Testing the tests: using pseudorandom number generators to improve empirical tests

Paul Leopardi

Mathematical Sciences Institute, Australian National University.
For presentation at MCQMC 2008
University of Montreal, Canada, July 2008.
Parts of this work were conducted at the School of Physics, University of Sydney.

July 2008

Outline of talk

- Original problem: archiving physics codes
- Solution: SUPRANGEN
- Next problem: anomalies in TestU01 results
- Problems in run test and solutions
- Problems in overlapping serial tests and solutions
- Final TestU01 results

Original problem: archiving physics codes

- Most often in Fortran.
- Different compilers use different RAND.
  - Test results unrepeatable except possibly as distributions.
- Poor generators for normal distributions, etc.

Solution: SUPRANGEN

- Archive source code for the PRNG along with physics code.
- Library of PRNGs, including Mersenne Twister and Brent Xorgens.
- Both 32-bit and 52-bit double precision $\mathbf{U[0,1]}$ generators.
- Normal distribution adaptor.
- Interfaces and implementations in C and Fortran.
- Interface to GSL RNG.

TestU01 batteries

TestU01 (L'Ecuyer, Simard, 2007)

- "Utilities for empirical statistical testing of uniform random number generators."
- Includes pseudoDIEHARD battery.
  - Based on "DIEHARD battery of tests of randomness" (Marsaglia 1995).

Typical use of TestU01 to test SUPRANGEN generators:

- For each of 4 seeds repeat pseudoDIEHARD 256 times.
- Test resulting sequence of p-values against $\mathbf{U[0,1]}$.

Null Hypotheses

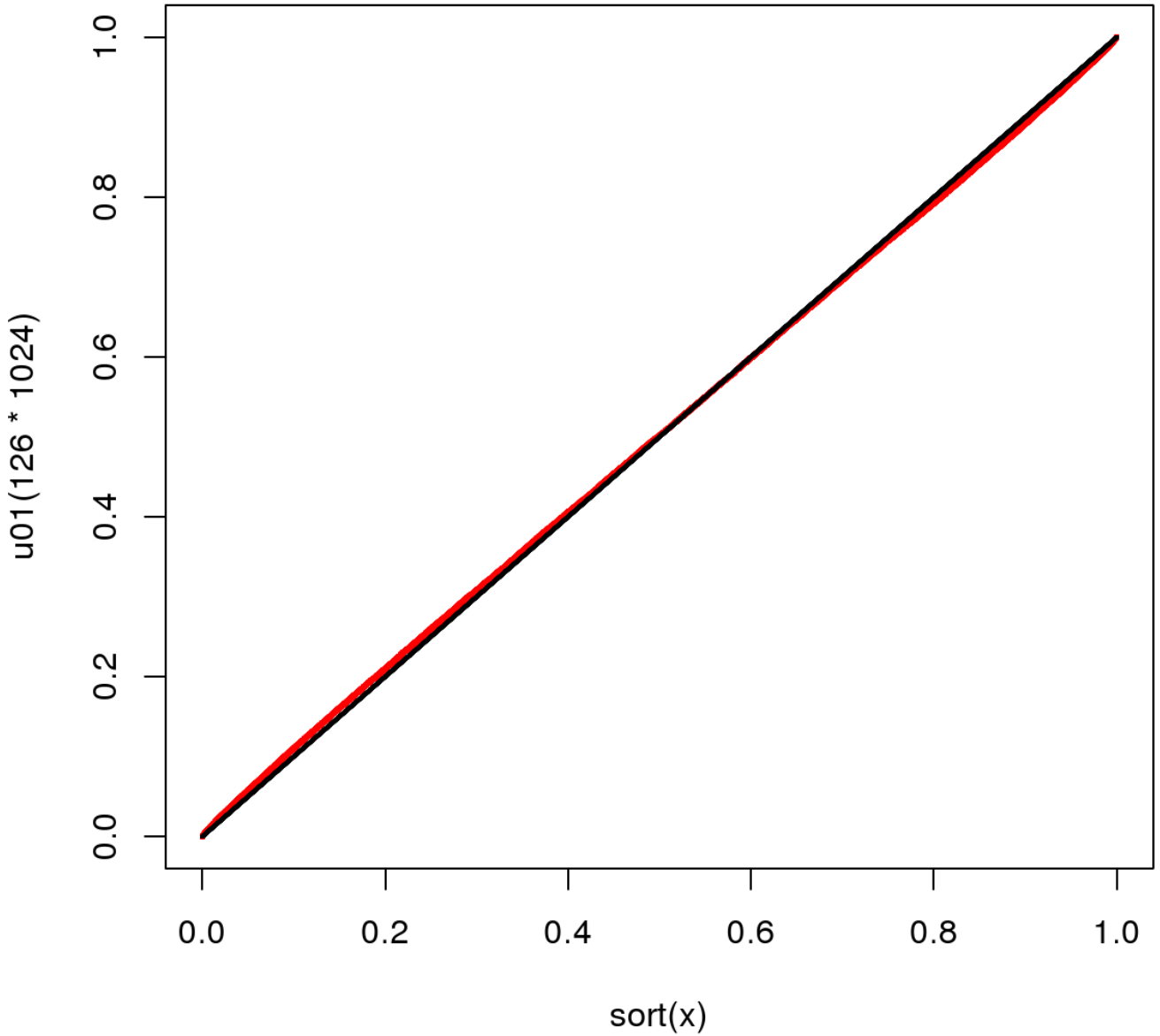
- A: Each test of a battery, when applied to a $\mathbf{U[0,1]}$ sequence, yields a p-value from $\mathbf{U[0,1]}$.
  - For tests using a statistic with a discrete distribution, A is not strictly true.
- B: Each PRNG under test generates a $\mathbf{U[0,1]}$ sequence, independent of seed, with different seeds yielding uncorrelated sequences.

- For a battery of tests on a single PRNG, it may be hard to distinguish failures of B from failures of A.
- Solution is to look for consistent failures of tests across multiple different "good enough" PRNGs.

