SJÄLVSTÄNDIGA ARBETEN I MATEMATIK

MATEMATISKA INSTITUTIONEN, STOCKHOLMS UNIVERSITET

# Versatility of the Coupon Collector's Problem

av

**Karl Forsbäck**

2022 - No K11

MATEMATISKA INSTITUTIONEN, STOCKHOLMS UNIVERSITET, 106 91 STOCKHOLM

# Versatility of the Coupon Collector's Problem

Karl Forsbäck

Självständigt arbete i matematik 15 högskolepoäng, grundnivå

Handledare: Pieter Trapman

2022

# Versatility of the Coupon Collector's Problem

Karl Forsbäck

March 29, 2022

## Abstract

The Coupon Collector's Problem asks the question of how many coupons, belonging to a set where the probability of all coupons is equal, must be purchased at random before at least one of each type has been acquired.

This thesis explores this specific problem, as well as the generalization to unequal probabilities, using two different methods. The traditional discrete approach, where one coupon is collected at each time unit, makes use of the geometric distribution and the harmonic series. The alternative way is to utilize the Poisson Process. There coupons are collected continuously at an independent probabilistic rate of 1 per time unit, where the times between coupons are exponentially distributed. An example calculating the expected number of draws needed to collect all coupons will be given for each method to illustrate the similarities and differences.

The discrete method is simpler to understand initially, however the continuous approach makes the generalization as well as a mathematical analysis of the problem significantly simpler. This will all be illustrated in a few examples as well.

Additionally, it covers two applications of the generalized problem, the Pokemon games and unfair dice. We will see how the games core mechanic, to discover unique creatures, can be explained and quantified using the Poisson Process method. Using the very same method we will show that the way numbers appear on any unfair dice may have a distinctly different distribution from a regular fair dice.

A very special thank you to my supervisor Pieter Trapman. Who both suggested the problem for this thesis, as well as helped me understand all the new and difficult concepts and formulae.
Additional thanks go out to my friends Alexander Ljungberg and Alexander Thurdin who took their time reading and commenting on the text.
Finally, a thank you to coffee.

# Contents

# 1

# Introduction

The purpose of this thesis is to investigate the Coupon Collector's problem, a classic problem of probability theory. It asks the question of how many coupons, belonging to a set where the probability of all coupons is equal, must be purchased at random before at least one of each type has been acquired. As well as its generalization, some applications and two methods that can be used to solve it.

First is the discrete method. It is simpler and require only a high-school level of mathematics to understand. However, it is quite limited and cannot be used to generalize the problem in an effective manner. The continuous method, Poisson Processes, is a much more powerful tool. More information about probability distributions is retained and most generalizations become easy, if not trivial, once you understand it. A thorough description and brief history of the Coupon Collector's Problem is given in Section 2.

Section 3 serves as an introduction to the mathematics used to solve the problem. We will cover three probability distributions, the Geometric, Poisson and Exponential distributions, and clarify some notation used later in the report. Additionally we will define the Poisson Process and prove that it is applicable to the problem at hand.

Section 4 and 5 are the heart of the thesis. In Section 4 we will solve the Coupon Collector's Problem using our two methods, and in Section 5 we will look at the theory of generalizing the problem further. We will see that it is very hard when using the discrete method, but no more difficult than the specific case if you use the continuous method.

In Section 6 we will look at two applications of the generalized problem. First is the Pokemon games, where we will see how the expected time until one of the main goals of the games is completed can be quantified. The second application pertains to unfair (or loaded) dice. Particularly we will see that the unfair dice we discuss takes, on average,longer than a fair dice to roll each side once.

# 2

# The Coupon Collector's problem

A classic formulation of the Coupon Collector's problem reads as follows:

*"Suppose a brand of cereal includes a coupon with every box purchased. There are $n$ different coupons, each one equally likely to be in any one box. If every draw is independent from the others, what is the average amount of boxes that must be purchased before one of each type of coupon has been collected?"* This formulation has been paraphrased from the description in section 1 of The Coupon Collector's Problem [FeSa14]

Though this formulation is obviously modernized to the twentieth century, the history of the Coupon Collector's problem goes all the way back to the early 1700s, when it was first described by A. De Moivre in *De Mensura Sortis* (*On the Measurement of Chance*). Since then it has been worked on and expanded by numerous people, including such prolific names as Laplace and Euler.

The Coupon Collector's Problem describes so called "collect all and win" contests. Where the goal is, rather self explanatory, to collect at least one of each type in a collection of things in order to "win". Unlike the name suggests, the Coupon Collector's Problem describes more than just coupons. Rolling a dice until you have rolled each number at least once is the exact same problem.

You can also generalize the problem to "coupons" with unequal probabilities. This is something that is nearly trivial when using the continuous method, but highly laborious and impractical in the discrete case. Something we will see in section 5.

The mathematical model that the problem describes is quite broadly applicable. In section 6 we will explore two simple applications of the generalized Coupon Collector's Problem, but there are many more. Including in electrical engineering and biology. For further information on this and the problem's history, see [FeSa14].

# 3

# Theoretical background

Before we can solve the Coupon Collector's Problem, we must first understand the theory behind the methods we will utilize.

