A (Probably) Exact Solution to the Birthday Problem

David Brink

July 2011

ABSTRACT. Given a year with $n \geq 1$ days, the *Birthday Problem* asks for the minimal number $\mathscr{X}(n)$ such that in a class of $\mathscr{X}(n)$ students, the probability of finding two students with the same birthday is at least 50 percent. We derive heuristically an exact formula for $\mathscr{X}(n)$ and argue that the probability that a counter-example to this formula exists is less than one in 45 billion. We then give a new derivation of the asymptotic expansion of Ramanujan's Q-function and note its curious resemblance to the formula for $\mathscr{X}(n)$.

1 Introduction

It is a surprising fact, apparently first noticed by Davenport around 1927, that in a class of 23 students, the probability of finding two students with the same birthday is more than 50 percent. This observation, and the multitude of related problems that it raises, have been discussed by many distinguished authors, including von Mises [19], Gamow [9, pp. 204–206], Littlewood [17, p. 18], Halmos [11, pp. 103–104], and Davenport [6, pp. 174–175]. The interested reader is referred to [22] for a historical account and to [12] to experience the problem experimentally.

Given a year with $n \geq 1$ days, the variant of the *Birthday Problem* studied here asks for the minimal number $\mathscr{X}(n)$ such that in a class of $\mathscr{X}(n)$ students, the probability of a birthday coincidence is at least 50 percent. In other words, $\mathscr{X}(n)$ is the minimal integer x such that

$$P(x) := \left(1 - \frac{1}{n}\right)\left(1 - \frac{2}{n}\right)\cdots\left(1 - \frac{x-1}{n}\right) \tag{1}$$

is less than or equal to $\frac{1}{2}$. Clearly, $\mathscr{X}(n)$ is at least 2 and at most $n+1$. The first 99 values of $\mathscr{X}(n)$, Sloane's A033810 [23], are given here:

n	1–2	3–5	6–9	10–16	17–23	24–32	33–42	43–54	55–68	69–82	83–99
$\mathscr{X}(n)$	2	3	4	5	6	7	8	9	10	11	12

The bounds

$$\sqrt{n + \tfrac{1}{4}} + \tfrac{1}{2} \leq \mathscr{X}(n) < \sqrt{2n \log 2 + \tfrac{1}{4}} + \tfrac{3}{2} \tag{2}$$

follow easily from $P(x) \geq 1 - \left(\frac{1}{n} + \frac{2}{n} + \cdots + \frac{x-1}{n}\right) = 1 - \frac{x^2-x}{2n}$ and $P(x) \leq \exp\left(-\frac{1}{n}\right)\exp\left(-\frac{2}{n}\right)\cdots \exp\left(-\frac{x-1}{n}\right) = \exp\left(-\frac{x^2-x}{n}\right)$.[1] Mathis [18] showed the better lower bound

$$\sqrt{2n\log 2 + \tfrac{1}{4}} - \tfrac{1}{2} < \mathscr{X}(n). \tag{3}$$

Ahmed and McIntosh [1] gave a short proof of the asymptotic approximation $\mathscr{X}(n) \sim \sqrt{2n\log 2}$.

In the following, we show

$$\frac{3 - 2\log 2}{6} < \mathscr{X}(n) - \sqrt{2n\log 2} \leq 9 - \sqrt{86\log 2}. \tag{4}$$

As will be seen, these bounds are optimal in the sense that $(3 - 2\log 2)/6 \approx 0.269$ is the infimum of the sequence $\mathscr{X}(n) - \sqrt{2n\log 2}$, while $9 - \sqrt{86\log 2} \approx 1.279$ is the maximum, taken for $n = 43$. Contrary to (2) and (3), the bounds (4) are sufficiently tight to give the exact value of $\mathscr{X}(n)$ in most cases, for example $\mathscr{X}(365) = 23$.

In general, it follows from (4) that $\mathscr{X}(n)$ always equals either $\lceil\sqrt{2n\log 2}\rceil$ or $\lceil\sqrt{2n\log 2}\rceil + 1$ where $\lceil x\rceil$ denotes the ceiling function. We prove that the formula

$$\mathscr{X}(n) = \left\lceil\sqrt{2n\log 2}\right\rceil \tag{5}$$

holds for a set of integers n with asymptotic density $(3 + 2\log 2)/6 \approx 0.731$, and moreover that

$$\mathscr{X}(n) = \left\lceil\sqrt{2n\log 2} + \frac{3 - 2\log 2}{6}\right\rceil \tag{6}$$

holds for "almost all" n, i.e., for a set of integers n with asymptotic density 1. Furthermore, we show that the formula

$$\mathscr{X}(n) = \left\lceil\sqrt{2n\log 2} + \frac{3 - 2\log 2}{6} + \frac{9 - 4(\log 2)^2}{72\sqrt{2n\log 2}} - \frac{2(\log 2)^2}{135n}\right\rceil \tag{7}$$

referred to in the title of this paper holds for all n up to 10^{18}, and we give a heuristic argument that the probability that a counter-example to this formula exists is less than one in 45 billion.

Finally, we use the method developed here to give a new derivation of the asymptotic expansion of Ramanujan's Q-function which has a curious resemblance to (7).

2 Power Sum Polynomials

It follows immediately from the definition that $\mathscr{X}(n)$ can be described as the minimal integer x such that

$$-\log P(x) \geq \log 2. \tag{8}$$

[1] Cf. [11, p. 104], but note that the inequality $P(x) < e^{-x^2/2n}$ given there is not correct.

