Journal of Applied Mathematics & Bioinformatics, vol.5, no.3, 2015, 1-14
ISSN: 1792-6602 (print), 1792-6939 (online)
Scienpress Ltd, 2015

How many trials does it take to collect all different *types* of a population with probability p ?

Aristides V. Doumas[1]

Abstract

Coupons are collected one at a time (independently and with replacement) from a population containing N distinct *types*. This process is repeated until all N different *types* (coupons) have been collected (at least once).
Recently, interesting results have been published regarding the asymptotics of the moments and the variance, of the number T_N of coupons that a collector has to buy in order to find all N existing different coupons as $N \to \infty$. Moreover, the limit distribution of the random variable T_N (appropriately normalized), has been obtained for a large class of coupon probabilities (see, [9], [10], and [11]).
This classical problem of probability theory has found a plethora of applications in many areas of science, and quite recently, it has been highly involved with cryptography. In this note we take advantage of

[1] Department of Mathematics, National Technical University of Athens,
Zografou Campus, 157 80 Athens, Greece.
5th Lyceum, Vironas, Athens, Greece.
E-mail: aris.doumas@hotmail.com

Article Info: *Received* : November 1, 2014. *Revised* : February 2, 2015.
Published online : September 15, 2015.

the above results and present in detailed various examples that illustrate problems similar to those one faces in the real life. We also conjecture on the minimum of the variance $V[T_N]$.

Mathematics Subject Classification: 60F05; 60F99
Keywords: Limit distributions; Urn problems; Cartophily problem; Coupon collecting

1 Introduction

Consider a population whose members are of N different *types* (e.g. baseball cards). For $1 \leq j \leq N$ we denote by p_j the probability that a member of the population is of type j, where $p_j > 0$ and $\sum_{j=1}^{N} p_j = 1$. The members of the population are sampled independently *with replacement* and their types are recorded. The so-called *"coupon collector problem"* deals with questions arising in the above procedure. T_N is, of course, a random variable. Some key quantities are the (rising) moments of the number T_N of trials it takes until all N types are detected (at least once), the variance, and (of course), the distribution of T_N.

The CCP pertains to the family of Urn problems among with other famous problems such as the *birthday problem* (initiated by Richard von Mises in 1932), and the *matching problem*, see, e.g., [8]. CCP has a long history starting from P.S. Laplace and A.De Moivre, (for details see, e.g., [9], [10], [11] and the references therein.) The problem became popular in the 1930's when the Dixie Cup company introduced a highly successful program by which children collected Dixie lids to receive "Premiums," beginning with illustrations of their favored Dixie Circus characters, and then Hollywood stars and major league baseball players (for the Dixie Cup company history see [31]). The information above explains, possibly, the fact that the entertaining term *cartophily* appeared in the title of two (relatively unknown) papers published in the first half of the past century (both in *Math. Gazette*, see [21], [18]).

The problem was highlighted in W. Feller's work [15] as a classic and important topic (see Section 2 below). Since then, CCP has attracted the attention of various researchers due to the fact that it has found many applications in many areas of science (computer science/search algorithms, mathematical programming, optimization, learning processes, engineering, ecology, as well as linguistics—see, e.g., [4]). Some representative applications (quite a few articles are now in the math arXiv) concern biology, (see, [25]), linguistics, (see, e.g., [19], [13], [26]), computer science (see, e.g., [24], [1], [17], [3], and, [6].) In particular, the work of *J. Bonneau* and *E. Shutova* [6], triggered the article **Computer passwords Speak, friend, and enter**, published in the *Economist, 24 March 2012.* CCP has attracted the interest of researchers working in *Cryptography* and is our belief that this interest will be increased in the near future, (see, e.g., [16] and [28]).

The outline of this note follows: In Subsection 2.1 we remind the reader well known results (from the sixties) for the simplest version of the problem, i.e., the case where the coupon probabilities are equal. In Subsection 2.2 we present recently published results for the general case of unequal probabilities. Finally, in Section 3 we present some examples by exploiting the above mentioned results.

