Reproducibility in computational science: a case study: Randomness of the digits of Pi

David H. Bailey* Jonathan M. Borwein† Richard Brent‡

Mohsen Reisi Ardali§

January 3, 2016

Abstract

Mathematical research is undergoing a transformation from a mostly theoretical enterprise to one that involves a significant amount of experimentation. Indeed, *computational and experimental mathematics* is now a full-fledged discipline with mathematics, and the larger field of *computational science* is now taking its place as an experimental discipline on a par with traditional experimental fields. In this new realm, *reproducibility* comes to the forefront as an essential part of the computational research enterprise, and establishing procedures to ensure and facilitate reproducibility is now a central focus of researchers in the field.

In this study, we describe our attempts to reproduce the results of a recently published paper by Reinhard Ganz, who concluded that the decimal expansion of π is not statistically random, based on an analysis of several trillion decimal digits provided by Yee and Kondo. While we are able to reproduce the specific findings of Ganz, additional statistical analysis leads us to reject his overall conclusion.

1 Introduction

Mathematics is undergoing a transformation from a mostly theoretical enterprise to one that involves a significant amount of experimentation. Indeed, *computational and experimental mathematics* is now a full-fledged discipline with mathematics. The larger field of *computational science*, which spans many different disciplines ranging from physics and engineering to the social sciences, is now taking its place as an experimental discipline on a par with traditional experimental fields. But concomitant with its increased stature as an experimental discipline, computational science has had to come to grips with the need to foster greater

*Lawrence Berkeley National Laboratory (retired), Berkeley, CA 94720, and University of California, Davis, Davis, CA 95616, USA. E-mail: david@davidhbailey.com.

†CARMA, University of Newcastle, Callaghan, NSW 2308, Australia. E-mail: jon.borwein@gmail.com.

‡Mathematical Sciences Institute, Australian National University, Canberra, ACT 2614, Australia. E-mail: Richard.Brent@anu.edu.au.

§CARMA, University of Newcastle, Callaghan, NSW 2308, Australia. E-mail: mohsen.reisi@uon.edu.au.

reproducibility in its research findings. Indeed, in most fields it is entirely appropriate to perform an experiment that is merely intended to reproduce another recent finding.

1.1 Reproducible computational science

Most experimental disciplines have long instituted both formal and informal guidelines and procedures for facilitating and ensuring reproducibility. Prospective physicists and chemists have been taught to keep a detailed logbook of their experimental work, including such minutiae as exact details of the experimental setup, the source of samples, a minute-by-minute (if necessary) log of everything that happens, data collection and analysis and more. Social scientists and medical scientists, troubled with apparent instances of "effects" being seen in data that later proved illusory, have instituted strict double-blind procedures and other methodologies to improve reproducibility.

In spite of these efforts, though, several fields have been stung in recent years by widely publicized instances of reproducibility failures. For example, in 2012 a pharmaceutical researcher at Amgen reported that he and his colleagues were unable to replicate 47 of 53 "landmark papers" about the incidence of cancer, even with some assistance from the original researchers [10, 17]. Similarly, a 2015 attempt by Brian Nosek and several colleagues at the University of Virgina failed to reproduce 60 out of 100 published studies in the field of psychology [14]. Several prominent journals have issued new guidelines for publication studies, including a disclosure and proper analysis of the statistical methods used [21]. Numerous writers, e.g., [18], have raised concern about the use and misuse of statistics in scientific research.

In the wake of these developments, research studies in the field of computational science have come under additional scrutiny. Unfortunately, due in part to the rapid growth of the field, computational science has not fostered a culture of reproducibility. In many cases, there are concerns that the results are not numerically reproducible — numerical sensitivities in a code that are minor in most applications suddenly become major problems when the code (and applications using the code) are scaled up in size to run on highly parallel supercomputers. At a deeper level, published papers in the field typically do not include full details of computational environment and the specific algorithms employed, nor do they, in most cases, offer the actual source code used and output files on a publicly accessible repository. And very few published papers attempt specifically to reproduce the findings of a previous computational study.