By (1) and the Taylor expansion of $\log(1-z)$, we may write

$$\begin{aligned} -\log P(x) &= -\log\left(1-\frac{1}{n}\right) - \log\left(1-\frac{2}{n}\right) - \cdots - \log\left(1-\frac{x-1}{n}\right) \\ &= \sum_{i=1}^{\infty} \frac{1^i}{i\cdot n^i} + \sum_{i=1}^{\infty} \frac{2^i}{i\cdot n^i} + \cdots + \sum_{i=1}^{\infty} \frac{(x-1)^i}{i\cdot n^i} \\ &= \sum_{i=1}^{\infty} \frac{1^i + 2^i + \cdots + (x-1)^i}{i\cdot n^i}. \end{aligned} \tag{9}$$

This derivation is due to Tsaban [24].

For $i \geq 1$, the sum $1^i + 2^i + \cdots + (x-1)^i$ can be expressed as

$$S_i(x) := \frac{1}{i+1} \sum_{t=0}^{i} \binom{i+1}{t} B_t x^{i+1-t} \tag{10}$$

where B_t is the t-th *Bernoulli number*, cf. [10]. It was this with this famous formula that Jakob Bernoulli, as related in his *Ars Conjectandi* [3], computed

$$1^{10} + 2^{10} + \cdots + 1000^{10} = 91409924241424243424241924242500$$

within half of a quarter of an hour! It appears that $S_i(x)$ is a polynomial of degree $i+1$ with leading coefficient $\frac{1}{i+1}$. The first few such *power sum polynomials* are

$$\begin{aligned} S_1(x) &= \tfrac{1}{2}x^2 - \tfrac{1}{2}x, \\ S_2(x) &= \tfrac{1}{3}x^3 - \tfrac{1}{2}x^2 + \tfrac{1}{6}x, \\ S_3(x) &= \tfrac{1}{4}x^4 - \tfrac{1}{2}x^3 + \tfrac{1}{4}x^2. \end{aligned}$$

Incidentally, the beautiful identity $1^3 + 2^3 + \cdots + x^3 = (1+2+\cdots+x)^2$ was discovered by the Indian mathematician Aryabhata in the fifth century AD, cf. [13].

Using (9) and (10), we can express $-\log P(x)$ by the series

$$L(x) := \sum_{i=1}^{\infty} \frac{S_i(x)}{i\cdot n^i} = \frac{x^2-x}{2n} + \frac{2x^3-3x^2+x}{12n^2} + \frac{x^4-2x^3+x^2}{12n^3} + \cdots \tag{11}$$

The function $P(x)$ is only defined for *integers* $x > 0$; hence $-\log P(x) = L(x)$ holds for such x only. It seems we can now define $L(x)$ for all *real* $x > 0$ and that, if we find a solution x to the equation $L(x) = \log 2$, we may conclude $\mathscr{X}(n) = \lceil x \rceil$. This idea, however, turns out not to work because (11) diverges for all real numbers $x > 0$ except $x = 1, 2, \ldots, n$. The reason is that for x fixed and i going to infinity, $S_i(x)$ grows much faster for non-integral values of x than for integral values. More precisely, the "normalized" functions

$$\frac{(2\pi)^{i+1}}{2i!} \cdot S_i(x)$$

converge (pointwise) to $\pm\sin(2\pi x)$ or $\pm(\cos(2\pi x) - 1)$ for i going to infinity in a fixed residue class modulo 4, cf. [7]. We will see later how to get around this problem by truncating $L(x)$.

For later use, we note the following lemma.

Lemma 1. *The power sum polynomials* $S_i(x)$ *satisfy*

$$S_i(x) < \tfrac{1}{i+1}x^{i+1} \text{ for all integral } x > 0 \tag{12}$$

as well as

$$S_i'(x) > 0 \text{ for all real } x > \tfrac{i}{2}. \tag{13}$$

Proof. The first inequality follows immediately from $S_i(x) = 1^i + \cdots + (x-1)^i$. To prove the second, write

$$S_i'(x) = \sum_{t=0}^{i} \binom{i}{t} B_t x^{i-t} = x^i - \tfrac{i}{2}x^{i-1} + \sum_{t=2}^{i} \binom{i}{t} B_t x^{i-t}.$$

Clearly, $x^i - \frac{i}{2}x^{i-1}$ is positive for $x > \frac{i}{2}$. The Bernoulli numbers B_t are zero for $t \geq 3$ odd, cf. [10]. For $t \geq 2$ even, they are given by Euler's remarkable formula

$$B_t = (-1)^{t/2+1} \cdot \zeta(t) \cdot \frac{2t!}{(2\pi)^t}$$

and hence satisfy

$$-\frac{B_{t+2}}{B_t} = \frac{\zeta(t+2)}{\zeta(t)} \cdot \frac{(t+2)(t+1)}{(2\pi)^2} < \frac{(t+2)(t+1)}{(2\pi)^2}.$$

From

$$\frac{\binom{i}{t+2}}{\binom{i}{t}} = \frac{(i-t)(i-t-1)}{(t+2)(t+1)}$$

thus follows

$$-\frac{\binom{i}{t+2} B_{t+2} x^{i-t-2}}{\binom{i}{t} B_t x^{i-t}} < \frac{(i-t)(i-t-1)}{(2\pi x)^2} < \left(\frac{i}{2\pi x}\right)^2 < 1$$

for t even, $2 \leq t \leq i-2$, and $x > \frac{i}{2\pi}$. It follows that $\sum_{t=2}^{i} \binom{i}{t} B_t x^{i-t}$ is non-negative for $x > \frac{i}{2\pi}$ since the terms have decreasing absolute values and alternating signs starting with plus. □

The assumption $x > \frac{i}{2}$ in (13) in not quite optimal but good enough for our purposes. Figure 1 shows a logarithmic plot of $S_{99}(x)$ and $\frac{1}{100}x^{100}$. It is striking how badly (12) fails for small, non-integral x (roughly less than $\frac{100}{2\pi e}$). For these values of x, $S_{99}(x)$ is approximated well by $\frac{1}{100}B_{100} \cdot (\cos(2\pi x) - 1)$.

