AN ABSTRACT OF THE THESIS OF

SUEY-HUEY TARNG for the degree of DOCTOR OF PHILOSOPHY in STATISTICS

presented on March 6, 1980

Title: ESTIMATION OF THE POPULATION TOTAL WHEN THE SAMPLE IS TAKEN FROM A LIST CONTAINING AN UNKNOWN AMOUNT OF DUPLICATION

Abstract approved: Redacted for privacy

G. David Faulkenberry

A frame contains a known number, N, of units, but the units are grouped into an unknown number of M distinct classes. A measurement y_j is associated with each class, and, based on the information obtained from a simple random sample of units from the frame, we wish to estimate the population total, $\Sigma_{j=1}^{M} y_j$, without knowing M. Several researchers have proposed methods for estimating M based on a sample. In this thesis five of these methods are generalized to obtain estimates of the population total.

Estimation of The Population Total When The Sample Is Taken From A List Containing An Unknown Amount of Duplication

by

Suey-huey Tarng

A THESIS

submitted to

Oregon State University

in partial fulfillment of
the requirements for the
degree of

Doctor of Philosophy

June 1980

APPROVED:

Redacted for privacy

Professor of Statistics

in charge of major

Redacted for privacy

Chairman of Department of Statistics

Redacted for privacy

Dean of Graduate School

Date thesis is presented March 6, 1980

Typed by Debbie Dudley for Suey-huey Tarng

Acknowlegement

The author would like to express her appreciation to Dr. David Faulkenberry who are her major professor, and Dr. David Birkes for their great amount of time, and encouragement provided. Thanks go to my parents and my friends -- Mr. Y.F. Suen and his wife for their encouragement.

TABLE OF CONTENTS

ESTIMATION OF THE POPULATION TOTAL WHEN THE SAMPLE IS TAKEN FROM A LIST CONTAINING AN UNKNOWN AMOUNT OF DUPLICATION

CHAPTER 1

INTRODUCTION

The problem considered here arose in connection with a sample survey of the owners of fishing licenses. The objective of the survey was to estimate the total number of fish caught. A list of fishing licenses was available from which to select a sample, but since it is possible for one individual to buy more than one license, the same fisherman could appear two or more times in the list. The presence of an unknown amount of duplication causes much difficulty. Two distinct conditions exist. One can either determine how many licenses each person in the sample has, or this cannot be determined. The estimate of the total number of fish caught for the first condition was obtained by Rao [14]. We shall consider only the estimation of the total number of fish caught for the second condition.

In an abstract setting, there is a list of a known number, N, of units (licenses) which is subdivided into an unknown number, M, of distinct classes, C_j, j=1, 2, ..., M (each fisherman represents a class of licenses). If the number of units in a class is R_j, then $\sum_{j=1}^{M} R_j = N$. The class of a unit is readily identifiable when the unit is examined. To each class, a measurement, y_j, (the number of fish caught by the fisherman) is associated. From a sample of size n, we wish to estimate the total of these measurements, $T = \sum_{j=1}^{M} y_j$, without knowing the R_j values for units in the sample. Several researchers have proposed methods for estimating the total number M of distinct classes. In this thesis we generalize five of these methods to obtain estimates of the population total, T. Note that in the special case when $y_j = 1$ for all j, the total is simply M.

The statistical methods used in this study can be classified as follows:

(A) Nonparametric models

(a) Sampling without replacement - Goodman's Method

Goodman offered an unbiased estimate of the total number M of distinct classes. In this thesis we generalize his estimate to find the unbiased estimate of the population total, T.

(b) Sampling with replacement - Good and Toulmin's Method, Harris' Method, and one of Efron and Thisted's Methods

Good, Toulmin, Efron, and Thisted obtained reasonable estimates of the total number M of distinct classes. Harris found approximations to the supremum and infimum of these estimates. We generalize these results to find estimates of the population total and approximations to the supremum and infimum of the estimates.

(B) Parametric Models

Sampling with replacement - Good and Rao's Method and one of Efron and Thisted's Methods

Good, Rao, Efron, and Thisted found reasonable estimates of the total number, M, of distinct classes by assuming gamma and/or beta distribution. We generalize these estimates to obtain estimates of the population total.

The performance of each method was tested on a set of simulated data.

1.1 Notation

We define the following notation:

N:	the list size
M:	the number of distinct classes of the list
C_j:	the jth class $(j=1, \ldots, M)$
y_j:	the measurement of the jth class
$T = \sum_{j=1}^{M} y_j$:	the total of the measurements of all classes
R_j:	the number of units in the jth class
q:	the maximum number of units contained in any class, i.e., $q = \max_{j=1, \ldots, M} R_j$
J_ℓ:	the collection of indices of all the classes consisting of ℓ elements, i.e., $J_\ell = \{j: R_j = \ell\}$
X_j:	the number of units in the jth class showing in the sample

$$Z_j^{(r)} = y_j I_{\{r\}}(X_j) \text{ where } I(.) \text{ is the indicator function}$$

$$= \begin{cases} y_j & \text{if the jth class has r units in the sample} \\ 0 & \text{otherwise} \end{cases}$$

Hence $\sum_{j=1}^{M} Z_j^{(r)}$ is the total of the measurements of all the classes having r units in the sample.

$$\delta_j = \begin{cases} 1 & \text{if the jth class shows in the sample} \\ 0 & \text{otherwise} \end{cases}$$

$$Y_j = \delta_j y_j = \begin{cases} y_j & \text{if the jth class shows in the sample} \\ 0 & \text{otherwise} \end{cases}$$

$T_s = \sum_{j=1}^{M} Y_j = \sum_{r=1}^{n} \sum_{j=1}^{M} Z_j^{(r)}$: the total of the measurements of all classes that show in the sample

T': the total of the measurements of all the units of the list, i.e., $T' = \sum_{j=1}^{M} R_j y_j$

T_s': the total of the measurements of the units of the sample, i.e., $T_s' = \sum_{r=1}^{n} r\left(\sum_{j=1}^{M} Z_j^{(r)}\right)$

d_i: unit i of the random sample for i=1, ... , n

P_j: the probability that the ith unit of the sample is in the jth class, i.e., $P_j = P_r\{d_i \varepsilon C_j\} > 0$ (not depending on i)

We regard the random sample of size n as being the basic sample. We imagine a second hypothetical sample of size tn. Since the estimates of the population total based on Good and Toulmin's method, Harris' method, and Efron and Thisted's method are the prediction of the population total that will be observed in the second sample of size N where $t = \frac{N}{n}$, we need the following notation:

$X_j^{(t)}$: the number of units of the jth class showing in the second sample of size tn

$$Z_j^{(r)}(tn) = y_j I_{\{r\}}(X_j^{(t)})$$

$$= \begin{cases} y_j & \text{if the jth class has r units in the sample of size tn} \\ 0 & \text{otherwise} \end{cases}$$

Hence $\sum_{j=1}^{M} Z_j^{(r)}(tn)$ is the total of the measurements of all the classes having r units in the sample of size tn.

$$\delta_j^{(t)} = \begin{cases} 1 & \text{if the jth class shows in sample of size tn} \\ 0 & \text{otherwise} \end{cases}$$

$$Y_j(tn) = y_j \delta_j^{(t)} = \begin{cases} y_j & \text{if the jth class shows in the sample of size tn} \\ 0 & \text{otherwise} \end{cases}$$

Hence $\sum_{j=1}^{M} Y_j(tn) = \sum_{r=1}^{n} \sum_{j=1}^{M} Z_j^{(r)}(tn)$ is the total of the measurements of all the classes in the second sample.

CHAPTER 2

GOODMAN'S METHOD

2.1 Introduction

In this chapter the sampling is done without replacement.

Goodman [8] offered the unbiased estimator $\sum_{i=1}^{n} A_i f_i$ of the total number M of distinct classes, where $A_i = 1 - (-1)^i \dfrac{[N - n + i - 1]^{(i)}}{n^{(i)}}$,

$$a^{(t)} = \begin{cases} a(a-1) \ldots (a-t+1) & \text{for } t > 0 \\ 1 & \text{for } t = 0 \end{cases}$$

and f_i = the number of classes containing i units in the sample. Knott [13] showed that by considering a second sample of size tn = N he got the same unbiased estimator of M. We generalize their results to find an unbiased estimator of the total $T = \sum_{j=1}^{M} y_j$. The unbiased estimator is $\sum_{r=1}^{n} A_r \left(\sum_{j=1}^{M} Z_j^{(r)} \right)$.

2.2 Derivations

In order to find the unbiased estimator of $T = \sum_{j=1}^{M} y_j$ we need:

Assumption: The sample size n is not less than the maximum number, q, of individuals contained in any one class.

This assumption is reasonable for our practical problems.

Lemma 2.1: $$E\left[\sum_{j=1}^{M} Z_j^{(r)}\right] = \sum_{\ell=r}^{q} \frac{\binom{\ell}{r}\binom{N-\ell}{n-r}}{\binom{N}{n}} \left(\sum_{j\in J_\ell} y_j \right)$$

Proof: $$E\left[\sum_{j=1}^{M} Z_j^{(r)}\right] = \sum_{j=1}^{M} y_j E\left[I_{\{r\}}(X_j)\right] = \sum_{j=1}^{M} y_j \frac{\binom{R_j}{r}\binom{N-R_j}{n-r}}{\binom{N}{n}}$$

$$= \sum_{R_j=r}^{q} \frac{\binom{R_j}{r}\binom{N-R_j}{n-r}}{\binom{N}{n}} \left(\sum_{j\epsilon J_{R_j}} y_j\right)$$

Using this lemma we obtain an unbiased estimator of T in the following theorem.

Theorem 2.1: Let $A_r = 1 - (-1)^r \frac{[N - n + r - 1]^{(r)}}{n^{(r)}}$,

$$\text{where } a^{(t)} = \begin{cases} a(a-1)\ \ldots\ (a-t+1) & \text{for } t > 0 \\ 1 & \text{for } t = 0 \end{cases}.$$

$$\text{Then } E\left[\sum_{r=1}^{n} A_r\left(\sum_{j=1}^{M} Z_j^{(r)}\right)\right] = \sum_{j=1}^{M} y_j\ .$$

Proof: $$E\left[\sum_{r=1}^{n} A_r\left(\sum_{j=1}^{M} Z_j^{(r)}\right)\right] = \sum_{r=1}^{n} A_r E\left[\sum_{j=1}^{M} Z_j^{(r)}\right]$$

$$= \sum_{r=1}^{n}\left[1 - (-1)^r\frac{[N - n + r - 1]^{(r)}}{n^{(r)}}\right]\left[\sum_{\ell=r}^{q} \frac{\binom{\ell}{r}\binom{N-\ell}{n-r}}{\binom{N}{n}} \sum_{j\epsilon J_\ell} y_j\right]$$

$$= \sum_{\ell=1}^{q}\left(\sum_{j\epsilon J_\ell} y_j\right)\left[\sum_{r=1}^{\ell}\left(1 - (-1)^r \frac{[N - n + r - 1]^{(r)}}{n^{(r)}}\right)\frac{\binom{\ell}{r}\binom{N-\ell}{n-r}}{\binom{N}{n}}\right]$$

$$= \Sigma_{\ell=1}^{q} \sum_{j\epsilon J_\ell} y_j = \Sigma_{j=1}^{M} y_j \text{ by lemma 2 of [8].}$$

An alternative derivation of the result in Theorem 2.1 can be obtained as follows:

Theorem 2.2: Suppose the statistics $W_1, W_2, \ldots, W_n$ are the solution of the system of linear equations

$$\Sigma_{j=1}^{M} Z_j^{(r)} = \Sigma_{\ell=r}^{n} \frac{\binom{\ell}{r}\binom{N-\ell}{n-r}}{\binom{N}{n}} W_\ell \text{ for } r = 1, 2, \ldots, n.$$

Then $E(W_\ell) = \Sigma_{j\varepsilon J_\ell} y_j$.

Proof: The same proof as Theorem 4 of [8].

Therefore $\Sigma_{\ell=1}^{n} W_\ell$ is an unbiased estimator of T.

There always exists a unique solution of the system of linear equations in Theorem 2.2 since the determinant of the coefficients of W_ℓ, $\ell=1, \ldots, n$ is not equal to zero. The following theorem shows that $\Sigma_{\ell=1}^{n} \ell W_\ell$ is an unbiased estimator of T', the sum of the measurements of all the units of the list.

Theorem 2.3: If $W_1, \ldots, W_n$ are as in Theorem 2.2,

Then $E\left(\Sigma_{\ell=1}^{n} \ell W_\ell\right) = T'$.

Proof: Recall T_s', the sum of the measurements of the units of the sample, and note that

$$T_s' = \Sigma_{r=1}^{n} r \left(\Sigma_{j=1}^{M} Z_j^{(r)}\right)$$

$$= \Sigma_{r=1}^{n} r\left[\Sigma_{\ell=r}^{n} \frac{\binom{\ell}{r}\binom{N-\ell}{n-r}}{\binom{N}{n}} W_\ell\right]$$

$$= \Sigma_{\ell=1}^{n} W_\ell \left[\Sigma_{r=1}^{\ell} r \frac{\binom{\ell}{r}\binom{N-\ell}{n-r}}{\binom{N}{n}} \right]$$

$$= \frac{n}{N} \Sigma_{\ell=1}^{n} \ell W_\ell .$$

Thus $\Sigma_{\ell=1}^{n} \ell W_\ell = \frac{N}{n} T_s'$, so

$$E\left(\Sigma_{\ell=1}^{n} \ell W_\ell \right) = T' .$$

In some of the later chapters the problem of estimating the total is considered as the prediction of the total of a second sample drawn from the same infinite population. Here we give the similar result for a second sample from a finite list. The following theorem gives an unbiased estimator of

$$E\left[\Sigma_{j=1}^{M} Z_j^{(r)}(tn) \right] \text{, for a second sample of size tn.}$$

Theorem 2.4: $E\left[\Sigma_{s=r}^{n} \frac{\binom{tn}{r}\binom{n-tn}{s-r}}{\binom{n}{s}} \left(\Sigma_{j=1}^{M} Z_j^{(s)} \right) \right] = E\left[\Sigma_{j=1}^{M} Z_j^{(r)}(tn) \right]$

Proof: Since $E\left[\sum_{j=1}^{M} Z_j^{(r)}(tn) \right] = \sum_{\ell=r}^{n} \frac{\binom{\ell}{r}\binom{N-\ell}{tn-r}}{\binom{N}{tn}} \left(\sum_{j\varepsilon J_\ell} y_j \right)$,

$$\text{Hence } E\left[\sum_{s=r}^{n} \frac{\binom{tn}{r}\binom{n-tn}{s-r}}{\binom{n}{s}} \left(\sum_{j=1}^{M} Z_j^{(s)} \right) \right] = \sum_{s=r}^{n} \frac{\binom{tn}{r}\binom{n-tn}{s-r}}{\binom{n}{s}} \left[\sum_{\ell=s}^{n} \frac{\binom{\ell}{s}\binom{N-\ell}{n-s}}{\binom{N}{n}} \left(\sum_{j\varepsilon J_\ell} y_j \right) \right]$$

$$= \sum_{\ell=r}^{n} \sum_{s=r}^{\ell} \frac{\binom{tn}{r}\binom{n-tn}{s-r}}{\binom{n}{s}} \frac{\binom{\ell}{s}\binom{N-\ell}{n-s}}{\binom{N}{n}} \left(\sum_{j\varepsilon J_\ell} y_j \right)$$

by lemma of [11]

$$= \sum_{\ell=r}^{n} \frac{\binom{\ell}{r}\binom{N-\ell}{tn-r}}{\binom{N}{tn}} \left(\sum_{j\varepsilon J_\ell} y_j\right) = E\left[\sum_{j=1}^{M} Z_j^{(r)}(tn)\right]$$

Remark:

(1) If tn = N (i.e. we sample the whole population), then

$$E\left[\sum_{j=1}^{M} Z_j^{(r)}(N)\right] = \sum_{j\varepsilon J_r} y_j .$$

In other words, $\sum_{s=r}^{n} \frac{\binom{N}{r}\binom{n-N}{s-r}}{\binom{n}{s}} \left(\sum_{j=1}^{M} Z_j^{(s)}\right)$ is an unbiased estimator of $\sum_{j\varepsilon J_r} y_j$.

