1-1-2005

Sequential procedure for test of uniformity in multinomial models

Hai Zhen
University of Nevada, Las Vegas

SEQUENTIAL PROCEDURE FOR TEST OF UNIFORMITY IN MULTINOMIAL MODELS

by

Hai Zhen

Bachelor of Science
Anhui University, Hefei China
1988

A thesis submitted in partial fulfillment
of the requirements for the

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

Graduate College
University of Nevada, Las Vegas
December 2005

UNLV
UNIVERSITY OF NEVADA LAS VEGAS

Thesis Approval

The Graduate College
University of Nevada, Las Vegas

November 17, 2005

The Thesis prepared by

Hai Zhen

Entitled

Sequential Procedure for Test of Uniformity in Multinomial Models

is approved in partial fulfillment of the requirements for the degree of

Master of Science

Examination Committee Chair

Dean of the Graduate College

Examination Committee Member

Examination Committee Member

Graduate College Faculty Representative

1017-53

ABSTRACT

Sequential Procedure for Test of Uniformity in Multinomial Models

by

Hai Zhen

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

In this thesis we deal with a sequential procedure for testing uniformity in a given multinomial distribution using inverse sampling. From a decision theoretic point of view, we devise an efficient stopping rule that satisfies a pre-determined P*-condition. Dirichlet distributions of Type II will be primarily used for developing the inverse-type sequential procedure based on the decision theoretic point of view. We assume a non-zero cell probability (parameter) for given multinomial models. In particular, we will be focusing on the equal probability configuration (EPC) among all feasible cell configurations. One of the main goals is to find optimal sample sizes that result from a desirable probability level, the probability of correct decision P{CD}, in testing uniformity in multinomial models. As an illustration, "wheel of fortune" will be considered to fit the developed model. Finally, the developed procedure will be discussed via Monte Carlo experimentation.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS

Despite of the corkscrew path of my academic pursuit and not a serene progressive arc of my thesis writing, I have been honored and humbled by the confidence that my advisor, Professor Hokwon Cho has shown to me. I would like to express my deepest gratitude to my advisor, who also has been my mentor and friend, for his warmest encouragement and patience throughout this undertaking. His excellence in both research and teaching will always be an example to me.

My thanks also go to my respectable committee members, Professors Malwane Ananda, Chih-Hsiang Ho, Sandra Catlin and Bernard Malamud for their influence during my graduate studies.

Last, but not least, I wish to thank my families for their love and support throughout my education. Without them, I would not be able to present this paper.

CHAPTER 1

INTRODUCTION

1.1 Motivation

The multinomial probability distribution is defined by specifying a several numbers of categories (or classes, or cells), say k ($< \infty$, frequently known or assumed), for outcomes, which is one of the most frequently occurring statistical phenomena in decision making problems. In those decision procedures, two most important aspects of statistical decision are error control to be minimized (or maximizing the statistical confidence) and sample sizes. For instance, when we are dealing with a testing problem (with two competing hypothesis), we wish to choose a test that guarantees the maximum power among all possible statistical tests. However, this is possible only when the (given) sample sizes are the same for all possible tests, and the power of the test depends upon the sample sizes that the researchers can obtain in the experiment. Moreover, the existing fixed-sample classical tests do not provide the optimal sample sizes for any statistical tests. Suppose for a given multinomial population we are interested in cell probabilities (p_i, $i = 1,2,..,k$) for all k categories and wish to test a hypothesis H_0: $p_i = 1/k$.

Then, our decision-making procedure must reflect the fact that an appropriate sample size ensures the statistical confidence in such testing.

In this thesis, we wish to approach the problem by taking a different sampling strategy, namely the inverse sampling procedure. In particular, in many applications where samples are costly, the inverse sampling method is frequently recommended not only to reduce costly sampling units, but also to optimize sample sizes in reaching a decision. For more details about sequential sampling schemes, see Govindarajulu (1999).
For an illustration we will use the well-known game – Wheel of Fortune, which has the number of categories ranging from 2 to 10 where the equal probability configuration (EPC) is assumed. The goal is then to find an optimal stopping time rule that satisfies the prescribed probability level P* to test of uniformity utilizing the Ranking and Selection Methodologies (RSM).

1.2 Wheel of Fortune

A feeling of adventure is an element of games. We compete against the uncertainty. The course of the game and its outcome change each time we play. The future remains in darkness. That is what keeps things entertaining and generates excitement.

The popularity of wheel of fortune proves this point over and over again. Random influences occur in games involving dice, wheel of fortune

and the mixing of cards. The course of a game, in accordance with its rules, is determined not only by the decision made by the players, but by the results of random processes. If the influence of chance dominates the decisions of the players, then we speak of a game of chance, therefore, we are always in quest of the equiprobability.

A wheel of fortune is a game by spinning the wheel that has several categories with specified outcomes such as amount of reward. A contestant usually takes the wheel for a spin and waits until it stops completely. Then, the observer declares the category (i.e. outcome) that the indicator points exactly one of possible categories.