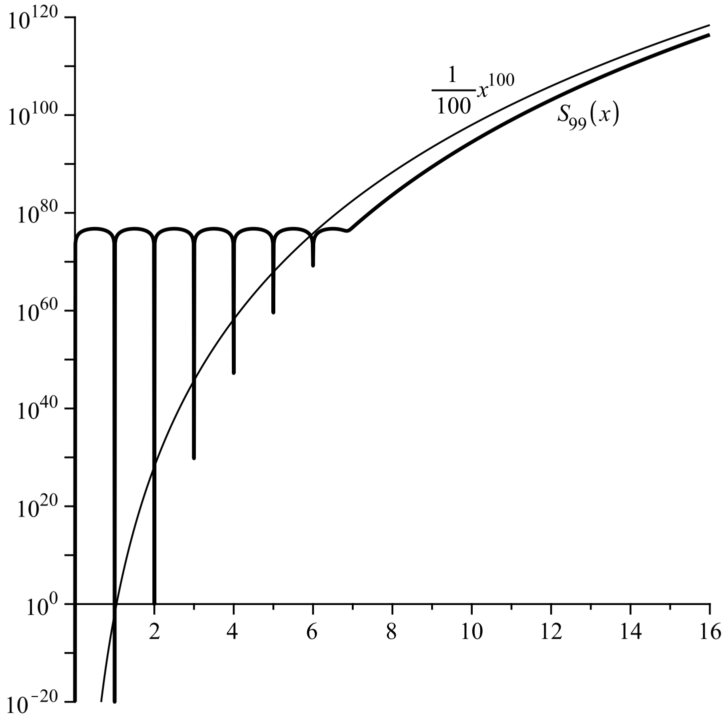

Figure 1.

3 The Numbers c_t

Lemma 2. *There is a unique sequence of real numbers* $c_0, c_1, c_2, \ldots$ *with* $c_0 > 0$ *such that the expression*

$$x_\infty = c_0\sqrt{n} + c_1 + \frac{c_2}{\sqrt{n}} + \frac{c_3}{n} + \frac{c_4}{n\sqrt{n}} + \cdots \tag{14}$$

formally *satisfies* $L(x_\infty) = \log 2$.

Proof. Recall the definition of $L(x)$ in (11). Inserting x_∞ into $L(x)$ gives

$$L(x_\infty) = \tfrac{1}{2}c_0^2 + \frac{c_0c_1 - \frac{1}{2}c_0 + \frac{1}{6}c_0^3}{\sqrt{n}} + \cdots \tag{15}$$

Write (15) as $L(x_\infty) = \sum_{t=0}^{\infty} a_t n^{-t/2}$. Clearly, there is a unique $c_0 > 0$ such that $a_0 = \log 2$. For $t > 0$, the coefficient a_t is a rational polynomial in $c_0, \ldots, c_t$. Also, the only term in a_t involving c_t is c_0c_t which comes from the term $\frac{1}{2}x^2$ in $S_1(x)$. Hence setting $a_t = 0$ leads to c_t being expressed as c_0^{-1} times a rational polynomial in $c_0, \ldots, c_{t-1}$ and the lemma follows. □

In order to compute the numbers c_t, one has to truncate $L(x)$ and x_∞ after a suitable finite number of terms. Let

$$L_t(x) := \sum_{i=1}^{t} \frac{S_i(x)}{i \cdot n^i}$$

and

$$x_t := c_0\sqrt{n} + c_1 + \cdots + \frac{c_{t-1}}{n^{t/2-1}}. \tag{16}$$

Then $L(x_\infty)$ and $L_t(x_t)$ agree on the first t terms, and thus

$$L_t(x_t) = \log 2 + O\left(\frac{1}{n^{t/2}}\right). \tag{17}$$

So if $c_0, \ldots, c_{t-1}$ are given, c_t is computed by setting the $(t+1)$-th coefficient of $L_{t+1}(x_{t+1})$ equal to zero. In this way, one gets

$$c_0 = \sqrt{2\log 2}, \quad c_1 = \frac{3 - 2\log 2}{6}, \quad c_2 = \frac{9 - 4(\log 2)^2}{72\sqrt{2\log 2}}, \quad c_3 = -\frac{2(\log 2)^2}{135} \quad \text{etc.}$$

Some more values are given here numerically:

t	c_t	t	c_t	t	c_t	t	c_t
0	1.177410	10	−0.007691	20	−0.033288	30	−4.600721
1	0.268951	11	−0.015160	21	−0.213123	31	−32.525344
2	0.083495	12	−0.018821	22	−0.457278	32	−78.432870
3	−0.007118	13	−0.017225	23	−0.647448	33	−120.668962
4	0.010599	14	−0.008059	24	−0.543744	34	−90.042864
5	0.015199	15	0.012161	25	0.234529	35	147.977172
6	0.012099	16	0.040053	26	2.061711	36	760.519219
7	0.011477	17	0.069337	27	4.872601	37	1763.143956
8	0.005595	18	0.085003	28	7.455238	⋮	
9	−0.001123	19	0.059623	29	6.364906	100	1.9×10^{28}

The expression (14) is a formal Laurent series in $n^{-1/2}$. It is not clear if it converges with the coefficients c_t defined in Lemma 2. In fact, a glance at the numerical values of the c_t makes convergence seem highly doubtful.

4 A Quasi-Asymptotic Formula

The following lemma expresses quantitatively that if $L_t(x)$ is close to $\log 2$ for some real x, then this x is close to $\mathscr{X}(n)$. Figure 2 shows $L(x)$ and $L_1(x)$ for $n = 10$.