(2) Note $\sum_{j=1}^{M} Y_j(tn) = \sum_{r=1}^{n} \sum_{j=1}^{M} Z_j^{(r)}(tn)$. An unbiased estimator of $E\left[\sum_{j=1}^{M} Y_j(tn)\right] = \sum_{r=1}^{n} E\left[\sum_{j=1}^{M} Z_j^{(r)}(tn)\right]$

is $\sum_{r=1}^{n} \sum_{s=r}^{n} \frac{\binom{tn}{r}\binom{n-tn}{s-r}}{\binom{n}{s}} \left[\sum_{j=1}^{M} Z_j^{(s)}\right] = \sum_{s=1}^{n} \left[1 - \frac{\binom{n-tn}{s}}{\binom{n}{s}}\right]\left[\sum_{j=1}^{M} Z_j^{(s)}\right]$.

(3) If tn = N, then an unbiased estimator of $T = \sum_{j=1}^{M} y_j$

is $\sum_{s=1}^{n} \left[1 - \frac{\binom{n-N}{s}}{\binom{n}{s}}\right]\left[\sum_{j=1}^{M} Z_j^{(s)}\right] = \sum_{s=1}^{M} A_s\left(\sum_{j=1}^{M} Z_j^{(s)}\right)$. Thus, Theorem 2.4 leads us to the same estimator of T as Theorem 2.1 does.

The following theorem shows the variance of the unbiased estimator $\Sigma_{r=1}^{n} A_r \left(\Sigma_{j=1}^{M} Z_j^{(r)}\right)$.

Theorem 2.5:

$$\mathrm{Var}\left[\sum_{r=1}^{n} A_r\left(\sum_{j=1}^{M} Z_j^{(r)}\right)\right] =$$

$$\sum_{r=1}^{n}\sum_{s=1}^{n} A_r A_s \left\{ \sum_{\substack{h=1 \\ v\varepsilon J_h \\ w\varepsilon J_\ell}}^{q} \sum_{\ell=1}^{q} \mathrm{Cov}\left(I_{\{r\}}(X_v),\ I_{\{s\}}(X_w)\right)\left(\sum_{j\varepsilon J_h} y_j\right)\left(\sum_{k\varepsilon J_\ell} y_k\right) \right.$$

$$\left. - \sum_{\substack{h=1 \\ v\varepsilon J_h \\ w\varepsilon J_h}}^{q} \mathrm{Cov}\left(I_{\{r\}}(X_v),\ I_{\{s\}}(X_w)\right)\left(\sum_{j\varepsilon J_\ell} y_j^2\right)\right\} + \sum_{r=1}^{n} A_r^2\left\{\sum_{\substack{h=1 \\ v\varepsilon J_h}}^{q} \mathrm{Cov}\left(I_{\{r\}}(X_v)\right)\left(\sum_{j\varepsilon J_h} y^2\right)\right\}$$

Proof: $$\mathrm{Var}\left[\sum_{r=1}^{n} A_r\left(\sum_{j=1}^{M} Z_j^{(r)}\right)\right] = \sum_{r=1}^{n}\sum_{s=1}^{n} A_r A_s\ \mathrm{Cov}\left(\sum_{j=1}^{M} Z_j^{(r)},\ \sum_{j=1}^{M} Z_j^{(s)}\right)$$

$$= \sum_{r=1}^{n}\sum_{s=1}^{n} A_r A_s \sum_{j}\sum_{k} y_j y_k\ \mathrm{Cov}\left(I_{\{r\}}(X_j),\ I_{\{s\}}(X_k)\right) ,$$

where

$$\mathrm{Cov}\left(I_{\{r\}}(X_j),\ I_{\{s\}}(X_k)\right) = \begin{cases} 0 & j=k \text{ and } r\neq s \\ \mathrm{Var}\left(I_{\{r\}}(X_j)\right) & j=k \text{ and } r=s \\ \mathrm{Cov}\left(I_{\{r\}}(X_j),\ I_{\{s\}}(X_k)\right) & j\neq k \end{cases}$$

2.3 Discussion

Since $W = \sum_{r=1}^{n} A_r\left(\sum_{j=1}^{M} Z_j^{(r)}\right)$, the unbiased estimator of T, can be negative, we consider other possible estimators of T.

(1) In many practical problems $\sum_{j=1}^{M} Z_j^{(r)}$ is small for $r \geq 3$, and a

reasonable estimator is $W' = A_1 \sum_{j=1}^{M} Z_j^{(1)} + A_2 \sum_{j=1}^{M} Z_j^{(2)}$

$$= \frac{N}{n} T'_s - \frac{N(N-1)}{n(n-1)} \sum_{j=1}^{M} Z_j^{(2)} .$$

(2) Another estimator sometimes used in $W'' = \frac{N}{n} T_s = \frac{N}{n} \sum_{r=1}^{n} \sum_{j=1}^{M} Z_j^{(r)}$.

It may be shown to overestimate when $q \neq 1$.

If the value of W is positive, then it is reasonable to use W as the estimator of T. If the value of W is negative, then we might consider W'. And if the value of W' is negative, we might prefer to use W" as the estimator of T, which is always positive.

2.4 Example

Consider a list of size N = 14,115 with M = 12,000 distinct classes, 9,885 of them having 1 unit and 2,115 of them having 2 units. Suppose the measurements y_j, j = 1, ... , 12,000, are from a Poisson distribution with mean 15. We simulated a sample of size n = 1,000 without replacement from such a population.

Let n_1 be the number of classes that occur once in the sample and let n_2 be the number of classes that occur twice in the sample. We obtained $n_1 = 968$, $n_2 = 16$, $\sum_{j=1}^{M} Z_j^{(1)} = 14,669$, $\sum_{j=1}^{M} Z_j^{(2)} = 56$. The unbiased estimate of $T = \sum_{j=1}^{M} y_j$ is $W = \frac{N}{n} \sum_{j=1}^{M} Z_j^{(1)} + \left[1 - \frac{(N-n+1)(N-n)}{n(n-1)}\right] \sum_{j=1}^{M} Z_j^{(2)} =$ 163,652. In this example, the measurements of y_j are actually random.

The expected value of T is 12,000 x 15 = 180,000. Using the expected value of the Poisson variables the variance of W is Var(W) = 89,166,177 and the standard deviation is 9,442.78. The relative standard deviation is 0.0577.

CHAPTER 3

GOOD AND TOULMIN'S METHOD

3.1 Introduction

In this chapter the sampling is done with replacement.

Good and Toulmin [7] considered the problem of sampling an infinite population and found an approximate relationship between $E[f_r(tn)]$ and $E[f_r]$ where f_r is the number of distinct classes which are represented exactly r times in the basic sample and $f_r(tn)$ is the number of distinct classes which are represented exactly r times in a second sample of size tn:

$$E[f_r(tn)] \simeq t^r \sum_{i=o}^{\infty} (-1)^i \binom{r+i}{r} (t-1)^i \, E\left(f_{r+i}\right) .$$

They they define an estimator of $E[f_r(tn)]$ by

$$\hat{f}_r(tn) = t^r \sum_{i=o}^{\infty} (-1)^i \binom{r+i}{r} (t-1)^i f_{r+i} \quad .$$

They use the approximation

$$Cov(f_r, f_s) \simeq \delta_{rs} E(f_r) - 2^{-r-s} \binom{r+s}{r} E\left[f_{r+s}(2n)\right]$$

to obtain

$$Var\left(\hat{f}_r(tn)\right) \simeq t^{2r} \left\{ \sum_{i=o}^{\infty} (t-1)^{2i} \binom{r+i}{r}^2 E\left(f_{r+i}\right) - \binom{2r}{r} (2t)^{-2r} E\left[f_{2r}(2tn)\right] \right\}$$

We generalize these derivations to obtain an approximate formula for $E\left[\sum_{j=1}^{M} Z_j^{(r)}(tn)\right]$ in terms of $E\left[\sum_{j=1}^{M} Z_j^{(r)}\right]$. From this we obtain an approximate formula for $E\left[\sum_{j=1}^{M} Y_j(tn)\right]$, which lead us to an estimator of $T = \sum_{j=1}^{M} y_j$. We also derive an approximate expression for the variance of this estimator.

3.2 Estimation of the Total Measurement T

Suppose that C_j is the jth class and d_i is the ith unit of the random sample. Hence

$$P_r\left\{d_i \in C_j\right\} = P_j > 0 \quad \text{for } j = 1, \ldots, M,\ i = 1, \ldots, n$$

$$\text{and } \sum_{j=1}^{M} P_j = 1.$$

Theorem 3.1: $$E\left[\sum_{j=1}^{M} Z_j^{(r)}(tn)\right] \approx t^r \sum_{i=o}^{I} (-1)^i (t-1)^i \binom{r+i}{r} E\left[\sum_{j=1}^{M} Z_j^{(r+i)}\right]$$

Where I is some integer such that $I \ll n-r$.

Proof: $$E\left[\sum_{j=1}^{M} Z_j^{(r)}(tn)\right] = \sum_{j=1}^{M} y_j \binom{tn}{r} P_j^r \left(1 - P_j\right)^{tn-r}$$

$$= \sum_{j=1}^{M} y_j \binom{tn}{r} P_j^r \left(1 - P_j\right)^{n-r} \left(1 + \frac{P_j}{1 - P_j}\right)^{-(t-1)n}$$

$$= \sum_{j=1}^{M} y_j \binom{tn}{r} P_j^r \left(1 - P_j\right)^{n-r} \sum_{i=o}^{\infty} \binom{-(t-1)n}{i} P_j^i \left(1 - P_j\right)^{-i}$$

$$= \sum_{i=o}^{\infty} \binom{tn}{r} \binom{-(t-1)n}{i} \sum_{j=1}^{M} y_j P_j^{r+i} \left(1 - P_j\right)^{n-(r+i)}$$

$$= \sum_{i=o}^{\infty} \frac{\binom{tn}{r}\binom{-(t-1)n}{i}}{\binom{n}{r+i}} E\left[\sum_{j=1}^{M} Z_j^{(r+i)}\right]$$

For $i \ll n-r$ we have $r+i \ll n$, and $i \ll (t-1)n$, so

$$\frac{\binom{tn}{r}\binom{-(t-1)n}{i}}{\binom{n}{r+i}} \simeq \frac{(tn)^r(-(t-1)n)^i(r+i)!}{r! \quad i! \quad n^{r+i}} = (-1)^i t^r (t-1)^i \binom{r+i}{r}$$

Hence, retaining only terms with $i \ll n-r$, we obtain

$$E\left[\sum_{j=1}^{M} Z_j^{(r)}(tn)\right] \simeq t^r \sum_{i=o}^{I} (-1)^i (t-1)^i \binom{r+i}{r} E\left[\sum_{j=1}^{M} Z_j^{(r+i)}\right] .$$

Corollary 3.1: $E\left[\sum_{j=1}^{M}\left(Z_j^{(r)}(tn)\right)^2\right] \simeq t^r \sum_{i=o}^{I} (-1)^i (t-1)^i \binom{r+i}{r} E\left[\sum_{j=1}^{M}\left(Z_j^{(r+i)}\right)^2\right]$

Proof: The same as that of Theorem 3.1.

Remark 3.1: (1) We define an estimator of $E\left[\sum_{j=1}^{M} Z_j^{(r)}(tn)\right]$ by

$$\widehat{\sum_{j=1}^{M} Z_j^{(r)}(tn)} = t^r \sum_{i=o}^{I} (-1)^i \binom{r+i}{r} (t-1)^i \left(\sum_{j=1}^{M} Z_j^{(r+i)}\right) .$$

(2) $E\left[\sum_{j=1}^{M} Y_j(tn)\right] = \sum_{j=1}^{M} y_j \left[1 - \left(1 - P_j\right)^{tn}\right] = \sum_{j=1}^{M} y_j$

$$- \sum_{j=1}^{M} y_j \left(1 - P_j\right)^{tn} \simeq \sum_{j=1}^{M} y_j \text{ for large } t$$

(3) $E\left[\sum_{j=1}^{M} Y_j(tn)\right] = E\left[\sum_{r=1}^{n} \sum_{j=1}^{M} Z_j^{(r)}(tn)\right] \simeq \sum_{r=1}^{n} t^r \sum_{i=o}^{I} (-1)^i \cdot$

$$(t-1)^i \binom{r+i}{r} E\left[\sum_{j=1}^{M} Z_j^{(r+i)}\right]$$

(4) Since $E\left[\sum_{j=1}^{M} Z_j^{(0)}(tn)\right] \simeq \sum_{i=o}^{I} (-1)^i (t-1)^i E\left[\sum_{j=1}^{M} Z_j^{(i)}\right]$,

$$E\left[\sum_{j=1}^{M} Y_j(tn)\right] = \sum_{r=1}^{n} E\left[\sum_{j=1}^{M} Z_j^{(r)}(tn)\right] = \sum_{j=1}^{M} y_j - E\left[\sum_{j=1}^{M} Z_j^{(0)}(tn)\right]$$

$$\simeq \sum_{j=1}^{M} y_j - E\left[\sum_{j=1}^{M} Z_j^{(0)}\right] - \sum_{i=1}^{I} (-1)^i (t-1)^i E\left[\sum_{j=1}^{M} Z_j^{(i)}\right]$$

$$= \sum_{r=1}^{n} E\left[\sum_{j=1}^{M} Z_j^{(r)}\right] - \sum_{i=1}^{I} (-1)^i (t-1)^i E\left[\sum_{j=1}^{M} Z_j^{(i)}\right] .$$

(5) Therefore, we can estimate $T = \sum_{j=1}^{M} y_j$ by

$\Sigma\hat{Y}_j(tn) = T_s - \sum_{i=1}^{I} (-1)^i (t-1)^i \left(\sum_{j=1}^{M} Z_j^{(i)}\right)$ when t is large.

However, the factor $(t-1)^i$ increases rapidly with i if $t > 2$ and attaches weight to terms for which $\sum_{j=1}^{M} Z_j^{(i)}$ is small. This is likely to produce a large percentage error when estimated from the basic sample. We follow Good and Toulmin in using a summation method to try to overcome this difficulty.

(6) In the case when the second sample is an enlargement of the basic one, the expectation of the new total measurement is approximately

$$(t-1) \sum_{j=1}^{M} Z_j^{(1)} - (t-1)^2 \sum_{j=1}^{M} Z_j^{(2)} + - \dots .$$

3.3 Variance of the Estimator of T

In this section we find the variance of

$$\sum_{j=1}^{M} \hat{Y}_j(tn) = T_s - \sum_{i=1}^{I} (-1)^i (t-1)^i \left(\sum_{j=1}^{M} Z_j^{(i)}\right) .$$

First, we find $\text{Cov}\left[\sum_{j=1}^{M} Z_j^{(r)}, \sum_{j=1}^{M} Z_j^{(s)}\right]$.

