5-2010

# General coupon collecting models and multinomial games

James Y. Lee
*University of Nevada Las Vegas*

GENERAL COUPON COLLECTING MODELS

AND MULTINOMIAL GAMES

by

James Y. Lee

Bachelor of Science
University of California, Los Angeles
2003

Master of Science
Creighton University
2005

A thesis submitted in partial fulfillment
of the requirements for the

**Master of Science in Mathematical Sciences**
**Department of Mathematical Sciences**
**College of Sciences**

**Graduate College**
**University of Nevada, Las Vegas**
**May 2010**

THE GRADUATE COLLEGE

We recommend the thesis prepared under our supervision by

**James Y. Lee**

entitled

**General Coupon Collecting Models and Multinomial Games**

be accepted in partial fulfillment of the requirements for the degree of

**Master of Science in Mathematical Sciences**

Hokwon Cho, Committee Chair

Malwane Ananda, Committee Member

Sandra Catlin, Committee Member

Evangelos Yfantis, Graduate Faculty Representative

Ronald Smith, Ph. D., Vice President for Research and Graduate Studies
and Dean of the Graduate College

**May 2010**

ABSTRACT

**General Coupon Collecting Models and Multinomial Games**

by

James Y. Lee

Hokwon Cho, Ph.D., Examination Committee Chair
Associate Professor of Mathematical Sciences
University of Nevada, Las Vegas

The coupon collection problem is one of the most studied problems in statistics. It is the problem of collecting $r$ ($r<\infty$) distinct coupons one by one from $k$ different kinds ($k<\infty$) of coupons. We note that this is equivalent to the classical occupancy problem which involves the random allocation of $r$ distinct balls into $k$ distinct cells. Although the problem was first introduced centuries ago, it is still actively investigated today. Perhaps its greatest feature is its versatility, numerous approaches, and countless variations. For this reason, we are particularly interested in creating a classification system for the many generalizations of the coupon collection problem. In this thesis, we will introduce models that will be able to categorize these generalizations. In addition, we calculate the waiting time for the models under consideration. Our approach is to use the Dirichlet Type II integral. We compare our calculations to the ones obtained through Monte Carlo simulation. Our results will show that our models and the method used to find the waiting times are ideal for solving problems of this type.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

# ACKNOWLEDGEMENTS

I would like to thank everyone who has helped me during this journey. I especially would like to thank the UNLV Department of Mathematical Sciences for giving me this opportunity to receive my degree. I am extremely grateful for the guidance and supervision of my advisor, Dr. Hokwon Cho. You have pushed me and taught me how to become a better person. I couldn’t have finished without your support. In addition, I would like to extend my gratitude to the various faculty from whom I have learned a great deal, especially the members of my thesis committee. Lastly, I would like to thank my family for never giving up on me and staying with me to see the finish. Special thanks to my wife Jinhee for always believing in me.

# CHAPTER 1

# INTRODUCTION

## 1.1 Motivation of the Problem

The coupon collecting problem is one of the most well known problems among probability and statistics. It has been studied extensively ever since it was first formulated by many mathematicians and statisticians (Hald, 1984). It is still actively studied. We assume that there are $k$ ($<\infty$) distinct coupons to collect and the probability of collecting a coupon of type $i$ ($i = 0,1,\ldots,k$) is non-zero, and coupons are obtained one at a time. We are interested in the waiting time that represents the number of coupons until we have collected one of each. Generalizations of this problem include collecting a subset, collecting at least two of each coupon, and many others. The goal of this thesis is to classify or model these generalizations into more detailed and appropriate categories. In addition, we will find the expected waiting time to collect the coupons for each model we introduce.

The coupon collection problem can be explained via a multinomial distribution. In general, the waiting time of a sequential decision problem such as the coupon collection problem can be found using the incomplete Dirichlet integrals. We will use Monte Carlo simulation and compare these to the expected waiting time.

## 1.2 Assumptions and Definitions

Suppose that there are $k$ ($<\infty$) distinct types of coupons to collect. Denote the probability of collecting a coupon of type $i$ by $p_i$, where $\sum_{i=1}^{k} p_i = 1.$ Then a complete set refers to all $k$ distinct objects in the set. A subset is any part of the complete set. A singleton is an object that appears once and only once in the set. For example, if we roll a fair six-sided die, then the best case scenario would be to see all six faces of the die exactly once. Any extra object beyond the complete set is referred to as a surplus.

## 1.3 Some Important Probability Distributions

In this section, we introduce some important probability distributions.

**Definition 1.3.1 (Binomial Distribution)** A random variable *X* is said to have a binomial probability distribution with parameters (*n*,*p*) if the probability mass function is given by:

$$P(X=m)=\begin{cases}\binom{n}{m}p^m(1-p)^{n-m}, m=0,1,\ldots n\\ 0,\ \text{otherwise.}\end{cases} \tag{1.1}$$

We denote this by $X\sim Bin(n,p),$ and

$$E(X)=np,\ \text{and}\ Var(X)=np(1-p).$$

**Definition 1.3.2 (Multinomial Distribution)** The Multinomial distribution is a generalization of the binomial distribution. Consider *n* independent trials $(n<\infty)$, where each trial results in one of *k* mutually exclusive outcomes, and each outcome has a probability $p_i$ of occurring, where $\sum_{i=1}^{k}p_i=1$ and $0\le p_i\le 1,\ i=1,\ldots,k.$ Let $Y_{i,n}$ be the number of outcomes falling in cell *i* ($1\le i\le k$) after *n* observations. It follows that

$0 \leq Y_{i,n} \leq n$ and $\sum_{i=1}^{k} Y_{i,n} = n.$ Then a random vector **Y** is said to have the multinomial probability mass function

$$P(Y_1 = y_1, \ldots Y_k = y_k) = \frac{n!}{y_1!y_2!\ldots y_k!} p_1^{x_1} p_2^{x_2} \ldots p_k^{x_k}. \tag{1.2}$$

with parameters $n$ and $\mathbf{p} = (p_1, \ldots, p_k)$.

