Parabola Volume 58, Issue 3 (2022)

Application of the Multiple Subset Coupon Collector Problem to COVID-19 testing

Pannawich Tangkitsiriroj[1], **Pim Chotnapalai**[2], **Supawich Trongdee**[3] and **Thanaporn Thanodomdech**[4]

1 Introduction

The coupon collector problem is a well-known problem in probability theory and addresses a situation where one collects items (coupons) in a finite collection of size n. These items arrive randomly and sequentially. The random appearance of items is shown in Figure 1. Note that we might collect the same coupon multiple times.

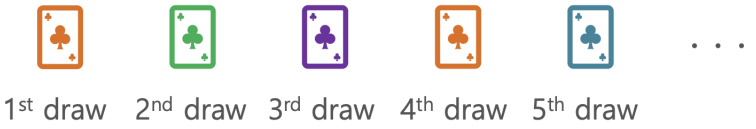

Figure 1: An illustration of the random appearance of items.

The essential objective question is "What is the expected number of coupons one needs to draw to complete the collection?" In the classical (and simplest) case, all items in the collection have an equal probability, and one of them arrives at some time. For this case, the question can be solved by considering that the numbers of coupons are geometrically distribute; see [2]. Hence, the expected number N of coupons drawn is

$$E[N] = n \sum_{i=1}^{n} \frac{1}{i}.$$

Over time, the problem has became far more complicated, with more conditions imposed, resulting in various alternatives as seen in [2] and other works, including the *multiple subset coupon collector problem* [1]. In this variant, coupons are collected at a constant integer rate of 1, i.e., the number of coupons is a constant for each time they are drawn. In this case, Chang and Ross [1] tackled the question by assuming that the entire sequence of drawings is a Poisson process. With this, they found that the mean

[1]Student at Kamnoetvidya Science Academy
[2]Student at Kamnoetvidya Science Academy
[3]Student at Kamnoetvidya Science Academy
[4]Teacher at Kamnoetvidya Science Academy

number of coupons needed to draw to complete at least one subset when only one coupon is randomly obtained at a time, which is

$$E[N] = \int_0^\infty \prod_{i=1}^{m} \Big[1 - \prod_{j \in S_i} (1 - e^{-p_j t})\Big] dt \,. \tag{1}$$

A similar problem appears in trading card games where cards are sealed in packets containing fixed numbers of cards. Moreover, coupons (in this case, cards) may have varying probabilities. The equal ones are of the same "rarity", which can be a subset of the entire collection.

Inspired by this situation, this paper studies the multiple subset coupon collector problem in which coupons in each subset have an equivalent probability of being drawn. The purpose of this study is to find how this case can be modified to find the expected number of coupons required to complete at least one rarity in the collection. This starts from a simple case with a rate of 1. Simulations are provided as illustrated examples and for verification.

A real-world application is also given that refers to the recent COVID-19 pandemic. While testing is important to identify the infected people, governments need to optimize resources and might not be able to test everyone. Our answers to the coupon collector problem are used to give estimates for appropriate numbers of test kits.

2 The expected number of draws and its variance

Throughout this paper, we consider the multiple subset coupon collector problem. We can group into subsets items that share some of the same attributes, such as rarities. These are Poissonised with rate 1, meaning that one item is drawn per time unit. Also, we assume that the divided subsets are mutually disjoint. We apply Poisson and exponential distributions, the definitions of which can be found in [3].

The expected value and variance of N, the number of coupons needed to be drawn to complete at least one of the subsets, are derived from the result of [1] as follows.

Proposition 1. *The expected number of draws in the m-rarity case is*

$$E[N] = \int_0^\infty \prod_{i=1}^{m} \Big[1 - (1 - e^{-p_i t})^{R_i}\Big] dt \,,$$

where p_i is the probability of getting card of rarity i, and R_i is the number of cards of rarity i, for $i \in \{1, 2, 3, \ldots, m\}$.

