Solution

Week 60 (11/3/03)

Cereal box prizes

First Solution: Assume that you have collected c of the colors, and let B_c be the number of boxes it takes to get the next color. The average value of B_c, which we will denote by $\overline{B}_c$, may be found as follows. The probability that a box yields a new color is $1 - c/N$, and the probability that it does not is c/N. The expected number of boxes to get the next prize is therefore

$$\overline{B}_c = 1\left(1-\frac{c}{N}\right) + 2\left(\frac{c}{N}\right)\left(1-\frac{c}{N}\right) + 3\left(\frac{c}{N}\right)^2\left(1-\frac{c}{N}\right) + \cdots. \qquad (1)$$

Letting $x \equiv c/N$, we have

$$\begin{aligned}
\overline{B}_c &= (1-x)(1+2x+3x^2+4x^3+\cdots) \\
&= (1-x)\Big((1+x+x^2+\cdots)+(x+x^2+\cdots)+(x^2+\cdots)+\cdots\Big) \\
&= (1-x)\left(\frac{1}{1-x}+\frac{x}{1-x}+\frac{x^2}{1-x}\right)+\cdots \\
&= \frac{1}{1-x}\,. \qquad (2)
\end{aligned}$$

Therefore,

$$\overline{B}_c = \frac{N}{N-c}\,. \qquad (3)$$

Note that the above $1 + 2x + 3x^2 + \cdots = 1/(1-x)^2$ result, for $0 \le x < 1$, may also be derived by taking the derivative of the equality $1 + x + x^2 + \cdots = 1/(1-x)$.

We see that (with $c = 0$) it takes one box, of course, to get the first color. Then (with $c = 1$) it takes an average of $N/(N-1)$ boxes to get the second color. Then (with $c = 2$) it takes an average of $N/(N-2)$ boxes to get the third color. This continues until (with $c = N-1$) it takes an average of N boxes to get the last color.

We wish to find the average, $\overline{B}$, of the total number of boxes needed to get all the colors. $\overline{B}$ can be viewed as the average of the sum of the waiting times for each new color. But this equals the sum of the average waiting times, which were calculated above. In other words, letting B be the total number of boxes in a particular trial, and letting $B_0, B_1, \ldots, B_{N-1}$ be the waiting times during that particular trial, we have (with a bar denoting average)

$$\begin{aligned}
\overline{B} &= \overline{(B_0+B_1+\cdots+B_{N-1})} \\
&= \overline{B}_0+\overline{B}_1+\cdots+\overline{B}_{N-1} \\
&= N\left(\frac{1}{N}+\frac{1}{N-1}+\cdots+\frac{1}{2}+1\right). \qquad (4)
\end{aligned}$$

For large N, this goes like

$$\overline{B} \approx N(\ln N + \gamma), \qquad (5)$$

where $\gamma \approx 0.577$ is Euler's constant.

Second Solution: We will calculate the probability, $P(n)$, that the final color is obtained in the nth box. The desired expectation value is then $\sum_0^\infty nP(n)$.

Claim: $P(n) = \sum_{k=1}^{N-1} (-1)^{k-1} \binom{N-1}{k-1} \left(1 - \frac{k}{N}\right)^{n-1} \qquad (n = 2, 3, \ldots).$

Proof: Let us first calculate the probability, $p(n)$, that you have obtained all the colors by the time you have looked in the nth box. $P(n)$ is then given by $P(n) = p(n) - p(n-1)$.

Assume that n boxes have been bought. Then there is a total of N^n equally likely possibilities for the way the colors can turn up in these n boxes. In finding $p(n)$, we will need to subtract from N^n the number of possibilities that do *not* have at least one prize of each color. We can count this number is the following way.

If (at least) color *1* is missing, then there are $(N-1)^n$ different combinations. Likewise for the situations where another color is missing. So there seem to be $N(N-1)^n$ combinations missing a color. However, we have double-counted some of the cases. For example, a combination which has (at least) *1* and *2* missing has been counted twice; there are $(N-2)^n$ of these. Likewise for all the other pairs of colors. So we must subtract off $\binom{N}{2}(N-2)^n$ combinations. But now a combination which has (at least) *1*, *2*, and *3* missing has not been counted at all (because we have included it three times, and then subtracted it off three times); there are $(N-3)^n$ of these. Likewise for the other triplets. So we must add on $\binom{N}{3}(N-3)^n$ combinations. Now, however, the combinations with (at least) *1*, *2*, *3*, and *4* missing have been counted $\binom{4}{1} - \binom{4}{2} + \binom{4}{3} = 2$ times. Likewise for the other quadruplets. So we must subtract off $\binom{N}{4}(N-4)^n$ combinations.

