A computer look at *N*!

JERRY SULLIVAN

Factorials are huge and widespread

The product over the integers 1 to N is written as $N!$, is called N factorial, and is defined as:

$$N! = 1 \times 2 \times 3 \times \ldots \times (N-1) \times N. \tag{1}$$

Factorials have two important characteristics, the frequency with which they occur in mathematics and the sheer size of their values. They occur often in combination problems, probablilty theory, power series expansions for many familiar functions, and so forth. Factorials are also common in the physics branch of statistical mechanics, where probability is applied to large collections of particles. 'Large collections' means numbers like $N = 10^{19}$ or 10^{22}, so $N!$ is an astromomical number. In Planck's paper that gave birth to quantum theory, he applied probability theory and factorials to the atoms in a 'blackbody'.

The sum over the integers 1 to N, $\{1 + 2 + 3 + \ldots + (N-1) + N\}$, can be simplified to $\frac{1}{2}N(N+1)$. However, there is no formula for $N!$ that is both exact *and* relatively easy to compute. Mathematicians have worked on the problem for nearly 300 years and have discovered many good estimates. In 1730, Leonard Euler found *symbolically* that for any integer N, $N!$ was equal to the Area (from $x = 0$ to ∞) under the curve $y_N(x) = x^N e^{-x}$:

$$\int_0^\infty x^N e^{-x} dx = N!, \tag{2}$$

where $e = 2.7182\ldots$ is Euler's number.

It turns out that the approximation process is considerably easier to visualise by making a simple change of variable, $x = t^2$, in the integral in (2). The expression for $N!$ is now

$$\int_0^\infty 2t^{(2N+1)} e^{-t^2} dt = N!. \tag{3}$$

The approach used here starts by first plotting the curves $y_N = 2t^{(2N+1)}e^{-t^2}$ for different values of N on a computer, to see what connection they have with $N!$. Because the y_N values are so large, even for small N, they have to be scaled to fit on the computer screen. Once the scaled curves are displayed, they look remarkably similar, leading to a simple first working approximation for $N!$. Next, plotting the ratio between $N!$ and the first approximation to $N!$ leads to a straightforward second approximation. This second estimate is accurate enough for many applications.

https://doi.org/10.1017/mag.2022.63 Published online by Cambridge University Press

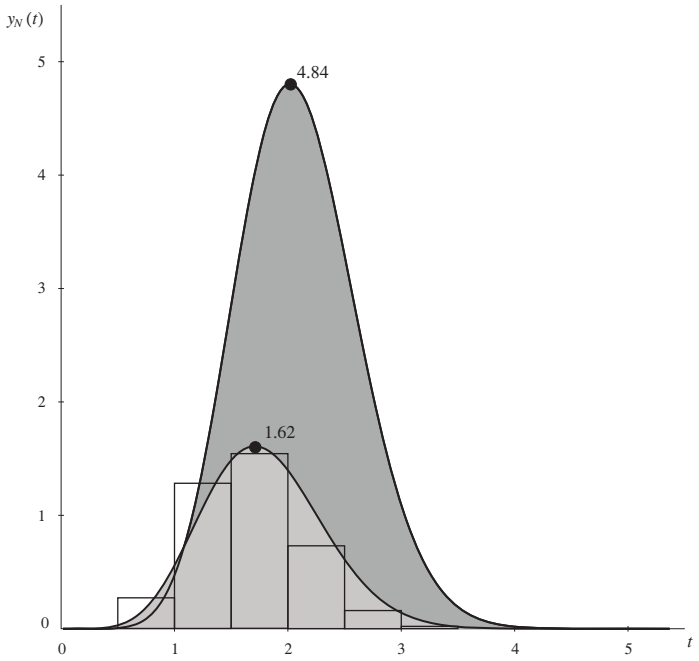

FIGURE 1: The area under the curves $y_N(t) = 2t^{(2N+1)}e^{-t^2}$ is shown for $N = 2$ and $N = 3$. The maximum values for each curve are shown by the circles.

