Refutation of claims such as “Pi is less random than we thought”.

George Marsaglia
Professor Emeritus
Florida State University

1 Introduction

We begin with an allegory, a fanciful tale in which you plan to assess the randomness of π. You use the digits of π to simulate the result of flipping a coin—0,1,2,3,4 for heads, 5,6,7,8,9 for tails. After 3000 flips, the number of heads should be close to 1500, with standard deviation $\sigma = \sqrt{npq} = \sqrt{3000/4} = 27.386$. You decide to grade the randomness of π according to how many σ's the heads count is from its expected value of 1500:

$No.of\sigma's$	$0-.1$	$.1-.5$	$.5-1$	$1-2$	$2-3$	$3-4$	>4
$Grade$	$A+$	A	$A-$	B	C	D	E

After 3000 ‘flips’, π produces 1454 heads, (true), some 1.679 σ's from the mean of 1500, so you give π a grade of B. You then use the popular congruential RNG $x_n = 69069x_{n-1} + 12345 \bmod 2^{32}$, with $x_0 = 22222$, designating heads whenever $x_n/2.^{32} < .5$, to find that the first 3000 ‘flips’ produce 1501 heads, (true), only .0365 σ's from the mean. Thus this congruential RNG rates an $A+$. You get a few more grades for π, getting, in effect, its GPA, compare it with similar, restricted testing on other sources, then, through your public relations office, you announce to the world that the randomness of π is not as good as that from certain other sources.

Sounds silly? Perhaps, but something much like this happened recently, leading to worldwide coverage to the effect that the randomness of π is not as good as that from other sources. The extent of worldwide coverage for the claim that π gets a poorer grade for randomness can be judged from over 400 internet sources, a few of which are:

http://news.uns.purdue.edu/html4ever/2005/050426.Fischbach.pi.html
http://www.physorg.com/news3886.html
http://www.theallineed.com/science/05050906.htm
http://science.slashdot.org/article.pl?sid=05/05/01/1759240&tid=228&tid=14
http://clearlyexplained.com/news/nature/2005/apr/1N2704_2005.html
http://www.news.uns.purdue.edu/html4ever/%202005/050426.Fischbach.pi.html
http://digg.com/technology/Pi:_Less_Random_Than_We_Thought
http://forevergeek.com/news/pi_is_not_the_best_random_number_generator.php
http://www.memagazine.org/backissues/june05/departments/computing/computing.html
http://www.alpheratz.net/murison/weblog/pi_seems_a_good_random_number_generator_but_not_always_the_best/
http://use.perl.org/~n1vux/journal/24408

The only difference is that instead of heads/tails, the digits of π were taken thirty at a time to form three uniform variates U_1, U_2, U_3 in [0,1), ten digits for each U. Then, after many such triples, the distance of the theoretical expected value from what amounts to the average of $(U_1-U_2)(U_2-U_3)$, measured in sigmas, was used to provide a few letter grades for π, then compared with those coming from several standard RNGs.

2 The claim

The above grading scheme for randomness was put forward by Tu and Fischbach,[2], from Purdue's Dept. of Physics, and used to compare the randomness of π with that from other sources.

Their unusual method of letter-grading randomness and making conclusions based on a few grades is contrary to usual methods for testing randomness, and the conclusions are not at all supported by application of their method to, for example, the first 960 million digits of π. Far more extensive testing of π,e,$\sqrt{2}$, as well as that of expansions of k/p for large primes p, are in reference [1]. The results of such extensive testing showed that the digits of π do as well in tests of randomness as do any other source. That and the discussion below suggest that the analysis in Tu and Fischbach,[2] is woefully inadequate.

Purdue is mentioned here because apparently its PR office initiated the claim that led to worldwide coverage, so it can share in the glory—or the shame.

Tu and Fischbach's analysis is based on what they call the Geometric Random Inner Product (GRIP) test, which uses the average value of the inner product of the differences between three random points in an n-cube.