## 3.1 Relevant Probability Distributions and Little-o

**Definition 3.1.** Geometric distribution. A Geometric random variable counts the number of tries leading up to, and including, the first success. If every attempt has, independently of one another, a probability p of success, then $X$ is a geometrically distributed random variable with parameter $p$, and the probability mass function is:

$$p_X(n) = P\{X = n\} = (1-p)^{n-1}p,\ n = 1, 2, 3, ...$$

The geometric distribution is appropriate if:

1. The modeled phenomena is a a sequence of independent trials.

2. Every trial can either succeed or fail.

3. The probability of success is the same for each trial.

**Definition 3.2.** Poisson Distribution. A random variable $Y$, taking the values $0, 1, 2, 3, ...$ is said to have a Poisson distribution with parameter $\lambda$ if, for $\lambda > 0$:

$$p_Y(i) = P\{Y = i\} = e^{-\lambda}\frac{\lambda^i}{i!},\ i = 0, 1, 2, ...,$$

which naturally defines a probability mass function since $\sum_{i=0}^{\infty} \dfrac{\lambda^i}{i!} = e^{\lambda}$. The Poisson distribution is intimately connected to the Exponential distribution. As will be shown in Section 3.3, the interarrival (or waiting) times of the Poisson Process are exponential random variables.

**Definition 3.3.** The Exponential distribution. A continuous random variable $Z$ is Exponentially distributed with parameter $\lambda > 0$ if:

$$P\{Z > t\} = \int_0^\infty \lambda e^{-\lambda t} dt = e^{-\lambda t},\ t > 0.$$

In addition, the exponential distribution is memoryless. A random variable is without memory if the probability for an event to take place, after a certain time $t$, is independent of how much time has already passed when the count starts, i.e:

$$\begin{aligned} &P\{x > s + t \mid x > t\} = P\{Z > s\} \\ \Longleftrightarrow &P\{Z > s + t\} = P\{Z > s\}P\{Z > t\}. \end{aligned}$$

It is easily verified that the aforementioned exponential obeys this relation:

$$\begin{aligned} &P\{Z > s + t\} = P\{Z > s\}P\{Z > t\} \\ \Rightarrow\ &e^{-\lambda(s+t)} = e^{-\lambda s}e^{-\lambda t}. \end{aligned}$$

It turns out that having an Exponential distribution is equivalent to being without memory for continuous random variables. See [Fel71], chapter 1.3, page 8.

**Definition 3.4.** Little-o notation. A function $f(.)$ is said to be $o(h)$ if:

$$\lim_{h \to\ 0} \frac{f(h)}{h} = 0.$$

That is, the function shrinks faster than its argument. Some properties of such functions are:

1. If the functions $f$ and $g$ are both $o(h)$ respectively, then so is $f + g$.

$$\lim_{h\to 0} \frac{f(h) + g(h)}{h} = \lim_{h\to 0} \frac{f(h)}{h} + \lim_{h\to 0} \frac{g(h)}{h} = 0 + 0 = 0.$$

2. If $f$ is $o(h)$ then so is $cf$ for $c \in \mathbf{R}$.

$$\lim_{h\to 0} \frac{cf(h)}{h} = c \lim_{h\to 0} \frac{f(h)}{h} = c \cdot 0.$$

3. From 1. and 2. follow that any finite linear combination of functions who are $o(h)$, is also $o(h)$.

As an example, $f(x) = x^2$ is $o(h)$, but $g(x) = e^x - 1$ is not.

$$\lim_{h \to 0} \frac{h^2}{h} = \lim_{h \to 0} h = 0,$$
$$\lim_{h \to 0} \frac{e^h - 1}{h} \approx \lim_{h \to 0} \frac{1}{h} \left( h + \frac{h^2}{2} + \frac{h^3}{6} + ... \right) = \lim_{h \to 0} 1 + \frac{h}{2} + \frac{h^2}{6} + ... = 1.$$

Where, in the second equation, the Maclaurin Series of $e^x$ was used to find the limit.

The "little-o" notation is useful in giving more precise statements than could otherwise be made.

$$P(t < x < t + h) \approx f(t)h,$$

should then be read as:

$$P(t < x < t + h) = f(t)h + o(h).$$

## 3.2 The Poisson Process

In order to understand what the Poisson Process is, we must first know what a counting process in general is. A counting process is simply any process counting the number of events that has happened. Examples include; cars passing an intersection, babies born at a certain hospital and paintings sold at auction.

A counting process cannot take negative values and can never go down. Meaning that while the total amount of customers a store has in a day can be described by a counting process, the number of people inside the store at any one time can not be. Because people also leave the store.

**Definition 3.5.** The Poisson Process. The counting process $N(t)$ is a Poisson Process with parameter $\lambda$ if:

1. $N(0) = 0$. When we start counting, we start at zero.

2. $N(t), t > 0$ has independent and stationary increments.

3. $P[N(t + h) - N(t) = 0] = 1 - \lambda h + o(h)$

4. $P[N(t + h) - N(t) = 1] = \lambda h + o(h)$.

5. $P[N(t + h) - N(t) \geq 2] = o(h)$. In the limit of $h \longrightarrow 0$, the probability of having more than one event in any one time interval is effectively zero.

