

Longest waiting time in the coupon collector's problem

Brennan C.^{*1}, Kariv, J.² and Knopfmacher, A.³

¹ The John Knopfmacher Centre for applicable Analysis and Number Theory,
University of the Witwatersrand, Wits 2050, Johannesburg, South Africa

² School of Statistics and Actuarial Science,

University of the Witwatersrand, Wits 2050, Johannesburg, South Africa

³ The John Knopfmacher Centre for applicable Analysis and Number Theory,
University of the Witwatersrand, Wits 2050, Johannesburg, South Africa

**Original Research
Article**

Received: XX December 20XX

Accepted: XX December 20XX

Online Ready: XX December 20XX

Abstract

The coupon collector's problem is a classical problem usually solved by summing independent but not identically distributed geometric random variables. These geometric random variables represent waiting times for the j -th coupon for $j = 1, 2, \dots$. In this paper we address the question about the longest of these waiting times. In particular, we evaluate the moments of the longest waiting time.

Keywords: coupon collector's problem, maximum, moments

2010 Mathematics Subject Classification: 60C05

1 Introduction

The coupon collector's problem is a well known problem in probability theory. Suppose a certain commercial brand distributes a coupon in every box of a particular product where there are in total n distinct coupons. A coupon collector wishes to collect all n coupons and does not know which coupon is in any particular box. The standard problem is to determine the expected total number of boxes to collect all coupons. To find the total waiting time, the usual method is to suppose that the collector already has $i - 1$ distinct coupons, he now wishes to obtain the i -th distinct coupon. We call the number of boxes he opens before he collects the i -th coupon X_i . Clearly X_i is a random geometric variable with $p = \frac{n-i+1}{n}$. The expected waiting time to collect all the coupons is nH_n where H_n is the n -th harmonic number and asymptotically this equals $n(\log n + \gamma)$. This is the standard coupon collecting problem.

Brennan, C.: E-mail: Charlotte.Brennan@wits.ac.za

The problem was first mentioned in 1709 by de Moivre in his collection of 26 problems related to games of chance titled *De Mensura Sortis, seu, de Probabilitate Eventuum in Ludis a Casu Fortuito Pendentibus*. Kendall and Babington-Smith (1938) mentioned the problem in relation to checking the randomness of their sampling numbers, Feller (1968) presented the question as a type of urn problem. Some new aspects to this problem can be found in Adler et al. (2003); Gadidov and Thomas (2015); Myers and Wilf (2003). We let M be the maximum value in the sequence (X_i) . The interpretation of this is the longest time one needs to wait until obtaining a new coupon. In this paper we will find the expected value, the variance and higher moments of M .

Some relevant references on samples of independent identically distributed geometric random variables and their applications can be found in Archibald et al. (2006); Grabner et al. (2000) and Kirschenhofer and Prodinger (1990, 1993); Louchard and Prodinger (2008) and Prodinger (1996); Szpankowski and Rego (1990).

2 Expected value of the longest waiting time

Let M be the longest waiting time, i.e., let $M = \max(X_i)$. In this section, we compute $\mathbb{E}[M]$, the expected value of M .

Theorem 2.1. *The expected value of the longest waiting time $\mathbb{E}[M]$ is asymptotic to*

$$\sum_{j=1}^{\infty} \left(\frac{2}{j(3j-1)} + \frac{2}{j(3j+1)} \right) n,$$

which is equal to $1.25519745 \dots n$.

Proof. For this we use the standard formula for the expected value

$$\mathbb{E}[M] = \sum_{i=0}^{\infty} P(M > i).$$

Thus

$$\begin{aligned} \mathbb{E}[M] &= \sum_{i=0}^{\infty} (1 - P(M \leq i)) \\ &= \sum_{i=0}^{\infty} \left(1 - \prod_{j=1}^n P(X_j \leq i) \right) \\ &= \sum_{i=0}^{\infty} \left(1 - \prod_{j=1}^n \sum_{k=1}^i \frac{n-j+1}{n} \left(\frac{j-1}{n} \right)^{k-1} \right) \\ &= \sum_{i=0}^{\infty} \left(1 - \prod_{j=1}^n \left[1 - \left(\frac{j-1}{n} \right)^i \right] \right) \\ &= \sum_{i=0}^{\infty} \left(1 - \prod_{j=0}^{n-1} \left[1 - \left(\frac{j}{n} \right)^i \right] \right) \end{aligned} \tag{2.1}$$

$$:= \sum_{i=0}^{\infty} f(i, n), \tag{2.2}$$

where

$$f(i, n) := 1 - \prod_{j=0}^{n-1} \left[1 - \left(\frac{j}{n} \right)^i \right]. \tag{2.3}$$

