Comment on JRM Problem 2680

Don Knuth, 09 January 2011

The following problem by Daniel Shine was published on pages 144–145 of the *Journal of Recreational Mathematics*, Volume 33:

> **2680.** One thousand loops of string, each of unit length, are placed in a box. One piece of string is removed at random. It is cut at a random location and put back into the box. This select-and-cut process is repeated 1000 times. All pieces, whether loops or not, are equally likely to be selected. After the 1000 repetitions, what is the average length of the pieces of string in the box?

A supposed "solution" appeared on page 151 of Volume 34, but it incorrectly assumed that division is a linear operation. So I wrote to the editor and pointed out this serious flaw. The editor quoted from my letter on page 161 of Volume 35.

Alas, when I reread those remarks after receiving that issue of the journal, I noted to my chagrin that I wasn't operating on all cylinders when I'd written that letter! My comments were correct, but they didn't nail the error. Namely, I'd observed that the true average number of pieces after 3 cuttings, P_3, was ≈ 1000.003, and that $P_2 + 1/P_2 \approx 1000.002$. But any reasonable reader would respond to that by saying, "So what?" The supposed value of P_3 in the published solution wasn't $P_2 + 1/P_2$, it was claimed to be $P_2 + 2/P_2$; and $P_2 + 2/P_2$ is also ≈ 1000.003.

I should really have said that $P_3 \approx 1000.002999998002$, while $P_2 + 1/P_2 \approx 1000.002999998000$. The error in the solver's formula for P_3 would then have been manifest.

Embarassed by my oversight, I decided to take a closer look at the problem, and I was surprised to see that some rather astonishing mathematical patterns arise. So I couldn't resist noting them down here.

Of course I decided to attack the problem with "modern" techniques of analysis, using generating functions. Let p_{nk} be the probability that, after n cuts, there are n nonloops and k total pieces (thus $k - n$ loops); and let's define

$$P_n(z) = \sum_{k=0}^{\infty} p_{nk} z^k = \sum_{k=1000}^{\infty} p_{nk} z^k.$$

Let Φ be the operator on generating functions that takes z^k into $(z^k - z^{k+1})/k$; namely

$$\Phi g(z) = (1-z) \int_0^z \frac{g(z)}{z}\, dz.$$

Then it's not difficult to verify that our generating function $P_n(z)$ is defined by the recurrence

$$P_0(z) = z^{1000};$$
$$P_{n+1}(z) = P_n(z) - n\Phi P_n(z) = (1 - n\Phi)P_n(z).$$

Consequently we have the nice "factored" formula

$$P_n(z) = (1-\Phi)(1-2\Phi)\ldots(1-(n-1)\Phi)z^{1000}.$$

Since Φ is a linear operator, it's natural to look for its eigenvalues. And a bit of experimentation reveals that there's a nice family of eigenfunctions $f_m(z)$ such $\Phi f_m(z) = f_m(z)/m$, for all $m \ge 1$, namely

$$f_m(z) = (1-z)(ze^{-z})^m.$$

To verify this one can note, for example, that $\Phi = (1-z)\vartheta^{-1}$ and $f_m(z) = \vartheta((ze^{-z})^m)/m$, where ϑ is the operator $zD = zd/dz$ discussed in Section 5.6 of my book *Concrete Mathematics*.

Thus we can write $P_0(z) = z^{1000}$ as a linear combination of the f's,

$$z^{1000} = \sum_{m=1000}^{\infty} a_m f_m(z); \qquad (*)$$

and it will follow that

$$P_n(z) = \sum_{m=1000}^{\infty} a_m(1-\Phi)(1-2\Phi)\dots(1-(n-1)\Phi)f_m(z).$$

These expressions must be understood as formal power series, not as convergent sums unless z is sufficiently small. For example, $(*)$ cannot be true when $z=1$, because $1^{1000}=1$ while each f_m has $f_m(1)=0$. But when $z=1-\epsilon$ we have $f_m(1-\epsilon)=\epsilon/e^{m+m\epsilon^2/2}+O(\epsilon^2)$, so the sum will converge if $a_m=O(e^m)$.

