コレクタの数理：
Coupon Collector 問題の Monte Carloシミュレーション

中　尾　泰　士*
Nakao Yasushi

1　はじめに

硬貨を入れてハンドルを回すとカプセルに入った玩具が出てくる機器「ガチャガチャ(1)」では，いくつかの種類の玩具がシリーズになっているのが一般的である。どの玩具が出てくるかは分からないため，コレクタ癖のある人はそれらすべてを収集するために何度も硬貨を投入することになる。また，「食玩」と呼ばれてスーパーマーケットの食品売場などで売られているものは，本来おまけであるはずの玩具の方が購買者の目当てになっている。箱を開けてみないと中の玩具が分からないものの場合，コレクタはシリーズすべてを集めるためにいくつも購入することになってしまう。このような仕組みで売られているシリーズものをすべて収集するためには，いったいいくつ購入する必要があるのか知りたいところだ。

このタイプの問題は，確率論においてしばしば「Coupon Collector 問題」と呼ばれているものである（たとえば，[1]Blom, Holst, & Sandell 1994 など)。この「Coupon Collector 問題」については，その最も簡単な場合において理論的な解を求めることが出来る（[2]Feller 1968)。

しかし，各 Coupon の出現確率が一様でない場合については理論的な取り扱いが難しい。本論はそのような一般化された「Coupon Collector 問題」を Monte Calro シミュレーションによって考察したものである。

2　古典的 Coupon Collector 問題

まず，問題の定式化を行おう。「Coupon Collector 問題」とは，

> 「N種類の Coupon c_i（$i = 1, 2, \cdots, N$）があり，その出現確率をそれぞれ p_iとする。1 回に 1 個の Coupon が得られるとき，すべての種類の Coupon を集めるために必要な回数はどれだけか」

という問題である。この問題において，ちょうど X 回目にすべての種類の Coupon が集まる確率を $W(X)$ としよう。明らかに，

$$W(X) = 0 \text{ for } X < N, \qquad (1)$$

である。また，Coupon の出現確率については当然ながら，

＊奈良産業大学情報学部　nakaoy@nara-su.ac.jp
(1)「ガシャポン」,「ガチャポン」などとも呼ばれる

$$\sum_{i=1}^{N} p_i = 1, \tag{2}$$

が成り立つ。

さて，「古典的」な場合とは，すべての Coupon の出現確率が等しい場合，すなわち，$p_i = 1/N$（$i = 1, 2, \cdots, N$）の場合を指す。このとき，$W(X)$ は，$X \geq N$ に対して，

$$W(X) = \sum_{j=0}^{N-1} (-1)^j {}_{N-1}\mathrm{C}_j \left(1 - \frac{1+j}{N}\right)^{X-1}, \tag{3}$$

で与えられる（[2] Feller 1968）。また，X の期待値 $E(X)$ は，

$$E(X) = N \sum_{i=1}^{N} \frac{1}{i}, \tag{4}$$

となる（[3] Baum & Billingsley 1965，[4] Dawkins 1991）。式 (3) と (4) の導出については付録 A を参照されたい。

3 Monte Calro シミュレーション

各 Coupon の出現確率が一様でない場合，問題の解析的な取り扱いは難しい。そのため，本論では Monte Calro シミュレーションを行うことで，この問題に対するいくつかの知見を引き出したい。シミュレーションは次のような手順で行った。

- 各 Coupon c_i に対して，その出現確率 p_i を設定する。
- 設定した出現確率 p_i に基づき，疑似乱数を用いて Coupon を 1 つずつ取り出すことをくり返す。
- すべての種類の Coupon が得られたら，その時の取り出し回数 X を記録する。
- この試行を1,000,000回くり返して，相対頻度から $W(X)$ を求める。

疑似乱数としては，簡便のため，Java 言語の java.lang.Math.random( ) メソッドを使用した。

3.1 古典的な場合

まず，「古典的」な場合について，$W(X)$ の理論的な分布 (3) とシミュレーションの結果を比較してみよう。図 1 は Coupon の種類が10種類（$N = 10$）の場合について，理論式 (3)（実線）の上に，シミュレーション結果のデータ点を重ねてプロットしたものである。