**Definition 1.3.3 (Beta Distribution)** The Beta distribution is a continuous distribution with the probability distribution function

$$f(x \mid \alpha, \beta) = \begin{cases} \dfrac{1}{B(\alpha, \beta)} x^{\alpha-1}(1-x)^{\beta-1}, \ 0 < x < 1, \ \alpha > 0, \ \beta > 0 \\ 0, \ \text{elsewhere}, \end{cases} \tag{1.3}$$

where $B(\alpha, \beta)$ represents the beta function,

$$B(\alpha, \beta) = \int_0^1 x^{\alpha-1}(1-x)^{\beta-1} dx,$$

and

$$E[X] = \frac{\alpha}{\alpha+\beta}, \text{ and } Var[X] = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}.$$

The Beta distribution is closely related to the binomial distribution. The role of the random variable is reversed in the binomial and the beta distribution. Second, consider a problem of the following: let $X \sim Bin(n, p)$ and we wish to calculate $P(X < m)$ or $P(X \geq m)$, such that we have

$$P(X \geq m) = 1 - P(X < m) = \sum_{k=m}^{n} \binom{n}{m} p^{m} (1-p)^{n-m}. \tag{1.4}$$

For large values of *m* and *n*, the computation is difficult to perform. However, note that we can do the following:

$$\sum_{k=m}^{n} \binom{n}{m} p^{k} (1-p)^{n-k} = \frac{n!}{(m-1)!(n-m)!} \int_{0}^{p} x^{m-1} (1-x)^{n-m} dx. \tag{1.5}$$

Thus the binomial pdf can be calculated using the beta function. Note that replacing *p* with *x* and *n*,*n*-*m* with *α*-1, *β*-1 is the Bin (*n*,*p*) distribution. That is, Beta (*α*-1,*β*-1) is equivalent to Bin (*n*,*p*).

**Definition 1.3.4 (Dirichlet Distribution)** Let $\nu_1, \nu_2, ..., \nu_{k+1} > 0$. A random vector $\boldsymbol{X} = (X_1, X_2, ..., X_k)$ is said to have a Dirichlet probability distribution with parameters ( $\nu_1, \nu_2, ..., \nu_{k+1}$ ) if the joint probability distribution function of **X** is given by:

$$f_x(x_1,\dots,x_k)=\begin{cases}\dfrac{\Gamma(v_1+\dots+v_{k+1})}{\Gamma(v_1)\cdots\Gamma(v_{k+1})}x_1^{v_1-1}\cdots x_k^{v_k-1}(1-\sum_{i=1}^{k}x_i)^{v_{k+1}-1}, v_i>0,\\ 0, \text{elsewhere},\end{cases} \tag{1.6}$$

We denote the Dirichlet distribution by $D(\nu_1,\nu_2,...;\nu_{k+1})$. When $k$ = 1, Eq. (1.6) reduces to $D(v_1,v_2)$, which is the beta $(\nu_1,\nu_2)$ distribution. Hence the Dirichlet distribution is a multivariate generalization of the beta distribution. It is interesting to note that the Dirichlet distribution is the conjugate prior of the multinomial distribution in Bayesian setting.

The incomplete Dirichlet integrals of type I and type II are generalizations of the incomplete beta function (in 1-dimension):

$$I_p(r,s)=\frac{\Gamma(r+s)}{\Gamma(r)\Gamma(s)}\int_0^p x^{r-1}(1-x)^{s-1}dx=\frac{\Gamma(r+s)}{\Gamma(r)\Gamma(s)}\int_0^{p/q}\frac{y^{r-1}dy}{(1+y)^{r+s}}, \tag{1.7}$$

where $0\le p\le 1$ and $q$ = 1- $p$.

For $b\ge 2$ dimensions, the first integral in Eq. (1.7) is called the Dirichlet integral of Type 1. It is defined by the following $b$-dimensional integral:

$$I_p^{(b)}(r,n)=\frac{\Gamma(n+1)}{\Gamma^{(b)}(r)\Gamma(n+1-br)}\int_0^p\dots\int_0^p(1-x_1-\dots-x_b)^{n-br}\prod_{i=1}^{b}x_i^{r-1}dx_i\,, \tag{1.8}$$

where $0 \le p \le 1/b,\ \ n \ge rb$. We can see that when $b = 1$, we arrive at the first integral in Eq. (1.7). When $b = 2$ , we can use the Type I integral to sum either tail of the binomial or negative binomial distribution. For more than two it gives rise as special case to the distribution of the minimum frequency in the multinomial.

The second integral in Eq. (1.7) can also be generalized for $b \ge 2$ dimensions. These are the Dirichlet integrals of Type II. The C integral is the lower –tail form and is given by

$$C_a^{(b)}(r,m) = \frac{\Gamma(m+br)}{\Gamma^b(r)\Gamma(m)} \int_0^a \ldots \int_0^a \frac{\prod_{i=1}^{b} x_i^{r-1} dx_i}{(1+x_1+\ldots+x_b)^{m+br}}, \tag{1.9}$$

where $a \ge 0$, $b$ is an integer and $m$, $b$, and $r$ are all positive. The D integral is the upper-tail form and is given by

$$D_a^{(b)}(r,m) = \frac{\Gamma(m+br)}{\Gamma^b(r)\Gamma(m)} \int_a^\infty \ldots \int_a^\infty \frac{\prod_{i=1}^{b} x_i^{r-1} dx_i}{(1+x_1+\ldots+x_b)^{m+br}}. \tag{1.10}$$

Note that in both cases if we set $b = 1$ then the C and D integrals are reduced to the second integral in Eq. (1.7). In two dimensions, the C and D functions can be used for the tail of the negative binomial distribution, and for $b \geq 2$ they represent the tails of of the negative multinomial distribution. They can be applied to the area of ranking and selection problems. They can also be applied to finding probabilities associated with counting cell problems. The Type I and Type II integrals can be used to calculate the waiting time for counting cell problems. The Type I integral would be used when the number of cells is fixed. The Type II integral would be used in a sequential sampling scheme where the number of cells is not fixed in advance. For the coupon collection problem and for this thesis, we are dealing only with the Type II integral. For more applications using the Dirichlet integrals, see Sobel, Uppuluri, and Frankowski, 1985.

