The Sticker Collector's Problem

M. A. Diniz, D. Lopes, A. Polpo, and L. E. B. Salasar




Márcio Alves Diniz (marcio.alves.diniz@gmail.com) is an assistant professor of statistics at Federal University of São Carlos (USFCar) in Brazil. He received his Ph.D. in statistics from the University of São Paulo and enjoys research in various fields. If you can relate (apparently) different problems mathematically, or solve a problem with a new a approach, then you will probably get his attention. **Danilo Lopes** (lopes@ufscar.br) is an assistant professor of statistics at USFCar. He received his Ph.D. in statistics at Duke University. Lopes likes to play video games and to listen to comedy podcasts and alternative rock songs. **Adriano Polpo** (polpo@ufscar.br) is an associate professor of statistics at USFCar. He received his Ph.D. in statistics from the University of São Paulo. He has been a visiting researcher at Florida State University, head of Department of Statistics at USFCar, and president of ISBrA, the Brazilian chapter of the International Society for Bayesian Analysis. Polpo likes to cook and enjoy good beer or wine. **Luis Ernesto Bueno Salasar** (luis.salasar@gmail.com) is an assistant professor of statistics at USFCar where he also received his Ph.D. He likes to play soccer and get together with his friends and family.

The FIFA World Cup is one of the biggest sports events in the world and occurs every four years. A related activity for many people (mainly children) is collecting stickers with photos of players of all the teams. The fun of buying, swapping, and playing games to get the missing stickers brings together friends and family. However, this can be a very expensive hobby.

Curious about the cost of this hobby, a journalist for the children's supplement of a major newspaper called one of the authors. She wanted to know the average number of packages needed to complete the collection, without considering swapping or other strategies. The collection has 649 stickers and each package has five stickers, guaranteed to be distinct. Being busy that morning, the author asked for some time and promised to call her later with the answer. After writing and running a simple Monte Carlo experiment in R, he provided the number 914. In the next day's newspaper, he found instead a number given by a professor from another university, 840.

Why the discrepancy? After sharing the problem with the other authors and double-checking the R code for the Monte Carlo estimate, we became more certain of 914 as the correct answer. Also, we soon realized that the question is almost as old as probability calculus; it is a generalization of the coupon collector's problem, first proposed by de Moivre in the early 18th century [**2**]. (See [**8**] for another generalization.)

This article is the result of our work on exploring different ways to solve the problem (and also understand why someone might suggest 840 as the answer). Specifically,

http://dx.doi.org/10.4169/college.math.j.47.4.255
MSC: 60G99

we consider the Monte Carlo approach, a "lottery" solution, and Markov chains. We believe that these various methods can be used as motivating examples in courses offered for undergraduate students of varying backgrounds, and also to motivate the curiosity about probability in younger students (principally using the Monte Carlo solution).

Monte Carlo solution

Monte Carlo simulations are widely used to find (approximate) solutions to probability problems, especially in statistical inference where closed solutions are not known or are very difficult to compute. We believe that Monte Carlo simulations have a good potential to motivate students in the classroom, especially by illustrating how elementary probability problems can be solved. The sticker collector's problem, as we call it, is particularly suitable for this.

The Monte Carlo solution is simply an algorithm: We simulate a number of stickers packages and count how many of them are necessary to fill one album (or collection) sequentially. After this procedure is repeated a given number of times, one simply computes the mean number of packages necessary to fill the albums.

The algorithm must reproduce the conditions described in the problem; here, the album has 649 stickers, each package has five distinct stickers, and the stickers are uniformly distributed among the packages. Its steps are as follows.

1. Let MyStickers be a vector with the labels (numbers) of stickers that you have. Initially it has length zero.
2. Take $k = 1$ where k is the number of packages.
3. Simulate one package by taking a sample of size five, without replacement, from an urn with 649 balls numbered from 1 to 649.
4. Check the stickers in the package of step 3 against those in MyStickers and include any stickers that are not yet present. Discard stickers you already have.
5. If MyStickers has length 649, then stop.
   Else increment k by one and go back to step 3.