2 Well known results

2.1 The case of equal probabilities

Naturally, the simplest case occurs when one takes

$$p_1 = \cdots = p_N = \frac{1}{N}. \tag{1}$$

There are three classical references for this version of the problem:

(*i*) **W. Feller's** well known work, *An Introduction to Probability Theory and Its Applications, Vol. I,* (1950), [15].[2]

(*ii*)**D. J. Newman's, L. Shepp's** paper, *The Double Dixie Cup problem*, (1960), [23], (where they answered the more general question: how long, on

[2]Readers may observe that general results for the classic CCP of [21] and [18], had appeared before Feller's book.

average does it take to obtain m complete sets of N coupons),

(*iii*) the famous paper of **P. Erdős, A. Rényi**, *On a classical problem of Probability theory*, (1961), where the limit distribution of the random variable T_N (appropriately normalized), turned out to be the standard Gumbel distribution (see [14]). The results for this case are:

$$E\left[\,T_N\,\right] = N H_N, \tag{2}$$

where (see, [15])

$$H_N = \sum_{m=1}^{N} \frac{1}{m} \tag{3}$$

(H_N is, sometimes, called the N-th harmonic number). Moreover,

$$E\left[\,T_N\,(T_N+1)\,\right] = N^2 \left(H_N^2 + \sum_{m=1}^{N} \frac{1}{m^2} \right). \tag{4}$$

By the celebrated Euler-Maclaurin Summation formula (see, [2]) one has

$$E\left[\,T_N\,\right] \sim N \ln N + \gamma N + \frac{1}{2} - \sum_{k=2}^{\infty} \frac{B_k}{k N^{k-1}}, \tag{5}$$

$$E\left[\,T_N\,(T_N+1)\,\right] = N^2 \left[(\ln N)^2 + 2\gamma \ln N + \gamma^2 + \frac{\pi^2}{6} + O\left(\frac{\ln N}{N}\right) \right], \tag{6}$$

where B_K are the well known Bernoulli numbers. Using the identity

$$V\left[\,T_N\,\right] = E\left[\,T_N(T_N+1)\,\right] - E\left[\,T_N\,\right] - E\left[\,T_N\,\right]^2, \tag{7}$$

one arrives at the expression

$$V\left[\,T_N\,\right] = \frac{\pi^2}{6} N^2 - N \ln N - (\gamma+1) N + O\left(\frac{\ln N}{N}\right). \tag{8}$$

The heaviest result is due to Erdős and Rényi, [14]. In the case of equal probabilities the authors stated and proved the following remarkable limit theorem (see also [12]):

$$P\left\{ \frac{T_N - N \ln N}{N} \le x \right\} \to \exp\left(-e^{-x}\right), \qquad N \to \infty. \tag{9}$$

♣ A nice computer simulation of the CCP in the case of equal probabilities is available from [32].

2.2 The general case of unequal coupon probabilities

For the general (and much more interesting) case of unequal sampling probabilities the first important result was given by Herman Von Schelling, in a German journal of a limited circulation as early as 1934 (even before W. Feller!). This paper, [29], remained practically unknown; besides the formulas were given without proofs. Later, Von Schelling, (1954), published a complete version of the first paper, in the *American Mathematical Monthly*, and obtained general formulae for the first and second moment of T_N, see [30]. In particular, he proved that:

$$E\left[\,T_N\,\right] = \sum_{\substack{J\subset\{1,\ldots,N\}\\ J\neq\emptyset}} \frac{(-1)^{|J|-1}}{\sum_{j\in J} p_j} = \sum_{m=1}^{N}(-1)^{m-1}\sum_{1\leq j_1<\cdots<j_m\leq N}\frac{1}{p_{j_1}+\cdots+p_{j_m}} \tag{10}$$

and

$$E\left[\,T_N\left(T_N+1\right)\right] = 2\sum_{\substack{J\subset\{1,\ldots,N\}\\ J\neq\emptyset}} \frac{(-1)^{|J|-1}}{\left(\sum_{j\in J} p_j\right)^2} \tag{11}$$

$$= 2\sum_{m=1}^{N}(-1)^{m-1}\sum_{1\leq j_1<\cdots<j_m\leq N}\frac{1}{\left(p_{j_1}+\cdots+p_{j_m}\right)^2}. \tag{12}$$

However, these expressions are not easy to handle, when N is large. This is why asymptotic estimates of $E[\,T_N\,]$ and $E\left[\,T_N\left(T_N+1\right)\right]$ are necessary. Of course, (7) holds. Before we continue we would like to make a conjecture.

