** Unknown Journal ****

A New Statistical Approach to Evaluating Random Number Generators

MELVIN J.HINICH*[1] AND EDUARDO M. A. M. MENDES†[2]

[1]*Applied Research Laboratories, University of Texas at Austin, Austin TX 78713-8029 USA*

[2]*Departamento de Engenharia Eletrônica, Universidade Federal de Minas Gerais, Av. Antônio Carlos 6627, Belo Horizonte, MG, Brasil, 31.270-901, Tel: +55 (31) 3449 4862, Fax: +55 (31) 3449 4850*

Abstract

In this paper, we propose the use of the bispectrum based tools to evaluate the statitiscal quality of a pseudo-random generator. Two well-know implementations of a pseudo-random generator and a new idea from physics were used as an example of how to use these statistical tools.

Keywords: random numbers, bispectrum, independence

1. INTRODUCTION

Algorithms are required to generate pseudo-random variates that mimic the statistical properties of the true stochastic model that is selected as the basis of the simulation. The typical procedure for generating a sequence of pseudo-random variates is to use an algorithm for generating uniform (0, 1) pseudo-independent variables and then transforming these (0, 1) variates to obtain a sequence with the desired statistical properties.

Pseudo-random uniform $(0,1)$ sequences are employed in a variety of applications such as signal encvryption, spread spectrum, randomization of experiments and in simulations. See, for instance, [1, 2, 3, 4] and references therein.

The most widely used method for generating (0, 1) pseudo-independent variables utilizes one or more congruential generators. A congruential generator whose seed is properly chosen will generate variates that will have a uniform distribution on the unit interval but the variates will not be independent since the generator is deterministic. The algorithm must produce variates whose statistical properties conform to the probabilistic implications of independence. For example the sample correlation function of a generated sequence should statistically be indistinguishable from the sample correlation function of an independent uniform (0, 1) process. The sample correlations should be statistically near zero for the sample size used in the test.

The higher-order statistical properties of available pseudo-random generators may not be sufficiently

*E-mail: hinich@mail.la.utexas.edu

†E-mail: emmendes@cpdee.ufmg.br

reliable to ensure that their lack of true independence does not cause significant deviations in the properties of simulated results from the true properties given the parameters used in the simulations. One should never take the statistical properties of pseudo-random generators for granted.

In this paper we demonstrate that the bispectrum based tests of zero bispectrum and linearity produced by Hinich and his collaborators can be used to benchmark pseudo-random generators. The bispectrum is the double Fourier transform of the bicovariance function $b_x(\tau_1,\tau_2) = Ex(t)x(t+\tau_1)x(t+\tau_2)$ of a stationary random process $x(t)$.

The statistical test proposed here follows the line of tests discussed in [4]. It is not a competing test but an additional test that can be included in the existing library of tests already distributed in the NIST's web site [4].

This paper is divided as follows. The background material is given in Sec. 2. Examples using two well-known implementations of a pseudo-random generator and a new idea from physics are given in Sec. 4. Sec. 5 summarizes the results presented in this paper.

2. BASIC STATISTICAL TESTS FOR PSEUDO-RANDOM VARIATES

The simulations that will report in the next section use a total of $N = 10^{10}$ pseudo-random uniform (0,1) variates. The first step in a statistical evaluation of a pseudo-random algorithm is to computer the following four statistics: 1) the mean μ, 2) the standard deviation σ, 3) the skewness $\gamma = \frac{\mu_3}{\sigma^3}$, and 4) the kurtosis $\kappa = \frac{\mu_4}{\sigma^4} - 3$. The mean of a uniform (0,1) is 0.5, $\sigma = 0.2887$, $\gamma = 0.$, and $\kappa = -1.2$.

The random number generators used in this work are: Intel random number generators (it comes with the C and Fortran compiler by Intel), the Matlab function rand and the dynamical system introduced by González and co-workers [5, 6] that are claimed to generate true random variates.

Table 1 shows the results for the three generators. The mean for the Intel and Matlab generators are correct to four decimals places. The approximate standard errors for the estimates are of order 10^{-5} and the sums in the computation of these statistics are in double precision. Thus we expect the results to be correct to four decimals at least. But the mean for the Gonzalez et al algorithm has an negative bias of –0.003 which is highly statistically significant for our large sample size.

The standard deviations for the Intel generator is correct to four decimals. The Gonzalez et al generator has a positive bias of 0.013 which is also highly statistically significant. The Matlab standard deviation has a significant positive bias of 0.0003 which is not much of a problem for most simulations but the generator needs improvement.