Theorem 3.2: For $rs \ll n$,

$$E\left[\left(\sum_{j=1}^{M} Z_j^{(r)}\right)\left(\sum_{j=1}^{M} Z_j^{(s)}\right)\right]$$

$$\simeq \delta_{rs} E\left[\sum_{j=1}^{M}\left(Z_j^{(r)}\right)^2\right] + E\left[\sum_{j=1}^{M} Z_j^{(r)}\right] E\left[\sum_{j=1}^{M} Z_j^{(s)}\right]$$

$$- \Sigma(-1)^u \frac{(r+s+u)!}{r!s!u!} E\left[\sum_{j=1}^{M}\left(Z_j^{(r+s+u)}\right)^2\right]$$

$$\simeq \delta_{rs} E\left[\sum_{j=1}^{M}\left(Z_j^{(r)}\right)^2\right] + E\left[\sum_{j=1}^{M} Z_j^{(r)}\right] E\left[\sum_{j=1}^{M} Z_j^{(s)}\right]$$

$$- 2^{-r-s}\frac{(r+s)!}{r!s!} E\left[\sum_{j=1}^{M}\left(Z_j^{(r+s)}(2n)\right)^2\right]$$

$$\text{where } \delta_{rs} = \begin{cases} 1 & \text{if } r = s \\ 0 & \text{otherwise} \end{cases}$$

Proof: $$E\left[\left(\sum_{j=1}^{M} Z_j^{(r)}\right)\left(\sum_{j=1}^{M} Z_j^{(s)}\right)\right] = E\left[\left(\sum_{j=1}^{M} y_j I_{\{r\}}(X_j)\right)\left(\sum_{j=1}^{M} y_j I_{\{s\}}(X_j)\right)\right]$$

$$= \sum_{j=1}^{M} \sum_{k=1}^{M} y_j y_k E\left[I_{\{r\}}(X_j) I_{\{s\}}(X_k)\right]$$

$$= \sum_{j=k} y_j^2 E\left[I_{\{r\}}(X_j) I_{\{s\}}(X_k)\right] + \sum_{j \neq k}\Sigma y_j y_k E\left[I_{\{r\}}(X_j) I_{\{s\}}(X_k)\right]$$

$$= \delta_{rs} \sum_{j=1}^{M} y_j^2 E\left[I_{\{r\}}(X_j)\right] + \sum_{j \neq k}\Sigma y_j y_k E\left[I_{\{r\}}(X_j) I_{\{s\}}(X_k)\right]$$

$$\text{where } \delta_{rs} = \begin{cases} 1 & \text{if } r=s \\ 0 & \text{if } r \neq s \end{cases}$$

$$= \delta_{rs} E\left[\sum_{j=1}^{M}\left(Z_j^{(r)}\right)^2\right] + \sum_{j \neq k}\sum y_j y_k \frac{n!}{r!s!(n-r-s)!} P_j^{\,r} P_k^{\,s}\left(1 - P_j - P_k\right)^{n-r-s}$$

$$= \delta_{rs} E\left[\sum_{j=1}^{M}\left(Z_j^{(r)}\right)^2\right] + \frac{n!}{r!s!(n-r-s)!}\left[\sum_j \sum_k y_j y_k P_j^{\,r} P_k^{\,s}\left(1 - P_j - P_k\right)^{n-r-s}\right.$$

$$\left. - \sum_j y_j^{\,2} P_j^{\,r+s}\left(1 - 2P_j\right)^{n-r-s}\right]$$

$$= \delta_{rs} E\left[\sum_{j=1}^{M}\left(Z_j^{(r)}\right)^2\right] + \frac{n!}{r!s!(n-r-s)!}\left[\sum_j \sum_k y_j y_k P_j^{\,r} P_k^{\,s}\left(\sum_{u=o}^{s}\binom{s}{u} P_j^{\,u}\right.\right. \cdot$$

$$\left.\left(1 - P_j\right)^{n-r-u}\right)\left(\sum_{v=o}^{r}\binom{r}{v} P_k^{\,v}\left(1 - P_k\right)^{n-s-v}\right) \cdot \left(\sum_{w=o}^{n-r-s}(-1)^w\right. \cdot$$

$$\left.\binom{n-r-s}{w} P_j^{\,w}\left(1 - P_j\right)^{-w} P_k^{\,w}\left(1 - P_k\right)^{-w}\right) - \sum_j y_j^{\,2} P_j^{\,r+s} \cdot \left(\sum_{u=o}^{n-r-s}\right. \cdot$$

$$\left.\left.(-1)^u \binom{n-r-s}{u} P_j^{\,u}\left(1 - P_j\right)^{n-r-s-u}\right)\right] \quad \text{by (26), (27) of [8]}$$

$$= \delta_{rs} E\left[\sum_{j=1}^{M}\left(Z_j^{(r)}\right)^2\right] + \frac{n!}{r!s!(n-r-s)!}\left[\sum_{u,v,w}(-1)^w \binom{s}{u}\binom{r}{v}\binom{n-r-s}{w}\right.$$

$$\cdot \sum_j \sum_k y_j y_k P_j^{\,r+u+w}\left(1 - P_j\right)^{n-r-u-w} P_k^{\,s+v+w}\left(1 - P_k\right)^{n-s-v-w}$$

$$\left. - \sum_{u=o}^{n-r-s}(-1)^u \binom{n-r-s}{u} \sum_j y_j^{\,2} P_j^{\,r+s+u}\left(1 - P_j\right)^{n-r-s-u}\right]$$

$$= \delta_{rs} E\left[\sum_{j=1}^{M}\left(Z_j^{(r)}\right)^2\right] + \frac{n!}{r!s!(n-r-s)!}\left[\sum_{u,v,w}(-1)^w \binom{s}{u}\binom{r}{v}\binom{n-r-s}{w}\right. \cdot$$

$$\left(\sum_j y_j P_j^{\,r+u+w}\left(1 - P_j\right)^{n-r-u-w}\right)\left(\sum_k y_k P_k^{\,s+v+w}\left(1 - P_k\right)^{n-s-v-w}\right)$$

$$\left. - \sum_{u=o}^{n-r-s}(-1)^u \binom{n-r-s}{u} \sum_j y_j^{\,2} P_j^{\,r+s+u}\left(1 - P_j\right)^{n-r-s-u}\right]$$

$$= \delta_{rs} E\left[\sum_{j=1}^{M}\left(Z_j^{(r)}\right)^2\right] + \frac{n!}{r!s!(n-r-s)!}\left[\sum_{u,v,w} \frac{(-1)^w \binom{s}{u}\binom{r}{v}\binom{n-r-s}{w}}{\binom{n}{r+u+w}\binom{n}{s+v+w}}\right.$$

$$\left. \cdot E\left[\sum_{j=1}^{M} Z_j^{(r+u+w)}\right] E\left[\sum_{j=1}^{M} Z_j^{(s+v+w)}\right] - \sum_{u=o}^{n-r-s} \frac{(-1)^u \binom{n-r-s}{u}}{\binom{n}{r+s+u}} \cdot\right.$$

$$E\left[\sum_{j=1}^{M}\left(Z_j^{(r+s+u)}\right)^2\right]$$

$$= \delta_{rs} E\left[\sum_{j=1}^{M}\left(Z_j^{(r)}\right)^2\right] + \sum_{u,v,w} (-1)^w \frac{(n-r-u-w)!(n-s-v-w)!(r+u+w)!(s+v+w)!}{(n-r-s-w)!n!u!v!w!(s-u)!(r-v)!}$$

$$\cdot E\left[\sum_{j=1}^{M} Z_j^{(r+u+w)}\right] E\left[\sum_{j=1}^{M} Z_j^{(s+v+w)}\right] - \sum_u (-1)^u \frac{(r+s+u)!}{r!s!u!}$$

$$\cdot E\left[\sum_{j=1}^{M}\left(Z_j^{(r+s+u)}\right)^2\right]$$

if u, v, w, r, s are all << n, then the coefficient in the first sum is $O((rs/n)^{u+v+w})$ and when u=v=w=0, use of Stirling's formula shows that it is $1+O(rs/n)$. Hence if rs << n it is proved.

Remark 3.2:

$$\mathrm{Cov}\left(\sum_{j=1}^{M} Z_j^{(r)}, \sum_{j=1}^{M} Z_j^{(s)}\right) \simeq \delta_{rs} E\left[\sum_{j=1}^{M}\left(Z_j^{(r)}\right)^2\right] - \sum_u (-1)^u$$

$$\frac{(r+s+u)!}{r!s!u!} E\left[\sum_{j=1}^{M}\left(Z_j^{(r+s+u)}\right)^2\right]$$

$$\simeq \delta_{rs} E\left[\sum_{j=1}^{M}\left(Z_j^{(r)}\right)^2\right] - 2^{-r-s}\binom{r+s}{r} E\left[\sum_{j=1}^{M}\left(Z_j^{(r+s)}(2n)\right)^2\right]$$

Theorem 3.3: $$\mathrm{Var}\left[\sum_{j=1}^{M} \hat{Z}_j^{(r)}(tn)\right] \simeq t^{2r}\left\{\sum_{i=o}^{I}(t-1)^{2i}\binom{r+i}{r}^2 E\left[\sum_{j=1}^{M}\left(Z_j^{(r+i)}\right)^2\right] - \binom{2r}{r}(2t)^{-2r}E\left[\sum_{j=1}^{M}\left(Z_j^{(2r)}(2tn)\right)^2\right]\right\}$$

where I is an integer such than I << n-r.

Proof: $$\mathrm{Var}\left[\sum_{j=1}^{M} \hat{Z}_j^{(r)}(tn)\right] = \mathrm{Var}\left[t^r\sum_{i=o}^{\infty}(-1)^i\binom{r+i}{r}(t-1)^i\left(\sum_{j=1}^{M}Z_j^{(r+i)}\right)\right]$$

$$= t^{2r}\left\{\sum_{i,k=o}^{\infty}(-1)^{i+k}(t-1)^{i+k}\binom{r+i}{r}\binom{r+k}{r}\mathrm{Cov}\left(\sum_{j=1}^{M}Z_j^{(r+i)}, \sum_{j=1}^{M}Z_j^{(r+k)}\right)\right\}$$

$$\simeq t^{2r}\left\{\sum_{i,k=o}^{\infty}(-1)^{i+k}(t-1)^{i+k}\binom{r+i}{r}\binom{r+k}{r}\left\{\delta_{ik}E\left[\sum_{j=1}^{M}\left(Z_j^{(r+i)}\right)^2\right] - 2^{-2r-i-k}\binom{2r+i+k}{r+i}E\left[\sum_{j=1}^{M}\left(Z_j^{(2r+i+k)}(2n)\right)^2\right]\right\}\right\}$$

$$= t^{2r}\left\{\sum_{i=o}^{\infty}(t-1)^{2i}\binom{r+i}{r}^2 E\left[\sum_{j=1}^{M}\left(Z_j^{(r+i)}\right)^2\right] - \sum_{\ell=o}^{\infty}(-1)^{\ell}(t-1)^{\ell}2^{-2r-\ell}E\left[\sum_{j=1}^{M}\left(Z_j^{(2r+\ell)}(2n)\right)^2\right]\frac{(2r+\ell)!}{\ell!r!r!}\cdot\sum_{i,k=o}^{i+k=\ell}\frac{(i+k)!}{i!k!}\right\}$$

$$= t^{2r}\left\{\sum_{i=o}^{\infty}(t-1)^{2i}\binom{r+i}{r}^2 E\left[\sum_{j=1}^{M}\left(Z_j^{(r+i)}\right)^2\right] - \sum_{\ell=o}^{\infty}(-1)^{\ell}(t-1)^{\ell}2^{-2r}\frac{(2r+\ell)!}{\ell!r!r!}E\left[\sum_{j=1}^{M}\left(Z_j^{(2r+\ell)}(2n)\right)^2\right]\right\}$$

$$\simeq t^{2r}\left\{\sum_{i=o}^{\infty}(t-1)^{2i}\binom{r+i}{r}^2 E\left[\sum_{j=1}^{M}\left(Z_j^{(r+i)}\right)^2\right]\right.$$

$$- (2t)^{-2r}\binom{2r}{r} E\left[\sum_{j=1}^{M}\left(Z_j^{(2r)}(2tn)\right)^2\right]\Bigg\}$$

Remark 3.3:

Since $\Sigma\hat{Y}_j(tn) = \Sigma y_j - \sum_{j=1}^{M} \hat{Z}_j^{(0)}(tn)$,

$$\mathrm{Var}\left(\sum_{j=1}^{M} \hat{Y}_j(tn)\right) = \mathrm{Var}\left(\sum_{j=1}^{M} \hat{Z}_j^{(0)}(tn)\right)$$

$$\simeq \sum_{i=0}^{\infty} (t-1)^{2i} E\left[\sum_{j=1}^{M}\left(Z_j^{(i)}\right)^2\right] - E\left[\sum_{j=1}^{M}\left(Z_j^{(0)}(2tn)\right)^2\right]$$

$$= \sum_{i=0}^{\infty} (t-1)^{2i} E\left[\sum_{j=1}^{M}\left(Z_j^{(i)}\right)^2\right] - \sum_{i=0}^{\infty} (-1)^i (2t-1)^i E\left[\sum_{j=1}^{M}\left(Z_j^{(i)}\right)^2\right]$$

$$= \sum_{i=1}^{\infty} (t-1)^{2i} E\left[\sum_{j=1}^{M}\left(Z_j^{(i)}\right)^2\right] - \sum_{i=1}^{\infty} (-1)^i (2t-1)^i E\left[\sum_{j=1}^{M}\left(Z_j^{(i)}\right)^2\right].$$

3.4 Summation of the Series

Euler's transformation with parameter q, generally called the (E, q) method, is a method of forcing series like $\sum_{i=1}^{\infty} (-1)^i (t-1)^i E\left[\sum_{j=1}^{M}\left(Z_j^{(i)}\right)\right]$, $\sum_{i=1}^{\infty} (t-1)^{2i} E\left[\sum_{j=1}^{M}\left(Z_j^{(i)}\right)^2\right]$, $\sum_{i=1}^{\infty} (-1)^i (2t-1)^i E\left[\sum_{j=1}^{M}\left(Z_j^{(i)}\right)^2\right]$, etc. to converge rapidly. This is to transform the series $\sum_{i=0}^{\infty} a_i$ into $\sum_{j=0}^{\infty} a_j^{(q)}$

where $a_j^{(q)} = \dfrac{1}{(q+1)^{j+1}} \sum_{i=0}^{j} \binom{j}{i} q^{j-i} a_i$.

First consider $\sum_{i=1}^{\infty} (-1)^i (t-1)^i E\left[\sum_{j=1}^{M} Z_j^{(i)}\right]$. In our example,

$E\left[\sum_{j=1}^{M} Z_j^{(r)}\right]$ generally decreases slowly for $r \geqq 2$ and so we will write

$$\sum_{i=1}^{\infty} (-1)^i (t-1)^i E\left[\sum_{j=1}^{M} Z_j^{(i)}\right] \simeq -(t-1)E\left[\sum_{j=1}^{M} Z_j^{(1)}\right]$$

$+ E\left[\sum_{j=1}^{M} Z_j^{(2)}\right](t-1)^2 \sum_{i=0}^{\infty} (-1)^i (t-1)^i$. We apply the (E, q) method to

$\sum_{i=0}^{\infty} (-1)^i (t-1)^i$. Define $a_i = (-1)^i (t-1)^i$

$$a_j^{(q)} = \frac{1}{(q+1)^{j+1}} \sum_{i=0}^{j} \binom{j}{i} q^{j-i} a_i = \frac{1}{q+1} \left(\frac{q-(t-1)}{q+1}\right)^j$$

$$\sum_{j=0}^{\infty} a_j^{(q)} = \frac{1}{t} .$$

Hence $\sum_{i=1}^{\infty} (-1)^i (t-1)^i E\left[\sum_{j=1}^{M} Z_j^{(i)}\right] \simeq -(t-1)E\left[\sum_{j=1}^{M} Z_j^{(1)}\right] + \frac{(t-1)^2}{t} E\left[\sum_{j=1}^{M} Z_j^{(2)}\right]$.