CHAPTER 2

PRELIMIARIES

2.1. Probability Distributions

Binomial Distribution

Binomial probability mass function takes the form

$$p_m = P\{X = m\} = \binom{n}{m} p^m (1-p)^{n-m}, m = 0,1,...,n, \qquad (2.1)$$

where $0 \leq p \leq 1$. We denote the binomial distribution by $X \sim \text{Bin}(n, p)$. By definition, we have that $E(X) = np$, $\text{Var}(X) = np(1-p)$.

Beta Distribution

The beta distribution is an absolutely continuous distribution whose probability density function defined on the interval [0,1] is given

$$f(x) = \frac{1}{B(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1} \qquad (2.2)$$

where $\alpha > 0$, $\beta > 0$, are parameters, and the beta function B (α, β) is

$$B(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$

The expected value and variance of a beta random variable X with parameters α and β are given by

$$E(X) = \frac{\alpha}{\alpha + \beta}$$

$$V(X) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$$

It is worthwhile to see the relationship between binomial distribution and beta distribution.

Let X be Bin *(n, p)*, from time to time, we need to calculate $P\{X< m\}$ or $P\{X \geq m\}$, then we have

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

It is not difficult to see that computation gets unmanageable for large values of m and n. Hence, we have the following

$$\sum_{k=m}^{n} \binom{n}{m} p^k (1-p)^{n-k} = \frac{n!}{(m-1)!(n-m)!} \int_0^p x^{m-1}(1-x)^{n-m}\,dx$$

from this, the binomial tail probability can be calculated through beta Function, furthermore, comparing these two pdfs, we can see beta distribution is nothing but a binomial distribution replacing p with x, and n, $n-m$ with $\alpha-1$, $\beta-1$.

Multinomial Distribution

The Multinomial distribution is a multivariate generalization of the binomial distribution discussed in Subsection 2.1. Suppose that n independent trials of same experiment are conducted, each having k mutually exclusive and exhaustive possible outcomes (or cells). Let p_i

$(0 < p_i < 1, \ \sum_{i=1}^{k} p_i = 1)$ denote the single-trial probability of the event associated with ith cells ($1 \le i \le k$), and let $Y_{i,n}$ be the number of outcomes falling in cell i ($1 \le i \le k$) after n observations have been taken. Then $0 \le Y_{i,n} \le n$ and $\sum_{i=1}^{k} Y_{i,n} = n$. The k–variate discrete random variable **Y** = ($Y_{1,n}, Y_{2,n}, \ldots Y_{k,n}$) has the probability mass function

$$P\{\mathbf{Y} = (y_1, y_2, \ldots, y_k)\} = \frac{n!}{\prod_{i=1}^{k} y_i!} \prod_{i=1}^{k} p_i^{y_i}, \qquad (2.3)$$

and we say **Y** has the multinomial distribution with parameters n and **p** = ($p_1, p_2, \ldots p_k$) .

Dirichlet Distribution

The Dirichlet distribution, being the multivariate generalization of Beta distribution, is also an absolutely continuous distribution with its probability density function of $(X_1, X_2, \ldots, X_n)$ given by

$$p_{X_1, X_2 \ldots, X_n}(x_1, x_2, \ldots, x_n) = \frac{\Gamma(a_1 + \ldots + a_{n+1})}{\Gamma(a_1) \cdots \Gamma(a_{n+1})} x_1^{a_1 - 1} \cdots x_n^{a_n - 1} (1 - \sum_{i=1}^{n} x_i)^{a_{n+1} - 1}, \qquad (2.4)$$

where $a_i > 0$, $i = 1,2,\ldots,n+1$ and $(x_1, \ldots, x_n) \in S_n$.

Wilks (1962) was the first to use the terminology "Dirichlet distribution" for random variable having the density function (2.4). He explained how these distributions arise in the construction of distribution-free tolerance intervals and other connections between the Dirichlet distributions and the theory of order statistics. Dirichlet distribution is also the conjugate prior of multinomial distribution in Bayesian statistics.

The Dirichlet distribution of type II can be derived by choosing random variable Y_i appropriately. If we consider the following transformation

$$Y_1 = \frac{X_1}{1 - X_1 - \cdots - X_n}, \ldots, Y_n = \frac{X_n}{1 - X_1 - \cdots - X_n}$$

and equivalently we have

$$X_1 = \frac{Y_1}{1 + Y_1 + \cdots + Y_n}, \ldots, X_n = \frac{Y_n}{1 + Y_1 + \cdots + Y_n}.$$

