FiveThirtyEight's December 20, 2019 Riddler

Emma Knight

December 26, 2019

This is my attempt to say what I can about the riddler from December 20th, 2019. To wit, here is the question:

Question 1. *You have N pairs of socks, all distinct and unmatched in your sock drawer. Every morning, you reach into your drawer and pull out socks one at a time until you find a match. On average, how many socks do you pull out?*

To state the answer to this question, I will introduce some notation. Let $f(N)$ be the average number of socks needed to pull out until you get a match. Additionally, let $g(N, k)$ be the probability that you need exactly k socks to get a match. Then we have the following:

Observation 2.

$$\lim_{N\to\infty} \frac{f(N)}{\sqrt{N}} = \sqrt{\pi} \approx 1.77245\ldots.$$

Here is the basic heuristic argument as to why $f(N) = \Theta(\sqrt{N})$: when one has pulled k socks, there are roughly $k^2/2$ different possible pairs of socks among the socks that have been pulled. There are roughly $2N^2$ different possible pairs of socks, and N matching pairs of socks. Assuming that the pairs of socks pulled are distributed randomly among all possible pairs of socks (which isn't true but is probably close enough to being true to be a reasonable assumption), then one expects there to be roughly a $\left(1 - \frac{1}{2N}\right)^{k^2/2} \approx e^{-\frac{k^2}{4N}}$ chance of not having a match. If $k \ll \sqrt{N}$, then it's vanishingly unlikely that you will get a match with only k socks, and if $k \gg \sqrt{N}$, then it's vanishingly unlikely that you won't have gotten a match already. Thus, most of the mass of the function $g(N, k)$ is when k is on the same order as $\sqrt{N}$ and so $f(N) = \Theta(\sqrt{N})$.

To dress this up further, let's give explicit formulae for f and g. In order to get a match at exactly sock k, one needs to not get matches at sock i for any $i < k$ and then get a match at sock k. If one hasn't gotten a match before sock i, then there are $2N - i + 1$ socks remaining, with $i - 1$ giving a match. Thus, the odds of not getting a match at sock i are $\left(\frac{2N-2i+2}{2N-i+1}\right)$ and the odds of getting a match at i are $\left(\frac{i-1}{2N-i+1}\right)$. Thus, one has that

$$g(N, k) = \left(\frac{k-1}{2N-k+1}\right) \prod_{i=1}^{k-1} \frac{2N-2i+2}{2N-i+1} = \left(\frac{k-1}{2N-k+1}\right) \prod_{i=0}^{k-2} \left(1 - \frac{i}{2N-i}\right).$$

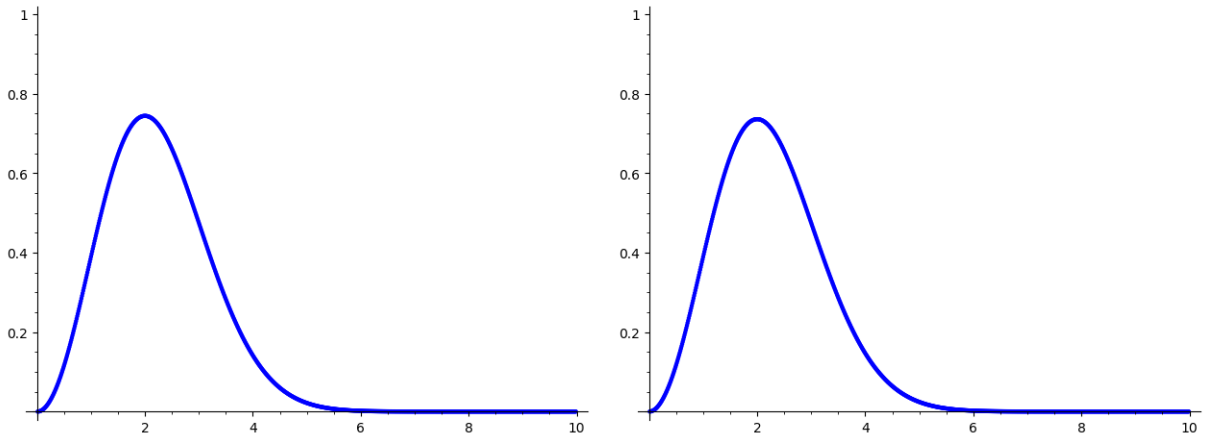
All that happened in the last equality was a little bit of rearrangement. Additionally, $f(N) = \sum_{k=2} N + 1 kg(N,k)$.