A 2013 workshop at the Institute for Computational and Experimental Research in Mathematics (ICERM) in Providence, Rhode Island, USA specifically addressed the issue of reproducibility in computational and experimental mathematics. The workshop report recommended broad changes in the field to promote a "culture change that will integrate computational reproducibility into the research process." The findings included these recommendations for publication of research results [19, 20]:

1. A precise statement of assertions to be made in the paper, together with a statement of the computational approach, and why it constitutes a rigorous test of the hypothesized assertions.
2. Complete statements of, or references to, the algorithms employed.
3. Details of software (both research and commercial software) used in the computation.

4. Details of the test environment, including hardware, system software and the number of processors utilized.
5. Details of data reduction and statistical analysis methods.
6. Discussion of the adequacy of numeric precision and grid resolution.
7. A valid summary of experimental results.
8. Verification and validation tests performed by the author(s).
9. Availability of computer code, input data and output data, with some reasonable level of documentation.

Of course these requests are not all equally relevant to each piece of research.

1.2 Reproducibility study: randomness of the digits of π

With this background, the present authors sought to perform a "case study" by analyzing the reproducibility of a recent paper in the computational mathematics field. The study that we selected is by Reinhard Ganz, who in a 2014 article [16], analyzed a large dataset of the decimal digits of π to see whether or not the decimal digits of π are statistically random. We were led to attempt to replicate his study for three reasons.

- Several colleagues had contacted various of the current authors us to ask our opinions of the study.
- The conclusion was striking given that, while π is clearly not random, we knew of no compelling prior evidence of statistical non-randomness the digits of π, but see [15].
- We were keen, in light of the discussion in Section 1.1, to attempt a full hearted replication/reproduction of a mathematical experiment.

Two of the present authors have previously published studies on the question of whether π or other constants are *b-normal*. By b-normal, for some positive integer b, we mean that every base-b digit appears, in the limit, with frequency $1/b$, that every pair of base-b digits appears, in the limit, with frequency $1/b^2$, and so on, with every m-long string of base-b digits appearing, in the limit, with frequency $1/b^m$.

As we will discuss in more detail in the next section, whether or not π (or almost any other irrational or transcendental constant that appears in the mathematical literature) is b-normal, either for $b = 2$ or $b = 10$, let alone for all bases simultaneously, remains an outstanding open problem in mathematics. Since the dawn of computing, mathematicians have wondered whether or not this is true; indeed, to gain insight on this question is a leading motivation for the many recorded computations of π and other mathematical constants. The question of whether π, in particular, is b-normal to any or all bases is among the oldest unanswered questions of mathematics.

Some of the previous research on this question by the present authors and colleagues include the following. In [6], one of the present authors and Richard Crandall (deceased December 2012) established that the question of whether constants such as $\pi, \log 2$ and others are 2-normal in each case reduces to a conjecture about the behavior of a closely related pseudorandom number generator. If the associated pseudorandom number generator can be proved to be uniform in $(0, 1)$, then normality is established. In [7], the same authors

proved normality for an uncountably infinite class of transcendental constants; for example, the authors proved that the constant (now known as Stoneham's constant)

$$\alpha_{2,3} = \sum_{n=0}^{\infty} \frac{1}{3^n 2^{3^n}} \; = \; 0.54188368\ldots \tag{1}$$

is provably 2-normal (this result was proven more simply, using an ergodic theory argument, in [9]). In a separate study by two of the present authors [5], it was shown that $\alpha_{2,3}$ is provably *not* 6-normal.

In a 2012 article [2], two of the present authors, together with Cristian Calude, Michael Dinneen, Monica Dumitrescu and Alex Yee, demonstrated that based on a Poisson process model and the first four trillion hexadecimal digits of π, it was "extraordinarily unlikely" that π is not asymptotically 16-normal. In a 2013 article [1], two of the present authors, together with Francisco Aragon Artacho and Peter Borwein, analyzed the normality of π, Stoneham's constant and others using graphical and statistical techniques. These authors showed, for example, that while Stoneham's constant is provably 2-normal, it fails to satisfy a stronger condition of statistical randomness.

