MAT^2

MATerials MATemàtics

Versió per a e-book del

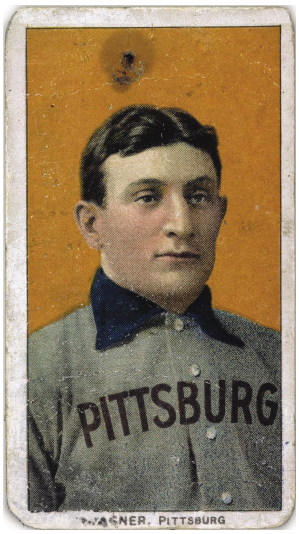
treball no. 2 del volum 2014

www.mat.uab.cat/matmat

The Coupon Collector's Problem

Marco Ferrante

Monica Saltalamacchia



In this note we will consider the following problem: how many coupons we have to purchase (on average) to complete a collection. This problem, which takes everybody back to his childhood when this was really "a problem", has been considered by the probabilists since the eighteenth century and nowadays it is still possible to derive some new results, probably original or at least never published. We will present some classic results, some new formulas, some alternative approaches to obtain known results and a couple of amazing expressions.

1 History

The coupon collector's problem is a classical problem in combinatorial probability. Its description is easy: consider one person that collects coupons and assume that there is a finite number, say N, of different types of coupons, that for simplicity we denote by the numbers $1, 2, \ldots, N$. These items arrive one by one in sequence, with the type of the successive items being independent random variables that assume the value k with probability p_k. When the probabilities p_k are constant (the equal probabilities case) we will usually face an easier problem, while when these probabilities are unequal the problem becomes more challenging, even if more realistic too.

Usually one is interested in answering the following questions: which is the probability to complete the collection (or a given subset of the collec-

tion) after the arrival of exactly n coupons ($n \geq N$)? Which is the expected number of coupons that we need to collect in order to complete the collection? How these probabilities and expected values change if we assume that the coupons arrive in groups of constant size or we are considering a group of friends that intends to complete m collections? In this note we will consider the problem of more practical interest, which is the (average) number of coupons that one needs to purchase in order to complete one or more than one collection in both the cases of equal and unequal probabilities. We will obtain explicit formulas for these expected values and, as suggested by the intuition, that the minimum expected number of purchases is needed in the equal case, while in the unequal case a very rare coupon can bring this expected number to tend to infinity.

We also present some approximation formulas for the results obtained since, most of all in the unequal probabilities case, even if the exact formulas appear easy and compact, they are computationally extraordinarily heavy. Some of the results in this note are probably original or at least never published before.

The history of the coupon collector's problem began in 1708, when the problem first appeared in *De Mensura Sortis (On the Measurement of Chance)* written by A. De Moivre. More results, due among others to Laplace and Euler (see [8] for a comprehensive introduction on this topic), were obtained in the case of constant probabilities, i.e. when $p_k \equiv \frac{1}{N}$ for any k.

In 1954 H. Von Schelling [10] first obtained the waiting time to complete a collection when the probability of collecting each coupon wasn't equal

and in 1960 D. J. Newman and L. Shepp [7] calculated the waiting time to complete two collections of coupons in the equal case. More recently, some authors have made further contribution to this classical problem (see e.g. L. Holst [4] and Flajolet et. al. [3]).

The coupon collector's problem has many applications, especially in electrical engineering, where it is related to the cache fault problem and it can be used in electrical fault detection, and in biology, where it is used to estimate the number of species of animals.

2 Single collection with equal probabilities

Assume that there are N different coupons and that they are equally likely, with the probability to purchase any type at any time equal to $\frac{1}{N}$. In

this section we will derive the expected number of coupons that one needs to purchase in order to complete the collection. We will present two approaches, the first of which presents in most of the textbooks in Probability, and we derive a simple approximation of this value.

2.1 The Geometric Distribution approach

Let X denote the (random) number of coupons that we need to purchase in order to complete our collection. We can write $X = X_1 + X_2 + \ldots + X_N$, where for any $i = 1, 2, \ldots, N$, X_i denotes the additional number of coupons that we need to purchase to pass from $i-1$ to i different types of coupons in our collection. Trivially $X_1 = 1$ and, since we are considering the case of a uniform distribution, it follows that when i distinct types of coupons have been collected, a new coupon purchased will

be of a distinct type with probability equal to $\frac{N-i}{N}$. By the independence assumption, we get that the random variable X_i, for $i \in \{2, \ldots, N\}$, is independent from the other variables and has a geometric law with parameter $\frac{N-i+1}{N}$. The expected number of coupons that we have to buy to complete the collection will be therefore

$$\begin{aligned}\mathbb{E}[X] &= \mathbb{E}[X_1] + \cdots + \mathbb{E}[X_N] \\ &= 1 + \frac{N}{N-1} + \frac{N}{N-2} + \cdots + \frac{N}{2} + N \\ &= N \sum_{i=1}^{N} \frac{1}{i} \ .\end{aligned}$$

2.2 The Markov Chains approach

Even if the previous result is very simple and the formula completely clear, we will introduce an alternative approach, that we will use in the following sections where the situation becomes more ob-

scure. If we assume that one coupon arrives at any unit of time, we can interpret the previous variables X_i as the additional time that we have to wait in order to collect the i-th coupon after $i-1$ different types of coupons have been collected. It is then possible to solve the previous problem by using a Markov Chains approach.

Let Y_n be the number of different types of coupons collected after n units of time and assume again that the probability of finding a coupon of any type at any time is $p = \frac{1}{N}$. Y_n will be clearly a Markov Chain on the state space $S = \{0, 1, \ldots, N\}$ with $|S| = N+1$ and it is immediate to obtain that its transition matrix is given by

$$P = \begin{pmatrix} 0 & 1 & 0 & \ldots & \ldots & \ldots & 0 \\ 0 & \frac{1}{N} & \frac{N-1}{N} & 0 & \ldots & \ldots & 0 \\ 0 & 0 & \frac{2}{N} & \frac{N-2}{N} & 0 & \ldots & 0 \\ \vdots & \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & \ldots & \ldots & \ldots & 0 & \frac{N-1}{N} & \frac{1}{N} \\ 0 & \ldots & \ldots & \ldots & 0 & 0 & 1 \end{pmatrix}.$$

Note that $\{N\}$ is the unique closed class and the state N is absorbing. So, in order to determine the mean time taken for X_n to reach the state N, which is equal to the expected number of coupons needed to complete the collection, we have to solve the linear system:

$$\begin{cases} k_N = 0 \\ k_i = 1 + \sum_{j \neq N} p_{ij} k_j, \quad i \neq N \ . \end{cases}$$

Using the transition matrix P, this system is equivalent to

$$\begin{cases} k_0 = k_1 + 1 \\ k_1 = \frac{1}{N} k_1 + \frac{N-1}{N} k_2 + 1 \\ k_2 = \frac{2}{N} k_2 + \frac{N-2}{N} k_3 + 1 \\ \vdots \\ k_{N-2} = \frac{N-2}{N} k_{N-2} + \frac{2}{N} k_{N-1} + 1 \\ k_{N-1} = \frac{N-1}{N} k_{N-1} + 1 \\ k_N = 0 \end{cases}$$

$$\iff \begin{cases} k_N = 0 \\ \frac{1}{N} k_{N-1} = 1 \\ \frac{2}{N} k_{N-2} = \frac{2}{N} k_{N-1} + 1 \\ \vdots \\ \frac{N-2}{N} k_2 = \frac{N-2}{N} k_3 + 1 \\ \frac{N-1}{N} k_1 = \frac{N-1}{N} k_2 + 1 \\ k_0 = k_1 + 1 \end{cases}$$

$$\iff \begin{cases} k_N = 0 \\ k_{N-1} = N \\ k_{N-2} = k_{N-1} + \frac{N}{2} = N + \frac{N}{2} \\ k_{N-3} = k_{N_2} + \frac{N}{3} = N + \frac{N}{2} + \frac{N}{3} \\ \vdots \\ k_2 = k_3 + \frac{N}{N-2} = N + \frac{N}{2} + \cdots + \frac{N}{N-3} + \frac{N}{N-2} \\ k_1 = k_2 + \frac{N}{N-1} = N + \frac{N}{2} + \cdots + \frac{N}{N-2} + \frac{N}{N-1} \\ k_0 = k_1 + 1 = N + \frac{N}{2} + \cdots + \frac{N}{N-2} + \frac{N}{N-1} + 1 \end{cases}$$

It follows that the waiting time to collect all N

coupons is given by

$$k_0 = N \sum_{i=1}^{N} \frac{1}{i} \ .$$

2.3 An approximation formula

For small values of N we can provide easily the computation of the expected number of coupons one has to buy to complete a collection:

N	1	2	3	4	5	6	7
$\mathbb{E}[X]$	1	3	5.50	8.33	11.42	14.70	18.15

For bigger values of N we can instead use the following well known approximation:

$$\sum_{i=1}^{N} \frac{1}{i} = \log(N) + \gamma + \frac{1}{2N} + O\left(\frac{1}{N^2}\right) \ ,$$

where $\gamma \approx 0.5772156649$ is the Euler–Mascheroni constant. We can approximate $\mathbb{E}[X]$ as $N \to +\infty$

by:

$$\mathbb{E}[X] = N\log(N) + N\gamma + \frac{1}{2} + O\left(\frac{1}{N}\right),\ N \to +\infty\ .$$

E.g. if $N = 100$, thanks to *Matlab* it's possible to evaluate the waiting time to complete the collection and to compare this with the asymptotic expansion. If $N = 100$, the exact evaluation gives $\mathbb{E}[X] = 518.7377 \pm 10^{-4}$, while the evaluation of $N\log(N) + N\gamma + \frac{1}{2}$ at $N = 100$ gives 518.7385 ± 10^{-4} as an approximation of the value of $\mathbb{E}[X]$. The exact and approximated values of $\mathbb{E}[X]$ up to $N = 100$ are plotted in Figure 1.

Note that the exact computation in this case can be carried out up to big values of N, but this will not be true in the case of unequal probabilities, that we will consider in the next section.

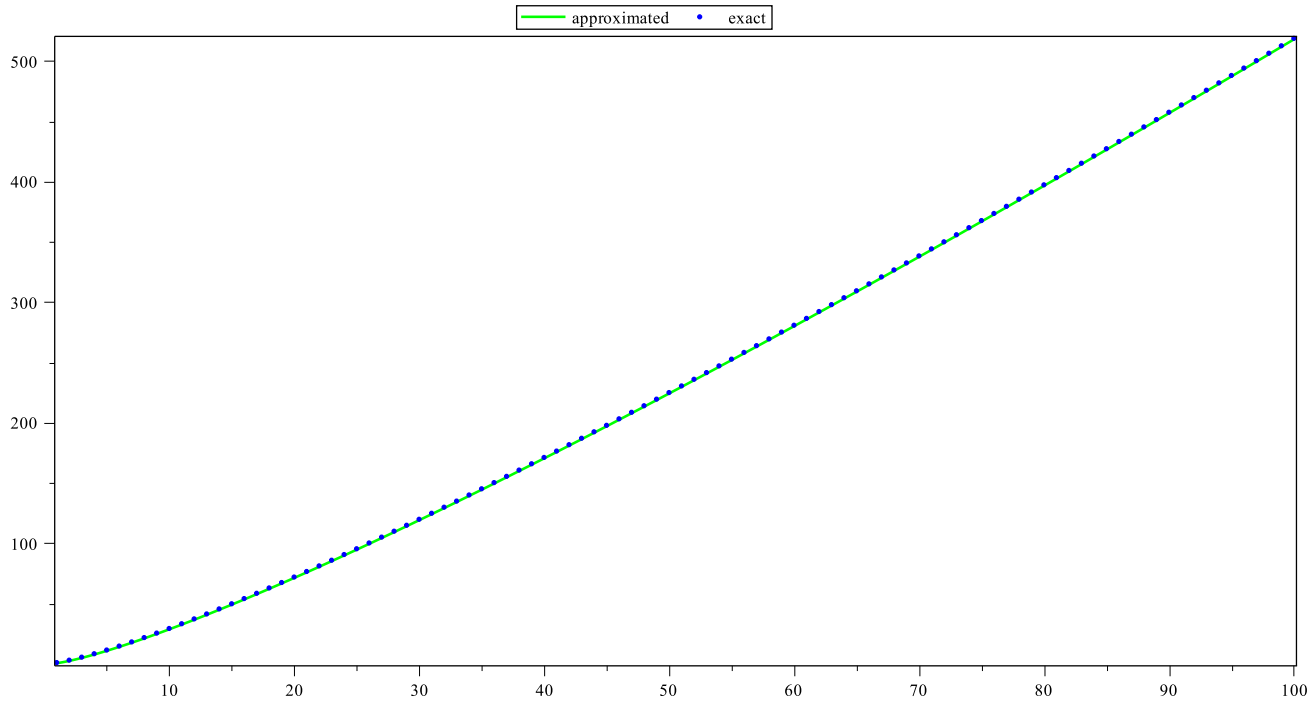


Figure 1: Exact an approximated values of $\mathbb{E}[X]$

3 Single collection with unequal probabilities

Consider now the general case in which there are N different coupons and the type of the successive items are independent random variables that assume the value k with probability p_k no more constant. Here we will have that the i-th coupon type can be purchased with probability $p_i \geq 0$,

with $p_1 + \cdots + p_N = 1$. This is clearly the most realistic case, since in any collection there are "rare" coupons that makes the average number of coupons needed bigger and bigger. Our goal here is to find the expected number of coupons that we have to buy to complete the collection and we will recall a known formula (see e.g. Ross [8]), present a new formula (obtained in [2]) and a possible approximation procedure, outlined in the case of the Mandelbrot distribution.