Conjecture. Under the constraints: $\sum_{j=1}^{N} p_j = 1$, $p_j > 0$, the variance $V\left[\,T_N\,\right]$ is a Schur convex function.

Remark. If the statement above is true, then by a well known property of Schur convex functions (see [20]), the variance $V\left[\,T_N\,\right] := V\left(p_1, p_2, \cdots, p_N\right)$ attains its minimum value when all its variables are equal. Then, the conjecture stated in [9] is also true.

The following set up was first introduced in [5] and then it has been adapted in [9], [10] and [11]. Let $\alpha = \{a_j\}_{j=1}^{\infty}$ be a sequence of strictly positive numbers. Then, for each integer $N > 0$, one can create a probability measure $\pi_N =$

$\{p_1, ..., p_N\}$ on the set of types $\{1, ..., N\}$ by taking

$$p_j = \frac{a_j}{A_N}, \qquad \text{where} \quad A_N = \sum_{j=1}^{N} a_j. \tag{13}$$

Notice that p_j depends on α and N, thus, given α, it makes sense to consider the asymptotic behavior of $E\left[T_N\right]$, $E\left[T_N\left(T_N+1\right)\right]$, and $V\left[T_N\right]$ as $N \to \infty$. Thus,

$$E\left[T_N\right] = E_N(A_N^{-1}\alpha) = A_N E_N(\alpha), \tag{14}$$

where

$$E_N(\alpha) := \sum_{\substack{J \subset \{1,...,N\} \\ J \neq \emptyset}} \frac{(-1)^{|J|-1}}{\sum_{j\in J} a_j} = \sum_{k=1}^{N} (-1)^{k-1} \sum_{1 \leq j_1 < \cdots < j_k \leq N} \frac{1}{a_{j_1} + \cdots + a_{j_k}} \tag{15}$$

$$= \int_0^\infty \left[1 - \prod_{j=1}^{N}\left(1 - e^{-a_j t}\right)\right] dt = \int_0^1 \left[1 - \prod_{j=1}^{N}\left(1 - x^{a_j}\right)\right] \frac{dx}{x}. \tag{16}$$

Similaly,

$$E\left[T_N(T_N+1)\right] = Q_N(A_N^{-1}\alpha) = A_N^2 Q_N(\alpha), \tag{17}$$

where

$$Q_N(\alpha) := 2 \sum_{\substack{J \subset \{1,...,N\} \\ J \neq \emptyset}} \frac{(-1)^{|J|-1}}{\left(\sum_{j\in J} a_j\right)^2} = 2\sum_{m=1}^{N} (-1)^{m-1} \sum_{1 \leq j_1 < \cdots < j_m \leq N} \frac{1}{\left(a_{j_1} + \cdots + a_{j_m}\right)^2} \tag{18}$$

$$= 2\int_0^\infty \left[1 - \prod_{j=1}^{N}\left(1 - e^{-a_j t}\right)\right] t\, dt = -2\int_0^1 \left[1 - \prod_{j=1}^{N}\left(1 - x^{a_j}\right)\right] \frac{\ln x}{x}\, dx. \tag{19}$$

Set

$$L_1(\alpha) := \lim_N E_N(\alpha) = \int_0^1 \left[1 - \prod_{j=1}^{\infty}(1 - x^{a_j})\right] \frac{dx}{x} \tag{20}$$

and

$$L_2(\alpha) := \lim_N Q_N(\alpha) = -2\int_0^1 \left[1 - \prod_{j=1}^{\infty}\left(1 - x^{a_j}\right)\right] \frac{\ln x}{x} dx. \tag{21}$$

