# Collecting coupons — A mathematical approach

John S. Croucher
*Macquarie University*
<john.croucher@gsm.mq.edu.au>

## Introduction

A special but common type of scenario is one in which a company has a promotion that is designed to make the customer purchase more of their product than they otherwise might. Although this can be aimed specifically at children, it really applies to all persons. The basic premise is that the company issues a "set" of different items or coupons and places one of the coupons in boxes of their product. The consumer does not know which of the coupons in the set they will get until they purchase the product and open the packaging. The questions of interest here are:

- What is the expected number of boxes that must be purchased before the complete set of coupons is obtained?
- What are the chances of obtaining the complete set for a given number of purchases?

This situation is sometimes known as the "coupon collector's problem" or "cereal box problem" (since the coupons are often a set of toys found in a packet of cereal) and the aim here is to analyse it generally and then demonstrate by using specific examples. It provides a most interesting instance of how mathematics can be used to analyse an everyday problem while illustrating the important concept of modelling. Specific instances of its application include coin collecting by Lu and Skiena (2000).

## Formulation of the problem

Suppose that a company decides to issue a set that consists of $n$ different coupons. These coupons are distributed equally among packets of its product so that if a single packet is purchased there is a $1/n$ chance it will contain a particular coupon. The key here is that the statistical distribution applicable is the *geometric distribution*, this being the one that deals with the expected time until the first success.

Formally, if there is a series of independent trials of process in which the

probability of a "success" at any one trial is a constant $p$, then the time to achieving the first "success" is given by the geometric distribution with parameter values:

$$\mu = \frac{1}{p} \tag{1}$$

$$\sigma^2 = \frac{1-p}{p} \tag{2}$$

In our problem, the first packet purchased will always contain one of the coupons. For each subsequent packet, there is a

$$\frac{n-1}{n}$$

chance that the coupon found will be different than the first, and so from (1) the expected number of packets that must be purchased to find a different coupon is

$$\frac{n}{n-1}$$

Once two coupons in the set are obtained, there is now a

$$\frac{n-2}{n}$$

chance that a subsequent packet will obtain a coupon different from the first two coupons.

In a similar way, from (1) the expected number of packets that must be purchased to find a different coupon is

$$\frac{n}{n-2}$$

The expected total number of packets, $E(n)$, that must be purchased to obtain the entire set of $n$ coupons is given by:

$$\begin{aligned} E(n) &= \frac{n}{n} + \frac{n}{n-1} + \frac{n}{n-2} + \frac{n}{n-3} + \ldots \frac{n}{2} + \frac{n}{1} \\ &= n\sum_{i=1}^{n} \frac{1}{i} \end{aligned} \tag{3}$$

It follows from (3) that:

$$\begin{aligned} E(n+1) &= (n+1)\sum_{i=1}^{n+1} \frac{1}{i} \\ &= \left(\frac{n+1}{n}\right) E(n) + 1 \end{aligned} \tag{4}$$

## Expected number of boxes to purchase

The recursive relationship in (4) makes it easy to calculate values of $E(n)$ for various values of $n$. Since, for example, $E(1) = 1$, then $E(2) = 2E(1) + 1 = 3$ and $E(3) = 1.5E(2) + 1 = 5.5$. Table 1 shows the values of $E(n)$ for values of $n$ up to 12.

*Table 1: Values of E(n), the expected number of boxes to be purchased.*

| $n$ | E($n$) |
|---|---|
| 1 | 1.0 |
| 2 | 3.0 |
| 3 | 5.5 |
| 4 | 8.3 |
| 5 | 11.4 |
| 6 | 14.7 |
| 7 | 18.2 |
| 8 | 21.7 |
| 9 | 25.5 |
| 10 | 29.3 |
| 11 | 33.2 |
| 12 | 37.2 |

While the values in Table 1 could be continued indefinitely, for larger values of $n$ we can use *Euler's approximation* for the partial sum of a harmonic series. That is:

$$\sum_{i=1}^{n} \frac{1}{i} \approx \log_e n + 0.5772 \qquad (5)$$

Substituting (5) into (3), an approximate value of E($n$) is given by:

$$\mathrm{E}(n) = n[\log_e n + 0.5772] \qquad (6)$$

The values in (6) can be easily found using a pocket calculator. To test the accuracy of the approximation, Table 2 shows the values given by (6) for $n = 1$ to 12 and a comparison can be made with the exact values in Table 1.

*Table 2. Approximate values of E(n) for n = 1 to 12.*

| $n$ | Approximate value of E($n$) |
|---|---|
| 1 | 0.6 |
| 2 | 2.5 |
| 3 | 5.0 |
| 4 | 7.9 |
| 5 | 10.9 |
| 6 | 14.2 |
| 7 | 17.7 |
| 8 | 21.3 |
| 9 | 25.0 |
| 10 | 28.8 |
| 11 | 32.7 |
| 12 | 36.7 |

By comparing the values in Tables 1 and 2, it can be seen that for values of $n$ greater than 2 the approximation is quite good. In fact, the absolute error made in estimating the value of $n$ in no case exceeds 0.50 (and in each case *underestimates* the true value). Using (6) we can also easily find excellent approximate values of E($n$) for large values of $n$. For example, E(20) ≈ 71 (exact is 71.95), E(50) ≈ 224 (exact is 225.0) and E(100) ≈ 518 (exact is 518.7). In both of these cases the absolute error does not exceed 1, again underestimating the true value.