The Poisson Process does not get its name from Siméon Denis Poisson directly. He was neither the discoverer nor a student of the process. The name comes from the relation it has with the Poisson Distribution, that the number of events in any time interval of length $t$ is a Poisson random variable with mean $\lambda t$, which in turn was derived by Poisson. See [Sti00].

**Theorem 3.6.** *If $N(t), t \geq 0$ is a Poisson Process with rate $\lambda$, then for all $s, t > 0$, $N(s+t) - N(s)$ is a random variable with a Poisson distribution and mean $\lambda t$.*

*Proof.* The methodology for this proof was borrowed from [Ros14], page 299, Theorem 5.1.

To start, we will derive the Laplace transform of $N(t)$, which is $E[e^{-uN(t)}]$. Set $g(t) = E[e^{-uN(t)}]$ and fix $u > 0$. We will now find an expression for the transform by deriving a differential equation.

$$\begin{aligned} g(t+h) &= E[e^{-uN(t+h)}] \\ &= E[e^{-u(N(t)+N(t+h)-N(t))}] \\ &= E[e^{-uN(t)}e^{-u(N(t+h)-N(t))}] \\ &= E[e^{-uN(t)}]E[e^{-u(N(t+h)-N(t))}] \\ &= g(t)E[e^{-u(N(t+h)-N(t))}]. \end{aligned}$$

Using point 3 to 5 of Definition 3.5, the second factor can be reformulated as:

$$\begin{aligned} E[e^{-u(N(t+h)-N(t)}] &= 1 - \lambda h + o(h) + e^{-u}(\lambda h + o(h)) + o(h) \\ &= 1 - \lambda h + e^{-u}\lambda h + o(h), \\ \Rightarrow g(t+h) &= g(t)(1 + \lambda h(e^{-u} - 1) + o(h)) \\ \iff \frac{g(t+h) - g(t)}{h} &= g(t)\lambda(e^{-u} - 1) + \frac{o(h)}{h}. \end{aligned}$$

In the limit as $h \to 0$, the left hand side becomes the definition of the derivative of g(t), and the expression has now become the sought after differential equation.

$$\begin{aligned} g'(t) &= g(t)\lambda(e^{-u} - 1) \\ \iff \frac{g'(t)}{g(t)} &= \lambda(e^{-u} - 1), \end{aligned}$$

where the left hand side now can be recognized as the derivative of $ln(g(t))$. Integration gives us:

$$\begin{aligned} ln(g(t)) &= \lambda(e^{-u}-1)t + C \\ \iff g(t) &= exp[\lambda t(e^{-u}-1) + C]. \end{aligned}$$

Since $N(0) = 0 \implies g(0) = E[e^{-uN(0)}] = E[1] = 1$, $C$ must be zero and we have:

$$g(t) = exp[\lambda t(e^{-u}-1)].$$

Let us now perform a Laplace transform on a Poisson Random Variable with mean $\lambda t$ specifically:

$$\begin{aligned} &E[e^{-uX}],\ P\{X = i\} = e^{-\lambda t}\frac{(\lambda t)^i}{i!}, i = 0, 1, 2, ... \\ &E[e^{-uX}] = \sum_{i=0}^{\infty} e^{-\lambda t}\frac{(\lambda t)^i}{i!}e^{-ui} \\ &\quad = e^{-\lambda t}\sum_{i=0}^{\infty}\frac{(\lambda t e^{-u})^i}{i!} \\ &\quad = e^{-\lambda t}e^{\lambda t e^{-u}} \\ &\quad = exp[\lambda t(e^{-u}-1)], \end{aligned}$$

which is the same as before. Since the Laplace transform uniquely determines the distribution, see [Ros14], we can conclude that $N(t)$ is a Poisson random variable with parameter $\lambda t$.

In order to prove that $N(s+t) - N(s)$ is a poisson random variable we fix $s$ and let $N_s(t) = N(s+t) - N(s)$ be the number of events in time $t$ when we start our count at time $s$ instead of 0. The argument is then analogous with the preceding one, and we have shown that the number of events of a counting process in a time interval of length t is a Poisson random variable with parameter $\lambda t$. □

## 3.3 Interarrival Time of a Poisson Random Variable

The interarrival time is the time between events in a counting process. Let $T_1$ denote the time of the very first event, then $T_2$ is the time between the first and second event. $T_3$ will be the time between the second and third event, and so on. $T_n, n = 1, 2, 3, ...$ is called the sequence of interarrival times, the sum of which is the total time elapsed.

To determine the distribution of the $T_n$ we can notice that the time of the first event $T_1$ is greater than $t$ if, and only if, no events take place in $[0, t]$. Therefore:

$$P\{T_1 > t\} = P\{N(t) = 0\} = e^{-\lambda t}.$$

That the probability is an exponential comes from the fact that the counting process must be memoryless by virtue of its independent and stationary increments, and as mentioned in Definition 3.3, the exponential is the only continuous probability distribution to possess that property. The probability that the second event happens after time $t$ is:

$$P\{T_2 > t\} = E[P\{T_2 > t\} \mid T_1].$$

But because of the memoryless property, $T_2$ is unaffected by $T_1$:

$$\begin{aligned} P\{T_2 > t \mid T_1 = s\} &= P\{N(t+s) - N(s) = 0 \mid T_1 = s\} \\ &= P\{N(t+s) - N(s) = 0\} \\ &= e^{-\lambda t}. \end{aligned}$$

We can therefore conclude that $T_1$ and $T_2$ are independent Exponential random variables both with means $\frac{1}{\lambda}$. Repeating the argument gives us that the sequence of interarrival times, $T_n, n = 1, 2, 3, ...$, are all identically distributed independent random variables with mean $\frac{1}{\lambda}$. This result will be a key factor when using the Poisson Process to solve the Coupon Collector's Problem.

