NATIONAL SECURITY AGENCY
FORT GEORGE G. MEADE, MARYLAND 20755-6000

Case: 102241
9 September 2026

JOHN GREENEWALD
27305 W LIVE OAK RD
STE 1203
CASTAIC CA 91384-4520

Dear John Greenewald:

This responds to your Freedom of Information Act (FOIA) request dated 6 August 2017, for "the NSA TECHNICAL JOURNAL article, "Coupon Collecting and Cryptology." A copy of your request is enclosed. Your request has been processed under the FOIA and the document responsive to your request is enclosed. Certain information, however, has been protected in this enclosure.

Some of the withheld information has been found to be currently and properly classified in accordance with Executive Order 13526. The information meets the criteria for classification as set forth in Subparagraph (c) of Section 1.4 and remains classified SECRET as provided in Section 1.2 of Executive Order 13526. The information is classified because its disclosure could reasonably be expected to cause serious damage to national security. Because the information is currently and properly classified, it is exempt from disclosure pursuant to the first exemption of the FOIA (5 U.S.C. Section 552(b)(1)). The information is exempt from automatic declassification in accordance with Section 3.3(b) sub-paragraph (1) of E.O. 13526.

In addition, this Agency is authorized by various statutes to protect certain information concerning its activities. We have determined that such information exists in this document. Accordingly, those portions are exempt from disclosure pursuant to the third exemption of the FOIA, which provides for the withholding of information specifically protected from disclosure by statute. The specific statutes applicable in this case are: Title 50 U.S. Code 3024(h); and Section 6, Public Law 86-36 (50 U.S. Code 3605).

Please be advised that the Agency reasonably foresees that disclosure of the withheld information would be harmful to an interest that is protected by the identified exemptions.

Since information was withheld from the enclosure, you may construe this as a partial denial of your request.

You may appeal this decision in the manner outlined below. NSA will endeavor to respond within 20 working days of receiving your appeal, absent any unusual circumstances.

- The appeal must be sent via U.S. postal mail, fax or electronic delivery (e-mail) to:

  NSA FOIA/PA Appeal Authority
  National Security Agency
  9800 Savage Road STE 6932
  Fort George G. Meade, MD 20755-6932

  Facsimile: 443-479-3612

  Email: FOIA_PA_Appeals@nsa.gov

- It must be postmarked or delivered electronically within 90 calendar days of the date of this letter. Appeals filed later than 90 days will not be addressed.

- Include the case number provided above.

- Describe with sufficient detail why you believe the denial of requested information was unwarranted.

You may contact the Office of Government Information Services (OGIS) at the National Archives and Records Administration to inquire about the FOIA mediation services they offer at:

Office of Government Information Services
National Archives and Records Administration
8601 Adelphi Rd - OGIS
College Park, MD 20740
ogis@nara.gov / 877-684-6448 / (Fax) 202-741-5769

You may also contact our FOIA Public Liaison at foialo@nsa.gov for any further assistance and to discuss any aspect of your request.

Sincerely,

Sally A. Nicholson

SALLY A. NICHOLSON
Chief, FOIA/PA Division
NSA Initial Denial Authority

Encls:
a/s

Coupon Collecting and Cryptology

BY MARIE CAPOZZA

~~Secret~~

1. INTRODUCTION

It is my intention to present in this paper some well-known and some new results found in the literature on coupon collecting and to point out some cryptologic applications. Sections 1 through 4 are unclassified, whereas section 5 on cryptologic applications and the bibliography and references are classified SECRET.

Coupon Collecting Defined.

Let us assume that a manufacturer encloses in each package of his product one coupon that is chosen randomly from a stock of K different coupons. A customer who collects these coupons may gather some of them several times before he completes a set. The problem is to determine the probability of obtaining r distinct of the K coupons in n packages, where $0 \leq r \leq K$.

This is the classical occupancy problem, which is concerned with the random distribution of a specified number of objects (n) in a given number of cells (K).

2. SOME PRELIMINARY RESULTS

Probability Function.

Given events $A_1, A_2, \ldots, A_K$, we wish to compute the probability of the event that at least one among the A_j occurs. That is, we look for $A = A_1 \cup A_2 \cup A_3 \cup \ldots \cup A_K$. It is not, however, sufficient to know the probability of the individual events A_j; we must also be concerned with all possible overlaps. Let $\{i_1, i_2, \ldots, i_r\}$ be r distinct integers from the set $\{1, 2, \ldots, K\}$. Then, for every pair (i_1, i_2), every triple (i_1, i_2, i_3), etc., we must know the probability of the joint occurrence of events A_{i_1} and A_{i_2}; the joint occurrence of events A_{i_1}, A_{i_2}, and A_{i_3}; etc. Let

$$p_{i_1} = P(A_{i_1}), p_{i_1 i_2} = P(A_{i_1} \cap A_{i_2}), \ldots, p_{i_1 i_2 \ldots i_K} = P(A_{i_1} \cap A_{i_2} \cap \ldots \cap A_{i_K}).$$