図 1 からは，$N = 10$の場合，$W(X)$ は $X = 23$　辺りに最頻値があり，$X \sim 40$より右に tail を引く分布となることが見てとれる。期待値 $E(X)$ は，$E(X) \simeq 29.29$（理論値）に対して，$E(X) \simeq 29.31$（シミュレーション）であった。

これらの比較から，若干の数値誤差を除いて，シミュレーションから $W(X)$ の分布についておおまかな傾向を知ることが出来そうである。

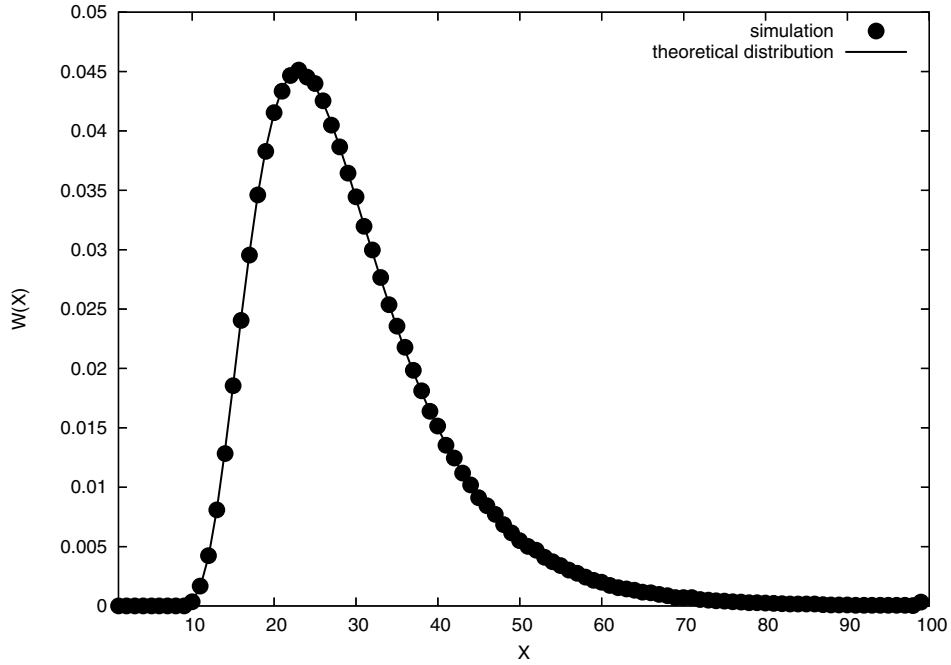

図1：$N=10$ の場合の $W(X)$。比較のために，理論的分布（実線）にシミュレーション結果の点をプロットしてある。

3.2　ある Coupon が稀少な場合

次に，ある Coupon が稀少な場合を考える。稀少な Coupon を c_1 とし，その出現確率p_1は，その他の Coupon c_i ($i=2,3,\cdots,N$) の出現確率に対して $1/K$ だけ小さいとする ($K\geq 1$)。すなわち，

$$\begin{cases} p_1 = \dfrac{p}{K}, \\ p_2 = p_3 = \cdots = p_N = p, \end{cases} \tag{5}$$

とする。このとき，式(2)より，

$$p_1 = \frac{1}{1+K(N-1)}, \tag{6}$$

$$p = \frac{K}{1+K(N-1)}, \tag{7}$$

となる。

稀少度をあらわすパラメータ K に対して，$W(X)$ の確率分布はどう変化するだろうか。シミュレーション結果を見てみよう。

図2は，いくつかの K の値に対して，シミュレーションで得られた$W(X)$ の分布をグラフにしたものである。Coupon の種類は $N=10$ に固定してある。

$K=1$ の場合は「古典的」な問題に帰着する。図2からは，K が大きくなるにつれて，分布の tail 部分の比率が相対的に大きくなっていく傾向が読み取れる。一方，分布の最頻値を与える X_*は Kの値にほとんどよらない。