Lemma 3. *Let $t \geq 1$ and $n \geq \max\{\frac{1}{4}(t-1)^2, 10\}$, and suppose x is a real number satisfying $\sqrt{n} + \frac{1}{2} < x < \sqrt{2n}$. Then $\mathscr{X}(n) = \lceil x + \epsilon \rceil$ with*

$$|\epsilon| < \sqrt{n}\left(|L_t(x) - \log 2| + \frac{4}{(t+1)(t+2)} \cdot \left(\frac{2}{n}\right)^{t/2}\right).$$

Proof. For any real $y > \sqrt{n}+\frac{1}{2}$, Lemma 1 and $n \geq \frac{1}{4}(t-1)^2$ give that $S_i'(y)$ is positive for $i = 1, \ldots, t$ and hence

$$L_t'(y) = \sum_{i=1}^{t} \frac{S_i'(y)}{i \cdot n^i} \geq \frac{S_1'(y)}{n} = \frac{y - \frac{1}{2}}{n} > \frac{1}{\sqrt{n}}. \tag{18}$$

It follows that the integer

$$x_{\max} := \left\lceil x + \sqrt{n} \cdot |L_t(x) - \log 2| \right\rceil$$

satisfies $L(x_{\max}) > L_t(x_{\max}) > \log 2$ and consequently $\mathscr{X}(n) \leq x_{\max}$.

To bound $\mathscr{X}(n)$ from below, note that for integral $z < \sqrt{2n}$, Lemma 1 and $n \geq 10$ give

$$\begin{aligned} L(z) - L_t(z) &= \sum_{i=t+1}^{\infty} \frac{S_i(z)}{i \cdot n^i} \\ &< \sum_{i=t+1}^{\infty} \frac{(\sqrt{2n})^{i+1}}{i(i+1) \cdot n^i} \\ &< \frac{2}{(t+1)(t+2)} \sum_{i=t+1}^{\infty} \left(\sqrt{\frac{2}{n}}\right)^{i-1} \\ &= \frac{2}{(t+1)(t+2)} \cdot \frac{1}{1-\sqrt{2/n}} \cdot \left(\frac{2}{n}\right)^{t/2} \\ &< \frac{4}{(t+1)(t+2)} \cdot \left(\frac{2}{n}\right)^{t/2}. \end{aligned} \tag{19}$$

Now consider the integer

$$x_{\min} := \left\lfloor x - \sqrt{n}\left(|L_t(x) - \log 2| + \frac{4}{(t+1)(t+2)} \cdot \left(\frac{2}{n}\right)^{t/2}\right) \right\rfloor.$$

If $x_{\min} \leq \sqrt{n} + \frac{1}{2}$, then $\mathscr{X}(n) > x_{\min}$ already by (2). If $x_{\min} > \sqrt{n} + \frac{1}{2}$, then (18) gives

$$L_t(x_{\min}) < L_t(x) - \frac{x - x_{\min}}{\sqrt{n}} < \log 2 - \frac{4}{(t+1)(t+2)} \cdot \left(\frac{2}{n}\right)^{t/2}$$

and hence $L(x_{\min}) < \log 2$ by (19). There follows $\mathscr{X}(n) > x_{\min}$. □

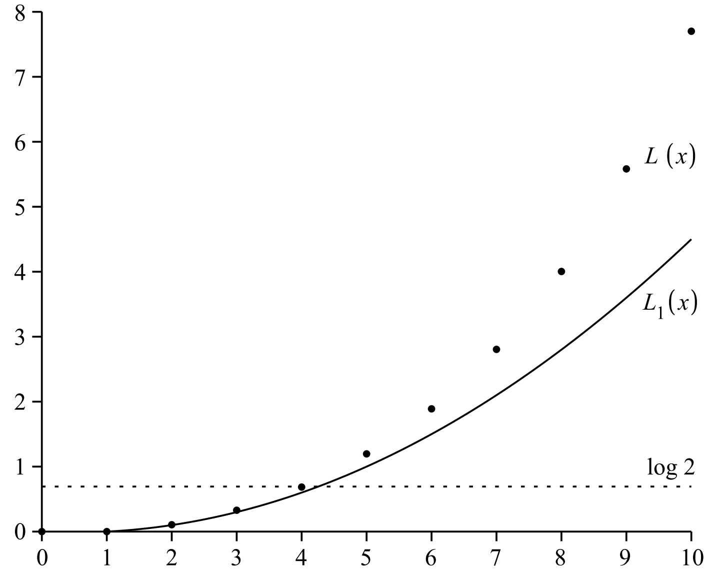

Figure 2.

Theorem 1. *With the constants $c_0, c_1, c_2, \ldots$ defined in Lemma 2, we have*

$$\mathscr{X}(n) = \left\lceil c_0\sqrt{n} + c_1 + \frac{c_2}{\sqrt{n}} + \frac{c_3}{n} + \cdots + \frac{c_{t-1}}{n^{t/2-1}} + O\left(\frac{1}{n^{(t-1)/2}}\right)\right\rceil$$

for any $t \geq 1$.

Proof. This follows immediately from Lemma 3 and (17) by taking $x = x_t$. Note that the condition $\sqrt{n} + \frac{1}{2} < x_t < \sqrt{2n}$ holds for sufficiently large n since $1 < c_0 < \sqrt{2}$. □

Theorem 1 represents $\mathscr{X}(n)$ by something which is very similar to an asymptotic series in the sense of Poincaré without actually being one because of the brackets denoting the ceiling function. For example, the coefficients in a true asymptotic series are uniquely determined [4], but it is not clear for how many of the c_t this is the case.