The order of subscripts is irrelevant, but we shall always write the subscripts in increasing order. Thus, the probability of events

A_2, A_6, and A_8 occurring simultaneously will have the unique representation p_{268}. Let

$$S_1 = \Sigma p_{i_1},\ S_2 = \Sigma p_{i_1 i_2},\ \ldots,\ S_r = \Sigma p_{i_1 i_2 \ldots i_r},\ \ldots,\ S_K = p_{12 \ldots K}.$$

Each summation is overall $i_1, i_2, \ldots, i_r$ such that $i_1 < i_2 < \ldots < i_r \leq K$. In the sums, then, each combination appears once and only once; hence, S_r has (K/r) terms, and S_K has $(K/K) = 1$ term, $p_{12 \ldots K}$, which is the probability of simultaneously realizing all K events.

THEOREM: The probability P of the realization of at least one among the events $A_1, A_2, \ldots, A_K$ is given by [1]

$$P = S_1 - S_2 + S_3 - \ldots \pm S_K. \tag{1}$$

In our collector's problem we are purchasing n packages with coupons of K distinct kinds distributed at random in the packages. Each arrangement, therefore, has probability $1/K^n$. We seek the probability $p_r(n,K)$ of not getting r of K kinds, i.e., of getting $(K-r)$ different coupons.[2]

Let A_j be the event that we do not get coupon type j, where j can take on values $1, 2, \ldots, K$. In this event, all n packages have coupons of the $(K-1)$ remaining types and this can happen in $(K-1)^n$ ways. Similarly, there are $(K-2)^n$ arrangements, if we do not get 2 preassigned types of coupon, etc.

Therefore,

$$p_{i_1} = \left(1 - \frac{1}{K}\right)^n,\ p_{i_1 i_2} = \left(1 - \frac{2}{K}\right)^n,\ \ldots,$$

$$p_{i_1 i_2 \ldots i_{K-1}} = \left(1 - \frac{K-1}{K}\right)^n,\ p_{12 \ldots K} = 0,$$

and for $1 \leq \nu \leq K$

$$S_\nu = \binom{K}{\nu}\left(1 - \frac{\nu}{K}\right)^n.$$

The probability that we are missing at least one type of coupon is given by (1) and therefore, the probability that all types are collected is

$$1 - P = 1 - S_1 + S_2 - \ldots \pm S_K.$$

---

[1] See [6], p. 89 for proof.

[2] H. von Schelling [23] treats the general case where the i^{th} coupon has probability Π_i and $\Pi_1 + \Pi_2 + \ldots + \Pi_K = 1$.

Thus

$p_o\ (n, K)$, the probability of getting all K different types of coupon in a sample of size n, is

$$\sum_{\nu=0}^{K} (-1)^\nu \binom{K}{\nu}\left(1 - \frac{\nu}{K}\right)^n \qquad (2)$$

Now consider the distribution in which exactly r types of coupon are not present. These r types can be chosen in (K/r) ways. Then the $(K-r)$ types collected are distributed among the n packages in $(K-r)^n\ p_o(n, K-r)$ ways. Dividing by K^n, we find the probability of finding exactly $(K-r)$ different types of coupons in n packages.

$$\begin{aligned} p_r\ (n, K) &= \binom{K}{r}\left(1 - \frac{r}{K}\right)^n p_o\ (n, K-r) \qquad (3) \\ &= \binom{K}{r} \sum_{\nu=0}^{K-r} (-1)^\nu \binom{K-r}{\nu}\left(1 - \frac{r+\nu}{K}\right)^n \\ &= \binom{K}{r} \frac{\Delta^{K-r}\ (0^n)}{K^n}. \end{aligned}$$ [3]

Table 1 provides a numerical illustration.

TABLE 1

Probabilities $p_r\ (n, 10)$

r	$n = 10$	$n = 18$
0	0.000 363	0.134 673
1	0.016 330	0.385 289
2	0.136 080	0.342 987
3	0.355 622	0.119 425
4	0.345 144	0.016 736
5	0.128 596	0.000 876
6	0.017 189	0.000 014
7	0.000 672	0.000 000
8	0.000 005	0.000 000
9	0.000 000	0.000 000

[3]The operator Δ has the following meaning assigned to it in the calculus of finite differences: $\Delta f\ (x) = f(x + 1) - f(x)$ and $\Delta^m\ 0^n$ is the result of placing $x = 0$ in $\Delta^m\ x^n$. See [7], Table XXII.

Mean Waiting Time.

Again we have K distinct coupons sampled with replacement. Because of repetition, a sample of n packages will, in general, contain fewer than K distinct coupons. As the number of packages purchased increases, new coupons will appear more and more rarely. We are interested in knowing how many packages of the product we need buy to obtain r distinct types of coupon, where $0 \leq r \leq K$.