For example, if (x_1, y_1, z_1), (x_2, y_2, z_2), (x_3, y_3, z_3) are three random points in the 0-centered 3-cube, the GRIP test is concerned with the average value of the inner product of the differences, $(x_1 - x_2, y_1 - y_2, z_1 - z_2)$ and $(x_2 - x_3, y_2 - y_3, z_2 - z_3)$, that is, the average value of

$$(x_1 - x_2)(x_2 - x_3) + (y_1 - y_2)(y_2 - y_3) + (z_1 - z_2)(z_2 - z_3).$$

(Most researchers interested in testing for randomness would be concerned with the *distribution* of that quantity, not merely its mean, but more on that later.)

Points in the 0-centered cube are obtained by converting uniform [0,1) variates U into uniform [-1,1] variates V by means of $V = 2U - 1$. Thus 90 successive digits of π, floated ten at a time, would provide nine uniform [0,1) variates $U_1, \ldots, U_9$, which converted to uniform [-1,1] variates $V_1, \ldots, V_9$, yield the three random points in the cube: (V_1, V_2, V_3), (V_4, V_5, V_6) and (V_5, V_6, V_7), with inner product of the differences:

$$(V_1 - V_4)(V_4 - V_7) + (V_2 - V_5)(V_5 - V_8) + (V_3 - V_6)(V_6 - V_9).$$

This immediately raises another question: If three uniform [0,1) variates U_1, U_2, U_3 are used to produce uniform variates $V_i = 2U_i - 1$, and the product $(V_2 - V_1)(V_3 - V_2)$ is formed, why go through such unnecessary conversions? Surely

$$(V_2 - V_1)(V_3 - V_2) = 4(U_2 - U_1)(U_3 - U_2),$$

and one might as well deal with the simpler expression $(U_2 - U_1)(U_2 - U_3)$, for which underlying theory is less messy. And because $(U_2 - U_1)(U_3 - U_2)$ is negative more often than positive, I prefer to use the form $(U_2 - U_1)(U_2 - U_3)$, and will derive its distribution below, ignoring the multiplier 4.

Tu and Fischbach report results of their GRIP test for 3- and 6-cubes, each of which amounts to the equivalent sum of three or six expressions of the type $(U_2 - U_1)(U_2 - U_3)$, and since the average of a set of averages is still an average, the whole thing amounts to finding the average of an unnecessarily complicated version of what is basically a collection of $(U_2 - U_1)(U_2 - U_3)$'s.

Thus, in spite of the allusions to geometric inner products and their complexities, the GRIP test can be reduced to getting the average value of $(U_2 - U_1)(U_2 - U_3)$ for a succession of triples of purported uniform [0,1) random variables U_1, U_2, U_3 produced by the source. And we will see that π produces quite satisfactory values for not only the mean of $(U_2 - U_1)(U_2 - U_3)$, but for the distribution, a much more difficult assessment.

Of course, if that average value is shown to differ significantly from its expected value—when the uniform variates are produced by ten successive digits from π—then there would be no need for further tests, and worldwide attention would be justified. But the Tu and Fischbach analysis does not come close to being persuasive.

. It is, in my opinion, nonsense to make conclusions based on a few sigma values, unless one or more happen to be extreme. And that raises the question of how to measure extremeness. The distribution of the number of σ's from the mean for the average of a large number of observations is generally taken to be standard normal. Since most of us do not carry more than a few points of the standard normal distribution in our heads, but we all carry the uniform distribution, it is customary to convert a sigma value to a p-value, having a uniform distribution in the interval $0 < p < 1$.

If s is the distance, in standard deviations, of the sample mean from its expected value, then $p = \int_{-\infty}^{s} e^{-x^2/2}\,dx/\sqrt{2\pi}$ should be uniform in [0,1). If a simulation is repeated enough times to give numerous p-values, it is customary to apply an Anderson-Darling or Kolmogorov test for uniformity of those p-values. If there are thousands of p-values, the classical $(\mathrm{OBS} - \mathrm{EXP})^2/\mathrm{EXP}$ chi-square test might be applied to the frequencies for intervals such as $p < .01, .01 < p < .02, .02 < p < .03$.

And p-values around .95 or even .98 are not unusual among a few random selections from the uniform distribution. It is p-values such as .99999 that lead to rejecting certain sources of purported randomness, not GPAs based on grading a few p-values according to their distance from .5.