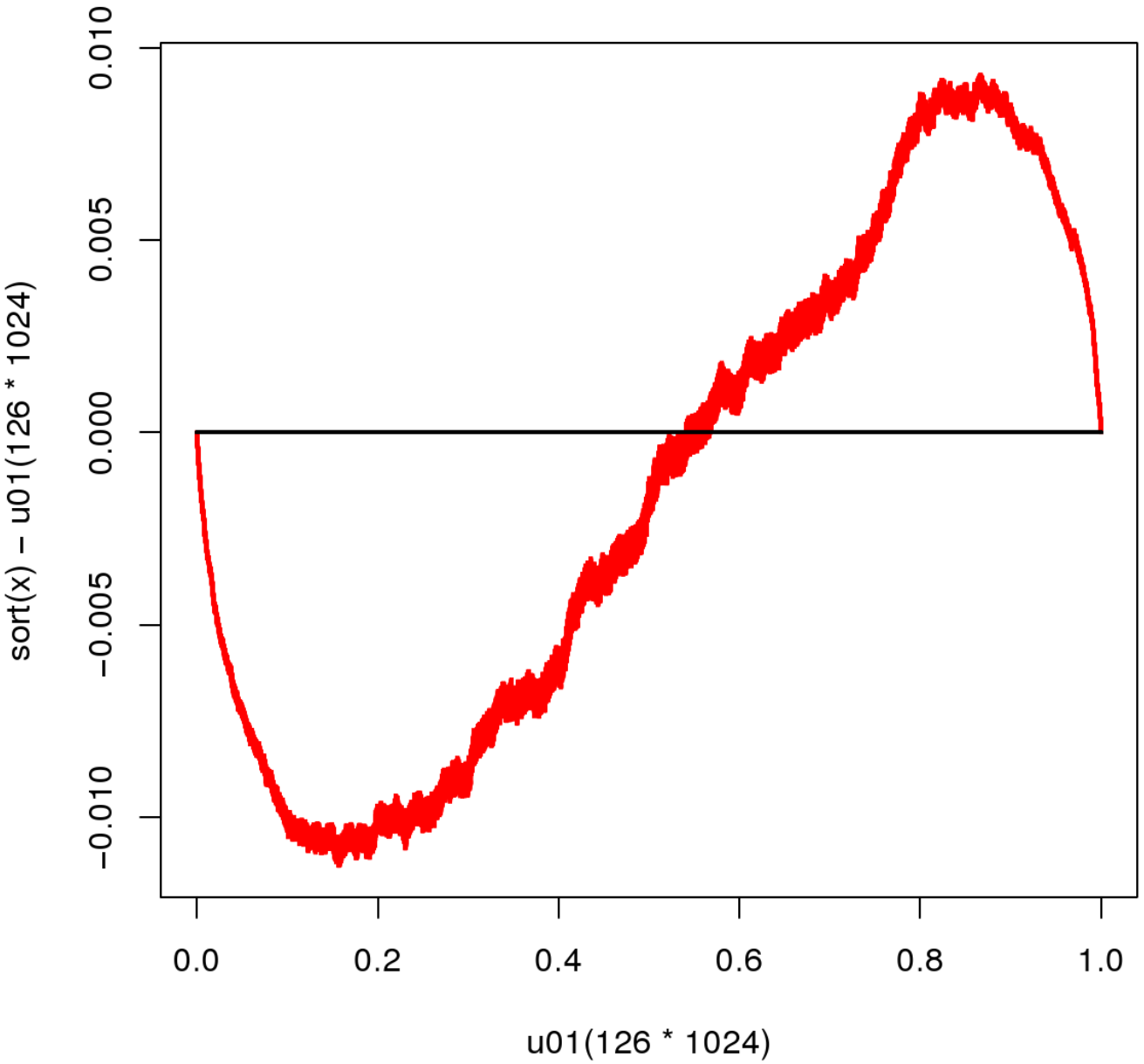
TestU01 0.6.1 pseudoDIEHARD results

- Used TestU01 0.6.1 pseudoDIEHARD battery 256 times for each of 4 seeds.
- Tried a number of generators including
  - Mersenne Twister MT19937 (Matsumoto, Nishimura, 1998, 2002),
  - XORGENS XOR4096 (Brent, 2006, 2007).
- Tested sequence of resulting p-values in R using one-sample Kolmogorov-Smirnov test with alternative hypothesis: two-sided. Results:
  - MT19937: D $= \mathbf{0.011}$, p-value $= \mathbf{5.951 \times 10^{-14}}$,
  - XOR4096: D $= \mathbf{0.0113}$, p-value $= \mathbf{1.221 \times 10^{-14}}$.

TestU01 0.6.1 pseudoDIEHARD p-values (XOR4096)



TestU01 0.6.1 pseudoDIEHARD p-values (XOR4096)



Run test (1981 version)

- Based on (Levene, Wolfowitz, 1944; Wolfowitz, 1944)

For a sequence of $\mathbf{n}$ random numbers from $\mathbf{U[0,1]}$, if $\mathbf{r_1}$ to $\mathbf{r_5}$ are the numbers of "runs up" of length 1 through 5 and $\mathbf{r_6}$ is the number of runs up of length 6 or more, then (Knuth, 1981) states that for "large $\mathbf{n}$",