3.1 The Maximum-Minimums Identity approach

Let us now denote by X_i the random number of coupons we need to buy to obtain the first coupon of type i. The waiting time to complete the collection is therefore given by the random variable $X = \max\{X_1, \ldots, X_N\}$. Note that X_i is a geometric random variable with parameter p_i (because

each new coupon obtained is of type i with probability p_i), but now these variables are no more independent. Since the minimum of X_i and X_j is the number of coupons needed to obtain either a coupon of type i or a coupon of type j, it follows that for $i \neq j$, $\min(X_i, X_j)$ is a geometric random variable with parameter $p_i + p_j$ and the same holds true for the minimum of any finite number of these random variables. To compute the expected value of the random variable X, we will use the *Maximum-Minimums Identity*, whose proof can be found e.g. in [8]:

$$\begin{aligned}
\mathbb{E}[X] &= \mathbb{E}\left[\max_{i=1,\dots,N} X_i\right] \\
&= \sum_i \mathbb{E}[X_i] - \sum_{i<j} \mathbb{E}[\min(X_i, X_j)] \\
&\qquad + \sum_{i<j<k} \mathbb{E}[\min(X_i, X_j, X_k)] - \dots \\
&\qquad \cdots + (-1)^{N+1} \mathbb{E}[\min(X_1, X_2, \dots, X_N)] \\
&= \sum_i \frac{1}{p_i} - \sum_{i<j} \frac{1}{p_i + p_j} + \sum_{i<j<k} \frac{1}{p_i + p_j + p_k} - \dots \\
&\qquad \cdots + (-1)^{N+1} \frac{1}{p_1 + \cdots + p_N} \ .
\end{aligned}$$

Recalling that

$$\int_0^{+\infty} \mathrm{e}^{-p\,x}\,dx = -\left.\frac{\mathrm{e}^{-p\,x}}{p}\right|_{x=0}^{x=+\infty} = \frac{1}{p}$$

and integrating the identity

$$1-\prod_{i=1}^{N}(1-\mathrm{e}^{-p_i\,x}) = \sum_i \mathrm{e}^{-p_i\,x} - \sum_{i<j} \mathrm{e}^{-(p_i+p_j)x} + \cdots + (-1)^{N+1}\mathrm{e}^{-(p_1+\cdots+p_N)x}$$

we get the useful equivalent expression:

$$\mathbb{E}[X] = \int_0^{+\infty} \left(1-\prod_{i=1}^{N}(1-\mathrm{e}^{-p_i\,x})\right) dx\ .$$

Remark 1. Let N be the number of different coupons we want to collect and assume that at each trial we find the i-th coupon with probability $p_i \geq 0$ such that $p_1 + \cdots + p_N = 1$. Our goal is to determine $\mathbb{E}[X]$, where X is the number of coupons we have to buy in order to complete our collection. Assume that the number of coupons bought up to

time t, say $X(t)$, has Poisson distribution with parameter $\lambda = 1$; let Y_i be the generic interarrival time, which represents the time elapsed between the $(i-1)$-th and the i-th purchase: Y_i has exponential distribution with parameter $\lambda = 1$. Note that X is independent from Y_i for all i; indeed, knowing the time interval between two subsequent purchases does not change the probability that the total number of purchases to complete the collection will be equal to a certain fixed value. Let Z_i be the time in which the coupon of type i arrives for the first time (hence $Z_i \sim \exp(p_i)$) and let $Z = \max\{Z_1, \ldots, Z_N\}$ be the time in which we have a coupon of each type (and so we complete the collection). Note that $Z = \sum_{i=0}^{N} Y_i$ and

$\mathbb{E}[X] = \mathbb{E}[Z]$, indeed:

$$\mathbb{E}[Z] = \mathbb{E}[\mathbb{E}[Z|X]] = \sum_k \mathbb{E}\left[\sum_{i=1}^{k} Y_i | X = k\right] \mathbb{P}(X = k)$$

$$= \sum_k \mathbb{E}\left[\sum_{i=1}^{k} Y_i\right] \mathbb{P}(X = k)$$

$$= \sum_k \sum_{i=1}^{k} \mathbb{E}[Y_i] \, \mathbb{P}(X = k)$$

$$= \sum_k k \, \mathbb{P}(X = k) = \mathbb{E}[X] \ .$$

It follows that it suffices to calculate $\mathbb{E}[Z]$ to get $\mathbb{E}[X]$. Since $Z = \max\{Z_1, .., Z_N\}$, we have:

$$F_Z(t) = \mathbb{P}(Z \leq t) = \mathbb{P}(Z_1 \leq t, \ldots, Z_N \leq t)$$

$$= \prod_{i=1}^{N} F_{Z_i}(t) = \prod_{i=1}^{N} (1 - \mathrm{e}^{-p_i \, t})$$

and then

$$\mathbb{E}[Z] = \int_0^{+\infty} \mathbb{P}(Z > t)\, dt$$
$$= \int_0^{+\infty} \left(1 - \prod_{i=1}^{N} (1 - \mathrm{e}^{-p_i\, t})\right) dt\ .$$

3.2 An alternative approach

In [2] has been observed that the coupon collector's problem is related to the following more general problem: given a discrete distribution, determine the minimum size of a random sample drawn from that distribution, in order to observe a given number of different records. The expected size of such a sample can be seen as an alternative method to compute the expected number of coupons needed to complete a collection.

Let $S = \{1, \ldots, N\}$ be the support of a given discrete distribution, $p = (p_1, \ldots, p_N)$ its discrete

density and assume that the elements are drawn randomly from this distribution in sequence; the random variables in the sample will be independent and the realization of each of these will be equal to k with probability p_k. Our aim is to compute the number of drawn one needs to obtain k different realizations of the given distribution. Let X_1 be the random number of drawn needed to have the first record, let X_2 be the number of additional drawn needed to obtain the second record and in general let X_i be the number of drawn needed to go from the $(i-1)$-th to the i-th different record in the sample for every $i \leq N$.

It follows that the random number $X_N(k)$ of drawn one needs to obtain k different records is equal to $X_1 + \cdots + X_k$ and $\mathbb{P}(X_N(k) < +\infty) = 1$.

Note that this problem is very close to the classical coupon collector's problem, but in that case the

random variables X_i denote the number of coupons one has to buy to go from the $(i-1)$-th to the i-th different type of coupon in the collection and $X_N(N)$ represents the number of coupons one has to buy to complete the collection.

In the case of uniform distribution, i.e. $p_k = \frac{1}{N}$ for any $k \in \{1, \ldots, N\}$, the random variable X_i, for $i \in \{2, \ldots, N\}$, has geometric law with parameter $\frac{N-i}{N}$, hence the expected number of drawn needed to obtain k different records is given by:

$$\mathbb{E}[X_N(k)] = 1 + \frac{N}{N-1} + \frac{N}{N-2} + \cdots + \frac{N}{N-k+1} .$$

Note that if $k = N$ we obtain the solution of the coupon collector's problem in the case of uniform distribution. When the probabilities p_k are unequal, to compute the expected value of the random variables X_i we have first to compute their expected values given the types of the preceding $i-1$ different records obtained. To simplify the

notation, define $p(i_1, \ldots, i_k) = 1 - p_{i_1} - \cdots - p_{i_k}$ for $k \leq N$ and different indexes $i_1, i_2, \ldots, i_k$. It can be proved that:

Proposition 1. *For any $k \in \{2, \ldots, N\}$, the expected value of X_k is given by:*

$$\mathbb{E}[X_k] = \sum_{i_1 \neq i_2 \neq \cdots \neq i_{k-1}=1}^{N} \frac{p_{i_1} \cdots p_{i_{k-1}}}{p(i_1)\, p(i_1, i_2) \ldots p(i_1, i_2, \ldots, i_{k-1})}$$

and therefore:

$$\begin{aligned} \mathbb{E}[X_N(k)] &= \sum_{s=1}^{k} \mathbb{E}[X_s] \\ &= 1 + \sum_{i_1=1}^{N} \frac{p_{i_1}}{p(i_1)} + \sum_{i_1 \neq i_2=1}^{N} \frac{p_{i_1}\, p_{i_2}}{p(i_1)\, p(i_1, i_2)} + \ldots \qquad (1) \\ &\quad + \sum_{i_1 \neq i_2 \neq \cdots \neq i_{k-1}=1}^{N} \frac{p_{i_1} \cdots p_{i_{k-1}}}{p(i_1)\, p(i_1, i_2) \ldots p(i_1, i_2, \ldots, i_{k-1})} \,. \end{aligned}$$

Remark 2. Note that, when $k = N$, the last expression represents an alternative way to compute the expected number of coupons needed to complete a collection. The proof of the equivalence

between the last expression with $k = N$ and the formula obtained using the Maximum-Minimums Identity is not trivial; furthermore, both expressions are not computable for large values of k.

Proof of Proposition 1: In order to compute the expected value of the variable X_k, we shall take the expectation of the conditional expectation of X_k given $Z_1, \ldots, Z_{k-1}$, where Z_i, for $i = 1, \ldots, N$, denotes the type of the i-th different coupon collected.

To simplify the exposition, let us start by evaluating $\mathbb{E}[X_2]$; we have immediately that $X_2|Z_1 = i$ has a (conditioned) geometric law with parameter $1 - p_i = p(i)$ and therefore $\mathbb{E}[X_2|Z_1 = i] = \frac{1}{p(i)}$. So

$$\begin{aligned}\mathbb{E}[X_2] &= \mathbb{E}[\mathbb{E}[X_2|Z_1]] \\ &= \sum_{i=1}^{N} \mathbb{E}[X_2|Z_1 = i]\, \mathbb{P}[Z_1 = i] = \sum_{i=1}^{N} \frac{p_i}{p(i)} \ .\end{aligned}$$

Let us now take $k \in \{3, \ldots, N\}$: it is easy to see that

$$\begin{aligned}\mathbb{E}[X_k] &= \mathbb{E}[\mathbb{E}[X_k|Z_1, Z_2, \ldots, Z_{k-1}]] \\ &= \sum_{i_1 \neq i_2 \neq \cdots \neq i_{k-1}=1}^{N} \mathbb{E}[X_k|Z_1 = i_1, , Z_2 = i_2, \cdots, Z_{k-1} = i_{k-1}] \\ &\qquad \times \mathbb{P}[Z_1 = i_1, ..., Z_{k-1} = i_{k-1}] \ .\end{aligned}$$

Note that $\mathbb{P}[Z_i = Z_j] = 0$ for any $i \neq j$. The conditional law of $X_k|Z_1 = i_1, Z_2 = i_2, \cdots, Z_{k-1} = i_{k-1}$, for $i_1 \neq i_2 \neq \cdots \neq i_{k-1}$, is that of a geometric random variable with parameter $p(i_1, \ldots, i_{k-1})$ and its conditional expectation is $p(i_1, \ldots, i_{k-1})^{-1}$. By the multiplication rule, we get

$$\begin{aligned}\mathbb{P}[Z_1 =& i_1, \ldots, Z_{k-1} = i_{k-1}] \\ &= \mathbb{P}[Z_1 = i_1] \times \mathbb{P}[Z_2 = i_2|Z_1 = i_1] \times \\ &\qquad \cdots \times \mathbb{P}[Z_{k-1} = i_{k-1}|Z_1 = i_1, \ldots, Z_{k-2} = i_{k-2}] \ .\end{aligned}$$

Note that, even though the types of the successive coupons are independent random variables, the random variables Z_i are not mutually independent. A simple computation gives, for any

$s = 2, \ldots, k-1$, that

$$\mathbb{P}[Z_s = i_s | Z_1 = i_1, \ldots, Z_{s-1} = i_{s-1}] = \frac{p_{i_s}}{1 - p_{i_1} - \ldots - p_{i_{s-1}}}$$

if $i_1 \neq i_2 \neq \cdots \neq i_{k-1}$ and zero otherwise. Recalling the compact notation $p(i_1, \ldots, i_k) = 1 - p_{i_1} - \ldots - p_{i_k}$, we then get

$$E[X_k] = \sum_{i_1 \neq i_2 \neq \cdots \neq i_{k-1} = 1}^{N} \frac{p_{i_1}\, p_{i_2} \cdots p_{i_{k-1}}}{p(i_1)\, p(i_1, i_2) \cdots p(i_1, i_2, \ldots, i_{k-1})}$$

and the proof is complete. □

3.3 An approximation procedure via the Heaps' law in natural languages

In [2] Ferrante and Frigo proposed the following procedure in order to approximate the expected number of coupons needed to complete a collection in the unequal case.