The Jacobian of this transformation is $(1 + Y_1 + \cdots + Y_n)^{-(n+1)}$, then we can derive the density function of $\mathbf{Y}=(Y_1, \ldots, Y_n)$ as (2.5) from (2.4)

$$p_{Y_1, Y_2, \ldots, Y_n}(y_1, y_2, \ldots, y_n) = \frac{\Gamma(a_1 + \ldots + a_{n+1})}{\Gamma(a_1) \cdots \Gamma(a_{n+1})} (y_1^{a_1 - 1} \cdots y_n^{a_n - 1}) / (1 + \sum_{i=1}^{n} y_i)^{\sum_{i=1}^{n+1} a_i} \quad . \qquad (2.5)$$

It is also called inverted Dirichlet distribution, a multivariate generalization of beta distribution of second kind, a number of examples of its applications can be found in Sobel (1985). This distribution (2.5) can be derived from independent gamma random variables (Wilks, 1962). Especially, the cdf of the Dirichlet distribution is named Incomplete Dirichlet type I integral which is a direct generalization of the incomplete beta distribution for the multinomial case (see Sobel, 1977). The two integrals are called *I* and *J* functions. They were intensely studied in Sobel (1977) and expressed as the following

$$I_p^{(b)}(\underset{\sim}{r}, n) = \frac{n!}{(n-R)! \prod_i^b \Gamma(r_i)} \int_0^{p_1} \int_0^{p_2} \ldots \int_0^{p_b} (1 - \sum_{i=1}^{b} x_i)^{n-R} \prod_{i=1}^{b} x_i^{r_i - 1} dx_i \qquad (2.6)$$

where we assume $0 < p_i < 1/b$, $n \geq R$, $R = \sum_{i=1}^{b} r_i$ and $m = n - R + 1$, see (2.7).

In general, this Dirichlet integral can be used with most of multinomial problems, especially in max and min frequency in homogeneous multinomial (Sobel, 1977).

The Incomplete Dirichlet Type II Integral

The cdf of Dirichlet type II distribution is named the incomplete Dirichlet type II integral, which is a direct generalization of incomplete beta distribution of second kind. A b-variate random vector $(X_1,...,X_b)$ is said to be a Dirichlet Type II distribution with parameters $(r_1,...,r_b;m) = (\vec{r},m)$ if the joint density function is given by

$$f_b(\vec{x},\vec{r},m) = \frac{\Gamma(m+s)}{\Gamma(m)\prod_{i=1}^{b}\Gamma(r_i)} \frac{\prod_{i=1}^{b} x_i^{r_i-1}}{(1+\sum_{i=1}^{b} x_i)^{s+m}}$$

over the b-dimensional positive orthant $R_b = \{(x_1, x_2,..., x_b); x_i \geq 0, i = 1,..., b\}$ and is zero outside R_b, $s = \sum_{i=1}^{b} r_i$ then we have $\vec{x} \sim D_2(\vec{r},m)$. based on this, two incomplete dirichlet type-II integrals C and D functions are defined as below:

$$C_{\vec{a}}^{(b)}(\vec{r},m) = \int_0^{a_b} ... \int_0^{a_1} f_b(\vec{x},\vec{r},m) \prod_{i=1}^{b} dx_i \qquad (2.7)$$

$$D_{\vec{a}}^{(b)}(\vec{r},m) = \int_{a_b}^{\infty} ... \int_{a_1}^{\infty} f_b(\vec{x},\vec{r},m) \prod_{i=1}^{b} dx_i \qquad (2.8)$$

where $\vec{a}=(a_1,...,a_b)$, $\vec{r}=(r_1,...,r_b)$, $a_1,...,a_b$ are nonnegative, $r_1,...,r_b,m$ are all positive and b is a positive integer. Then, Olkin and Sobel (1965) have shown the following

$$P(E_1)=C_{\underset{\sim}{a}}^{(b)}(\underset{\sim}{r},m), \tag{2.9}$$

$$P(E_2)=D_{\underset{\sim}{a}}^{(b)}(\underset{\sim}{r},m), \tag{2.10}$$

where $a_i=\frac{p_i}{p_{b+1}},(i=1,2,...,b)$, (b+1)st cell as counting cell.

E_1 and E_2 represent minimum and maximum frequency events in its own right from those cells at stopping time.

$E_1=\{f_1\geq r_1,...,f_b\geq r_b$ at stopping time when f_{b+1} = m for the first time}.

$E_2=\{f_1<r_1,...,f_b<r_b$ at stopping time when f_{b+1} = m for the fist time}.

A detailed interpretation of the above expression can be found in Cho (2003).

2.2 Ranking and Selection Methodology (RSM)

Ranking and Selection is one of statistical methodologies in statistical multiple decision theory, which is commonly described by selecting (a set of) the best or largest cell(s) from the categories in a given multinomial population.