5 Explicit Bounds

Next, we investigate what constants lie behind the O-symbol in Theorem 1 for small t. Take, say, $t = 6$ and suppose $n \geq 10$. Then $n > \frac{1}{4}(t-1)^2$ and $\sqrt{n} + \frac{1}{2} < x_6 < \sqrt{2n}$ as required by Lemma 3. Write

$$L_6(x_6) = \sum_{i=1}^{6} \frac{S_i(x_6)}{i \cdot n^i} = \log 2 + \frac{a_7}{n^3} + \frac{a_8}{n^{7/2}} + \cdots + \frac{a_{41}}{n^{20}}$$

by inserting (16) into (10). Computing the real constants $a_7, \ldots, a_{41}$ numerically gives

$$|L_6(x_6) - \log 2| \leq \left(|a_7| + \frac{|a_8|}{\sqrt{10}} + \cdots + \frac{|a_{41}|}{10^{17}}\right)\frac{1}{n^3} < \frac{1}{10n^3}.$$

Also note

$$\frac{4}{(t+1)(t+2)} \cdot \left(\frac{2}{n}\right)^{t/2} = \frac{4}{7n^3}.$$

It now follows from Lemma 3 that

$$\mathscr{X}(n) = \left\lceil c_0\sqrt{n} + c_1 + \frac{c_2}{\sqrt{n}} + \frac{c_3}{n} + \frac{c_4}{n\sqrt{n}} + \frac{c_5}{n^2} + \frac{\epsilon_6}{n^2\sqrt{n}} \right\rceil \tag{20}$$

with $|\epsilon_6| < 1$ for all $n \geq 10$. It can easily be checked that (20) also holds for $n < 10$, but this is of secondary concern to us since the bound on ϵ_6 is rather weak, and (20) is used only as a stepping-stone to the sharper results in the following theorem.

Theorem 2. *With $c_0, \ldots, c_4$ as defined in Lemma 2, the following formulas hold for all $n \geq 1$:*

(a) $\mathscr{X}(n) = \left\lceil c_0\sqrt{n} + \epsilon_1 \right\rceil$ *with* $c_1 < \epsilon_1 < 0.28$,

(b) $\mathscr{X}(n) = \left\lceil c_0\sqrt{n} + c_1 + \frac{\epsilon_2}{\sqrt{n}} \right\rceil$ *with* $0.083 < \epsilon_2 < c_2$,

(c) $\mathscr{X}(n) = \left\lceil c_0\sqrt{n} + c_1 + \frac{c_2}{\sqrt{n}} + \frac{\epsilon_3}{n} \right\rceil$ *with* $c_3 < \epsilon_3 < -0.007$,

(d) $\mathscr{X}(n) = \left\lceil c_0\sqrt{n} + c_1 + \frac{c_2}{\sqrt{n}} + \frac{c_3}{n} + \frac{\epsilon_4}{n\sqrt{n}} \right\rceil$ *with* $c_4 < \epsilon_4 < 0.011$.

Proof. Let ϵ_6 be as in (20) and put

$$\begin{aligned}
\epsilon_1 &= c_1 + \frac{c_2}{\sqrt{n}} + \frac{c_3}{n} + \frac{c_4}{n\sqrt{n}} + \frac{c_5}{n^2} + \frac{\epsilon_6}{n^2\sqrt{n}}, \\
\epsilon_2 &= c_2 + \frac{c_3}{\sqrt{n}} + \frac{c_4}{n} + \frac{c_5}{n\sqrt{n}} + \frac{\epsilon_6}{n^2}, \\
\epsilon_3 &= c_3 + \frac{c_4}{\sqrt{n}} + \frac{c_5}{n} + \frac{\epsilon_6}{n\sqrt{n}}, \\
\epsilon_4 &= c_4 + \frac{c_5}{\sqrt{n}} + \frac{\epsilon_6}{n}.
\end{aligned}$$

The four statements then hold with each ϵ_i within the stated bounds for all sufficiently large n, i.e., for n greater than 100, 200, 10000, and 5000, respectively. The remaining cases are checked by explicit computation. □

The same proof shows that the upper bound on ϵ_1 can be strengthened to $\epsilon_1 < 0.277$ for $n \neq 43$. This, together with $c_1 < \epsilon_1$, gives the inequalities (4) mentioned in the introduction. It follows from Theorem 2.b and the fact that $c_0\sqrt{n} + c_1$ is dense modulo 1 that $\mathscr{X}(n) - c_0\sqrt{n}$ has limit inferior c_1 and limit superior $1 + c_1$.

6 Uniform Distribution

Let $\{x\}$ denote the fractional part of x. A sequence of real numbers a_n, $n = 1, 2, \ldots$, is called *uniformly distributed modulo 1* if, for every interval I contained in the unit interval, the number

$A(N)$ of indices $n = 1, \ldots, N$ with $\{a_n\} \in I$ satifies

$$\frac{A(N)}{N} \to |I| \text{ for } N \to \infty,$$

i.e., the set of indices n with $\{a_n\} \in I$ has asymptotic density $|I|$.

Lemma 4 below is due to van der Corput; it is proved in [5] and [16] using *Weyl's Criterion.* The following proof is perhaps more direct.

Lemma 4. *Let a_n be a sequence of real numbers. Suppose the sequence $\Delta a_n := a_{n+1} - a_n$ is decreasing with $\Delta a_n \to 0$ and $n \cdot \Delta a_n \to \infty$ for $n \to \infty$. Then a_n is uniformly distributed modulo 1.*