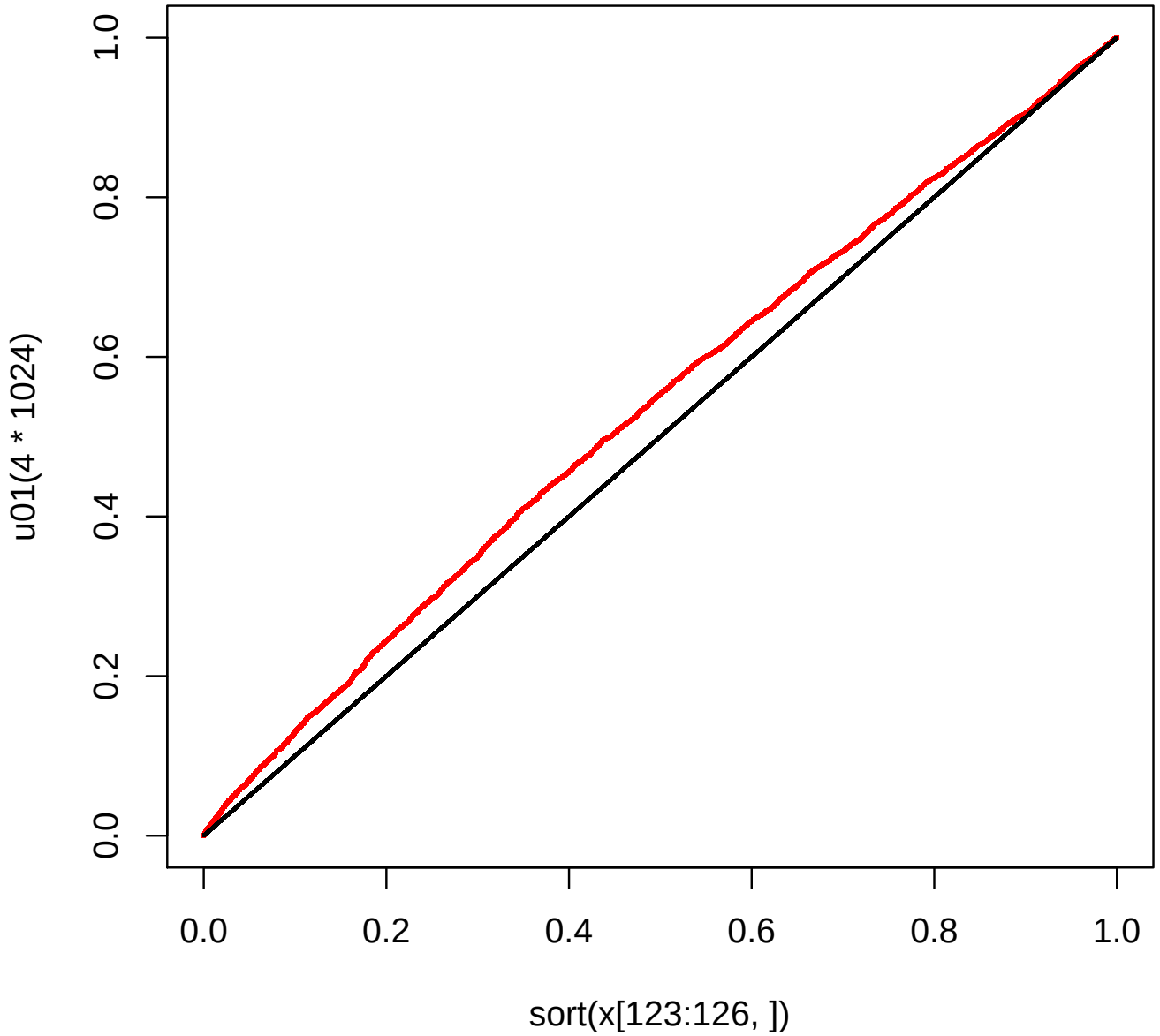
$$\mathbf{V} := \frac{\mathbf{1}}{\mathbf{n}}(\mathbf{r} - \mathbf{nb})^{\mathbf{T}}\mathbf{A}(\mathbf{r} - \mathbf{nb})$$

should approach a χ^2 distribution with 6 degrees of freedom, where $\mathbf{A}$ is a constant matrix and $\mathbf{b}$ is a constant vector.

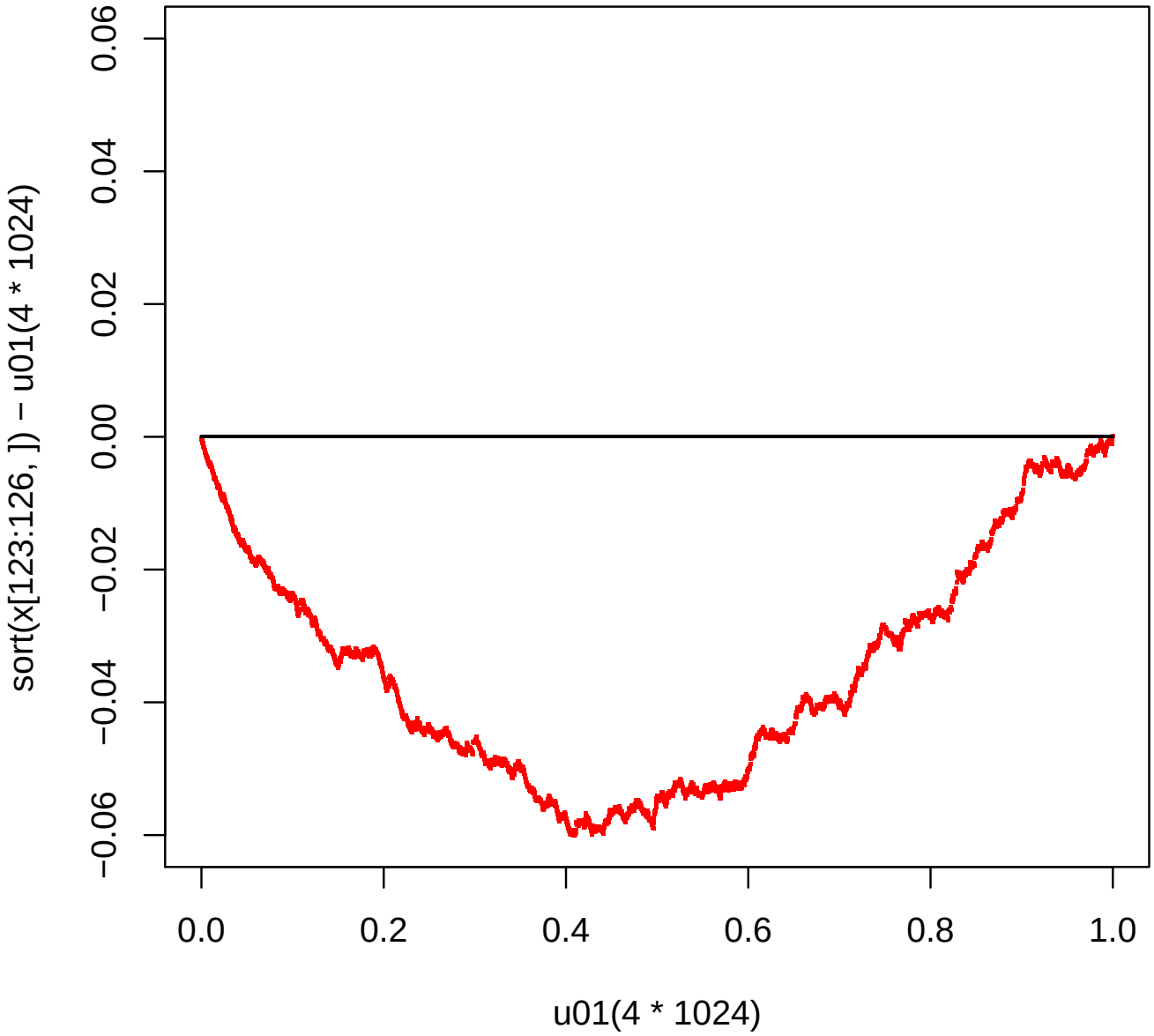
Run test in TestU01 0.6.1

- TestU01 0.6.1 implements Run test as per (Knuth 1981).
- pseudoDIEHARD includes 4 Run tests, 2 up, 2 down, each using $\mathbf{n = 10000}$ numbers.
- pseudoDIEHARD 256 $\times$ 4 seeds results in:
  - MT19937: D $= \mathbf{0.0415}$, p-value $= \mathbf{1.542 \times 10^{-06}}$,
  - XOR4096: D $= \mathbf{0.0602}$, p-value $= \mathbf{2.534 \times 10^{-13}}$.