Let T be the number of packages necessary to complete the collection. We are interested in the expected value of T so we simply repeat the above procedure for a great number of albums and take the arithmetic mean of the final values of k. To provide an example, we ran the algorithm for 10,000 "albums" and obtained 913.3 as an estimate of the expected number of packages necessary to complete one album. It should be stressed that, as a stochastic approximation, the algorithm generally yields a different answer each time it is performed (the answer 914 mentioned before comes from rounding to the smallest integer above 913.3).

From this work, an instructor could proceed to the concept of Monte Carlo error and how its estimate, the standard deviation of the generated samples, is a measure of the accuracy of the estimate of $E(T)$. As shown in Figure 1, one can plot the probability of needing more than k packages to complete the album.

Survival function

Deriving an exact solution to this problem illustrates the inclusion-exclusion principle and the combinatorics methods suitable for "equally probable" sampling schemes. This section is based on [**1, 4, 7**].

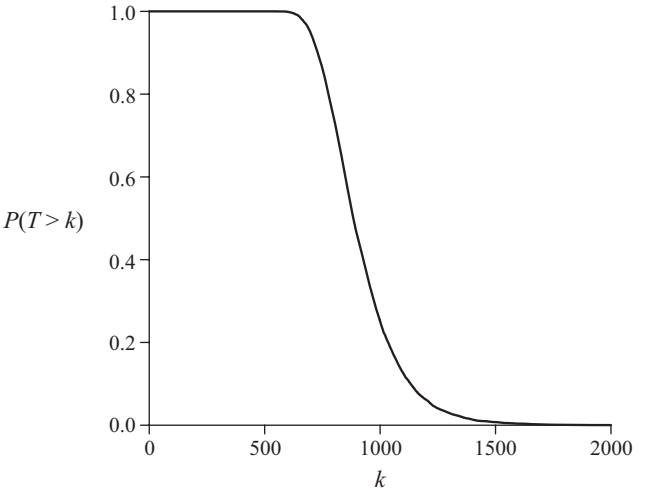

Figure 1. Probability that the album will be completed with more than k packages as estimated from 10,000 Monte Carlo simulations.

To simplify the notation, let N be the total number of stickers in a collection (649 in our case). The easiest way to proceed is to determine the probability that more than k packages are needed to complete the album, that is, the survival function $P(T > k)$. This event is equivalent to the event "at least one of the N stickers is not selected in the first k packages bought" which is equivalent to the following logical sum: "sticker 1 is not selected in the first k packages" or "sticker 2 is not selected in the first k packages" or ... or "sticker 649 is not selected in the first k packages." Denoting the event "sticker i is not selected in the first k packages" by $A_{i,k}$, we have

$$P(T > k) = P\left(\bigcup_{i=1}^{N} A_{i,k}\right).$$

Event $A_{i,k}$ occurs if and only if sticker i is not selected in the first package, nor in the second, ..., nor in the kth package. Therefore, using the independence between packages, we have $P(A_{i,k}) = p_1^k$ where

$$p_r = \binom{N-r}{n} \Big/ \binom{N}{n}$$

is the probability that no sticker from a subset of r stickers is found in the jth package with n stickers each. In our problem, $n = 5$, and we always assume that $n < N$. Note that p_r neither depends on the particular subset nor the particular package j.