In [9] the authors obtained a dichotomy, namely that both $L_1(\alpha)$, $L_2(\alpha)$ are positive numbers **(CASE I)**, or they are both infinite **(CASE II)**. In particular they are both finite, if and only if there exist a $\xi \in (0,1)$ such that

$$\sum_{j=1}^{\infty} \xi^{a_j} < \infty. \tag{22}$$

The results we have for the **(CASE I)** are the following (see, again [9]): If $L_i(\alpha) < \infty$, $i \in \{1,2\}$, then, as $N \to \infty$,

$$E\left[T_N\right] = A_N L_1(\alpha)\left[1+o(1)\right] \tag{23}$$

$$E\left[T_N(T_N+1)\right] = A_N^2 L_2(\alpha)\left[1+o(1)\right] \tag{24}$$

$$V\left[T_N\right] = A_N^2\left[L_2(\alpha) - L_1(\alpha)^2\right] + o\left(A_N^2\right). \tag{25}$$

The heaviest result is the following limit theorem:

$$P\left\{\frac{T_N}{A_N} \le y\right\} \to \prod_{j=1}^{\infty}\left(1 - e^{-a_j y}\right), \qquad N \to \infty. \tag{26}$$

Notice that here the limiting distribution depends on the choice of the sequence $\{a_j\}$.
Regarding **(CASE II)** the authors restricted the class of sequences $\alpha = \{a_j\}$:

$$a_j = \frac{1}{f(j)}, \qquad \text{where } f \in C^3, \qquad f(x) > 0, \qquad f'(x) > 0,$$

and, as $x \to \infty$,

$$f(x) \to \infty, \quad \frac{f'(x)}{f(x)} \to 0, \quad \frac{f''(x)/f'(x)}{[f'(x)/f(x)} = O(1), \quad \frac{f'''(x)\,f(x)^2}{f'(x)^3} = O(1). \tag{27}$$

These conditions are satisfied by several common functions. In particular, $f(\cdot)$ belongs to the class of positive and strictly increasing $C^3(0,\infty)$ functions, which grow to ∞ (as $x \to \infty$) slower than exponentials, but faster than powers of logarithms. The results for the **(CASE II)** are the following:

$$E[T_N] = A_N f(N)\left[\frac{1}{\delta} + \ln\delta + \gamma - \delta\ln\delta + (\omega(N) - \gamma)\,\delta + O\left(\delta^2\ln^2\delta\right)\right], \tag{28}$$

$$E[T_N(T_N+1)] = A_N^2 f(N)^2\left\{\frac{1}{\delta^2} + \frac{2\ln\delta}{\delta} + \frac{2\gamma}{\delta} + \ln^2\delta + 2\,(\gamma - 1)\ln\delta\right.$$

$$+ \left(2\omega(N) + \gamma^2 + \frac{\pi^2}{6} - 2\gamma\right) + O\left(\delta \ln^2 \delta\right) \Bigg\} . \tag{29}$$

$$V\left[T_N\right] \sim \frac{\pi^2}{6} \, A_N^2 \, f(N)^2 = \frac{\pi^2}{6} \cdot \frac{1}{p_N^2} = \frac{\pi^2}{6} \cdot \frac{1}{\min_{1 \le j \le N}\{p_j\}^2}, \tag{30}$$

where

$$\delta(N) = \left[\ln\left(\frac{f(N)}{f'(N)}\right)\right]^{-1}$$

and

$$\omega(N) := -2 + \frac{f''(N)/f'(N)}{f'(N)/f(N)}.$$

(The leading behavior of the rising moments of the r.v. T_N has been obtained in [11]). What is even more important is a limit theorem which gives to the results above a tangible sense. It has been proved that as $N \to \infty$

$$P\left\{\frac{T_N - b_N}{k_N} \le y\right\} \longrightarrow \exp(e^{-y}), \qquad \text{for all } \ y \in \mathbb{R}, \tag{31}$$

where

$$b_N = A_N f(N) \left[\ln\left(f(N)/f'(N)\right) - \ln\ln\left(f(N)/f'(N)\right)\right] \text{ and } k_N = A_N f(N). \tag{32}$$

Notice that the limiting distribution is Gumbel, independently of the choice of $f(x)$(!).

Observation. It is notable that the authors in [9] took advantage of well known but general limit theorems of [22], in order to prove (26) and (31).

