Proposal for an enhanced autocorrelation test for random number generators

Antoine Levotre[1], Cécile Dumas[2], Philippe Elbaz-Vincent[1]

[1]Univ. Grenoble Alpes, CNRS, IF, 38000 Grenoble, France

[2]Univ. Grenoble Alpes, CEA, LETI, DSYS, CESTI, F-38000 Grenoble

Introduction

Randomness is a cornerstone of the security of cryptographic protocols, with random numbers being used for countermeasures, ephemeral data, or key generation. As such, numbers produced by random number generators (RNG) must have excellent statistical properties. In particular, generated number must be perfectly uncorrelated to prevent an attacker from gaining an advantage on the cryptographic system based on the knowledge of previously generated numbers.

A methodology to ensure that numbers have satisfactory statistical properties is to apply statistical tests on the output of the RNG. On the matter of correlations, one such standard test is the Autocorrelation test, which was present in the German standard AIS 31 until version 2.35 [1].

In our work, we provide an analysis of this test to prove that, while inexpensive, it is not optimal for the evaluation of correlations in bit sequences, and can even fail when no correlation exists in the sequence. We then propose an alternative autocorrelation test, which will be proven to be optimal for the evaluation of Pearson autocorrelations in bit sequences.

We then further expand our correlation testing methodology with the use of the partial autocorrelation function (PACF). The use of this function corrects a drawback of our first proposed test, and allows for a more direct link between detected statistical anomalies and the parameters of the tested RNG.

The models we present in this work are all **stationary**, as per the considerations of both the AIS 31 and the NIST SP 800-90, which demand that stochastic models for RNG are stationary.

Result and discussions

Analysis of the Autocorrelation test from the AIS 31.

The autocorrelation test from the AIS 31 applies the following statistic $\mathcal{A}_k$ on a bit sequence $(b_i)_{1\leqslant i\leqslant n}$:

$$\mathcal{A}_k = \sum_{i=k}^{N+k} b_{i-k} \oplus b_i.$$

In the standard, $1 \leqslant k \leqslant 5000$, $N = 5000$ and the test fails if $\mathcal{A}_k \notin [2326, 2674]$.

The statistic has a complexity of $O(N)$, which makes it very efficient to use. However, we show that this test can fail even when a sequence presents no correlation (in particular, no Pearson autocorrelation).

The Pearson autocorrelation of lag k of a stationary discrete stochastic process $\{B_i\}$ is expressed as:

$$\rho_k = \frac{1}{\sigma^2}\mathbb{E}\left[(B_i - \mu)(B_{i-k} - \mu)\right]$$

with μ and σ^2 the expectancy and variance of the process $\{B_i\}$. $\rho_k \in [-1, 1]$ and $\rho_k = 0$ means the process suffers from no autocorrelation of lag k, while $\rho_k \approx -1$ or $\rho_k \approx 1$ means strong correlations exist.

For a sequence built as the concatenation of independent random drawings of bits, such that $P(B_i = 1) = 0.8$, the expectancy of $\mathcal{A}_k$ is $\mathbb{E}[\mathcal{A}_k] = 5000 \times 2 \times P(B_i = 1) \times P(B_i = 0) = 1600$. The test is then largely expected to fail, but the Pearson autocorrelation of such a sequence is $\rho_k = 0$.

For this reason, the statistic $\mathcal{A}_k$ is not optimal for the evaluation of correlations in bit sequences, and we propose a new statistic, which we will call the Enhanced autocorrelation statistic.

Enhanced autocorrelation statistic

The issue with the statistic $\mathcal{A}_k$ is that it is based on the joint probability $P(B_i, B_{i-k}) = P(B_i|B_{i-k}) \times P(B_i)$, which causes some uncorrelated sequences to make the test fail due to the presence of $P(B_i)$.

For a stationary binary process $\{B_i\}$, we can prove that the Pearson autocorrelation of lag k of the process is tied only to $P(B_i|B_{i-k})$, and more specifically that:

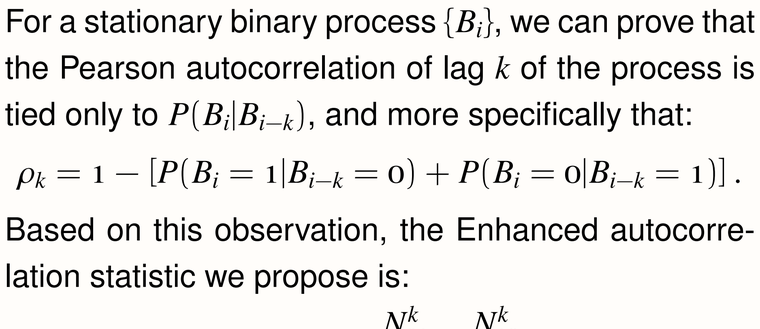
$$\rho_k = 1 - \left[P(B_i = 1|B_{i-k} = 0) + P(B_i = 0|B_{i-k} = 1)\right].$$