Remark 3.4:

Recall the estimator $\sum_{j=1}^{M} \hat{Y}_j(tn)$ in Remark 3.1.(5). The summation in that expression has upper limit I. Let us, however, change the upper limit to ∞ and then use Euler's transformation to obtain

$$E\left[\sum_{j=1}^{M} \hat{Y}_j(tn)\right] \simeq \sum_{r=1}^{n} E\left[\sum_{j=1}^{M} Z_j^{(r)}\right] + (t-1)E\left[\sum_{j=1}^{M} Z_j^{(1)}\right]$$

$$- \frac{(t-1)^2}{t} E\left[\sum_{j=1}^{M} Z_j^{(2)}\right] .$$

We previously argued that $\sum_{j=1}^{M} \hat{Y}_j(tn)$ is a reasonable estimator of T when t is large, say $t = \frac{N}{n}$. We now see that another expression for a reasonable estimator of T is

$$\sum_{j=1}^{M} \tilde{Y}_j(tn) = \sum_{r=1}^{n} \sum_{j=1}^{M} Z_j^{(r)} + \left(\frac{N}{n} - 1\right) \sum_{j=1}^{M} Z_j^{(1)} - \frac{\left(\frac{N}{n} - 1\right)^2}{\frac{N}{n}} \sum_{j=1}^{M} Z_j^{(2)} .$$

If $\sum_{j=1}^{M} Z_j^{(r)} = 0$ for $r \geq 2$ (this is nearly true in many examples), then

$\sum_{j=1}^{M} \tilde{Y}_j(N) = \sum_{j=1}^{M} \hat{Y}_j(N) = \frac{N}{n} T_s'$, which is the natural estimator of the

population total when there is no duplication.

To obtain an approximate expression for the variance of

$\sum_{j=1}^{M} \hat{Y}_j(tn)$, now consider $\sum_{i=1}^{\infty} (t-1)^{2i} E\left[\sum_{j=1}^{M} \left(Z_j^{(i)}\right)^2\right]$ and $\sum_{i=1}^{\infty} (-1)^i (2t-1)^i E\left[\sum_{j=1}^{M} \left(Z_j^{(i)}\right)^2\right]$.

In our examples, $E\left[\sum_{j=1}^{M} \left(Z_j^{(i)}\right)^2\right]$ is nearly constant for $r \geqq 2$, and so we write

$$\sum_{i=1}^{\infty} (t-1)^{2i} E\left[\sum_{j=1}^{M} \left(Z_j^{(i)}\right)^2\right] = (t-1)^2 E\left[\sum_{j=1}^{M} \left(Z_j^{(1)}\right)^2\right] + E\left[\sum_{j=1}^{M} \left(Z_j^{(2)}\right)^2\right] \sum_{i=2}^{\infty} (t-1)^{2i} .$$

Applying the (E, q) method to $\sum_{i=2}^{\infty} (t-1)^{2i}$, we obtain

$$\sum_{i=1}^{\infty} (t-1)^{2i} E\left[\sum_{j=1}^{M} \left(Z_j^{(i)}\right)^2\right] \simeq (t-1)^2 E\left[\sum_{j=1}^{M} \left(Z_j^{(1)}\right)^2\right] +$$

$$\frac{(t-1)^2}{1 - (t-1)^2} E\left[\sum_{j=1}^{M} \left(Z_j^{(2)}\right)^2\right] .$$

Also, we can write

$$\sum_{i=1}^{\infty} (-1)^i (2t-1)^i E\left[\sum_{j=1}^{M} \left(Z_j^{(i)}\right)^2\right] = -(2t-1) E\left[\sum_{j=1}^{M} \left(Z_j^{(1)}\right)^2\right]$$

$$+ E\left[\sum_{j=1}^{M} Z_j^{(2)}\right] \sum_{i=2}^{\infty} (-1)^i (2t-1)^i \quad .$$

Applying the (E, q) method to $\sum_{i=2}^{\infty}(-1)^i(2t-1)^i$, we obtain

$$\sum_{i=1}^{\infty}(-1)^i(2t-1)^i E\left[\sum_{j=1}^{M}\left(Z_j^{(i)}\right)^2\right] \simeq -(2t-1)E\left[\sum_{j=1}^{M}\left(Z_j^{(1)}\right)^2\right]$$

$$+\frac{(2t-1)^2}{t}E\left[\sum_{j=1}^{M}\left(Z_j^{(2)}\right)^2\right].$$

Remark 3.5:

Using Euler's transformation

$$\mathrm{Var}\left[\sum_{j=1}^{M}\hat{Y}_j(tn)\right] \simeq t^2E\left[\sum_{j=1}^{M}\left(Z_j^{(1)}\right)^2\right] + \frac{4t^2-10t+5}{2(2-t)}E\left[\sum_{j=1}^{M}\left(Z_j^{(2)}\right)^2\right].$$

To obtain an approximate expression for variance of $\sum_{j=1}^{M}\tilde{Y}_j(tn)$, now consider $\sum_{i=0}^{\infty}(-1)^i\binom{2+i}{2}(t-1)^iE\left[\sum_{j=1}^{M}\left(Z_j^{(2+i)}\right)^2\right]$ and $\sum_{i=0}^{\infty}(-1)^i\binom{1+i}{1}(t-1)^iE\left[\sum_{j=1}^{M}\left(Z_j^{(1+i)}\right)^2\right]$. In our example, $E\left[\sum_{j=1}^{M}\left(Z_j^{(r)}\right)^2\right]$, $\binom{1+i}{1}$, and $\binom{2+i}{2}$ generally decrease slowly and so we write

$$\sum_{i=0}^{\infty}(-1)^i\binom{2+i}{2}(t-1)^iE\left[\sum_{j=1}^{M}\left(Z_j^{(2+i)}\right)^2\right] \simeq E\left[\sum_{j=1}^{M}\left(Z_j^{(2+i)}\right)^2\right]\sum_{i=0}^{\infty}(-1)^i(t-1)^i.$$

and

$$\sum_{i=0}^{\infty}(-1)^i\binom{1+i}{1}(t-1)^iE\left[\sum_{j=1}^{M}\left(Z_j^{(1+i)}\right)^2\right] \simeq E\left[\sum_{j=1}^{M}Z_j^{(1)}\right]$$

$$-\ 2(t-1)E\left[\sum_{j=1}^{M}\left(Z_j^{(2)}\right)^2\right]\sum_{i=0}^{\infty}(-1)^i(t-1)^i.$$

Applying the (E, q) method to $\sum_{i=0}^{\infty}(-1)^i(t-1)^i$, we obtain

$$\sum_{i=0}^{\infty} (-1)^i \,{}^{2+i}_{\;\;2}\, (t-1)^i E \sum_{j=1}^{M} Z_j^{(2+i)\;2}$$

$$\approx \frac{1}{t} E \sum_{j=1}^{M} Z_j^{(2)\;2}, \text{ and}$$

$$\sum_{i=0}^{\infty} (-1)^i \,{}^{1+i}_{\;\;1}\, (t-1)^i E \sum_{j=1}^{M} Z_j^{(1+i)\;2} \approx E \sum_{j=1}^{M} Z_j^{(1)} - \frac{2(t-1)}{t} E \sum_{j=1}^{M} Z_j^{(2)} .$$

Remark 3.6:

$$\mathrm{Var}\left[\sum_{j=1}^{M} \tilde{Y}_j(tn)\right] = \mathrm{Var}\left[\sum_{j=1}^{n} Z_j^{(0)}\right] + (t-1)^2\mathrm{Var}\left[\sum_{j=1}^{M} Z_j^{(1)}\right] + \frac{(t-1)^4}{t^2}\mathrm{Var}\left[\sum_{j=1}^{M} Z_j^{(2)}\right]$$

$$- 2(t-1)\mathrm{Cov}\left[\sum_{j=1}^{M} Z_j^{(0)}, \sum_{j=1}^{M} Z_j^{(1)}\right] + 2\frac{(t-1)^2}{t}\mathrm{Cov}\left[\sum_{j=1}^{M} Z_j^{(0)}, \sum_{j=1}^{M} Z_j^{(2)}\right]$$

$$- 2\frac{(t-1)^3}{t}\mathrm{Cov}\left[\sum_{j=1}^{M} Z_j^{(1)}, \sum_{j=1}^{M} Z_j^{(2)}\right] .$$

Without considering Euler's transformation we obtain

$$\mathrm{Var}\left[\sum_{j=1}^{M} Z_j^{(0)}\right] \approx - \sum_{i=1}^{\infty} (-1)^i E\left[\sum_{j=1}^{M}\left(Z_j^{(i)}\right)^2\right]$$

$$\mathrm{Var}\left[\sum_{j=1}^{M} Z_j^{(1)}\right] \approx E\left[\sum_{j=1}^{M}\left(Z_j^{(1)}\right)^2\right] - 2 \sum_{i=0}^{\infty} (-1)^i \binom{2+i}{2} E\left[\sum_{j=1}^{M}\left(Z_j^{(2+i)}\right)^2\right]$$

$$\mathrm{Var}\left[\sum_{j=1}^{M} Z_j^{(2)}\right] \approx E\left[\sum_{j=1}^{M}\left(Z_j^{(2)}\right)^2\right] - 6 \sum_{i=0}^{\infty} (-1)^i \binom{4+i}{4} E\left[\sum_{j=1}^{M}\left(Z_j^{(4+i)}\right)^2\right]$$

$$\mathrm{Cov}\left[\sum_{j=1}^{M} Z_j^{(0)}, \sum_{j=1}^{M} Z_j^{(1)}\right] \approx - \sum_{i=0}^{\infty} (-1)^i (i+1) E\left[\sum_{j=1}^{M}\left(Z_j^{(i+1)}\right)^2\right]$$

$$\mathrm{Cov}\left[\sum_{j=0}^{M} Z_j^{(0)}, \sum_{j=1}^{M} Z_j^{(2)}\right] \approx - \sum_{i=0}^{\infty} (-1)^i \binom{2+i}{2} E\left[\sum_{j=1}^{M}\left(Z_j^{(2+i)}\right)^2\right]$$

$$\mathrm{Cov}\left[\sum_{j=1}^{M} Z_j^{(1)}, \sum_{j=1}^{M} Z_j^{(2)}\right] \approx - 3 \sum_{i=0}^{\infty} (-1)^i \binom{3+i}{3} E\left[\sum_{j=1}^{M}\left(Z_j^{(3+i)}\right)^2\right] .$$

With the use of Euler's transformation we obtain

$$\mathrm{Var}\left[\sum_{j=1}^{M} Z_j^{(0)}\right] \approx E\left[\sum_{j=1}^{M}\left(Z_j^{(1)}\right)^2\right] - E\left[\sum_{j=1}^{M}\left(Z_j^{(2)}\right)^2\right]$$

$$\mathrm{Var}\left[\sum_{j=1}^{M} Z_j^{(1)}\right] \approx E\left[\sum_{j=1}^{M}\left(Z_j^{(1)}\right)^2\right] - E\left[\sum_{j=1}^{M}\left(Z_j^{(2)}\right)^2\right]$$

$$\mathrm{Var}\left[\sum_{j=1}^{M} Z_j^{(2)}\right] \approx E\left[\sum_{j=1}^{M} Z_j^{(2)}\right] - 6\sum_{i=0}^{\infty}(-1)^i\binom{4+i}{4} E\left[\sum_{j=1}^{M}\left(Z_j^{(4+i)}\right)^2\right]$$

$$\mathrm{Cov}\left[\sum_{j=1}^{M} Z_j^{(0)}, \sum_{j=1}^{M} Z_j^{(1)}\right] \approx -E\left[\sum_{j=1}^{M}\left(Z_j^{(1)}\right)^2\right] + E\left[\sum_{j=1}^{M}\left(Z_j^{(2)}\right)^2\right]$$

$$\mathrm{Cov}\left[\sum_{j=1}^{M} Z_j^{(0)}, \sum_{j=1}^{M} Z_j^{(2)}\right] \approx -\frac{1}{2} E\left[\sum_{j=1}^{M}\left(Z_j^{(2)}\right)^2\right]$$

$$\mathrm{Cov}\left[\sum_{j=1}^{M} Z_j^{(1)}, \sum_{j=1}^{M} Z_j^{(2)}\right] \approx -3\sum_{i=0}^{\infty}(-1)^i\binom{3+i}{3} E\left[\sum_{j=1}^{M}\left(Z_j^{(3+i)}\right)^2\right].$$

3.5 Example

Consider a list of size N = 14,115 with M = 12,000 distinct classes, 9,885 of them having 1 unit and 2,115 of them having 2 units. Suppose the measurements y_j, j = 1, ... , 12,000, are from a Poisson distribution with mean 15. We simulated a sample of size n = 1,000 with replacement such a population.

Let n_1 be the number of classes that occur once in the sample, let n_2 be the number of classes that occur twice in the sample, and let n_3 be the number of classes that occur three times in the sample. We obtained n_1 = 900, n_2 = 47, n_3 = 2, $\sum_{j=1}^{M} Z_j^{(1)}$ = 13,461, $\sum_{j=1}^{M} Z_j^{(2)}$ = 671, $\sum_{j=1}^{M} Z_j^{(3)}$ = 33, $\sum_{j=1}^{M}\left(Z_j^{(1)}\right)^2$ = 214,613, $\sum_{j=1}^{M}\left(Z_j^{(2)}\right)^2$ = 10,157, and

$\sum_{j=1}^{M}\left(Z_j^{(3)}\right)^2 = 549.$ By remark 3.1.(5),

$$\sum_{j=1}^{M} \hat{Y}_j(tn) = 33t^3 - 770t^2 + 14902t - 66 \text{ (see Figure 3.1)}$$

$$= 149{,}734 \text{ when } t = {}^N/_n = 14.115.$$

Therefore, we obtain the estimate of $T = \sum_{j=1}^{M} y_j$ is 149,735 without considering Euler's transformation. If its variance is obtained by Remark 3.3 (i.e. without using Euler's transformation), then

$\hat{Var}\left[\sum_{j=1}^{M} \hat{Y}_j(N)\right] = 3{,}138{,}255{,}014.82$, its standard deviation is 56,020.13 and its relative standard deviation is .3741. If its variance is obtained by Remark 3.5 (i.e. using Euler's transformation), then