Proof. First note that a_n is strictly increasing with $a_n \to \infty$. Let I be any interval contained in the unit interval. We may assume that $p = |I|$ satisfies $0 < p < 1$. Let J be the interval between I and $I+1$, and put $q = |J| = 1-p$. We may also assume that the first element a_1 of the sequence is in I. For $i \geq 0$, let A_i and B_i be the numbers of indices n such that $a_n \in I+i$ and $a_n \in J+i$, respectively. From $\Delta a_n \to 0$ it follows that $A_i \to \infty$ and $B_i \to \infty$. We have to show that the fractions

$$u_i := \frac{A_0 + A_1 + \cdots + A_i}{A_0 + B_0 + A_1 + B_1 + \cdots + A_i}, \; v_i := \frac{A_0 + A_1 + \cdots + A_i}{A_0 + B_0 + A_1 + B_1 + \cdots + B_i}$$

both converge to p. For any i with $B_i \neq 0$, let m be the index such that a_m is the last element of the sequence in $J + i$. Then

$$B_i \leq \frac{q}{\Delta a_m} + 1 \quad \text{and} \quad A_{i+1} \geq \frac{p}{\Delta a_m} - 1 \tag{21}$$

since the sequence Δa_n is decreasing. The first inequality of (21) and the assumption $n \cdot \Delta a_n \to \infty$ give

$$\frac{B_i}{A_0 + B_0 + A_1 + B_1 + \cdots + B_i} = \frac{B_i}{m} \leq \frac{q}{m \cdot \Delta a_m} + \frac{1}{m} \to 0 \quad \text{for} \quad i \to \infty$$

and thus

$$u_i - v_i \to 0 \quad \text{for} \quad i \to \infty. \tag{22}$$

It also follows from (21) that

$$\frac{A_{i+1} + 1}{B_i - 1} \geq \frac{p}{q}, \quad \liminf_{i \to \infty} \frac{A_{i+1}}{B_i} \geq \frac{p}{q}, \quad \liminf_{i \to \infty} \frac{A_1 + \cdots + A_{i+1}}{B_0 + \cdots + B_i} \geq \frac{p}{q}$$

and hence

$$\liminf_{i \to \infty} u_i \geq p. \tag{23}$$

An analogous argument using the last element in $I + i$ gives

$$\limsup_{i \to \infty} v_i \leq p. \tag{24}$$

It now follows from (22), (23), and (24) that $u_i \to p$ and $v_i \to p$ for $i \to \infty$ as required. □

Note that the sequence $a_n = \log n$ is (correctly) excluded by the assumption $n \cdot \Delta a_n \to \infty$. For

Δa_n decreasing, the two assumptions $\Delta a_n \to 0$ and $n \cdot \Delta a_n \to \infty$ can be stated compactly as $\frac{1}{n} \prec \Delta a_n \prec 1$. It is easy to see that $\Delta a_n \prec 1$ is equivalent to $a_n \prec n$, and that $\frac{1}{n} \prec \Delta a_n$ implies $\log n \prec a_n$. However, $\log n \prec a_n \prec n$ does *not* guarantee uniform distribution. Consider, for example, the sequence given by $a_2 = 2$ and $\Delta a_n = 2^{-2^i}$ for $2^{2^i} \le n < 2^{2^{i+1}}$ and $i \ge 0$. It satisfies $\Delta a_n = \frac{1}{n}$ infinitely often and consequently is not uniformly distributed, but it is a nice exercise to show that $a_n > \sqrt[3]{n}$ for all $n \ge 2$.

7 Exact Formulas

Theorem 3. *For every $n \ge 1$, the following four statements hold:*

$$\mathscr{X}(n) = \left\lceil c_0\sqrt{n} \right\rceil \quad or \quad \mathscr{X}(n) = \left\lceil c_0\sqrt{n} \right\rceil + 1, \tag{25}$$

$$\mathscr{X}(n) = \left\lceil c_0\sqrt{n} + c_1 \right\rceil \quad or \quad \mathscr{X}(n) = \left\lceil c_0\sqrt{n} + c_1 \right\rceil + 1, \tag{26}$$

$$\mathscr{X}(n) = \left\lceil c_0\sqrt{n} + c_1 + \frac{c_2}{\sqrt{n}} \right\rceil \quad or \quad \mathscr{X}(n) = \left\lceil c_0\sqrt{n} + c_1 + \frac{c_2}{\sqrt{n}} \right\rceil - 1, \tag{27}$$

$$\mathscr{X}(n) = \left\lceil c_0\sqrt{n} + c_1 + \frac{c_2}{\sqrt{n}} + \frac{c_3}{n} \right\rceil \quad or \quad \mathscr{X}(n) = \left\lceil c_0\sqrt{n} + c_1 + \frac{c_2}{\sqrt{n}} + \frac{c_3}{n} \right\rceil + 1. \tag{28}$$

Let (25.I)–(28.I) and (25.II)–(28.II) denote the first and second alternatives of (25)–(28), respectively. Then (25.I) holds for a set of integers n with asymptotic density $1 - c_1 \approx 0.731$, whereas (26.I), (27.I), and (28.I) hold for a set of integers n with asymptotic density 1.

Proof. The four statements (25)–(28) follow directly from Theorem 2. As before, let $\{x\}$ denote the fractional part of x. It then follows from Theorem 2.b that (25.I) holds precisely when

$$\left\{c_0\sqrt{n}\right\} \le 1 - \left(c_1 + \frac{\epsilon_2}{\sqrt{n}}\right). \tag{29}$$

The sequence $\{c_0\sqrt{n}\}$ is uniformly distributed in the unit interval by Lemma 4 above. The right-hand side of (29) converges to $1 - c_1$. Therefore, the set of integers n satisfying (29) has asymptotic density $1 - c_1$. Similarly, (26.I), (27.I), and (28.I) all hold, say, when

$$\left\{c_0\sqrt{n} + c_1\right\} \le 1 - \frac{1}{\sqrt{n}}. \tag{30}$$

From the uniform distribution of the left-hand side of (30), it follows that the set of integers n satisfying (30) has asymptotic density 1. □

Conjecture 1. *(a) The number of positive integers $n \le x$ for which (26.II) holds is asymptotically equal to $2c_2\sqrt{x} \approx 0.167\sqrt{x}$.*

(b) The number of positive integers $n \le x$ for which (27.II) holds is asymptotically equal to $-c_3 \log x \approx 0.007 \log x$.