Proof. Let S_i be the rarity subsets having cardinality R_i. Each rarity has a uniform probability, i.e., each coupon in subset S_i has probability p_i, so the proposition follows from Equation (1). □

Proposition 2. *The variance of the number of coupons in the m-rarity case is*

$$Var(N) = 2 \int_0^\infty t \prod_{i=1}^{m} \Big[1 - (1 - e^{-p_i t})^{R_i}\Big] dt - E[N](1 + E[N]) \,.$$

Proof. In [1] it was calculated that

$$Var(N) = Var(T) - E[N] = 2\int_0^\infty t \prod_{i=1}^{m} \Big[1 - \prod_{j \in S_i} (1 - e^{-p_j t})\Big] dt - E[N](1 + E[N])$$

where T is the time that at least one of the subsets is present in the collection. With the same argument as in the proof of Proposition 1, we obtain

$$Var(N) = 2\int_0^\infty t \prod_{i=1}^{m} \Big[1 - (1 - e^{-p_i t})^{R_i}\Big] dt - E[N](1 + E[N])\,.$$

□

By Proposition 1, we can describe the common three-rarity case as follows.

Corollary 3. *The expected number of draws in three-rarity case is*

$$E[N] = \int_0^\infty \Big[1 - (1 - e^{-p_r t})^R\Big]\Big[1 - (1 - e^{-p_u t})^U\Big]\Big[1 - (1 - e^{-p_c t})^C\Big] dt\,,$$

where p_r, p_u and p_c are the probabilities of getting rare, uncommon and common card, respectively. Here, R, U and C are the numbers of rare, uncommon and common cards, respectively.

3 Verification by Simulation Using Python

To verify our theoretical model and its results above, we simulated an actual event using Python. The simulation starts by applying a random function to represent the drawing of cards. The remainder from dividing the random number by a constant is used to classify the card type according to probability. Once one subset is completed, the drawing time is returned. The program repeats this process millions of times and calculates the average number, which should align with the theoretical expected value, and variance.

For the example case, we suppose that the collection contains 2 rare cards with probability $p_r = 2.5\%$ each, 4 uncommon cards with $p_u = 5\%$ each and 5 common cards with $p_c = 15\%$ each. Figure 2 shows graphs representing the expectation and variance from 1 million draws to 100 million draws.

As illustrated by these graphs, the experimental values run toward the theoretical ones as the number of draws increases in both expected number and variance. This can imply that the theoretical results are reliable especially when the number of draws grow large. In the limit, the expected value for N is 13.6447. That means a collector needs to buy 14 cards, on average, to complete a collection.

4 Application to COVID-19 Testing

In 2019, a type of virus was introduced to humans and changed the world ever since. For three long years, the pandemic of COVID-19 has caused people to lose, grieve and

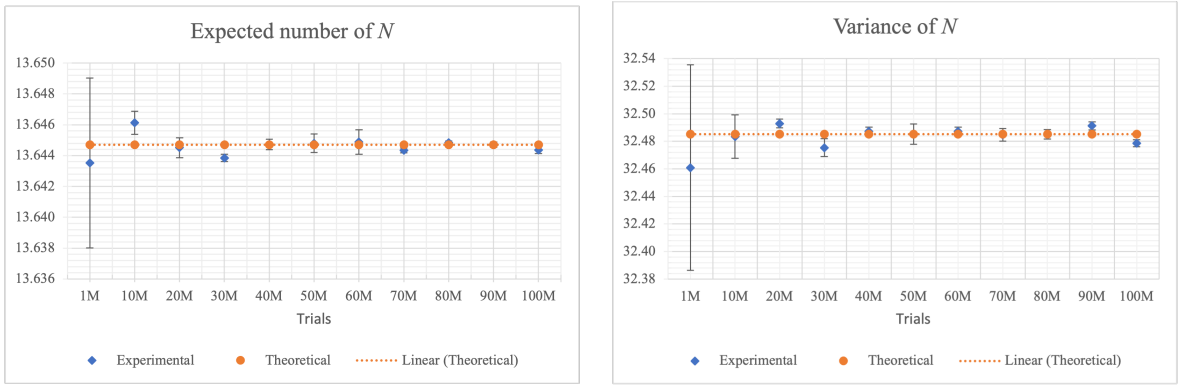

Figure 2: (left) the expectation $E[N]$ versus the number of draws and (right) the variance $Var(N)$ versus the number of draws for the example case.