# CHAPTER 2

# COUPON COLLECTING MODELS

## 2.1 The Coupon Collection Problem

The coupon collection problem was essentially first seen in literature in 1708 when introduced by the French mathematician De Moivre (Hald, 1984). Since then, the coupon collection problem has gone by many different names. One might have seen it as the occupancy or urn problem, the random allocation problem, or the birthday problem (Holst, 1986). The basis of these problems is the same. Suppose an urn contains $r$ different balls, and balls are drawn with replacement until $k$ balls have been drawn at least $m$ times each. Let $n$ equal the number of balls drawn. The coupon collection problem deals with the number of balls drawn until $k$ different balls have been drawn. The occupancy problem finds the number of balls drawn after $n$ draws have been made. The random allocation of cells involves placing $n$ balls independently into $r$ cells. The birthday problem seeks to find how many people are needed to get a duplication of birthdays (letting $r$ equal 365 and the balls as days of the year). One can see that these problems are equivalent! Many attempts have been made to calculate the waiting times using different methods. Feller (1950) showed that simple combinatorics could be used to solve the birthday problem. Johnson and Kotz (1977) calculated the waiting time for the coupon collection problem applying Stirling numbers of the second. Kolchin, et al (1978) used generating functions to find the waiting time. Holst (1986) took the approach of using the Poisson process to calculate the waiting time. He also introduced asymptotic results in the same paper. These results form the classical coupon collection problem, where coupons are collected one at a time, and the probability of collecting a coupon of any type is uniform. There are, however, many other variations of the problem. Von Schelling (1934) calculated the waiting time when the probability of collecting each coupon was not uniform. Norman and Shepp (1960) calculated the waiting time to collect two complete sets of coupons. Stadje (1990) found the waiting time when collecting multiple coupons at once. More recently, Chang and Ross (2007) showed that the Poisson Process could be used to determine the waiting time for collecting multiple subsets of coupons. May (2008) used generating functions to find the waiting time for collecting quotas of coupons.

## 2.2 Formulation of Basic Models

We can see that there are many variations of the coupon collection problem. What if we were to establish some sort of order to the problem? Imagine a system where we can classify any statistical game and/or process into one of the models. This would simplify the problem greatly and set up a common structure that can be used by any interested individuals. This thesis is the first attempt to formally organize the coupon collection problem into its various. In addition, we will use one singular approach the find the expected waiting time. This will further simplify the problem. Our approach is to use the Dirichlet distributions. First, we investigate the cell configuration in our models. These will provide the basic structure of our models.

### 2.2.1 Cell Configuration in a Multinomial Model

A multinomial model involves collecting $r$ coupons from a set of $k$ coupons. Suppose we assume that all $k$ coupons have the same probability of being selected, $p_i = 1/k$, $i = 1, 2, \ldots, k$. This is referred to as the Equal Probable Configuration, or EPC (Cho, 2003). For example, consider a problem of tossing a fair six-sided die. One wishes to know the expected number of rolls until all six faces of the die appear at least once. This is equivalent to the classic coupon collection problem! Note that the die is rolled only once each time, and that the probability of rolling any single number is the same for all numbers. Figure 2.1 gives the cell configuration and probabilities for this scenario.

$$\underbrace{\boxed{1/6}\quad\boxed{1/6}\quad\boxed{1/6}\quad\boxed{1/6}\quad\boxed{1/6}\quad\boxed{1/6}}_{\text{6 cells}}$$

**Figure 2.1**. Cell structure and probabilities for 6 cells under the EPC.

For the general case under the EPC with *n* coupons, the probability of obtaining a couple of type *i*, $p_i$, where $i = 1, 2, \ldots$ k, is $1/k$. Figure 2.2 gives the cell configuration and probabilities for *k* cells under the EPC.

$$\underbrace{\boxed{1/\mathrm{k}}\quad\boxed{1/\mathrm{k}}\quad\boxed{1/\mathrm{k}}\quad\boxed{1/\mathrm{k}}\ \ldots\ \boxed{1/\mathrm{k}}}_{k\text{ cells}}$$

**Figure 2.2** Cell structure and probabilities under the EPC for *k* cells.

The second configuration is the Single Slippage Configuration, or SSC (Cho, 2010). Consider the ordered cell probabilities $p_{[1]} < p_{[2]} = \ldots = p_{[k-1]} = p_{[k]}$, where $p_{[j]}$ represents the $j$th ordered cell probability, $j = 1, 2, \ldots, k$. Slippage occurs when the cell probabilities are not all equally likely. We denote the slippage by $\Delta$, where $p_{[1]} = p_{[j]} - \Delta$, $j = 1, 2, \ldots, k$. Hence in the single slippage model, one of the cell probabilities is different from all the others. The difference in probability between this smallest cell and all other cells is defined as slippage. It should be that a coupon with the smallest probability, $p_{[1]}$, plays a major role to determine the waiting time N. We will show that as $p_{[1]}$ decreases, N increases

For example, consider a loaded 6-side die, where the probability of rolling one of the numbers, say 1, is $\varepsilon$, while the probability of rolling any of the remaining five numbers is the same. Figure 2.3 gives the cell configuration and probabilities for single slippage for $k = 6$ with one cell, where $\varepsilon = 1/10$.

$$\underbrace{\boxed{9/50}\quad\boxed{9/50}\quad\boxed{9/50}\quad\boxed{9/50}\quad\boxed{9/50}}_{\text{5 cells}}\quad\underbrace{\boxed{1/10}}_{\text{1 cell}}$$

**Figure 2.3** Single slippage configuration with one cell for $k = 6$, $\Delta = 4/50$.

In general, the SSC for $k$ +1 cells has $p_{[1]} = \varepsilon$ ( > 0), with all other cells having equal probability of being selected. Figure 2.4 gives the cell configuration and probabilities for single slippage with one cell under general conditions.