So it was with considerable interest that the present authors read a 2014 paper by Reinhard Ganz, published in the journal *Experimental Mathematics* [16]. Ganz, based on a dataset of the first several trillion decimals digits of π provided to him by Yee and Kondo, concluded that these digits are *not* statistically random, to a fairly impressive confidence level. If Ganz's result is correct, it could be landmark study in the field. At the very least, Ganz's result and methods well deserve to be reproduced by other researchers, and to be extended to larger datasets if possible.

1.2.1 The status of Pi

For three other recent articles on the number theoretic and computational status of π we refer the reader to [3, 12, 13].

2 Ganz's study and our present analysis

The purpose of Reinhard E. Ganz's paper [16] is to examine whether the decimal expansion of π is statistically random. In [16] a statistical test is introduced and employed to analyse the first 10^{13} digits of π. In what follows, we briefly describe the test and show that the results are replicable, and finally, we present our additional numerical experiments showing that the statistical test is not robust.

We note that while Ganz gives a fairly explicit description of his experiment, the current authors were not able to fully digest, and so attempt to replicate his methodology, without several communications between us and Ganz.

2.1 Ganz's original experiment

The null hypothesis introduced in Ganz's paper, which is referred to as $Null$, is that the decimal expansion of π behaves like a realization of a sequence of mutually independent, identically distributed (iid) random variables on $\{0, 1, ..., 9\}$. To test $Null$, the full set of digits are divided into 222 disjoint blocks, each containing 5×10^9 consecutive non-overlapping

9-tuples of digits so that $222 \times 5 \times 10^9 \times 9 = 9.99 \times 10^{12}$. These are then converted to 222 binary values, $X_i, i = 1, ..., 222$, according to whether the block contained more than 2.3×10^6 9-tuples with at least five identical consecutive digits (for example, 157777704).

$Null^1$ is then defined as that *the binary stream X_i is Bernoulli process with* $Pr(X_i = 0) = Pr(X_i = 1) = 0.5$. Afterwards, lengths $L_j(X_i)$ (L_j for short) are determined for consecutive runs of X_i , where a j-th run is defined as the succession of L_j identical values $X_i = 0$ or $X_i = 1$. For example, if 00100011 is a consecutive run of X_i, then $L_1 = 2$, $L_2 = 1$, $L_3 = 3$, and $L_4 = 2$. $Null^2$ is then defined as that the random variable L_j has the geometric distribution with scale parameter $q = 0.5$.

To show that distribution L_j is independent of the value and position of X_i, the author introduces the difference $d_k(L_j, L_{j+1}) = L_j - L_{j+1}$ (or d_k for short) and the differences $D_k(L_j, L_{j+\lfloor r/2 \rfloor}) = L_j - L_{j+\lfloor r/2 \rfloor}$ (or D_k for short), where $1 \leq j \leq \lfloor r/2 \rfloor$, r is the number of L_j, and $\lfloor . \rfloor$ denotes the floor function. Finally, $Null^3$ is defined to be that d_k and D_k follow the symmetric Laplace distribution with parameter $p = 1 - q = 0.5$.

In a nutshell, the author of [16] believes that the randomness of digits ($Null$) implies that X_i is an iid random variable ($Null^1$) which implies that the random variable L_j has the geometric distribution with scale parameter $q = 0.5$ ($Null^2$) which eventually implies $Null^3$. Therefore, he believes that the rejection of $Null^3$ would lead to rejection of $Null$ as well.