Let us consider a text written in a natural language: the Heaps' law is an empirical law which describes the portion of the vocabulary which is

used in the given text. This law can be described by the following formula

$$\mathbb{E}[R_m(n)] \sim K\, n^{\beta}$$

where $R_m(n)$ is the (random) number of different words presents in a text consisting of n words and taken from a vocabulary of m words, while K and β are free parameters determined empirically. In order to obtain a formal derivation of this empirical law, van Leijenhorst and van der Weide in [5] have considered the average growth in the number of records, when elements are drawn randomly from some statistical distribution that can assume exactly m different values. The exact computation of the average number of records in a sample of size n, $\mathbb{E}[R_m(n)]$, can be easily obtained using the following approach. Let $S = \{1, 2, \ldots, m\}$ be the support of the given distribution, define $X = m - R_m(n)$ the number of values in S not observed and denote

by A_i the event that the record i is not observed. It is immediate to see that $\mathbb{P}[A_i] = (1-p_i)^n$, $X = \sum_{i=1}^{m} \mathbf{1}_{A_i}$ and therefore that

$$\mathbb{E}[R_m(n)] = m - \mathbb{E}[X] = m - \sum_{i=1}^{m} (1-p_i)^n \ . \quad (2)$$

Assuming now that the elements are drawn randomly from the Mandelbrot distribution, van Leijenhorst and van der Weide obtain that the Heaps' law is asymptotically true as n and m go to infinity and $n \ll m^{\theta-1}$, where θ is one of the parameters of the Mandelbrot distribution (see [5] for the details).

It is possible to relate this problem with the previous one: assume that we are interested in the minimum number $X_m(k)$ of elements that we have to draw randomly from a given statistical distribution in order to obtain k different records. This is clearly strictly related to the previous problem

and at first sight one expects that the technical difficulties would be similar. However, we have just proved that the computation of the expectation of $X_m(k)$ is much more complicated. The formula that we obtained before is computationally hard and we are able to perform the exact computation in the environment R just for distributions with a support of small cardinality. An approximation procedure can be obtained in the special case of the Mandelbrot distribution, widely used in the applications, making use of the asymptotic results proved in [5] in order to derive the Heaps' law.

The exact formula we obtained in the previous section is nice, but it is tremendously heavy to compute as soon as the cardinality of the support of the distribution becomes larger then 10. The number of all possible ordered choices of indexes sets in-

volved in (1) increases very fast with k leading the objects hard to handle with a personal computer. For this reason it would be important to be able to approximate this formula, at least in some cases of interest, even if its complicated structure may suggest that it could be quite difficult in general. In this section we shall consider the case of the Mandelbrot distribution, which is commonly used in the Heaps' law and other practical problems. Applying the results proved in [5], we present here a possible strategy to approximate the expectation of $X_m(k)$ and present some numerical approximation in order to test our procedure. Let us consider $R_m(n)$ and $X_m(k)$: these two random variables are strictly related, since $[R_m(n) > k] = [X_m(k) < n]$, for $k \leq n \leq m$. However, we have seen that the computation of their expected values is quite different. With an abuse of language, we could say that the

two functions $n \mapsto \mathbb{E}[R_m(n)]$ and $k \mapsto \mathbb{E}[X_m(k)]$ represent one the "inverse" of the other. In order to confirm this statement, let us consider the case studied in [5], i.e. let us assume to sample from the Mandelbrot distribution. Fixed three parameters $m \in \mathbb{N}$, $\theta \in [1, 2]$ and $c \geq 0$, we shall assume that $S = \{1, \ldots, m\}$ and

$$p_i = a_m(c+i)^{-\theta} \ , \ a_m = \left(\sum_{i=1}^{m} (c+i)^{-\theta} \right)^{-1} . \quad (3)$$

We implement both the expressions (2) and (1) using the environment R. We set the parameters of the Mandelbrot distribution to be $c = 0.30$ and $\theta = 1.75$. Using (1), we compute the expected number $\mathbb{E}[X_m(k)]$ of elements we have to draw randomly from a Mandelbrot distribution in order to obtain k different records, for three levels of m, being m the vocabulary size, i.e the maximum size of different words. In brackets we show the expected

number of different words in a random selection of exactly $E[X_m(k)]$ elements, computed using (2). Results are collected in Table 1. We see that the number of different words we expect in a text size of dimension $E[X_m(k)]$ is close to the value of k and this supports our statement about the connection between $\mathbb{E}[R_m(n)]$ and $\mathbb{E}[X_m(k)]$. As underlined before, we can compute these expectations only for small values of k.

At the same time, since $\mathbb{E}[R_m(n)] \leq m$, it is clear that our statement that $n \mapsto \mathbb{E}[R_m(n)]$ and $k \mapsto \mathbb{E}[X_m(k)]$ represent one the "inverse" of the other could be valid just for values of k small with respect to m. This idea arises also from Table 1, but in order to confirm this we shall compare the two functions for larger values of m. Since our formula is not computable for values larger than 10, we shall perform a simulation to obtain its approx-

		Vocabulary size		
		$m = 5$	$m = 8$	$m = 10$
number of different words	$k = 2$	2.80 (1.97)	2.63 (2.00)	2.57 (2.01)
	$k = 3$	6.08 (2.87)	5.17 (2.95)	4.93 (2.97)
	$k = 4$	12.42 (3.76)	9.01 (3.90)	8.31 (3.92)
	$k = 5$	28.46 (4.59)	14.81 (4.84)	13.04 (4.88)
	$k = 6$	-	23.95 (5.77)	19.68 (5.84)
	$k = 7$	-	39.96 (6.69)	29.21 (6.80)
	$k = 8$	-	77.77 (7.55)	43.66 (7.74)

Table 1: Expected text size in order to have k different words taken from a vocabulary of size m

imated values. In Figure 2 we compare the values of the two functions for $m = 100$ and for values of k ranging from 1 to m. Again, we suppose that the elements are drawn from a Mandelbrot distribution with the same value of c and θ. The two functions are close up to $k = 90$, while for larger values of k the distance between the two values increases. Thanks to these results, we propose the

following approximation strategy: the main result proved in [5] is that

$$\mathbb{E}[R_m(n)] \sim \alpha\, n^{\beta}$$

when $n, m \to \infty$ with validity region $n \ll m^{\theta-1}$, where $\beta = \theta^{-1}$ and $\alpha = a_\infty^{\beta}\,\Gamma(1-\beta)$, where $a_\infty = \lim_{m\to\infty} a_m$ (see expression (3)). Assuming that for values of $n \ll m^{\theta-1}$, $n \mapsto \mathbb{E}[R_m(n)]$ and $k \mapsto \mathbb{E}[X_m(k)]$ could represent one the "inverse" of the other, we get

$$\mathbb{E}[X_m(k)] \sim \left(\frac{k}{\alpha}\right)^{\theta}$$

with validity region $k \ll \tau$, where τ is the approximated value of k for which $\mathbb{E}[X_m(k)] = m^{\theta-1}$. In order to test our approximation scheme, we shall take the same value of the constants as before, $m = 500$, $k = 1, \ldots, 60$. Figure 3 shows the results: we obtain a very good correspondence between the

simulated values and the approximation curve in the range of applicability $k \ll 25$.

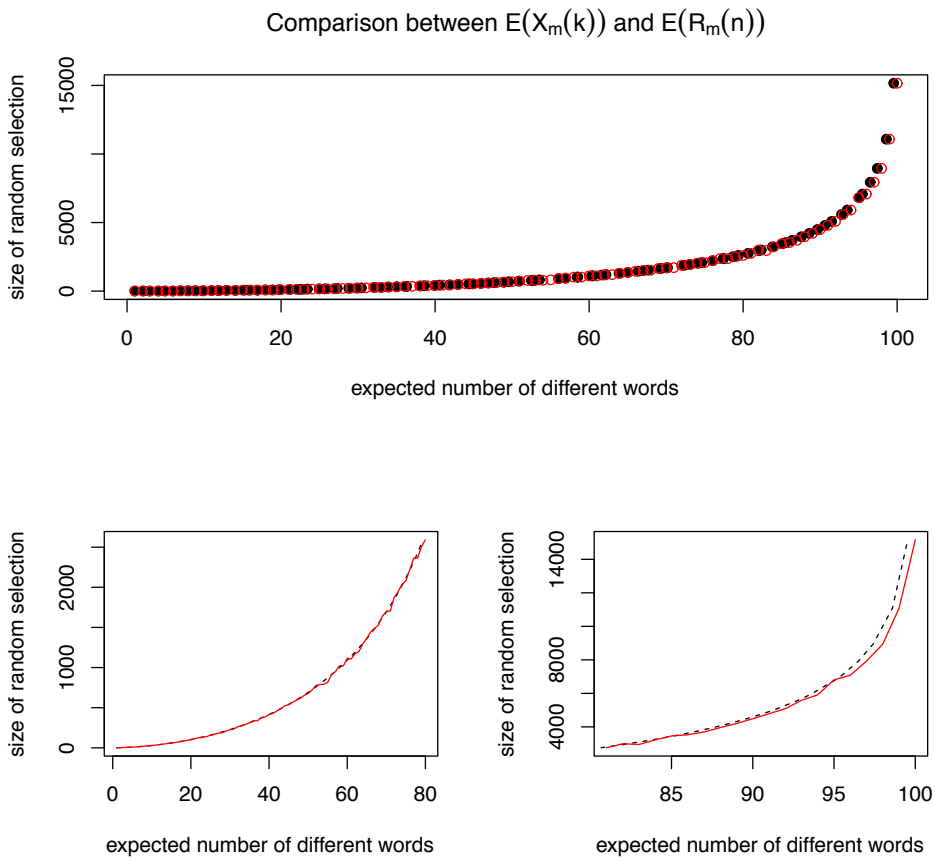


Figure 2: Comparison between $\mathbb{E}[X_m(k)]$ (filled red circles) and $\mathbb{E}[R_m(n)]$ (solid black circles) for $m = 100$ and $k = 1, \ldots, 100$ (main figure). Zoom: comparison between $\mathbb{E}[X_m(k)]$ (solid red line) and $\mathbb{E}[R_m(n)]$ (dashed black line) for $k = 1, \ldots, 80$ (left) and $k = 81, \ldots, 100$ (right).

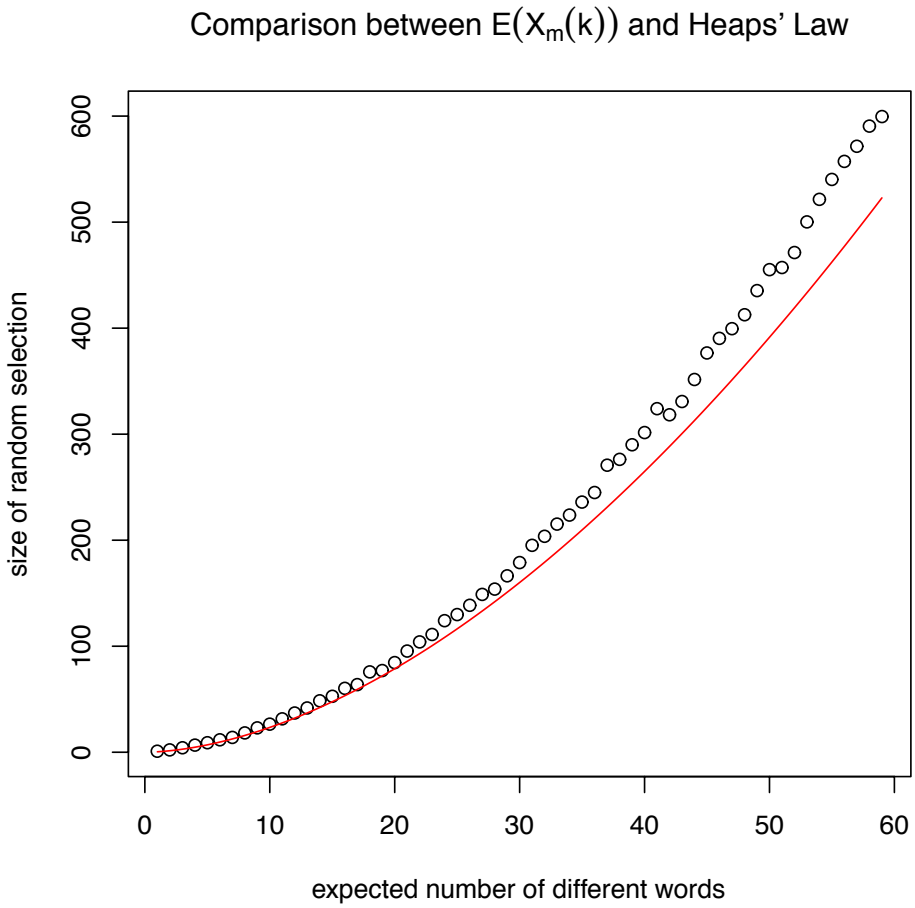


Figure 3: Comparison between $\mathbb{E}[X_m(k)]$ (filled black circles) and $(k/\alpha)^\theta$ (solid red line) for $m = 500$ and $k = 1, \ldots, 60$.