$$\underbrace{\boxed{(1-\varepsilon)/k}\quad\boxed{(1-\varepsilon)/k}\quad\boxed{(1-\varepsilon)/k}\quad \ldots \boxed{(1-\varepsilon)/k}}_{k \text{ cells}}\quad\underbrace{\boxed{\quad\varepsilon\quad}}_{1 \text{ cell}}$$

**Figure 2.4** Single slippage configuration with one cell, $\Delta = [1 - \varepsilon(1+k)] / k$.

Since the smallest cell probability, $p_{[1]}$, is less than or equal to the probability of any other cell, it must be that when $k \leq 10$, $\Delta$ must be less than 1/10. If $\Delta = 1/10$ then the configuration is the EPC. Similarly, when $k \leq 20$, $\Delta$ must be less than 1/20, etc…

It may also be that more than one cell shares this slippage property. It is still considered single slippage. Figure 2.5 gives the cell configuration and probabilities when $k = 6$.

$$\underbrace{\boxed{1/5}\quad\boxed{1/5}\quad\boxed{1/5}\quad\boxed{1/5}}_{4 \text{ cells}}\quad\underbrace{\boxed{1/10}\quad\boxed{1/10}}_{2 \text{ cells}}$$

**Figure 2.5** Single slippage with two cells ($k = 6$, $\Delta = 1/10$).

Finally, the third configuration is the Multiple Slippage Configuration, or MSC (Cho, 2010). In this scenario, there is more than one slippage in the configuration. Consider the ordered cell probabilities $p_{[1]} =\ldots= p_{[j]} < p_{[j+1]} = \ldots\ p_{[l]} < p_{[l+1]} =\ldots= p_{[k]}$. Note that when $k = l$, the MSC reduces to the SSC. Again, the cells that have lower slippage properties play a major role in determining the waiting time N. Figure 2.6 gives the cell configuration and probabilities for double slippage when $k$ =6.

$$\underbrace{\boxed{17/80}\quad \boxed{17/80}\quad \boxed{17/80}\quad \boxed{17/80}}_{\text{4 cells}}\quad \underbrace{\boxed{1/10}}_{\text{1 cell}}\quad \underbrace{\boxed{1/20}}_{\text{1 cell}}$$

**Figure 2.6** Double slippage with one cell each; $\Delta_1 = 1/20$, $\Delta_2 = 9/80$.

It is generally the case that in the MSC, more than one cell shares the slippage property. Consider Figure 2.7 for the case when $k = 6$.

$$\underbrace{\boxed{1/4}\quad \boxed{1/4}\quad \boxed{1/4}}_{\text{3 cells}}\ \underbrace{\boxed{1/10}\quad \boxed{1/10}}_{\text{2 cells}}\quad \underbrace{\boxed{1/20}}_{\text{1 cell}}$$

**Figure 2.7** Double Slippage under OMM for MSC; $\Delta_1 = 1/20$, $\Delta_2 = 3/20$.

### 2.2.2 One Multinomial Model

We now introduce the One Multinomial Model, or OMM, where coupons are drawn from a single set. For example, collecting a set of baseball cards, or cards from a deck, or game pieces of a popular board game. In this model, the configuration of the cell probabilities of obtaining each coupon in the set will come from either the EPC, SSC, or MSC.

### 2.2.3 Compound Multinomial Model

Consider a scenario where coupons are collected from different sets. We call this the Compound Multinomial Model, or CMM. For example, suppose we are tossing two dice simultaneously and we are interested in seeing the face numbers at least once regardless of which die it is on. Similarly, suppose we are collecting football cards and baseball cards. The setup is as follows. Supposed we have $k$ distinguishable types of coupons, where P(coupon of type $i$) = $p_i$, i = 1,2,…,$k$, and $\sum_{i=1}^{k} p_i = 1$, and more generally, $\sum_{i=1}^{k} p_i \leq 1$.

Let $C_i$ denote the number of coupons of type $i$, $i$ = 1,2,…,$k$, ($k < \infty$). Let $N = \sum_{i=1}^{k} C_i$ be the total number of distinct coupons among the $k$ types, which may include types not being collected. Then E(WT) = E($WT_1$) + E($WT_2$).

### 2.2.4 Customer's Choice Model

The Multinomial Models mentioned previously share the idea that one complete set must be collected. However, it may not always be desirable to collect a complete set. Perhaps we want to collect the basketball cards of only our favorite players. We call this subset selection. Or maybe we would like to collect two complete sets of coupons. Also consider the scenario where we only want one coupon in the whole set. For this reason, we call this category the Customer Choice Model. This is more of an umbrella model as it covers scenarios where we are not collecting complete sets.

A more flexible way to represent the expected waiting time under the CCM is by $E(WT|S_i)$, where $S_i$ is a subset of the set of coupons S, $i = 1,2,\ldots, k$. Consider the following; suppose one wants to see all sides of each of several type of dice, and we assume that we have one of each type. However, it may be that one can be tossing them separately, and the player decides which one to toss. Or, it may be that the player tosses one of each type simultaneously. Also consider the scenario where the player tosses the dice separately with a random mechanism for selecting the type of die to be tossed. The expected waiting time would fall under the CCM. This type of problem can be related to various waiting time problems where the die represents different sets of coupons and the faces represent the coupons to be collected.

Similarly, consider the problem of observing the faces of die regardless of which die it is on; here we toss the dice simultaneously so that we can observe two more different numbers (i.e. faces of the die) in one toss.

# CHAPTER 3

# EXPECTED WAITING TIMES

In this chapter, we calculate the expected waiting times for each model under consideration. We determine how many coupons we need to collect in order to have a complete set. It is important to note that in our models the coupons are collected one at a time. Since we do not know in advance how many we need to collect, we continue until we have achieved our goal. This is called an inverse type sequential sampling scheme. Hence we can use the Dirichlet Type II integral to calculate the expected waiting times. In the scenario where the number of coupons to be collected are fixed, we use the Dirichlet Type I integral.