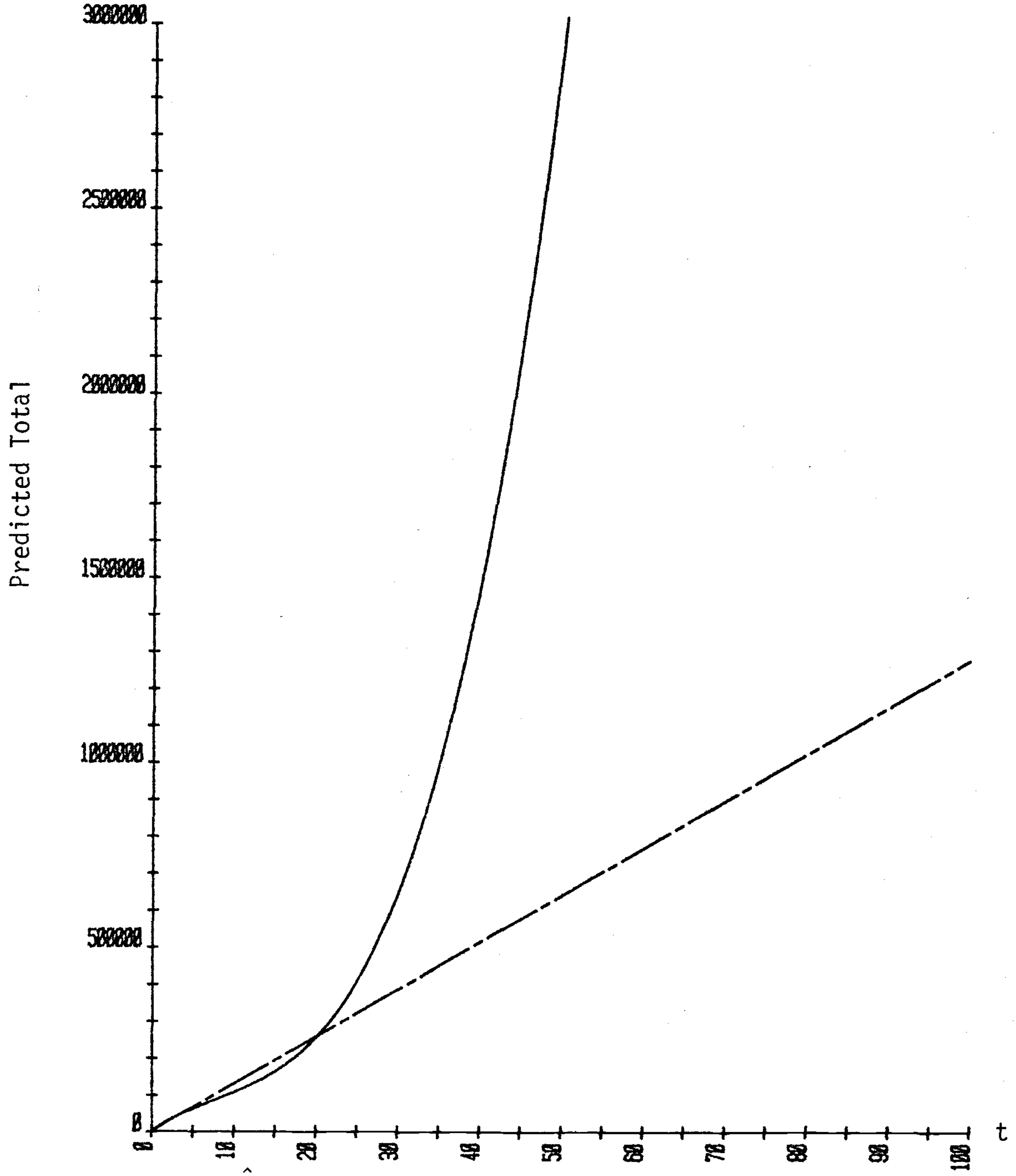
$\hat{Var}\left[\sum_{j=1}^{M} \hat{Y}_j(N)\right] = 42{,}481{,}045.82$, its standard deviation is 6,517.75 and its relative standard deviation is .0435. Using Remark 3.4 we obtain

$$\sum_{j=1}^{M} \tilde{Y}_j(tn) = 12{,}790t - 671/t + 2046 \quad \text{(see Figure 3.1)}$$

$$= 182{,}529 \text{ when } t = {}^N/_n = 14.115.$$

Therefore, we obtain the estimate of $T = \sum_{j=1}^{M} y_j$ is 182,529 with Euler's transformation. Using Remark 3.6 without using Euler's transformation, we find that the variance of the estimates is 41,158,599.42, its standard deviation is 6,415.50, and its relative standard deviation is .0351. Using Euler's transformation we find its variance is 42,645,357.32, its standard deviation is 6530.34, and its relative standard deviation is .0358.

Figure 3.1



——— : $\sum_{j=1}^{M} \hat{Y}_j(1000t)$ is the prediction of $\sum_{j=1}^{M} Y_j(1000t)$ without Euler's transformation

— - — : $\sum_{j=1}^{M} \tilde{Y}_j(1000t)$ is the prediction of $\sum_{j=1}^{M} Y_j(1000t)$ with Euler's transformation

This figure shows the predicted population totals with and without Euler's transformation based on a sample of size 1000 where the Y_j's are from a Poisson distribution with mean 15.

CHAPTER 4

HARRIS' METHOD

4.1 Introduction

In this chapter samples are taken with replacement.

In Chapter 3 we found that the estimator of $\sum_{j=1}^{M} y_j$ using Euler's transformation gives a reasonably good answer in our examples. Harris [10] gives us a check on the accuracy of this estimator. His approach offers approximations of the supremum and infimum of $E\left[\sum_{j=1}^{M} Y_j(tn)\right]$ which for large t is approximately equal to $T = \sum_{j=1}^{M} y_j$. If an estimate of T falls wihtin these bounds, we can regard it as reasonable (from this rather conservative viewpoint).

Define d to be the number of distinct classes observed in the sample and d(tn) to be the number of distinct classes which would be observed in a second sample of size tn. Harris [10] showed

$$E[d(tn)] \simeq E(d) + E(f_1)\int_0^{\infty} \frac{1-e^{-(t-1)x}}{x}\, d\,G(x)$$

and

$$\int x^r dG(x) \simeq \frac{(r+1)!E(f_{r+1})}{E(f_1)}$$

where f_r is as in Section 3.1 and G is a constructed cumulative distribution function. Harris computed the supremum and infimum of E[d(tn)] taken over all cumulative distribution functions whose first k moments are specified by $\int x^r dG(x)$.

Now we generalize his computations to obtain the supremum and infimum of $E\left[\sum_{j=1}^{M} Y_j(tn)\right]$.

4.2 Derivations

Lemma 4.1: For large n we have

$$\text{(i)}\quad E[T_s] = \sum_{j=1}^{M} y_j\left[1 - \left(1 - P_j\right)^n\right] \simeq \sum_{j=1}^{M} y_j\left[1 - e^{-nP_j}\right],$$

and

$$\text{(ii)}\quad E\left[\sum_{j=1}^{M} Z_j^{(r)}\right] = \sum_{j=1}^{M} y_j \binom{n}{r} P_j^{\,r}\left(1 - P_j\right)^{n-r} \simeq \sum_{j=1}^{M} y_j \frac{(nP_j)^r e^{-nP_j}}{r!}.$$

Proof:

$$\text{(i)}:\quad \left|\frac{\sum_{j=1}^{M} y_j\left[1 - \left(1 - P_j\right)^n\right] - \sum_{j=1}^{M} y_j\left[1 - e^{-nP_j}\right]}{\sum_{j=1}^{M} y_j\left[1 - e^{-nP_j}\right]}\right|$$

$$\leqq \sup_j \frac{y_j\left[e^{-nP_j} - \left(1 - P_j\right)^n\right]}{y_j\left[1 - e^{-nP_j}\right]}$$

$$= \sup_j \frac{e^{-nP_j} - \left(1 - P_j\right)^n}{1 - e^{-nP_j}}$$

By Harris' proof on p. 545 [10], we know

$$\sup_j \frac{e^{-nP_j} - \left(1 - P_j\right)^n}{1 - e^{-nP_j}} \to 0 \quad \text{as } n \to \infty$$

(ii) As stated by Harris, $\binom{n}{r} \simeq \frac{n^r}{r!} \exp\left[-\frac{r(r-1)}{2n}\right]$ and

$(1 - P)^{n-r} \simeq \exp\left[-(n-r)P - \frac{(n-r)P^2}{2}\right]$ for $P < 1$.

Hence, we have

$$\sum_{j=1}^{M} \frac{y_j (nP_j)^r e^{-nP_j}}{r!} - \sum_{j=1}^{M} y_j \binom{n}{r} P_j^{\,r} (1 - P_j)^{n-r}$$

$$= \sum_{j=1}^{M} \frac{y_j (nP_j)^r e^{-nP_j}}{r!} - \sum_{j=1}^{M} y_j \frac{n^r e^{\frac{-r(r-1)}{2n}} P_j^r e^{-(n-r)P_j - (n-r)P_j^2}}{r!}$$

$$= \sum_{j=1}^{M} \frac{y_j (nP_j)^r e^{-nP_j}}{r!} \left\{ 1 - \exp\left[rP_j - \frac{r(r-1)}{2n} - \frac{(n-r)}{2} P_j^2 - \cdots\cdots \right] \right\}$$

(a) If $P \geqq {}^1/_{n^{2/3}}$, then

$$\sum_{P_j \geqq {}^1/_{n^{2/3}}} \frac{y_j (nP_j)^r}{r!} e^{-nP_j} \left\{ 1 - \exp\left[rP_j - \frac{r(r-1)}{2n} - \frac{(n-r)}{2} P_j^2 - \cdots\cdots \right] \right\}$$

$$\leq \sum_{P_j \geqq {}^1/_{n^{2/3}}} \frac{y_j (nP_j)^r}{r!} e^{-nP_j} \leq \frac{\left(\max_j y_j\right) n^{\frac{r+2}{3}} e^{-n^{1/3}}}{r!} \rightarrow 0$$

as $n \rightarrow \infty$.

(b) If $P < {}^{1}/_{n^{2/3}}$, then

$$\frac{\displaystyle\sum_{P_j < {}^{1}/_{n^{2/3}}} \frac{y_j (nP_j)^r e^{-nP_j}}{r!}\left\{1 - \exp\left[rP_j - \frac{r(r-1)}{2n} - \frac{(n-r)}{2} P_j^2 - \ldots\right]\right\}}{\displaystyle\sum_{P_j < {}^{1}/_{n^{2/3}}} \frac{y_j (nP_j)^r e^{-nP_j}}{r!}}$$

$$\leq \sup_{P_j < {}^{1}/_{n^{2/3}}} \frac{\frac{y_j (nP_j)^r e^{-nP_j}}{r!}\left\{1 - \exp\left[rP_j - \frac{r(r-1)}{2n} \frac{(n-r)}{2} P_j^2 - \ldots\right]\right\}}{\frac{y_j (nP_j)^r e^{-nP_j}}{r!}}$$

$$= \sup_{P_j < {}^{1}/_{n^{2/3}}} \left\{1 - \exp\left[rP_j - \frac{r(r-1)}{2n} - \frac{(n-r)}{2} P_j^2 - \ldots\right]\right\}$$

$$= 1 - e^{o\left({}^{1}/_{n^{2/3}}\right)} . \qquad \square$$

Now we have by lemma 4.1.(i)

$$E\left[\sum_{j=1}^{M} Y_j(tn)\right] = \sum_{j=1}^{M} y_j\left[1 - \left(1 - P_j\right)^{tn}\right] \simeq \sum_{j=1}^{M} y_j\left[1 - e^{-tnP_j}\right]$$

which is

$$= \sum_{j=1}^{M} y_j\left(1 - e^{-nP_j}\right) + \sum_{j=1}^{M} y_j\left(e^{-nP_j} - e^{-tnP_j}\right)$$

$$\simeq E(T_s) + \sum_{j=1}^{M} y_j e^{-nP_j}\left[1 - e^{-(t-1)nP_j}\right]$$

$$\simeq E(T_s) + E\left[\sum_{j=1}^{M} Z_j^{(1)}\right] \frac{\sum_{j=1}^{M} y_j\left(nP_j\right)e^{-nP_j}\left[\frac{1 - e^{-(t-1)nP_j}}{nP_j}\right]}{\sum_{j=1}^{M} y_j\left(nP_j\right)e^{-nP_j}}$$

Define $F(c) = \frac{\sum_{nP_j \leq c} y_j nP_j e^{-nP_j}}{\sum_{j=1}^{M} y_j nP_j e^{-nP_j}}$. One readily observes that F(c)

is a cumulative distribution function, and it depends on the unknown parameters $(y_1, y_2, \ldots, y_M, P_1, P_2, \ldots, P_M)$. We have just shown that

Theorem 4.1:

$$E\left[\sum_{j=1}^{M} Y_j(tn)\right] \simeq E(T_s) + E\left[\sum_{j=1}^{M} Z_j^{(1)}\right]\int_0^{\infty} \frac{1 - e^{-(t-1)x}}{x}\, dF(x).$$

Remark 4.1:

(1) We can follow the procedure of Harris to obtain upper and lower bounds of $\int_0^{\infty} \frac{1 - e^{-(t-1)x}}{x}\, d\,F(x)$ for any cumulative distribution function F with given values of the first k moments. By substituting those bounds in the equation of Theorem 4.1, and also substituting T_s for $E(T_s)$ and $\sum_{j=1}^{M} Z_j^{(1)}$ for $E\left[\sum_{j=1}^{M} Z_j^{(1)}\right]$, we obtain

upper and lower bounds of $E\left[\sum_{j=1}^{M} Y_j(tn)\right]$.

(2) To apply the procedure of Harris (see Section 4 and 5 of [10]) we only need to specify the moments $\mu_r = \int_0^\infty x^r \, d\,F(x)$. Since F(x) is unknown, we use the approximation

$$m_r = \frac{(r+1)! \sum_{j=1}^{M} Z_j^{(r+1)}}{\sum_{j=1}^{M} Z_j^{(1)}} \quad \text{because } \mu_r = \frac{\sum_{j=1}^{M} y_j \left(nP_j\right)^{r+1} e^{-nP_j}}{\sum_{j=1}^{M} y_j nP_j e^{-nP_j}}$$

$$\simeq \frac{(r+1)!\,E\left[\sum_{j=1}^{M} Z_j^{(r+1)}\right]}{E\left[\sum_{j=1}^{M} Z_j^{(1)}\right]} .$$

(3) The bounds for $E\left[\sum_{j=1}^{M} Y_j(tn)\right]$ can be used as bounds for T if t is large. As indicated in Remark 3.4, $t = {}^N/_n$ seems to be a good choice for t. The following theorem shows that the estimator $\sum_{j=1}^{M} \hat{Y}_j(tn)$ in Chapter 3 is the same as the $\sum_{j=1}^{M} \hat{Y}_j(tn)$ above if we replace I by ∞.

Theorem 4.2:

$$\sum_{j=1}^{M} \hat{Y}_j(tn) = T_s + \left(\sum_{j=1}^{M} Z_j^{(1)}\right) \int_0^\infty \frac{1 - e^{-(t-1)x}}{x} \, d\,F(x)$$

$$= T_s - \sum_{i=1}^{\infty} (-1)^i (t-1)^i \left(\sum_{j=1}^{M} Z_j^{(i)} \right)$$

Proof:

Harris showed (see p. 540 of [10])

$$\int_0^{\infty} \frac{1 - e^{-(t-1)x}}{x} \, d\,F(x) = \int_0^{\alpha-1} \int_0^{\infty} e^{-tx} \, d\,F(x)\, dt$$

where $\int_0^{\infty} e^{-tx} \, d\,F(x)$ is the moment generating function of (-X).