3 Examples

In this section we present some examples which illustrate the limit theorems (9), (26), and (31). We believe, that the connection, as well as, possible applications with areas such as cryptology, arise naturally. We begin with a warm up example from W. Feller, [15].

Example 1. What is the probability that in a village of 2190 ($= 6 \cdot 365$) people all birthdays are presented? Is the answer much different for 1825 ($= 5 \cdot 365$) people?

Here $N = 365$, and $N \ln N = 2153$. Hence, (9) yields

$$\begin{aligned} P\left(T_{365} \leq 2190\right) &= P\left(\left(T_{365} - 2153\right)/365 \leq 37/365\right) \\ &\approx \exp(-e^{-0.1014}) = \exp(-0.9036) = 0.4051. \end{aligned}$$

On the other hand

$$\begin{aligned} P\left(T_{365} \leq 1825\right) &= P\left(\left(T_{365} - 2153\right)/365 \leq -328/365\right) \\ &\approx \exp(-e^{0.8986}) = \exp(-2.4562) = 0.085. \end{aligned}$$

The following is an entertaining example; in fact it is an inverse type of Example 1.

Example 2. A Ph.D. student is the first who visits every day the library of his university. The library has 200 numbered lockers from 1 to 200. Each day the librarian gives the student, with probability 1/200, a (numbered) key, so that the student's personal things may be safe, while he studies. What should be the duration of this Ph.D. program, so that with probability 0.90, our student will borrow all 200 keys at least once (consider that the library is 365 days per year open and that the student is willing to study every day).

In this case $N = 200$ and $N \ln N = 1059.66$. Assume that the answer is λ years. We have (from (9))

$$\begin{aligned} P\left(T_{200} \leq \lambda \cdot 365\right) &= P\left(\left(T_{200} - 1059.66\right)/200 \leq \left(\lambda \cdot 365 - 1059.66\right)/200\right) \\ &\approx \exp(-e^{-\alpha}) = 0.90, \end{aligned}$$

where $\alpha = \left(\lambda \cdot 365 - 1059.66\right)/200$.
Thus, $\alpha = -\ln\left[-\ln(0.90)\right] = 2.25037$, hence with probability 0.90, this Ph.D. program should last no less than 4.14 years.

We continue with a general example:

Example 3. It is well known that 100 different types of fish live in the Lake Michigan (U.S.A.), (see, e.g., [33]). Assume that a fisherman always gets a fish[3] and that after every success he releases the fish into the lake. His goal is to get a complete set of all 100 types of fish (at least once). What should be

[3] He is that capable !

the minimum number of trials, so that with probability 0.90 he gets a complete set, in each one of the three following cases:

(i) $a_j = 1/j,\ j = 1, 2, \cdots, 100$. This case is the so–called *standard Zipf distribution.*

(ii) $a_j = j,\ j = 1, 2, \cdots, 100$. This is the well known *Linear case.*

(iii) $a_j = 1$, i.e., $p_j = 1/100$. This case has already been studied (see Subsection 2.1 and Examples 1–2 of Section 3).

► (i) We have, $\sum_{j=1}^{\infty} \xi^{1/j} = \infty$, for all $\xi \in (0, 1)$. In addition, the function $f(x) = x$ satisfies conditions (27). Hence, the sequence a_j falls into **Case II**. We have:
$N = 100,\ f(N) = N,\ \ln\left(f(N)/f'(N)\right) = \ln N$, and $A_N = H_{100} = 5.18738$. Hence, (32) yields $b_{100} = 1596.67,\ k_{100} = 518.738$. Assume that the answer is a trials. We have

$$\begin{aligned} P\left(T_{100} \le a\right) &= P\left(\left(T_{100} - 1596.67\right)/518.738 \le \left(a - 1596.67\right)/518.738\right) \\ &\approx \exp(-e^{-\lambda}) = 0.90, \end{aligned}$$

where $\lambda = \left(a - 1596.67\right)/518.738$.
So that $\lambda = -\ln\left[-\ln\left(0.90\right)\right] = 2.25037$. Thus, with probability 0.90 one needs at least **2,765** trials to collect all 100 different types of fish. (The exact calculation yields $a = 2,764.07$ trials).