3.4 Comparison between the equal and unequal cases

Let us denote by $X_{(\frac{1}{N},\dots,\frac{1}{N})}$ the random number of coupons we have to buy to complete the collection using the uniform probability and by $X_{(p_1,\dots,p_N)}$ the number of coupons in the unequal case. We have already calculated their expectations:

$$\mathbb{E}[X_{(\frac{1}{N},\dots,\frac{1}{N})}] = N\sum_{i=1}^{N}\frac{1}{i}\ ,$$

$$\mathbb{E}[X_{(p_1,\dots,p_N)}] = \sum_{i}\frac{1}{p_i} - \sum_{i<j}\frac{1}{p_i+p_j} + \cdots + (-1)^{N+1}\frac{1}{p_1+\cdots+p_N}\ .$$

Now we introduce some standard notation that will lead us to conclude that it is harder to collect all kinds of coupons if there is some bias for the probability of appearance of coupons. For a distribution $p = (p_1,\dots,p_N)$, let $p_{[j]}$ be the j-th largest value

of $\{p_1, \ldots, p_N\}$, that is $p_{[1]} \geq p_{[2]} \geq \cdots \geq p_{[N]}$. We say that a distribution $p = (p_1, \ldots, p_N)$ is majorized by a distribution $q = (q_1, \ldots, q_N)$, and we write $p \prec q$, if $\sum_{i=1}^{k} p_{[i]} \leq \sum_{j=1}^{k} q_{[j]}$ for all $1 \leq k \leq N-1$. Let us prove that $\left(\frac{1}{N}, \ldots, \frac{1}{N}\right) \prec p$ for any distribution $p = (p_1, \ldots, p_N)$. Indeed, since p is a distribution, we have that $p_{[1]} \geq \frac{1}{N}$. Let now assume that there exists $1 < k \leq N-1$ such that $\sum_{i=1}^{k} p_{[i]} < \frac{k}{N}$. This implies that $\sum_{i=k+1}^{N} p_{[i]} > \frac{N-k}{N}$, which in turn implies that there exists $j \in \{k+1, \ldots, N\}$ such that $p_{[j]} > \frac{1}{N}$, which leads to a contradiction with the fact that $\sum_{i=1}^{k} p_{[i]} < \frac{k}{N}$.

We say that the symmetric function $f(p)$ defined on a distribution is Schur convex (resp. concave) if $p \prec q \implies f(p) \leq f(q)$ (resp. $f(p) \geq f(q)$). Finally, we say that a random variable X is stochastically smaller than a random variable Y if

$\mathbb{P}(X > a) \leq \mathbb{P}(Y > a)$ for all real a. The following results have been proved in [6]:

Theorem 1. *The probability* $\mathbb{P}(X_p \leq n)$ *is a Schur concave function of* p.

Corollary 1. *If* $p \prec q$, *then* X_p *is stochastically smaller than* X_q. *In particular:* $X_{(\frac{1}{N},\ldots,\frac{1}{N})}$ *is stochastically smaller than* X_p *for all* p.

Corollary 2. *The expectation* $\mathbb{E}[X_p]$ *is a Schur convex function of* p. *In particular:* $\mathbb{E}[X_{(\frac{1}{N},\ldots,\frac{1}{N})}] \leq \mathbb{E}[X_{(p_1,\ldots,p_N)}]$ *for all* p.

3.5 Examples

- $N = 2$, $\mathbb{E}[X_{(\frac{1}{2},\frac{1}{2})}] = 3$

- $p_1 = \frac{1}{3},\ p_2 = \frac{2}{3}$

$$\mathbb{E}[X_{(p_1,p_2)}] = \frac{1}{p_1} + \frac{1}{p_2} - \frac{1}{p_1+p_2}$$
$$= 3 + \frac{3}{2} - 1 = \frac{7}{2} = 3.5$$

- $p_1 = \frac{1}{6},\ p_2 = \frac{5}{6}$

$$\mathbb{E}[X_{(p_1,p_2)}] = \frac{1}{p_1} + \frac{1}{p_2} - \frac{1}{p_1+p_2}$$
$$= 6 + \frac{6}{5} - 1 = \frac{31}{5} = 6.2$$

- $N = 3,\ \mathbb{E}[X_{(\frac{1}{3},\frac{1}{3},\frac{1}{3})}] = 5.5$

  - $p_1 = \frac{1}{4},\ p_2 = \frac{1}{2},\ p_3 = \frac{1}{4}$

$$\mathbb{E}[X_{(p_1,p_2,p_3)}] = \frac{1}{p_1} + \frac{1}{p_2} - \frac{1}{p_1+p_2} - \frac{1}{p_1+p_3} - \frac{1}{p_2+p_3} + \frac{1}{p_1+p_2+p_3}$$
$$= \frac{19}{3} \approx 6.33$$

- $p_1 = \frac{1}{6},\ p_2 = \frac{4}{6},\ p_3 = \frac{1}{6}$

$$\begin{aligned}\mathbb{E}[X_{(p_1,p_2,p_3)}] = &\frac{1}{p_1} + \frac{1}{p_2} - \frac{1}{p_1+p_2} \\ &- \frac{1}{p_1+p_3} - \frac{1}{p_2+p_3} \\ &+ \frac{1}{p_1+p_2+p_3} \\ = &\frac{91}{10} = 9.1\ .\end{aligned}$$

In Figure 4 we show the value of $\mathbb{E}[X]$ for different choices of p_1 ($p_2 = 1-p_1$) in the case $N = 2$, while in Figure 5 we show some level curves of $\mathbb{E}[X]$ for different choices of p_1 and p_2 ($p_3 = 1-p_1-p_2$) in the case $N = 3$.

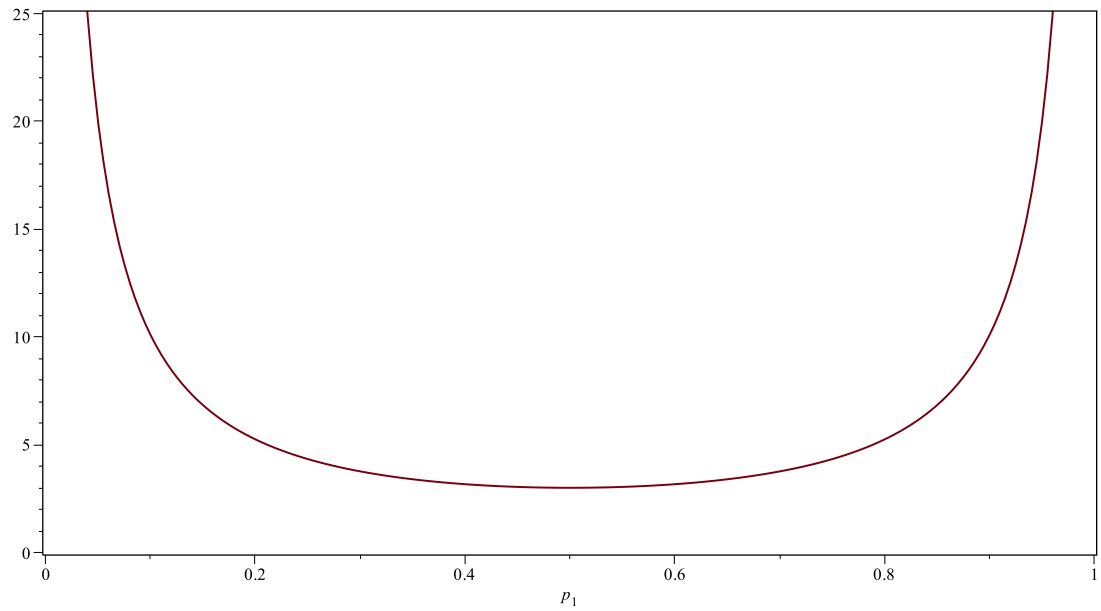


Figure 4: Values of $\mathbb{E}[X]$ in terms of p_1 when $N = 2$

4 Coupons in groups of constant size

Consider now the case of coupons which arrives in groups of constant size g, with $1 < g < N$, independently and with unequal probabilities; assume that each group does not contain more than one coupon of any type, hence the total number of groups will be $\binom{N}{g}$ and each group A can be identified with a vector $(a_1, \ldots, a_g) \in \{1, \ldots, N\}^g$ with $a_i < a_{i+1}$ for $i = 1, \ldots, g-1$.

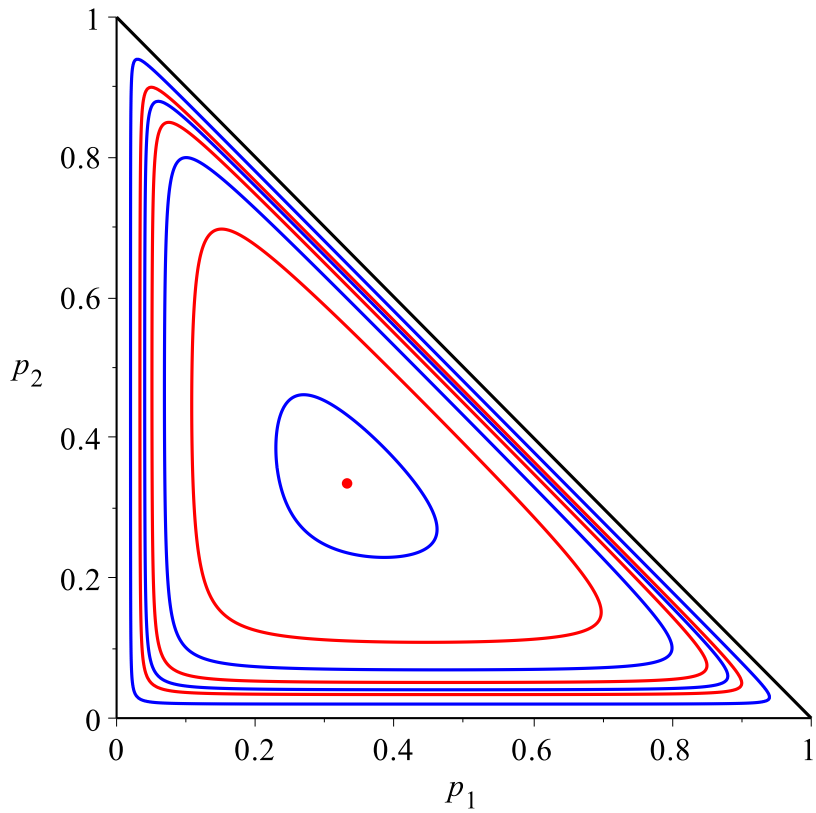


Figure 5: Level curves coresponding to $\mathbb{E}[X] = 50, 30, 25, 20, 15, 10, 6, 5.5$ in terms of p_1 and p_2 when $N = 3$

Moreover, assume that the type of successive groups of coupons that one collects form a sequence of independent random variables.

To study the unequal probabilities case, we first order the groups according to the lexicographical order (i.e. $A = (a_1, \ldots, a_g) < B = (b_1, \ldots, b_g)$ if there exists $i \in \{1, \ldots, g-1\}$ such that $a_s = b_s$ for $s < i$ and $a_i < b_i$).
We denote by $q_i,\ i \in \left\{1, \ldots, \binom{N}{g}\right\}$, the probability to purchase (at any given time) the i-th group of coupons, according to the lexicographical order, and given $k \in \{1, \ldots, N-g\}$ we denote by $q(i_1, \ldots, i_k)$ the probability to purchase a group of coupons which does not contain any of the coupons $i_1, \ldots, i_k$.