2.2 Ganz's experiment replicated

Ganz uses $\theta = \frac{T}{\sqrt{V}}$ (for details on T and V refer to [16]) as a standard normalized statistic for testing $H_0 : \theta = 0$ versus $H_1 : \theta \neq 0$, so that the test rejects $Null^3$ at the significant level $\alpha = 0.001$ if $|\theta| > t_{df;0.9995}$, where df denotes the degrees of freedom of the t distribution. It is claimed that use of $t_{df;0.9995}$, for the corresponding t-distribution, instead of $z_{0.9995}$ (z_δ denotes the δth quantile of the standard normal distribution) provides additional protection of the α level, with only minor reduction in the power of the test.

Our values for Ganz's test statistics are $\theta(d_k) = 3.933$ and $\theta(D_k) = 3.496$ ($> t_{52;0.9995} = 3.488$, where $n = 54$). These numbers are identical to those reported in the paper — suggesting that the null hypothesis is indeed to be rejected.

2.3 Ganz's experiment extended

Our results so far show the correctness of those presented in the paper, but do not assess the statistical test itself. To do so, we decided to run the test by varying the number of blocks with the same set of digits so as to validate the robustness of the statistical test used in [16].

This time, we divide the 10^{13} digits into $213, 217, 226, 231, 236, 241, 246, 252$ or 258 blocks respectively. The reason for this selection is that our digits are saved in text files each containing 10^8 digits, and it is easier to store an integer number of text files in each block for the sake of simplicity in coding. In addition, the number of digits in each block must be a multiple of 9, since we are counting the 9-tuples. Hence, $nB = \lfloor \frac{10^{13}}{9 \times 10^8 \times p} \rfloor$, where nB is the number of blocks and $\beta \in \{43, 44, \ldots, 52\}$. When $\beta = 50$, then $nB = 222$ which is the case presented in the original paper.

2.4 Our Findings

The θ values presented in Table 1 show that only when $nB = 222$ is the null hypothesis rejected. This reveals the lack of robustness of the statistical test in [16]. Figures 1 and 2 illustrate respectively the distribution of d_k and of D_k for all variations of nB. These figures show that the distributions are fairly symmetric in most cases while in few cases like $nB = 222$ appear skew. The overall results convince us that the statistical test is not robust, and therefore, cannot be used to confirm the author's claim: "the decimal expansion of π is not statistically random".

Heedful of Carl Sagan's observation that extraordinary claims require extraordinary evidence, we wonder why Ganz did not attempt confirmatory experiments. Once the digits and statistical tests were prepared, the cost (in time) of our additional experiments was not great. Rather than doing this, Ganz compared his experiment to that of a quantum random sequence and concluded:

> This result of our test is consistent with recent results in [15, p. 10]. In that study, bit strings extracted from the binary expansion of π as well as bit strings produced by PRNGs from software packages Mathematica 6 and Maple 11 were compared to sequences of quantum random bits obtained by different physical experiments; a set of randomness tests inspired by algorithmic information theory revealed statistically significant differences between the distributions of the data obtained from π and those obtained from quantum systems.

nB	n	$\theta(d_k)$	$\theta(D_k)$	$t_{(n-2,0.9995)}$	Rejected, $\theta(d_k)$	Rejected, $\theta(D_k)$
213	58	1.212	1.405	3.473	×	×
217	53	1.970	1.565	3.492	×	×
222	**54**	**3.933**	**3.496**	**3.488**	✓	✓
226	61	0.239	1.348	3.463	×	×
231	57	2.985	1.417	3.476	×	×
236	64	3.440	2.393	3.454	×	×
241	64	0.371	1.515	3.454	×	×
246	64	-0.171	3.068	3.454	×	×
252	63	0.728	1.739	3.457	×	×
258	68	1.538	-1.015	3.444	×	×