For distinct elements $i_1, i_2, \ldots, i_s$ of $\{1, \ldots N\}$, we have

$$\begin{aligned} P(A_{i_1,k}) &= p_1^k, \\ P\left(A_{i_1,k} \cap A_{i_2,k}\right) &= p_2^k, \\ &\vdots \\ P\left(A_{i_1,k} \cap A_{i_2,k} \cap \cdots \cap A_{i_s,k}\right) &= p_s^k. \end{aligned}$$

With these probabilities we can compute $P(T > k)$: Using the inclusion-exclusion principle, we have, for $k \geq 0$,

$$\begin{aligned} P(T > k) &= P\left(\bigcup_{i=1}^{N} A_{i,k}\right) \\ &= \sum_{i_1=1}^{N} P(A_{i_1,k}) - \sum_{i_1 \neq i_2} P(A_{i_1,k} \cap A_{i_2,k}) + \cdots \\ &\quad + (-1)^{N-1} P(A_{1,k} \cap A_{2,k} \cap \cdots \cap A_{N,k}) \\ &= \binom{N}{1} p_1^k - \binom{N}{2} p_2^k + \cdots + (-1)^{N-1} p_N^k \\ &= \sum_{i=1}^{N} (-1)^{i-1} \binom{N}{i} p_i^k. \end{aligned} \tag{1}$$

For nonnegative discrete random variables, there is a simple relationship between the expectation and the survival function. With some algebra, one can show that

$$E(T) = \sum_{i=1}^{N} (-1)^{i-1} \binom{N}{i} \frac{1}{1 - p_i}. \tag{2}$$

Using Python, we easily approximated (2) to be 913.108. From (1) we can also find the probability function of T for $k \geq 1$:

$$\begin{aligned} P(T = k) &= P(T > k - 1) - P(T > k) \\ &= \sum_{i=1}^{N} (-1)^{i-1} \binom{N}{i} p_i^{k-1} (1 - p_i). \end{aligned} \tag{3}$$

The probability and the survival functions are very useful because they allow the computation of theoretical quantiles for T which can be compared to the quantiles found by the Monte Carlo experiment.

By (3), the probability function of T is a linear combination of probability functions of geometric random variables (with different probabilities of "failure," p_i). It is actually an affine combination, as the coefficients/weights sum to one. This problem can be compared with the coupon collector's problem where each package has just one sticker and T is a simple sum of geometric random variables. Hence, from (2), the moments of T are also a linear combination of the respective geometric moments.

Recursive calculation

Here we change our previous notation slightly and denote by T_i the number of packages needed to fill the album when we already have i different stickers. Let t_i be the expected value of T_i. We are particularly interested in t_0, but it might be helpful to find a general formula for any t_i. Trivially we have $t_N = 0$. Smaller initial collection sizes, however, require some work.

Let π_{ij} be the probability that, after buying one package, we go from i to j different stickers. Under the assumption that stickers are uniformly distributed among packages, π_{ij} is the hypergeometric probability given by

$$\pi_{ij} = \frac{\binom{N-i}{j-i}\binom{i}{n-j+i}}{\binom{N}{n}}$$

for $n \leq i \leq j \leq \min(i+n, N)$ (recall that n is the number of stickers in a package and N is the total number of stickers). Certainly our first package contains exactly n new stickers, so $\pi_{0n} = 1$. For any other values of the parameters, $\pi_{ij} = 0$ for the first package.

The law of total expectation should be stressed in a basic probability course: For any $M \in \mathbb{N}$, if $B_0, B_1, \ldots, B_M$ partition the sample space, then

$$t_i = E(T_i) = \sum_{j=0}^{M} E(T_i|B_j)P(B_j),$$

following from the law of total probabilities [**9**]. Setting B_j for $j = 0, \ldots, N$ as the event "after buying only one package when we already have i different stickers, we now have j different stickers" gives $P(B_j) = \pi_{ij}$ and hence

$$t_i = \sum_{j=i}^{\min(i+n,N)} E(T_i|B_j)\pi_{ij}.$$

Now $E(T_i|B_j)$ is the expected number of packages needed to fill the album when we start with i stickers and are sure that the next package will increase our initial collection to j stickers (or keep it at i stickers if $j = i$). This expectation equals the expected number of packages when we start with j stickers plus the first package, i.e., $E(T_i|B_j) = E(T_j + 1) = 1 + t_j$. Therefore,

$$t_i = \sum_{j=i}^{\min(i+n,N)} \pi_{ij}(1 + t_j) = 1 + \sum_{j=i}^{\min(i+n,N)} \pi_{ij}t_j.$$