How is Euler's area under the integral related to factorials?

Figure 1 shows what the unscaled curves $y_N(t) = 2t^{(2N+1)}e^{-t^2}$ and $N = 2$ and 3 look like. The curves for all values of N have the same general shape, with a single maximum and decreasing to 0 as $t \to 0$ or ∞. For any N, calculus shows that the maximum value is

$$\mathrm{YMAX}_N = \mathrm{YMAX}(t_\mathrm{M}) = 2\left[\frac{N+\frac{1}{2}}{e}\right]^{N+\frac{1}{2}},$$

which occurs when $2t_\mathrm{M}^2 = 2N + 1$ or $t_M = \sqrt{N+\frac{1}{2}}$. What does the area under these curves have to do with 2! and 3! ? First of all, the ratio of their maximum values, $4.84/1.62 \approx 2.986$ nearly satisfies the definition of $N!$, $N!/(N-1)! = N$. This simple observation is essentially the key to all that follows; the maximum value of Euler's (modified) curve for any N contains almost all the information needed to evaluate $N!$. Table 1 shows this property holds for the first few integers.

https://doi.org/10.1017/mag.2022.63 Published online by Cambridge University Press

N	2	3	4	5	6	7	8	9	10
R	1.979	2.986	3.989	4.991	5.993	6.994	7.995	8.995	9.996

TABLE 1: The ratio $R = \dfrac{\text{YMAX}_N}{\text{YMAX}_{N-1}}$ calculated for $N = 2$ to 10.

Secondly, the identity in (3) can be confirmed, at least for small N, by calculating the area numerically using a simple midpoint approximation to sum the areas of 12 rectangles, from $t = 0$ to 6. This process is shown in Figure 1; add up the areas under the rectangles to approximate the $N = 2$ curve. The rectangle widths all equal $\frac{1}{2}$; the heights are computed at $t = 0.25, 0.75, \ldots, 5.25, 5.75$. The area under the $y_N(t) = 2t^{(2N-1)}e^{-t^2}$ curves can be computed with surprising accuracy by this process. (Some rectangle heights are so small that they blend in with the horizontal axis.) The results are shown in Table 2.

N	2	4	6	8	10
$N!$	2	24	720	40,320	3,628,000
$N!_{\text{app}}$	2.00016	24.00003	720.0001	40,320	3,628,000

TABLE 2: $N!$ values are compared to results from approximating the area ($N!_{\text{APP}}$) under the curves $y_N(t) = 2t^{2N+1}e^{-t^2}$ using midpoint rectangles such as those shown in Figure 1.

Finally, the identity in (3) is formally proved by using the identity $2t^{(2N+1)}e^{-t^2} = \left(2te^{-t^2}\right)t^{2N} = -\dfrac{d\left(e^{-t^2}\right)}{dt}t^{2N}$ and repeated integration by parts:

$$\int_0^\infty 2t^{(2N+1)}e^{-t^2}dt = N\int_0^\infty 2t^{(2N-1)}e^{-t^2}dt = N(N-1)\int_0^\infty 2t^{(2N-3)}e^{-t^2}, \ldots,$$

where $\int_0^\infty 2t^{(1)}e^{-t^2}dt = 1$.

Scaling the curves on a computer screen: first approximation to $N!$

The maximum values of the $y_N(t)$ curves for $N = 5$, 10 and 20 are 1×10^2, 3×10^6 and 2×10^{18}. Because the maximum values increase so dramatically, in order to compare these curves on a computer monitor it makes sense to divide each curve by its maximum value. Figure 2 shows what the scaled curves $y_N(t)/\text{YMAX}_N$ for $N = 5$, 10 and 20 look like. The scaled curves are very similar to each other and to the familiar bell-shaped (Gaussian normal) curve.

https://doi.org/10.1017/mag.2022.63 Published online by Cambridge University Press