Let $\boldsymbol{X}\sim M(n, p_i)$, i = 1,2,...,k and denote the ordered p_i's

$$p_{[1]}\leq p_{[2]}\leq ...\leq p_{[k]}.$$

Consider the problem of selecting the cell associated with the largest probability $p_{[k]}$. To approach a multiple decision problem, there are two classical formulations of the problem; one is the indifference zone approach due to Bechhofer (1954) and, Bechhofer and Sobel (1954), who began as an alternative to the traditional analysis of variance, the other is the subset selection approach due to Gupta (1956) and, Gupta and Sobel (1957). These two approaches basically became the Ranking and Selection group of methodologies (RSM) we know today and which continue to develop. For applications, the former one can be used to the multiple comparisons in the analysis of experimental design, while the subset selection approach can be used to the general selecting procedures in multiple decision problems. For more details, refer the following two pioneering monographs; Gibbons, Olkin and Sobel (1977), and Gupta and Panchapakesan (1979). We will only focus on the indifference zone formulation in this thesis.

2.3. Indifference Zone Formulation

In 1954, Bechhofer introduced the concept of Ranking and Selection. He describes a problem in which the goal is to select the population with the largest mean for some population statistic from a set of t normal populations. Thus the goal of the indifference zone approach is to select a single cell and claim that it has cell probability $p_{[k]}$. Consider a procedure which selects the cell associated with $p_{[k]} = \max_{1 \le i \le t} \{p_i\}$, the

statistical issue is to determine the minimum sample size n to be used in the procedure to guarantee a pre-specified probability of correctly identifying the cell associated with $p_{[k]}$, a "Correct Selection (CS)" for this formulation. In addition, we may be indifferent in the selection of a cell when two cells are nearly the same. To quantify this, define δ to be indifference zone, if $p_{[k]} - p_{[k-1]} < \delta$, we will be indifferent to choosing $p_{[k]}$ or $p_{[k-1]}$. Therefore, the probability of correct selection

$$P\{CS\} = P\{p_{[k]} > p_{[i]} \mid p_{[k]} - p_{[i]} > \delta\} \geq P^*, \forall i \neq k$$

where $\{\delta, P^*\}$ are pre-specified. Since $P\{CS\} = 1/k$ can be achieved by simply choosing a cell at random, then $1/k < P^* < 1$ is required. If we use parameter ratio, we can have the following expression: given $\delta > 0$ and $0 < \alpha < 1$, choose the smallest n such that

$$P\{CS\} \geq 1 - \alpha$$

for all $\mathbf{p} = \{p_{[k]} \geq \delta p_{[k-1]}\}$, since we are regarded as indifferent to which cell is selected when $p_{[k]} - p_{[k-1]} < \delta$. Similarly, the sample size n is chosen so that

$$\inf_{\underset{\sim}{P}} P\{CS \mid R\} = \lim_{\underset{\sim}{P} \to \underset{\sim}{P}_{EPC}} P\{CS \mid R\}$$

for the case of LFC and EPC

In this Section, we presented some basic distributions, their notations, their properties, and interrelationships, which play fundamental roles in deriving our subsequent study method.

CHAPTER 3

DIRICHLET TYPE II INTEGRALS AND PROCEDURE

3.1 Formulation

A test that is commonly related to the goodness of fit test is the Chi-squared test for testing cell probabilities in multinomial distribution. Karl Pearson, in the paper of 1900, which in many ways opened the door to the modern era of statistical inference, proposed the test statistic

$$\chi_0^2 = \sum_{i=1}^{k} \frac{(o_i - \xi_i)^2}{\xi} = \sum_{i=1}^{k} \frac{o_i^2}{\xi_i} - n,$$

where k indicates disjoint cells, with o_i and ξ_i being observed and expected frequency, and $\sum o_i = \sum \xi_i = n$ to test the goodness-of-fit hypothesis (H_o:p=p0 vs. H_a :p≠p0). It is clear a large discrepancies between the observed and expected cell counts will result in larger values of χ^2, which roughly is the sum of squares of standardized distances from the expected counts under the null hypothesis (H_o: p=p0).

In our paper, we will consider a new testing procedure alternative to standard χ^2–test trying to consider a multinomial models where the cells have a natural ordering (Ranking and Selection).

Observations are taken one at a time from a multinomial distribution with k cells, and we arrange them in the order of their magnitude $p_{[1]} \leq p_{[2]} \leq \ldots \leq p_{[k]}$. The goal is to select a cell $p_{[1]}$, referred to as "worst" cell, so that the probability requirement to be satisfied is expressed in terms of the ratio δ of $p_{[k]}$ to $p_{[1]}$ and two pre-assigned constants P^* and δ^* such that $1/k<P^*<1$, $\delta^* >1$, we define the probability of correct decision of observing all k cells by stopping time using our inverse sampling procedure as $P\{CD|R\}$, we select the worst cell that satisfies the probability requirement

$$P\{CD|R\} \geq P^* \qquad (3.1)$$

Let R denote our inverse sampling procedure, we observe c observations from counting cell $p_{[i]}$ after observing c from any other cell (i= 1,2,...k) the integer $c = C(k, P^*, \delta^*)$ is chosen in advance in such a way that Eq. (3.1) is satisfied. This c value is our stopping rule, or stopping constant.