We proceed using the substitution $\alpha = \frac{j}{n}$ and defining the function $g(\alpha) := \lim_{n \rightarrow \infty} f(\alpha n, n)$, thus

$$\begin{aligned} g(\alpha) &= \lim_{n \rightarrow \infty} \left[1 - \prod_{j=0}^{n-1} \left[1 - \left(\frac{j}{n} \right)^{\alpha n} \right] \right] \\ &= \lim_{n \rightarrow \infty} \left[1 - \prod_{j=1}^n \left[1 - \left(\frac{n-j}{n} \right)^{\alpha n} \right] \right] \\ &= 1 - \prod_{j=1}^{\infty} [1 - e^{-j\alpha}]. \end{aligned} \tag{2.4}$$

The convergence of $f(\alpha n, n)$ to $g(\alpha)$ is fast and is illustrated below

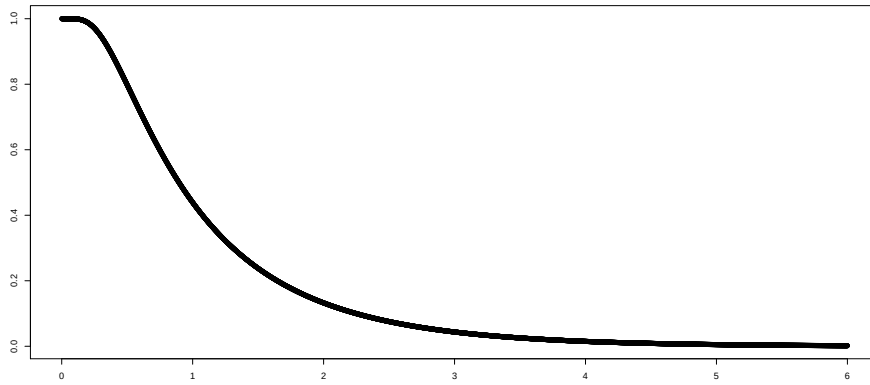


Figure 1: Graph of $f(\alpha n, n)$ against α for $n = 10$

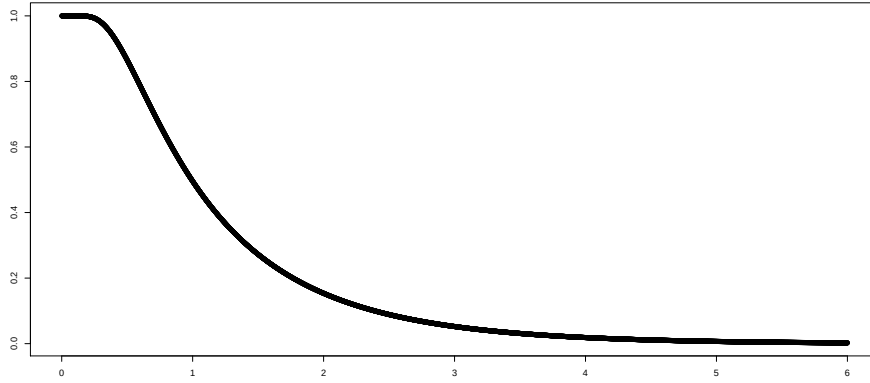


Figure 2: Graph of $g(\alpha)$ against α

To evaluate the above expression we use the following lemma

Lemma 2.2. *Euler's pentagonal number theorem, (see Andrews (1998), Corollary 1.7 p11 and Comtet (1974) for proof).*

$$\prod_{n=1}^{\infty} (1 - q^n) = \sum_{j=0}^{\infty} (-1)^j q^{\frac{j(3j-1)}{2}} + \sum_{j=0}^{\infty} (-1)^j q^{\frac{j(3j+1)}{2}} = 1 + \sum_{j=1}^{\infty} (-1)^j q^{\frac{j(3j-1)}{2}} (1 + q^j).$$

We use the lemma to obtain

$$\begin{aligned} g(\alpha) &= 1 - \prod_{j=1}^{\infty} [1 - e^{-j\alpha}] \\ &= \sum_{j=1}^{\infty} (-1)^{j+1} e^{-\alpha \frac{j(3j-1)}{2}} (1 + e^{-\alpha j}) \\ &= \sum_{j=1}^{\infty} (-1)^{j+1} e^{-\alpha \frac{j(3j-1)}{2}} + \sum_{j=1}^{\infty} (-1)^{j+1} e^{-\alpha \frac{j(3j+1)}{2}}. \end{aligned} \quad (2.5)$$

Recall

$$\mathbb{E}[M] = \sum_{i=0}^{\infty} f(i, n).$$

Thus

$$\lim_{n \rightarrow \infty} \frac{\mathbb{E}[M]}{n} = \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=0}^{\infty} f(i, n) = \int_0^{\infty} g(\alpha) d\alpha.$$