► (ii). We have, $\sum_{j=1}^{\infty} (\frac{1}{2})^j < \infty$. Hence, this case falls into **CASE I**. We have $N = 100$ and $A_N = \sum_{j=1}^{100} j = 5,050$. As we will see **more than 12,000 trials are needed**. Indeed,

$$\begin{aligned} P\left(T_{100} \le 12,000\right) &= P\left(T_{100}/5,050\right) \le \left(12,000/5,050\right) \\ &\approx \prod_{j=1}^{\infty} \left(1 - e^{-2.37624j}\right) = 0.898477 \approx 0.90. \end{aligned}$$

► (iii). This is an example similar to Example 2. We mention, again, this case in order to compare with the results of (i), (ii). We have $N = 100$ and $N \ln N = 460.517$. Assume that the answer is b trials. From the corresponding limiting theorem (see, (9)) we get

$$\begin{aligned} P\left(T_{100} \leq b\right) &= P\left(\left(T_{100}-460.517\right)/100 \leq \left(b-460.517\right)/100\right) \\ &\approx \exp(-e^{-u}) = 0.90, \end{aligned}$$

where $u = (b - 460.517)/100$.
Hence, $u = -\ln\left[-\ln(0.90)\right] = 2.25037$, so that with probability 0.90, at least **686** trials are needed. (The exact calculation yields $b = 685.554$ trials).

Regarging the standard deviation we have the following approximations for each case respectively:

$$\sqrt{V[\,T_{N_{Zipf}}\,]} \approx \sqrt{\frac{\pi^2}{6}N^2\left(\ln N\right)^2} \overset{(30)}{=} \sqrt{348,851} \approx 590.64$$

$$\sqrt{V[\,T_{N_{Linear}}\,]} \approx \sqrt{\left(L_2(\alpha) - L_1(\alpha)^2\right)A_N^2} \overset{(25)}{=} \sqrt{20,945,800} \approx 4,576.66$$

$$\sqrt{V[\,T_{N_{Equal}}\,]} \approx \sqrt{\frac{\pi^2}{6}N^2} \overset{(8)}{=} \sqrt{1,6449.3} \approx 128.26.$$

Acknowledgments. As already mention, all the theoretical results of Subsection 2.2 come from my joint work with Professor Vassilis G. Papanicolaou, [9], [10]. I am deeply indebted to Vassilis for our collaboration and endless talks.

References

[1] S. Abraham, G. Brockman, S. Sapp and A.P. Godbole, Omnibus sequences, Coupon collection, and missing word counts, arXiv:0905.4517v1 [math.PR], (2009).

[2] C.M. Bender and S.A. Orszag, *Advanced Mathematical Methods for Scientists and Engineers I: Asymptotic Methods and Perturbation Theory*, Springer-Verlag, New York, 1999.

[3] J. Du Boisberranger, D. Gardy and Y. Ponty, The weighted words collector, *DMTCS Proc. AofA 23rd Intern. Meeting on Probabilistic, Combinatorial, and Asymptotic Methods for the Analysis of Algorithms*, Montreal, Canada, (2012).

[4] A. Boneh and M. Hofri, The Coupon Collector Problem Revisited–a Survey of Engineering Problems and Computational Methods, *Comm. Statist. Stochastic Models*, **13**(1), (1997), 39–66.

[5] S. Boneh and V.G. Papanicolaou, General Asymptotic Estimates for the Coupon Collector Problem, *Journal of Computational and Applied Mathematics*, **67**(2), (Mar. 1996), 277–289.

[6] J. Bonneau and E. Shutova, Linguistic properties of multi-word passphrases, *USEC'12: Workshop on Usable Security*, Kralendijk, Bonaire, The Netherlands, Featured in The Economist, 24 March 2012 under the title: Computer passwords Speak, friend, and enter.

[7] R.K. Brayton, On the asymptotic behavior of the number of trials necessary to complete a set with random selection, *Journal of Mathematical Analysis and Applications*, **7**, (1963), 31–61.

[8] P. Diaconis and S. Holmes, A Bayesian peek into Feller volume I, *Sankhyā*, Special issue in memory of D. Basu, **64** Ser. A (3, part 2), (2002), 820–841.