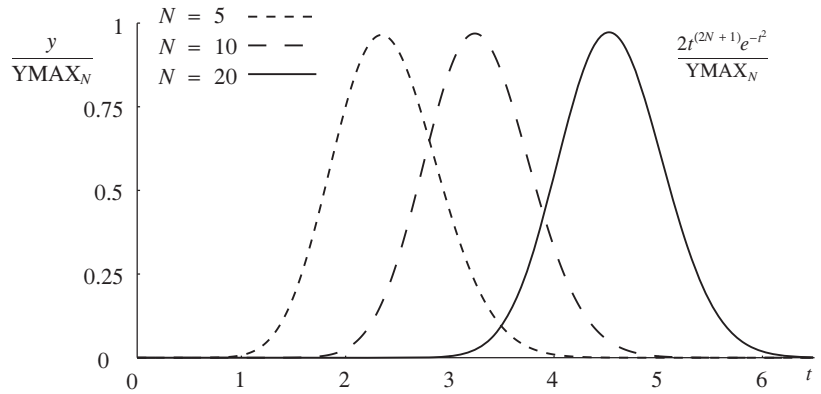

FIGURE 2: The $y_N(t) = 2t^{(2N+1)}e^{-t^2}$ curves for 5!, 10! and 20! are each scaled by their maximum value YMAX_N : 10^2, 3×10^6 and 2×10^{18}, respectively

To further show the similarity of the curves, we simply shift them so that the maximum values from each curve are plotted at the same point on the horizontal axis. Figure 3 shows the same three scaled curves, but now shifted and overlaid. This is equivalent to plotting them against the variable $T = t - \sqrt{N + \frac{1}{2}}$. The similarity of the three curves is strikingly clear. The curves are now exceptionally close to each other (there really are three curves). Figure 3 suggests that as N increases, the curves approach what we shall term a *universal* curve, which will be discussed shortly.

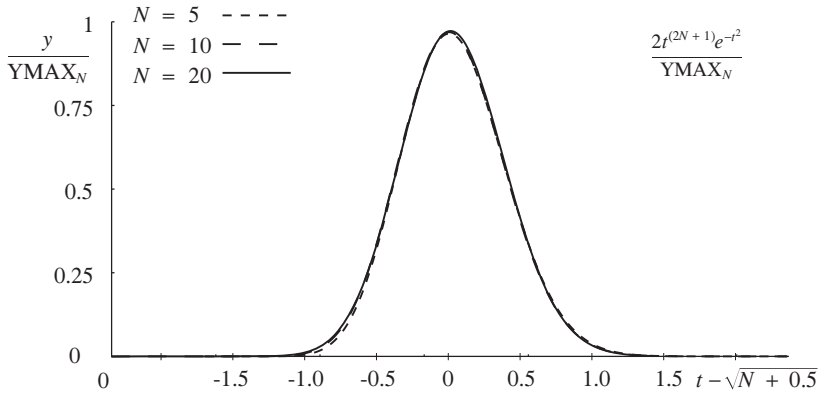

FIGURE 3: The same curves as in Figure 2, $y_N(t) = 2t^{(2N+1)}e^{-t^2}$ curves for $N = 5,\ 10$ and 20 and scaled by their maximum values, but now overlaid such that their maximum values coincide with each other.

To obtain a first approximation for $N!$, we now unravel the process that generated Figure 3. To produce the *universal* curve, we divided the y axis of each $N!$ curve by $2\left[(N + \frac{1}{2})/e\right]^{N+\frac{1}{2}}$. Figure 3 suggests that all scaled $N!$

https://doi.org/10.1017/mag.2022.63 Published online by Cambridge University Press

curves fall effectively on this same curve. To compute any particular $N!$ back from the universal curve, we simply reverse the process and multiply the area under the universal curve by $2\left[(N + \frac{1}{2})/e\right]^{N+\frac{1}{2}}$. So, the first approximation for calculating $N!$ efficiently is

$$N! \approx C_1 \left\{2\left[\frac{N + \frac{1}{2}}{e}\right]^{N+\frac{1}{2}}\right\}, \tag{4}$$

where C_1 is a coefficient that measures the area under the universal curve in Figure 3.