suffer as it keeps inevitably going. Mass testing is one strong candidate for rescuing the world from this tragedy, except that testing costs a lot of money and resource. The mass testing solution also comes with great plastic waste that we are not yet ready to handle. It is therefore important to be able to find a good balance between the benefits and drawbacks. That is, it is important to calculate the least optimal number of tests needed to be conducted to search for all people with the disease.

We apply our results above to provide an estimate for such an optimal number. We categorize residents into two non-overlapping subsets: those who are infected by COVID-19 and those who are not. Populations are defined as R_i and R_u, respectively. If a particular country or city has a total population of n, then

$$n = R_i + R_u \,.$$

To find R_i, the number of people infected on a certain day is predicted by analyzing the trend of the last seven days. For example, 1,000 means that 1,000 people are predicted to be carrying the disease. Then, the percentage of daily COVID-19 tests that are positive, p, is used to determine the probability of detecting infected people. If a high proportion of the COVID-19 tests is positive, there is also a high probability to meet infected people the next day. At this point, the probability p_i of the infected people and the probability p_u of those uninfected can be calculated by

$$p_i = \frac{p}{R_i} \quad \text{and} \quad p_u = \frac{1-p}{R_u} \,.$$

By Proposition 1 with $m = 2$, we obtaine the expected number of residents that must be tested to discover those with virus as follows:

$$E[N] = \int_0^\infty \left[1 - (1 - e^{-p_i t})^{R_i}\right]\left[1 - (1 - e^{-p_u t})^{R_u}\right] dt \,. \qquad (\star)$$

By this formula, we obtain a very valuable figure, the number of tests that have to be conducted. In a world where we have limited resources and cannot inconsiderately create any other excessive waste, this result will contribute to society in a way that cuts down the number of infections along with developing a sustainable green world.

5 Simulation data for COVID-19 cases in three countries

This section aims to apply the formula derived in the preceding section to a real-life COVID-19 pandemic, which occurs in diverse ways around the world. In countries with severe outbreaks, there are many confirmed cases and positive results. On the other hand, some countries have only a few confirmed cases and the majority of the results are negative. Thailand, Germany and the Netherlands were chosen to represent mild, moderate and severe COVID-19 spread in this simulation. The data in Table 1 were collected on 5 March 2022, from ourworldindata.org.

Country	Confirmed cases (per 10,000)	Share of positive results (%)
Thailand	3	36.5
Germany	19	50.7
Netherlands	33	69.8

Table 1: The confirmed cases per 10,000 people and share of positive tests in Thailand, Germany and the Netherlands on 5 March 2022.

Values for all variables can be determined in Table 2, where each country's population number has been normalised to $n = 10,000$ people.

Country	R_i	R_u	p	p_i	p_u
Thailand	3	9997	36.5	12.17	0.006350905
Germany	19	9981	50.7	2.67	0.004939385
Netherlands	33	9967	69.8	2.12	0.003029999

Table 2: The values of R_i, R_u, p, p_i and p_u in Thailand, Germany and the Netherlands

Consequently, we substitute all values in $(\star)$ to theoretically compute $E[N]$, while Python simulation proceeds with trials varying from 100,000 to 1 million. Each of them is repeated five times, and the average represents the experimental value. The results from the simulation of each country are shown below. The graphs in Figure 3, Figure 4 and Figure 5 depict experimental value and theoretical value of 100,000 to 1 million trials in Thailand, Germany and the Netherlands, respectively.