On the first purchase we collect one type.

Let X_j be the number of drawings necessary to get the next, $(j+1)st$ different coupon. Then,

$$\overline{S}_r = 1 + X_1 - X_2 + \ldots X_{r-1}, \tag{4}$$

where $\overline{S}_r$ is the sample size necessary to get r of K types of coupon.

Once we have j different coupons, the probability of a new type is

$$p_j = \frac{K - j}{K}. \tag{5}$$

The number, X_j, of drawings up to and including the drawing of a new element is $1+Y_j$, where random variable Y_j is the number of failures preceding the first success in Bernoulli trials with $p_j = (K-j)/K$ and $q_j = j/K$. Y_j has probability function $q_j^i \, p_j$ and $E\ (Y_j) = q_j/p_j$.

$$E\ (X_j) = E\ (1 + Y_j) = 1 + q_j/p_j = 1 + \frac{j/K}{1-j/K} = K/(K-j) \tag{6}$$

and therefore

$$\begin{aligned} E\ (\overline{S}_r) &= E\ (1 + X_1 + \ldots + X_{r-1}) = 1 + \frac{K}{K-1} + \frac{K}{K-2} + \\ &\qquad \ldots + \frac{K}{K-r+1} \\ &= K\left[\frac{1}{K} + \frac{1}{K-1} + \ldots + \frac{1}{K-r+1}\right]. \end{aligned} \tag{7}$$

For $r = K$ we have

$$E\ (\overline{S}_K) = K\left[\frac{1}{K} + \frac{1}{K-1} + \ldots + 1\right], \tag{8}$$

the expected number of drawings to exhaust the entire population. Suppose that we have decimal digits. Then

$$E\ (\overline{S}_{10}) = 29.29$$

is the expected length of the random digit sequence in order to have all 10 digits represented. However, note that

$$E\,(\overline{S}_5) = 6.46.$$

This means that we can expect to get half of the population in about 6 or 7 drawings, whereas the second half of the population requires 23 more drawings, on the average.

For sampling from random letters of alphabet size 26, i.e., $K = 26$,

$$E\,(\overline{S}_{26}) = 100.20.$$[4]

We have the following approximation:

$$\frac{1}{K} + \frac{1}{K-1} + \ldots + \frac{1}{K-r+1} = \sum_{\nu=1}^{K} \frac{1}{\nu} - \sum_{\nu=1}^{K-r+1} \frac{1}{\nu}$$

$$= \log K + \frac{1}{2} + \frac{1}{2K} + \int_1^K \frac{P_1(x)}{x^2}\,dx - \Big[\log(K-r+1)$$

$$+ \frac{1}{2} + \frac{1}{2(K-r+1)} + \int_1^{K-r+1} \frac{P_1(x)}{x^2}\,dx\Big],$$

which for K very large

$$\approx \log K - \log(K-r+1) = \log \frac{K}{K-r+1}.$$

Thus

$$E\,(\overline{S}_r) \approx K \log \frac{K}{K-r+1}. \tag{9}$$

In particular, when $r = K$

$$E\,(\overline{S}_K) \approx K \log K. \tag{10}$$

The variance is given by

$$\text{Var}\,(X_j) = \text{Var}\,(1 + Y_j) = q_j/p_j^2$$

$$= \frac{j/K}{\left(\frac{K-j^2}{K}\right)} = K\left[\frac{j}{(K-j)^2}\right], \tag{11}$$

where Y_j is a random variable with probability function $q_j^i p_j$.

[4]See Table 3 on page 00 for the waiting time to collect all K types of coupon where $1 \leq K \leq 50$.

Since all random variables X_i are independent,

$$\mathrm{Var}\ (\overline{S}_r) = \mathrm{Var}\ (1 + X_1 + X_2 + \ldots + X_{r-1}) \tag{12}$$

$$= \mathrm{Var}\ (X_1) + \mathrm{Var}\ (X_2) + \ldots + \mathrm{Var}\ (X_{r-1})$$

$$= \frac{K}{(K-1)^2} + \frac{2K}{(K-2)^2} + \ldots + \frac{(r-1)\ K}{(K-r+1)^2}$$

$$= K\left[\frac{1}{(K-1)^2} + \frac{2}{(K-2)^2} + \ldots + \frac{(r-1)}{(K-r+1)^2}\right]^{5} \tag{13}$$

or equivalently

$$\mathrm{Var}\ (\overline{S}_r) = K^2\left[\frac{1}{K^2} + \frac{1}{(K-1)^2} + \ldots + \frac{1}{(K-r+1)^2}\right]$$
$$- K\left[\frac{1}{K} + \frac{1}{K-1} + \ldots + \frac{1}{K-r+1}\right]^{6} \tag{14}$$

From Yoxall's form for the variance with $r = K$ and for K very large,

$$K^2\left[\frac{1}{K^2} + \ldots + 1\right] \approx K^2\left[1 + \frac{1}{2^2} + \frac{1}{3^2} + \ldots\right] \approx \frac{K^2\pi^2}{6}$$

and $$\left[\frac{1}{K} + \frac{1}{K-1} + \ldots + 1\right] \approx \log K.$$

Combining these, we have

$$\mathrm{Var}\ (\overline{S}_K) \approx \frac{K^2\ \pi^2}{6} - K \log K \tag{15}$$

and as $K \to \infty$

$$\frac{\mathrm{Var}\ (\overline{S}_K)}{K^2} = \frac{\pi^2}{6}. \tag{16}$$

From Table 2 we see that most of the variance is due to the last coupon.