The skewness estimates for both the Intel and Matlab generators are correct to four decimal standard but the Gonzalez et al skewness has a highly significant bias of 0.0094.

Finally the kurtosis estimates for both the Intel and Matlab generators are correct whereas the Gonzalez et al estimate has a significant negative bias of –0.003.

These results show that the Gonzalez et al generator fails to match the first four moments of the uniform (0,1) distribution for the large sample simulation that we ran and thus it is not a credible algorithm for applications that require statistical precision of the pseudo-random variates.

We now turn to the use of bispectrum analysis to address the quality of the these random number generators advancing the work of [7].

Generator	μ	σ	γ	κ
Intel	0.5000	0.2887	0.0000	-1.2000
Gonzalez	0.4970	0.2900	0.0094	-1.2003
Matlab	0.5000	0.2890	0.0000	-1.2000

Table 1: Mean, Standard Deviation, Skewness and Kurtosis for the random generators

2.1. The Bispectrum Analysis

Before proceeding to the description of the tests sued in this work, it is important to define what is meant by a stochastic *linear* process. A random sampled process $\{x(t_n)\}$ is linear if it is of the form $x(t_n) = \sum_{k=-\infty}^{\infty} h(t_{n-k})\varepsilon(t_k)$ where $\{\varepsilon(t_n)\}$ is a sequence of independent and identically distributed random variables, $\sum_{k=-\infty}^{\infty}|h(t_k)| < \infty$ and $t_n = n\tau$ for a fixed sampling rate τ^{-1}. Using signal processing terminology $\{x(t_n)\}$ is the output of a stable linear filter whose impulse response is $\{h(t_n)\}$ and whose input is the *pure white noise* process $\{\varepsilon(t_n)\}$.

The signal's bispectrum is

$$B(\omega_1,\omega_2) = \sum_{\tau_1=-\infty}^{\infty}\sum_{\tau_2=-\infty}^{\infty} c_{xxx}(\tau_1,\tau_2)\exp\left[-i2\pi(f_1\tau_1+f_2\tau_2)\right] \tag{1}$$

where $B(\omega_1,\omega_2)$ is the bispectrum of the signal [8] and $c_{xxx}(\tau_1,\tau_2)$ is the bicorrelation.

The bispectrum is computed using conventional nonparametric methods. When computing the bispectrum we take advantage of two properties: 1) Bispectrum values are approximately normally distributed [9], and 2) Bispectral estimators are approximately independent across frequencies [10].

Suppose that we have a sample $\{x(1),...,x(N)\}$that we partition into $P = [N/L]$ non-overlapping frames of length L where the last frame is dropped if it has less than L observations. The pth frame is $\{x_p(1),...,x_p(L)\} = \{x((p-1)L+1),...,x(pL)\}$. The discrete Fourier transform of the pth frame is $X_p(k) = \sum_{t=1}^{L} x_p(t)\exp\left(-i2\pi\frac{kt}{L}\right)$ and the periodogram of the mth frame is $\frac{1}{L}\left|X_p(k)\right|^2 = \frac{1}{L}X_p(k)X_p(-k)$. Because $N \simeq LP$ the frame-averaged estimate of the spectrum at frequency $\omega_k = \frac{2\pi k}{L}$ is

$$\hat{S}(\omega_k) = \frac{1}{N}\sum_{p=1}^{P}\left|X_p(k)\right|^2 \tag{2}$$

Then $E\left[\hat{S}_x(f_k)\right] = S(\omega_k) + O\left(\frac{1}{L}\right)$ where the error term of order $1/L$ is due to the frame windowing of the spectrum, and the variance of the estimate for large values of L and P is $\frac{1}{P}S^2(\omega_k)$.

Similarly, the frame-averaged estimate of the bispectrum at frequencies $(\omega_{k_1}, f_{\omega_2})$ is

$$\hat{B}(\omega_{k_1},\omega_{k_2}) = \frac{1}{N}\sum_{p=1}^{P} X_p(k_1)X_p(k_2)X_p(-k_1-k_2) \tag{3}$$

with $E\left[\hat{B}(\omega_{k_1},\omega_{k_2})\right] = B(\omega_{k_1},\omega_{k_2}) + O\left(\frac{1}{L}\right)$ and variance for large L and P expressed as $\frac{L}{P}S(\omega_{k_1})S(\omega_{k_2})S(\omega_{k_1}+\omega_{k_2})$

The ideas briefly laid here are the base for the Hinich test for zero bispectrum and linearity of stationary time-series [9].