To compute the probabilities $q(i_1, \ldots, i_k)$'s,

note that by the defined ordering it holds that:

$$q(1) = \sum_{i=\binom{N-1}{g-1}+1}^{\binom{N}{g}} q_i \ , \quad q(2) = \sum_{i=\binom{N-1}{g-1}+\binom{N-2}{g-1}+1}^{\binom{N}{g}} q_i$$

and in general:

$$q(1,2,\ldots,k) = \begin{cases} \displaystyle\sum_{i=\binom{N-1}{g-1}+\cdots+\binom{N-k}{g-1}+1}^{\binom{N}{g}} q_i, & \text{if } k \leq N-g \\ 0 & \text{otherwise.} \end{cases}$$

For any permutation $(i_1, \ldots, i_N)$ of $(1, \ldots, N)$, one first reorders the q_i's according to the lexicographical order of this new alphabet and then compute

$$q(i_1,i_2,\ldots,i_k) = \begin{cases} \displaystyle\sum_{i=\binom{N-1}{g-1}+\cdots+\binom{N-k}{g-1}+1}^{\binom{N}{g}} q_i, & \text{if } k \leq N-g \\ 0 & \text{otherwise.} \end{cases}$$

Remark 3. There are many conceivable choices for the unequal probabilities q_i's. For example, we

can assume that one forms the groups following the strategy of the draft lottery in the American professional sports, where different proportion of the different coupons are put together and we choose at random in sequence the coupons, discarding the eventually duplicates, up to obtaining a group of k coupons. Or, more simply, we can assume that the i-th coupon will arrive with probability p_i and that the probability of any group is proportional to the product of the probabilities of the single coupons contained.

Let us start with the case of uniform probabilities, i.e.

$$q_i = \frac{1}{\binom{N}{g}}$$

for any i, and let us define the following set of random variables:

$$V_i = \left\{ \begin{array}{l} \text{number of groups to purchase to obtain} \\ \text{the first coupon of type } i \end{array} \right\}.$$

These random variables have a geometric law with parameter

$$1 - \frac{\binom{N-1}{g}}{\binom{N}{g}} \ ,$$

therefore the random variables $\min(V_i, V_j)$ have geometric law with parameter $1 - \frac{\binom{N-2}{g}}{\binom{N}{g}}$ and the random variables $\min(V_{i_1}, \ldots, V_{i_{N-g}})$ have geometric law with parameter $1 - \frac{1}{\binom{N}{g}}$; the minimum of more random variables, i.e. $\min(V_{i_1}, \ldots, V_{i_k})$ for $k > N - g + 1$, will be equal to the constant random variable 1. Applying the *Maximum-Minimums Principle*, we obtain the expected number of groups of coupons that we have to buy to

complete the collection:

$$
\begin{aligned}
&\mathbb{E}[\max(V_1,\dots,V_N)] \\
&\quad = \sum_{1\le i\le N}\mathbb{E}[V_i] - \sum_{1\le i\le j\le N}\mathbb{E}[\min(V_i,V_j)] + \dots \\
&\qquad \cdots + (-1)^{N-g+1} \sum_{0\le i_1<i_2<\cdots<i_{N-g}\le N}\mathbb{E}[V_{i_1},\dots,V_{i_{N-g+1}}] \\
&\qquad\quad + (-1)^{N-g+2} \sum_{0\le i_1<i_2<\cdots<i_{N-g+1}\le N} 1 + \cdots + (-1)^{N+1} \\
&\quad = \binom{N}{1}\frac{1}{1-\frac{\binom{N-1}{g}}{\binom{N}{g}}} - \binom{N}{2}\frac{1}{1-\frac{\binom{N-2}{g}}{\binom{N}{g}}} \\
&\qquad + \binom{N}{3}\frac{1}{1-\frac{\binom{N-3}{g}}{\binom{N}{g}}} - \cdots \\
&\qquad \cdots + (-1)^{N-g+1}\binom{N}{N-g}\frac{1}{1-\frac{1}{\binom{N}{g}}} \\
&\qquad\quad + \sum_{1\le k\le g}(-1)^{N-g+k+1}\binom{N}{N-g+k}\,.
\end{aligned}
$$

This result has been first proved by W. Stadje in [9] with a different technique, but with the present approach developed in [1] it can be easily general-

ized to the unequal probabilities case as follows:

Proposition 2. *The expected number of groups of coupons that we need to complete the collection, in the case of unequal probabilities q_i, is equal to:*

$$\sum_{1\le i\le N}\frac{1}{1-q(i)}-\sum_{1\le i\le j\le N}\frac{1}{1-q(i,j)}+\sum_{0\le i<j<l\le N}\frac{1}{1-q(i,j,l)}-\cdots\cdots+(-1)^{N-g+1}\sum_{0\le i_1<i_2<\cdots<i_{N-g}\le N}\frac{1}{1-q(i_1,\ldots,i_{N-g})}+\sum_{1\le k\le g}(-1)^{N-g+k+1}\binom{N}{N-g+k}.$$

In the following Table we list the mean number of groups of coupons of constant size g that we have to buy to complete a collection of N coupons

in the case of uniform probabilities.

	$g=1$	$g=2$	$g=3$	$g=4$	$g=5$
$N=5$	11.4167	5.3254	3.2222	2.2500	1
$N=10$	29.2897	14.1234	9.0462	6.4877	4.9340
$N=15$	49.7734	24.2783	15.7651	11.4960	8.9232
$N=20$	71.9548	35.3047	23.0766	16.9532	13.2709
$N=30$	119.8496	59.1591	38.9208	28.7952	22.7145
$N=40$	171.1417	84.7376	55.9299	41.5211	32.8717
$N=50$	224.9201	111.6713	73.7602	54.8777	43.5889

5 Multiple collections with equal probabilities

Let us now consider the following generalization: m siblings collect coupons with the aim to complete m collections. When a coupon arrives, the oldest sibling checks his collection and if the coupon is already present, passes this to the second oldest sibling and so on. What is the expected number of coupons needed to complete these m collections

in this collaborative setting? For sure this number will be smaller than m times the average number of coupons needed to complete a single collection, but how difficult even if possible will be the exact computation? In the equal case we will present the exact solution, known in the literature since 1960, and for the unequal case we will present an upper bound and the exact solution via the Markovian approach. Both these results are to the best of our knowledge original, or at least difficult to be found in the existing literature.

5.1 General solution

The solution to the problem of determining the mean time to complete m sets of N equally likely coupons was found in 1960 by D. J. Newman and L. Shepp.

Assume we want to collect m sets of N equally

likely coupons and let p_i be the probability of failure of obtaining m sets up to and including the purchase of the i-th coupon. Then, denoting by X the random number of coupons needed to complete m sets, its expectation $\mathbb{E}[X]$ is equal to $\sum_{i=0}^{+\infty} p_i$. Note that $p_i = \frac{M_i}{N^i}$, where M_i is the number of ways that the purchase of the i-th coupon can fail to yield m copies of each of the N coupons in the set. If we represent the coupons by $x_1, \ldots, x_N$, then $M_i = (x_1 + \cdots + x_N)^i$ expanded and evaluated at $(1, \ldots, 1)$ after all the terms have been removed which have each exponent for each variable larger than $m-1$. Consider m fixed and introduce the following notation: if $P(x_1, \ldots, x_N)$ is a polynomial (or a power series) we define $\{P(x_1, \ldots, x_N)\}$ to be the polynomial (or series) resulting when all terms having all exponents greater or equal to m have been removed. In terms of this notation, p_i equals

to $\frac{\{(x_1+\cdots+x_N)^i\}}{N^i}$ evaluated at $x_1 = \cdots = x_N = 1$. Define

$$S_m(t) = \sum_{k=0}^{m-1} \frac{t^k}{k!}$$

and consider the expression

$$F = \mathrm{e}^{x_1+\cdots+x_N} - (\mathrm{e}^{x_1} - S_m(x_1)) \times \cdots \times (\mathrm{e}^{x_N} - S_m(x_N)) .$$

Note that F has no terms with all exponents greater or equal to m, but F doesn't have all terms of $\mathrm{e}^{x_1+\cdots+x_N}$ with at least one exponent smaller than m; it follows that

$$F = \{\mathrm{e}^{x_1+\cdots+x_N}\} = \sum_{i=0}^{\infty} \frac{\{(x_1+\cdots+x_N)^i\}}{i!} .$$

Remembering that

$$\mathbb{E}[X] = \sum_{i=0}^{\infty} p_i = \sum_{i=0}^{\infty} \frac{\{(x_1+\cdots+x_N)^i\}}{N^i}$$

at $x_1 = \cdots = x_N = 1$, and using the identity

$$N \int_0^{+\infty} \frac{t^i}{i!} \mathrm{e}^{-Nt}\, dt = \frac{1}{N^i}$$

it follows that

$$\sum_{i=0}^{\infty} \frac{\{(x_1 + \cdots + x_N)^i\}}{N^i}$$

$$= \sum_{i=0}^{\infty} \left(\{(x_1 + \cdots + x_N)^i\} \, N \int_0^{+\infty} \frac{t^i}{i!} \, \mathrm{e}^{-N\,t} \, dt \right)$$

$$= N \int_0^{+\infty} \sum_{i=0}^{\infty} \left(\{(x_1 + \cdots + x_N)^i\} \, \frac{t^i}{i!} \right) \mathrm{e}^{-N\,t} \, dt$$

$$= N \int_0^{+\infty} \Big[\mathrm{e}^{t\,(x_1 + \cdots + x_N)} - \big(\mathrm{e}^{t\,x_1} - S_m(t\,x_1) \big) \times \ldots$$

$$\cdots \times \big(\mathrm{e}^{t\,x_N} - S_m(t\,x_N) \big) \Big] \, \mathrm{e}^{-N\,t} \, dt \ .$$

We easily obtain, evaluating the previous expressions for $x_1 = \cdots = x_N = 1$,

$$\mathbb{E}[X] = N \int_0^{+\infty} \Big[1 - \big(1 - S_m(t)\, \mathrm{e}^{-t} \big)^N \Big] \, dt \ .$$

So in order to determine the explicit value, we have to substitute the corresponding value of $S_m(t)$ and integrate.

It can be seen that for large values of m, $\mathbb{E}[X]$ is asymptotic to mN. Indeed, let Y_1^k denotes the

(random) number of coupons needed to obtain the first coupon of type k, and, for any $i = 2, \ldots, m$, Y_i^k denotes the additional coupons that we need to purchase to pass from $i-1$ to i coupon of type k in our collection. These random variables are independent and geometrically distributed of parameter $1/N$. So, by the Strong Law of Large Numbers, we get that

$$\frac{Y_1^k + \ldots + Y_m^k}{m} \to \mathbb{E}[Y_1^1] = N$$

for any $k \in \{1, \ldots, N\}$. Since

$$X = \max_{k=1,\ldots,N} \{Y_1^k + \ldots + Y_m^k\}$$

we get that X is asymptotic to mN for m large. On the contrary, for m fixed it can be proved that:

$$\mathbb{E}[X] = N\left[\log(N) + (m-1)\log\log(N) + C_m + o(1)\right], \quad N \to +\infty \ .$$

Let us now see some examples, where $G(m, N)$ will denote the expected number of coupons needed to complete m sets of N different coupons.

5.1.1 Examples

- $m = 2,\ N = 2$

$$G(2,2) = 2\int_0^{+\infty} \left[1 - \left(1 - (1+t)\,\mathrm{e}^{-t}\right)^2\right] dt$$
$$= 2\int_0^{+\infty} \left[2(1+t)\,\mathrm{e}^{-t} - (1+t)^2\,\mathrm{e}^{-2t}\right]$$

Integrating by parts we get:

$$G(2,2) = \frac{11}{2} = 5.5\ .$$

Recall that $G(1,2) = 3$, so $2G(1,2) = 6 > 5.5 = G(2,2)$.

- $m = 2,\ N = 3$

$$G(2,3) = 3\int_0^{+\infty} \left[1 - \left(1 - (1+t)\,\mathrm{e}^{-t}\right)^3\right] dt$$
$$= 3\int_0^{+\infty} \left[(1+t)^3\,\mathrm{e}^{-3t} - 3\,(1+t)^2\,\mathrm{e}^{-2t} + 3\,(1+t)\,\mathrm{e}^{-t}\right] dt\ .$$

Integrating by parts we get:

$$G(2,3) = \frac{347}{36} \approx 9.64\ .$$

Since $G(1,3) = 5.5$, we get $2 \cdot G(1,3) = 11 > G(2,3) \approx 9.64$.