Table 1: θ values for our suite of ten experiments

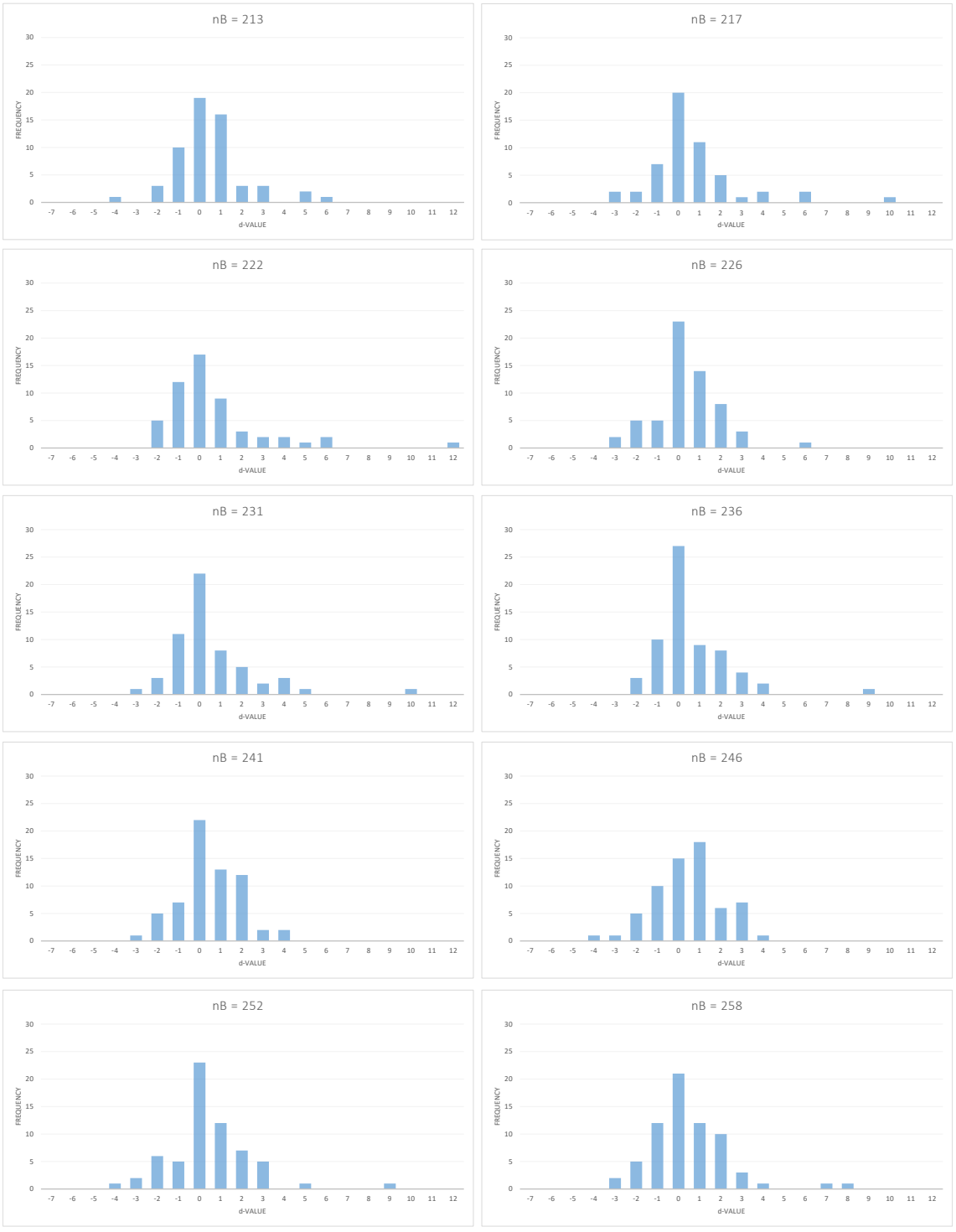

Figure 1: Distribution of d_k for our suite of experiments.