Therefore, we compute

$$\begin{aligned} \int_0^\infty g(\alpha) d\alpha &= \sum_{j=1}^\infty (-1)^{j+1} \int_0^\infty \left[e^{-\alpha \frac{j(3j-1)}{2}} + e^{-\alpha \frac{j(3j+1)}{2}} \right] d\alpha \\ &= \sum_{j=1}^\infty (-1)^{j+1} \left[\frac{2}{j(3j-1)} e^{-\alpha \frac{j(3j-1)}{2}} + \frac{2}{j(3j+1)} e^{-\alpha \frac{j(3j+1)}{2}} \right]_0^\infty \\ &= \sum_{j=1}^\infty \left(\frac{2}{j(3j-1)} + \frac{2}{j(3j+1)} \right). \end{aligned} \quad (2.6)$$

This expression tends to the value $1.25519745 \dots$. □

3 Variance of the longest waiting time

In this section, we find the variance of the longest waiting time. Our result is obtained from

$$\mathbb{E}[M^2] - (\mathbb{E}[M])^2$$

and asymptotically we have

Theorem 3.1. *The variance of the longest waiting time $Var(M)$ is asymptotically equal to $0.82132279 \dots n^2$ as $n \rightarrow \infty$.*

Proof. We first find $\mathbb{E}[M^2]$, the second moment about the origin. This is given by the well known formula,

$$\mathbb{E}[M^2] = \sum_{i=0}^\infty (2i+1)P(M > i).$$

So in our case,

$$\begin{aligned} \lim_{n \rightarrow \infty} \frac{\mathbb{E}[M^2]}{n^2} &= \lim_{n \rightarrow \infty} \sum_{i=0}^\infty \frac{2i+1}{n^2} P(M > i) \\ &= \lim_{n \rightarrow \infty} \left[\sum_{i=0}^\infty \frac{2i}{n^2} P(M > i) + \sum_{i=0}^\infty \frac{1}{n^2} P(M > i) \right] \\ &= \lim_{n \rightarrow \infty} \sum_{i=0}^\infty \frac{2i}{n^2} P(M > i). \end{aligned}$$

We proceed with our calculations using a similar argument to the one in Section 2 and obtain

$$\begin{aligned} \lim_{n \rightarrow \infty} \frac{\mathbb{E}[M^2]}{n^2} &= \int_0^\infty 2\alpha g(\alpha) d\alpha \\ &= \sum_{j=1}^\infty (-1)^{j+1} \int_0^\infty \left[2\alpha e^{-\alpha \frac{j(3j-1)}{2}} + 2\alpha e^{-\alpha \frac{j(3j+1)}{2}} \right] d\alpha \\ &= 8 \sum_{j=1}^\infty (-1)^{j+1} \left(\frac{1}{j^2(3j-1)^2} + \frac{1}{j^2(3j+1)^2} \right) \end{aligned} \quad (3.1)$$

which numerically tends to $2.396843429 \dots$.

Using (3.1) – (2.6)² we obtain

$$8 \sum_{j=1}^{\infty} (-1)^{j+1} \left(\frac{1}{j^2(3j-1)^2} + \frac{1}{j^2(3j+1)^2} \right) - \left(\sum_{j=1}^{\infty} \left(\frac{2}{j(3j-1)} + \frac{2}{j(3j+1)} \right) \right)^2$$

which is asymptotically equal to $0.82132279 \cdots n^2$ as $n \rightarrow \infty$. $\square$

4 Higher moments of the longest waiting time

In this section, we find the higher moments for the longest waiting time. We use the formula for the k -th moment

$$\mathbb{E}[M^k] = \sum_{i=0}^{\infty} ((i+1)^k - i^k) P(M > i).$$

So in our case

$$\begin{aligned} \lim_{n \rightarrow \infty} \frac{\mathbb{E}[M^k]}{n^k} &= \int_0^{\infty} k \alpha^{k-1} g(\alpha) d\alpha \\ &= \sum_{j=1}^{\infty} (-1)^{j+1} \int_0^{\infty} \left[k \alpha^{k-1} e^{-\alpha \frac{j(3j-1)}{2}} + k \alpha^{k-1} e^{-\alpha \frac{j(3j+1)}{2}} \right] d\alpha. \end{aligned} \quad (4.1)$$

In order to obtain the above integral we first compute

$$\int_0^{\infty} \alpha^k e^{-b\alpha} d\alpha = \frac{1}{b^k} \int_0^{\infty} (b\alpha)^k e^{-k\alpha} d\alpha = \frac{k!}{b^{k+1}}.$$