## 3.4 Splitting of Poisson processes

Consider a Poisson process with rate $\lambda > 0$ where events are, independently of one another, either type 1 with probability $p$ or type 2 with probability $1 - p$. For example, imagine flipping a coin. Every flip is an event in the counting process and can be either heads with probability $\frac{1}{2}$ or tails with probability $\frac{1}{2}$. Label these as type 1 and type 2 events respectively, and let $N_1(t)$ and $N_2(t)$ be the number of type 1 or 2 events in the interval $[0, t]$. Naturally, $N_1(t) + N_2(t) = N(t)$. It can be shown that $N_1(t)$ and $N_2(t)$, for $t \geq 0$, are independent Poisson Processes with rates $\lambda p$ and $\lambda(1 - p)$ respectively. That this statement applies to an arbitrary number of event types follows naturally. Label then the probabilities $p_1, p_2, ..., p_n$, for $n$ types of events. In the classic Coupon Collector's Problem, where all probabilities are equal, $p_1 = p_2 = ... = p_n = \frac{1}{n}$. For an informative argument and proof, see [Ros14], page 304, Proposition 5.2.

This property is not only useful, but necessary when using the Poisson Process to solve the Coupon Collector's Problem. Since every type of coupon drawn will be a type 1, 2, 3, and so on, event, with respective processes $N_X(t)$ and rates $\lambda p_X$, where $X$ stands for the event type.

# 4

# Solving the Coupon Collector's Problem

## 4.1 Discrete time

In discrete time we think of the problem exactly as it is formulated in Section 2. We draw a coupon and see if it is a new one, or not. If we draw one we already have, we try again. If it is a new one, we add it to the collection and start looking for the rest that we have yet to find. This continues until we have all $n$ coupons.

**Definition 4.1.** Discrete interarrival times. Let $T$ be the total time taken to complete the collection of $n$ different coupons. Here "time" stands for "number of draws", to stay consistent with the continuous case. Then let $t_i$ be the time it takes to collect the $i$-th coupon and let $p_i$ be the probability of success, per draw, to collect a new coupon after $i-1$ coupons have already been collected. $t_i$ is the interarrival time of the $i$-th coupon. The $t_i$ are geometric random variables with parameter $p_i$.

*Proof.* Let us show that the Coupon Collectors Problem follow the criteria of Definition 3.1:

1. Every draw is an attempt to acquire a new coupon. Per definition, every draw is independent of the last. Just like a roll of the dice.

2. Drawing a new coupon counts as a success and not doing so counts as a failure.

3. If $i-1$ coupons have already been drawn, then until we actually draw the $i$-th coupon, the probability of a success doesn't change.

Therefore, as stated in Definition 4.1, every individual $t_i$ is a geometrically distributed random variable. □

The total time $T$ is the sum of all the interarrival times $t_i$. $T = t_1 + t_2 + ... + t_n$. By virtue of the linearity of expectation values we can conclude that:

$$\begin{aligned} E[T] &= E[t_1 + t_2 + ... + t_n] \\ &= E[t_1] + E[t_2] + ... + E[t_n]. \end{aligned}$$

Since every $t_i$ is a geometric random variable, its expected value is $\frac{1}{p_i}$, the reciprocal of its probability.

While the probability of finding a new coupon is unchanged until we actually do, it does change when that happens. The probability of finding a new one on the very first draw is 1. Now that we have one the probability of finding any one we have yet to draw is $\frac{n-1}{n}$. The probability of finding a third is then $\frac{n-2}{n}$, and so on. Note that the probability is not "per draw", it is "per unique coupon we already have" or "per new coupon to be collected", however you prefer to think about it. We can express this as $p_i = \frac{n-i+1}{n}$, where $i = 1, 2, ..., n$ represents the $i$-th coupon to be collected after $i - 1$ already has been. The final expression for the expected value of draws needed is then:

$$\begin{aligned} E[T] &= E[t_1] + E[t_2] + ... + E[t_n] \\ &= \frac{n}{n} + \frac{n}{n-1} + ... + \frac{n}{1} \\ &= n \sum_{i=1}^{n} \frac{1}{i}. \end{aligned}$$

Which is simply the Harmonic series up to n, multiplied by n. As an example, consider a 20-sided dice. There are 20 sides and therefore $n = 20$:

$$n \sum_{i=1}^{n} \frac{1}{i} = 20 \sum_{i=1}^{20} \frac{1}{i} = 20 \left( \frac{1}{1} + \frac{1}{2} + \frac{1}{3} + ... + \frac{1}{20} \right) \approx 72. \tag{1}$$

It takes, on average, 72 rolls before each side has been rolled at least once. Since we are talking about discrete attempts in the end, I will round to the nearest whole number.