Run test 0.6.1 p-values (XOR4096)



Run test 0.6.1 p-values (XOR4096)



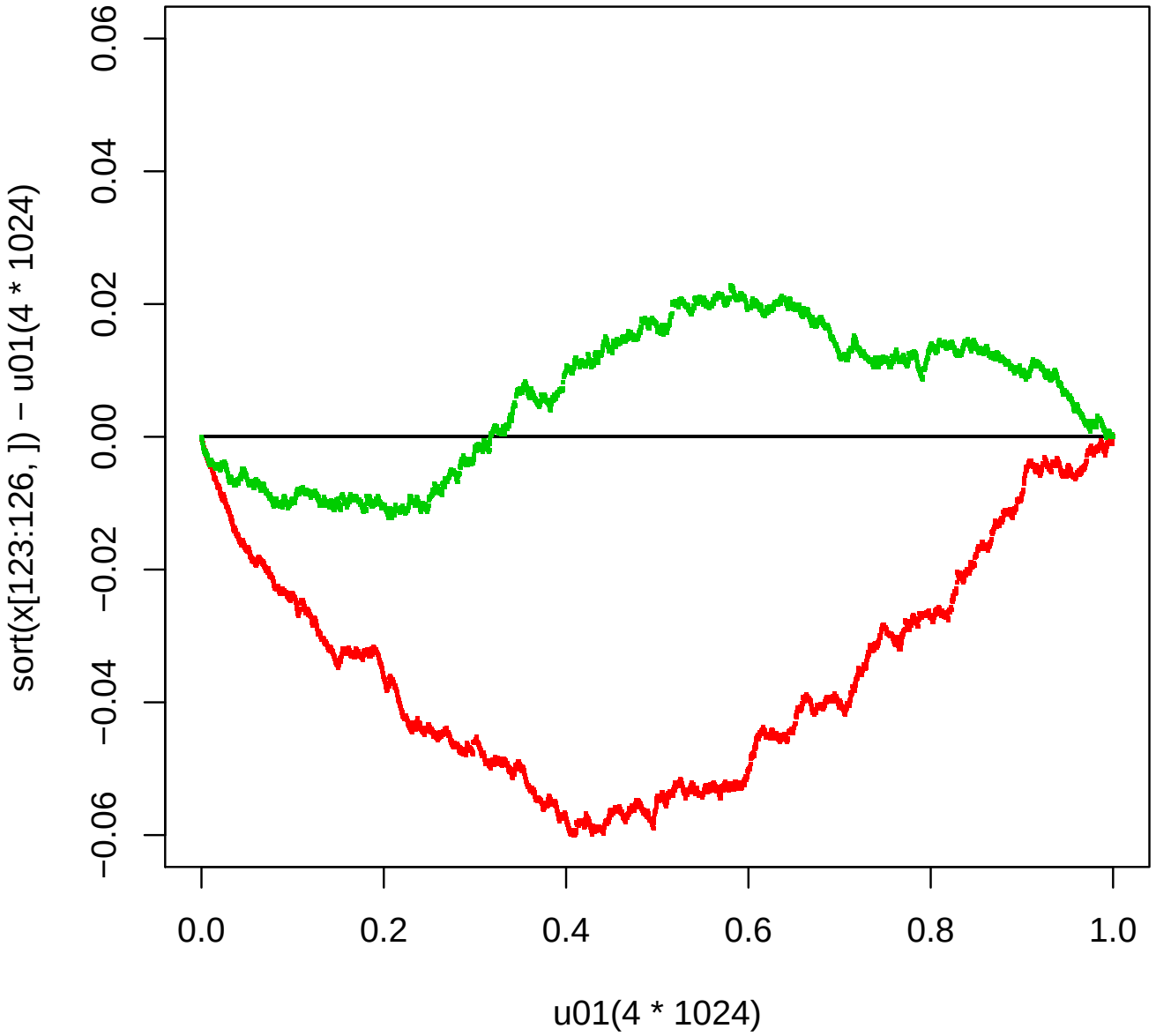
Run test (1998 version) in TestU01 1.2.1

In (Knuth, 1998) $\mathbf{V}$ is

$$\mathbf{V} := \frac{1}{\mathbf{n-6}}(\mathbf{r}-\mathbf{nb})^{\mathsf{T}}\mathbf{A}(\mathbf{r}-\mathbf{nb}).$$

- TestU01 1.2.1 implements Run test as per (Knuth 1998), with a more accurate $\mathbf{A}$.
- pseudoDIEHARD 256 $\times$ 4 seeds results in:
  - MT19937: D $= \mathbf{0.0142}$, p-value $= \mathbf{0.3790}$,
  - XOR4096: D $= \mathbf{0.0229}$, p-value $= \mathbf{0.02718}$.

Fixed Run test p-values (XOR4096)



Overlapping serial ("monkey") tests (1993 version)

- Of 126 p-values generated by pseudoDIEHARD, 82 come from three overlapping serial tests (Marsaglia, Zaman, 1993):
- OPSO: 23, OQSO: 28, DNA: 31.