## 3.1 E[WT] for the Classic Coupon Collection Problem

In the classic case, there are $k$ coupons to collect, each with equal probability of being selected. From elementary probability theory, the expected waiting time E(WT) is given by

$$\text{E(WT)} = k(1+\frac{1}{2}+\frac{1}{3}+\ldots+\frac{1}{k}) \ , \tag{3.1}$$

and can also be approximated by

$$\text{E(WT)} \approx k(\log\ k+\kappa), \tag{3.2}$$

where $\kappa = 0.57721566$ is Euler's constant.

We can also make the best use of the Dirichlet Type II C-Integral to arrive at the same answer. The beta distribution can be used to calculate the lower and upper tail probabilities of the binomial distribution. In the same way, the Dirichlet Type II integral can be used to calculate the lower and upper tail of the multinomial distribution. In fact, the C integral, which we use here, is used to calculate the probability that the last coupon reaches its quota. Using the $r^{\text{th}}$ ascending factorial moment given in Sobel et al (1977), we can obtain the first two moments of the waiting time, E(WT) and E($\text{WT}^2$).

$$\mu^{[\gamma]} = \frac{b\Gamma(r+\gamma)}{\Gamma(r)\,p^{\gamma}} C_a^{(b-1)}(r, r+\gamma), \tag{3.3}$$

where $b$ is the number of cells, $r$ is the common quota, $a$ = 1, and the C-integral is the same as introduced before. The first ascending factorial moment when $\gamma$ = 1 is the mean $\mu$, and the variance can be obtained from the relation

$$\sigma^2 = \mu^{[2]} - \mu(\mu+1). \tag{3.4}$$

For example, under the Equal Probable Configuration with $b$ = 6, $\gamma$ = 1, $a$ =1 and $r$ =1, the first factorial moment (and hence also the expected waiting time) becomes

$$\mu^{[1]} = \frac{6 \bullet \Gamma(2)}{\Gamma(1) \bullet \frac{1}{6}} \bullet \frac{\Gamma(7)}{\Gamma^5(1)\Gamma(2)} \bullet \int_0^1\int_0^1\int_0^1\int_0^1\int_0^1 \frac{dx_1 dx_2 dx_3 dx_4 dx_5}{(1 + x_1 + x_2 + x_3 + x_4 + x_5)^7}$$

The result is 14.6998. Note that if we used Eq. (3.1) then the result is 14.7. This agrees with the expected value obtained using the Dirichlet integral! An approximation using Eq. (3.2) gives 14.2138. To calculate the variance, we first need to calculate the second ascending factorial moment.

$$\mu^{[2]} = \frac{6\Gamma(3)}{\Gamma(1)(1/6)^2} \bullet \frac{\Gamma(8)}{\Gamma^5(1)\Gamma(3)} \bullet \int_0^1\int_0^1\int_0^1\int_0^1\int_0^1 \frac{dx_1 dx_2 dx_3 dx_4 dx_5}{(1 + x_1 + x_2 + x_3 + x_4 + x_5)^8}.$$

The result is 269.78. Now we can find the variance using Eq. (3.2):

$$\sigma^2 = \mu^{[2]} - \mu^{[1]}(1 + \mu^{[1]}) = 38.99.$$

The D integral, which we do not use, can be used to calculate the expected waiting time until the first cell reaches its quota. We simply replace the C integral with the D integral in the ascending factorial.

The following table gives the waiting times for collecting one complete set of coupons, where the number of coupons in the set range from 2 to 20. Their corresponding standard deviations are also given. The cell configuration is the Equal Probable Configuration.

**Table 3.1**: Expected Waiting Time for OMM under EPC

| # of coupons | Cell Configuration | E(WT) | Stdev |
|---|---|---|---|
| 2 | (1/2,1/2) | 3.0000 | 1.4142 |
| 3 | (1/3,1/3,1/3) | 5.5000 | 2.4898 |
| 4 | (1/4,1/4,1/4,1/4) | 8.3331 | 3.8007 |
| 5 | (1/5,1/5,…..,1/5) | 11.4180 | 5.0161 |
| 6 | (1/6,1/6,…….,1/6) | 14.6998 | 6.2442 |
| 7 | (1/7,1/7,………,1/7) | 18.0280 | 7.6173 |
| 8 | (1/8,1/8,………..,1/8) | 21.7428 | 8.7185 |
| 9 | (1/9,1/9,………..…,1/9) | 25.4607 | 9.9629 |
| 10 | (1/10,1/10,……….,1/10) | 29.2896 | 11.2110 |
| 12 | (1/12,1/12,……..…,1/12) | 37.2385 | 13.7156 |
| 15 | (1/15,1/15,…………..1/15) | 49.7734 | 17.4878 |
| 20 | (1/20,1/20,……………,1/20) | 71.9547 | 23.8015 |

As the number of coupons increases, the expected waiting time also increases. Similarly, the standard deviation of the waiting time also increases.

## 3.2 E[WT] for Basic Models

We are interested in the waiting times for models when the probabilities are non-uniform, such as in the slippage case. We can use the $r^{\text{th}}$ ascending factorial moment again to calculate the waiting times, but it is in a different from than in the EPC case. It is as follows:

$$\mu^{[\gamma]} = \sum_{\alpha=1}^{b} \frac{\Gamma(r_\alpha + \gamma)}{\Gamma(r_\alpha)\, p_\alpha^\gamma} C_{\underset{\rightarrow}{a}_\alpha}^{(b-1)}(\underset{\rightarrow}{r}_\alpha, r_\alpha + \gamma), \tag{3.5}$$

where the C integral is as above, and

$$\underset{\rightarrow}{a}_\alpha = \left(\frac{p_1}{p_\alpha}, \ldots, \frac{p_{\alpha-1}}{p_\alpha}, \frac{p_{\alpha+1}}{p_\alpha}, \ldots, \frac{p_b}{p_\alpha}\right) \tag{3.6}$$

represents the ratio of the cell probabilities. In the EPC case, since the probability of collecting any coupon in the set is the same, *a* was simply 1 and it was not necessary to represent *a* as a vector. Now, ***a*** is a vector because the probabilities are not all uniform. In this thesis, we have calculated the single slippage case. Consider the configuration where $k = 2$, $p_{[1]} = 1/10$, and $p_{[2]} = 9/10$. The first ascending factorial moment would be of the form

$$\mu^{[1]} = \frac{\Gamma(2)}{\Gamma(1)(1/10)} \bullet \frac{\Gamma(3)}{\Gamma(1)\Gamma(2)} \bullet \int_0^9 \frac{dx_1}{(1+x_1)^3} + \frac{\Gamma(2)}{\Gamma(1)(9/10)} \bullet \frac{\Gamma(3)}{\Gamma(1)\Gamma(2)} \bullet \int_0^{1/9} \frac{dx_1}{(1+x_1)^3}.$$

The answer is 10.1110. The following table gives the expected waiting time for single slippage with one cell. When *k* is less than or equal to ten, the smallest cell probability is equal to 1/10. The slippage values are also given. It represents the difference between the smallest cell probability and the probability of collecting any of the other $k$ - 1 coupons in the set (since the probability of collecting any of the remaining $k$ – 1 coupons is uniform).