To further illustrate how the expected time grows with $n$. let us find its asymptotic behavior. That the following inequality holds true should not be surprising:

$$\int_0^n \frac{1}{x+1}dx \;\leq\; \sum_{i=1}^{n}\frac{1}{x} \;\leq\; 1+\int_1^n \frac{1}{x}dx$$
$$\Rightarrow \log(n+1) \;\leq\; \sum_{i=1}^{n}\frac{1}{x} \;\leq\; 1+\log(n).$$

Since the sum is neatly nestled between two logarithmic functions , we can say that the expected time, $n\sum_{i=1}^{n}\frac{1}{x}$, grows approximately like $n\log(n)$. Figure (1) shows a graph of the two functions as well as the first eight terms of the sum.

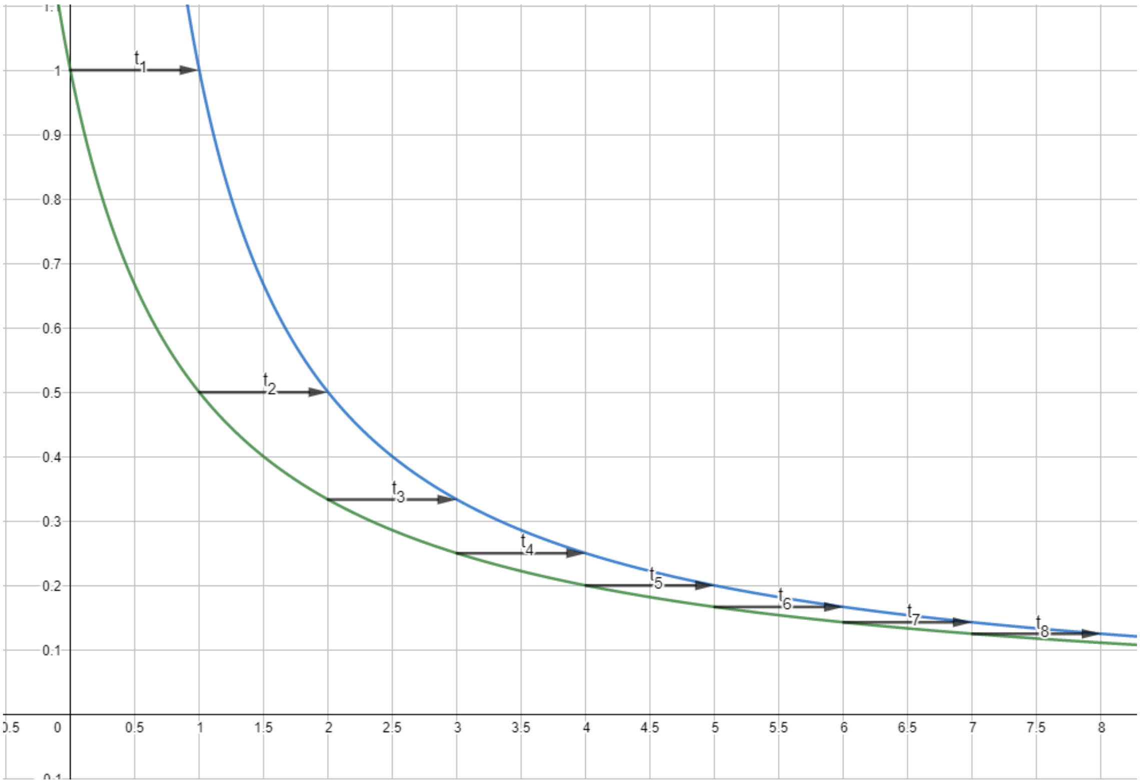

Figure 1: Green line $= \frac{1}{x+1}$, blue line $= \frac{1}{x}$, $t_1$ through $t_8$ represent the terms in the summation. Made with GeoGebra.

One final thing we are going to do is to calculate the variance of the Coupon Collector's Problem in discrete time. Or, rather, we are going to set an upper bound for it. The variance of a geometric random variable is $\frac{1-p}{p^2}$. See, for example, [Ros14], page 40 for a definition on how to find an expression for the variance. Using our expression for the probability of drawing a new coupon from before, $p_i = \frac{n-i+1}{n}$, we can express the variance as:

$$\begin{aligned} Var[T] &= Var[t_1 + t_2 + ... + t_n] \\ &= Var[t_1] + Var[t_2] + ... + Var[t_n] \\ &= \frac{1-p_1}{p_1^2} + \frac{1-p_2}{p_2^2} + ... \frac{1-p_n}{p_n^2}. \end{aligned}$$

Now we ignore the numerator. Since it is smaller than one and larger than zero, ignoring it only makes our answer larger, which is okay because we are trying to set an upper bound. Doing so gives us that:

$$\begin{aligned} Var[T] &< \frac{n^2}{n^2} + \frac{n^2}{(n-1)^2} + ... + \frac{n^2}{1^2} \\ &= n^2 \sum_{k=1}^{n} \frac{1}{k^2} \\ &< n^2 \frac{\pi^2}{6}, \end{aligned}$$

since $\sum_{k=1}^{\infty} \frac{1}{k^2} = \frac{\pi^2}{6}$. An upper bound for the variance is then that it grows like $n^2$. For our previously taken example of $n = 20$, the variance is bound by:

$$20^2 \cdot \frac{\pi^2}{6} \approx 658.$$

Here, some simplifying steps have been taken, and you can certainly do better in finding an upper bound for the variance of the total time taken.