Second approximation for $N!$

A direct way to estimate the coefficient C_1 is to divide the exact $N!$ value by the term $\left\{2\left[\frac{N + \frac{1}{2}}{e}\right]^{N+\frac{1}{2}}\right\}$, as given by (4). Table 3 shows the C_1 estimates for $N = 5$, 10, 15 and 20. C_1 is not an exact constant; this is also shown by the fact that the scaled curves for $N = 5$, 10 and 20 in Figure 3 do not coincide precisely. As N increases, the difference between consecutive C_1 values decreases, suggesting that C_1 approaches a limiting value as N gets large. One way to estimate the limit is to plot C_1 against the variable $1/(N + \frac{1}{2})$. As N goes from 1 to ∞, $1/(N + \frac{1}{2})$ only varies between 2/3 and 0, greatly compressing the horizontal axis and making the value of the limit more apparent.

N	$N!/N!_{(\text{app})}$	C_1
5	120 / 96.47	1.2439
10	$3.6288 \times 10^6 / 2.9069 \times 10^6$	1.2484
15	$1.3077 \times 10^{12} / 1.0462 \times 10^{12}$	1.2500
20	$2.4329 \times 10^{18} / 1.9451 \times 10^{18}$	1.2508

TABLE 3: For $N = 5$, 10, 15 and 20, the coefficient $C_1 = N!/N!_{(\text{app})}$ is computed. $\left\{2\left[\frac{N + \frac{1}{2}}{e}\right]^{N+\frac{1}{2}}\right\}$ is labelled $N!_{(\text{app})}$

Figure 4 shows the plotted values of C_1 against $1/(N + \frac{1}{2})$. Two features are apparent; the values appear to converge as $1/(N + \frac{1}{2}) \to 0$, and they lie on a (nearly) straight line. The values converge to $C_\infty = 1.2533$ as $1/(N + \frac{1}{2}) \to 0$. The second approximation is now

$$N! \approx 1.2533\left\{2\left[\frac{N + \frac{1}{2}}{e}\right]^{N+\frac{1}{2}}\right\} = 2.5066\left[\frac{N + \frac{1}{2}}{e}\right]^{N+\frac{1}{2}}.$$

Using C_1 values at $N = 10$ and $N = 20$ (7-digit accuracy) to estimate the slope of the line, the value is $-1/24.09$. A more detailed analysis shows the value multiplying $1/\left(N + \frac{1}{2}\right)$ to be $-1/24$, resulting in a more accurate estimate for $N!$.

https://doi.org/10.1017/mag.2022.63 Published online by Cambridge University Press

The limiting value of C_1 is actually $C_\infty = \sqrt{\pi/2} = 1.2533\ldots$, so our final estimate here for $N!$ is

$$N! \approx \sqrt{2\pi}\left[\frac{N + \frac{1}{2}}{e}\right]^{N+\frac{1}{2}}\left[1 - \frac{1}{24\left(N + \frac{1}{2}\right)}\right]. \tag{5}$$

This is very similar to Stirling's approximation for $N!$, discovered within a year of Euler's integral definition for $N!$.