3. Zero Bispectrum and Linearity tests

Specific statistical properties of an estimate of the bispectrum are now discussed in order to understand the logic behind the tests [9] of linearity and zero bispectrum (The Fortran program written by Hinich, available, upon request, finds K for whatever band is selected. In [11], $q = 0.8$ is used but a more robust test uses the $q = 0.9^{th}$ quantile based upon numerous tests of the method on various real and artificial data.).

Let $\{x(t_n)\}$ denote a zero mean strictly stationary random process that is bandlimited and sampled at a rate sufficient to avoid aliasing with $t_n = n\tau$. To simplify notation let τ =1. The bicorrelation of the process is $c_{xxx}(\tau_1,\tau_2) = Ex(n)x(n+\tau_1)x(n+\tau_2)$ and its bispectrum is the two-dimensional Fourier transform $B_x(\omega_1,\omega_2) = \sum_{n_1=-\infty}^{\infty}\sum_{n_2=-\infty}^{\infty} c_{xxx}(\tau_1,\tau_2)\exp[-i(\omega_1\tau_1+\omega_2\tau_2)]$. For further details see [9, 10].

For linear processes, it follows that $B_x(\omega_1,\omega_2) = \mu_{3\varepsilon}H(\omega_1)H(\omega_2)H(-\omega_1-\omega_2)$ where $H(\omega)$ is the Fourier transform of h(k), and $\mu_{3\varepsilon} = E\varepsilon^3(n)$ is the skewness of $\varepsilon(n)$. Since the skewness of an uniform distribution is zero, the bispectrum of the pseudo-random sequence should be zero.

In what follows a detailed description of the tests for zero bispectrum and linearity is given.

3.1. Testing Zero Bispectrum

If the density of the noise variates $\{\varepsilon(n)\}$ is symmetric about its mean (zero) then the skewness is zero, and its bispectrum will not be statistically significant different from zero. The Hinich test statistic [9] to test for input symmetry is the sum, S, over the $V(\omega_1,\omega_2)$ for the K bifrequencies. Since the bispectral estimates are approximately independent across the bifrequency grid, S will be approximately distributed as a $\chi^2_{2M}(0)$. The non-central parameter is zero for the null hypothesis since the skewness is zero. Let $F(s)$ denote the cumulativa distribution function of the $\chi^2_{2M}(0)$ and let $U = F(S)$, then the statistic U has an uniform $0,1$ distribution under the null hypothesis that the bispectrum is zero. The null hypothesis is rejected if U is greater than a threshold that is determined by the size probability required by the user

3.2. Testing Linearity

Let $\hat{B}_e(\omega_1,\omega_2)$ denote the estimate of the bispectrum of the residuals at bifrequency (ω_1,ω_2) using a resolution bandwidth of Δ. Using Theorem 5.3.1 of [12] it can be shown that the real and imaginary parts $N^{\frac{1}{2}}\Delta[\hat{B}(\omega_1,\omega_2) - B(\omega_1,\omega_2)]$ are independently distributed and gaussian with mean zero and variance $\sigma_e^2/2$ as N goes to infinity. Thus the large sample distribution of the *normalized skewness function* defined by $V(\omega_1,\omega_2) = 2N\Delta^2\sigma_e^{-6}[B(\omega_1,\omega_2)]^2$ is $\chi^2_2(\lambda)$, a chi squared with two degree-of-freedom and non-centrality parameter $\lambda = 2N\Delta^2\sigma_e^{-6}[\mu_{3e}]^2$ for each bifrequency for the null hypothesis of independence. This parameter is estimated by $\hat{\lambda}$, the average skewness $V(\omega_1,\omega_2)$ for all bifrequencies in the bispectrum's principal domain.

Let $F(\upsilon|\lambda)$ denote the cumulative distribution function of a $\chi^2_2(\lambda)$ random variable and let $U(\omega_1,\omega_2) = F[V(\omega_1,\omega_2)|\lambda]$. The normalized skewness values are then transformed into uniform (0,1) variates under the null hypothesis. Then the modified Hinich test for linearity (independence of the residuals) is to compute the q^{th} quantile, Q of the sorted U statistics for all K bifrequencies in the principal domain, where the user selects q. If the whole bandwidth up to the folding frequency is used then there are approximately K=$\frac{1}{16\Delta^2}$ bifrequencies in the principal domain [11]. The q^{th} quantile is approximately gaussian with mean q and variance σ^2=q(1-q)/K under the null hypothesis, and we use the q=0.9^{th} quantile [11].