**Table 3.2**: Expected Waiting Time for OMM under SSC ($p_{[1]} = 1/10$)

| # of coupons | Cell Configuration | Slippage | E(WT) | Stdev |
|---|---|---|---|---|
| 2 | (1/10,9/10) | 4/5 | 10.1110 | 9.4106 |
| 3 | (1/10,9/20,9/20) | 7/20 | 10.6969 | 8.9966 |
| 4 | (1/10,3/10,3/10,3/10) | 1/5 | 11.8970 | 8.4725 |
| 5 | (1/10,9/40,9/40,…,9/40) | 1/8 | 13.7001 | 8.0662 |
| 6 | (1/10,9/50,9/50,..…,9/50) | 4/50 | 16.0320 | 7.9864 |
| 8 | (1/10,9/70,9/70,….…,9/70) | 2/70 | 22.0231 | 9.4094 |
| 10 | (1/10,1/10,………...…,1/10) | 0 | 29.2896 | 11.2110 |

As the number of coupons increases, the value of the slippage decreases. For $k$ = 10, there is no slippage. This configuration falls under the EPC!

What would happen if the smallest cell probability is smaller than 1/10? Table 3.3 gives the expected waiting time for single slippage, but now we see that the smallest cell probability is 1/20.

**Table 3.3**: Expected Waiting Time for OMM under SSC ($p_{[1]}$ = 1/20)

| # of coupons | Cell Configuration | Slippage | E(WT) | Stdev |
|---|---|---|---|---|
| 2 | (1/20,19/20) | 9/10 | 20.0526 | 19.4436 |
| 3 | (1/20,19/40,19/40) | 17/40 | 20.3483 | 19.2323 |
| 4 | (1/20,19/60,19/60,19/60) | 4/15 | 20.9979 | 18.7457 |
| 5 | (1/20,19/80,19/80,…,19/80) | 3/16 | 22.0415 | 18.0489 |

We observe that as the smallest cell probability decreases, the expected waiting time increases. If we are collecting two complete sets under the EPC, we use Eq. (3.1) again, but in this case $r$ = 2. The calculation becomes only a little more complicated.

**Table 3.4**: Expected Waiting Time for OMM under the EPC, 2 Complete Sets

| # of coupons | Cell Configuration | E(WT) | Stdev |
|---|---|---|---|
| 2 | (1/2,1/2) | 5.500 | 1.3426 |
| 3 | (1/3,1/3,1/3) | 9.6357 | 3.2557 |
| 4 | (1/4,1/4,1/4,1/4) | 14.1926 | 4.0767 |
| 5 | (1/5,1/5,…….,1/5) | 19.0413 | 6.1056 |
| 6 | (1/6,1/6,………,1/6) | 24.1338 | 7.5425 |
| 10 | (1/10,1/10,…….,1/10) | 46.2295 | 13.3007 |

CHAPTER 4

# SUMMARY AND CONCLUSION

## 4.1 Monte Carlo Simulation

We carry out Monte Carlo simulation to verify our models. The results of the simulation for some of our models are given in the following tables. Table 4.1 gives the results for the One Multinomial Model under the Equal Probable Configuration. The average stopping time is given by E(WT), which is the expected waiting time, and it's variability is given as the standard error, denoted by S.E. The standard error is the standard deviation divided by the number of simulations. Calculations were performed for $k$ = 2 to 20, where each row is based on 10000 experiments.

**Table 4.1**: One Multinomial Model, EPC

| # of coupons | Cell Configuration | E(WT) | S.E. |
|---|---|---|---|
| 2 | (1/2,1/2) | 3.0243 | 0.0145 |
| 3 | (1/3,1/3,1/3) | 5.4784 | 0.0261 |
| 4 | (1/4,1/4,1/4,1/4) | 8.3468 | 0.0376 |
| 5 | (1/5,1/5,…..,1/5) | 11.4323 | 0.0509 |
| 6 | (1/6,1/6,…...,1/6) | 14.7326 | 0.0634 |
| 7 | (1/7,1/7,…….,1/7) | 18.0971 | 0.0741 |
| 8 | (1/8,1/8,……...,1/8) | 21.7320 | 0.880 |
| 9 | (1/9,1/9,………,1/9) | 25.6202 | 0.0993 |
| 10 | (1/10,1/10,……,1/10) | 29.1361 | 0.1113 |

| | | | |
|---|---|---|---|
| 11 | (1/11,1/11,……,1/11) | 33.2630 | 0.1236 |
| 12 | (1/12,1/12,…….,1/12) | 37.2011 | 0.1375 |
| 13 | (1/13,1/13,……..,1/13) | 41.4696 | 0.1493 |
| 14 | (1/14,1/14,………,1/14) | 45.5014 | 0.1601 |
| 15 | (1/15,1/15,……….,1/15) | 49.9203 | 0.1800 |
| 16 | (1/16,1/16,……..…,1/16) | 54.2025 | 0.1872 |
| 17 | (1/17,1/17,…………,1/17) | 58.6064 | 0.2011 |
| 18 | (1/18,1/18,………….,1/18) | 62.6443 | 0.2073 |
| 19 | (1/19,1/19,………..…,1/19) | 67.1957 | 0.2238 |
| 20 | (1/20,1/20,……………,1/20) | 71.7870 | 0.2378 |

These values are very close to the exact values which we calculated using the Dirichlet integrals. We observe that as the number of cells increases, so too does the expected waiting time. This is important in our discussion as we have shown that our simulation provides an accurate validation of our computed results.