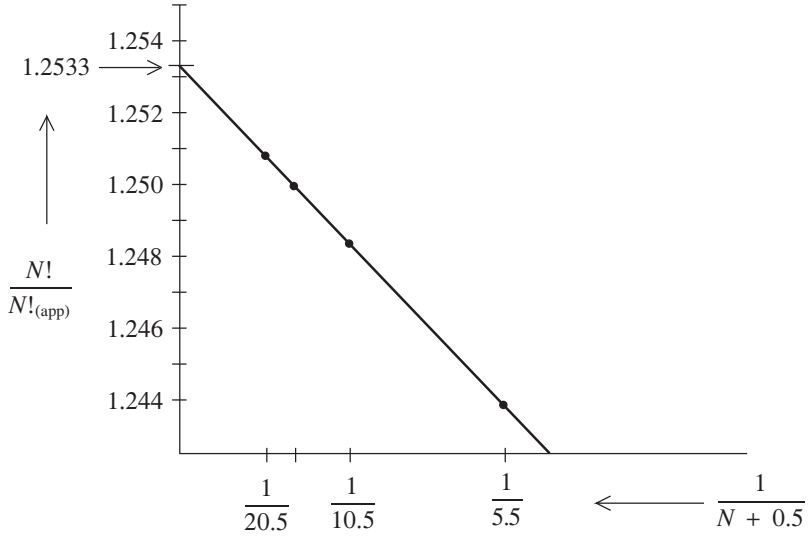

FIGURE 4: The coefficient C_1 plotted against the variable $1/(N + 0.5)$, as $N \to \infty$. The line is an actual straight line (drawn by hand).

How accurate is the approximation?

Table 4 shows the accuracy of the approximation in (5), for values of $N = 5$, 10 and 20, and also for 70!, which displays an *Overflow* message on my hand calculator. The goodness-of-fit depends on the criterion. The approximation for 10! is 36 less than the actual value but is also 99.999% of the actual value.

5!	120	≈	119.995
10!	3,628,800	≈	3,628,764
20!	$2.43290\ldots \times 10^{18}$	≈	$2.43289\ldots \times 10^{18}$
70!	$1.1978571\ldots. \times 10^{100}$	≈	$1.1978569\ldots. \times 10^{100}$

TABLE 4: Comparison of $N!$ to the approximation given by (5), shown by the (≈) symbol.

The approximation works very well, especially considering the range of the

https://doi.org/10.1017/mag.2022.63 Published online by Cambridge University Press

numbers involved. There are better approximations (remember that mathematicians have been working on this for almost 300 years), but this is sufficient for even fairly advanced work.

Universal curve and the area under it

The scaled curves in Figure 3 closely resemble Gaussian normal curves, so I assumed that the universal curve had the form $U(T) = \exp(-aT^2)$, with $T = (t - t_{\mathrm{M}}) = \left[t - \sqrt{N + \frac{1}{2}}\right]$. As $N \to \infty$, $T \to -\infty$, so the limits of integration for the universal curve are $-\infty$ to ∞. I first determined the constant a by this method. The numerical work from Figure 4 (and about 300 years of mathematical analysis) shows that the area under the Gaussian-like curve $\approx 1.2533\ldots$. The integral $\int_{-\infty}^{\infty} e^{-at^2} dt = \sqrt{\pi / a}$, so $\sqrt{a} = \sqrt{\pi}/1.2533 \approx 1.41423$. From this, $a = 1.41423^2 \approx 2.00005$, which is where the original choice of $\exp(-2T^2)$ comes from. This is a global way of showing that $a = 2$. Figure 5 shows the scaled curves for $N = 5$ and 15 plotted against the universal curve $U(T) = \exp(-2T^2)$. As N gets larger, the scaled factorial curves approach the universal curve.