K によって期待値 $E(X)$ がどのように変化するかをプロットしたものが図3である。K が大きくなるにつれて，期待値 $E(X)$ はほぼ直線的に増大する。シミュレーションした範囲では，大きな Kに対して，

$$E(X) \sim K(N-1)+1 = \frac{1}{p_1}, \tag{8}$$

のような近似が出来そうである。

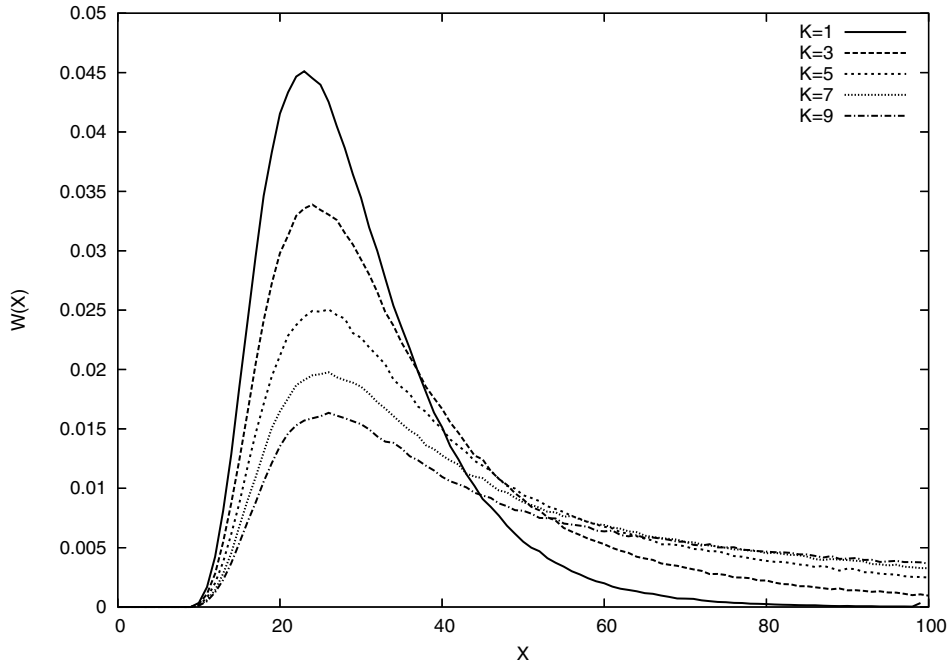

図 2：稀少パラメータ K を変化させた時の $W(X)$ の分布。$K=1$ が「古典的」な場合に相当する。すべて $N=10$ の場合。

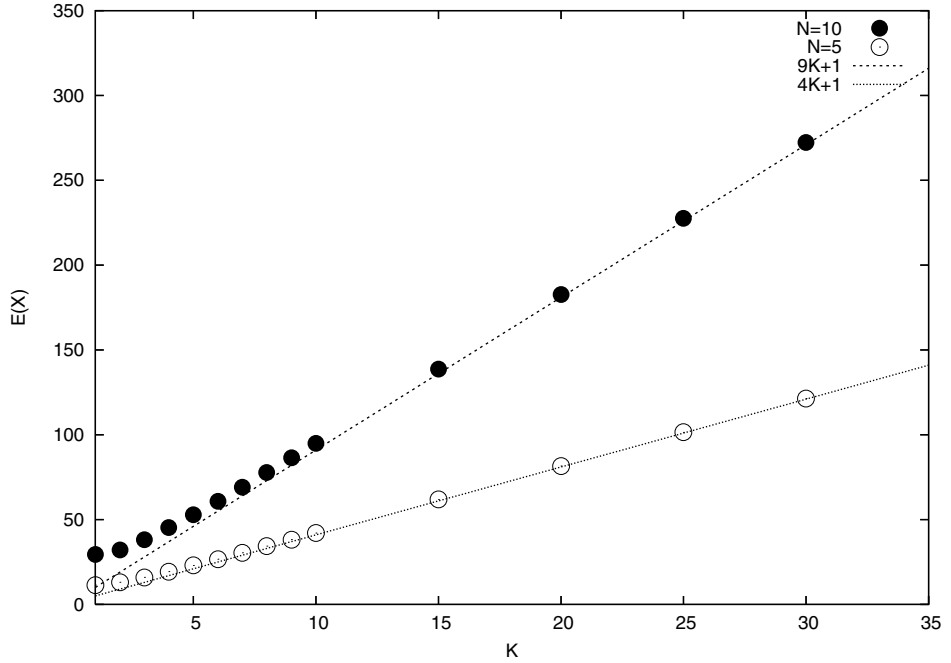

図 3：稀少パラメータ K に対する期待値 $E(X)$。$N=10$ の場合と $N=5$ の場合。それぞれについて，近似直線 (8) を重ねてある。

3.3 ある Coupon が頻出する場合

今度は逆に，Coupon c_1 がその他の Coupon に比べて出やすい場合を考えよう。これは，式(5)において，$0 < K \leq 1$ とすればよい。K が小さくなればなるほど，Coupon c_1 が頻出するような状況に相当する。

この場合のシミュレーション結果を図4と図5に示す。図4からは，K が小さくなるにつれ，分布 $W(X)$ の最頻値を与える X_* が次第に右にシフトしていくことが見てとれる。$K \geq 1$ の場合に，K の値によって X_* がほとんど変化しないのと対照的である。K が小さくなると X_* が右にシフトすることに対応して，期待値 $E(X)$ は増大していく（図5）。