Since $\mu_r \simeq \dfrac{(r+1)!E\left[\sum_{j=1}^{M} Z_j^{(r+1)}\right]}{E\left[\sum_{j=1}^{M} Z_j^{(1)}\right]}$,

we have

$$\int_0^{\infty} e^{-tx} \, d\,F(x) \simeq \sum_{r=0}^{\infty} \frac{(-1)^r (r+1) \sum_{j=1}^{M} Z_j^{(r+1)} \quad t^r}{\sum_{j=1}^{M} Z_j^{(1)}} .$$

Upon integrating $\int_0^{\infty} e^{-tx} \, d\,F(x)$ term by term, we get

$$\left(\sum_{j=1}^{M} Z_j^{(1)} \right) \int_0^{\infty} \frac{1 - e^{-(t-1)x}}{x} dF(x) = \sum_{r=0}^{\infty} (-1)^r \left(\sum_{j=1}^{M} Z_j^{(r+1)} \right) (t-1)^{r+1}$$

$$= \sum_{i=1}^{\infty} (-1)^i (t-1)^i \left(\sum_{j=1}^{M} Z_j^{(i)} \right) .$$

4.3 Example

This is the same example as that in the last chapter. By Remark 4.1.(2) we get

$$m_1 = 2! \sum_{j=1}^{M} Z_j^{(2)} \Bigg/ \sum_{j=1}^{M} Z_j^{(1)} = .0996954$$

$$m_2 = 3! \sum_{j=1}^{M} Z_j^{(3)} \Bigg/ \sum_{j=1}^{M} Z_j^{(2)} = .0147092$$

When we do not consider the addition of any moment constraint (i.e., k=0), we have

$$\sup \sum_{j=1}^{M} Y_j(tn) = T_s + \left(\sum_{j=1}^{M} Z_j^{(1)} \right) \lim_{x \to 0} \frac{1 - e^{-(t-1)x}}{x}$$

$$= \sum_{j=1}^{M} Y_j + (t-1) \sum_{j=1}^{M} Z_j$$

$$= 14165 + 13461(t-1)$$

$$= 190{,}706 \text{ when } t = {}^{N}/_{n} = 14.115$$

$$\inf \sum_{j=1}^{M} Y_j(tn) = T_s + \left(\sum_{j=1}^{M} Z_j^{(1)} \right) \lim_{b \to \infty} \frac{1 - e^{-(t-1)b}}{b} = \sum_{j=1}^{M} Y_j$$

$$= 14165.$$

The lower bound 14,165 seems quite conservative because, as noted in Section 2.4, the (expected) value of T is 180,000. If we add the first moment constraint m_1, then using Theorem 9 in [10], we conclude that

$$\inf \sum_{j=1}^{M} Y_j(tn) = 149186.2748 - 135021.2748e^{-.0996956(t-1)}$$

$$= 112{,}663.8231 \quad \text{when } t = 14.115.$$

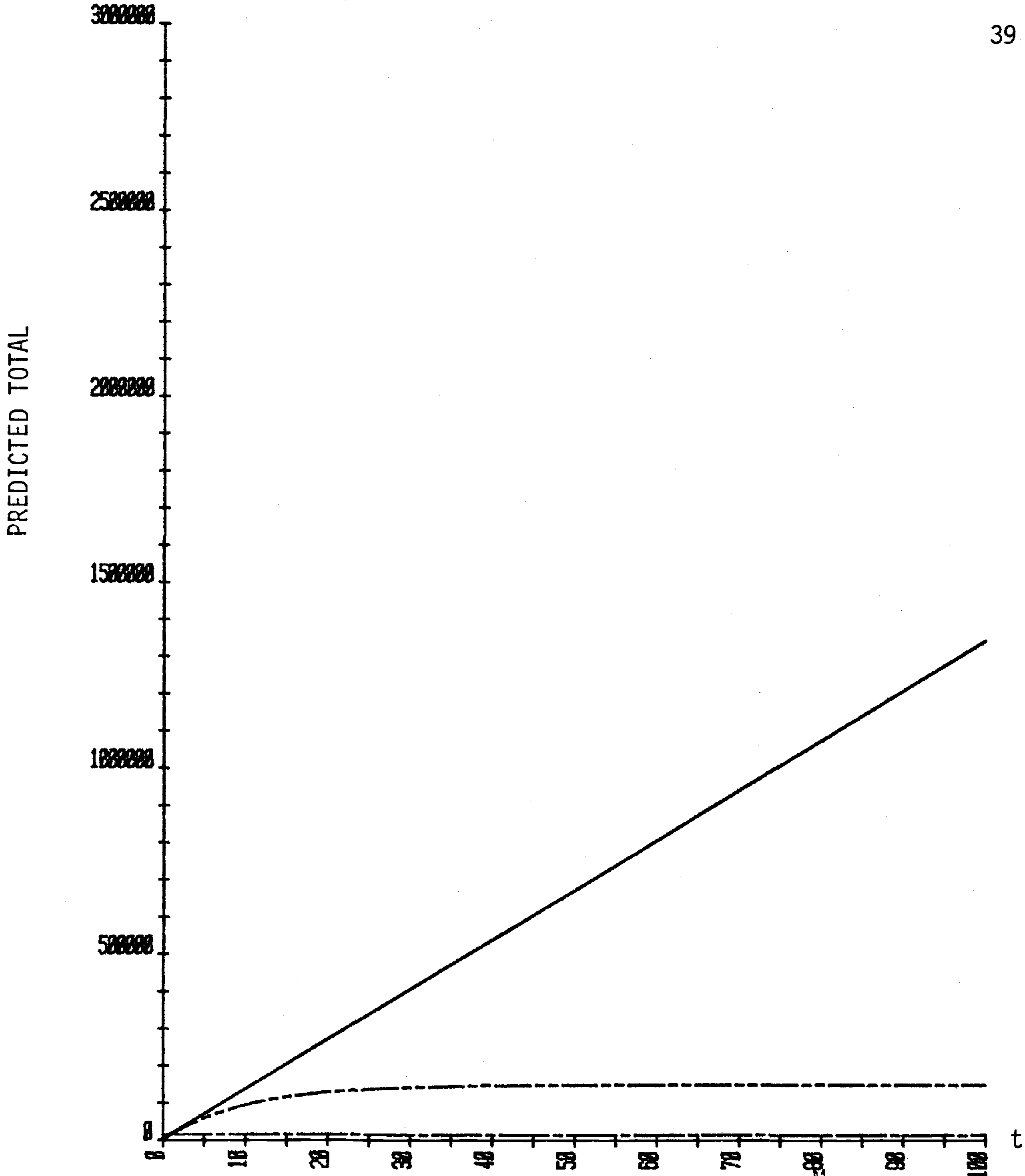
If we add the second moment constraint m_2, then using Theorem 9 in [10], we conclude that

$$\sup \sum_{j=1}^{M} Y_j(tn) = \left\{ \frac{m_2 - m_1^2}{m_2} \lim_{x \to 0} \frac{1 - e^{-(t-1)x}}{x} + \frac{m_1^2}{m_2} \cdot \frac{1 - e^{-(t-1)\frac{m_2}{m_1}}}{\frac{m_2}{m_1}} \right\} \left(\sum_{j=1}^{M} Z_j^{(1)} \right) + \sum_{j=1}^{M} Y_j$$

$$= 71448.54382 + 4365.250075t - 61648.79308$$

$$= 119{,}795 \qquad \text{when } t = 14.115.$$

From Theorem 9 of [10] the extremum which is attained for any moment constraint $(m_1, \ldots, m_r)$ is not improved by the addition of the $(r+1)$st moment constraint. Since $\sum_{j=1}^{M} \hat{Y}_j(N) = 149{,}734$ and $\sum_{j=1}^{M} \tilde{Y}_j(N) = 182{,}529$ are between 14,165 and 190,706, the bounds for k=0 make our estimator appear reasonable. But this is not true if we use the upper bound for k=2. Our feeling is that the bounds for $k \geq 1$ involve too many approximations to be accurate.



Approximations of the supremum and infimum of $\sum_{j=1}^{M} Y_j(1000t)$

$\sup \sum_{j=1}^{M} Y_j(1000t)$ without moment constraint

$\inf \sum_{j=1}^{M} Y_j(1000t)$ without moment constraint

$\inf \sum_{j=1}^{M} Y_j(1000t)$ with the first moment constraint

This figure shows the approximations of the supremum and infimum of population total based on a sample of size 1000 where y_j's are from a Poisson distribution with mean 15.

CHAPTER 5

GOOD AND RAO'S METHOD

5.1 Introduction

In this chapter sampling is done with replacement.

From Chapter 3 we have the model

$$\text{(M1)}\quad E\left[\sum_{j=1}^{M} Z_j^{(r)} \,\middle|\, P_j,\ j=1,2,\ \ldots\ ,M\right] = \sum_{j=1}^{M} y_j \binom{n}{r} P_j^{\,r}\left(1 - P_j\right)^{n-r}$$

and

$$E\left[T_s \,\middle|\, P_j,\ j=1,2,\ \ldots\ ,M\right] = \sum_{j=1}^{M} y_j\left[1 - \left(1 - P_j\right)^n\right],$$

or when n is large enough from Chapter 4 we have

$$\text{(M2)}\quad E\left[\sum_{j=1}^{M} Z_j^{(r)} \,\middle|\, \lambda_j,\ j=1,2,\ \ldots\ ,M\right] \simeq \sum_{j=1}^{M} y_j \frac{e^{-\lambda_j}\lambda_j^{\,r}}{r!}$$

where $\lambda_j = nP_j$. Also

$$E\left[T_s \,\middle|\, \lambda_j,\ j=1,2,\ \ldots\ ,M\right] \simeq \sum_{j=1}^{M} y_j\left[1 - e^{-\lambda_j}\right].$$

As prior distributions for $P_1, P_2, \ldots, P_M$ and $\lambda_1, \lambda_2, \ldots, \lambda_M$ we take beta distribution and gamma distributions respectively. We calculate the posterior means of $\sum_{j=1}^{M} Z_j^{(r)}$ and T_s, which involve the parameters of the prior distribution. In dealing with the model M2 (with $y_j = 1$ for all j), Rao [13] offered the pseudo method of moments to estimate the parameters of the gamma distribution. We extend this method to model M1 and to arbitrary y_j. The expression for the posterior mean leads to an estimator of T.

5.2 Derivations for M1

Let $f(P;\alpha,\beta) = \frac{1}{B(\alpha,\beta)} P^{\alpha-1}(1-P)^{\beta-1}$, $0 \leqq P \leqq 1$, be the density

f a beta distribution such that $\frac{\alpha+\beta}{\alpha} = M$.

Therefore

$$E_P E\left[\sum_{j=1}^{M} Z_j^{(r)} \middle| P_j,\ j=1,\ 2,\ \ldots\ ,\ M\right] = \sum_{j=1}^{M} y_j \binom{n}{r} \int_0^1 P^r(1-P)^{n-r} f(p;\alpha,\beta)dp$$

$$= \binom{n}{r} \frac{B(\alpha+r,\ \beta+n-r)}{B(\alpha,\beta)} \left(\sum_{j=1}^{M} y_j\right), \text{ and}$$

$$E_P E\left[T_s \middle| P_j,\ j=1,\ 2,\ \ldots\ ,\ M\right] = \sum_{j=1}^{M} y_j \int_0^1 \left(1 - (1-P)^n\right) f(P;\alpha,\beta)dp$$

$$= \left[1 - \frac{B(\alpha,\beta+n)}{B(\alpha+1,\ \beta)}\right]\left(\sum_{j=1}^{M} y_j\right)$$

If we can estimate α and β, then we can form the following estimators of $\sum_{j=1}^{M} y_j$

$$T_1(M1,r) = \frac{\sum_{j=1}^{M} Z_j^{(r)}}{\binom{n}{r} \frac{B(\hat{\alpha}+r,\ \hat{\beta}+n-r)}{B(\hat{\alpha},\hat{\beta})}} \quad \text{for all } r \tag{5.1}$$

or
$$T_2(M1) = \frac{T_s}{\frac{B(\hat{\alpha},\ \hat{\beta}+n)}{B(\hat{\alpha}+1,\ \hat{\beta})}} \tag{5.2}$$

Let f_r be the frequency of the classes represented by r individuals, i.e., $f_r = \sum_{j=1}^{M} I_{\{r\}}(X_j)$. Then

$$E\left[f_r \middle| P_j,\ j=1,\ 2,\ \ldots,\ M\right] = \sum_{j=1}^{M} \binom{n}{r} P_j^{\,r} \left(1 - P_j\right)^{n-r},\ \text{so}$$

$$E_P E\left[f_r \middle| P_j,\ j=1,\ 2,\ \ldots,\ M\right] = \binom{n}{r} \frac{B(\alpha+r,\ \beta+n-r)}{B(\alpha,\beta)} .$$

5.2.1 Pseudo Method of Moments for Estimating α and β

Let S denote the number of classes observed and R the number of individuals observed. Then

$$S = \sum_{r=1}^{n} f_r\ , \qquad R = \sum_{r=1}^{n} r f_r$$

and

$$E_P E(S) = \sum_{r=1}^{n} \binom{n}{r} \frac{B(\alpha+r,\ \beta+n-r)}{B(\alpha,\ \beta)} \tag{5.3}$$

$$E_p E(R) = \sum_{r=1}^{n} r \binom{n}{r} \frac{B(\alpha+r,\ \beta+n-r)}{B(\alpha,\ \beta)} . \tag{5.4}$$

Consider the equations obtained by equating the observed values of S and R to their expectations. If these equations can be solved, we use the solutions as estimates $\hat{\alpha}$ and $\hat{\beta}$ of α and β.

5.2.2 Variances of the estimators of $\sum_{j=1}^{M} y_j$

(I) Find the variance of $\hat{T}_1(M1, r)$:

The variance of $\hat{T}_1(M1, r)$ is

$$\mathrm{Var}\left[\hat{T}_1(M1, r)\right] \simeq a_r^2 \,\mathrm{Var}(S) + b_r^2 \,\mathrm{Var}(R) + c_r^2 \,\mathrm{Var}\left(\sum_{j=1}^{M} Z_j^{(r)}\right)$$

$$+ 2a_r b_r \,\mathrm{Cov}\,(S, R) + 2a_r c_r \,\mathrm{Cov}\left(S, \sum_{j=1}^{M} Z_j^{(r)}\right)$$

$$+ 2b_r c_r \,\mathrm{Cov}\left(R, \sum_{j=1}^{M} Z_j^{(r)}\right). \tag{5.5}$$

Since $R = n$, $\mathrm{Var}(R) = \mathrm{Cov}(S, R) = \mathrm{Cov}\left(R, \sum_{j=1}^{M} Z_j^{(r)}\right) = 0.$

To find $\mathrm{Var}(S)$, $\mathrm{Var}\left(\sum_{j=1}^{M} Z_j^{(r)}\right)$, and $\mathrm{Cov}\left(S, \sum_{j=1}^{M} Z_j^{(r)}\right)$, we use the following formulas.

From Remark 3.2 we have

$$\mathrm{Cov}\left(\sum_{j=1}^{M} Z_j^{(r)}, \sum_{j=1}^{M} Z_j^{(s)}\right) \simeq \delta_{rs} E\left[\sum_{j=1}^{M} \left(Z_j^{(r)}\right)^2\right] - 2^{-r-s}\binom{r+s}{r} E\left[\sum_{j=1}^{M} \left(Z_j^{(r+s)}(2n)\right)^2\right]. \tag{5.6}$$

From (30) of [7]

$$\mathrm{Cov}(f_r, f_s) \simeq \delta_{rs} E(f_r) - 2^{-r-s}\binom{r+s}{r} E\left(f_{r+s}(2n)\right) \tag{5.7}$$

and by the same proof we get

$$\mathrm{Cov}\left(\sum_{j=1}^{M} Z_j^{(r)}, f_s\right) \simeq \delta_{rs} E\left[\sum_{j=1}^{M} Z_j^{(r)}\right] - 2^{-r-s} E\left[\sum_{j=1}^{M} Z_j^{(r+s)}(2n)\right] . \tag{5.8}$$

The following is to derive it.

Define $g_r(\alpha, \beta, \omega) = \dfrac{\omega B(\alpha, \beta)}{\binom{n}{r} B(\alpha+r, \beta+n-r)}$ and note that

$\hat{T}(M1,r) = g_r(\hat{\alpha}, \hat{\beta}, \hat{\omega})$ where $\hat{\omega} = \sum_{j=1}^{M} Z_j^{(r)}$. Then

$$dg_r = \frac{\partial g_r}{\partial \alpha} d\alpha + \frac{\partial g_r}{\partial \beta} d\beta + \frac{\partial g_r}{\partial \omega} d\omega$$

$$= \frac{\omega}{\binom{n}{r}} \frac{B_\alpha(\alpha, \beta)B(\alpha+r, \beta+n-r) - B_\alpha(\alpha+r, \beta+n-r)B(\alpha, \beta)}{[B(\alpha+r, \beta+n-r)]^2} d\alpha$$

$$+ \frac{\omega}{\binom{n}{r}} \frac{B_\beta(\alpha, \beta)B(\alpha+r, \beta+n-r) - B_\beta(\alpha+r, \beta+n-r)B(\alpha, \beta)}{[B(\alpha+r, \beta+n-r)]^2} d\beta$$

$$+ \frac{B(\alpha, \beta)}{\binom{n}{r} B(\alpha+r, \beta+n-r)} d\omega \, .$$

Define

$$S(\alpha, \beta) = \sum_{r=1}^{n} \binom{n}{r} \frac{B(\alpha+r, \beta+n-r)}{B(\alpha, \beta)}$$

$$R(\alpha, \beta) = \sum_{r=1}^{n} r \binom{n}{r} \frac{B(\alpha+r, \beta+n-r)}{B(\alpha, \beta)}$$

and note that $S(\hat{\alpha}, \hat{\beta}) = S$ and $R(\hat{\alpha}, \hat{\beta}) = R$.