In our case we only consider a given number of categories with equally probable distinct cells. Suppose X_1, X_2,...,X_k be a sequence of independent observed frequencies from a multinomial population with k equally probable categories, we conduct a sequential sampling procedure by taking samples one at a time (ISP: inverse sampling procedure) seeking an efficient stopping rule, and sampling under the precision of pre-assigned P^*. According to the research results from Cho (2003), his initial minimal assumption of one cell probability is $\varepsilon = 1/10 = 0.1$, he investigated the following cell configuration scenarios

a) Equal Probability Configuration (EPC)

b) Least Favorable Configuration (LFC)

c) Multiple Slippage Configuration (MSC)

and calculated the minimum frequency on the counting cell for each case, c is used here as a stopping rule to cut short our inverse sampling procedure. In other words, we carry out our sampling and stop as soon as we observe c. Based on the research result tabulated in Cho (2003), we use c derived from the P* requirement as a stopping rule for any given number of cells k, count frequency of each cell X_i and P*, the table gives out the stopping constant c such that $P(CD \mid R) \geq P^*$ (95% and 99%). Then we use c to calculate the expected number of trials needed to halt our sampling experiment under EPC. We are mainly interested in the expected number of sample size $E(N_c)$ required to perform the test. Suppose there is a sequence of independent multinomial trials, for instance, the spin of the wheel of fortune one at a time, with outcomes $X_1, X_2, \ldots$ where X_i takes value of frequency for i–th trial, with unknown probability p_i, then a prefixed value p_0, in this case, will be $1/k$, the problem is to test

$$H_0 : p_i = p_0 \quad \text{against } H_a : p_i \text{ are not all equal to } p_0$$

Our approach for solving this is to consider N_c sample observations $X_1, X_2, \ldots X_{N_c}$ such that

$$N_0 = \min\left\{k : \sum_{i=2}^{k} X_i = n_0 \mid \min_{2 \leq i \leq k} X_i = c\right\}$$

that is, we take the sample up to N_0 for pre-assigned positive number $E(N_c|EPC)$. Then the left-tailed test based on N_0 is appropriate and is given by

Reject H_0 if $N_0 < E(N_c|EPC)$ and accept H_0 otherwise,

where $E(N_c|EPC)$, a positive number derived from the c value, is the expected number of observation under EPC, hence the level of significance can be guaranteed by 1– P*.

3.2 A compound CD integral, an inverse sampling test procedure and c value explained

If we generalize the C and D functions, the problem will become a compound multinomial model, that is, when a specified counting cell reaches at frequency m and some (k-1) specified cells have frequency bigger or equal than r , the remaining (k-t) cells being smaller than r, the probability distribution can be written as

$$CD_{\underset{\sim}{a}}^{(t-1),(k-t)}(r,m)=\frac{\Gamma[m+(k-1)r]}{\Gamma^{k-1}(r)\Gamma(m)}\int_0^{a_1}\cdots\int_0^{a_{t-1}}\int_{a_t}^{\infty}\cdots\int_{a_{k-1}}^{\infty}\frac{\prod_{i=1}^{t-1}x_i^{r-1}dx_i}{[1+\sum_{i=1}^{t-1}x_i]^{m+(t-1)r}} \qquad (3.2)$$

where $\underset{\sim}{a}=(a_1,...,a_{t-1},a_t,...a_{k-1})$ and $a_j=p_j/p_0$ is the ratio of the jth cell probability to the probability of the counting cell. This expression has been established by Sobel, Uppuluri, and Frankowski (1985).

A further generalization of the CD integral can be found in Cho (2003), they are called generalized multiple CD-integrals.

3.2.1 A proposed inverse sampling procedure to test for homogeneity in multinomial models and stopping constant *c*

In this section, we will propose a new test procedure using inverse sampling procedure with an efficient stopping rule.

Let $X_1, X_2, ...X_k$ be a random sample from multinomial distribution with parameters $p_1, p_2, ...p_k$ with joint probability function

$$f(x_1, x_2, ...x_k) = \binom{N}{x_1, x_2, ...x_k} p_1^{x_1} ... p_k^{x_k}$$

where $\sum_{i=1}^{k} x_i = N, \sum_{i=1}^{k} p_i = 1$ and let $X_{(1)} \le X_{(2)} \le ... \le X_{(k)}$ be the order statistic obtained by arranging them in increasing order of magnitude. In this paper, we are interested in testing, as we mentioned in the introduction, the homogeneity hypothesis

$$H_0 : p_1 = p_2 = ... = p_k = 1/k.$$

the standard χ^2–test for the H_0 is based on a fixed sample size N, with

test statistic $$T = k/N \sum_{i=1}^{k} (X_i - N/k)^2$$

with its distribution is approximately χ^2 , *k–1* degrees of freedom under *H0* for large N. (see Hogg and Craig, 1995)

In the inverse sampling procedure, observations are taken one at a time and the sampling is terminated when the count in any one of the cells reaches a specified number *c*. Let $X_1, X_2, ...X_k$ denote the cell counts at termination. For a fixed value of *k* and P*, we first choose values of *c* and

N_0 respectively. We then take one observation at a time and continue sampling until the first time when one of the following three events happens:

(a) $p_{(k)} - p_{(k-1)} \geq \delta^*$ (this is the condition to reject H_0)

(b) $X_{(1)} = c$

(c) N_0 observations have been taken (from $E(N_c \mid EPC)$).