There is nothing wrong with the methodology we used here, except that it retains little to no information about the probability distribution. Meaning it can only effectively answer the original question. For more versatility, we must look to another method.

## 4.2 Continuous time

Let us now look at the Coupon collectors Problem in continuous time. We want to find $E[N]$, the expected value of the total number of draws. However, in this case instead of drawing coupons discretely and analyzing the probabilities of success, we draw coupons at random times chosen in accordance with a Poisson Process with rate $\lambda = 1$.

Just like the discrete case we have $n$ different types of coupons, $j = 1, 2, ..., n$ and with equal probabilities. As discussed in Section 3.3, these are represented by individual Poisson Processes. Let $X_j$ denote the time of the first event of the $j$-th process. The larger the value, the later the event of drawing that particular coupon for the first time happens. The largest of the $X_j$ must therefore be the time when the last coupon is drawn. Let then $X = \text{max}_{1 \leq j \leq n} X_j$ be the time when a complete collection is amassed. By virtue of the $X_j$:s being independent random variables with exponential distributions, the probability of completing the collection before time $t$ (the cumulative distribution function) is given by the product:

$$
\begin{align}
P\{X \leq t\} &= P\{X_j \leq t, j = 1, 2, ..., n\} \tag{2}\\
&= \prod_{j=1}^{n} (1 - e^{-p_j t}) \tag{3}\\
&= (1 - e^{-\frac{t}{n}})^n. \tag{4}
\end{align}
$$

In the final step I have used the fact that all probabilities are equal and that there are n factors. This here is the probability density function of the whole Poisson Process for the Coupon Collector's Problem. To find the expectation value we simply integrate $P\{X > t\} = 1 - P\{X < t\}$ with respect to $t$:

$$
E[X] = \int_0^\infty (1 - (1 - e^{-\frac{1}{n}t})^n) dt. \tag{5}
$$

Evaluating the integral is a tedious affair of repeatedly multiplying parentheses together to get a sum of exponentials and then integrating. For small $n$ this is manageable by hand, but for large $n$ it is best left to a computer. An example calculation for $n = 4$ is given in section 5.2. The final step is to relate $E[X]$ to $E[N]$. Let $T_i$ be the $i$:th interarrival time of the Poisson Process. Then, naturally, the total time X is:

$$
X = \sum_{i=1}^{N} T_i.
$$

Since the $T_i$:s are independent exponentials with rate 1, they all have expectation 1 respectively, and by the linearity of expectation values and the standard relation $E[X] = E[E[X \mid N]]$, see [Wol10]:

$$
\begin{aligned}
&E[X \mid N] = E[\sum_{i=1}^{N} T_i] = NE[T_i] = N \\
&\implies E[X] = E[N].
\end{aligned}
$$

We can therefore conclude that the expected value of the total number of coupons needed is given by Equation (5).

Let us look at the same example as in section 4.1, the 20-sided dice. Since there are 20 sides $n = 20$:

$$E[X] = \int_0^\infty (1 - (1 - e^{-\frac{1}{20}t})^{20})\, dt \approx 72. \tag{6}$$

As stated previously, evaluating the integral is a tedious calculation. Performing it will reveal that, unsurprisingly, the answer is the same as in equation 1.

The Poisson Process has also given us a way to more accurately express the variance. $F(t)$ stands for the cumulative distribution function.

$$
\begin{aligned}
Var(X) &= E[X^2] - E[X]^2 \\
&= \int_0^\infty 2t\,(1 - F(t))\, dt - \left(\int_0^\infty (1 - F(t))dt\right)^2 \\
&= \int_0^\infty 2t\left(1 - (1 - e^{-\frac{t}{n}})^n\right) dt - \left(\int_0^\infty (1 - (1 - e^{-\frac{t}{n}})^n)dt\right)^2.
\end{aligned}
$$

Continuing to use the 20-sided dice as an example:

$$\int_0^\infty 2t\left(1 - (1 - e^{-\frac{t}{20}})^{20}\right) dt - \left(\int_0^\infty (1 - (1 - e^{-\frac{t}{20}})^{20})dt\right)^2 \approx 638.$$

Again, the evaluation of the integral is long and tedious and best left to a computer. Remember that the upper bound on the variance for $n = 20$ we found using the discrete method was 658. Which is larger than 638, but not much larger.

# 5

# Generalizations

## 5.1 Unequal probabilities

One natural generalization of the Coupon Collector's Problem is coupons with unequal probabilities. While it is certainly doable using the discrete approach, the Poisson Process makes this nearly trivial if you have an understanding of the specific case with all equal probabilities.