What are those mystery coefficients a_m? It turns out, somewhat miraculously, that there's a closed form,

$$a_m = m^{m-1000}/(m-1000)!.$$

For example, $a_{1000}=1$, $a_{1001}=1001$, $a_{1002}=502002$, $a_{1003}=1009027027/6$, etc. These coefficients seem to grow at an alarming rate, but eventually they settle down somewhat; Stirling's approximation tells us that $a_{m+1000}=(m+1000)^m/m! \approx (1+1000/m)^m e^m/\sqrt{2\pi m} \approx e^{1000+m}/\sqrt{2\pi m}$. So we have in fact

$$a_m = O(m^{-1/2}e^m),$$

and the power series $\sum a_m f_m(z)$ actually converges for $|z|<1$.

(Where did that magic formula for a_m come from? Well, there's a general theory that if $g(z) = \sum a_m(ze^{-z})^m$, then $g(T(z))=\sum a_m z^m$, where $T(z)$ is the wonderful "tree function" that is described, for example, in my book *Fundamental Algorithms*. In this case we take $g(z)=z^{1000}/(1-z)$, and we use the fact that the coefficients of $T(z)^{1000}/(1-T(z))$ have a known form because of the combinatorial properties of free trees.)

The goal of Problem 2680 is to discover the expected value of the average piece length after 1000 cuttings. The total length is 1000, so the average piece length is $1000/k$ when there are k pieces. Thus the answer to the problem is $1000\sum_{k=1000}^{\infty} p_{nk}/k$.

Given a generating function $g(z)=\sum_{k=1}^{\infty} g_k z^k$, let $\varphi g = \sum_{k=1}^{\infty} g_k/k$. The answer we seek can be expressed conveniently in terms of this φ operator, as $1000\,\varphi P_{1000}$.

So what is φf_m? It turns out to be e^{-m}/m, because we know that $f_m(z)=\vartheta((ze^{-z})^m)/m$, and we may set $z=1$ after taking ϑ^{-1} of this function.

Now comes the *coup de grâce:*

$$\begin{aligned}
\varphi P_{1000} &= \varphi(1-\Phi)(1-2\Phi)\dots(1-999\Phi)z^{1000}\\
&= \varphi(1-\Phi)(1-2\Phi)\dots(1-999\Phi)\sum_{m=1000}^{\infty} a_m f_m(z)\\
&= \sum_{m=1000}^{\infty} a_m \varphi(1-\Phi)(1-2\Phi)\dots(1-999\Phi)f_m(z)\\
&= \sum_{m=1000}^{\infty} a_m \varphi(1-1/m)(1-2/m)\dots(1-999/m)f_m(z)\\
&= \sum_{m=1000}^{\infty} a_m (1-1/m)(1-2/m)\dots(1-999/m)\,\varphi f_m\\
&= \sum_{m=1000}^{\infty} a_m (m!/m^{1000})(e^{-m}/m)\\
&= \sum_{m=1000}^{\infty} m^{m-2001} m!/((m-1000)!^2 e^m). \qquad (**)
\end{aligned}$$

The terms of this final sum grow as $m^{-3/2}$, so it converges.

A pure mathematician will now stop, because we've expressed the answer as a convergent sum of "known" quantities. But a practical man might not be real happy with $(**)$, because my pure math side must admit that the convergence is *real slow:* One needs to evaluate something like 10^{2N} terms in order to get N decimal places of accuracy.

Therefore let's turn now to brute force, instead of using such a fancy analysis. After all, P_{1000}/z^{1000} is just a polynomial of degree 999. With little difficulty we can evaluate it accurately by using interval arithmetic, thus finding all the probabilities p_{nk} for $n = 1000$ and $1000 \le k < 2000$. These probabilities increase from about 4×10^{-433} for $k = 1000$ to about .0339 for $k = 1414$, then decrease to about 2×10^{-600} for $k = 1999$. Further exploration shows that $p_{(1000)(1400)} \approx .0168$ and $p_{(1000)(1428)} = .0166$, showing that the distribution is sort of bell-shaped about its mode.

The numerical value of $1000\,\varphi P_{1000}$ is thereby found to be

$$0.7072836536930390137543010297424017779615987765012863769040537599123236,$$

correct to 70 decimal places. For comparison, the value of $1/\sqrt{2}$ is

$$0.7071067811865475244008443621048490392848359376884740365883398689953662.$$

There remain some interesting questions, which I leave as exercises for the reader: How can $(**)$ be evaluated by asymptotic methods? Why is it fairly close to $1/(1000\sqrt{2})$? More precisely, if 1000 is replaced by N in that sum, and if 2001 is replaced by $2N + 1$, how close is the result to $1/(N\sqrt{2})$, as $N \to \infty$?

Stanford University
10 January 2011