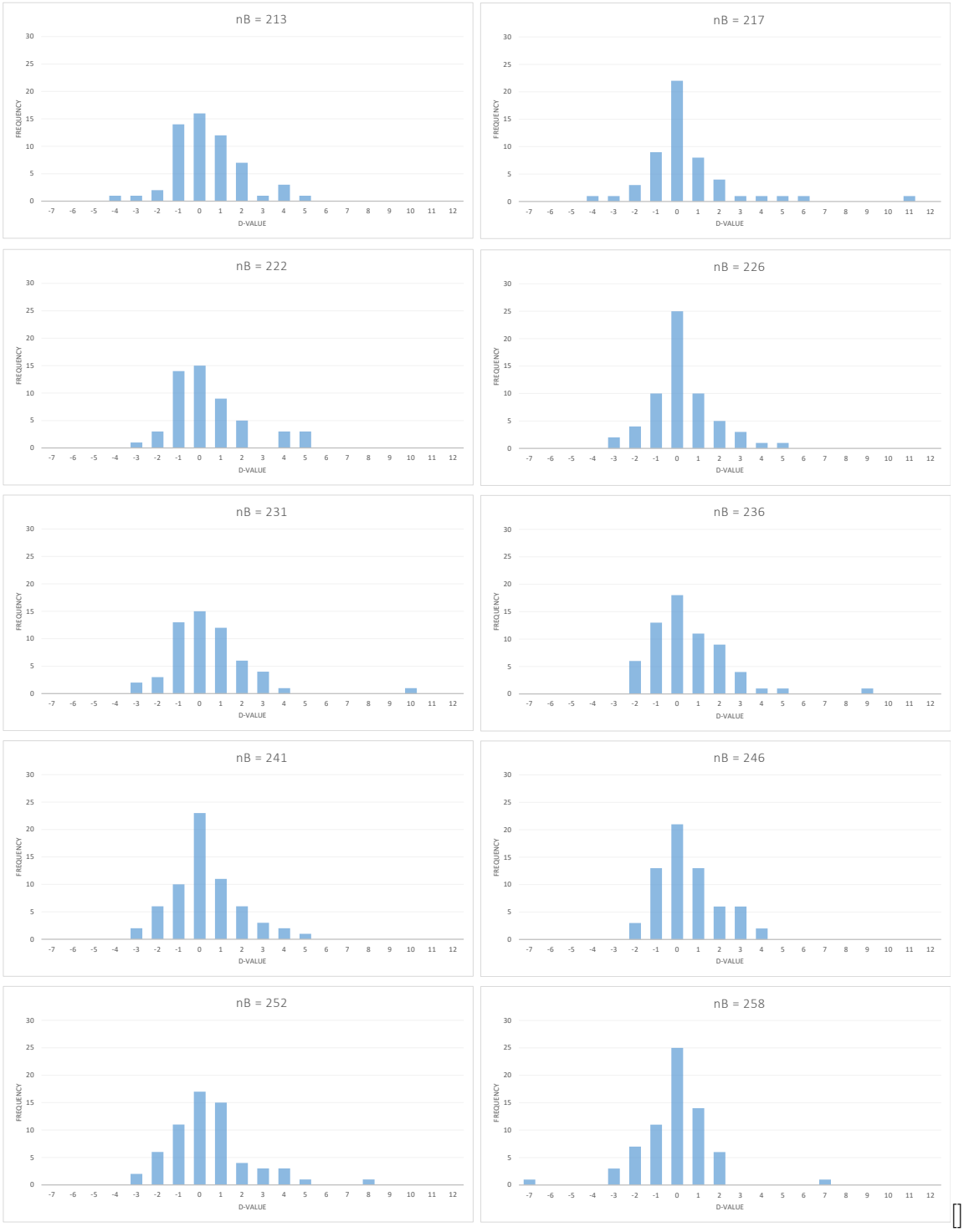

Figure 2: Distribution of D_k for our suite of experiments.

3 Conclusions and recommendations

Now that we have completed this case study, let us revisit the recommendations we recorded in Section 1.1.