We have

$$dS = \sum_{r=1}^{n} \binom{n}{r} \frac{B_\alpha(\alpha+r,\ \beta+n-r)B(\alpha,\ \beta) - B_\alpha(\alpha,\ \beta)B(\alpha+r,\ \beta+n-r)}{[B(\alpha,\ \beta)]^2}\ d\alpha$$

$$+ \sum_{r=1}^{n} \binom{n}{r} \frac{B_\beta(\alpha+r,\ \beta+n-r)B(\alpha,\ \beta) - B_\beta(\alpha,\ \beta)B(\alpha+r,\ \beta+n-r)}{[B(\alpha,\ \beta)]^2}\ d\beta$$

$$dR = \sum_{r=1}^{n} r \binom{n}{r} \frac{B_\alpha(\alpha+r,\ \beta+n-r)B(\alpha,\ \beta) - B_\alpha(\alpha,\ \beta)B(\alpha+r,\ \beta+n-r)}{[B(\alpha,\ \beta)]^2}\ d\alpha$$

$$+ \sum_{r=1}^{n} r \binom{n}{r} \frac{B_\beta(\alpha+r,\ \beta+n-r)B(\alpha,\ \beta) - B_\beta(\alpha,\ \beta)B(\alpha+r,\ \beta+n-r)}{[B(\alpha,\ \beta)]^2}\ d\beta\ .$$

In other words, we get

$$\begin{pmatrix} dS \\ dR \end{pmatrix} = J \begin{pmatrix} d\alpha \\ d\beta \end{pmatrix}$$

where

$$J = \begin{bmatrix} \sum_{r=1}^{n} \binom{n}{r} \psi_\alpha^{(r)}(\alpha,\ \beta) & \sum_{r=1}^{n} \binom{n}{r} \psi_\beta^{(r)}(\alpha,\ \beta) \\ \sum_{r=1}^{n} r \binom{n}{r} \psi_\alpha^{(r)}(\alpha,\ \beta) & \sum_{r=1}^{n} r \binom{n}{r} \psi_\beta^{(r)}(\alpha,\ \beta) \end{bmatrix}$$

$$\psi_\alpha^{(r)}(\alpha,\ \beta) = \frac{B_\alpha(\alpha+r,\ \beta+n-r)B(\alpha,\ \beta) - B_\alpha(\alpha,\ \beta)B(\alpha+r,\ \beta+n-r)}{[B(\alpha,\ \beta)]^2}$$

$$\psi_\beta^{(r)}(\alpha,\ \beta) = \frac{B_\beta(\alpha+r,\ \beta+n-r)B(\alpha,\ \beta) - B_\beta(\alpha,\ \beta)B(\alpha+r,\ \beta+n-r)}{[B(\alpha,\ \beta)]^2}$$

Solving for $d\alpha$ and $d\beta$ in terms of dS and dR we obtain

$$dg_r = a_r dS + b_r dR + c_r d\omega$$

Where a_r, b_r and c_r are suitable functions of α and β. Then the asymptotic variance of $g(\hat{\alpha}, \hat{\beta}, \hat{\omega})$, using the formula (6a.2.9) on page 322 in [12], is obtained as stated.

(II) Find the variance of $\hat{T}_2(M1)$:

In order to get $Var(\hat{T}_2(M1))$ we need for formulas (5.6), (5.7), and (5.8) and $Var(T_s) \simeq - \sum_{i=1}^{\infty} (-1)^i E\left[\sum_{j=1}^{M}\left(Z_j^{(i)}\right)^2\right]$.

The approach to find $Var(\hat{T}_2(M1))$ is the same as that of (I) except $\omega = T_s$ and

$$\psi_\alpha(\alpha, \beta) = \frac{\partial}{\partial\beta} \frac{\omega B(\alpha+1, \beta)}{B(\alpha, \beta+n)}$$

$$\psi_\beta(\alpha, \beta) = \frac{\partial}{\partial\beta} \frac{B(\alpha+1, \beta)}{B(\alpha, \beta+n)} .$$

5.3 Example of M1

For the example of Section 3.5, the equations of the pseudo method of moments estimators for α and β are

$$949 = \binom{1000}{1} \frac{B(\alpha+1, \beta+999)}{B(\alpha, \beta)} + \binom{1000}{2} \frac{B(\alpha+2, \beta+998)}{B(\alpha, \beta)} + \binom{1000}{3} \frac{B(\alpha+3, \beta+997)}{B(\alpha, \beta)}$$

$$1{,}000 = \binom{1000}{1} \frac{B(\alpha+1, \beta+999)}{B(\alpha, \beta)} + 2\binom{1000}{2} \frac{B(\alpha+2, \beta+998)}{B(\alpha, \beta)} + 3\binom{1000}{3} \frac{B(\alpha+3, \beta+997)}{B(\alpha, \beta)} .$$

Unfortunately, there do not exist solutions for α and β. That is, the method of moments does not work in this example.

5.4 Derivations for M2

We have

$$E\left[\sum_{j=1}^{M} Z_j^{(r)} \middle| \lambda_j, j=1, 2, \ldots, M\right] \approx \sum_{j=1}^{M} y_j \frac{e^{-\lambda_j}\lambda_j^{\,r}}{r!}$$

and

$$E\left[\sum_{j=1}^{M} Y_j \middle| \lambda_j, j=1, 2, \ldots, M\right] \approx \sum_{j=1}^{M} y_j\left[1 - e^{-\lambda_j}\right].$$

Suppose that λ_1, λ_2, ... , and λ_M can be approximated by a gamma distribution with density

$$\frac{1}{\Gamma(\alpha)\beta^{\alpha}} \lambda^{\alpha-1} e^{-\lambda/\beta} d\lambda \ .$$

Hence

$$E_\lambda E\left[\sum_{j=1}^{M} Z_j^{(r)} \middle| \lambda_j, j=1, 2, \ldots, M\right] = \frac{\Gamma(\alpha+r)}{r!\Gamma(\alpha)} \frac{1}{(1+\beta)^{\alpha}} \left(\frac{\beta}{1+\beta}\right)^r \left(\sum_{j=1}^{M} y_j\right)$$

and

$$E_\lambda E\left[T_s \middle| \lambda_j, j=1, 2, \ldots, M\right] = \left[1 - \frac{1}{(1+\beta)^{\alpha}}\right]\left(\sum_{j=1}^{M} y_j\right)$$

If we can estimate α and β, then we can form the following estimators of $\sum_{j=1}^{M} y_j$:

$$\hat{T}_1(M2,r) = \frac{\sum_{j=1}^{M} Z_j^{(r)}}{\frac{\Gamma(\hat{\alpha}+r)}{r!\Gamma(\hat{\alpha})} \frac{1}{(1+\hat{\beta})^{\hat{\alpha}}} \left(\frac{\hat{\beta}}{1+\hat{\beta}}\right)^r} \quad \text{for all } r \tag{5.9}$$

or

$$\hat{T}_2(M2,r) = \frac{T_s}{1 - \frac{1}{(1+\hat{\beta})^{\hat{\alpha}}}} \ . \tag{5.10}$$

Since

$$E_\lambda E\left[f_r \middle| \lambda_j,\ j=1,\ 2,\ \ldots,\ M\right] = M\,\frac{\Gamma(\alpha+r)}{r!\,\Gamma(\alpha)}\,\frac{1}{(1+\beta)^\alpha}\left(\frac{\beta}{1+\beta}\right)^r$$

$$= \tau\,\frac{\Gamma(\alpha+r)}{r!\,\Gamma(\alpha)}\,\frac{1}{(1+\beta)^\alpha}\left(\frac{\beta}{1+\beta}\right)^r \quad \text{where } \tau = M\alpha\ ,$$

we can find estimators of α , β, and τ in terms of the f_r.

5.4.1 Pseudo Method of Moments for Estimating α, β, and τ

Define $S = \sum_{r=1}^{n} f_r$, $R = \sum_{r=1}^{n} rf_r$ and $U = \sum_{r=1}^{n} r^2 f_r$. Then

$$E_\lambda E(S) = \tau\,\frac{\left[1 - (1+\beta)^{-\alpha}\right]}{\alpha} \tag{5.11}$$

$$E_\lambda E(R) = \tau\beta \tag{5.12}$$

$$E_\lambda E(U) = \tau\beta(1 + \beta + \alpha\beta)\ . \tag{5.13}$$

Equating observed values of S, R, and U to their expectations, we obtain estimates $\hat{\alpha}$, $\hat{\beta}$, and $\hat{\tau}$ (if the solutions exist) of α, β, and τ.

5.4.2 Variances of the estimators of $\sum_{j=1}^{M} y_j$

(I) Find the variance of $\hat{T}_1(M_2,r) = \dfrac{\sum_{j=1}^{M} Z_j^{(r)}}{\dfrac{\Gamma(\hat{\alpha}+r)}{r!\,\Gamma(\hat{\alpha})}\dfrac{1}{(1+\hat{\beta})^{\hat{\alpha}}}\left(\dfrac{\hat{\beta}}{1+\hat{\beta}}\right)^r}$:

Define $g_r(\alpha,\ \beta,\ \tau,\ \omega) = \dfrac{\omega}{\dfrac{\Gamma(\alpha+r)}{r!\,\Gamma(\alpha)}\dfrac{1}{(1+\beta)^\alpha}\left(\dfrac{\beta}{1+\beta}\right)^r}$ and note that

$\hat{T}_1(M2,r) = g_r(\hat{\alpha},\ \hat{\beta},\ \hat{\tau},\ \hat{\omega})$ where $\hat{\omega} = \sum_{j=1}^{M} Z_j^{(r)}$. Then

$$dg_r = \frac{\partial g_r}{\partial \alpha} d\alpha + \frac{\partial g_r}{\partial \beta} d\beta + \frac{\partial g_r}{\partial \tau} d\tau + \frac{\partial g_r}{\partial \omega} d\omega \qquad (5.14)$$

where

$$\frac{\partial g}{\partial \alpha} = \omega r! (1+\beta)^{\alpha} \left(\frac{1+\beta}{\beta}\right)^{r} \left\{ \frac{\Gamma'(\alpha)\Gamma(\alpha+r) - \Gamma'(\alpha+r)\Gamma(\alpha)}{[\Gamma(\alpha+r)]^2} + \frac{\Gamma(\alpha)}{\Gamma(\alpha+r)} \ln(1+\beta) \right\}$$

$$\frac{\partial g}{\partial \beta} = \omega \frac{r!\Gamma(\alpha)}{\Gamma(\alpha+r)} (1+\beta)^{\alpha-1} \left(\frac{1+\beta}{\beta}\right)^{r-1} \left\{ \alpha\left(\frac{1+\beta}{\beta}\right) - \frac{r}{\beta^2} (1+\beta) \right\}$$

$$\frac{\partial g}{\partial \tau} = 0$$

$$\frac{\partial g}{\partial \omega} = \frac{r!\Gamma(\alpha)}{\Gamma(\alpha+r)} \alpha (1+\beta)^{\alpha-1} \left(\frac{1+\beta}{\beta}\right)^{r} .$$

Define

$$S(\alpha, \beta, \tau) = \tau \frac{\left[1 - (1+\beta)^{-\alpha}\right]}{\alpha}$$

$$R(\alpha, \beta, \tau) = \tau\beta$$

$$U(\alpha, \beta, \tau) = \tau\beta(1+\beta + \alpha\beta)$$

and note that $S(\alpha, \beta, \tau) = S$, $R(\alpha, \beta, \tau) = R$, and $U(\alpha, \beta, \tau) = U$. We have

$$\begin{bmatrix} dS \\ dR \\ dU \end{bmatrix} = J_1 \begin{bmatrix} d\alpha \\ d\beta \\ d\tau \end{bmatrix}$$

where

$$J_1 = \begin{bmatrix} \frac{\tau}{\alpha^2}\left\{-1 + (1+\beta)^{-\alpha}[1 + \log(1+\beta)]\right\} & \tau(1+\beta)^{-\alpha-1} & \frac{1-(1+\beta)^{-\alpha}}{\alpha} \\ 0 & \tau & \beta \\ \tau\beta^2 & \tau(1+2\beta+2\alpha\beta) & \beta(1+\beta+\alpha\beta) \end{bmatrix} .$$

Solving for $d\alpha$, $d\beta$, and $d\tau$ in terms of dS, dR, and dU we obtain

$$dg_r = a_r dS + b_r dR + c_r dU + d_r d\omega$$

where a_r, b_r, c_r, and d_r are suitable functions of α, β, τ, and ω. Then the asymptotic variance of $g(\hat{\alpha}, \hat{\beta}, \hat{\tau}, \hat{\omega})$, using the formula (6a.2.9) on page 322 in [12], is

$$\begin{aligned}\mathrm{Var}(\hat{T}_1(M2, r)) = {} & a_r^2 \mathrm{Var}(S) + b_r^2 \mathrm{Var}(R) + c_r^2 \mathrm{Var}(U) \\ & + d_r^2 \mathrm{Var}\left(\sum_{j=1}^{M} Z_j^{(r)}\right) + 2a_r b_r \mathrm{Cov}(S, R) + 2a_r c_r \mathrm{Cov}(S, U) \\ & + 2a_r d_r \mathrm{Cov}\left(S, \sum_{j=1}^{M} Z_j^{(r)}\right) + 2b_r c_r \mathrm{Cov}(R, U) + 2b_r d_r \mathrm{Cov}\left(R, \sum_{j=1}^{M} Z_j^{(r)}\right) \\ & + 2c_r d_r \mathrm{Cov}\left(U, \sum_{j=1}^{M} Z_j^{(r)}\right) . \end{aligned} \tag{5.15}$$

From [13] on page 136 we get

$$\mathrm{Cov}\begin{bmatrix} S \\ R \\ U \end{bmatrix} = \begin{bmatrix} \dfrac{\tau[(1+\beta)^{-\alpha} - (2+\beta)^{-\alpha}]}{\alpha} & \tau\beta(1+\beta)^{-\alpha-1} & \tau\beta(1+\beta)^{-\alpha-2}(2+\alpha+\beta) \\ \tau\beta(1+\beta)^{-\alpha-1} & \tau\beta & \tau\beta[1+2\beta(\alpha+1)] \\ \tau\beta(1+\beta)^{-\alpha-2}(2+\alpha+\beta) & \tau\beta[1+2\beta(\alpha+1)] & \tau\beta[4+3\beta(\alpha+1)+4\beta^2(\alpha+1)(\alpha+2)] \end{bmatrix} \tag{5.16}$$

Remark 5.1:

(1) $\displaystyle\sum_{j=1}^{M} Z_j^{(r)}(tn) = t^r \sum_{i=0}^{\infty} (-1)^i \binom{r+i}{r} (t-1)^i \left(\sum_{j=1}^{M} Z_j^{(r+i)}\right)$ by Remark 3.1

If we consider Euler's transformation assuming that $\sum_{j=1}^{M} Z_j^{(r)}$ decreases slowly after the first term, then

$$\sum_{j=1}^{M} Z_j^{(1)}(tn) \simeq t\sum_{j=1}^{M} Z_j^{(1)} - 2(t-1)\sum_{j=1}^{M} Z_j^{(2)} \qquad (5.17)$$

and

$$\sum_{j=1}^{M} Z_j^{(r)}(tn) \simeq t^{r-1}\sum_{j=1}^{M} Z_j^{(r)} \quad \text{when } r \geqq 2. \qquad (5.18)$$

(2) Since $\mathrm{Cov}\left(S, \sum_{j=1}^{M} Z_j^{(r)}\right) = \mathrm{Cov}\left(M - f_0, \sum_{j=1}^{M} Z_j^{(r)}\right) = -\mathrm{Cov}\left(f_0, \sum_{j=1}^{M} Z_j^{(r)}\right) = 2^{-r}E\left[\sum_{j=1}^{M} Z_j^{(r)}(2n)\right]$,

$$\hat{\mathrm{Cov}}\left(S, \sum_{j=1}^{M} Z_j^{(r)}\right) = \sum_{i=0}^{\infty} (-1)^i \binom{r+i}{r} \sum_{j=1}^{M} Z_j^{(r+i)} \quad \text{without Euler's transformation}$$

$$\text{or } \hat{\mathrm{Cov}}\left(S, \sum_{j=1}^{M} Z_j^{(r)}\right) = \begin{cases} \sum_{j=1}^{M} Z_j^{(1)} - \sum_{j=1}^{M} Z_j^{(2)} & \text{when } r=1 \\ \frac{1}{2}\sum_{j=1}^{M} Z_j^{(r)} & \text{when } r \geqq 2. \end{cases} \quad \text{with Euler's transformation}$$

(3) $\mathrm{Cov}\left(R, \sum_{j=1}^{M} Z_j^{(r)}\right) = 0$ for all r since R = n.