These use an alphabet of size α, form a string of length $\mathbf{n = 2^{21}}$ by taking $\mathbf{n \times \log_2 \alpha}$ bits from a PRNG, and examine the $\mathbf{n - t + 1}$ overlapping words of length $\mathbf{t}$. Number of missing words should be normal with mean μ and variance σ^2:

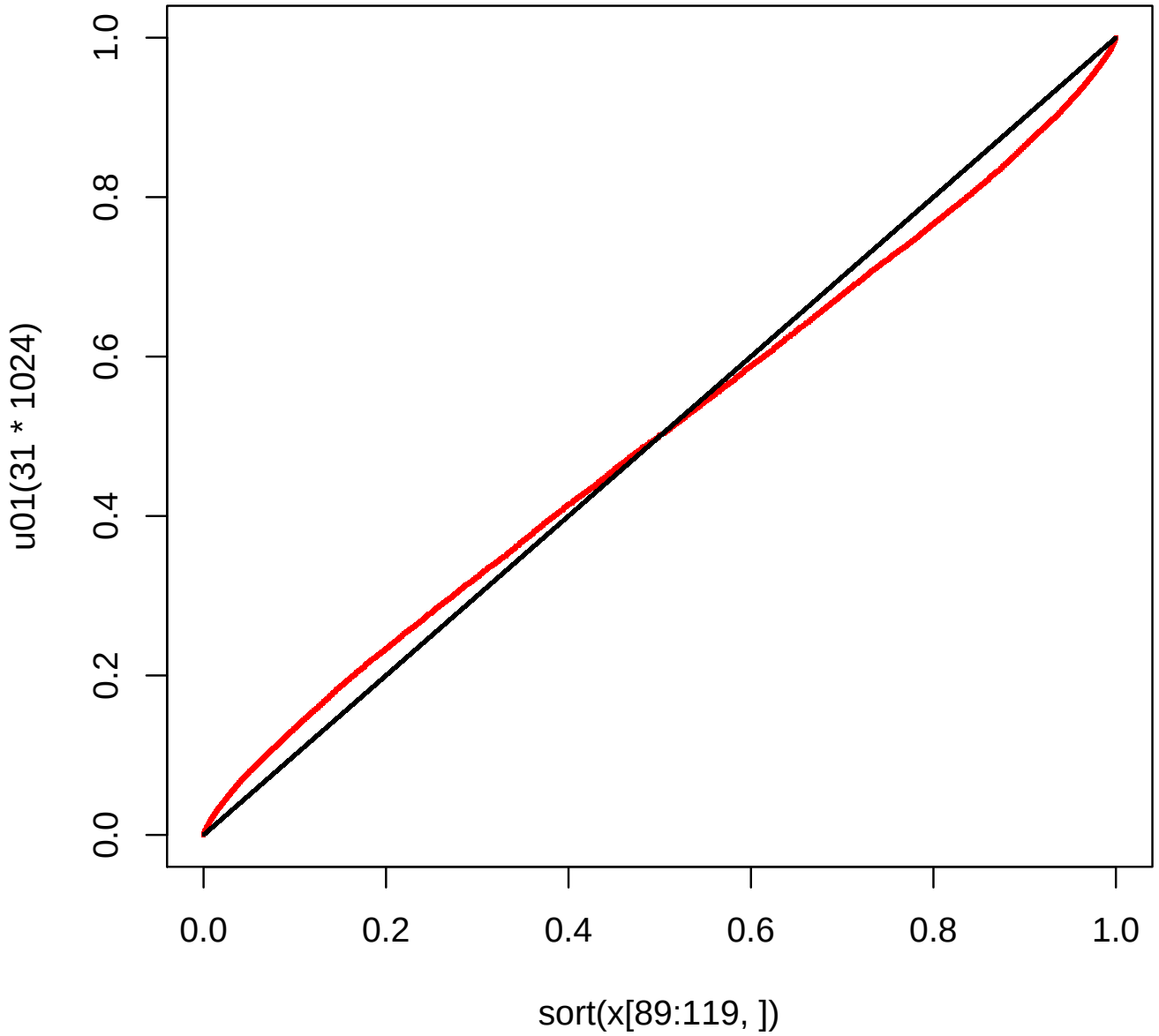
- OPSO: $\alpha = \mathbf{2^{10}}$, $\mathbf{t = 2}$: $\mu = \mathbf{141909.4653}$, $\sigma = \mathbf{290.27}$.
- OQSO: $\alpha = \mathbf{2^5}$, $\mathbf{t = 4}$: $\mu = \mathbf{141909.4737}$, $\sigma = \mathbf{290}$.
- DNA: $\alpha = \mathbf{4}$, $\mathbf{t = 10}$: $\mu = \mathbf{141910.5378}$, $\sigma = \mathbf{290}$.

Overlapping serial tests in TestU01 0.6.1

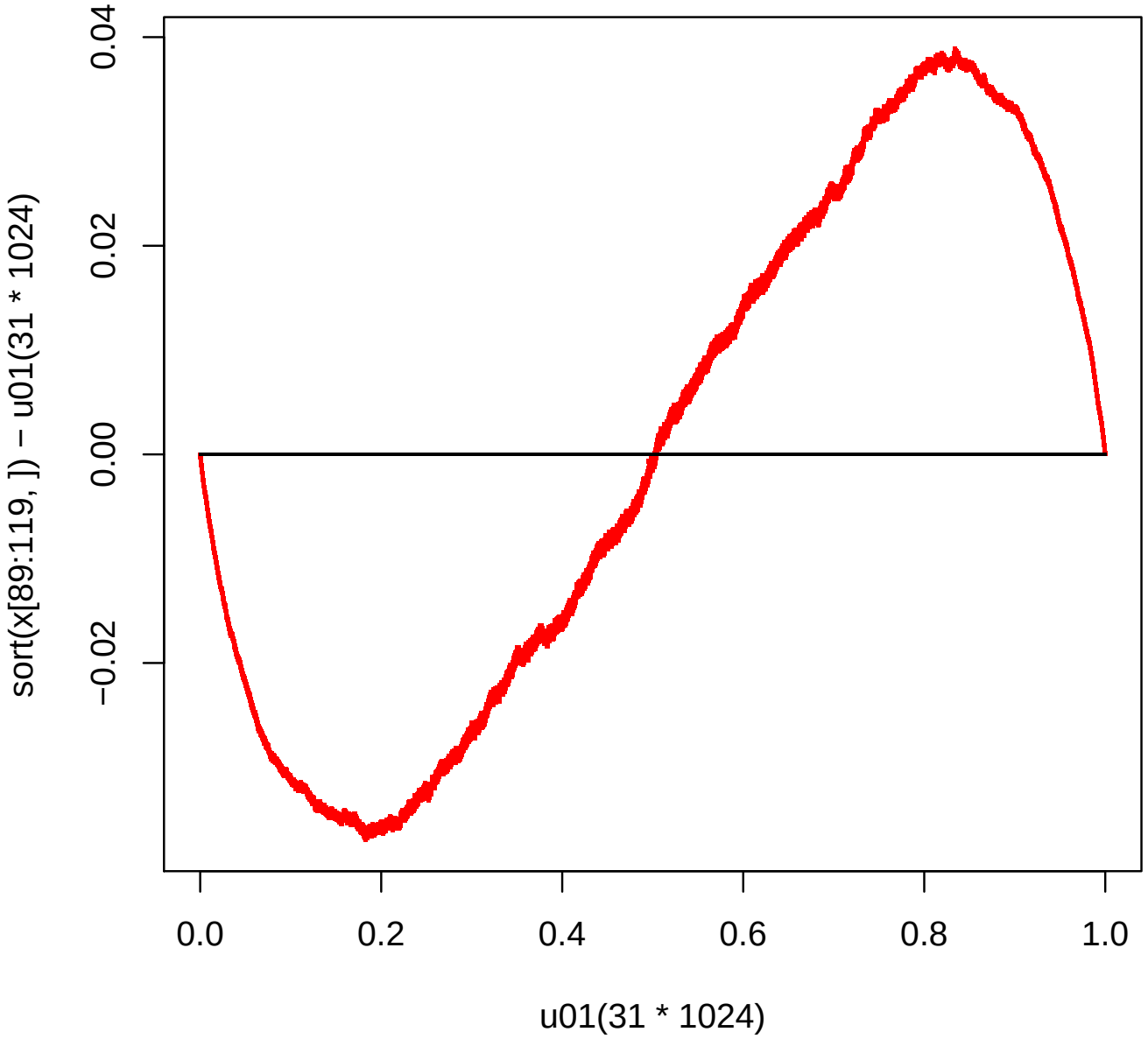
- TestU01 0.6.1 has $\sigma = 2^{10}\sqrt{e^{-2} - 3e^{-4}} \simeq 290.3331$.
- pseudoDIEHARD 256 $\times$ 4 seeds results in:

		D	p-value
OPSO	MT19937	**0.0073**	**0.1680**
	XOR4096	**0.0051**	**0.569**
OQSO	MT19937	**0.0061**	**0.24**
	XOR4096	**0.0085**	**0.03311**
DNA	MT19937	**0.0374**	$< 2.2 \times 10^{-16}$
	XOR4096	**0.0389**	$< 2.2 \times 10^{-16}$

DNA Test 0.61 p-values (XOR4096)



DNA Test 0.61 p-values (XOR4096)



Overlapping serial tests (1995 version)

(Marsaglia, 1995) has the revised values:

- OPSO: $\mu = \mathbf{141909.60}, \quad \sigma = \mathbf{290.46}.$
- OQSO: $\mu = \mathbf{141909.4737}, \ \sigma = \mathbf{295}.$
- DNA: $\mu = \mathbf{141910.5378}, \ \sigma = \mathbf{339}.$

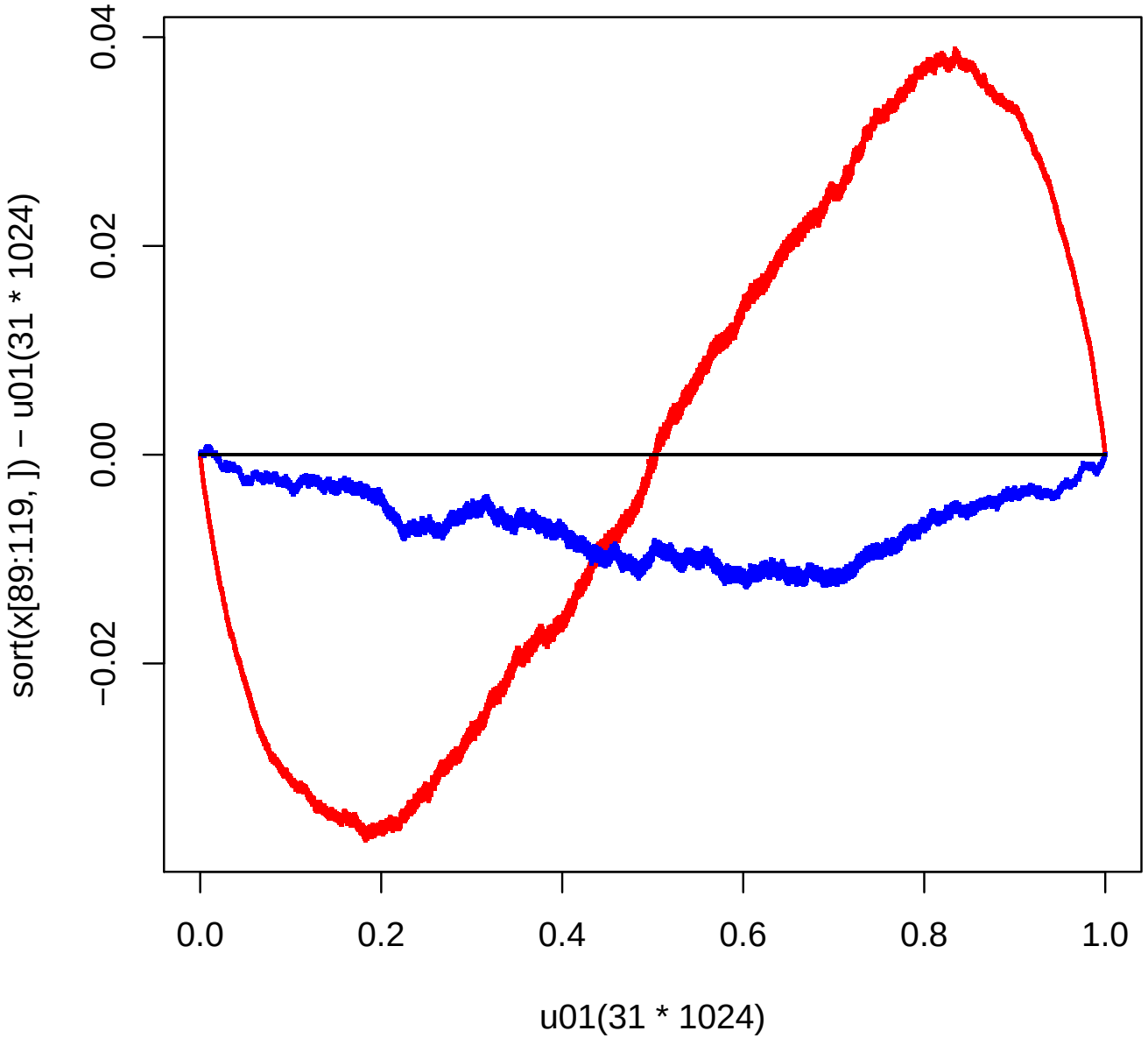
Overlapping serial tests in TestU01 1.2.1