[5] See Feller [6], p. 224.

[6] Yoxall's form for the variance [28].

TABLE 2
Variance when $K = 10$

r	Var (S_r)
2	0.1234
3	0.4359
4	1.0482
5	2.1593
6	4.1593
7	7.9093
8	15.6970
9	35.6870
10	125.6870

3. DISTRIBUTION OF THE FIRST REPEAT[7]

We are to continue random placement of balls in K cells until for the first time a ball is placed in a cell already occupied, i.e., a duplication of coupon occurs. The problem is to find the mean waiting time until the first duplication occurs.

First, we find the probability that the process terminates at the n^{th} step, i.e., we get a duplication at the n^{th} trial. Let $\vec{J} = (j_1, j_2, \ldots, j_n)$ denote the $1^{st}, 2^{nd}, \ldots, n^{th}$ ball placed in cell number $j_1, j_2, \ldots, j_n$. We have $1 \leq j_i \leq K$, since there are K categories.

The first $(n-1)$ are all different and the last one is the same as one of the first $(n-1)$. Each coupon is selected with equal probability; therefore the probability of any particular $\vec{J}$ is $(1/K)^n$. We can select the first $(n-1)$ components of the vector in $K\,(K-1)\,(K-2)\ldots(K-n+2)$ ways and the last component j_n, being a duplication of one of the first $(n-1)$ components, can be selected in $(n-1)$ ways.

Therefore,

$$P \text{ (duplication at the } n^{th} \text{ trial)} = \frac{(K)_{n-1}\,(n-1)}{K^n}$$

$$= \frac{K(K-1)\ldots(K-n+2)\,(n-1)}{K^n}$$

$$= \left(\frac{K}{K}\right)\left(\frac{K-1}{K}\right)\cdots\left(\frac{K-n+2}{K}\right)\left(\frac{n-1}{K}\right)$$

$$= \left(1-\frac{1}{K}\right)\left(1-\frac{2}{K}\right)\cdots\left(1-\frac{n-2}{K}\right)\left(\frac{n-1}{K}\right) = q_n \qquad (17)$$

[7] See [19].

P (duplication at 1^{st} trial) $= 0$

P (duplication at 2^{nd} trial) $= 1/K$

P (process lasts longer than n trials) $= P$ (1^{st} n are all different)

$$= 1 - (q_1 + q_2 + \ldots + q_1) = \frac{(K)_n}{K^n} = p_n \qquad (18)$$

$$p_{K+j} = \frac{(K)_{K+j}}{K^{K+j}} = 0 \text{ for } j \geq 1$$

and

$$q_1 + q_2 + \ldots + q_{K+1} = 1.$$

Now we find the mean.

$$E\,(N) = \sum_{n=2}^{K+1} n\, q_n = \sum_{n=2}^{K+1} \frac{n\,K!}{(K-n+1)!} \frac{n-1}{K^n}$$

$$= \sum_{n=0}^{K+1} \frac{K!}{(K-n+1)!} \frac{n\,(n-1)}{K^n}.$$

Since

$$\binom{K}{j} = \binom{K-1}{j} + \binom{K-1}{j-1},$$

we have

$$E\,(N) = \sum_{n=0}^{K+1} \binom{K-1}{n-2} \frac{n!}{K^{n-1}}$$

$$= \sum_{n=0}^{K+1} \binom{K}{n-1} \frac{n!}{K^{n-1}} - \sum_{n=0}^{K+1} \binom{K-1}{n-1} \frac{n!}{K^{n-1}}$$

$$= \sum_{n=0}^{K} \frac{K!}{(K-n)!} \frac{n+1}{K^n} - \sum_{n=0}^{K} \frac{K!}{(K-n)!} \frac{n}{K^n}$$

$$= \sum_{n=0}^{K} \frac{K!}{(K-n)!} \frac{1}{K^n}$$

$$= \frac{K!}{K!} \frac{1}{K^0} + \frac{K!}{(K-1)!} \frac{1}{K} + \cdots + \frac{K!}{(K-K)!} \frac{1}{K^K}$$

$$= 1 + 1 + \left(\frac{K-1}{K}\right) + \ldots + \left(\frac{K-1}{K}\right)\left(\frac{K-2}{K}\right) \ldots \left(\frac{1}{K}\right)$$

$$= 2 + \left(1 - \frac{1}{K}\right) + \left(1 - \frac{1}{K}\right)\left(1 - \frac{2}{K}\right) + \ldots + \left(1 - \frac{1}{K}\right)\left(1 - \frac{2}{K}\right) \ldots \left(1 - \frac{K-1}{K}\right),$$

the mean waiting time to the first repeat.