Suppose we apply the Tu and Fischbach approach in its more efficient form, finding the average value of $(U_2 - U_1)(U_2 - U_3)$ for successive triples U_1, U_2, U_3 produced by π, each U obtained by floating ten digits. But rather than doing this a few times and using the $A+, A, A-, B, C, D, E$ grade system for the σ values that result, we instead convert each such value to a p-value and see if the resulting p-values can be considered the realization of set of independent uniform [0,1) variates.

Here is the result for 32 such tests, each finding the average value of $(U_2 - U_1)(U_2 - U_3)$ for one million triples U_1, U_2, U_3 formed from 30 consecutive digits of π. The average value v is distanced $s = (v - \frac{1}{12})/\sigma$ from its expected value of $\frac{1}{12}$, with $\sigma = \sqrt{\frac{19}{720}/1000000} = .1624465724\text{e-}3$, and each s is converted to a p-value by means of $p = \int_{-\infty}^{s} e^{-x^2/2}\,dx/\sqrt{2\pi}$.

p-values from 32 "Purdue" tests on 960 million digits of π.

.69691	.54896	.85987	.16279	.35252	.15646	.55476	.63224
.90927	.02556	.14053	.40180	.01492	.45216	.63008	.85735
.44559	.12281	.81445	.06868	.06561	.76278	.78647	.04832
.56177	.86149	.97462	.75406	.73671	.89946	.02035	.61826,

Anderson-Darling test for uniformity on those 32 values: p=.598548.

Those 32 p-values are quite consistent with having come uniformly from the interval [0,1). There is no need for them to be near the .5 that would provide a high GPA under the Purdue grading system. Indeed, they should show the variation, some near 0, (.01492), some near 1, (.97462), that we expect from uniform samples from [0,1), and in fact we would be hard pressed to create a more satisfactory set of 32 p-values than this set we get from the first 960 million digits of π.

3 A more rigorous application of the GRIP test

The GRIP test is essentially concerned with the mean value of $Z = (U_2 - U_1)(U_2 - U_3)$ for uniform [0,1) variates U_1, U_2, U_3. The standard way to assess the apparent randomness of a sequence of independent realizations $Z_1, Z_2, Z_3, \ldots$ of a random variable Z is to use the distribution function of Z: $F(z) = \Pr(Z < z)$, and test the sequence $F(Z_1), F(Z_2), F(Z_3), \ldots$ for uniformity.

We may find $F(z)$ as follows:
Let S_1, S_2, S_3, S_4 be the spacings induced by sorting the three uniform variates U_1, U_2, U_3. A little thought will lead to the conclusion that the random variable $Z = (U_2 - U_1)(U_2 - U_3)$ is distributed as $-S_1S_2$ with probability 1/3, or as $S_1(S_1 + S_2)$ with probability 2/3. Thus, if $F(z) = \Pr(Z < z)$ then F is a piecewise function on $-\frac{1}{4} < z < 1$, a mixture of 1/3 of the distribution of $-S_1S_2$ and 2/3 of the distribution of $S_1(S_1 + S_2)$.

To find $\Pr(S_1S_2 > a)$, we integrate the density of S_1 times the conditional probability that $S_2 > a/x$, given that $S_1 = x$:

$$\Pr(S_1S_2 > a) = \int_0^1 3(1-x)^2 \Pr(S_2 > \frac{a}{x} \,|\, S_2 = x)\, dx.$$

Given $S_1 = x$, the conditional variate $S_2/(1-x)$ behaves as s, where s is the first of three spacings induced by two uniform variates, and $\Pr(s > b) = (1-b)^2$. Thus

$$\Pr(S_2 > \frac{a}{x} \,|\, S_1 = x) = \Pr(\frac{s}{1-x} > \frac{a}{x}) = (1 - \frac{(1-x)a}{x})^2,$$

and we integrate over x's for which $0 < (1-x)a/x < \frac{1}{4}$, that is,

$$\frac{1-\sqrt{1-4x}}{2} < x < \frac{1-\sqrt{1+4x}}{2}.$$

.