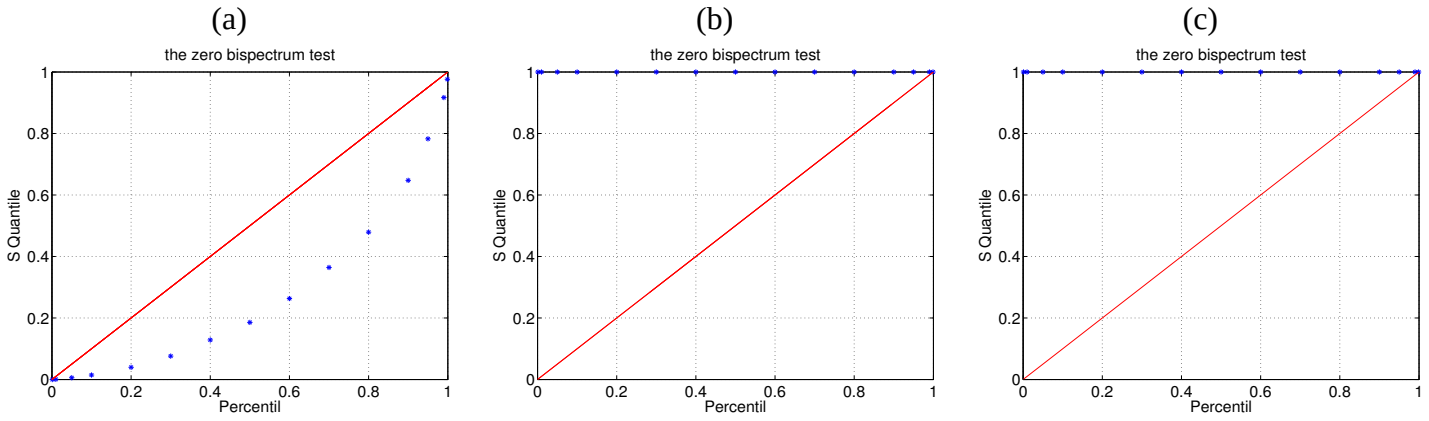

Figure 1: Zero Bispectrum S results: (a) Intel random number generator, (b) Gonzalez model and (c) Matlab rand function.

Using these estimates of the mean and variance, the asymptotic gaussian distribution the 5% critical value for the one tailed test of linearity is easily found to be $0.9 + 0.492/\sqrt{K}$. If the 0.9^{th} quantile is larger than this value the null hypothesis of linearity is rejected at the 5% size level. Thus under the null hypothesis 5% of such statistics would be larger than the above value.

4. DATA ANALYSIS AND SIMULATION

We ran 5,000,000 numbers in sequence with 2000 diferent seeds for each one of three generators listed in section 2. Then we computed the following quantiles for the S statistic: 0.001, 0.01, 0.05, 0.1 0.2 ,0.3 0.4, 0.5, 0.6, 0.7 0.8 0.9, 0.95, 0.99, 0.999. The quantiles for the Q are exact the same.

Figure 1 shows the plot of the quantiles for the zero bispectrum S statistic. The 45% degree line from zero to 1 is the expected value of the quantiles. If the generator is producing independent uniform variates, we expect that the sample quantiles should lie on or very close to the expected line. For the Intel random number generator, the qauntiles are below of the expected which means that they are too small. For the Matlab and González the S statistics are just above zero and so the first quantile is one and therefore all quantilea are one.

For Q statistic quantiles (See Figure 2), the Intel results are also below the expected value line but a bit closer than the zero bispectrum results. For the Gonzaález generator, the quantile Q results are lower than the Intel's. For the Matlab generator, the first quantiles jumps to one and the others stay at one. None of the results are statisfactory using the bispectrum based measuring tools.

5. CONCLUSIONS

The two bispectrum based tests will detect that a pseudo random sequence is not an indenpendent distributed random process and therefore it does not comform to the statistical implications of indenpendence. Pseudo random numbers are not statistically independent, but the idea is to generate pseudo random numbers that can fool the data mining algorithms as if they are pure noise.

We used two well-known random generators and a new idea from physics as an example on how to use the bispectrum tools to evaluate the statistical quality of a pseudo random number generator.