Based on this observation, the Enhanced autocorrelation statistic we propose is:

$$\mathcal{A}^\star_k = 1 - \frac{N^k_{01}}{N_0} - \frac{N^k_{10}}{N_1}$$

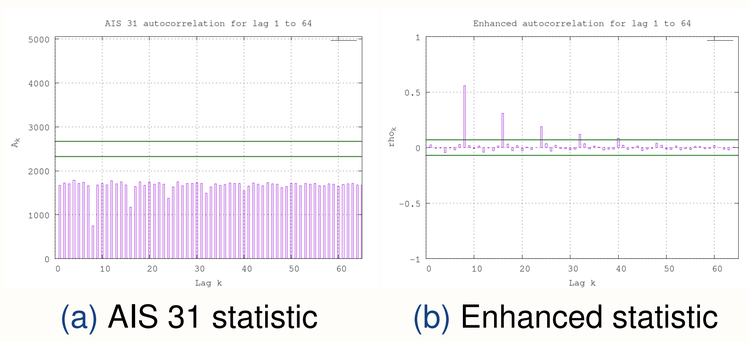
with, N_x the number of bits x, and N^k_{xy} the number of pairs $(b_{i-k} = x, b_i = y)$ in the sequence. The complexity of this statistic is $O(N)$ (with $N + k$ the total number of bits in the tested sequence) and we can prove that its expectancy is exactly the Pearson autocorrelation of lag k of $\{B_i\}$. It is thus optimal for the evaluation of the autocorrelation of a binary sequence.

To compare the behavior of our new statistic with the one from the AIS 31, we applied both on various simulated sequences. The first sequence is the one we mentioned in the previous subsection, which presents a strong global bias (denoted by the term $p_1 = P(B_i = 1)$), but no autocorrelation. The second sequence presents both a strong bias, and a strong Pearson autocorrelation of lag 8.



(a) AIS 31 statistic (b) Enhanced statistic

Figure 1: AIS 31 vs enhanced autocorrelation statistic, $N = 5000$, $p_1 = 0.8$, $\rho_8 = 0$



(a) AIS 31 statistic (b) Enhanced statistic

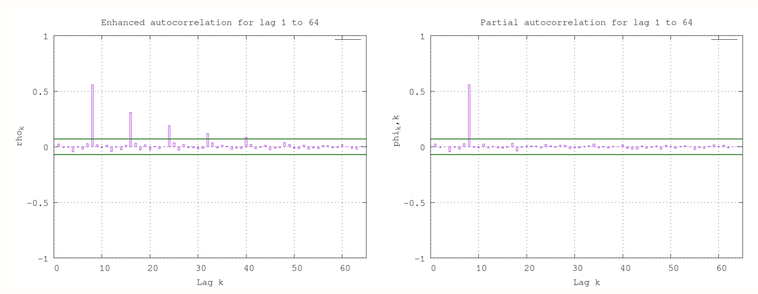
Figure 2: AIS 31 vs enhanced autocorrelation statistic, $N = 5000$, $p_1 = 0.8$, $\rho_8 = 0.6$

Figure 1, shows that, while a strong bias alone can indeed make the Autocorrelation test from the AIS 31 fail for any lag k, it does not impact our enhanced statistic at all (see Fig. 1b). Fig 2 then shows that, in the presence of both a global bias and correlations (here of lag $k = 8$), our enhanced statistic is able to properly characterize the correlation, and is still not impacted by the bias.

However, despite only simulating a single correlation phenomenon (of lag 8), our statistic highlights multiple correlation anomalies, for every lag multiple of 8. We can indeed prove that with our statistic, a correlation phenomenon of lag k will translate into an infinite set of detected anomalies, such that the amplitude of the anomaly detected for a lag mk is ρ^m_k.

Partial Autocorrelation Function

To make a more direct link between the output of the statistic and the original correlation phenomena which impact the tested RNG, we then looked for a way to eliminate the "propagated" anomalies. One way to achieve this was to use the partial autocorrelation function (PACF) [2], a tool designed with the explicit objective of getting rid of propagations in the correlation terms of a time series. More specifically, the PACF is based on an autoregressive representation of the terms ρ_k, in which $\rho_k = \phi_{kk-1}\rho_1 + ... + \phi_{k1}\rho_{k-1} + \phi_{kk}$. The term $\phi_{k,k}$ is the PACF of order k, and can be interpreted as the part of the amplitude of ρ_k which is not due to the propagation of any ρ_j, $j < k$. The term $\phi_{k,k}$ is computed as the quotient of two matrices of the terms $(\rho_j)_{1\leqslant j\leqslant k}$, as detailed in [2] (Sect. 3.2.5).



(a) Enhanced statistic (b) PACF

Figure 3: Enhanced autocorrelation statistic vs PACF, $N = 5000$, $p_1 = 0.8$, $\rho_8 = 0.6$

Figure 3 shows that the computation of $\phi_{k,k}$ indeed eliminates the propagated correlation anomalies, while still perfectly characterizing the original phenomenon at lag 8.

Summary and conclusions

In this work, we proposed a potential replacement for the Autocorrelation statistic from the AIS 31, which keeps the low complexity of the original statistic, but is optimal for the evaluation of the Pearson autocorrelation of stationary binary processes. While the Autocorrelation statistic is not present anymore in version 3.0 of the AIS 31, it still seems interesting to have a simple, yet efficient tool to measure correlations, alongside the more complex statistics now present in the German standard.

We also propose the use of the partial autocorrelation function to eliminate the propagated correlation phenomena which were highlighted by our first statistic, in order to have a more direct link between the statistic output and the defects of the tested random number generator.

References

[1] M.Peter and W.Schindler. A proposal for functionality classes for random number generators. Online. Available at: https://www.bsi.bund.de, 2022.

[2] G. E. P. Box, G. M. Jenkins, G. C. Reinsel, and G. M. Ljung. Times series analysis : Forecasting and control. Fifth edition. Wiley, 2015.