In general, if we have done this procedure up to $(k-1)$-tuplets, then the combinations missing (at least) k of the colors have been counted T times, where

$$T = \binom{k}{1} - \binom{k}{2} + \cdots + (-1)^k \binom{k}{k-1}. \tag{6}$$

However, the binomial expansion gives

$$\begin{aligned} 0 &= (1-1)^k \\ &= 1 - \binom{k}{1} + \binom{k}{2} + \cdots + (-1)^{k-1}\binom{k}{k-1} + (-1)^k \\ &= 1 - T + (-1)^k. \end{aligned} \tag{7}$$

Therefore, $T = 2$ for even k, and $T = 0$ for odd k. So we have either overcounted by one, or undercounted by one. Hence, the total number of combinations missing at least one color is

$$\binom{N}{1}(N-1)^n - \binom{N}{2}(N-2)^n + \cdots + (-1)^N \binom{N}{N-1}(1)^n. \tag{8}$$

To obtain the probability, $p(n)$, that you have collected all the colors by the nth box, we must subtract this number from N^n, and then divide the result by N^n. We obtain

$$p(n) = \sum_{k=0}^{N-1} (-1)^k \binom{N}{k} \left(1 - \frac{k}{N}\right)^n, \qquad (n = 1, 2, 3, \ldots). \tag{9}$$

Remark: $p(n)$ must of course equal zero, for $n = 1, 2, \ldots, N-1$, because you need to buy at least N boxes to get all the colors. To show this explicitly, first note that $p(n)$ may be written as

$$p(n) = \sum_{k=0}^{N} (-1)^k \binom{N}{k} \left(1 - \frac{k}{N}\right)^n, \tag{10}$$

where we have let the sum run up to $k = N$, because the $k = N$ term is zero anyway. It is therefore sufficient to demonstrate that $\sum_{k=0}^{N} (-1)^k \binom{N}{k} k^m = 0$, for $m = 0, 1, \ldots N-1$, because this identity will make all the separate terms arising from the binomial expansion of $(1 - k/N)^n$ in eq. (10) equal to zero on their own. You can prove this identity by taking successive derivatives of the relation $(1-x)^N = \sum (-1)^k \binom{N}{k} x^k$, and setting $x = 1$.

The probability, $P(n)$, that the final color is obtained in the nth box is

$$\begin{aligned} P(n) &= p(n) - p(n-1) \\ &= \sum_{k=0}^{N-1} (-1)^k \binom{N}{k} \left(1 - \frac{k}{N}\right)^{n-1} \left(\left(1 - \frac{k}{N}\right) - 1\right) \\ &= \sum_{k=0}^{N-1} (-1)^k \frac{N!}{k!(N-k)!} \left(1 - \frac{k}{N}\right)^{n-1} \left(-\frac{k}{N}\right) \\ &= \sum_{k=1}^{N-1} (-1)^{k-1} \binom{N-1}{k-1} \left(1 - \frac{k}{N}\right)^{n-1}, \qquad (n = 2, 3, \ldots). \end{aligned} \tag{11}$$

$P(n)$ is zero for $n = 2, 3, \ldots, N-1$. And $P(1)$ is of course also zero, but it cannot be expressed as in eq. (11), because eq. (9) is not valid for $n = 0$. ■

Having found $P(n)$, we can write the average number, $\overline{B}$, of required boxes as

$$\overline{B} = \sum_{n=2}^{\infty} n \sum_{k=1}^{N-1} (-1)^{k-1} \binom{N-1}{k-1} \left(1 - \frac{k}{N}\right)^{n-1}. \tag{12}$$

We have let the sum over n range from 2 to ∞, instead of N to ∞, because it will simplify our calculations a bit (and all the terms from $n = 2$ to $n = N-1$ are zero). Switching the order of the sums, and performing the sum over n by using a technique similar to the one used in eq. (2), we obtain

$$\begin{aligned} \overline{B} &= \sum_{k=1}^{N-1} (-1)^{k-1} \binom{N-1}{k-1} \left(\frac{N^2}{k^2} - 1\right) \\ &= \sum_{k=1}^{N-1} (-1)^{k-1} \frac{(N-1)!}{(k-1)!(N-k)!} \left(\frac{(N+k)(N-k)}{k^2}\right) \\ &= -\sum_{k=1}^{N-1} (-1)^k \binom{N-1}{k} \left(1 + \frac{N}{k}\right) \end{aligned}$$

$$\begin{aligned}
&= -\sum_{k=1}^{N-1} (-1)^k \binom{N-1}{k} - \sum_{k=1}^{N-1} (-1)^k \binom{N-1}{k} \frac{N}{k} \\
&= -\Big((1-1)^{N-1} - 1\Big) - N \sum_{k=1}^{N-1} (-1)^k \binom{N-1}{k} \frac{1}{k}\,. \\
&= 1 - N \sum_{k=1}^{N-1} (-1)^k \binom{N-1}{k} \frac{1}{k}\,. \qquad (13)
\end{aligned}$$

At this point, we need the following claim.