(c) There are no *positive integers n for which (28.II) holds.*

We give some arguments in favour of Conjecture 1. It follows from Theorem 2.c that (26.II) holds precisely when

$$\left\{c_0\sqrt{n}+c_1\right\} > 1-\left(\frac{c_2}{\sqrt{n}}+\frac{\epsilon_3}{n}\right). \tag{31}$$

Heuristically, the "probability" that (31) holds for any given n is $c_2/\sqrt{n}+\epsilon_3/n$. Hence, the expected number of integers $n \le x$ for which (31) holds is

$$\sum_{n=1}^{x}\left(\frac{c_2}{\sqrt{n}}+\frac{\epsilon_3}{n}\right) \sim 2c_2\sqrt{x}.$$

Similarly, (27.II) holds precisely when

$$\left\{c_0\sqrt{n}+c_1+\frac{c_2}{\sqrt{n}}\right\} < -\left(\frac{c_3}{n}+\frac{\epsilon_4}{n\sqrt{n}}\right),$$

and the expected number of integers $n \le x$ for which this happens is

$$-\sum_{n=1}^{x}\left(\frac{c_3}{n}+\frac{\epsilon_4}{n\sqrt{n}}\right) \sim -c_3\log x.$$

Even though this argument suggests that there are infinitely many counter-examples to (27.I), it also suggests that these counter-examples are extremely rare. For example, the expected number of such integers n less than $\exp(-c_3^{-1}) \approx 10^{61}$ is only one, and indeed the author has been unable to find any.

Finally, (28.I) holds whenever

$$\left\{c_0\sqrt{n}+c_1+\frac{c_2}{\sqrt{n}}+\frac{c_3}{n}\right\} \le 1-\frac{0.011}{n\sqrt{n}}, \tag{32}$$

and in the same heuristic sense as above, the expected number of integers n for which (32) fails is

$$\sum_{n=1}^{\infty}\frac{0.011}{n\sqrt{n}} = 0.011\cdot\zeta(1.5) \approx 0.029.$$

8 Computations

In order to compute $\mathscr{X}(n)$, it is rather slow to use the definition (1) directly. It is much faster to use first (28.I) and then (32) in order to validate the result. With this method, the author has computed for various values of x the number of integers $n \le x$ for which (25.II) to (28.II) hold:

x	(25.II)	(26.II)	(27.II)	(28.II)
10^2	28	3	0	0
10^4	2684	16	0	0
10^6	269087	169	0	0
10^8	26897996	1633	0	0
10^{10}	2689542678	16697	0	0
10^{12}	268951262882	166699	0	0
10^{14}	26895096291945	1669923	0	0
10^{16}	2689509420037294	16702323	0	0
10^{18}	268950939911815654	166985401	0	0

These results fit nicely with Theorem 3 and Conjecture 1. The computations also confirmed that (32) holds for all $n \le 10^{18}$. Given this fact, the expected number of integers for which (32) fails decreases to

$$\sum_{n=10^{18}+1}^{\infty} \frac{0.011}{n\sqrt{n}} < \frac{1}{45{,}000{,}000{,}000}.$$

This motivates the statement from the introduction that the probability that a counter-example to (7) exists is less than one in 45 billion.

It is a famous open problem similar to (32) to prove that the inequality

$$\left\{\left(\frac{3}{2}\right)^n\right\} \le 1 - \left(\frac{3}{4}\right)^n$$

holds for all $n \ge 2$. This conjecture is closely related to *Waring's Problem* and has been verified for all n up to 471,600,000 [15].

9 Ramanujan's Q-function

Recall the definition of $\mathscr{X}(n)$ as the minimal number of students in a class such that the probability of finding two students with the same birthday is at least 50 percent. Now suppose the students enter the class one at a time, and let the stochastic variable T be the number of students present when the first birthday coincidence occurs. Then T takes values $2, \ldots, n+1$ and has tail probabilities $\mathbf{P}(T > x) = P(x)$ given by (1) for $x = 1, 2, \ldots$ Our function $\mathscr{X}(n)$ is the *median* of T. The *expected value* can be written as $Q(n) + 1$ with

$$Q(n) := \sum_{x=1}^{n} P(x) = 1 + \left(1 - \frac{1}{n}\right) + \left(1 - \frac{1}{n}\right)\left(1 - \frac{2}{n}\right) + \cdots \tag{33}$$

This function is called *Ramanujan's Q-function* because of its connection to the following problem posed by Ramanujan in *Journal of the Indian Mathematical Society* in 1911, cf. [20, 21]: *Shew that*

$$\tfrac{1}{2}e^n = 1 + \frac{n}{1!} + \frac{n^2}{2!} + \cdots + \frac{n^n}{n!}\Theta,$$

where Θ *lies between* $\frac{1}{2}$ *and* $\frac{1}{3}$.[2] In the 1912 issue, Ramanujan showed

$$\Theta(n) = \frac{n!e^n}{2n^n} - Q(n) = \frac{n!e^n}{2n^n} + 1 - \int_0^\infty e^{-x}\left(1+\frac{x}{n}\right)^n dx,$$

gave the asymptotic series

$$\Theta(n) = \frac{1}{3} + \frac{4}{135n} - \frac{8}{2835n^2} - \frac{16}{8505n^3} + \cdots,$$

and posed a refined problem. The definition (33) is due to Knuth who derived rigorously the asymptotic expansion of $Q(n)$, cf. [14, pp. 112–118]. See also [2, 8] for the history and ultimate solution of Ramanujan's problems.

