The Cereal Box Prize Distribution

Evan Ott
UT EID: eao466

April 15, 2017

Abstract

In this paper, I'll introduce the Cereal Box Prize Distribution (more formally, the multigeometric distribution) (or the Ott distribution if I'm feeling particularly narcissistic). First, I'll start with a simple case, and gradually abstract the distribution.

Motivation

I began working on this distribution in February 2016 on my statistics blog, Quick Ventures. The idea is to investigate how many boxes of cereal you have to buy to get all the prizes inside. More formally, if there are k prizes, with probabilities $\pi_1, \pi_2, \ldots, \pi_k$:

1. What's the probability of having all the prizes afterbuying n boxes?
2. Similarly, what are the mean, median, mode, and any other interesting quantities?

Prior Work

In my initial blog post, I considered this problem for $k = 2$ prizes. The easiest way to think about this conceptually is as a mixture of geometric distributions. We buy a single box, and see which prize is inside. If we find prize 1, then we now just need to find prize 2, and vice-versa. In other words, the following two definitions are equivalent:

$$X \sim \text{Multigeometric}((\pi_1, \pi_2)) = \text{Multigeometric}((\pi_1, 1 - \pi_1)) \tag{1}$$

$$X - 1 = Y \sim \pi_1 \text{Geometric}(1 - \pi_1) + (1 - \pi_1)\text{Geometric}(\pi_1) \tag{2}$$

The slight transformation in the second line is needed simply to account for having to open the first box. As such, last time, I showed that the expected value is:

$$\mathbb{E}[X] = 1 + \mathbb{E}[Y] = 1 + \pi_1 \frac{1}{1 - \pi_1} + (1 - \pi_1)\frac{1}{\pi_1} = \frac{1}{\pi_1(1 - \pi_1)} - 1 \tag{3}$$

This is quite an elegant solution for the mean. However, this doesn't provide much insight for more complex situations.

For example, take $k = 3$. It's not obvious how to construct a similar mixture distribution. For $k = 2$, we can benefit from the pigeonhole principle in that we will find one of the prizes in the first box, so all we have to do is wait to find the second prize. For $k = 3$ or greater, we have to handle a complicated infinitely tall ternary tree of boxes. And only after we find two prizes do we have it easy. That's where I stopped.

$k = 3$

I decided to jump ahead to figure out the $k = 3$ case. I could have computed the median, mode, etc. for $k = 2$ from the mixture distribution above, but seeing as I needed additional complexity, I thought going one step further would be illuminating.

Here, we can start with a similar problem. Let X be distributed according to a multinomial distribution with parameters n and $\vec{\pi} = (\pi_1, \pi_2, \pi_3)$. Using this formulation, we can answer the first motivating question. We need to compute $\mathbb{P}(X_1 \geq 1, X_2 \geq 1, X_3 \geq 1)$. We have:

$$\begin{aligned}
X &\sim \text{Multinomial}(n, \vec{\pi}) \\
f_X(X) &= \frac{n!}{X_1!X_2!X_3!}\pi_1^{X_1}\pi_2^{X_2}\pi_2^{X_2} \\
\mathbb{P}(X_1 \geq 1, X_2 \geq 1, X_3 \geq 1) &= 1 - \mathbb{P}(X_1 = 0) - \mathbb{P}(X_2 = 0) - \mathbb{P}(X_3 = 0) \\
&\quad + \mathbb{P}(X_1 = 0, X_2 = 0) + \mathbb{P}(X_1 = 0, X_3 = 0) + \mathbb{P}(X_2, X_3 = 0)
\end{aligned}$$

In other words, just do some combinatorics to see how often each component of X is zero, then make sure we don't double-count the times when two components are zero (because $X_1 + X_2 + X_3 = n$, at least one component is *always* non-zero).