We can find the asymptotic expansion for the waiting time by introducing the gamma function integral:

$$
\begin{aligned}
E\ (N) &= \sum_{n=0}^{K} \frac{K!}{(K-n)!} \frac{1}{K^n} \\
&= \sum_{n=0}^{K} \binom{K}{n} \frac{n!}{K^n} \\
&= \sum_{n=0}^{K} \binom{K}{n} \frac{1}{K^n} \int_0^\infty e^{-x} x^n \, dx \\
&= \int_0^\infty e^{-x} \left[\sum_{n=0}^{K} \binom{K}{n} \left(\frac{x}{K}\right)^n \right] dx \\
&= \int_0^\infty e^{-x} \left(1 + \frac{x}{K}\right)^K dx \\
&= K \int_0^\infty e^{-Ky} (1+y)^K \, dy \\
&= K \int_0^\infty e^{-K\,[y - \log(1+y)]} \, dy.
\end{aligned}
$$

Make the substitution $t = y - \log\ (1 + y)$.

$$E\ (N) = K \int_0^\infty e^{-Kt} \left(\frac{dy}{dt}\right) dt$$

$$t = y - \log\ (1+y) - y^2/2 - y^3/3 + y^4/4 \ldots = y^2/2\,[1 - 2/3\ y + 1/2\, y^2 \ldots]$$

$$t^{1/2} = y/\sqrt{2}\ [1 - 2/3\ y + 1/2\ y^2 \ldots]^{1/2}$$

$$y = \sqrt{2}\, t^{1/2} + 2/3\ t + \frac{\sqrt{2}}{18} t^{3/2} - \frac{2}{135} t^2 \ldots$$

$$\frac{dy}{dt} = \frac{\sqrt{2}}{2} t^{-1/2} + 2/3 + \frac{\sqrt{2}}{12} t^{1/2} - \frac{4}{135} t \ldots$$

Integrating, we obtain

$$
\begin{aligned}
E\ (N) &= K \int_0^\infty e^{-Kt} \left[\frac{\sqrt{2}}{2} t^{-1/2} + 2/3 + \frac{\sqrt{2}}{12} t^{1/2} - \frac{4}{135} t + \ldots \right] dt \\
&= K \left[\frac{\sqrt{2\pi K}}{2K} + \frac{2}{3K} + \frac{\sqrt{2}}{12} \frac{\sqrt{\pi}}{2K\ K} - \frac{4}{135} \frac{1}{K^2} + \ldots \right] \\
&= \frac{\sqrt{2\pi K}}{2} + \frac{2}{3} + \frac{\sqrt{2\pi K}}{24K} - \frac{4}{135} + 0\left(\frac{1}{K^{3/2}}\right),
\end{aligned}
$$

where $0(\ )$ is the large 0-notation. Since

$$E\ (N^2)\ =\ 2K\ +\ E\ (N),$$

we have

$$\begin{aligned}\text{Var}\ (N) &= 2K + E\ (N) - (E\ (N))^2 \\ &= K\left(2 - \frac{\pi}{2}\right) - \frac{\sqrt{2\pi K}}{6} + \frac{8 - 3\pi}{36} + \frac{17\sqrt{2\pi}}{1080\ K} + 0\left(\frac{1}{K}\right).\end{aligned}$$

Now we find the asymptotic distribution as $K^{\to\infty}$. Using Stirling's formula, P (duplication at n^{th} trial)

$$= \frac{K!\ (n-1)}{(K-n+1)!\ K^n}$$

$$\approx \frac{K^{\ +1/2}\, e^{-K}\, \sqrt{2\pi}\ (n-1)}{(K-n+1)!^{K-n+3/2}\, e^{-K+n-1}\, \sqrt{2\pi}\ K^n}.$$

Since $(n-1) \sim \sqrt{K}$, let $(n-1) = \sqrt{K}x$. Then

$$P\ (n = \sqrt{K}x) = \frac{K^{K+1/2}\, e^{-K}\, \sqrt{K}x}{(K - \sqrt{K}x)^{K-\sqrt{K}x+1/2}\, e^{-K+\sqrt{K}x}\, K^{\sqrt{K}x+1}}$$

$$= \frac{x}{\sqrt{K}}\ \frac{1}{e^{\sqrt{K}x}\left(1 - \frac{x}{\sqrt{K}}\right)^{K-\sqrt{K}x+1/2}}.$$