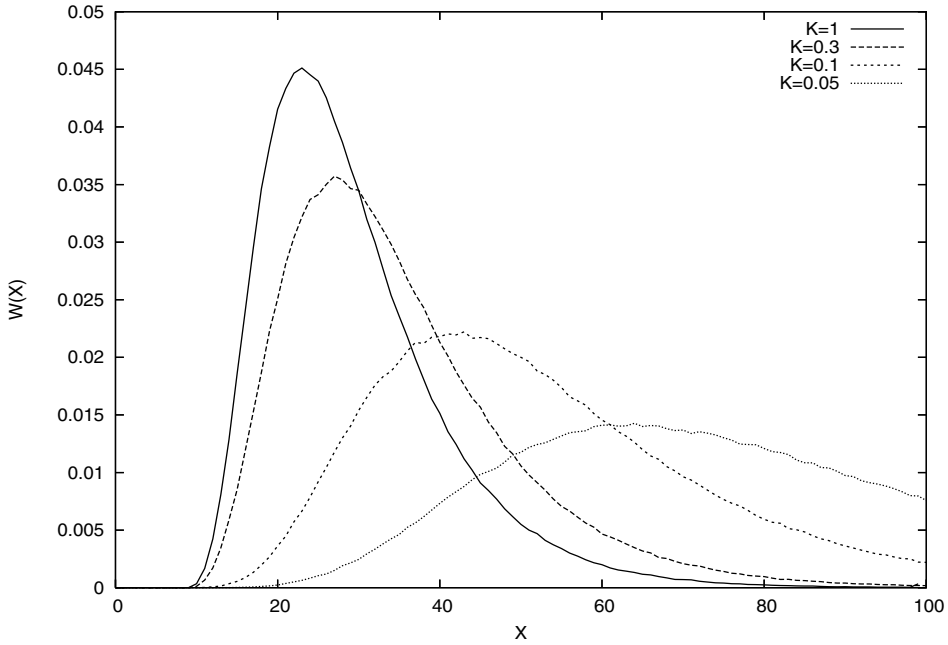

図4：稀少パラメータ K を変化させた時の $W(X)$ の分布。すべて，$N = 10$ の場合。

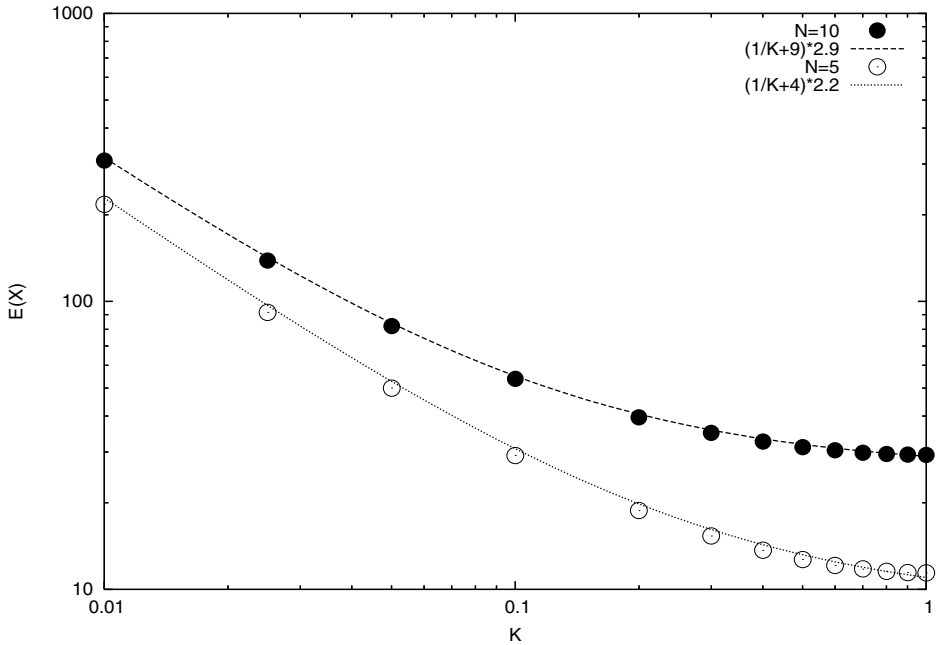

図5：稀少パラメータ K に対する期待値 $E(X)$。$N = 10$ の場合と $N = 5$ の場合。それぞれについて，近似曲線を重ねてある。横軸・縦軸とも対数目盛である。

4 考察

以上の結果について考察してみよう。まず，稀少度をあらわすパラメータ K が1以上の場合についてである。K が大きくなると稀少な Coupon c_1 の出現確率は減少する。その結果，比較的早く c_1 を引いた「幸運な」グループと，なかなか c_1 を引けない「不幸な」グループとに分かれる。この分割を $W(X)$ の最頻値を与える X_* を基準にしておこなえば，

- $N \leq X \leq X_*$ですべての Coupon がそろう「幸運な」グループ，
- $X_* < X$でないとそろわない「不幸な」グループ，

となる。シミュレーション結果が示していることは，K が増大するにつれ，両グループを分ける境界 X_* は変化しないものの，「幸運な」グループの比率が減少し，「不幸な」グループの比率が増大するということである。その結果，期待値 $E(X)$ は K とともに増大する。

一方，$K<1$ の場合は，K が減少するにつれて，分布 $W(X)$ の最頻値を与える X_* が大きくなっていく。これは「不幸な」グループの比率が増大するというよりはむしろ，「幸運な」グループにもより大きなコストを強いるようになることを意味する。ある Coupon c_1 を頻出させるということは，相対的にその他のすべての Coupon を稀少にするということになり，Coupon を集めることが難しくなるわけだ。

さて，はじめに述べた「ガチャガチャ」などでのアイテムの収集問題に戻ろう。シミュレーションからは，$K=1$ のとき，すなわち，各アイテムが等確率で出現する場合が，すべてのアイテムを収集するのにかかる平均コストが最も小さくなることが分かった。そして，稀少なアイテムが存在すれば，それが稀少（$K \gg 1$）であればあるほど，収集にかかる平均コストは増大する。アイテムを提供する側はアイテムをどの程度稀少にするかを決定することが出来るが，稀少度に関する情報は収集する側には知らされないことが多い。その場合，コレクタはどうすればよいだろうか。

たとえば，$K \geq 1$ のときの $W(X)$ の最頻値 X_* がほとんど変化しないことを利用して，収集を続けるか停止するかの1つの判断基準とすることが可能であろう。すなわち，