Trials	1	2	3	4	5	Experimental	Theoretical	Error (%)
100000	15.05464	15.07655	15.02633	15.06144	15.10199	15.064190	15.0644	0.001394015
200000	15.05866	15.05029	15.06017	15.08240	15.08089	15.066481	15.0644	0.013814025
300000	15.05538	15.06867	15.09295	15.05541	15.07633	15.069749	15.0644	0.035505342
400000	15.07179	15.06687	15.04234	15.06399	15.08435	15.065669	15.0644	0.008420515
500000	15.05437	15.07642	15.07093	15.03421	15.07230	15.061648	15.0644	0.018265580
600000	15.06999	15.06183	15.05486	15.05288	15.07815	15.063542	15.0644	0.005693334
700000	15.07264	15.05659	15.06071	15.06005	15.06871	15.063742	15.0644	0.004366017
800000	15.06410	15.06930	15.07862	15.07338	15.08051	15.073182	15.0644	0.058293062
900000	15.04918	15.07684	15.06278	15.09061	15.06331	15.068543	15.0644	0.027499712
1000000	15.07090	15.05964	15.04575	15.05503	15.05231	15.056727	15.0644	0.050931999

Table 3: The experimental value, theoretical value and error in Thailand

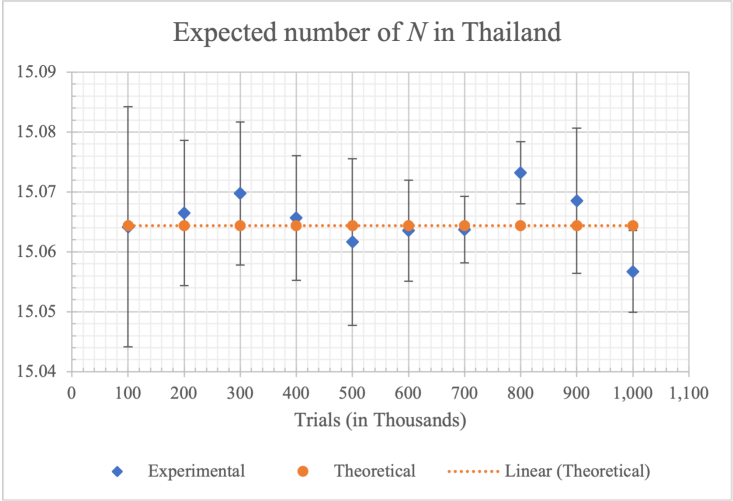

Figure 3: Experimental and theoretical values of 100,000 to 1 million trials, Thailand.

Trials	1	2	3	4	5	Experimental	Theoretical	Error (%)
100000	134.949	134.900	134.913	134.840	134.921	134.9046	134.855	0.03678025
200000	134.668	134.881	134.774	134.890	134.714	134.7854	134.855	0.05161099
300000	134.868	134.795	134.952	134.867	134.935	134.8834	134.855	0.02105966
400000	134.863	134.875	134.966	134.828	134.859	134.8782	134.855	0.01720366
500000	134.919	134.810	134.783	134.800	134.849	134.8322	134.855	0.01690705
600000	134.922	134.865	134.930	134.759	134.848	134.8648	134.855	0.00726707
700000	134.941	134.821	134.863	134.822	134.824	134.8542	134.855	0.00059323
800000	134.811	134.903	134.820	134.847	134.880	134.8522	134.855	0.00207630
900000	134.795	134.868	134.883	134.780	134.894	134.8440	134.855	0.00815691
1000000	134.873	134.824	134.819	134.812	134.943	134.8542	134.855	0.00059323

Table 4: The experimental value, theoretical value and error in Germany

As shown in the graphs in Figure 3, Figure 4 and Figure 5, experimental values converge to theoretical values as the trials increase, except in Thailand where there is more deviation at the end. This can indicate that the formula is acceptable to use with COVID-19 testing.