5.2 The Markov chains approach

As an alternative method, that we will develop in the last section in the case of unequal probabilities, let us consider the case $m = 2$ making use of the Markov Chains approach.
Let $X_n,\ n \geq 0$, be the random variable that denotes the number of different coupons in the collections after n trials; $\{X_n,\ n \geq 0\}$ is a Markov

chain on the state space $S = \{(i,j) : i,j \in \{0,1,\ldots,N\}, i \geq j\}$ and easily $|S| = \frac{(N+1)(N+2)}{2}$. The transition probabilities are given by:

$$(0,0) \longrightarrow (1,0) \qquad \text{with probability } 1$$

$$(i,j) \longrightarrow \begin{cases} (i+1,j) & \text{with probability } \frac{N-i}{N} \\ (i,j+1) & \text{with probability } \frac{i-j}{N} \\ (i,j) & \text{with probability } \frac{j}{N} \end{cases} \quad \text{with } i \geq j$$

$$(N,N) \longrightarrow (N,N) \qquad \text{with probability } 1 \ .$$

Note that the state (N,N) is absorbing, while all other states are transient, and if $i = N$ the transitions are given by:

$$(N,j) \longrightarrow \begin{cases} (N,j+1) & \text{with probability } \frac{N-j}{N} \\ (N,j) & \text{with probability } \frac{j}{N} \ . \end{cases}$$

To find the waiting time to complete the collections we have to determine $k_{(0,0)}^{(N,N)}$ solving a linear system, as in the one-collection case. For large values

of N we will need the help of a computer software, but for small values the computation can be carried out as shown in the following example.

Let us take $N = 3$. In this case the state space is

$$S = \{(0,0),(1,0),(1,1),(2,0),(2,1),(2,2),(3,0),(3,1),$$
$$(3,2),(3,3)\}$$

with $|S| = \frac{(N+1)(N+2)}{2} = 10$ and the transition matrix is given by:

$$P = \begin{pmatrix}
0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & \frac{1}{3} & \frac{2}{3} & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & \frac{1}{3} & 0 & \frac{2}{3} & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & \frac{2}{3} & 0 & \frac{1}{3} & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & \frac{1}{3} & \frac{1}{3} & 0 & \frac{1}{3} & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & \frac{2}{3} & 0 & 0 & \frac{1}{3} & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{3} & \frac{2}{3} & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{2}{3} & \frac{1}{3} \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1
\end{pmatrix}.$$

To determine the waiting time to complete the collections we have to solve the following linear sys-

tem:

$$\begin{cases} k_{(0,0)} = k_{(1,0)} + 1 \\ k_{(1,0)} = \frac{1}{3}k_{(1,1)} + \frac{2}{3}k_{(2,0)} + 1 \\ k_{(1,1)} = \frac{1}{3}k_{(1,1)} + \frac{2}{3}k_{(2,1)} + 1 \\ k_{(2,0)} = \frac{2}{3}k_{(2,1)} + \frac{1}{3}k_{(3,0)} + 1 \\ k_{(2,1)} = \frac{1}{3}k_{(2,1)} + \frac{1}{3}k_{(2,2)} + \frac{1}{3}k_{(3,1)} + 1 \\ k_{(2,2)} = \frac{2}{3}k_{(2,2)} + \frac{1}{3}k_{(3,2)} + 1 \\ k_{(3,0)} = k_{(3,1)} + 1 \\ k_{(3,1)} = \frac{1}{3}k_{(3,1)} + \frac{2}{3}k_{(3,2)} + 1 \\ k_{(3,2)} = \frac{2}{3}k_{(3,2)} + 1 \\ k_{(3,3)} = 0 \end{cases}$$

$$\Longrightarrow \begin{cases} k_{(3,3)} = 0 \\ k_{(3,2)} = 3 \\ k_{(3,1)} = \frac{9}{2} = 4.5 \\ k_{(3,0)} = \frac{11}{2} = 5.5 \\ k_{(2,2)} = 6 \\ k_{(2,1)} = \frac{27}{4} = 6.75 \\ k_{(2,0)} = \frac{22}{3} \approx 7.33 \\ k_{(1,1)} = \frac{33}{4} = 8.25 \\ k_{(1,0)} = \frac{311}{36} \approx 8.64 \\ k_{(0,0)} = \frac{347}{36} \approx 9.64 \end{cases}$$

6 Multiple collections with unequal probabilities

To the best of our knowledge, the case of multiple collections under the hypothesis of unequal probabilities has never been considered in the literature. This is clearly a challenging problem and it is difficult to present nice or at least exact expressions that represent the expected number of coupons needed. In the next section, we will see why the *Maximum-Minimums approach* does not work in this case, but we will derive at least an upper bound for the exact expected value. In the second section we will see how, at least theoretically, the Markovian approach is still applicable, but to obtain some explicit formulas, also in the easy cases, we need to make use of some software

like *Mathematica*. To conclude, we present a couple of examples solved explicitly and derive through simulation the value for higher values of the number of different coupons and several collections. We provide the simulation in the cases of the uniform distribution and the Mandelbrot distribution.

6.1 The Maximum-Minimums approach

Assume that we want to complete m sets of N coupons in the case of unequal probabilities.
Let X_1 be the random number of items that we need to collect to obtain m copies of the first coupon of type 1, let X_2 be the number of items that we need to collect to obtain m copies of the first coupon of type 2 and in general let X_i be the number of items that we need to collect to obtain m copies of the first coupon of type i; the waiting time to complete the collections is given by the

random variable $X = \max(X_1, \ldots, X_N)$.
To compute $\mathbb{E}[X]$ we can use the *Maximum-Minimums Identity*, as in the single collection case:

$$\mathbb{E}[X] = \sum_i \mathbb{E}[X_i] - \sum_{i<j} \mathbb{E}[\min(X_i, X_j)] + \sum_{i<j<k} \mathbb{E}[\min(X_i, X_j, X_k)] + \ldots \cdots + (-1)^{N+1} \mathbb{E}[\min(X_1, X_2, \ldots, X_N)] \ .$$

Note that the random variable X_i has negative binomial law with parameters (m, p_i), hence $\mathbb{E}[X_i] = \frac{m}{p_i}$, but now it isn't possible to compute the exact law of the random variables $\min_{i<j}(X_i, X_j)$, $\min_{i<j<k}(X_i, X_j, X_k), \ldots$. However, it is possible to prove that, for $i \neq j$:

$$\mathbb{E}[T(i, j; 2)] \leq \mathbb{E}[\min(X_i, X_j)] \leq \mathbb{E}[Z(i, j; 2)] \ ,$$

where $T(i,j;2)$ has negative binomial law with parameters $(m, p_i + p_j)$ and $Z(i,j;2)$ has negative binomial law with parameters $(m+1, p_i + p_j)$.

In general, for $2 \leq k \leq N$:

$$\begin{aligned}\mathbb{E}[T(i_1, i_2, \ldots, i_k; k)] \leq \mathbb{E}[\min(X_{i_1}, X_{i_2}, \ldots, X_{i_k})] \\ \leq \mathbb{E}[Z(i_1, i_2, \ldots, i_k; k)] \ ,\end{aligned}$$

where $T(i_1, i_2, \ldots, i_k; k)$ has negative binomial law with parameters $(m, p_{i_1} + \cdots + p_{i_k})$ and $Z(i_1, i_2, \ldots, i_k; k)$ has negative binomial law with parameters $(m(N-1)+1, p_{i_1} + \cdots + p_{i_k})$.

Using this consideration, we can give an upper

bound for $\mathbb{E}[X]$:

$$\begin{aligned}\mathbb{E}[X] &\leq \sum_i \mathbb{E}[X_i] - \sum_{i<j} \mathbb{E}[T(i,j;2)] \\ &\quad + \sum_{i<j<k} \mathbb{E}[Z(i,j,k;3)] \\ &\quad - \sum_{i<j<k<l} \mathbb{E}[T(i,j,k,l;4)] + \dots \\ &= \sum_i \frac{m}{p_i} - \sum_{i<j} \frac{m}{p_i+p_j} + \sum_{i<j<k} \frac{m\,(N-1)+1}{p_i+p_j+p_k} \\ &\quad - \sum_{i<j<k<l} \frac{m}{p_i+p_j+p_k+p_l} + \dots\end{aligned}$$

6.1.1 Examples

In the case of uniform probabilities, we have:

- $m = 1,\ N = 2,\ p_1 = p_2 = \frac{1}{2}$:

$$3 = \mathbb{E}[X] \leq \frac{1}{p_1} + \frac{1}{p_2} - \frac{1}{p_1+p_2} = 2+2-1 = 3$$

- $m = 1,\ N = 3,\ p_1 = p_2 = p_3 = \frac{1}{3}$:

$$\begin{aligned} 5.5 = \mathbb{E}[X] &\leq \frac{1}{p_1} + \frac{1}{p_2} + \frac{1}{p_3} \\ &\quad - \left(\frac{1}{p_1 + p_2} + \frac{1}{p_1 + p_3} + \frac{1}{p_2 + p_3} \right) \\ &\quad + \frac{1 \cdot 2 + 1}{p_1 + p_2 + p_3} \\ &= 3 \cdot 3 - 3 \cdot \frac{3}{2} + 3 = \frac{15}{2} = 7.5 \end{aligned}$$

- $m = 2,\ N = 2,\ p_1 = p_2 = \frac{1}{2}$:

$$5.5 = \mathbb{E}[X] \leq \frac{2}{p_1} + \frac{2}{p_2} - \frac{2}{p_1 + p_2} = 2 \cdot 2 \cdot 2 - 2 = 6$$

- $m = 2,\ N = 3,\ p_1 = p_2 = p_3 = \frac{1}{3}$:

$$\begin{aligned} 9.64 \approx \mathbb{E}[X] &\leq \frac{2}{p_1} + \frac{2}{p_2} + \frac{2}{p_3} \\ &\quad - \left(\frac{2}{p_1 + p_2} + \frac{2}{p_1 + p_3} + \frac{2}{p_2 + p_3} \right) \\ &\quad + \frac{2 \cdot 2 + 1}{p_1 + p_2 + p_3} \\ &= 3 \cdot 2 \cdot 3 - 3 \cdot 2 \cdot \frac{3}{2} + 5 = 14 \end{aligned}$$

In the case of unequal probabilities, we have:

- $m = 2,\ N = 2,\ p_1 = \frac{1}{3},\ p_2 = \frac{2}{3}$:

$$\mathbb{E}[X] \leq \frac{2}{\frac{1}{3}} + \frac{2}{\frac{2}{3}} - \frac{2}{\frac{1}{3} + \frac{2}{3}} = 6 + 3 - 2 = 7$$

- $m = 2,\ N = 2,\ p_1 = \frac{1}{6},\ p_2 = \frac{5}{6}$:

$$\mathbb{E}[X] \leq \frac{2}{\frac{1}{6}} + \frac{2}{\frac{5}{6}} - \frac{2}{\frac{1}{6} + \frac{5}{6}} = 12 + \frac{12}{5} - 2 = \frac{62}{5} = 12.4$$

- $m = 2,\ N = 3,\ p_1 = \frac{1}{2},\ p_2 = \frac{1}{3},\ p_3 = \frac{1}{6}$:

$$\begin{aligned}
\mathbb{E}[X] \leq\ & \frac{2}{\frac{1}{2}} + \frac{2}{\frac{1}{3}} + \frac{2}{\frac{1}{6}} \\
& - \left(\frac{2}{\frac{1}{2} + \frac{1}{3}} \frac{2}{\frac{1}{2} + \frac{1}{6}} + \frac{2}{\frac{1}{3} + \frac{1}{6}} \right) \\
& + \frac{2 \cdot 2 + 1}{\frac{1}{2} + \frac{1}{3} + \frac{1}{6}} \\
& = 4 + 6 + 12 - \frac{12}{5} - 3 - 4 + 5 \\
& = \frac{88}{5} = 17.6
\end{aligned}$$

- $m = 2,\ N = 3,\ p_1 = \frac{1}{4},\ p_2 = \frac{3}{8},\ p_3 = \frac{3}{8}$:

$$\begin{aligned}
\mathbb{E}[X] &\leq \frac{2}{\frac{1}{4}} + \frac{2}{\frac{3}{8}} + \frac{2}{\frac{3}{8}} \\
&\quad - \left(\frac{2}{\frac{1}{4} + \frac{3}{8}} + \frac{2}{\frac{1}{4} + \frac{3}{8}} + \frac{2}{\frac{3}{8} + \frac{3}{8}} \right) \\
&\quad + \frac{2 \cdot 2 + 1}{\frac{1}{4} + \frac{3}{8} + \frac{3}{8}} \\
&= 8 + 2 \cdot \frac{16}{3} - 2 \cdot \frac{16}{5} - \frac{8}{3} + 5 \\
&= \frac{73}{5} = 14.6 \ .
\end{aligned}$$

6.2 The Markov Chains approach

We can solve the case $m = 2$ with unequal probabilities using the Markov Chains, but to compute the expected time to complete the collections we have to use *Mathematica*, because the cardinality of the state space S grows rapidly.
If $N = 2$, let i_1, i_2 represent the coupons of the first collector and let j_1, j_2 represent the coupons of the second collector, hence the general state of the

Markov Chain will be $(i_1, i_2|j_1, j_2)$, and the state space is given by:

$$S = \{(0,0|0,0), (1,0|0,0), (0,1|0,0), (1,1|0,0),$$
$$(1,0|1,0), (0,1|0,1), (1,1|1,0), (1,1|0,1), (1,1|1,1)\} .$$

Let p_1 be the probability of getting a coupon of type 1 and let $p_2 = 1 - p_1$ be the probability of getting a coupon of type 2; the transition matrix is given by:

$$P = \begin{pmatrix} 0 & p_1 & p_2 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & p_2 & p_1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & p_1 & 0 & p_2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & p_1 & p_2 & 0 \\ 0 & 0 & 0 & 0 & p_1 & 0 & p_2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & p_2 & 0 & p_1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & p_1 & 0 & p_2 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & p_2 & p_1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} .$$

Solving the linear system

$$\begin{cases} k_{(1,1|1,1)} = 0 \\ k_i = 1 + \sum_{j \neq (1,1|1,1)} p_{ij} k_j, \quad i \neq (1,1|1,1) \end{cases}$$

it follows that the expected number of coupons needed to complete the collections is

$$k_{(0,0|0,0)} = \frac{2\,(p_1^4 - 2\,p_1^3 + p_1 - 1)}{p_1\,(p_1 - 1)}\ .$$

Letting p_1 vary from 0 to 1 we obtain the plot in Figure 6.