TestU01 1.2.1 uses

- For OPSO: $\sigma = 2^{10}\sqrt{e^{-\lambda} - (1+\lambda)e^{-\lambda}} \simeq 290.3332$, where $\lambda := (2^{21} - 1)/2^{10} \simeq 2.0$ (Rukhin, 2002);
- (Marsaglia, 1995) values of σ for OQSO and DNA.
- pseudoDIEHARD 256 $\times$ 4 seeds results in:

		D	p-value
OPSO	MT19937	**0.0066**	**0.2498**
	XOR4096	**0.0066**	**0.2613**
OQSO	MT19937	**0.0093**	**0.0141**
	XOR4096	**0.0109**	**0.0020921**
DNA	MT19937	**0.0109**	**0.001052**
	XOR4096	**0.0127**	**6.802×10^{-5}**

Almost fixed DNA Test p-values (XOR4096)



What went wrong?

OPSO, OQSO and DNA tests in TestU01 use words on a cycle of length $\mathbf{n}$ rather than a string, giving $\mathbf{n}$ not $\mathbf{n - t + 1}$ words of length $\mathbf{t}$. Using (Edlin, Zeilberger, 2000) and (Rivals, Rahmann, 2003) the corresponding $\boldsymbol{\mu}$ is:

- OPSO: $\mathbf{141909.19461972381.}$
- OQSO: $\mathbf{141909.19452590772.}$
- DNA: $\mathbf{141909.18458308319.}$

Corresponding $\boldsymbol{\sigma}$ is not yet known.
TestU01 0.6.1 uses $\boldsymbol{\mu = 2^{20}e^{-2} \simeq 141909.329955}$.
TestU01 1.2.1 (incorrectly) changes $\boldsymbol{\mu}$ to

- OPSO: $\mathbf{141910.329955.}$
- OQSO: $\mathbf{141912.329955.}$
- DNA: $\mathbf{141918.329955.}$

Value of $\boldsymbol{\lambda}$ used in OPSO in 1.2.1 is also wrong and should be $\mathbf{2}$.

Overlapping serial tests (2008 version)

Values calculated using (Noonan, Zeilberger, 1999), (Rivals, Rahmann, 2003) and (Rahmann, Rivals, 2003):

- OPSO: $\mu = \mathbf{141909.3299550069}$, $\sigma = \mathbf{290.4622634038}$.
- OQSO: $\mu = \mathbf{141909.6005321316}$, $\sigma = \mathbf{294.6558723658}$.
- DNA: $\mu = \mathbf{141910.4026047629}$, $\sigma = \mathbf{337.2901506904}$.

- Calculation of σ for OPSO uses 6 generating functions;
- OQSO uses 55; DNA uses 4592.

Fixed overlapping serial tests in TestU01 use a string of length $\mathbf{n}$ and the values above.

Exact variance of missing words in strings

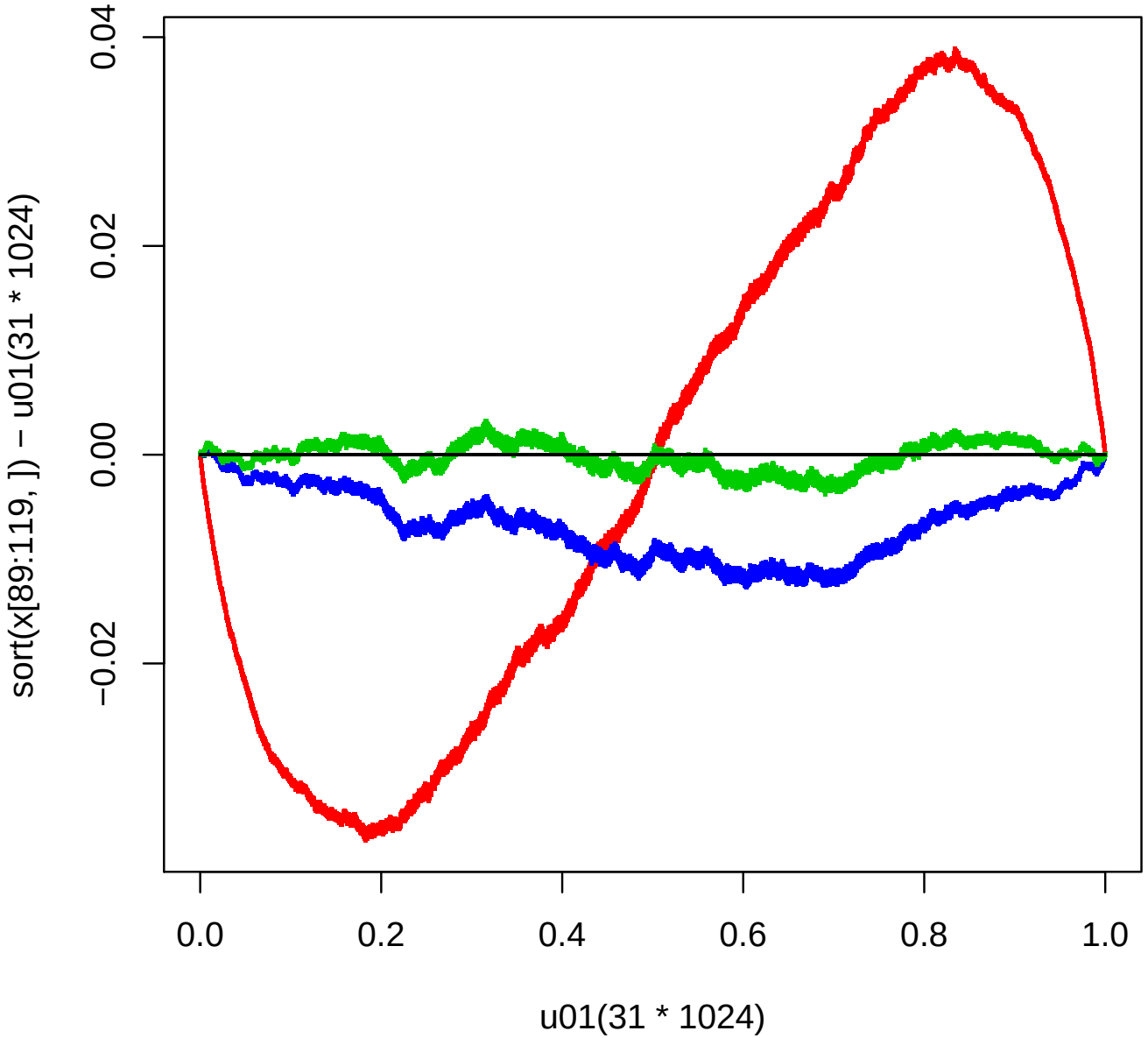
- OQSO: "I don't know, and doubt that I ever will know, the true variance. There are just too many kinds of pairs of 4-letter words to undertake finding all the necessary generating functions." (Marsaglia, 1995)
- DNA: "It appears a formidable task to find the exact variance for the DNA test." (Marsaglia, 1995)
- General: "Characterize and efficiently enumerate $\mathbf{2} \times \mathbf{2}$, and more generally, $\mathbf{k} \times \mathbf{k}$ matrices of correlation vectors between $\mathbf{k}$ pairwise different [words], and find the number of such matrices. Compute the number of $\mathbf{k}$-tuples of words that share a given correlation matrix." (Rahmann, Rivals, 2003)