To get an approximate value for $kg(N,k)$ for k on the order of $\sqrt{N}$, let's make some substitutions. Assume that $k = x\sqrt{N}$. Then

$$\begin{aligned} kg(N,k) &= k\left(\frac{k-1}{2N-k+1}\right)\prod_{i=0}^{k-2}\left(1-\frac{i}{2N-i}\right) \\ &\approx \frac{x^2}{2}\prod_{i=0}^{k-2}\left(1-\frac{i}{2N}\right) \\ &\approx \frac{x^2}{2}\prod_{i-0}^{k-2}\left(1-\frac{1}{2N}\right)^i \\ &= \frac{x^2}{2}\left(1-\frac{1}{2N}\right)^{(k-2)(k-1)/2} \\ &\approx \frac{x^2}{2}e^{-x^2/4}. \end{aligned}$$

Now, notice that $f(N)/\sqrt{N}$ is just a Riemann sum for the integral $I = \int_0^\infty \frac{x^2}{2}e^{-x^2/4}dx$. Letting $u = -x$ and $v = e^{-x^2/4}$, one gets that $I = \int_0^\infty e^{-x^2/4}dx$. A quick substitution gives that $I = 2\int_0^\infty e^{-x^2}dx = \int_{-\infty}^\infty e^{-x^2}dx = \sqrt{\pi}$, and so one sees that $f(N)/\sqrt{(N)}$ should tend to $\sqrt{\pi}$.

Numerically, it does! Getting sagemath to compute $f(200000)/\sqrt{200000}$, one gets 1.77245495868951 and asking about $\sqrt{\pi}$, one gets 1.77245385090552, accurate to 5 decimal places. Additionally, getting a plot of the points $(k/90, g(k, 8100))$ for k between 1 and 300 and a graph of $\frac{x^2}{2}e^{-x^2/4}$, one gets:



This is all pretty convincing that there isn't that much error in the approximations that were made and that the observation is actually correct.

However, this does not constitute a fully rigorous proof. To, wit:

Proof of Observation 2. As a prelude to the proof, we will be using big-O notation. Keeping the $k = x\sqrt{N}$ notation, we will have all of the constants in the big-Os implicitly depend on x and be true as $N \to \infty$.

First, we need to estimate $kg(N,k)$. We have $kg(N,k) = \left(\frac{k(k-1)}{2N-k+1}\right)\prod_{i=0}^{k-2}\left(1-\frac{i}{2N-i}\right)$. Now, $\left|\frac{k(k-1)}{2N-k+1} - \frac{k(k-1)}{2N}\right| = \frac{|k(k-1)^2|}{2N(2N-k+1)} = O(N^{-1/2})$, and $\frac{k(k-1)}{2N} = \frac{x^2}{2} + O(N^{-1/2})$, so the term outside the product is $\frac{x^2}{2} + O(N^{-1/2})$.

To compute the term inside the product, we will use $y = \exp(\ln(y))$ for y positive. Now, one has that

$$\begin{aligned}
\ln\left(\prod_{i=0}^{k-2}\left(1-\frac{i}{2N-i}\right)\right) &= \sum_{i=0}^{k-2}\ln\left(1-\frac{i}{2N-i}\right)\\
&= \sum_{i=0}^{k-2} -\frac{i}{2N-i} + O(N^{-1})\\
&= \sum_{i=0}^{k-2} -\frac{i}{2N} - \frac{i^2}{2N(2N-i)} + O(N^{-1})\\
&= \sum_{i=0}^{k-2} -\frac{i}{2N} + O(N^{-1})\\
&= -\frac{(k-1)(k-2)}{4N} + O(N^{-1/2})\\
&= -\frac{x^2}{4} + O(N^{-1/2})
\end{aligned}$$

and so one gets $\prod_{i=0}^{k-2}\left(1-\frac{i}{2N-i}\right) = e^{-x^2/4}e^{O(N^{-1/2})} = e^{-x^2/4}(1+O(N^{-1/2})) = e^{-x^2/4}+O(N^{-1/2})$. Thus, $kg(N,k) = (\frac{x^2}{2} + O(N^{-1/2}))(e^{-x^2/4} + O(N^{-1/2})) = \frac{x^2}{2}e^{-x^2/4} + O(N^{-1/2})$. Denote $h(x) = \frac{x^2}{2}e^{-x^2/4}$.

Now, choose $\alpha \gg 0$ so that $\int_0^\alpha h(x)dx$ is within $\frac{\epsilon}{3}$ of $\int_0^\infty h(x)dx = \sqrt{\pi}$. One has $\frac{1}{\sqrt{N}}\sum_{k<\alpha\sqrt{N}} kg(N,k) = \frac{1}{\sqrt{N}}\sum_{k<\alpha\sqrt{N}}\left(h\left(\frac{k}{\sqrt{N}}\right) + O(N^{-1/2})\right) = \frac{1}{\sqrt{N}}\sum_{k<\alpha\sqrt{N}}\left(h\left(\frac{k}{\sqrt{N}}\right)\right) + O(N^{-1/2})$. As $N \to \infty$, the sum approaches $\int_0^\alpha h(x)dx$, so for $N \gg 0$, $\frac{1}{\sqrt{N}}\sum_{k<\alpha\sqrt{N}} kg(N,k)$ is within $\frac{2\epsilon}{3}$ of $\sqrt{\pi}$.