Event (a) has been controlled by our indifference zone method, event (b) has also been controlled by the P*, event (c) can be derived from (a) and (b). Therefore, our testing procedure will be:

Once event (c) happens, then we stop sampling and accept H_0

Insofar as decision theory is concerned, Cho (2003), utilizing multiple decision theory, made his detailed study of seeking the expression of P(WD) in terms of both DC-integrals (the inverse side of CD-integrals) and generalized multiple CD-integral under the assumption of one missing cell and minimum cell probability taking value of *1/10*. The probability of making wrong decision P(WD) is as follows

$$P(WD) = t[(\frac{9}{9+t})^c + (\frac{t}{9+t})^c]C^{t-1}_{\frac{9}{9+t}}(c;c) + \frac{t(t-1)}{2^c}[C^{(t-2)}_{1/2}(c;c) - \sum_{i=1}^{c}(\begin{matrix}2c-1-i \\ c-1\end{matrix})[(\frac{t}{18+t})^{c-i}(\frac{18}{18+t})^c C^{t-2}_{\frac{9}{18+t}}(c;2c-i)] \quad (3.3)$$

where t is the number of the cells observed, c, the frequency of counting cell, the most important element in this thesis, is called the stopping constant which is actually same as r in the *CD*–integral. It is this very

decision rule that we use to stop our sampling. Before we go further, we would like to sideline c to clarify its real meaning and function in this thesis, c as a point estimator, serves as an approximation to the true, but unknown , values of integral of interest, it guarantee a specified (<5%) bound on error. As we previous mentioned, the development of c value is based on multiple decision theory, Cho (2003), in which he uniformly restricted the probability of making wrong decision within 0.05. This is same as putting the probability of making wrong decision under a tight control, the various configuration of cells themselves are taken into consideration when the Eq. (3.3) was developed. Therefore, the Eq. (3.3) indicates the probability of making a wrong decision, where the observed cells $t = k - 1$. Our purpose of research is to develop an efficient way of testing the fairness of wheel of fortune. In our model the cell number is given and an assumed equal probability of each cell is imposed. Therefore, in order to be 95% sure that our test is correct while using this c value, we have to re-examine table Cho (2003). It reveals fact that a case of given 2 cells equals to that of 2 cells are observed from making correct decision point of view, i.e. in order to use the c value to come to a correct decision, we have to initially start with t equal to 2, which correspond to the decision rule c value of 13 which will guarantee us at least 95% sure of making a correct decision. This dual aspect of the table of c value, due to its robust nature, allows us not only to estimate the true number of cells but to use it as a decision rule to stop sampling

under a given cell number case as well. We can express P(*CD*) through P(*WD*), where P{CD} is the probability of making correct decision observing all cells, P{WD} is the probability of wrong decision being made when we stop sampling without seeing all the cells. It can be expressed as following

$$P(CD) = 1 - P(WD) \tag{3.4}$$

for the development of c value, refer to Cho (2003).

3.3 Expected Number of Observations and Testing Under EPC

It is not difficult to notice that under EPC, the expected number of observations will turn out to be the best case insofar as the sample size is concerned, in other words, the number of observations will reach its minimum size compared to that of all the other cell configurations, making it the best and most economical case to have.

Due to its equiprobable cell configuration, our calculation of the expected number of observations will be drastically simplified. Cho (2003) has shown the expression of the equation as

$$E(N_c \mid EPC) = k^2 c C_1^{k-1}(c, c+1) \tag{3.5}$$

where c is the our minimum frequency of each cell, it also is our pre-assigned stopping constant derived from the last section. Then we can do our testing based on the those pre-assigned number, P*, c and $E(N_c \mid EPC)$.

CHAPTER 4

NUMERICAL STUDIES

4.1 Monte Carlo Simulation

For a simulation study, Monte Carlo experimentation is carried out in order to illustrate the behavior and the performance of the stopping rule in the proposed sequential procedure. The results of the Monte Carlo simulation, based on the stopping rule are summarized in the following Tables 3.1 to 3.4, which include the number of categories k, the stopping constant c (in fact, this is the minimum cell frequency, obtained LFC), the average of the optimal stopping time, E(N), the average sample number *Avg.#*, standard error of the average sample size *s.e.*, and the average observed coverage probability P(*CD*) in the experiments. Each row in the table corresponds to 10,000 independent experiments.