Our quantity of interest, t_0, can be easily found in a recursive manner by solving the following linear system. (This can be considered as a discrete-time Markov chain.)

$$\begin{aligned}
&t_N = 0,\\
&t_i = \frac{1}{1-\pi_{ii}}\left(1 + \sum_{j=i+1}^{\min(i+n,N)} \pi_{ij}t_j\right) \qquad \text{for } n \leq i \leq N-1,\\
&t_0 = 1 + t_n.
\end{aligned}$$

Applying the hypergeometric transition probabilities to this linear system with $N = 649$ and $n = 5$ gives the sequence of values shown in Figure 2. The expected number of packages that will fill a blank album is the first value of this sequence, $t_0 = 913.108$. It is worth mentioning that all the moments of T_i, not only the first,

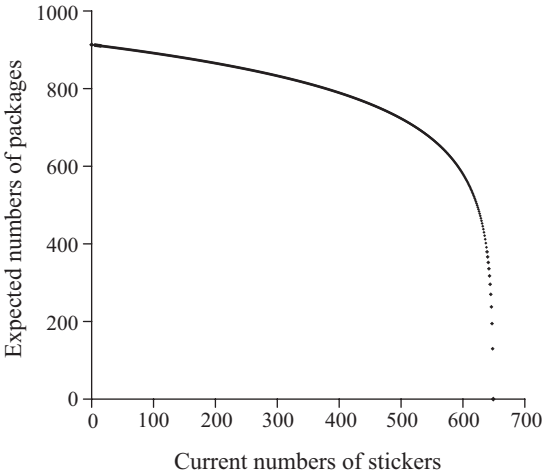

Figure 2. Expected number of packages needed to complete the album with $N = 649$ stickers as a function of the collection current size.

can be computed similarly, i.e., recursively solving a linear system of moments of the same order.

Similarly, we can find a recursive solution for $S_i(k) = P(T_i > k)$ for $k \geq 0$ and $i = 0, n, \ldots, N-1$, the survival function of T_i. It follows from the the law of total probability that

$$S_i(k) = P(T_i > k) = \sum_{j=i}^{\min(i+n,N)} P(T_i > k|B_j)\pi_{ij}.$$

The conditional probability $P(T_i > k|B_j)$ is the likelihood that, starting with i stickers, more than k packages are needed to complete the collection, given that the first package will increase our initial collection to j stickers. This conditional probability is $P(T_j > k-1)$ since after the first package is bought, there are j different stickers. Notice that $T_N = 0$, therefore $P(T_N > k-1) = 0$ for $k \geq 1$. Thus, for $n \leq i < N$ we have the following recursive formula:

$$S_i(k) = \sum_{j=i}^{\min(i+n,N)} P(T_j > k-1)\pi_{ij} = \sum_{j=i}^{\min(i+n,N-1)} S_j(k-1)\pi_{ij} \tag{4}$$

for $k \geq 1$ and the initial condition $S_i(0) = 1$.

For the case $i = N-1$ (i.e., only one sticker is missing), (4) reduces to $S_{N-1}(k) = (1-n/N)^k$ for $k \geq 1$ and, with the initial condition $S_{N-1}(0) = 1$, implies that

$$S_{N-1}(k) = \left(1 - \frac{n}{N}\right)^k$$

for $k \geq 0$. That is, T_{N-1} is a geometric random variable with probability of success n/N. In fact, T_{N-1} must be geometrically distributed since it can be regarded as the number of independent Bernoulli trials until the first success occurs (namely, the remaining sticker is collected).