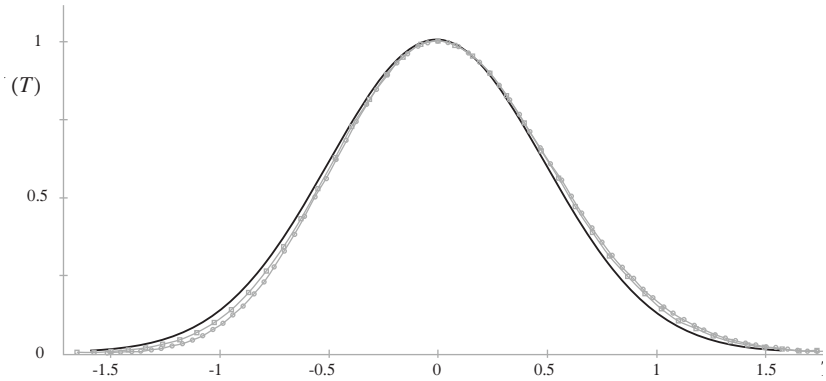

FIGURE 5: The broken curves are the scaled curve values for $N = 5$ (circles) and $N = 15$ (squares). The solid curve is the universal curve, $U(T) = \exp(-2T^2)$, where $T = (t - t_M)$.

The constant a can also be determined locally by showing that as $N \to \infty$, the sequence of scaled curves $\frac{y_N(t)}{y_N(t_{\mathrm{M}})} \to \exp(-2T^2)$, for any T.

$$\frac{y_N(t)}{y_N(t_{\mathrm{M}})} = \frac{2t^{(2N+1)}e^{-t^2}}{2t_M^{(2N+1)}e^{-t_{\mathrm{M}}^2}} = \frac{(t_{\mathrm{M}} + T)^{(2N+1)}\, e^{-(t_{\mathrm{M}}+T)^2}}{t_{\mathrm{M}}^{(2N+1)}e^{-t_{\mathrm{M}}^2}}$$

$$= \left[1 + \left(\frac{T}{t_{\mathrm{M}}}\right)^{2N+1}\right] e^{-(T^2 + 2Tt_{\mathrm{M}})} = \left[1 + \left(\frac{T}{\sqrt{N + \frac{1}{2}}}\right)^{2N+1}\right] e^{-\left(T^2 + T\sqrt{N+\frac{1}{2}}\right)}.$$

https://doi.org/10.1017/mag.2022.63 Published online by Cambridge University Press

The question is, for what value of a does

$$\left[1 + \frac{T}{\sqrt{N + \frac{1}{2}}}\right]^{(2N+1)} e^{-\left[T^2 + 2T\sqrt{N+\frac{1}{2}}\right]} \to e^{-aT^2}$$

as $N \to \infty$?

It is simpler if we separate the terms by first taking the natural logarithm of the scaled factorial, find the limit as $N \to \infty$, then exponentiate the result.

$$(2N + 1)\ln\left[1 + \frac{T}{\sqrt{N + \frac{1}{2}}}\right] - \left[T^2 + 2T\sqrt{N + \tfrac{1}{2}}\right] \to \left(-aT^2\right)$$

as $N \to \infty$?

Dividing by $-T^2$ gives

$$1 + \frac{2\left(\frac{T}{\sqrt{N+\frac{1}{2}}}\right) - 2\ln\left[1 + \frac{T}{\sqrt{N+\frac{1}{2}}}\right]}{\left(\frac{T}{\sqrt{N+\frac{1}{2}}}\right)^2} \to a \text{ as } N \to \infty. \tag{6}$$

Figure 6 shows this expression plotted against the variable $1/\sqrt{N + \frac{1}{2}}$, for various values of T ranging from $T = -3$ to $+3$. The N values go from 10^4 to 10^8. Convergence to the limit $a = 2$ as $N \to \infty$ or $1/\sqrt{N + \frac{1}{2}} \to 0$ is evident. Although not shown, for larger values of T the (nearly) straight line convergence to 2 is still evident; it just occurs at larger values of N.

The formal proof that the sequence in expression (6) converges to the limit $a = 2$, for all T, requires the Cesàro-Stolz Lemma. This lemma is a discrete version of L'Hôpital's rule; it uses differences instead of derivatives. However, in this particular case, using the lemma tends to disguise the result rather than explain it, so I use a simpler approach which still captures the essentials of the argument.