As K becomes large

$$\log\left[\frac{1}{e^{\sqrt{K}x}\left(1 - \frac{x}{\sqrt{K}}\right)^{K-\sqrt{K}x+1/2}}\right] \approx \frac{x^2}{2}$$

$$P\ (n = \sqrt{K}x) \approx \frac{x}{\sqrt{K}}\, e^{-x^2/2}$$

$$P\ (n/\sqrt{K} = x) = x\, e^{-x^2/2},$$

and

$$\sum_{A \le \frac{n}{\sqrt{K}} \le B} \frac{K!\ (n-1)}{(K-n+1)!\ K^n} \approx \int_A^B x\, e^{-x^2/2}\, dx = e^{-A^2/2} - e^{-B^2/2}.$$ [8]

Hence we have asymptotically

$$P\left(\frac{n}{\sqrt{K}} \ge A\right) = e^{-A^2/2}.$$

[8] This is the Rayleigh distribution with parameter $\alpha = 1$. See [17], p. 181.

4. LIMITING DISTRIBUTION.

For large K, use of (3) for computation becomes exceedingly tedious. Several authors have stated without proof that when n/K is suitably restricted the limiting distribution of the number of missing coupons as n, K approach infinity is normal. The following will very briefly indicate that a proof of this is due to Irving Weiss [27].

Asymptotic normality is shown by using moments rather than probabilities. These moments are obtained by defining a simple random variable over our sample space. It can be shown that the moments converge to the moments of the normal distribution and from this it follows (by a theorem in Uspensky [24]) that the distribution of the random variable converges uniformly to the normal distribution

$$\text{Let } \alpha = \frac{n}{K} \qquad \alpha > 0.$$

We define a random variable X_{ij} as follows:

$$\begin{aligned} X_{ij} &= 1 \text{ if coupon } j \text{ is drawn on the } i^{th} \text{ draw} \\ &= 0 \text{ otherwise} \\ j &= 1, \ldots, K \\ i &= 1, \ldots, n. \end{aligned}$$

The K elements are sampled one at a time and then replaced.

We define another random variable

$$\begin{aligned} X_j &= 1 \text{ if } \sum_{i=1}^{n} X_{ij} = 0 \\ &= 0 \text{ otherwise.} \end{aligned}$$

That is, X_j is one if in drawing a sample of size n, the j^{th} coupon never appears. Assuming that all K coupons are equally likely to be drawn and that the n trials are independent, we have

$$P\,(X_{ij} = 0) = 1 - \frac{1}{K}$$

$$P\,(X_j = 1) = E\,(X_j) = \left(1 - \frac{1}{K}\right)^n.$$

Also as

$$P\,(X_{ij} \text{ or } X_{ik} = 1) = \frac{2}{K},$$

then

$$P\,(X_{ij} \text{ and } X_{ik} = 0) = \left(1 - \frac{2}{K}\right)$$

and

$$P\ (X_j X_k = 1) = E\ (X_j X_k) = \left(1 - \frac{2}{K}\right)^n;$$

likewise

$$E\ (X_{i_1} X_{i_2} \ldots X_{i_r}) = \left(1 - \frac{r}{K}\right)^n.$$

If we put

$$X = \sum_{i=1}^{K} X_i ,$$

we find

$$E\ (X) = K\left(1 - \frac{1}{K}\right)^n = K\left(1 - \frac{1}{K}\right)^{\alpha K}$$

$$\lim_{K \to \infty} \frac{E\ (X)}{K} = e^{-\alpha}$$

X is the number of distinguishable coupons that have not appeared in the above sampling process, while $(E(X))/K$ is the proportion of such coupons that have not appeared. We see that as K becomes infinite, $E(X)$ becomes infinite, but $(E(X))/K$ approaches a limit.

The central moments of the normal distribution are:

$$\mu_k = (\sigma^2)^{k/2}\, 1 \cdot 3 \cdot 5 \ldots \cdot (k-1) \qquad \text{for } k \text{ even}$$
$$= 0 \qquad \text{for } k \text{ odd}.$$

It is shown that X defined above has an asymptotically normal distribution by showing that

$$\lim_{K \to \infty} \frac{\mu_k}{(\sigma^2)^{k/2}} = 1 \cdot 3 \ldots (k-1) \text{ for } k \text{ even,} \tag{19}$$
$$= 0 \text{ for } k \text{ odd.}$$

Begin by computing

$$\sigma^2 = E\ (X^2) - [E\ (X)]^2$$
$$= K\left(1 - \frac{1}{K}\right)^{\alpha K} + K\ (K-1)\left(1 - \frac{2}{K}\right)^{\alpha K} - K^2\left(1 - \frac{1}{K}\right)^{2\alpha K}.$$

Each of the above three terms on the right is analytic in the annular region

$$R = \{K : 0 < n \le K \le \overline{R} < \infty\}.$$

Also, a finite sum of analytic functions is analytic. Consequently, σ^2 has a unique Laurent series expansion in powers of K. This implies that no matter how the series is evaluated, we will always obtain the same result.