1. A precise statement of assertions to be made in the paper, together with a statement of the computational approach, and *why it constitutes a rigorous test of the hypothesized assertions*.
   - In our view, the highlighted phrase is missing from [16]. Without such a discussion it is hard to dismiss the possibility that the use of '222' was the result of cherry-picking the data or, more charitably, a lucky fluke.
   - Equally importantly, as argued in [22], a statistical correlation or a t-test without some substantial argument is far from convincing.
2. Complete statements of, or references to, the algorithms employed.
   - As noted, greater precision — or pseudocode — regarding Ganz's original experiment would have made our job easier.
3. Details of software (both research and commercial software) used in the computation.
   - Ganz gives limited information indicating that both *Maple 11* and *Mathematica 6* were used. Since neither package is open source and both versions are quite old, this is only partially reassuring.
   - In our case, all codes were written in C++ within a Linux environment to read the digits from zipped text files, to count the 9-tuples, and to extract values for X_i, L_j, d_k, D_k, and θ. The frequencies of d_k and D_k and their charts were provided using Excel.
4. Details of the test environment, including hardware, system software and the number of processors utilized.
   - The coding and computations were done within a Linux environment on CARMA[1] server computer with 189 GB of memory and 24 CPUs each Intel(R) Xeon(R) 3.47 GHz. For each nB we utilised two CPUs at the same time each occupied 0.9% of the memory for about 100 hours.
5. Details of data reduction and statistical analysis methods.
   - This was generally appropriate.
6. Discussion of the adequacy of numeric precision and grid resolution.
   - This is not applicable.
7. A valid summary of experimental results.
   - This was generally appropriate in [16].
8. Verification and validation tests performed by the author(s).
   - We feel that both Ganz's tests and ours have been adequately described above.
9. Availability of computer code, input data and output data, with some reasonable level of documentation.

[1]Computer-Assisted Research Mathematics and its Applicationa, at the University of Newcastle, Callaghan, NSW 2308, Australia

- In our case, we have provided a zipped version of our code, along with a 'read me' file at `https://www.carma.newcastle.edu.au/jon/repro-pi.zip`. One recommendation in [19, 20] was to use a standard community code repository, such as Github, for this purpose. We are assessing this situation and may also place it in a similar place.

We remark, in passing, that the entirety of our replication efforts appear to be comparable to, if not greater, than the original study. This appears to be the case in other replication efforts in the field as well [19, 20].

In summary, while the current authors could reproduce the main finding in [16], their subsequent more extensive tests failed to replicate the phenomenon. In light of this, and the failure of the author of [16] to provide a substantive justification for his test that would not be shared by our generalizations, we conclude that the evidence given in [16] that "the decimal expansion of π is not statistically random" is unconvincing.

Acknowledgement The authors wish to thank Alex Yee for providing the digits needed for our work and Reinhard E. Ganz for his willingness to respond to our queries – and so to participate in this attempt at replication. They also thank Dr. David Allingham for his technical assistance with the computations.

References

[1] Francisco J. Aragon Artacho, David H. Bailey, Jonathan M. Borwein and Peter B. Borwein, "Walking on real numbers," *Mathematical Intelligencer*, **35** (2013), 42–60.

[2] David H. Bailey, Jonathan M. Borwein, Cristian S. Calude, Michael J. Dinneen, Monica Dumitrescu and Alex Yee, "An empirical approach to the normality of pi," *Experimental Mathematics*, **21** (2012), 375–384.

[3] David H. Bailey, Jonathan M. Borwein, Andrew Mattingly, and Glenn Wightwick, "The computation of previously inaccessible digits of π^2 and Catalan's constant." *Notices of the AMS*. **60**(7) (2013), 844–854.

[4] David H. Bailey, Jonathan M. Borwein and Victoria Stodden, "Facilitating reproducibility in scientific computing: Principles and practice" in Harald Atmanspacher and Sabine Maasen, eds, *Reproducibility: Principles, Problems, Practices*, John Wiley and Sons, New York, to appear, 2016.

[5] David H. Bailey and Jonathan M. Borwein, "Nonnormality of Stoneham constants," *Ramanujan Journal*, **29** (2012), 409–422; `DOI10.1007/s11139-012-9417-3`.