We give here a new derivation of the asymptotic expansion of $Q(n)$ based on (11). Use $P(x) = \exp(-L(x))$ and the Taylor expansion of the exponential function to write

$$\begin{aligned} P(x) &= \exp\left(-\frac{x^2}{2n}\right)\cdot\exp\left(\frac{x}{2n} - \sum_{i=2}^{\infty}\frac{S_i(x)}{i\cdot n^i}\right) \\ &= \exp\left(-\frac{x^2}{2n}\right)\cdot\left(1+\frac{x}{2n}+\frac{-4x^3+9x^2-2x}{24n^2}+\cdots\right) \end{aligned}$$

for $x = 1, \ldots, n$. To bound the tail sums, Lemma 1 gives

$$\sum_{i=t}^{\infty}\frac{S_i(x)}{i\cdot n^i} < \frac{1}{t}\cdot\frac{x^{t+1}}{n^t}$$

so that for example

$$P(x) = \exp\left(-\frac{x^2}{2n}\right)\cdot\left(1+\frac{x}{2n}+O\left(\frac{x^3}{n^2}\right)\right).$$

Euler's summation formula [10] gives the asymptotic series

$$\sum_{x=1}^{n} x^i e^{-x^2/n} = I - \tfrac{1}{2}0^i + O\left(n^{-m}\right)$$

for $i \geq 0$ even, and

$$\sum_{x=1}^{n} x^i e^{-x^2/n} = I + \sum_{\substack{t=i+1\\ t\text{ even}}}^{m}\frac{(-1)^{(t+i-1)/2}B_t}{t((t-i-1)/2)!}n^{-(t-i-1)/2} + O\left(n^{-(m-i+1)/2}\right)$$

for $i > 0$ odd. Here m is an arbitrary (even) integer, 0^0 is defined as 1, and I is the integral

$$\int_0^\infty x^i e^{-x^2/n}dx = \tfrac{1}{2}\Gamma\left(\frac{i+1}{2}\right)n^{(i+1)/2}$$

which can be evaluated using the familiar properties $\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$, $\Gamma(1) = 1$, and $\Gamma(i+1) = i\cdot\Gamma(i)$ of the Gamma function. Putting everything together gives

$$\begin{aligned} Q(n) = \sqrt{\frac{\pi n}{2}} - \frac{1}{3} + \frac{1}{12}\sqrt{\frac{\pi}{2n}} - \frac{4}{135n} + \frac{1}{288}\sqrt{\frac{\pi}{2n^3}} + \frac{8}{2835n^2} - \frac{139}{51840}\sqrt{\frac{\pi}{2n^5}} \\ + \frac{16}{8505n^3} - \frac{571}{2488320}\sqrt{\frac{\pi}{2n^7}} - \frac{8992}{12629925n^4} + \frac{163879}{209018880}\sqrt{\frac{\pi}{2n^9}} + O\left(\frac{1}{n^5}\right). \end{aligned}$$

[2] The original has x rather than n.

References

[1] Ahmed, S.E., McIntosh, R.J., An asymptotic approximation for the birthday problem, *Crux Math.* **26**, 151–155 (2000)

[2] Berndt, B.C., Rankin, R.A. (editors), *Ramanujan: Essays and Surveys*, American Mathematical Society, Providence (2001)

[3] Bernoulli, J., *Ars Conjectandi*, Basel (1713). Translation: *Wahrscheinlichkeitsrechnung*, Verlag von Wilhelm Engelmann, Leipzig (1899)

[4] de Bruijn, N.G., *Asymptotic Methods in Analysis*, 3rd ed., Dover, New York, (1970)

[5] van der Corput, J.G., Diophantische Ungleichungen. I. Zur Gleichverteilung modulo Eins, *Acta. Math.* **56**, 373–456 (1931)

[6] Davenport, H., *The Higher Arithmetic*, 8th ed., Cambridge University Press, Cambridge (2008)

[7] Dilcher, K., Asymptotic behaviour of Bernoulli, Euler, and generalized Bernoulli polynomials, *J. Approx. Theory* **49**, 321–330 (1987)

[8] Flajolet, P., Grabner, P.J., Kirschenhofer, P., Prodinger, H., On Ramanujan's Q-function, *J. Comput. Appl. Math.* **58**, 103–116 (1995)

[9] Gamow, G., *One, Two, Three... Infinity*, Viking, New York (1947)

[10] Graham, R., Knuth, D.E., Patashnik, O., *Concrete Mathematics*, 2nd. ed., Addison Wesley, Reading (1994)

[11] Halmos, P.R., *I Want to Be a Mathematician*, Springer, New York (1985)

[12] Holmes, S., *Random Birthday Applet.* Available electronically at http://www-stat.stanford.edu/∼susan/surprise/Birthday.html

[13] Katz, V.J., *A History of Mathematics*, Harper Collins, New York (1993)

[14] Knuth, D.E., *The Art of Computer Programming*, vol. I, 2nd. ed., Addison Wesley, Reading (1973)

[15] Kubina, J.M., Wunderlich, M.C., Extending Waring's conjecture to $471,600,000$, *Math. Comp.* **55**, 815–820 (1990)

[16] Kuipers, L., Niederreiter, H., *Uniform Distribution of Sequences*, John Wiley & Sons, New York (1974)

[17] Littlewood, J.E., *A Mathematician's Miscellany*, Methuen, London (1953)

[18] Mathis, F.H., A generalized birthday problem, *SIAM Review* **33**, 265–270 (1991)

[19] von Mises, R., Ueber Aufteilungs- und Besetzungs-Wahrscheinlichkeiten, *Rev. Fac. Sci. Univ. Istanbul* **4**, 145–163 (1939)

[20] Ramanujan, S., Question 294, *J. Indian Math. Soc.* **3**, 128 (1911)

[21] Ramanujan, S., On Question 294, *J. Indian Math. Soc.* **4**, 151–152 (1912)

[22] Singmaster, D., *Sources in Recreational Mathematics*. Published electronically at www.g4g4.com/MyCD5/SOURCES/SOURCE4.DOC

[23] Sloane, N.J.E., *On-Line Encyclopedia of Integer Sequences*. Published electronically at http://oeis.org/

[24] Tsaban, B., Bernoulli numbers and the probability of a birthday surprise, *Discrete Appl. Math.* **127**, 657–663 (2003)