Furthermore, average errors or the average gap between experimental and theoretical data of all trials are compared in Thailand, Germany and the Netherlands in Table 6. As the table shows, average errors decrease in Thailand, Germany and the Netherlands, respectively. In other words, errors is inversely related to the outbreak severity. The lowest error found in the Netherlands has the most severe spread. In contrast, mild spread in Thailand makes the highest error. To conclude, our derived formula is more accurate in a severe situation, where many people get infected, and tests are positive.

The implementation of ($\star$), our derived formula, in the highly-spread situation would give an accurate number of tests to perform. Consequently, this can provide estimates for optimal numbers of antigen test kits, medical staff, testing time and location. These allow governments and local sectors to tackle COVID-19 more efficiently.

As the optimal number of tests to perform is calculated with the data on 5 March 2022, it is compared with the actual number of tests on the following day collected from

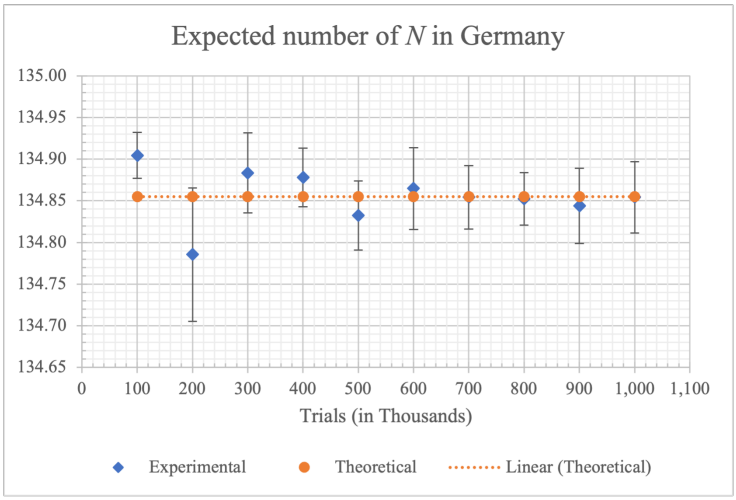

Figure 4: Experimental and theoretical values of 100,000 to 1 million trials, Germany.

Trials	1	2	3	4	5	Experimental	Theoretical	Error (%)
100000	193.25750	193.24863	193.17501	193.27331	193.36184	193.2270467	193.31	0.000429121
200000	193.38768	193.34355	193.30271	193.23541	193.41084	193.3446433	193.31	0.000179211
300000	193.29017	193.55209	193.23308	193.29612	193.23044	193.3584444	193.31	0.000250605
400000	193.44604	193.35842	193.23667	193.16352	193.29813	193.3470433	193.31	0.000191627
500000	193.38047	193.28161	193.37648	193.27272	193.15773	193.3461840	193.31	0.000187181
600000	193.30133	193.32104	193.37238	193.25787	193.47164	193.3315822	193.31	0.000111646
700000	193.31394	193.25460	193.11048	193.35778	193.43344	193.2263386	193.31	0.000432784
800000	193.34429	193.27340	193.28859	193.26602	193.33463	193.3020950	193.31	4.08929E-05
900000	193.41607	193.26684	193.25693	193.26201	193.22976	193.3132822	193.31	1.69791E-05
1000000	193.39342	193.25777	193.30466	193.31575	193.32798	193.3186157	193.31	4.45692E-05

