The TestU01 Bigcrush, an emperor without clothes?

Introduction

The TestU01 (see Wikipedia for details) is generally regarded as the "gold standard" for the statistical testing of random number generators. For a long time, we also believed this to be true. Today we have to think of the fairy tale "The emperor without clothes"! In fact, today we would like to declare in all openness: The emperor (gold standard) is standing there without clothes, in other words: The expectations we have all placed in him are not being fulfilled! That is why we are publishing on our home page all the information we have gained from carrying out 250,000 BigCrush tests. Let the public find out whether our judgment is justified.

We would like to emphasize that we are only the reporters of TestU01's "sins", so to speak, and we cannot perform a deeper analysis of the incorrect mathematical results due to time constraints. Therefore, we provide all the data. But what we can shed more light on, as old IT pioneers, are the technical errors in the programming. It is difficult to estimate how many judgments are based on statements by TestU01 that may have been incorrect.

Strange results in the Standard Deviation (Bernoulli process)

Many tests are useful and probably work as intended. However, some have gone completely off the rails, whether due to mathematical systematics or errors in implementation remains to be seen. Every statistician and every mathematician or researcher in the natural sciences will surely agree with us when we mention the following facts: An analysis of the SD(Bernoulli process, v*p*q) results in 4010 values with an SD of over 10, out of a total of 128068 values (254 p-values x 101 -202, due to no 177 and 179 in MT19937). The expected value in every bin is 500. P-values number 197 and 199 take the cake with this distribution:

SD of bin .50:	P-v No 197	P-v No 199
AHS-RNG-d	527.08945	536.52825
AHS-RNG-s	531.67401	536.16868
AHS-RNG-t	527.49397	521.06660
MT19937	538.50590	528.97721
XOshiro256**	530.68519	530.95487

It is known that with 128068 values and a statistically correct distribution, about 68.27 % of the SD values are expected to be above -1 and below +1. In our evaluation, however, it is only 58.044 %. Under these circumstances, how could TestU01 be explained as a reference in the statistical testing of RNGs? The answer is probably: because 18 years ago there was no better software package available for testing RNGs. In computer science, however, this time span is half an eternity, and it seems essential to us to publish an up-to-date description of the situation.

Mixing discrete math with continuous math error-prone

Those who have been involved in testing RNGs for some time will probably remember the discussion in 2006 about a mathematical error in one of the 16 NIST tests. In a paper by Song-Ju Kim, Ken Umeno and Akio Hasegawa, it was shown that mixing discrete math with

continuous math led to a situation where it became impossible to pass the test. The mixing created a discrepancy in the distribution: Two events fell into one bin, but three fell into another. Although the authors suggested a correction, NIST removed this test from the suite, and since then there have only been 15 tests in the NIST test suite. It is astonishing that the authors of TestU01 have not yet taken this fact into account, because the May 16, 2013 edition of the "user's guide" still mentions 16 NIST tests.

We ask this question because the same problem occurs in the statistics from 3 to 12, for example. The bin with p=.50 receives twice the expected value, but in these 10 tests the bin with p=.49 always comes up empty. This discrepancy also occurs several times with the other bins, even as no other still remains empty. If you remark that there are some cases where 11 counts need to fall in only 10 bins, no good statistical distribution is possible An evaluation of the highest p-values of the five generators revealed the following: Of the 10 highest p-values, the generator d has 7, the s has 6, the t has 7, the m has 7 and the x has 6 values of p-v number 11 or 12! Of the 30 highest p-values, the generator has d 13, the s 16, the t 22, the m 17 and the x 14 values. This is despite the fact that p-v numbers 11 and 12 do not even provide one percent of the p-v. According to our estimates, the values calculated by TestU01 are 2 or 3 powers of ten too high!