For $i = 0$ we have $S_0(k) = S_n(k-1)$ if $k \geq 1$ and $S_0(0) = 1$, since the only possible transition starting from 0 is to n. Thus, in order to find $S_0(k)$, the survival function

of $T = T_0$, we need to solve (4) recursively for $i = N-1, N-2, \ldots, n$. Applying this backward procedure we find 731, 886, and 1129 for the 0.1, 0.5, and 0.9 quantiles of $T = T_0$, respectively. Figure 3 presents the 0.1, 0.5, and 0.9 quantiles for all the T_i. Note that the expected value of the number of packages needed to complete the collection is 913.108.

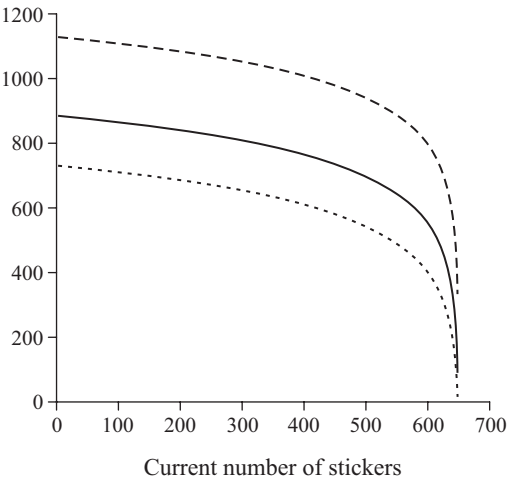

Figure 3. Quantiles 0.1 (dotted), 0.5 (solid), and 0.9 (dashed) for the additional number of stickers needed to complete the collection as a function of the collection's current size.

Asymptotic approximations

We have seen by Monte Carlo approximation and two exact methods that the answer to the reporter's question is 914. How is it that another professor gave the (unfortunately published) answer of 840?

In the original coupon collector's problem, each package has a single sticker. Thus, the number of packages T needed to complete the collection is a sum of geometric random variables with varying probability of finding a new sticker, and the expected value of this variable is simply the sum of the expected values of such geometric random variables,

$$E(T) = N\left(1 + \frac{1}{2} + \cdots + \frac{1}{N}\right).$$

The series inside parentheses is the Nth harmonic number. In our problem, since each package has five different stickers, an approximation is $E(T)/5$. What are approximations for the Nth harmonic number? Often it is roughly approximated by $\ln(N)$ and indeed, for $N = 649$, one computes $[649 \cdot \ln(649)]/5 \approx 840.51$. However, a different approximation is given by

$$N\left(1 + \frac{1}{2} + \cdots + \frac{1}{N}\right) \approx N\ln(N) + \gamma N + \frac{1}{2}$$

where $\gamma \approx 0.57721$ is the Euler–Mascheroni constant [**3**]. For $N = 649$, this approximation gives $E(T)/5 \approx 915.53$, much closer to the solutions found above.

Comparisons

Which of the methods we have discussed is preferable for given n and N? As expected, the exact methods have drawbacks: Using the survival function can push against a computer's limitations in representing large numbers and requires attention to avoid numerical issues. Using recursive calculation requires a more intricate derivation.

Table 1. Expected value of the number of packages (each with n distinct stickers) needed to complete the collection of 649 stickers given by approximation methods (Monte Carlo simulation, the log approximation for the harmonic numbers, the Euler–Mascheroni approximation for the harmonic numbers) and their respective absolute relative errors.

n	exact	MC	log	E–M	error MC	error log	error E–M
1	4577.7	4577.6	4202.6	4577.7	0.0000	0.0819	0.0000
5	913.1	913.3	840.5	915.5	0.0002	0.0795	0.0027
10	455.0	455.5	420.3	457.8	0.0011	0.0764	0.0060
61	72.0	72.1	68.9	75.0	0.0016	0.0432	0.0421
62	70.8	70.8	67.8	73.8	0.0005	0.0426	0.0429
100	42.7	42.8	42.0	45.8	0.0024	0.0154	0.0725
500	5.3	5.3	8.4	9.2	0.0024	0.5858	0.7273
649	1.0	1.0	6.5	7.1	0.0000	5.4754	6.0534

The Monte Carlo method is very simple to implement but it takes some time to evaluate; it is known as a "slow" procedure. Both asymptotic approximations are very simple to evaluate. Table 1 presents the exact values and results of the three approximations along with their absolute relative error for several values of n, the number of stickers in each package. The asymptotic approximations showed interesting results. The method presented in the newspaper (labeled log in Table 1) is not a good approximation if the number of stickers in the package is small when compared to the approximation with the Euler–Mascheroni constant, but this switches after 61 stickers per package. Neither of the asymptotic approximations is good when the number of stickers per package is large; Figure 4 shows the relative error of the three methods.