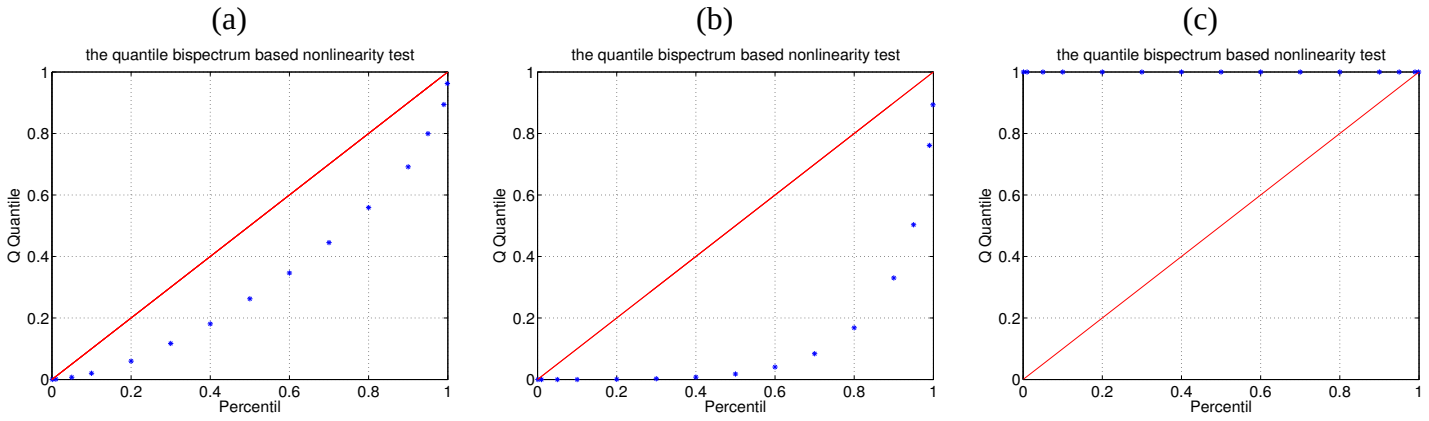

Figure 2: Linearity *Q* results: (a) Intel random number generator, (b) Gonzalez model and (c) Matlab rand function.

Finally, we conjecture that by the use of higher-order cumulant tests one can always detect that man-made sequences with symmetric density functions are not stochastic.

References

1. James Godman, Abram P. Dnacy, and Anatha P. Chandrakasan. An energy/security scable encryption processor using an embedded variable voltage dc/dc converter. *IEEE Journal of Sold Circuits*, 13(11):2799–1809, 1998.

2. F. James. Chaos and randomness. *Chaos, Solitons and Fractals*, 6:221–226, 1992.

3. J. Orlin Grabble. Michael riconoscinto on encryption. *http://www.orlingrablle.com/riscono.htm*, 1997.

4. A. Rukhin, J. Soto, J. Nechvatal, M. Smid, E. Barker, S. Leigh, M. Levenson, M. Vangel, D. Banks, A. Heckert, J. Dray, and S. Vo. A statistical test suite fo random and pseudorandom number generators for cryptographic applications. *NIST Special Publication 800-22*, 2001.

5. J. A. González, L. I. Reyes, J. J. Suárez, L. E. Guerrero, and G. Gutiérrez. A mechanism for randomness. *Physics Letters A*, 295:25–34, 2002.

6. J. A. González, L. I. Reyes, J. J. Suárez, L. E. Guerrero, and G. Gutiérrez. Chaos-induced true randomness. *Physica A*, 316:259–288, 2002.

7. John W. Dalle Molle, Melvin J Hinich, and Douglas J. Morrice. Higher-order cumulant spectral-based statistical tests of pseudo-random variate generators. In J. J. Swain, D. Goldsman, H. C. Crain, and J. R. Wilson, editors, *Proceedings of the IEEE Winter Simulation Conference*, pages 618–624, 1992.

8. M. J Hinich and C. S. Clay. Application of discrete fourier transform in the estimation of power spectra, coherence and bispectra of geophysical data. *Reviews of Geophysics*, 6(3):347–363, 1968.

9. M. J. Hinich. Testing for Gaussianity and Linearity of a Stationary Time Series. *Journal of Time Series Analysis*, 3(3):169–176, 1982.

10. M. J. Hinich and H. Messer. On the Principal Domain of the Discrete Bispectrum of a Stationary Signal. *IEEE Trans. on Signal Processing*, 43(9):2130–2134, 1995.

11. R.A. Ashley, D.M. Patterson, and M.J. Hinich. A Diagnostic Test for Nonlinear Serial Dependence in Time Series Fitting Errors. *Journal of Time Series Analysis*, 7:165–178, 1986.

12. D. Brillinger. *Time Series, Data Analysis and Theory*. Holt, Rinehart, and Winston, New York, 1975.