Using the following formulae,

$$\left(1 - \frac{j}{K}\right)^{kK} = e^{kK} \log\left(1 - \frac{j}{K}\right) = e^{-kj}\left(1 + \frac{j}{2K} + \frac{j^2}{3K^2} + \dots\right)$$

$$= e^{-kj}\left(1 - \frac{kj}{2K} - \frac{1}{K^2}\left(\frac{kj^3}{3} - \frac{k^2 j^4}{8}\right)\right) + o\left(\frac{1}{K^2}\right),$$

where $o(\ )$ is the small o-notation, we obtain

$$\sigma^2 = K\ (e^{-\alpha} - (\alpha + 1)\ e^{-2\alpha}) + 0\ (1) \qquad (20)$$

$$= K\ \sigma^2_0 + 0\ (1).$$

In somewhat analogous manner, it can be shown that for k even the k^{th} central moment, μ_k, can be uniquely expanded in R as a Laurent series of the following form:

$$\mu_k = P_o\, K^{k/2} + P_1\, K^{k/2-1} + \dots$$

$$= P_o\, K^{k/2} + 0\ (K^{k/2-1}), \qquad (21)$$

where

$$P_o = (\sigma^2_o)^{k/2} \quad 1 \cdot 3 \dots (k-1).$$

For k odd it can be shown that

$$\mu_k = Q_o\, K^{(k-1)/2} + Q_1\, K^{(k-3)/2} + \dots . \qquad (22)$$

Using (20), (21), and (22), formula (19) will follow.

It is well known $\mu_k/((\sigma^2)^{k/2})$ is the k^{th} central moment of a normalized (or standardized) random variable obtained from X by

$$Z = \frac{X - E\ (X)}{\sigma}.$$

5. CRYPTOLOGIC APPLICATIONS

Coupon Collector's Test for Randomness.[9]

R. E. Greenwood tested the digits of π under the null hypothesis that they have random properties. Using the 2035 decimal approximation to π given by G. W. Reitwiesner[10], Greenwood began with the initial position 3 and found that 33 positions were required to collect all ten digits. Beginning again at position 34, he found that 18 posi-

[9] [9] and [11]

[10] ibid.

tions were needed to collect another complete set. Proceeding in this manner, 67 sequences of complete sets were obtained.

The observed results were then compared with the expected results using a chi-square test and no significant difference was noted. W. Jones[11] performed a similar test on the digits in the expansion of e and obtained the same result.

Statistical Inference: Coupon Collector's Test for the Number of Code Groups.

Suppose that there are K different code groups. Suppose that we have a message of n code groups of which r are different, $r \leq n$, since we may have repeats. If K, n, and r are known, we may want to find the mean and variance of the message size n required to give all (or all but $K-r$) of the code groups. Or, if K is unknown we may want to estimate K from the values of n and r.

In both cases, we assume that the underlying population is flat, a condition that is not easily satisfied.

(b)(3)-P.L. 86-36

Coupon Collecting and Key Card Generation.

Key card generation is carried out in [redacted] using the method of coupon collecting.

Suppose we have a 12 x 12 field on a key card in which we are to place punches in the following manner:

Place a punch at random in row 12, then proceed to place a punch at random in row 11, 10, . . . , 1 with the requirement that there be one and only one punch per column and one and only one punch per row. If in our random placement of a punch we hit a forbidden position, do not punch this position, but count this as a trial, and try again to place a punch at random in that row.

We wish to know how many trials we may expect to make before all 12 rows and columns have a punch, i.e., the waiting time until we collect all 12 coupons.

If it takes too many trials to complete our matrix, then our "random" process of position selecting is not truly random since it has a tendency to select one or more positions. Of major concern to [redacted] is how to set upper and lower bounds on the number of trials that should be allowed in filling in one card or group of cards. The method now used is to test 10 cards at a time using upper and lower bounds that have been set empirically.

---

[11] ibid.

(b) (1)
(b) (3) - 50 USC 3024(h)
(b) (3)-P.L. 86-36

[redacted]

Note: There are three papers written for NSA in the list of references ([1], [2], and [20]) that have not been discussed in this paper and that deserve special note.

In [1], Baum and Billingsley present four theorems dealing with the asymptotic distribution of W_K, where W_K is the drawing on which for the first time a_K+1 distinct population elements have been sampled, $0 \leq a_K < K$.

Chase and Rubin [2] discuss the approximation of the distribution function of sums of independent random variables distributed as

$$S_r = 1 + X_1 + X_2 + \ldots + X_{r-1}.$$

In [20] Rubin and Zidek obtain an exact distribution for S_r and then derive an asymptotic approximation to this distribution for the case $r = K$.

TABLE 3

Expected waiting time until all K coupons are collected, where K is the number of categories.