[6] David H. Bailey and Richard E. Crandall,"On the random character of fundamental constant expansions," *Experimental Mathematics*, **10** (Jun 2001), 175–190.

[7] David H. Bailey and Richard E. Crandall, "Random generators and normal numbers," *Experimental Mathematics*, **11** (2002), 527–546.

[8] David H. Bailey, Robert F. Lucas and Samuel W. Williams, ed., *Performance Tuning of Scientific Applications*, CRC Press, Boca Raton, FL, 2011.

[9] David H. Bailey and Michal Misiurewicz, "A strong hot spot theorem," *Proceedings of the American Mathematical Society*, **134** (2006), 2495–2501.

[10] C. Glenn Begley and Less M. Ellis, "Drug development: Raise standards for preclinical cancer research," *Nature*, **483** (29 Mar 2012), 531–533. See http://www.nature.com/nature/journal/v483/n7391/full/483531a.html.

[11] Jonathan M. Borwein and David H. Bailey, *Mathematics by Experiment: Plausible Reasoning in the 21st Century* A.K. Peters Ltd, 2004, ISBN: 1-56881-136-5. Second edition, 2008.

[12] Jonathan M. Borwein, "The life of Pi," Extended and updated version of "La vita di pi greco," volume 2 of *Mathematics and Culture, La matematica: Problemi e teoremi*, Guilio Einaudi Editori, Turino, Italian, 2008) (French, in press), p. 532–561 of *From Alexandria, Through Baghdad: Surveys and Studies in the Ancient Greek and Medieval Islamic Mathematical Sciences in Honor of J.L. Berggren*, Sidoli, Nathan; Van Brummelen, Glen (Eds.) Springer-Verlag 2014.

[13] Jonathan M. Borwein and Scott T. Chapman, "I prefer pi: A brief history and anthology of articles in the American Mathematical Monthly," *American Mathematical Monthly*, **122** (March 2015), 195–216.

[14] Benedict Carey, "Many psychology findings not as strong as claimed, study says," *New York Times*, 27 Aug 2015. http://www.nytimes.com/2015/08/28/science/many-social-science-findings-not-as-strong-as-claimed-study-says.html.

[15] C. S. Calude, M. J. Dinneen, M. Dumitrescu, and K. Svozil. "Experimental evidence of quantum randomness incomputability," *Physical Review A* **82** (2010), 1–8.

[16] Reinhard E. Ganz, "The decimal expansion of π is not statistically random," *Experimental Mathematics*, **23** (2014), 99–104.

[17] George Johnson, "New truths that only one can see," *New York Times*, 20 Jan 2014. See http://www.nytimes.com/2014/01/21/science/new-truths-that-only-one-can-see.html.

[18] Garry Smith, *Standard Deviations: Flawed Assumptions, Tortured Data, and Other Ways to Lie With Statistics,* Overlook/Duckworth, New York, 2014.

[19] Victoria Stodden, David H. Bailey, Jonathan M. Borwein, Randall J. LeVeque, William Rider and William Stein, "Setting the default to reproducible: Reproducibility in computational and experimental mathematics," manuscript, 2 Feb 2013. http://www.davidhbailey.com/dhbpapers/icerm-report.pdf.

[20] Victoria Stodden, Jonathan Borwein and David H. Bailey, "'Setting the default to reproducible' in computational science research," *SIAM News*, 46 (June 2013), 4–6. http://sinews.siam.org/DetailsPage/tabid/607/ArticleID/351/Setting-the-Default-to-Reproducible-in-Computational-Science-Research.aspx.

[21] Richard Van Noorden, "Major scientific journal joins push to screen statistics in papers it publishes," *Scientific American*, 6 Jul 2014. See http://www.scientificamerican.com/article/major-scientific-journal-joins-push-to-screen-statistics-in-papers-it-publishes1/.

[22] S. T. Ziliak and D. N. McCloskey, *The Cult of Statistical Significance: How the Standard Error Costs Us Jobs, Justice, and Lives (Economics, Cognition, and Society)*, University of Michigan Press (2008).