[9] A.V. Doumas and V.G. Papanicolaou, The Coupon Collector's Problem Revisited: Asymptotics of the Variance, *Adv. Appl. Prob.*, **44**(1), (2012), 166–195.

[10] A.V. Doumas and V. G. Papanicolaou, Asymptotics of the Rising Moments for the Coupon Collector's Problem, *Electron. J. Prob.*, **18**(41), (2012), 1–15.

[11] A.V. Doumas and V. G. Papanicolaou, The Siblings of the Coupon Collector, 2012, submitted.

[12] R. Durrett, *Probability: Theory and Examples*, Third Edition, Duxbury Advanced Series, Brooks/Cole—Thomson Learning, Belmont, CA, USA, 2005.

[13] B. Efron and R. Thisted, Estimating the number of unsen species: How many words did Shakespeare know?, *Biometrika*, **63**(3), (1976), 435–447.

[14] P. Erdős and A. Rényi, On a classical problem of probability theory, *Magyar. Tud. Akad. Mat. Kutató Int. Kőzl.*, **6**, (1961), 215–220.

[15] W. Feller, *An Introduction to Probability Theory and Its Applications*, Vol. I, First Edition, John Wiley & Sons, Inc., New York, 1950.

[16] P. Garrett, Cryptographic primitives, *Proceedings of Symposia in Applied Mathematics*, Baltimore, Maryland, American Mathematical Society, **62**, (2003), 1–63.

[17] C.M. Goldie, R. Cornish and C.L. Robinson, Computer Assisted Assessment: How many questions are enough?, arXiv:math.PR/1002.2114, (2010).

[18] H.J. Goodwin, On cartophily and motor cars, *Math. Gazette*, **33**(305), (1949), 169–171.

[19] D. Mc Neil, Estimating an Author's Vocabulary, *J. American Statistical Assosiation*, **68**, (1973), 92–96.

[20] A.W. Marshall and I. Olkin, *Inequalities: Theory of Majorization and its Applications*, Academic Press, New York, 1979.

[21] F.G. Maunsell, A problem in cartophily, *Math. Gazette*, **22**(251), (1938), 328–331.

[22] P. Neal, The Generalised Coupon Collector Problem, *J. Appl. Prob.*, **45**, (2008), 621–629.

[23] D.J. Newman and L. Shepp, The Double Dixie Cup problem, *Amer. Math. Monthly*, **67**, (1960), 58–61.

[24] Luke O'Connor, Entropy Bounds for Traffic Confrmation, *IACR, Technical Report*, **365**, (2008).

[25] A. Poon, B.H. Davis and L. Chao, The Coupon collector and the suppressor mutation: Estimating the number of compensatory mutations by maximum likelihood, *Genetics*, **170**(3), (2005), 1323–1332.

[26] A. Rafferty, T. L. Griffiths and D. Klein, Convergence bounds for language evolution by iterated learning, *Proceed. 31st Annual Conference of the Cognitive Science Society*, (2009).

[27] S. Ross, *A First Course in Probability*, Seventh Edition Pearson Prentice Hall, 2006.

[28] Y. Sasaki, Y. Li, H. Sakamoto and Kazuo Sakiyama, Coupon Collector's Problem for Fault Analysis against AES - High Tolerance for Noisy Fault Injections, *Proceedings of Financial Cryptography*, Japan, Spinger, (2013), 213–220.

[29] H.V. Schelling, Auf der Spur des Zufalls, *Deutsches Statistisches Zentralblatt*, **26**, (1934), 137–146.

[30] H.V. Schelling, Coupon Collecting for Unequal Probabilities, *Amer. Math. Monthly*, **61**(5), (1954), 306–311.

[31] The Dixie Cup Company History:
http://academicmuseum.lafayette.edu/special/dixie/company.html.

[32] A Coupon Collector Problem simulation in the case of equal probabilities:
www-stat.stanford.edu/∼susan/surpise/Collector.html.

[33] Fish of Lake Michigan, University of Wisconsin Sea Grant Institute:
http://www.seagrant.wisc.edu/greatlakesfish/LakeMichFishIndex.html.