- アイテムの種類 N に対して，すべてのアイテムが等確率で出現する場合の理論式(3)から X_* を算出する。
- X_* 回を目処にアイテムの収集を続ける。
- 不幸にして X_* 回までにすべてのアイテムが集まらなければ，そこで収集を止める。

実際には，K の値が増大するにつれ，X_* 回までにアイテムが集まる可能性は低くなるが，K についての情報を持たない者が設定できる1つの基準として利用できそうだ。もちろん，そのアイテムの収集に非常に執着していてどれだけコストをかけても構わない人は，集まるまで続ければよいのだが。

一方，本論で考えた $K<1$ の場合は，インターネット上のファイル交換ネットワークにおける問題と関連づけることができる。いわゆる Winny に代表されるような Peer-to-Peer ファイル交換ネットワークにおいては，著作権侵害，情報漏洩などが問題となっている。その対策として，いくつかの企業（たとえば，[5] MediaDefender，[6] MediaSentry など）は次のようなサービスを提供している(2)

- 守りたいファイル（情報）に対して，偽物を作成。

---

(2) 大量の偽情報によって特定の情報を守る手法は“Poisoning”と呼ばれる。本文であげた「MediaDefender」や「MediaSentry」の他にも，「Overpeer」という米国の企業が2002年の半ばからサービスを開始したが，2005年末に会社は閉鎖された。日本でも「Whizzy R&D」という企業が「コンテンツシェルタ」というサービス名で2004年末からサービスを開始したものの，2007年9月現在，同社のWebサイトは閉鎖されている。この手法を用いたビジネスは現実的には苦戦しているようである。

- その偽物をファイル交換ネットワークに大量に放流する。
- その結果，ファイル交換を行うユーザがあるファイルをダウンロードしようとすると，偽物のファイルをつかまされる可能性が高くなり，なかなか本物のファイルに当たらなくなる。
- 結局，ユーザはファイル交換をあきらめ，ファイル（情報）が守られる。

この方法は，まさに本論でシミュレートした $K<1$ の場合に相当する。大量の偽ファイル（頻出Coupon）によって，その他のファイルの入手を困難にするというしくみだ。$K<1$ の場合は，比較的「幸運な」人にもより大きなコストを強いる結果になっていたことを想起しよう。

筆者は，2007年度から「生活の中の数学」という科目を担当することになった。本論で考察したことは，実はその授業で取り上げる話題として考え始めたものである。筆者の授業を受講してくれている学生諸氏にこの場をかりて感謝したい。

【参考文献】

[1] Blom, G., Holst, L., & Sandell, D.,“Problems and Snapshots from the World of Probability”, Sec. 7.5, Springer-Verlag, 1994

[2] Feller, W.,“An Introduction to Probability Theory and Its Applications”(vol. 1, 3rd ed., rev.), Sec. II-11, Sec. IV-2, John Wiley & Sons, 1968

[3] Baum, L. E., Billingsley, P.,“Asymptotic Distributions for the Coupon Collector's Problem”, Annals of Mathematical Statistics, vol.36, pp.1835-1839, 1965

[4] Dawkins, B.,“Siobhan's Problem: The Coupon Collector Revisited”, The American Statistician, vol.45, pp.76-82, 1991

[5] MediaDefender, Inc., http://www.mediadefender.com/

[6] MediaSentry, Inc., http://www.mediasentry.com/

付録A　古典的 Coupon Collector 問題の解析

ここでは，主に［2］Feller (1968)にしたがって，古典的 Coupon Collector 問題を解析的に取り扱う。

r 個の玉を N 個の箱に入れる問題を考えよう。その入れ方は N^r 通りある。そのうち，A_i（$i=1, 2, \cdots, N$）を i 番目の箱が「空」となる事象とすると，少なくともどの箱かが空になる事象は，

$$A = A_1 \cup A_2 \cup \cdots \cup A_N, \tag{A-1}$$

である。そして，その確率は，

$$\begin{aligned} P(A) = {} & P(A_1) + P(A_2) + \cdots + P(A_N) \\ & - P(A_1 \cap A_2) - P(A_1 \cap A_3) - \cdots \\ & + P(A_1 \cap A_2 \cap A_3) + \cdots \\ & - P(A_1 \cap A_2 \cap A_3 \cap A_4) - \cdots \\ & + - \cdots, \end{aligned} \tag{A-2}$$