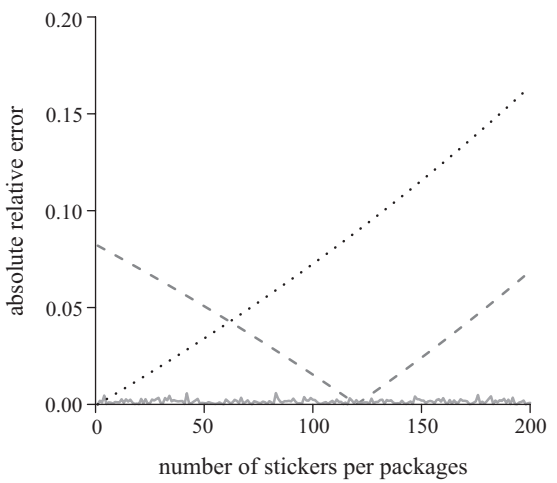

Figure 4. Absolute relative error for three approximation methods: Monte Carlo simulation (solid), the approximation with the Euler–Mascheroni constant (dotted), and the log approximation presented in the newspaper (dashed).

We recommend all of this material as suitable topics for basic and intermediary probability courses. Also, we advocate the Monte Carlo approach for computational statistics courses, whose literature, with a few exceptions [**5, 6**], thus far lacks this kind of example to motivate teaching algorithms and simulation methods.

Acknowledgment. The authors wish to thank their department colleagues who motivated them to write this paper. Adriano Polpo thanks CNPq-Brazil (308776/2014-3) for financial support.

Summary. We present a generalization of the coupon collector's problem called the sticker collector's problem. We cover four different ways to handle the problem, illustrating the results with the stickers of the FIFA World Cup album. This material could be used as motivating examples in undergraduate courses.

References

1. I. Adler, S. M. Ross, The coupon subset collection problem. *J. Appl. Probab.* **38** (2001) 737–746, http://dx.doi.org/10.1239/jap/1005091036.
2. A. de Moivre, De Mensura Sortis seu; de Probabilitate Eventuum in Ludis a Casu Fortuito Pendentibus, *Philos. Trans.* **27** (1710) 213–264, http://dx.doi.org/10.1098/rstl.1710.0018.
3. L. Euler, De Progressionibus Harmonicis Observationes, *Comm. Acad. Sci. Petrop.* **7** (1740) 150–161, http://eulerarchive.maa.org/pages/E043.html.
4. J. G. Leite, C. A. B. Pereira, F. W. Rodrigues, Waiting time to exhaust lottery numbers. *Comm. Statist. Theory Methods* **22** (1993) 301–310, http://dx.doi.org/10.1080/03610929308831019.
5. P. J. Nahin, *Dueling Idiots and Other Probability Puzzlers*. Princeton Univ. Press, Princeton, 2000.
6. ———, *Digital Dice: Computational Solutions to Practical Probability Problems*. Princeton Univ. Press, Princeton, 2008.
7. W. Stadje, The collector's problem with group drawings. *Adv. Appl. Probab.* **22** (1990) 866–882, http://dx.doi.org/10.2307/1427566.
8. H. von Schelling, Coupon collecting for unequal probabilities. *Amer. Math. Monthly* **61** (1954) 306–311, http://dx.doi.org/10.2307/2307466.
9. N. A. Weiss, *A Course in Probability*. Addison–Wesley, Boston, 2006.