## Variation and probability

As well as calculating the expected number of boxes to purchase in order to collect all coupons, it is also useful to determine the variation of this number. To do this, we consider the variance of the geometric distribution given in (2) and apply it to the case where n coupons are to be collected. The variance in the sum of the time taken is the sum of the variances of the individual times. This yields:

$$\begin{aligned}\text{Variance of total time} = \sigma^2 &= \frac{\left(1-\frac{n-1}{n}\right)}{\left(\frac{n-1}{n}\right)} + \frac{\left(1-\frac{n-2}{n}\right)}{\left(\frac{n-2}{n}\right)} + \ldots + \frac{\left(1-\frac{1}{n}\right)}{\left(\frac{1}{n}\right)} \\ &= \frac{1}{n-1} + \frac{2}{n-2} + \frac{3}{n-3} + \ldots + \frac{n-1}{1} \\ &= \sum_{i=1}^{n-1} \frac{i}{n-i} \qquad (7)\end{aligned}$$

From (7), the variance $\sigma^2$ and hence the value of the standard deviation $\sigma$ (by taking the square root) can easily be found. The values of $\mu$ and $\sigma$ for a selection of values of $n$ is shown in Table 3.

*Table 3. The mean and standard deviation of the number of packets that must be bought for several values of n.*

| $n$ | $\mu$ | $\sigma$ |
|---|---|---|
| 5 | 11.4 | 2.53 |
| 10 | 29.3 | 4.32 |
| 20 | 72.0 | 7.21 |
| 30 | 119.8 | 9.48 |
| 50 | 225.0 | 13.23 |

The values in Table 3 can be used to calculate the approximate probability that the number of boxes that will have to purchased to collect all $n$ coupons will be less than or exceed a specified value. This involves using the *Central Limit Theorem*, which suggests the total number of boxes that must be purchased follows an approximate normal distribution with the mean and standard deviation values shown in Table 3. The approximation works better for large values of $n$ (say, at least 20) but still gives a rough estimate for lower values. In particular, a 95% confidence interval for the number of boxes can be found by taking 1.96 standard deviations either side of the mean. The results are shown (to the nearest integer) in Table 4.

*Table 4. Upper and lower bounds on the number of packets that must be bought for several values of n.*

| $n$ | Lower bound | Average | Upper bound |
|---|---|---|---|
| 5 | 6 | 11 | 16 |
| 10 | 21 | 29 | 38 |
| 20 | 58 | 72 | 86 |
| 30 | 101 | 120 | 138 |
| 50 | 199 | 225 | 251 |

For example, from Table 4 it follows that if 10 coupons are to be collected, the average number of boxes that will need to be purchased is about 29. However, you would be lucky to get them all in less than 21 boxes and unlucky if it took you more than 38 boxes.

## Remarks

The above problem has many applications to real life where there are those who collect items as a hobby, including stamp and coin collectors who wish to complete a set. It is also relevant for those companies who reward all those who complete a set of coupons that may be redeemed for a prize. In some cases there may be a competition for the first people to obtain a complete set, a situation that can lead to swapping of coupons for those who have multiples of the same one. In such circumstances, however, it is quite likely that there is not an equal number of coupons in the boxes, with the "rare" ones only appearing in a few of them. There have been several introductory papers written on this type of problem such as those by Litwiller and Duncan (1992), Berry and Maull (1992) and Wilkins (1999). An interesting variation is provided by Ilan and Ross (2000), in that rather than collecting individual coupons at each time point they obtain a random subset of coupons. The problem of interest here is to determine the expected number of subsets needed until each coupon is contained in at least one of these subsets. A more advanced consideration is given by Myers and Wilf (2003), who calculate the probability that two competing collectors will complete the set with the same number of purchases. Other, more complex, variations are provided by Baum and Billingsley (1969) and Sheutzow (2002).

## References


Adler, I., Oren, S. & Ross, S., (2003). The coupon collector's problem revisited. *Journal of Applied Probability, 40*(2), 513–518.

Baum, L. E. & Billingsley, P. (1969). Asymptotic distributions for the coupon collector's problem. *Annals of Mathematical Statistics, 36*, 1835–1839.

Berry, J. & Maull, W. (1997). Modeling the coupon collector's problem. *Teaching Statistics, 19*(2), 43–46.

Litwiller, B. H. & Duncan, D. R. (1992). Prizes in cereal boxes: An application of probability. *School Science and Mathematics, 92*(4), 193–195.

Lu, S. & Skiena, S. (2000). Filling a penny album. *Chance, 13*(2), 25–28.

Myers, A. N. & Wilf, H. S. (2003). Some new aspects of the coupon collector's problem. *SIAM Journal of Discrete Mathematics, 17*(1), 1–17.

Sheutzow, M. (2002). Asymptotics for the maximum in the coupon collector's problem. *Math. Scientist, 27*, 85–90.

Wilkins, J. L. M. (1999). Cereal box problem revisited. *School Science and Mathematics, 99*(3), 193–195.