となる。表記を簡単にするため，1つの箱が空になる確率を S_1，2つの箱が空になる確率を S_2 などとすると，これは，

$$P(A) = S_1 - S_2 + S_3 - S_4 + - \cdots \pm S_N, \tag{A-3}$$

と表現できる。$j \leq N$ に対して，

$$S_j = {}_N\mathrm{C}_j \left(1 - \frac{j}{N}\right)^r, \tag{A-4}$$

であるから，すべての箱が空でない確率は，

$$P_0(r, N) = 1 - P(A) = \sum_{j=0}^{N} (-1)^j {}_N\mathrm{C}_j \left(1 - \frac{j}{N}\right)^r, \tag{A-5}$$

となる。この式を用いて，「ちょうど1個の箱が空になる」確率 $P_1(r, N)$ を求める。r 個の玉を $N-1$ 個の箱に，どの箱も空でないように入れる方法は，$(N-1)^r P_0(r, N-1)$ だけあるから，

$$\begin{aligned} P_1(r, N) &= {}_N\mathrm{C}_1 \left(1 - \frac{1}{N}\right)^r P_0(r, N-1) \\ &= N \sum_{j=0}^{N-1} (-1)^j {}_{N-1}\mathrm{C}_j \left(1 - \frac{1+j}{N}\right)^r, \end{aligned} \tag{A-6}$$

と表せる。

以上より，$X-1$ 回目までに，ちょうど1つの Coupon を除いて，その他はすべて集まっている確率 $P_1(X-1,N)$ が得られる。X 回目に残りの1つの Coupon を得られる（確率は $1/N$）とすれば，$W(X)$ の理論式(3)が導かれる。

次に，式(4)を導こう。X_i を $i-1$ 個の Coupon が集まっているときに新たに別の Coupon を得るのに必要な回数とすると，

$$X = \sum_{i=1}^{N} X_i, \tag{A-7}$$

と表せる。新しい Coupon が見つかる確率は，

$$\rho_i = 1 - \frac{i-1}{N}, \tag{A-8}$$

であり，確率変数 X_i はパラメータ ρ_i の幾何確率分布となる。その期待値は，

$$\begin{aligned} E(X_i) &= \sum_{k=1}^{\infty} k(1-\rho_i)^{k-1}\rho_i = \sum_{k=1}^{\infty}\sum_{j=1}^{k}(1-\rho_i)^{k-1}\rho_i = \sum_{j=1}^{\infty}\sum_{k=j}^{\infty}(1-\rho_i)^{k-1}\rho_i \\ &= \sum_{j=1}^{\infty}(1-\rho_i)^{j-1} = \frac{1}{\rho_i}, \end{aligned} \tag{A-9}$$

である。ここで，無限等比級数の和の公式を使用している。

X_i（$i = 1, 2, \cdots, N$）はそれぞれ独立な確率変数であるから，確率変数 X の期待値は，

$$\begin{aligned} E(X) &= E\left(\sum_{i=1}^{N} X_i\right) = \sum_{i=1}^{N} E(X_i) = \sum_{i=1}^{N} \frac{1}{\rho_i} = \sum_{i=1}^{N} \frac{N}{N-i+1} \\ &= N \sum_{i=1}^{N} \frac{1}{i}, \end{aligned} \tag{A-10}$$

である。これで(4)式が導かれた。関数 $1/x$ は単調減少であるから，

$$\sum_{i=2}^{N} \frac{1}{i} \leq \int_{1}^{N} \frac{dx}{x} \leq \sum_{i=1}^{N} \frac{1}{i}, \tag{A-11}$$

より，$E(X)$ は，

$$N \ln N \leq E(X) \leq N(\ln N + 1), \tag{A-12}$$

と評価される。