Results of fixed overlapping serial tests

- pseudoDIEHARD 256 $\times$ 4 seeds results in:

		D	p-value
OPSO	MT19937	**0.0076**	**0.134**
	XOR4096	**0.0058**	**0.4157**
OQSO	MT19937	**0.006**	**0.2456**
	XOR4096	**0.008**	**0.05186**
DNA	MT19937	**0.0034**	**0.8589**
	XOR4096	**0.0038**	**0.7527**

Fixed DNA Test p-values (XOR4096)

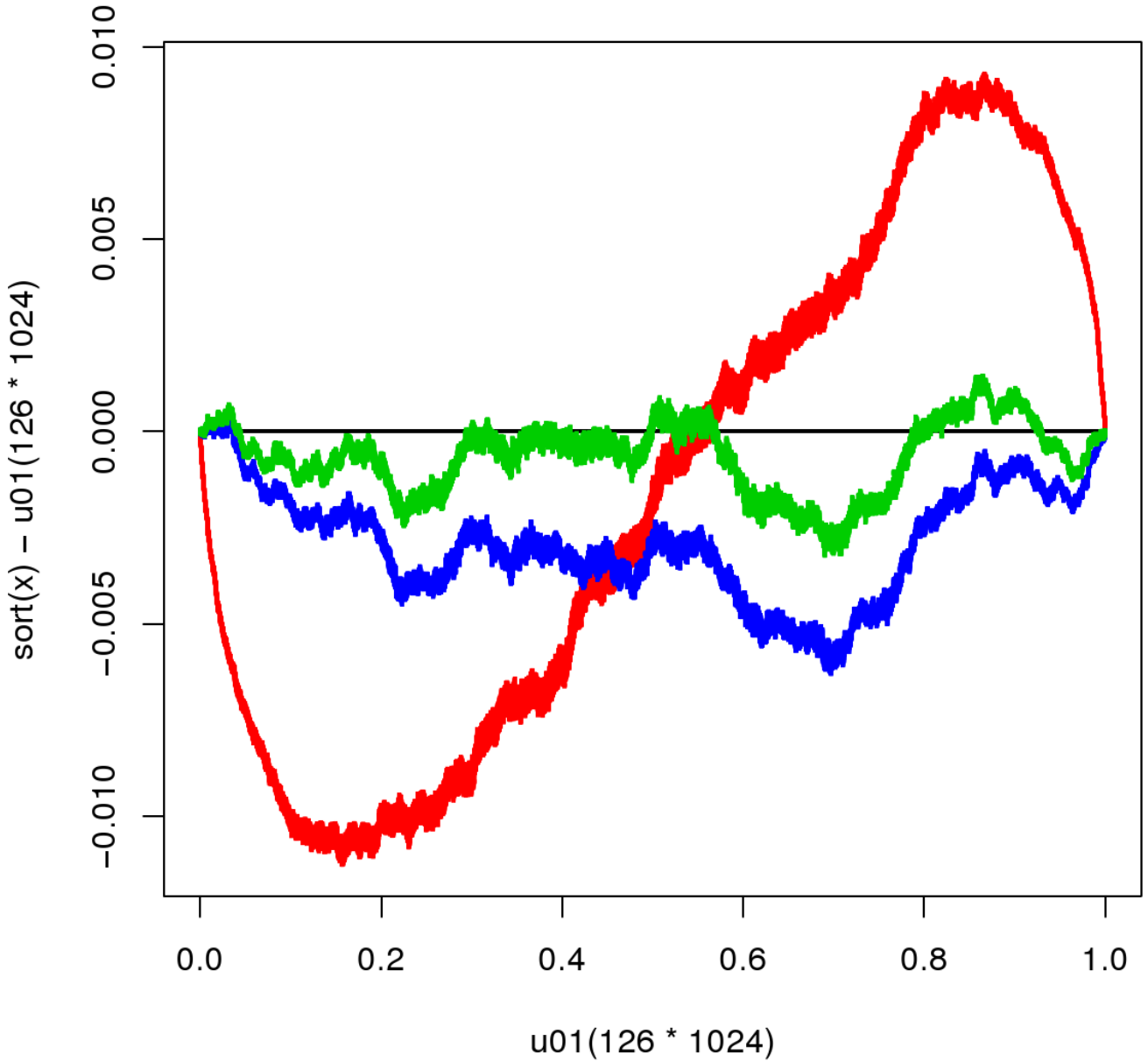


pseudoDIEHARD results

- pseudoDIEHARD 256 $\times$ 4 seeds results in:

Version		D	p-value
0.6.1	MT19937	**0.011**	$\mathbf{5.951 \times 10^{-14}}$
	XOR4096	**0.0113**	$\mathbf{1.221 \times 10^{-14}}$
1.2.1	MT19937	**0.0056**	**0.0006376**
	XOR4096	**0.0063**	$\mathbf{7.01 \times 10^{-5}}$
Fixed	MT19937	**0.0025**	**0.3982**
	XOR4096	**0.0032**	**0.1352**

pseudoDIEHARD p-values (XOR4096)



Acknowledgements

- Richard Simard, University of Montreal
- Sacha van Albada, Peter Drysdale, et al. at Complex Systems, School of Physics, University of Sydney
- Art Owen, Stanford University
- Jörg Arndt, Sylvain Foret, John Maindonald, Judy-anne Osborn, et al. at Mathematical Sciences Institute, Australian National University