### 5.1.1 Discrete Time

In the case of discrete time I will start by quoting my supervisor Pieter Trapman: *"It will be a combinatorial hell."*

Suppose you have two types of coupons, $n = 2$, but with unequal probabilities ($p_1 = 0.75$ and $p_2 = 0.25$). Depending on which is drawn first, the expected time until completion will differ. Say that coupon 1 is drawn first, then there is a $\frac{1}{4}$ chance of drawing coupon 2 at any subsequent draw. Since this is a geometric random variable with parameter $p_2 = \frac{1}{4}$, the expected time until completion is $E_1[T] = \frac{1}{p_2} = 4$. Say now that we draw coupon 2 first instead. Then there is now a $\frac{3}{4}$ chance of drawing coupon 1 during subsequent draws. By the same argument then, the expected time until completion is $E_2[T] = \frac{1}{p_1} = \frac{4}{3}$. The first option happens 75 percent of the time, and the second happens 25 percent of the time. Conditioning on this, as well as the fact that one draw has already been performed to get there, the expression for the total expected time is:

$$E[T] = p_1 E_1[T] + p_2 E_2[T] + 1 = \frac{3}{4} \cdot 4 + \frac{1}{4} \cdot \frac{4}{3} + 1 = \frac{13}{3} \approx 4.$$

Consider now four types of coupons, with probabilities $p_1 = \frac{1}{10}, p_2 = \frac{2}{10}, p_3 = \frac{3}{10}$ and $p_4 = \frac{4}{10}$. If coupon 1 is drawn first then there is a $\frac{9}{10}$ chance of drawing a new one thereafter, if coupon 2 is drawn first then there is a $\frac{8}{10}$ chance, and so on for the remaining two. In all four cases, the probability of drawing a third new coupon depends again on which new coupon was drawn after the first. There are three possibilities for the second coupon, for each of theses there are two possibilities for the third coupon, and finally which is the fourth becomes obvious after this point. This means there are 24 possible paths. For each of them you must calculate the expected time for each step and weigh it based on which coupons have been drawn previously. I will not perform this calculation. It would be long and dull and by now it should be clear why we need a different approach when dealing with unequal probabilities.

### 5.1.2 Continuous Time

As a contrast to discrete time, the general answer when using the continuous method is no more difficult than the specific case of equal probabilities. In fact, it has already been given by equation (3). Simply take the factors as they are and integrate the expression.

As an example, consider the same problem of four coupons with unequal probabilities stated in section 5.1. If the probabilities are $p_1 = \frac{1}{10}, p_2 = \frac{2}{10}, p_3 = \frac{3}{10}$ and $p_4 = \frac{4}{10}$, the integral becomes:

$$\begin{aligned}
E[T] &= \int_0^\infty \left(1 - (1 - e^{-\frac{t}{10}})(1 - e^{-\frac{2t}{10}})(1 - e^{-\frac{3t}{10}})(1 - e^{-\frac{4t}{10}})\right) \, dt \\
&= \int_0^\infty \left(e^{\frac{-t}{10}} + e^{\frac{-2t}{10}} - 2e^{\frac{-5t}{10}} + e^{\frac{-8t}{10}} + e^{\frac{-9t}{10}} - e^{-t}\right) dt \\
&= \left[e^{-t} \; - \frac{10}{9}e^{\frac{-9t}{10}} - \frac{10}{8}e^{\frac{-8t}{10}} + 4e^{\frac{-5t}{10}} - \frac{10}{2}e^{\frac{-2t}{10}} - 10e^{\frac{-9t}{10}}\right]_0^\infty \\
&= 10 + 5 - 4 + \frac{5}{4} + \frac{10}{9} - 1 \\
&= \frac{445}{36} \approx 12.
\end{aligned}$$

The problem which we all but abandoned in the previous section for being far too cumbersome, has now become manageable.

## 5.2 K Coupons of Each Type

Another generalization is to collect coupons until you have not just one, but two or more of each type. Obviously this is going to, on average, take longer than collecting only one of each. How much longer is not obvious however. Luckily, this question was answered in 1960 by Donald J. Newman [New60]. He uses the approximation that the expected time for $K = 1$ goes like $n \log(n)$, especially for large $n$. The conclusion he comes to is that the first set takes $n \log(n)$ attempts on average, and all subsequent sets takes an additional $n \log(\log(n))$ tries. The full expression is:

$$E_K[T] = n \log(n) + n(K-1)\log(\log(n)) + o(t). \tag{7}$$

Just as we thought obvious it takes longer to collect $k$ of each coupon, but not much longer, since the logarithm of the logarithm grows incredibly slowly.

As an example, consider again the 20-sided dice. It took on average 72 rolls until each side had been rolled once. Suppose we wanted to roll each side a total of 20 times ($K = 20$). Using Equation (7) we get:

$$\begin{aligned} E_K[T] &= n \log(n) + n(K-1)\log(\log(n)) + o(t) \\ &= 20\log(20) + 20(20-1)\log(\log(20)) + o(t) \\ &\approx 59.915 + 380 \cdot 1.0972 \\ &\approx 477. \end{aligned}$$

One thing of note here are that 60 is a bit off from 72, suggesting that 20 is not a large number of coupons. The main point, however, that further coupons do not take that much extra time to acquire, is highlighted.

# 6

# Applications

## 6.1 The Pokemon Games

The rather well known series of video games called Pokemon is a perfect example of the generalized Coupon Collector's Problem. The series has been going on for almost 30 years and has evolved substantially since its inception. However, the core element of the games has remained the same. To discover and collect all the different kinds of creatures available. The games slogan, "Gotta catch 'em all", is all but equivalent to the "Collect all and win" description of the Coupon Collectors Problem given in section 2.

The game is divided into several areas. Each area has certain creatures available to collect, with each type of creature having its own probability of appearing any time there is an encounter. Naturally it might be in the interest of players to know the expected number of encounters before they can hope to be done.