Set $v \equiv 1/\sqrt{N + \frac{1}{2}}$; as $N \to \infty$, $v \to 0$. Now evaluate the expression

$$Q = 1 + \frac{2\left[vT - \ln(1 + vT)\right]}{(vT)^2}, \text{ as } v \to 0. \tag{7}$$

This continuous function is a model for the actual sequence. As $v \to 0$, both the numerator, $2\left[vT - \ln(1 + vT)\right]$, and the denominator, $(vT)^2$, also tend to zero. Using L'Hôpital's rule we find that

$$\frac{d}{dv}\left\{2\left[vT - \ln(1 + vT)\right]\right\} = \frac{2vT^2}{1 + vT}, \tag{8}$$

$$\frac{d}{dv}\left\{(vT)^2\right\} = 2vT^2. \tag{9}$$

Dividing (8) by (9) results in $1/(1 + vT)$. As $v \to 0$, the term $1/(1 + vT) \to 1$, for all T, so

$$1 + 1 = 2 \to a \text{ and } \ln\left[\frac{y_N(t)}{y_N(t_M)}\right] \to -2T^2 \text{ as } N \to \infty.$$

https://doi.org/10.1017/mag.2022.63 Published online by Cambridge University Press

The scaled curve $\dfrac{y_N(t)}{y_N(t_M)} \to e^{-2T^2}$ as $N \to \infty$.

Both chains of reasoning, global and local, reproduce what Figure 5 suggests is true.

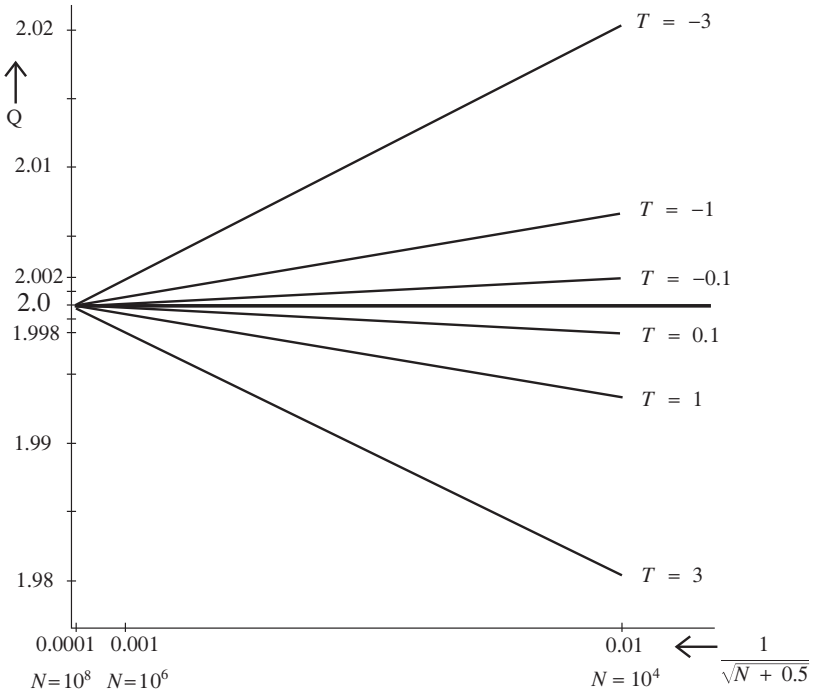

FIGURE 6: Let $v \equiv 1/\sqrt{N + \frac{1}{2}}$. The quantity in expression (6) from the text, $Q = 1 + 2\left[vT - \ln(1 + vT)\right]/(vT)^2$, is plotted against the variable $1/\sqrt{N + \frac{1}{2}}$, for values of N ranging from 10^4 to 10^8. Convergence to the limit 2 is shown for several values of T ranging from −3 to +3.

10.1017/mag.2022.63 © The Authors, 2022
Published by Cambridge University Press on behalf of The Mathematical Association

JERRY SULLIVAN
3006 Homewood Parkway, Kensington, MD 20895, USA
e-mail: *jerry.thomas.sullivan@gmail.com*