The Summary mess

The "Summary" at the end of the report shows the p-values of the statistics that lie outside a range of [0.001 - 0.999]. All others are said to be "passed". The numbering in the "Summary" is the number that corresponds to the 106 statistics. However, the third line of the summary reads: Number of statistics: 160. A simple transposed number that nobody noticed for 18 years? For the sake of clarity, we have numbered each individual line with the output of a P-Value consecutively, resulting in 254 individual values. This decision is based on the realization that each evaluation of a p-value is actually a test of the quality of the RNG. In any case, the specification of the statistic number would have required a two-level numbering, e.g. statistic 105.2. This negligence results in the following serious error: When listing the exceedances of the specified value range, only the last exceedance from a specific statistic is listed, although there are several, sometimes larger, deviations in the same statistic. Different statistics have up to 7 p-values. Such negligence in the quality control of important software, that is generally used as a decision-making aid in the selection of RNGs for scientific research, must be reviewed.

It can certainly be argued that the "Summary" only states that the following tests have at least one value outside the specified limits. But by reading the statment: "The following tests gave p-values outside [0.001, 0.9990]", one logically expects all outliers to be listed. Under no circumstances do you expect only the last one to be mentioned, regardless of whether it is the largest or not.

The IDQuantique Report No 2 as example for the discased error.

As indicated, we have run two evaluations of IDQuantique physical random numbers "hors concours". You can download these reports under "Downloads". It was not planned this way, but it turned out that these two reports are a good example of the fact we are criticizing here. In the second report, for example, we find the statistic 50 in the "Summary":

50 SampleProd, t = 8 1.1e-4

In contrast to this, however, is the expression of this statistic: (line 1937)

svaria_SampleProd test:

N = 40, n = 10000000, r = 0, t = 8

Kolmogorov-Smirnov+ statistic = D+ : 0.38

p-value of test : 7.0e-6 ***

Kolmogorov-Smirnov- statistic = D- : 4.84e-3

p-value of test : 0.9942

Anderson-Darling statistic = A2 : 8.08

p-value of test : 1.1e-4 ***

As you can easily see, the worse value is simply ignored.

In the two IDQ tests, 18 and 14 values are not mentioned in the “Summary”, but these are marked as “outside” with five stars in the text. They are not the last ones from a specific statistic.

Anderson-Darling a good choice?

In our opinion, the application of the Anderson-Darling statistic is a serious mistake. The reasoning behind our opinion is as follows: Two contradictory goals are mixed up here, in that on the one hand the general goal is the representation in a p-value of (0,1), corresponding to the probability for the randomness of the measured result, but on the other hand it is then emphasized (users guide, May 16,2013, p. 107) : "Why Anderson-Darling ? Because ... , and the Anderson.Darling test is particularly sensitive to detect that type of behavior." These are two contradictory definitions, and the result can be read in the analysis, 250000 BigCrushs, at p-values 11 and 12! You try to represent the probability as fair as possible, or you emphasize one part of the values to better detect the differences. But both targets can’t be combined.

Concerning the rounding

Unfortunately, the round is carried out without a system. Decisions are also made before the rounding. The notation “The following tests gave p-values outside [0.001, 0.9990];” means, according to our interpretation, that the limit points 0.001 and 0.9990 of the distance are in the “normal range” and therefore not “outside”. However, there are p-values of 0.9990 with ***** marking, which are also listed in the summary, although they do not belong there. On the other hand, there are also 0.9990 that do not receive five stars. This leads to the

assumption that the decision outside/not outside is already made before the round. This is questionable programming to say the least.

Are ”outsiders” some time not reported in the Summary?

We are currently still checking all 250,000 reports, as statistics have also emerged where “outsiders” have appeared in the details, but these were not listed in the “Summary” of the statistics. We will publish the result in a few days and announce it on our BILLBOARD. However, we would like to emphasize that this programming error did not affect our analyses. Our p-values were all extracted from the detailed texts and not from the summaries.

Conclusion

If anyone has doubts about the facts we are pointing here, we recommend to download the 254 pages book “250000 BigCrushs”, in Landscape Din A3, available on the navigation point “Downloads”. It will certainly be a pleasure to pass through the detailed presentation on the facts and statistics, figures from five different RNGs side by side, one page for every p-value number. You will certainly agree that you had never before seen such a presentation, and you will easily accept our remarks.