As an example. Near the beginning Pokemon Red (one of the first games in the series) you can encounter a type A creature with probability 0.45, a type B creature with probability 0.40, and a type C creature with a probability of 0.15. Using Equation (3) we can express the cumulative distribution function as:

$$\prod_{j=1}^{n}(1-e^{-p_j t}) = (1-e^{-0.45t})(1-e^{-0.40t})(1-e^{-0.15t}).$$

Which, upon integration, yields the expected time until you have encountered all three kinds.

$$E[X] = \int_0^\infty [1-(1-e^{-0.45t})(1-e^{-0.40t})(1-e^{-0.15t})]\,dt = 8.$$

## 6.2 Loaded Dice

A loaded, or weighted, dice is one which has been manipulated to more often land with a particular side facing up than would be expected of a regular fair dice. For example, a standard six-sided dice will roll any particular number between one and six with an equal probability of $\frac{1}{6}$ for all of them. Suppose a six-sided dice was made heavier on the side with one dot, making it more likely to show the side with six. The sides showing two through five will still be as probable as one another, while the one will be the least likely. What is the expected number of rolls needed to "collect" all numbers if $p(1) = \frac{1}{12}$, the probabilities of two through five are all $\frac{1}{8}$, and $p(6) = \frac{5}{12}$? Again, using Equation (3) and then integrating gives the expected number of rolls:

$$E[X] = \int_0^\infty [1 - (1 - e^{-\frac{t}{12}})(1 - e^{-\frac{t}{8}})^4(1 - e^{-\frac{5t}{12}})]\, dt = 20.$$

For a fair six-sided dice:

$$E[X] = \int_0^\infty [1 - (1 - e^{-\frac{t}{6}})^6]\, dt = 15.$$

Evidently the unfair dice takes longer to complete a full set. Whether or not this is generally true will be discussed later in Section 7.1.2.

# 7

# Conclusion and Discussion

The Coupon Collector's Problem is a classical, and versatile problem with its roots in the eighteenth century. It can be appreciated by anyone from high-school students to professors of mathematics. It has many real world applications, only a few of which were touched upon in this thesis.

We have seen that computing the expected time until completion can be done either using discrete geometric random variables, or by the Poisson Process, which uses the exponentially distributed mean of the Poisson distribution. The first method is simpler and avoids some mathematical notation that might scare away the more casual reader. However, the Poisson Process offers a deeper understanding of the problem and is necessary to generalize the problem effectively. In fact, the generalized problem is no more difficult than the specific case when using this method.

## 7.1 Further Problems

The main problem in this thesis was to compute the expected value of the total time taken for various examples of the Coupon Collector's Problem. Naturally, there are more questions to be asked about the base problem, as well as further questions about the examples that were explored.

### 7.1.1 Further Exploration of the Pokemon Games

The Pokemon Games go much further than the simple model presented in Section 6.1. Some complicating properties and mechanics of the games are:

1. The same type of creature can appear in multiple areas.

2. Some creatures may turn into other types of creatures, but not back. $A \to B$, $B \not\to A$.

3. Any time you encounter a creature, there is not a 100 percent chance you can collect it. The amount of tries it takes is also random with varying rates of success between creature types. Additionally, if it takes too many tries, you may run out of resources and be forced to abandon it and return later.

If someone intends to complete the goal of acquiring all creatures with as little work as possible, this certainly complicates things. Point 1 tells us that we should look in places with as little overlap with previous areas as possible. Point 2 tells us that we should search for the type A creature and not type B. Point 3 is the most complex. As the game progresses there are tools that the player may acquire that improve that rate of success. Therefore it may be beneficial to leave certain areas and creatures for later. When you have the proper tools and will not run the risk of running out.

In order to properly analyze this problem, a deep dive into the games is necessary. Much too deep to fit in any further capacity here.

### 7.1.2 Loaded Dice

The unsatisfying answer to the question of whether or not all unfair dice take longer before rolling one of each number than their fair counterparts, is that I do not know. To prove it you would probably have to analyze Equation (3) with the probabilities $p_j$ of rolling a particular number as variables, somehow. Consider this an open problem to be explored in the future.

## 7.2 Improvements and Alterations

No project or person is perfect, and this thesis is no exception. There are things I would have done differently with hindsight in mind.

The main thing is that I should have studied the theory much, much more than I did before I started writing the first draft. I thought I had a good, solid understanding of the problem. However I was woefully unaware of both the depth with which one can analyze the Coupon Collector's Problem, as well as the rigour of the mathematics behind it. This lead to most of my actual understanding coming first as the project was nearing its end.

# References

[Fel71] William Feller. Introduction to Probability Theory and its Applications, 1971.

[FeSa14] Marco Ferrante and Monica Saltalamacchia. The Coupon Collector's Problem, 2014.

[New60] Donald J. Newman. The Double Dixie Cup Problem. The American Mathematical Monthly, Vol. 67, No. 1, 1960.

[Ros14] Sheldon Ross. Introduction to Probability Models, 11th edition, 2014.

[Sti00] David Stirzaker. Advice to hedgehogs, or, constants can vary. The Mathematical Gazette, 2000.

[Wol10] Robert L. Wolpert. Conditional Expectation. Institute of Statistics and Decision Sciences, 2010.