We observe that both the expected optimal sample sizes and average sample numbers increase as the number of categories increases. We also see that the coverage probabilities are uniformly higher than the pre-assigned desirable probability P*. From this, we conclude that the numerical results indicate the small sample behavior and provide support for the suggested procedure. The next four tables are the results for P* = 0.95 and P* = 0.99, respectively. In addition, we present the

simulation results using several multiple slippage configurations for k = 4 and k = 6.

Table 3.1: Sequential Estimation of optimum sample size in Multinomial distribution using optimal stopping rule under EPC

$P\{CD\} \geq P^*$, P^*=0.95, ε=1/10

# of expt =10000 each row

k	C	E(N)	Avg.#	s.e.	P(CD)
2	13	30.0873	30.0423	.6876	.9650
3	8	31.3416	31.3045	.9425	.9606
4	6	34.6096	34.6236	1.1598	.9651
5	5	39.0148	39.0786	1.3178	.9645
6	4	40.7894	40.9034	1.4445	.9639
7	4	49.2179	49.2729	1.5838	.9609
8	4	57.8517	57.7773	1.6951	.9592
9	4	66.8694	66.6121	1.7992	.9580

Table 3.2

k	C	E(N)	Avg.#	s.e.	P(CD)
2	13	30.0873	30.0334	.6836	.9602
3	8	31.3416	31.2465	.9352	.9646
4	6	34.6096	34.6205	1.1555	.9631
5	5	39.0148	39.0475	1.3089	.9618
6	4	40.7894	40.9166	1.4604	.9604
7	4	49.2179	49.3128	1.5741	.9598
8	4	57.8517	58.0985	1.6688	.9581
9	4	66.8694	66.7224	1.7964	.9620

Table 3.3

k	C	E(N)	Avg.#	s.e.	P(CD)
2	13	30.0873	30.0186	.6894	.9587
3	8	31.3416	31.3317	.9332	.9640
4	6	34.6096	34.5447	1.1500	.9603
5	5	39.0148	38.9436	1.3076	.9621
6	4	40.7894	40.9192	1.4641	.9607
7	4	49.2179	49.3846	1.5669	.9592
8	4	57.8517	58.0223	1.6767	.9569
9	4	66.8694	66.7343	1.7893	.9564

Table 3.4: Sequential Estimation of optimum sample size in Multinomial distribution using optimal stopping rule under EPC

$P\{CD\} \geq P^*$, $P^*=0.99$, $\varepsilon=1/10$

# of expt =10000 each row

k	C	E(N)	Avg.#	s.e.	P(CD)
2	20	45.9212	44.9479	.6545	.9920
3	13	48.4064	48.4155	.9366	.9910
4	9	48.9326	48.9855	1.1197	.9930
5	7	51.3034	51.3127	1.2841	.9938
6	6	56.1073	56.1591	1.4177	.9934
7	5	58.6282	58.6465	1.5450	.9923
8	5	68.7021	68.7484	1.6468	.9925
9	5	79.0234	79.0271	1.7870	.9917

From Table 3.1-3.4, we observe that both the expected optimal sample sizes and the average sample numbers monotonically increase as the number of categories increases. The Monte Carlo simulation values of the average stopping time are very close to the expected number of stopping time, E(N), we computed. We also see that the coverage probabilities are uniformly higher than the prescribed desirable P*-condition. This provides a substantial amount of numerical evidence for us to conclude

that the proposed procedure performs satisfactorily in testing uniformity for given multinomial models.

4.2 Illustration – Wheel of Fortune

For an illustration, we use the result in the previous section. Let's take a wheel that has six categories, so $k = 6$. For simplicity, consider EPC, namely all categories have the same probability $p_i = 1/6$.

Let the random vector $\mathbf{X} = (X_1, X_2,\ldots,X_k)$ follow the multinomial distribution with corresponding probability vector $\mathbf{p} = (p_1, p_2,\ldots, p_k)$ whose components are positive and add to one. For testing the equi-probable model in a given multinomial distribution, we hypothesize the null model and specify the corresponding alternative as follows:

$$H_o: p_i = 1/k \text{ for all } i, \quad \text{vs.} \quad H_a: \text{Not all } p_i \text{ are equal.}$$

Then, we wish to test the fairness of a die. From the Tables 3.1-3.3 we find the stopping constant $c = 4$ and use this observed minimum positive frequency (MPF) = 4 for the stopping rule and rejection rule. Then we reject the null hypothesis of uniformity if we get to observed MPF = 4 without seeing all the six categories. In other words, every frequency at stopping time should be either zero or at least 4. Therefore the level of significance becomes $1-P^* < 0.05$ if we choose $P^* = 0.95$ as our level of confidence. In fact, we can compute that the probability of Type I error for this case is $1-0.9618 = 0.0322$ which is smaller than 0.05. Moreover, when the stopping value for $P^* = 0.99$, $c = 6$, the Type I error is less than 0.01.