The following table represents the simulation results under the Single Slippage Configuration. For $k$ from 1 to 10, the smallest cell probability is 1/10. For $k$ from 11 to 20, the smallest cell probability is 1/20. The average stopping time, the slippage in the cell configuration, and standard errors are also given, where each row is based on 10000 experiments.

**Table 4.2:** One Multinomial Model, SSC

| # of coupons | Cell Configuration | Slippage | E(WT) | S.E. |
|---|---|---|---|---|
| 2 | (1/10,9/10) | 4/5 | 10.1960 | 0.0949 |
| 3 | (1/10,9/20,9/20) | 7/20 | 10.7986 | 0.0895 |
| 4 | (1/10,3/10,3/10,3/10 | 1/5 | 11.9805 | 0.0863 |
| 5 | (1/10,9/40,……,9/40) | 1/8 | 13.6515 | 0.0796 |
| 6 | (1/10,9/50,…..….,9/50) | 2/25 | 15.9404 | 0.0789 |
| 7 | (1/10,3/20,………..,3/20) | 1/20 | 18.8859 | 0.0840 |
| 8 | (1/10,9/70,…….……,9/70) | 2/70 | 22.1072 | 0.0899 |
| 9 | (1/10,9/80,……..…..…,9/80) | 1/80 | 25.63847 | 0.1014 |
| 10 | (1/10,1/10,…………...…,1/10) | 0 | 29.1685 | 0.1121 |
| 11 | (1/20,19/200,19/200,…,19/200) | 9/200 | 36.0395 | 0.1596 |
| 12 | (1/20,19/220,19/220,…,19/220) | 2/55 | 39.2519 | 0.1591 |
| 13 | (1/20,19/240,19/240,…,19/240) | 7/240 | 43.0009 | 0.1689 |
| 14 | (1/20,19/260,19/260,…,19/260) | 3/130 | 46.4897 | 0.1714 |
| 15 | (1/20,19/280,19/280,…,19/280) | 1/56 | 50.6792 | 0.1802 |
| 16 | (1/20,19/300,19/300,…,19/300) | 1/75 | 54.4838 | 0.1917 |
| 17 | (1/20,19/320,19/320,…,19/320) | 3/320 | 58.7457 | 0.2014 |
| 18 | (1/20,19/340,19/340,…,19/340) | 1/170 | 63.0875 | 0.2134 |
| 19 | (1/20,19/360,19/360,…,19/360) | 1/360 | 67.1244 | 0.2272 |
| 20 | (1/20,1/20,………………,1/20) | 0 | 72.0891 | 0.2376 |

Under this more likely scenario, as the slippage value increases, so too does the expected waiting time. In short, the waiting time depends on the smallest cell probability. Note that when $k$ = 10 and when $k$ = 20, the configuration is the Equally Likely Configuration. Another feature of the OMM under the SSC reveals that as the number of coupons increases, the slippage decreases. Thus, when we compare the expected waiting time of the SSC to the EPC, we note that the waiting times will eventually converge as the number of coupons increases.

Table 4.3 gives the results when the smallest cell probability is 1/20 for $k$ from 2 to 5. As the smallest cell probability decreases, the expected waiting time increases. Table 4.4 gives the expected waiting time under the EPC for collecting two complete sets.

**Table 4.3:** One Multinomial Model, SSC

| # of coupons | Cell Configuration | Slippage | E(WT) | S.E. |
|---|---|---|---|---|
| 2 | (1/20,19/20) | 9/10 | 19.8218 | 0.1937 |
| 3 | (1/20,19/40,19/40) | 17/40 | 20.6559 | 0.1954 |
| 4 | (1/20,19/60,…,19/60) | 4/15 | 20.9611 | 0.1897 |
| 5 | (1/20,19/80,…….,19/80) | 3/16 | 22.1296 | 0.1836 |
| 6 | (1,20,19/100,……..,19/100) | 7/50 | 23.4503 | 0.1729 |
| 8 | (1/20,19/140,……….,19/140) | 3/35 | 27.6800 | 0.1657 |
| 10 | (1/20,19/180,…………,19/180) | 1/18 | 32.8003 | 0.1571 |

**Table 4.4:** One Multinomial Model under EPC, 2 complete sets

| # of coupons | Cell Configuration | E(WT) | S.E. |
|---|---|---|---|
| 2 | (1/2,1/2) | 5.4983 | 0.0180 |
| 3 | (1/3,1/3,1/3) | 9.6342 | 0.0329 |
| 4 | (1/4,1/4,1/4,1/4) | 14.1858 | 0.0471 |
| 5 | (1/5,1/5,…..,1/5) | 19.1057 | 0.0613 |
| 6 | (1/6,1/6,……,1/6) | 24.0318 | 0.0745 |
| 7 | (1/7,1.7,…….,1.7) | 29.4078 | 0.0741 |
| 8 | (1/8,1.8,…..…,1.8) | 35.0712 | 0.1052 |
| 9 | (1/9,1/9,………,1/9) | 40.4835 | 0.1187 |
| 10 | (1/10,1/10,……,1/10) | 46.1340 | 0.1328 |
| 11 | (1/11,1/11,….…,1/11) | 52.2246 | 0.1486 |
| 12 | (1/12,1/12,…..…,1/12) | 58.1540 | 0.1629 |
| 13 | (1/13,1/13,……..,1/13) | 64.2499 | 0.1770 |
| 14 | (1/14,1/14,………,1/14) | 70.3018 | 0.1923 |
| 15 | (1/15,1/15,……….,1/15) | 76.1953 | 0.2026 |
| 16 | (1/16.1/16,…………,1/16) | 82.7703 | 0.2165 |
| 17 | (1/17,1/17,………….,1/17) | 88.7004 | 0.2294 |
| 18 | (1/18,1/18,…………..,1/18) | 95.6422 | 0.2462 |
| 19 | (1/19,1/19,…………….,1/19) | 101.8835 | 0.2587 |
| 20 | (1/20,1/20,……………..,1/20) | 108.8387 | 0.2776 |

## 4.2 Summary of Results

These results will prove useful when faced with a situation that mirrors the coupon collection problem. The method used to calculate the waiting times present another means on top of the ones already discovered. We do not wish to compare our methods; instead, we show the versatility of the coupon collection problem and its many approaches.