Consider now the case $N = 3$, let i_1, i_2, i_3 represent the coupons of the first collector and let j_1, j_2, j_3 represent the coupons of the second collector, hence the general state of the Markov Chain will be $(i_1, i_2, i_3|j_1, j_2, j_3)$, and the state space is

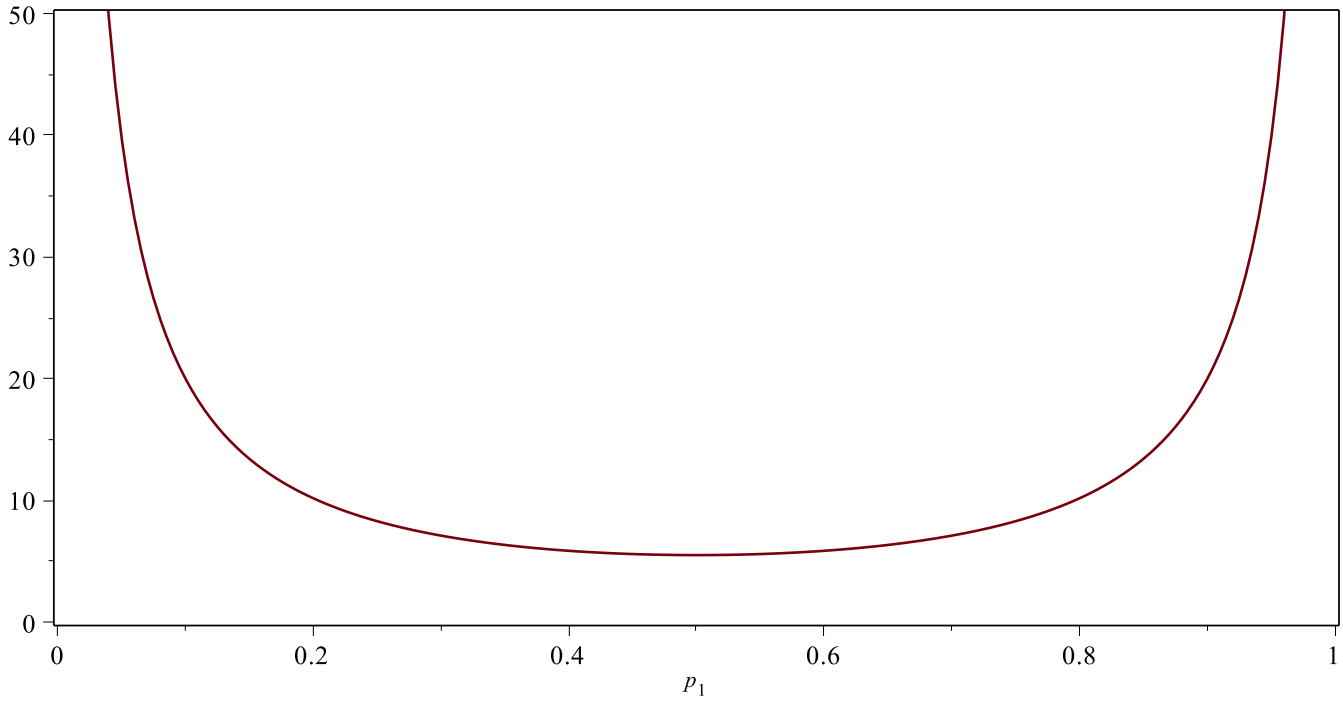


Figure 6: Expected number of coupons when $m = 2$, $N = 2$ in terms of p_1

given by:

$$
\begin{aligned}
S = \{&(0,0,0|0,0,0),(1,0,0|0,0,0),(0,1,0|0,0,0),(0,0,1|0,0,0),\\
&(1,0,0|1,0,0),(1,1,0|0,0,0),(1,0,1|0,0,0),(0,1,0|0,1,0),\\
&(0,1,1|0,0,0),(0,0,1|0,0,1),(1,1,0|1,0,0),(1,0,1|1,0,0),\\
&(1,1,0|0,1,0),(1,1,1|0,0,0),(1,0,1|0,0,1),(0,1,1|0,1,0),\\
&(0,1,1|0,0,1),(1,1,0|1,1,0),(1,1,1|1,0,0),(1,0,1|1,0,1),\\
&(1,1,1|0,1,0),(1,1,1|0,0,1),(0,1,1|0,1,1),(1,1,1|1,1,0),\\
&(1,1,1|1,0,1),(1,1,1|0,1,1),(1,1,1|1,1,1)\}\ .
\end{aligned}
$$

Let p_1 be the probability of getting a coupon of type 1, let p_2 be the probability of getting a coupon of type 2 and let $p_3 = 1 - p_1 - p_2$ be the probability of getting a coupon of type 3; the transition matrix has order 27 and solving the linear system

$$\begin{cases} k_{(1,1,1|1,1,1)} = 0 \\ \\ k_i = 1 + \displaystyle\sum_{j \neq (1,1,1|1,1,1)} p_{ij} k_j, \quad i \neq (1,1,1|1,1,1) \end{cases}$$

we get the expected waiting time to complete the

collections:

$$
\begin{aligned}
k_{(0,0,0|0,0,0)} = 2 \Bigg[& 1 - \frac{1}{1-p_1} + \frac{1}{p_1} - \frac{1}{1-p_2} \\
& + \frac{1}{p_2} + \frac{1}{1-p_1-p_2} + \frac{8\,p_1^3\,p_2\,(1-p_1-p_2)}{(1-p_1)^3} \\
& - \frac{3\,p_1^4\,p_2\,(1-p_1-p_2)}{(1-p_1)^3} - \frac{1}{p_1+p_2} \\
& + p_1^2 \left(-1 + \frac{1}{(1-p_2)^3} - \frac{6\,p_2\,(1-p_1-p_2)}{(1-p_1)^3} \right. \\
& \qquad \left. + \frac{1}{(p_1+p_2)^3} \right) \\
& + p_1 \left(1 - \frac{1}{(1-p_2)^2} - \frac{1}{(p_1+p_2)^2} \right) \Bigg] \ .
\end{aligned}
$$

Letting p_1 vary from 0 to 1 and p_2 from 0 to $1-p_1$ (to satisfy the constraint $p_1+p_2 \leq 1$) we obtain the plot of the level sets of the above function in Figure 7; the minimum of this surface corresponds to the point $(p_1, p_2) = (\frac{1}{3}, \frac{1}{3})$.

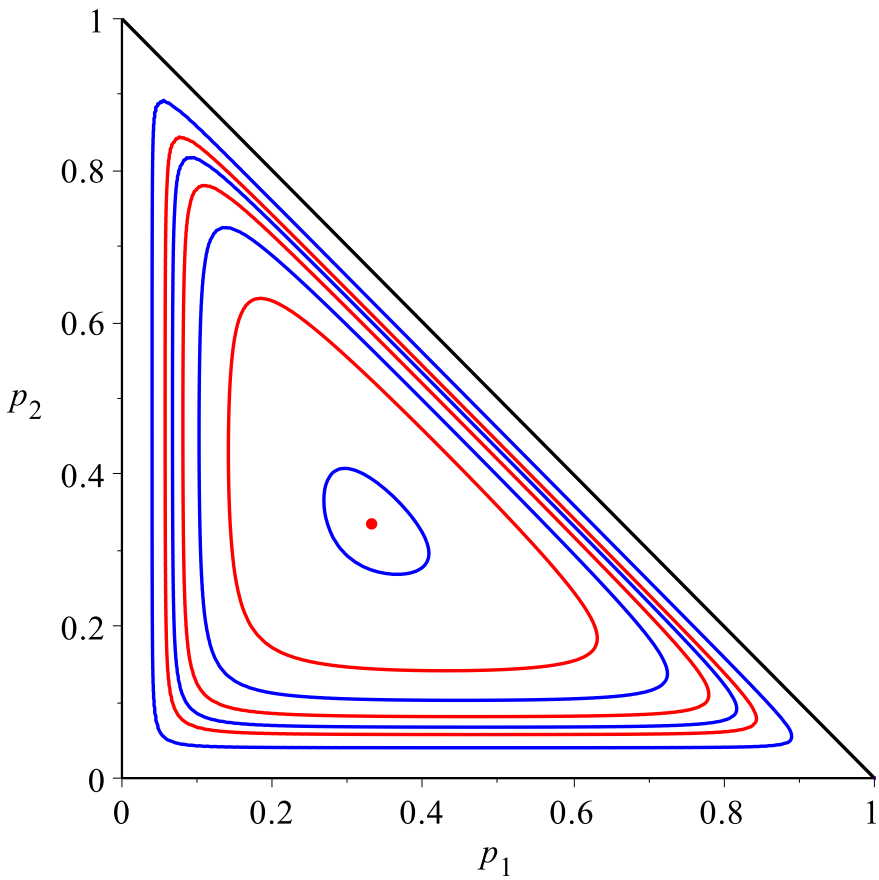


Figure 7: Level sets of the function giving the expected number of coupons when $m = 2$, $N = 3$ ($k = 50, 35, 30, 25, 20, 15, 10, 347/36 \approx 9.64$)

6.3 Examples

- if $m = 2,\ N = 2$:
  - if $p_1 = \frac{1}{2},\ p_2 = 1 - p_1 = \frac{1}{2}$: $\mathbb{E}[X] = \frac{11}{2} = 5.5 \leq 6$
  - if $p_1 = \frac{1}{3},\ p_2 = 1 - p_1 = \frac{2}{3}$: $\mathbb{E}[X] = \frac{59}{9} \approx 6.56 \leq 7$
  - if $p_1 = \frac{1}{6},\ p_2 = 1 - p_1 = \frac{5}{6}$: $\mathbb{E}[X] = \frac{1091}{90} \approx 12.12 \leq 12.4$
- if $m = 2,\ N = 3$:
  - if $p_1 = \frac{1}{3},\ p_2 = \frac{1}{3},\ p_3 = \frac{1}{3}$: $\mathbb{E}[X] = \frac{347}{36} \approx 9.64 \leq 14$
  - if $p_1 = \frac{1}{2},\ p_2 = \frac{1}{3},\ p_3 = \frac{1}{6}$: $\mathbb{E}[X] = \frac{240307}{18000} \approx 13.35 \leq 17.6$
  - if $p_1 = \frac{1}{4},\ p_2 = \frac{3}{8},\ p_3 = \frac{1}{3}$: $\mathbb{E}[X] = \frac{492697}{48000} \approx 10.26 \leq 14.6$

Note that the upper bound that we have found using the *Maximum-Minimums Identity* holds.

6.4 Numerical simulation

It is clear that in the XXI Century we cannot end this section without a numerical simulation. This allows us to see what happens beyond the Pillars of Hercules of our exact, even often not very useful, explicit formulas.