This provides the distribution of the product of the two spacings S_1, S_2. For $0 < a < \frac{1}{4}$,

$$G(a) = \Pr(S_1S_2 < a) = 1 - \int_{(1-\sqrt{1-4a})/2}^{(1+\sqrt{1+4a})/2} 3(1-x)^2(1-(1-x)a/x)^2\, dx$$

$$G(a) = \Pr(S_1S_2 < a) = 1 - (1+8a)\sqrt{1-4a} + 6a\ln(\frac{1+\sqrt{1-4a}}{1-\sqrt{1-4a}}).$$

A similar procedure permits development of the distribution of $S_1(S_1 + S2)$,
say $H(a) = \Pr\left(S_1(S_1 + S_2) < a\right)$. We have, for $0 < a < 1$,

$$H(a) = 1 - \Pr\left(S_1(S_1+S_2) > a\right) = 1 - \int_0^1 3(1-x)^2 \Pr(S_2 > a/x - x \,|\, S1 = x)\, dx$$

$$= 1 - \left(\int_a^{\sqrt{a}} 3(1-x)^2(1-(a/x-x)/(1-x))^2\, dx + \int_{\sqrt{a}}^1 3(1-x)^2\, dx\right),$$

which reduces to

$$H(a) = 4a^{3/2} - 3a\ln(a) - 3a.$$

Since the random variable $Z = (U_2 - U_1)(U_2 - U_3)$ takes the value $-S_1S_2$ with probability 1/3, or the value $S_1(S_1+S_2)$ with probability 2/3, the distribution of Z, $F(z) = \Pr(Z < z)$, is a piecewise function on $-\frac{1}{4} < z < 1$:

$$F(z) = \Pr(Z < z) = \begin{cases} \frac{1}{3}\left((1-8z)\sqrt{1+4z} + 6z\ln(\frac{1+\sqrt{1+4z}}{1-\sqrt{1+4z}}\right) \text{ for } -\frac{1}{4} < z < 0 \\ \\ \frac{1}{3} + \frac{2}{3}\left(4z^{3/2} - 3z\ln(z) - 3z\right) \text{ for } 0 \le z \le 1. \end{cases}$$

We then have what might be called the proper application of the test of Tu and Fischbach,[2] to the first 960 million digits of π: Test $F(Z_1), F(Z_2), F(Z_3), \ldots$ for uniformity, where the Z's are formed as $(U_2 - U_1)(U_2 - U_3)$, each triple U_1, U_2, U_3 coming from successive sets of 30 digits from the decimal expansion of π.

Since we will have one million values to test for uniformity, difficult to keep in a table, it seems better to keep counts for the number of times

$$0 \le F(Z) < .001,\ .001 \le F(Z) < .002, , \ldots, .999 \le F(Z) < 1,$$

The counts for each cell should average 1000 with variation measured by the usual

$$X = \sum(\text{OBS-EXP})^2/\text{EXP} \text{ and } p = \int_{-\infty}^{(X-n)/\sqrt{2n}} e^{-x^2/2}\,dx/\sqrt{2\pi},$$

because X should be a χ^2_{999} value with degrees of freedom $n = 999$ so large that $(X-n)/\sqrt{2n}$ should be standard normal.

p-values, tests on 960 million digits of π, stringent "Purdue" test.

.08128	.65829	.58260	.37389	.82416	.56824	.17004	.51861
.79397	.12175	.47818	.01786	.41497	.09200	.81099	.10526
.55498	.95764	.20014	.61680	.42587	.77222	.56740	.53158
.27139	.82007	.54886	.64193	.17467	.44898	.92642	.49688

Anderson-Darling test for uniformity on those 32 values: p=.309841.

As with the test on only the average value of $(U_2{-}U_1)(U_2{-}U_3)$, this test on $F\left((U_2 - U_1)(U_2 - U_3)\right)$ again supports the suitability of the expansion of π as a source of independent random digits.

But so do the much more extensive and difficult-to-pass tests in reference [1], the "difficult-to-pass" designation given because they make some RNGs consistently produce p-values such as .99999. Yet π sails through all of them.

References

[1] George Marsaglia, On the randomness of Pi and other decimal expansions,
InterStat, #5, October, 2005.

[2] Shu-Ju Tu and Ephraim Fischbach, A study on the randomness of π,
International Journal of Modern Physics C, **16**, No. 2, 281–294, 2005.