Table 5: The experimental value, theoretical value and error in the Netherlands

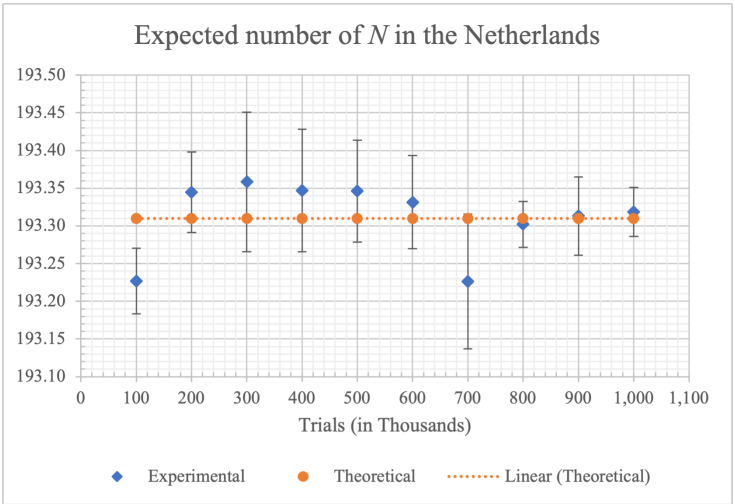

Figure 5: Experimental and theoretical values of 100,000 to 1 million trials, Netherlands

ourworldindata.org in Table 7. As shown in that table, the real-world data point out that these three countries conduct fewer COVID-19 tests than they should. Therefore,

Country	Confirmed cases (per 10,000)	Share of positive results (%)	Average error (%)
Thailand	3	36.5	0.02242
Germany	19	50.7	0.01622
Netherlands	33	69.8	0.00019

Table 6: The comparison of confirmed cases per 10,000 people, share of positive tests and errors among Thailand, Germany and the Netherlands.

Country	Optimal number of tests (per 10,000)	Actual number of tests (per 10,000)
Thailand	15	9
Germany	134	32
Netherlands	193	50

Table 7: The comparison of optimal and actual number of tests being performed among Thailand, Germany and the Netherlands.

local sectors might use the most recent number of confirmed cases and share of positive results to make an appropriate daily decision for regional COVID-19 testing. In addition, governments can use large-scale data to organize the distribution and long-term supply of test kits and medical personnel.

6 Discussion and Conclusion

The approach to considering the convergence of the experimental data in Figure 3, Figure 4 and Figure 5 is not stated in the previous section. The experimental data in Figure 4 and Figure 5 are convergent. However, the experimental data in Figure 3 appears to diverge at the end. This situation can happen because each experimental data is obtained by the Python simulation. Hence, it is a random event that sometimes generate outliers. That is the reason why we also include the mean deviation error bar in the graphs. By considering the range of error bar, the experimental data in Figure 3 is acceptable for convergence.

In this paper, we combined a multiple subset coupon collector problem with rarities in which we supposed that all coupons in a subset have the same occurrence probability. We proposed a formula for the expected number of draws and variance. Then, the formula is applied with COVID-19 testing to find the expected number of tests that have to be conducted. The numerical simulations are performed in three countries with varying levels of outbreak severity. The findings demonstrate that implementing a formula in a severe outbreak will result in a precise number of tests, which is a benefit for resource and time management.

The formula can be extended to other sampling events in which non-overlapping subsets of the population have their own rarities, implying that each individual in the subset has the same probability of being picked. For instance, it is possible to adapt with bird watching without capturing or quality control.

Acknowledgements

We wish to extend our thanks to Kamnoetvidya Science Academy, especially the Department of Mathematics and Computer Science for allowing us to conduct this research based on our interest and for giving us beneficial advice all along. We have gained a lot of knowledge and skills from this project and hope to integrate them into our future work. We would like to thank the referees for their comments and suggestions on the manuscript. In addition, we would like to thank Assoc. Prof. Ratinan Boonklurb for further reviewing and allowing us to share our findings via this journal.

References

[1] K.C. Chang and S.M. Ross, The multiple subset coupon collecting problem, *Probability in the Engineering and Informational Sciences* **21** (2007), 435–440.

[2] M. Ferrante and M. Saltalamacchia, The coupon collector's problem, *Materials Matemàtics* **2014** (2014), 35 pp.

[3] S.M. Ross, *A First Course In Probability*, Pearson Education Limited, 2020.