Thus our expression in (4.1) yields

Theorem 4.1. *The k -th moment of the longest waiting times satisfies*

$$\lim_{n \rightarrow \infty} \frac{\mathbb{E}[M^k]}{n^k} = \sum_{j=1}^{\infty} (-1)^{j+1} \frac{k! 2^k}{j^k} \left[\frac{1}{(3j-1)^k} + \frac{1}{(3j+1)^k} \right].$$

So for example for $k = 3$ and 4 we have

$$\sum_{j=1}^{\infty} (-1)^{j+1} \left[\frac{48}{j^3(3j-1)^3} + \frac{48}{j^3(3j+1)^3} \right]$$

and

$$\sum_{j=1}^{\infty} (-1)^{j+1} \left[\frac{384}{j^4(3j-1)^4} + \frac{384}{j^4(3j+1)^4} \right]$$

respectively.

In the table below we list the approximate numerical values for the moments about the origin, moments about the mean, and normalised moments about the mean, for $k = 1, 2, 3, 4, 5$ where σ is the standard deviation of M :

k	$\lim_{n \rightarrow \infty} \frac{\mathbb{E}[M^k]}{n^k}$	$\lim_{n \rightarrow \infty} \frac{\mathbb{E}[(M-\mu)^k]}{n^k}$	$\lim_{n \rightarrow \infty} \frac{\mathbb{E}[(M-\mu)^k]}{\sigma^k n^k}$
1	1.25520	0	0
2	2.39684	0.82133	1
3	6.68903	1.61867	2.17465
4	25.45310	7.07975	10.49519
5	123.70507	34.41198	56.28919

Table 1: Summary of numerical values of moments

4.1 Remarks on the distribution

We observe that $\frac{M}{n}$ tends to a particular distribution which is the distribution of the maximum of a collection of independent exponential random variables with means $1, \frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \dots, \frac{1}{k}$ and so on.

To see this, note that the last waiting time is a geometric random variable with mean n and if we divide by n it is well approximated by an exponential random variable with mean 1. Similarly the second last waiting time is geometric random variables with mean $\frac{n}{2}$ and again after dividing by n it is approximated by an exponential with mean $\frac{1}{2}$. The pattern is repeated for the third last waiting time and so on.

References

- Adler, I. Oren S. and Ross S. (2003). The coupon-collector's problem revisited, *Journal of Applied Probability*, 40, 513–518.
- Andrews, G. (1998). *Theory of partitions*, *Encyclopedia of Mathematics and its applications*, Addison-Wesley, vol. 2.
- Archibald, M. Knopfmacher, A. and Prodinger, H. (2006). The number of distinct values in a geometrically distributed sample. *European Journal of Combinatorics*, 27, 1059–1081.
- Comtet, C. (1974). *Advanced Combinatorics*, Reidel.
- Feller, W. (1968). *An introduction to probability theory and its applications*. I, Wiley, New York.
- Gadidov, A. and Thomas, M. (2015). The coupon collector problem: The case of two collections, *Collaborative Mathematics and Statistics Research*, Springer proceedings in Mathematics and Statistics, 109, 21–29.

- Grabner, P. Knopfmacher, A. and Prodinger, H. (2000). Combinatorics of geometrically distributed random variables: Run statistics. *LATIN 2000, Lecture Notes in Computer Science*, 1776, 457–462.
- Kendall, M. and Babington-Smith, B. (1938). Randomness and random sampling numbers, *J. R. Stat. Soc.*, 101, 147–166.
- Kirschenhofer, P. and Prodinger, H. (1990). On the analysis of probabilistic counting. *Lecture Notes in Mathematics, Number-theoretic Analysis*, 1452, 117–120.
- Kirschenhofer, P. and Prodinger, H. (1993). A result in order statistics related to probabilistic counting. *Computing*, 51, 15–27.
- Louchard, G. and Prodinger, H. (2008). On gaps and unoccupied urns in sequences of geometrically distributed random variables. *Discrete Math*, 308, 1538–1562.
- Myers, A. and Wilf, H. (2003). Some new aspects of the coupon collector's problem, *SIAM Journal of Discrete Mathematics*, 17(1), 1–17.
- Prodinger, H. (1996). Combinatorics of geometrically distributed random variables: Left-to-right maxima. *Discrete Mathematics*, 153, 253–270.
- Szpankowski, W. and Rego, V. (1990). Yet another application of a binomial recurrence. *Order statistics. Computing*, 43, 401–410.

©2011 Brennan, C. Kariv, J. & Knopfmacher, A.; This is an Open Access article distributed under the terms of the Creative Commons Attribution License <http://creativecommons.org/licenses/by/2.0>, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.