We can compare the EPC for the One Multinomial Model compared to the Customer Choice Model of obtaining two complete sets. For example, say one would like to collect two complete sets under the OMM as opposed to simply collecting two complete sets under the Customer Choice Model. Note that these are different problems. By the Customer Choice Model, the expected waiting time is equal to the sum of the individual waiting times. Thus E(WT) = 14.6998 + 14.6998 = 29.3996, while the expected waiting time under the Customer Choice Model is 24.0318. Why are these values different? In the Customer Choice Model the expectation of collecting two complete sets is conditional on obtaining the first set.

## 4.3 Future Study

We can apply the coupon collection problem to card games. Card games are essentially waiting time problems in that we keep playing until we have collected the winning hand. There are many variations to card games such as poker, pai gow, and baccarat, but essentially they are waiting time decisions. For future study or as an illustration, we could apply the coupon collection problem to a card game in order to find the waiting time necessary to be dealt a given hand.

We can apply decision theoretic approaches to the coupon collection problem as well. For example, suppose there is a cost associated with collecting a coupon, as in the McDonald's Monopoly game. Then we can ask, what is the optimal waiting time based on a loss function? This certainly would interest economists and basic consumer behavior. We hope that with this thesis, many more ideas and applications would be discovered.

## 4.4 Conclusion

In this thesis, we have studied various types of the coupon collection problem and suggested a classification system by introducing the slippage. Any waiting time problem could fit into one of our models. We have calculated the expected waiting times for our models using Dirichlet integrals and have given numerical evidence that our calculations are accurate by Monte Carlo simulation.

REFERENCES


[1] Balakrishnan, N. and Nevzorov, V.B. (2003). *A Primer on Statistical Distributions*. Wiley, New Jersey.

[2] Blom, G. and Holst, L. and Sandell, D. (1994). *Problems and Snapshots from the World of Probability*. Springer, New York.

[3] Casella, G. and Berger, R. (2002) *Statistical Inference*. 2nd Edition. Duxbury, Pacific Grove.

[4] Chang, K.C. and Ross, S. (2007). The Multiple Subset Coupon Collection Problem. *Probability in the Engineering and Information Sciences*, **21**, 435-440.

[5] Cho, H. (2003). Inverse-Type Sampling Procedure for Estimating the Number of Multinomial Classes. *Sequential Statistics*, **22**, 307-324

[6] Cho, H. (2010). *Testing Multinomial Categories with Sequential Sampling: Is a Wheel of Fortune Fair?* Unpublished manuscript.

[7] Feller, W. (1950) *An Introduction to Probability Theory and It's Applications.* Wiley, New York.

[8] Gibbons, J.D. and Olkin, I and Sobel, M. (1977) *Selecting and Ordering Populations*: *A New Methodology*. Wiley, New York.

[9] Hald, A. (1984) Commentary on 'De Mensura Sortis. *International Statistical Review*, **52**, 229-262.

[10] Holst, L. (1986). On Birthday, Collectors', Occupancy, and Other Classical Urn Problems. *International Statistical Review*, **54**, 15-27.

[11] Johnson, N. and Kotz, S. (1977) *Urn Models and Their Application*. Wiley, New York.

[12] Kolchin, V.F. and Sevast'yanov, B.A. and Chistayov, V.P. (1978). *Random Allocations*. V.H. Winston and Sons, Washington, D.C.

[13] Kotz, S. and Balakrishnan, N. and Johnson, N. (2000). *Continuous Multivariate Disributions.* Wiley, New York.

[14] May, R., (2008). Coupon Collecting With Quotas. *The Electronic Journal of Combinatorics*, **15**, 1-7.

[15] Maunsell, F.G. (1938). A Problem In Cartophily. *The Mathematical Gazette*. **22,** 328-331.

[16] Newman, D.J. and Shepp, L. (1960) The Double Dixie-Cup Problem. *American Mathematical Monthly*, **67**, 58-61.

[17] Smith, D.K. (2005) Decision Making and Card Collecting. *IMA Journal of Management Mathematics.* **16**, 89-97

[18] Sobel, M and Uppuluri, V.R.R. and Frankowski, K. (1977). *Selected Tables in Mathematical Statistics, Volume IV – Dirichlet Integrals of Type I.* Edited by the Institute of Mathematical Statistics. American Mathematical Society, Providence.

[19] Sobel, M. and Uppuluri, V.R.R. and Frankowski, K. (1985). *Selected Tables in Mathematical Statistics, Volume IX – Dirichlet Integrals of Type II and Their Applications.* Edited by the Institute of Mathematical Statistics. American Mathematical Society, Providence.

[20] Stadje, W. (1990) The Collector's Problem With Group Drawings, *Advanced Applied Probability*, **22**, 866-882.

[21] Von Schelling, H. (1954). Coupon Collecting for Unequal Probabilities, *American Mathematical Monthly*, **61**, 306-311.

[22] Wilks, Samuel. (1962). *Mathematical Statistics.* Wiley, New York.

VITA

Graduate College
University of Nevada, Las Vegas

James Y. Lee

Degrees:

Bachelor of Science in Applied Mathematics, 2003
University of California, Los Angeles

Master of Science in Atmospheric Sciences, 2005
Creighton University

Thesis Title: General Coupon Collection Models and Multinomial Games

Thesis Examination Committee:

Chairperson, Professor Hokwon Cho, Ph.D.
Committee Member: Professor Malwane Ananda, Ph.D.
Committee Member: Professor Sandra Catlin, Ph.D.
Graduate Faculty Representative: Professor Evangelos Yfantis, Ph.D. Computer Science