All that is left is to estimate the sum of the remaining terms: $\frac{1}{\sqrt{N}}\sum_{k>\alpha\sqrt{N}} kg(N,k)$. Now, for

$k > \alpha\sqrt{N}$, we have:

$$\begin{aligned} kg(N,k) &= k\left(\frac{k-1}{2N-k+1}\right)\prod_{i=0}^{k-2}\left(1-\frac{i}{2N-i}\right) \\ &= (h(\alpha)+O(N^{-1/2}))\frac{k(k-1)}{\alpha^2(2N-k+1)}\prod_{i>\alpha\sqrt{N}}^{k-2}\left(1-\frac{i}{2N-i}\right) \\ &\leq (h(\alpha)+O(N^{-1/2}))\frac{k(k-1)}{N}\prod_{i>\alpha\sqrt{N}}^{k-2}\left(1-\frac{\alpha\sqrt{N}}{2N}\right) \\ &= (h(\alpha)+O(N^{-1/2}))\frac{k(k-1)}{N}\left(1-\frac{\alpha}{2\sqrt{N}}\right)^{k-2-\lfloor\alpha\sqrt{N}\rfloor} \end{aligned}$$

We now split the sum into medium terms and big terms. Looking at $\alpha\sqrt{N} < k < 2\alpha\sqrt{N}$, the term $(h(\alpha)+O(N^{-1/2}))\frac{k(k-1)}{N}\left(1-\frac{\alpha}{2\sqrt{N}}\right)^{k-2}\left(1-\frac{\alpha}{2\sqrt{N}}\right)^{-\lfloor\alpha\sqrt{N}\rfloor}$ is at most

$$(h(\alpha)+O(N^{-1/2}))16\alpha^2\left(1-\frac{\alpha}{2\sqrt{N}}\right)^{k}\left(1-\frac{\alpha}{2\sqrt{N}}\right)^{-\lfloor\alpha\sqrt{N}\rfloor}$$

(assuming that $\alpha < \sqrt{N}$ which is true for $N \gg 0$), and summing that up is at most $(h(\alpha)+O(N^{-1/2}))16\alpha^2\sum_{\ell\geq 0}\left(1-\frac{\alpha}{2\sqrt{N}}\right)^{\ell} = (h(\alpha)+O(N^{-1/2}))32\alpha\sqrt{N}$ with $\ell = k-\lfloor\alpha\sqrt{N}\rfloor$, so

$$\frac{1}{\sqrt{N}}\sum_{\alpha\sqrt{N}\leq k\leq 2\alpha\sqrt{N}} kg(N,k) \leq (h(\alpha)+O(N^{-1/2}))32\alpha.$$

If $k > 2\alpha\sqrt{N}$, then $k(k-1) < 5(k-\lfloor\alpha\sqrt{N}\rfloor)(k-1-\lfloor\alpha\sqrt{N}\rfloor)$, so summing the terms up we get at most $\sum_{k>2\alpha\sqrt{N}}(h(\alpha)+O(N^{-1/2}))5\frac{(k-\lfloor\alpha\sqrt{N}\rfloor)(k-1-\lfloor\alpha\sqrt{N}\rfloor)}{N}\left(1-\frac{\alpha}{2\sqrt{N}}\right)^{k-2-\lfloor\alpha\sqrt{N}\rfloor}$. Again, replacing $k-\lfloor\alpha\sqrt{N}\rfloor$ with ℓ and summing over ℓ, we get at most $(h(\alpha)+O(N^{-1/2}))\frac{5}{N}\sum_{\ell\geq 0}(\ell)(\ell-1)\left(1-\frac{\alpha}{2\sqrt{N}}\right)^{(}\ell-2) = (h(\alpha)+O(N^{-1/2}))\frac{40\sqrt{N}}{\alpha^3}$, so one has

$$\frac{1}{\sqrt{N}}\sum_{k>2\alpha\sqrt{N}} kg(N,k) \leq (h(\alpha)+O(N^{-1/2}))\frac{40}{\alpha^3}.$$

Adding the two terms together, we get

$$\frac{1}{\sqrt{N}}\sum_{k>\alpha\sqrt{N}} kg(N,k) \leq (h(\alpha)+O(N^{1/2}))\left(32\alpha+\frac{40}{\alpha^3}\right),$$

which after choosing α and then N sufficiently large, we can arrange to be less than $\frac{\epsilon}{3}$.

Putting it all together, one has that

$$\begin{aligned}
\left|\frac{f(N)}{\sqrt{N}} - \sqrt{\pi}\right| &= \left|\left(\left(\frac{1}{\sqrt{N}}\sum_{k<\alpha\sqrt{N}} kg(N,k)\right) - \sqrt{\pi}\right) + \left(\frac{1}{\sqrt{N}}\sum_{k>\alpha\sqrt{N}} kg(N,k)\right)\right| \\
&\leq \left|\left(\frac{1}{\sqrt{N}}\sum_{k<\alpha\sqrt{N}} kg(N,k)\right) - \sqrt{\pi}\right| + \left|\frac{1}{\sqrt{N}}\sum_{k>\alpha\sqrt{N}} kg(N,k)\right| \\
&< \frac{2\epsilon}{3} + \frac{\epsilon}{3} \\
&= \epsilon.
\end{aligned}$$

Thus, one has that $\frac{f(N)}{\sqrt{N}} \to \sqrt{\pi}$ as $N \to \infty$. □