K	$E(K)$	K	$E(K)$
1	1.00	26	100.20
2	3.00	27	105.06
3	5.50	28	109.96
4	8.25	29	114.90
5	11.42	30	119.85
6	14.70	31	124.84
7	18.15	32	129.86
8	21.74	33	134.94
9	25.46	34	140.01
10	29.29	35	145.15
11	33.22	36	150.30
12	37.24	37	155.47
13	41.34	38	160.66
14	45.53	39	165.87
15	49.77	40	171.16
16	54.00	41	176.42
17	58.48	42	181.73
18	62.91	43	187.05
19	67.41	44	192.41
20	71.96	45	197.78
21	76.55	46	203.18
22	81.20	47	208.59
23	85.88	48	214.03
24	90.62	49	219.47
25	96.40	50	225.05

BIBLIOGRAPHY AND REFERENCES

[1] L. E. Baum and P. Billingsley, "Asymptotic Distributions for the Coupon Collector's Problem," SCAMP No. 27/64, September, 1964.

[2] G. Chase and H. Rubin, "Distribution of a Sum of Waiting Times in Coupon Collecting," Department of Statistics, Stanford University, Technical Report No. 109, August 12, 1965.

[3] *CPMC—Collected Papers in Mathematical Cryptology*, pp. 6, 41, 141–142, 144–146, 295–298, 305–306, 383, 524, 1081, 1141, 1509, 1792.

[4] H. Cramer, *Mathematical Methods of Statistics*, Princeton University Press, 1946.

[5] [redacted]

[6] W. Feller, *An Introduction to Probability and Its Applications*, Vol. 1.

[7] R. A. Fisher and F. Yates, *Statistical Tables for Biological Agricultural and Medical Research*, Oliver and Boyd, London and Edinburgh, 1948, Table XXII.

[8] [redacted]

[9] R. E. Greenwood, "Coupon Collector's Test for Random Digits," *Mathematical Tables and Other Aids to Computation*, Vol. IX, No. 49, January, 1955, pp. 1–5.

[10] R. E. Greenwood, "Coupon Collector's Problem and Statistical Inference as to Number of Code Groups," AFSA–34, April, 1952.

[11] R. E. Greenwood, "Coupon Collector's Test for Randomness," NSA–34, August, 1953, S32500.

[12] M. G. Kendall and B. B. Smith, "Randomness and Random Sampling Numbers," *Journal of the Royal Statistical Society*, Vol. 101, 1938, pp. 137–166.

[13] M. Kupperman, "Moments of the Distribution of the First Repeat," NSA–341, July, 1955.

[14] J. G. Mauldon, "A Problem in Coupon Collecting," SCAMP No. 26/64, September, 1964.

[15] F. G. Maunsell, "A Problem in Cartophily," *Mathematical Gazette*, Vol. 22, 1938, p. 328; Vol. 23, 1939, p. 77.

[16] [redacted]

[17] E. Parzen, *Modern Probability Theory and Its Application*, John Wiley and Sons, Inc. New York, 1960, pp. 76–86.

[18] G. Polya, "Eine Wahrscheinlichkeitsaufgabe zur Kundenwerbung," *Zeitschrift für Angewandte Mathematik und Mechanik*, Vol. 10, 1930, pp. 96–97.

[19] O. S. Rothaus, "Distribution of the First Repeat," NSA–341, July, 1955.

[20] H. Rubin and J. Zidek, "A Waiting Time Distribution Arising from the Coupon Collector's Problem," Department of Statistics, Stanford University, Technical Report No. 107, August 5, 1965.

[21] H. Rubin and R. Sitgreaves, "Probability Distributions Related to Random Transformations of a Finite Set," Technical Report No. 19A, January 22, 1954, Applied Mathematical and Statistics Laboratory, Stanford University.

[22] [redacted]

[23] H. von Schelling, "Auf der Spur des Zufalls," *Deutsches Statistisches Zentralblatt*, Vol. 26, 1934, pp. 137–146. Also, "Coupon Collecting for Unequal Probabilities," *American Mathematical Monthly*, Vol. 61, 1954, pp. 306–311.

[24] J. V. Uspensky, *Introduction to Mathematical Probability*, McGraw-Hill Book Co., New York, 1937.

(b)(1)
(b)(3) - 50 USC 3024(h)
(b)(3)-P.L. 86-36

[25] G. C. Wall, "Estimation of the Size of a Pack from the Number of Repeated Cards," *CPMC*, pp. 144–147, 10 February 1949.

[26] S. S. Wilks, *Mathematical Statistics*, John Wiley and Sons, Inc., New York, 1962, p. 28.

[27] I. Weiss, "Limiting Distributions in Some Occupancy Problems," Applied Mathematics and Statistics Laboratory, Stanford University, 15 July 1955.

[28] A. L. Yoxall, "The Mean and Variance in Maunsell's Cigarette Card Problem," *CPMC*, pp. 141–142, 7 December 1948.