Claim: $\displaystyle\sum_{k=1}^{M} (-1)^k \binom{M}{k} \frac{1}{k} = -\left(1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{M}\right).$

Proof: We will prove this by using a technique similar to the one mentioned in the above Remark, except that now we will find it necessary to take an integral instead of a derivative, because we want to produce a k in the denominator. Starting with the binomial expansion, we have

$$\begin{aligned}
\sum_{k=0}^{M} (-1)^k \binom{M}{k} x^k &= (1-x)^M \qquad (14) \\
\Longrightarrow \quad \int \sum_{k=1}^{M} (-1)^k \binom{M}{k} x^{k-1}\, dx &= \int \frac{1}{x}\Big((1-x)^M - 1\Big)\, dx \\
\Longrightarrow \quad \sum_{k=1}^{M} (-1)^k \binom{M}{k} \frac{x^k}{k} &= -\int \frac{1-(1-x)^M}{1-(1-x)}\, dx \\
\Longrightarrow \quad \sum_{k=1}^{M} (-1)^k \binom{M}{k} \frac{x^k}{k} &= -\int \Big(1 + (1-x) + \cdots + (1-x)^{M-1}\Big)\, dx \\
\Longrightarrow \quad \sum_{k=1}^{M} (-1)^k \binom{M}{k} \frac{x^k}{k} &= (1-x) + \frac{1}{2}(1-x)^2 + \cdots + \frac{1}{M}(1-x)^M + C,
\end{aligned}$$

where C is a constant of integration. Setting $x = 0$ in the last line yields

$$C = -\left(1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{M}\right). \qquad (15)$$

Setting $x = 1$ in the last line then proves the claim. ■

Using the claim in eq. (13) gives

$$\begin{aligned}
\overline{B} &= 1 + N\left(\frac{1}{N-1} + \cdots + \frac{1}{2} + 1\right) \\
&= N\left(\frac{1}{N} + \frac{1}{N-1} + \cdots + \frac{1}{2} + 1\right), \qquad (16)
\end{aligned}$$

in agreement with eq. (4).

Now let us maximize $P(n)$, as given in eq. (11). For convenience, we will instead maximize $P(n+1)$, so that there will be a simpler "n" in the exponent in eq. (11), instead of "$n-1$". Of course, such a distinction is irrelevant in the limit of large N. Let us assume for the moment (we will justify this below) that the values of k that contribute significantly to the sum in eq. (11) are small enough so that we may use the approximation,

$$\left(1-\frac{k}{N}\right)^n \approx \left(e^{-n/N}\right)^k. \tag{17}$$

The expression for $P(n+1)$ in eq. (11) then becomes (letting the sum run up to $k=N$, since the $k=N$ term is zero)

$$\begin{aligned} P(n+1) &\approx e^{-n/N}\sum_{k=1}^{N}(-1)^{k-1}\binom{N-1}{k-1}\left(e^{-n/N}\right)^{k-1} \\ &= e^{-n/N}\sum_{j=0}^{N-1}(-1)^{j}\binom{N-1}{j}\left(e^{-n/N}\right)^{j} \\ &= e^{-n/N}\left(1-e^{-n/N}\right)^{N-1} \\ &\approx e^{-n/N}e^{-(N-1)e^{-n/N}}. \end{aligned} \tag{18}$$

To maximize $P(n)$, we therefore want to minimize $n/N+(N-1)e^{-n/N}$. Taking the derivative of this with respect to n, we find that the minimum is achieved when

$$n_{\max} = N\ln(N-1) \approx N\ln N, \tag{19}$$

where we have dropped terms of order 1. Using this result in eq. (18), we see that the maximum value of $P(n)$, which is obtained at $n \approx N\ln N$, equals $1/(Ne)$.

Remarks:

1. Let us now justify eq. (17). Using

$$\ln\left(1-\frac{k}{N}\right)^n = n\ln\left(1-\frac{k}{N}\right) \approx n\left(-\frac{k}{N}-\frac{k^2}{2N^2}-\cdots\right), \tag{20}$$

we have

$$\left(1-\frac{k}{N}\right)^n \approx e^{-nk/N}e^{-nk^2/2N^2}. \tag{21}$$

Therefore, eq. (17) is valid if $nk^2/N^2 \ll 1$. For n near $N\ln N$, which is the general size of n we are concerned with, this requires that $k \ll \sqrt{N/\ln N}$. But for large N and for $n \approx N\ln N$, the first few terms in the second (and hence first) line of eq. (18) dominate the sum. This is true because the binomial coefficient goes like $N^j/j!$, and the $(e^{-n/N})^j$ term goes like $1/N^j$ (using $n \approx N\ln N$). Hence, the the jth term goes like $1/j!$. Therefore, only the first few terms contribute, so the relevant k values easily satisfy the bound $k \ll \sqrt{N/\ln N}$.