For the scenarios where a single component is zero, the other two components live on the simplex $X_i + X_j = n$. That lets us compute the solution:

$$\begin{aligned}
\mathbb{P}(X_1 \geq 1, X_2 \geq 1, X_3 \geq 1) &= 1 - \sum_{X_1+X_2=n} f_X(X_1, X_2, 0) - \sum_{X_2+X_3=n} f_X(0, X_2, X_3) - \sum_{X_1+X_3=n} f_X(X_1, 0, X_3) \\
&\quad + f_X(0,0,n) + f_X(0,n,0) + f_X(n,0,0) \\
&= 1 - \sum_{X_1+X_2=n} \frac{n!}{X_1!X_2!}\pi_1^{X_1}\pi_2^{X_2} - \sum_{X_2+X_3=n} \frac{n!}{X_2!X_3!}\pi_2^{X_2}\pi_3^{X_3} \\
&\quad - \sum_{X_1+X_3=n} \frac{n!}{X_1!X_3!}\pi_1^{X_1}\pi_3^{X_3} + \pi_1^n + \pi_2^n + \pi_3^n \\
&= 1 - \sum_{X_1=0}^{n} \binom{n}{X_1}\pi_1^{X_1}\pi_2^{n-X_1} - \sum_{X_2=0}^{n} \binom{n}{X_2}\pi_2^{X_2}\pi_3^{n-X_2} \\
&\quad - \sum_{X_3=0}^{n} \binom{n}{X_3}\pi_1^{n-X_3}\pi_3^{X_3} + \pi_1^n + \pi_2^n + \pi_3^n
\end{aligned}$$

Now, for large values of n, this quantity starts becoming intractible. Note that the second through fourth terms above are *not* the PDF of a binomial distribution. They are slightly different in that any two components of $\vec{\pi}$ need not add up to one.

$\pi_1 + \pi_2 + \pi_3 = 1$

Perhaps unsurprisingly, when all three prizes are equally likely, the comutation of $\mathbb{P}(X_1 \geq 1, X_2 \geq 1, X_3 \geq 1)$ simplifies dramatically:

$$\begin{aligned}
\mathbb{P}(X_1 \geq 1, X_2 \geq 1, X_3 \geq 1) &= 1 - \sum_{X_1=0}^{n} \binom{n}{X_1}\frac{1}{3^n} - \sum_{X_2=0}^{n} \binom{n}{X_2}\frac{1}{3^n} - \sum_{X_3=0}^{n} \binom{n}{X_3}\frac{1}{3^n} + 3\frac{1}{3^n} \\
&= 1 - 3\frac{1}{3^n}\sum_{k=0}^{n}\binom{n}{k} + 3\frac{1}{3^n} \\
&= 1 - \frac{2^n}{3^{n-1}} + \frac{1}{3^{n-1}}
\end{aligned}$$

Because there's an implicit n in the equation above, let's change notation slightly to make it easier to work with:

$$F(n) = \mathbb{P}(X_1 \geq 1, X_2 \geq 1, X_3 \geq 1) = 1 - \frac{2^n - 1}{3^{n-1}}$$

where $F(n)$ is defined for any integer $n \geq k = 3$. However, we can note that as n tends toward infinity, $F(n)$ tends to one, and is never greater than 1. It seems an awful lot like a discrete version of a CDF. First, let's ensure monotonicity:

So, if we were to extend F to the whole real line with properties:

- $F(x) = 0$ for any $x < k$

- $F(x) = F(\lfloor x \rfloor)$

then F is actually a proper CDF. That indicates that we can find the PDF (defined on integers greater than or equal to 3, for simplicity):

$$f(n) = F(n) - F(n-1) = \frac{2^n - 1 - 2}{3^{n-1}}$$

And now, we have everything we could need for the equally-likely three-prize Cereal Box Prize distribution.

Let's take a look at some of its properties:

n	3	4	5	6	7	8
$f(n)$	2/9	2/9	14/81	10/81	62/729	14/243
$F(n)$	2/9	4/9	50/81	20/27	602/729	644/729