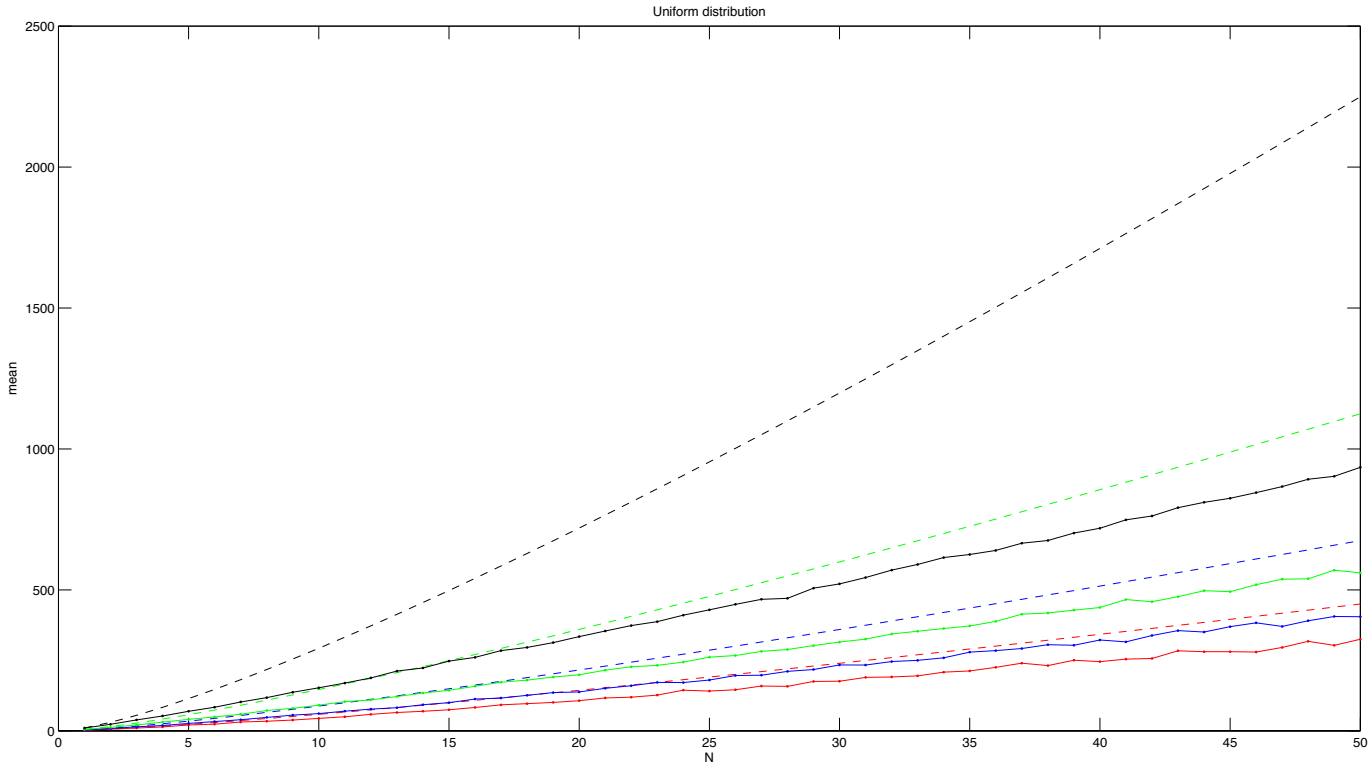


Figure 8: Uniform distribution

Uniform	$m = 1$	$2 \cdot (m = 1)$	$m = 2$	$3 \cdot (m = 1)$	$m = 3$
$N = 1$	1	2	2	3	3
$N = 2$	3	6	5.7	9	8.04
$N = 3$	5.5	11	9.78	16.5	13.97
$N = 4$	8.3333	16.667	13.88	25	18.66
$N = 5$	11.4167	22.8333	20.8	34.25	26.2
$N = 10$	29.2897	58.5794	44.05	87.869	61.06
$N = 15$	49.7734	99.5469	74.76	149.3203	99.73
$N = 20$	71.9548	143.9096	106.75	215.8644	137.75
$N = 30$	119.8496	239.6992	175.82	359.5488	233.68
$N = 40$	171.1417	342.2834	245.39	513.4252	322.16
$N = 50$	224.9603	449.9205	324.96	674.8808	404.68
$N = 100$	518.7378	1037.5	735.06	1556.2	927.55

Uniform	$m = 1$	$5 \cdot (m = 1)$	$m = 5$	$10 \cdot (m = 1)$	$m = 10$
$N = 1$	1	5	5	10	10
$N = 2$	3	15	12.35	30	23.21
$N = 3$	5.5	27.5	20.8	55	38.49
$N = 4$	8.3333	41.6667	30.72	83.3333	52.65
$N = 5$	11.4167	57.0833	40.69	114.1667	69.35
$N = 10$	29.2897	146.4484	91.16	292.8968	152.75
$N = 15$	49.7734	248.8672	143.56	497.7343	247.51
$N = 20$	71.9548	359.774	198.93	719.5479	334.09
$N = 30$	119.8496	599.2481	314.97	1198.5	521.22
$N = 40$	171.1417	855.7086	437.25	1711.4	718.86
$N = 50$	224.9603	1124.8	560.89	2249.6	934.86
$N = 100$	518.7378	2593.7	1220.4	5187.4	1986.2

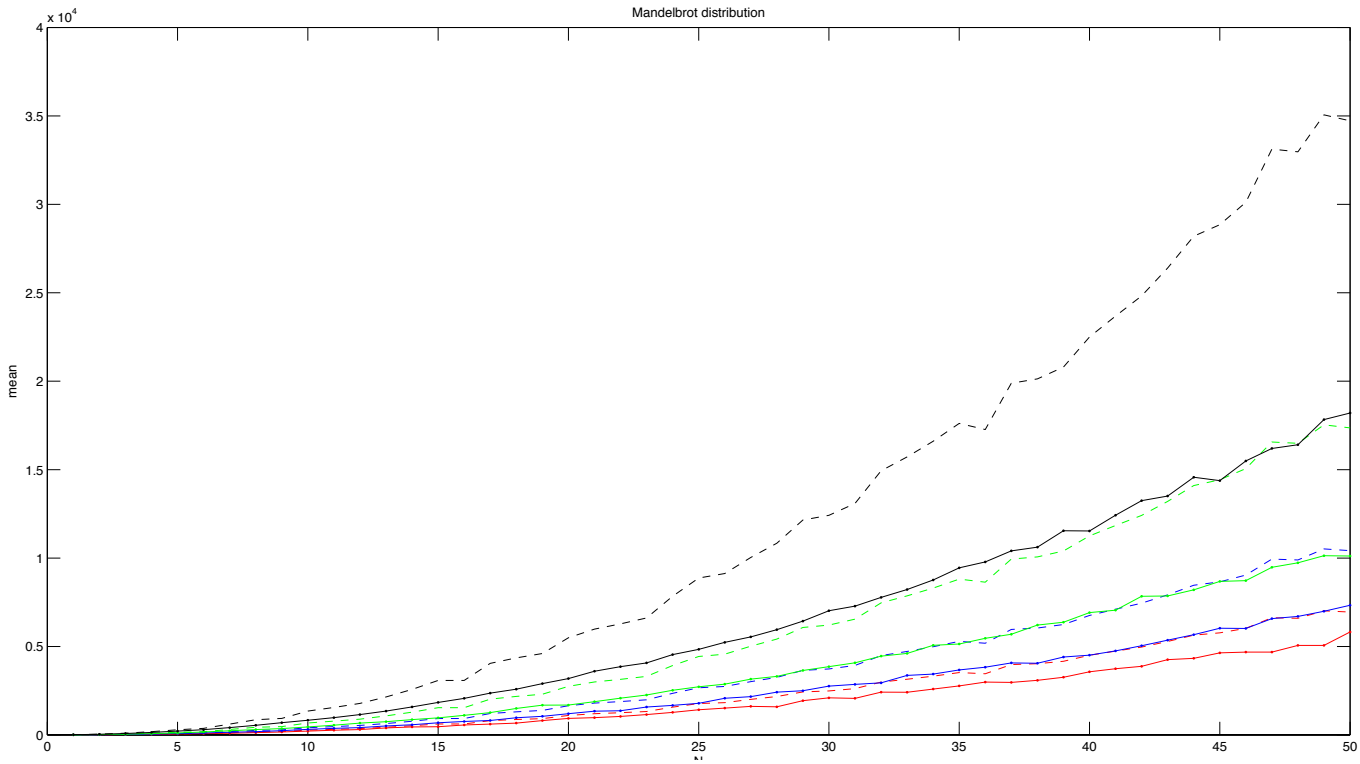


Figure 9: Mandelbrot distribution

Mandelbrot	$m = 1$	$2 \cdot (m = 1)$	$m = 2$	$3 \cdot (m = 1)$	$m = 3$
$N = 1$	1	2	2	3	3
$N = 2$	4.09	8.18	7.72	12.27	11.72
$N = 3$	9.39	18.78	19.55	28.17	25.99
$N = 4$	17.98	35.96	32.14	53.94	46.87
$N = 5$	28.56	57.12	52.91	85.68	68.89
$N = 10$	134.34	268.68	215.41	403.02	310.24
$N = 15$	307.77	615.54	462.25	923.31	685.5
$N = 20$	549.26	1098.3	930.76	1647.8	1196.6
$N = 30$	1241.9	2483.8	2096.8	3725.6	2753.9
$N = 40$	2250.5	4500.9	3565.2	6751.4	4507.1
$N = 50$	3472.2	6944.5	5820.6	10417	7332.7

Mandelbrot	$m = 1$	$5 \cdot (m = 1)$	$m = 5$	$10 \cdot (m = 1)$	$m = 10$
$N = 1$	1	5	5	10	10
$N = 2$	4.09	20.45	17.92	40.9	34.96
$N = 3$	9.39	46.95	42.72	93.9	79.46
$N = 4$	17.98	89.9	73.69	179.8	138.48
$N = 5$	28.56	142.8	111	285.6	215.68
$N = 10$	134.34	671.7	445.88	1343.4	827.25
$N = 15$	307.77	1538.8	959.4	3077.7	1832.8
$N = 20$	549.26	1098.3	930.76	1647.8	1196.6
$N = 30$	1241.9	6209.4	3860.6	12419	7018.5
$N = 40$	2250.5	11252	6913.1	22505	11531
$N = 50$	3472.2	17361	10118	34722	18202

In figures 8 and 9 we have considered the present case with several cooperative collectors and an arbitrary number of coupons that arrives according to the uniform probability distribution or to a Mandelbrot distribution with parameters $c = 0.30$ and $\theta = 1.75$. To see how big will be in this case the expected number of coupons to be collected in order to complete all the collections, we have simulated a big number of virtual collections. Then we have evaluated the arithmetic mean of the number of coupons needed in these simulated collections to complete the sets.

We also have plotted the numerical simulated values (solid line) and compared them with m times the expected number of coupons needed to complete a single collection (dashed line), again trough a simulation, to see how much money we save in the case of a cooperative collection; in par-

ticular, we have plotted in red the case $m = 2$, in blue the case $m = 3$, in green the case $m = 5$ and in black the case $m = 10$.

From these computations we observe that if $N = 50$, in the case of the uniform distribution we can save about the 27% of our money in the case of $m = 2$ and that this increases up to a 58% when $m = 10$, while in the case of the Mandelbrot distribution we can save about the 16% of our money in the case of $m = 2$ and about the 47% when $m = 10$.

References

[1] M. Ferrante, N. Frigo, *A note on the coupon-collector's problem with multiple arrivals and the random sampling*, arXiv:1209.2667v2, 2012.

[2] M. Ferrante, N. Frigo, *On the expected num-*

ber of different records in a random sample, arXiv:1209.4592v1, 2012.

[3] P. Flajolet, D. Gardy, L. Thimonier, *Birthday paradox, coupon collectors, caching algorithms and self-organizing search*, Discrete Appl. Math., 39:207-229, 1992.

[4] L. Holst, *On birthday, collectors', occupancy and other classical urn problems*, Internat. Statist. Rev., 54:15-27, 1986.

[5] D. van Leijenhorst, Th. van der Weide, *A formal derivation of Heaps' Law*, Information Sciences, 170: 263-272, 2005.

[6] T. Nakata, *Coupon collector's problem with unlike probabilities*, Preprint, 2008.

[7] D. J. Newman, L. Shepp, *The double dixie cup problem*, American Mathematical Monthly, 67:58-61, 1960.

[8] S. Ross, *A first course in probability*, 9th Edition, Pearson, 2012.

[9] W. Stadje, *The collector's problem with group drawings*, Advances in Applied Probability, 22:866-882, 1990.

[10] H. von Schelling, *Coupon collecting for unequal probabilities*, American Mathematical Monthly, 61:306-311, 1954.

Marco Ferrante
Dipartimento di Matematica
Università degli Studi di Padova
Padova, Italy
ferrante@math.unipd.it

Monica Saltalamacchia

Dipartimento di Matematica

Università degli Studi di Padova

Padova, Italy

monicasaltalamacchia@tiscali.it

Publicat el 25 de magi de 2014