Note also that the step in going from the third to fourth line in eq. (18) is valid because $e^{-n/N} \approx 1/N$ is sufficiently small.

2. For large N, the average number of boxes, and the point where $P(n)$ is maximum (given in eqs. (5) and (19), respectively) are

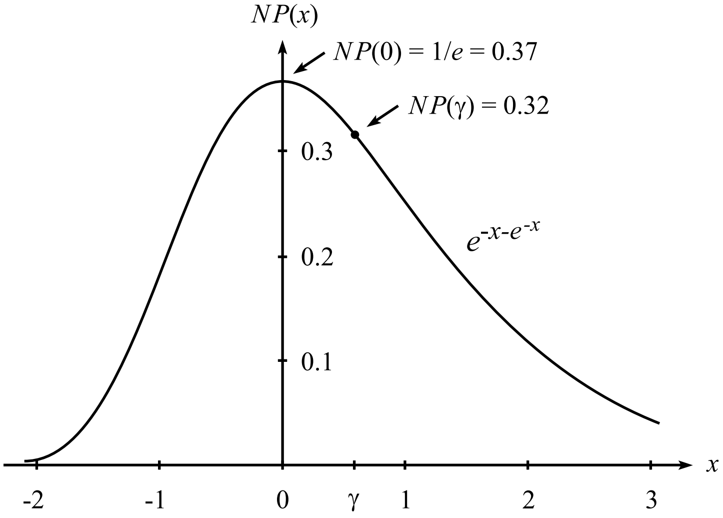
$$\overline{B} = N(\ln N + \gamma) \qquad \text{and} \qquad n_{\max} = N \ln N. \tag{22}$$

   The difference between these is $\gamma N \approx (0.577)N$. For example, when $N = 100$, the maximum occurs at $N \ln N \approx 460$, while the average number is $N \ln N + \gamma N \approx 518$.

3. If we let $n \equiv N \ln N + xN$, then we can rewrite eq. (18), to leading order in N, as

$$NP(x) \approx e^{-x-e^{-x}}. \tag{23}$$

   A plot of $NP(x)$ vs. x is shown below. This is simply a plot of $NP(n)$ vs. n, centered at $N \ln N$, and with the horizontal axis on a scale of N.



   For all large N, we obtain the same shape. Virtually all of the graph is contained within a few units of N from the maximum. Compared to a $P(n)$ vs. n graph, this graph has its vertical axis multiplied by N and its horizontal axis multiplied by $1/N$, so the total integral should still be 1. And indeed,

$$\int_{-\infty}^{\infty} e^{-x-e^{-x}} dx = e^{-e^{-x}} \Big|_{-\infty}^{\infty} = 1. \tag{24}$$

   How much area under the curve lies to the left of, say, $x = -2$? Letting $x = -a$ to be general, the integral in eq. (24) gives an area (in other words, a probability) of $e^{-e^{-x}}|_{-\infty}^{-a} = e^{-e^{a}}$. This decreases very rapidly as a grows. Even for $a = 2$, we find that there is a probability of only 0.06% of obtaining all the colors before you hit the $n = N \ln N - 2N$ box.

   How much area under the curve lies to the right of, say, $x = 3$? Letting $x = b$ to be general, we find an area of $e^{-e^{-x}}|_{b}^{\infty} = 1 - e^{-e^{-b}} \approx 1 - (1 - e^{-b}) = e^{-b}$. This decreases as b grows, but not as rapidly as in the above case. For $b = 3$, we find that there is a probability of 5% that you haven't obtained all the colors by the time you hit the $n = N \ln N + 3N$ box.

4. You might be tempted to fiddle around with a saddle-point approximation in this problem. That is, you might want to approximate $P(n)$ as a Gaussian around its maximum at $n_{\text{max}} \approx N \ln N$. However, this will *not* work in this problem, because for any (large) N, $P(n)$ will always keep its same lopsided shape. The average will always be a significant distance (namely γN, which is comparable to the spread of the graph) from the maximum, and the ratio of the height at the average to the height at the maximum will always be

$$\frac{P(\gamma)}{P(0)} = \frac{e^{-\gamma - e^{-\gamma}}}{e^{-0-e^{-0}}} \approx 0.87. \tag{25}$$