We present the simulation results of two examples for slippage configurations:

Table 3.5 Optimal Stopping time for Spinning of a Wheel of Fortune
k = 4 MSC = 1/8 & 3/8, stopping values from LFC
configuration (1/8,3/8,1/8,3/8)

P*=0.95 ε = 1/10 # of expt = 10000 each

k	c	Avg	s.e	P(CD)
4	6	58.8361	2.2640	.9552
4	6	59.0613	2.2655	.9552
4	6	58.7527	2.2530	.9543
4	6	58.8414	2.2593	.9554

Table 3.6 Optimal Stopping time for Spinning of a Wheel of Fortune
k =6 MSC = 1/9 & 2/9, stopping values from LFC
configuration (1/9,2/9,1/9,2/9,1/9,2/9)

P*= 0.95 ε = 1/10 # of expt = 10000 each

k	c	Avg	s.e	P(CD)
6	4	51.9099	2.1419	.9609
6	4	52.0990	2.1520	.9604
6	4	52.2573	2.1237	.9607
6	4	51.9201	2.1232	.9616

CHAPTER 5

CONCLUSION

In the thesis we have studied an inverse-type sequential procedure to obtain optimal sample sizes for testing the hypothesis H_0: $p_i = 1/k$, $i = 1,2,\ldots,k$. The traditional Chi-squared test performs very well for fixed-sample size, but the test does not tell us about the optimal sample size for the test. The proposed procedure for testing is to optimize the stopping time (or sample size) by controlling the prescribed P*-condition under the decision theoretic framework.

We have extended to testing for a slippage model with two different cell probabilities such as multiple slippage configurations, and presented the results for further development.

REFERENCES

1. Bechhofer, R. (1954). A Single-sample multiple decision procedure for ranking means of normal populations with known variances. *The Annals of Mathematical Statistics*, Vol. 25, pp. 16-39.

2. Bechhofer, R. and Sobel, M. (1954). A Single-sample multiple decision procedure for ranking variances of normal populations. *The Annals of Mathematical Statistics*, Vol. 25, pp. 273-289.

3. Cho, H. (2003). Inverse-Type Sampling Procedure for Estimating the Number of Multinomial Classes. *Sequential Analysis*. Vol. 22, pp. 307-324.

4. Gibbons, J., Olkin, I., and Sobel, M. (1977). *Selecting and Ordering Population: A New Statistical Methodology*. John Wiley & Sons, New York.

5. Govindarajulu, Z. (1999). The Elements of Sampling Theory and Methods. Prentice-Hall, Inc., New Jersey.

6. Greenwood, P. and Nikulin, M. (1996). *A Guide to Chi-Squared Testing*. John Wiley & Sons, New York.

7. Gupta, S. (1956). On a decision rule for a problem in ranking means. Institute of Statistics Mimeo Series No. 150, University of North Carolina, Chapel Hill, North Carolina.

8. Gupta, S. (1965). On some multiple decision (selection and ranking) rules. *Technometrics*, Vol. 7, pp. 225-245.

9. Gupta, S. and Panchapakesan, S. (1979). *Multiple Decision Procedures: Theory and Methodology of Selecting and Ranking Populations*. John Wiley & Sons, New York, New York.

10. Gupta, S. and Sobel, M. (1957). On a statistic which arises in selection and ranking problems. *The Annals of Mathematical Statistics*, Vol. 28, pp. 957-967.

11. Olkin, I. and Sobel, M. (1965). Integral expression for tail probabilities of the multinomial and negative multinomial distributions, *Biometrika*. Vol. 52, pp. 167-179.

12. Sobel, M., Uppuluri, V. and Frankowski, K. (1977). *Selected Tables in Mathematical Statistics, Vol. IV -Dirichlet distribution Type I.* Edited by IMS. American Mathematical Society, Providence, Rhode Island.

13. Sobel, M., Uppuluri, V. and Frankowski, K. (1985). *Selected Tables in Mathematical Statistics, Vol. IX -Dirichlet Integrals of Type II and Their Applications.* Edited by IMS. American Mathematical Society, Providence, Rhode Island.

14. Wilks, S. (1962). *Mathematical Statistics.* John Wiley & Sons, New York.

VITA

Graduate College
University of Nevada, Las Vegas

Hai Zhen

Local Address:
4386 Escondido St. #215
Las Vegas, NV 89119

Home Address:
144–80 Sanford Ave. #6H
Flushing, NY 11355

Degrees:
Bachelor of Science, Industrial Automation, 1988
Anhui University, Hefei, China

Thesis title: Sequential Procedure for Test of Uniformity in Multinomial Models

Thesis Examination Committee:
Chairperson, Professor Hokwon A. Cho, Ph.D
Committee Member, Professor Malwane Ananda, Ph.D
Committee Member, Professor Chih-Hsiang Ho, Ph.D
Committee Member, Professor Sandra Catlin, Ph.D
Graduate Faculty Representative, Professor Bernard Malamud, Ph.D