(4) Since $\mathrm{Cov}\left(U, \sum_{j=1}^{M} Z_j^{(r)}\right) = \mathrm{Cov}\left(\sum_{s=0}^{n} s^2 f_s, \sum_{j=1}^{M} Z_j^{(r)}\right) = \sum_{s=0}^{n} s^2 \mathrm{Cov}\left(f_s, \sum_{j=1}^{M} Z_j^{(r)}\right) = \sum_{s=1}^{n} s^2 \left\{\delta_{rs} E\left[\sum_{j=1}^{M} Z_j^{(r)}\right] - 2^{-r-s} \cdot E\left[\sum_{j=1}^{M} Z_j^{(r+s)}(2n)\right]\right\}$, we have

$$\hat{\mathrm{Cov}}\left(U, \sum_{j=1}^{M} Z_j^{(r)}\right) = r^2 \sum_{j=1}^{M} Z_j^{(r)} - \sum_{s=1}^{n} s^2 \sum_{i=0}^{n} (-1)^i \binom{r+s+i}{r} \cdot$$

$$\left(\sum_{j=1}^{M} Z_j^{(r+s+i)}\right) \text{ without Euler's transformation}$$

$$\text{or } \hat{\mathrm{Cov}}\left(U, \sum_{j=1}^{M} Z_j^{(r)}\right) = r^2 \sum_{j=1}^{M} Z_j^{(r)} - \frac{1}{2} \sum_{s=1}^{n} s^2 \left(\sum_{j=1}^{M} Z_j^{(r+s)}\right)$$

with Euler's transformation.

(5) From Remark 3.2(1) we have

$$\hat{\mathrm{Var}}\left(\sum_{j=1}^{M} Z_j^{(r)}\right) = \sum_{j=1}^{M}\left(Z_j^{(r)}\right)^2 - 2^{-2r}\binom{2r}{r}\sum_{j=1}^{M}\left(Z_j^{(2r)}(2n)\right)^2$$

$$= \sum_{j=1}^{M}\left(Z_j^{(r)}\right)^2 - \binom{2r}{r}\sum_{i=0}^{\infty}(-1)^i\binom{2r+i}{2r}\left[\sum_{j=1}^{M}\left(Z_j^{(2r+i)}\right)^2\right]$$

without Euler's transformation.

$$\text{or } \hat{\mathrm{Var}}\left(\sum_{j=1}^{M} Z_j^{(r)}\right) = \sum_{j=1}^{M}\left(Z_j^{(r)}\right)^2 - \frac{1}{2}\binom{2r}{r}\sum_{j=1}^{M}\left(Z_j^{(2r)}\right)^2$$

with Euler's transformation assuming that $\sum_{j=1}^{M}\left(Z_j^{(r)}\right)^2$ decreases slowly after the first term.

(II) Find the variance of $\hat{T}_2(M2) = \dfrac{T_s}{1 - \dfrac{1}{(1+\hat{\beta})^{\hat{\alpha}}}}$:

Define $g(\alpha, \beta, \tau, \omega) = \dfrac{\omega}{1 - \dfrac{1}{(1+\beta)^{\alpha}}}$ and note that

$\hat{T}_2(M2) = g(\hat{\alpha}, \hat{\beta}, \hat{\tau}, \hat{\omega})$ where $\hat{\omega} = T_s$. Then

$$dg = \frac{\partial g}{\partial \alpha} d\alpha + \frac{\partial g}{\partial \beta} d\beta + \frac{\partial g}{\partial \tau} d\tau + \frac{\partial g}{\partial \omega} d\omega \tag{5.19}$$

where

$$\frac{\partial g}{\partial \alpha} = \frac{-\omega(1+\beta)^{\alpha} \ln(1+\beta)}{[(1+\beta)^{\alpha} - 1]^2}$$

$$\frac{\partial g}{\partial \beta} = \frac{-\alpha\omega(1+\beta)^{\alpha-1}}{[(1+\beta) - 1]^2}$$

$$\frac{\partial g}{\partial \tau} = 0$$

$$\frac{\partial g}{\partial \omega} = \frac{(1+\beta)^{\alpha}}{(1+\beta)^{\alpha} - 1} \quad .$$

Using the same approach as (I) we get

$$\partial g = a\partial S + b\partial R + c\partial U + d\partial\hat{\omega} \tag{5.20}$$

where a, b, c and d are suitable functions of α, β, τ, and ω and

$$\text{Var}\,(\hat{T}_2(M2)) = a^2\text{Var}(S) + b^2\text{Var}(R) + c^2\text{Var}(U) + d^2\text{Var}(T_s)$$

$$+ 2ab\,Cov(S, R) + 2ac\,Cov(S, U) + 2ad\,Cov(S, T_s) + 2bc\,Cov(R, U) + 2bd\,Cov(R, T_s) + 2cd\,Cov(U, T_s)$$

where

$$Cov(S, T_s) = \sum_{r=1}^{n} Cov\left(S, \sum_{j=1}^{M} Z_j^{(r)}\right)$$

$$Cov(R, T_s) = 0$$

$$Cov(U, T_s) = \sum_{r=1}^{n} Cov\left(U, \sum_{j=1}^{M} Z_j^{(r)}\right).$$

5.5 Example of M2

We now apply this method to the example in Section 3.5. We have

$$\hat{\tau}\,\frac{[1 - (1+\hat{\beta})^{-\hat{\alpha}}]}{\hat{\alpha}} = 949 ,$$

$$\hat{\tau}\hat{\beta} = 1{,}000 \quad \text{, and}$$

$$\hat{\tau}\hat{\beta}(1+\hat{\beta}+\hat{\alpha}\hat{\beta}) = 1{,}106.$$

The solutions are

$$\begin{cases} \hat{\alpha} = 8.78268266064 \\ \hat{\beta} = .01083547363 \\ \hat{\tau} = 92287.45906 \end{cases} \quad \text{or} \quad \begin{cases} \hat{\alpha} = -.00000057585 & \text{(not reasonable)} \\ \hat{\beta} = .10600006104 \\ \hat{\tau} = 9433.956832 . \end{cases}$$

For $r=1$, $$\hat{T}_1(M2, r=1) = \frac{\sum_{j=1}^{M} Z_j^{(1)}}{\frac{\Gamma(\hat{\alpha}+1)}{(\hat{\alpha})}\;\frac{\hat{\beta}}{(1+\hat{\beta})^{\hat{\alpha}+1}}} = 157{,}177$$

for r=2, $\hat{T}_1(M2, r=2) = \dfrac{\sum_{j=1}^{M} Z_j^{(2)}}{\dfrac{\Gamma(\hat{\alpha}+2)}{2!\Gamma(\hat{\alpha})} \dfrac{\hat{\beta}^2}{(1+\hat{\beta})^{\hat{\alpha}+2}}} = 149{,}431$, and

for r=3, $\hat{T}_1(M2, r=3) = \dfrac{\sum_{j=1}^{M} Z_j^{(3)}}{\dfrac{\Gamma(\hat{\alpha}+3)}{3!\Gamma(\hat{\alpha})} \dfrac{\hat{\beta}^3}{(1+\beta)^{\alpha+3}}} = 190{,}747.$

Also,

$$\hat{T}_2(M2) = \frac{\sum_{j=1}^{M} Y_j}{1 - \dfrac{1}{(1+\hat{\beta})^{\hat{\alpha}}}} = 156{,}847$$

Now let us consider the variance $\hat{Var}(\hat{T}_1(M2, r))$

$$\mathrm{Cov}\begin{bmatrix} S \\ R \\ U \end{bmatrix} = \begin{bmatrix} 9536.16 & 899.92 & 9609.16 \\ 899.92 & 999.98 & 1211.97 \\ 9609.16 & 1211.97 & 4367.23 \end{bmatrix}$$

$$J_1^{-1} = \begin{bmatrix} -.0109521933736 & .016007251633 & -.005069705794336 \\ .00001213084646318 & -.0001307821596695 & .000107839046779 \\ -103.3903909322 & 1206.182399682 & -918.4826883093 \end{bmatrix}$$

when r=1

$a_1 = 19.936073931$ $\qquad b_1 = 1438.915688$

$c_1 = -1318.114736$ $\qquad d_1 = 101.45170575$

$$\hat{\text{Cov}}\left(S, \sum_{j=1}^{M} Z_j^{(1)}\right) = \begin{cases} 12{,}218 & \text{without Euler's transformation} \\ 12{,}790 & \text{with Euler's transformation} \end{cases}$$

$$\hat{\text{Cov}}\left(U, \sum_{j=1}^{M} Z_j^{(1)}\right) = \begin{cases} 11{,}822 & \text{without Euler's transformation} \\ 13{,}059.5 & \text{with Euler's transformation} \end{cases}$$

$$\hat{\text{Var}}\left(\sum_{j=1}^{M} Z_j^{(1)}\right) = \begin{cases} 197{,}593 & \text{without Euler's transformation} \\ 204{,}456 & \text{with Euler's transformation} \end{cases}$$

Therefore

$$\text{Var}(\hat{T}_1(M2, r=1)) = \begin{cases} 3.532533918 \times 10^9 & \text{without Euler's transformation} \\ 3.274515433 \times 10^9 & \text{with Euler's transformation} \end{cases}$$

The relative standard error is

$$\begin{cases} .38 & \text{without Euler's transformation} \\ .36 & \text{with Euler's transformation} \end{cases}$$

when r=2

$a_2 = 20.754798748$ $b_2 = 2907.842194$

$c_2 = -2646.960626$ $d_2 = 1934.924667$

$$\hat{\text{Cov}}\left(S, \sum_{j=1}^{M} Z_j^{(2)}\right) = \begin{cases} 572 & \text{without Euler's transformation} \\ 335.5 & \text{with Euler's transformation} \end{cases}$$

$$\hat{\text{Cov}}\left(U, \sum_{j=1}^{M} Z_j^{(2)}\right) = \begin{cases} 2,585 & \text{without Euler's transformation} \\ 2,667.5 & \text{with Euler's transformation} \end{cases}$$

$$\hat{\text{Var}}\left(\sum_{j=1}^{M} Z_j^{(2)}\right) = 10,157 \quad \text{with and without Euler's transformation}$$

Therefore

$$\hat{\text{Var}}(\hat{T}_1(M2, r=2) = \begin{cases} 3.104794347 \times 10^{10} & \text{without Euler's transformation} \\ 3.018387282 \times 10^{10} & \text{with Euler's transformation} \end{cases}$$

The relative standard error is

$$= \begin{cases} 1.18 & \text{without Euler's transformation} \\ 1.16 & \text{with Euler's transformation} \end{cases}$$

when r=3

a_3 = 8.967069573 b_3 = 5706.407682

c_3 = -5167.194706 d_3 = 50,221.67689

$$\hat{\text{Cov}}\left(S, \sum_{j=1}^{M} Z_j^{(3)}\right) = \begin{cases} 33 & \text{without Euler's transformation} \\ 16.5 & \text{with Euler's transformation} \end{cases}$$

$$\hat{Cov}\left(U, \sum_{j=1}^{M} Z_j^{(3)}\right) = 297 \text{ with and without Euler's transformation}$$

$$\hat{Cov}\left(\sum_{j=1}^{M} Z_j^{(3)}\right) = 549 \text{ with and without Euler's transformation}$$

Therefore

$$\hat{Var}(\hat{T}_1(M2, r=3)) = \begin{cases} 1.307477378 \times 10^{12} & \text{without Euler's transformation} \\ 1.307376523 \times 10^{12} & \text{with Euler's transformation} \end{cases}$$

The relative standard error is

$$= \begin{cases} 5.99 & \text{without Euler's transformation} \\ 5.99 & \text{with Euler's transformation} \end{cases} .$$

Now let us consider $\hat{Var}(\hat{T}_2(M2))$

Since $a = 19.961152284$ $\quad b = 1522.798549$

$c = -1393.979799$ $\quad d = 11.072835296$

$$\text{and } \hat{Cov}(S, T_s) = \begin{cases} 12{,}823 & \text{without Euler's transformation} \\ 13{,}142 & \text{with Euler's transformation} \end{cases}$$

$$\hat{Cov}(U, T_s) = \begin{cases} 14{,}704 & \text{without Euler's transformation} \\ 16{,}024 & \text{with Euler's transformation} \end{cases}$$

$$\hat{Cov}(T_s) = \begin{cases} 208{,}299 & \text{without Euler's transformation} \\ 215{,}162 & \text{with Euler's transformation} \end{cases}$$

$$\hat{Var}(\hat{T}_2(M2)) = \begin{cases} 4.76078734 \times 10^9 & \text{without Euler's transformation} \\ 4.721020597 \times 10^9 & \text{with Euler's transformation} \end{cases}$$

The relative standard error is

$$\begin{cases} .44 & \text{without Euler's transformation} \\ .44 & \text{with Euler's transformation} \end{cases}$$

These calculations are summarized in Table 5.1.

From the information above in this case we would choose the estimate of $\sum_{j=1}^{M} y_j$

to be $\hat{T}_1(M2,\ r=1) = 157{,}177$

with the relative standard error is .36.

	estimated population total	estimated variance		relative standard error	
		without Euler's transformation	with Euler's transformation	without Euler's transformation	with Euler's transformation
$\hat{T}_1(M2,r=1)$	157,177	3.532533918×10^9	3.274515443×10^9	.38	.36
$\hat{T}_1(M2,r=2)$	149,431	$3.104794347 \times 10^{10}$	$3.018387282 \times 10^{10}$	1.18	1.16
$\hat{T}_1(M2,r=3)$	190,747	$1.307477378 \times 10^{12}$	$1.307376523 \times 10^{12}$	5.99	5.99
$\hat{T}_2(M2)$	156,847	4.76078734×10^9	4.721020597×10^9	.44	.44

Table 5.1: Estimated population total, estimated variance, and relative standard error.

CHAPTER 6

EFRON AND THISTED'S METHOD

6.1 Introduction

In this chapter we still consider sampling with replacement. Efron and Thisted [2] tried to find a reasonable estimator of $d(\infty)$ supposing that $E(f_r) = H\int \frac{e^{-\lambda}\lambda^x}{x!}\, dG(\lambda)$ for some distribution G. If $G(\lambda)$ is a gamma distribution with parameters α, β, then an estimator of $d(tn)$ is

$$\hat{d}(tn) = \begin{cases} \frac{f_1}{\gamma\alpha}\left[1 - \frac{1}{(1+\gamma t)^{\alpha}}\right] & \text{if } \alpha > 0 \\ \frac{f_1}{\gamma}\log(1+\gamma t) & \text{if } \alpha = 0 \end{cases}$$

where $\gamma = \frac{\beta}{1+\beta}$.

He also found other possible estimators.

(1) $\hat{d}(tn) = \sum_{x=1}^{\infty} (-1)^{x+1} f_x t^x$, or

if Euler's transformation is considered, then

$$\hat{d}(tn) = \sum_{y=1}^{X_0} \xi_y u^y \text{ where } \xi_y = \sum_{x=1}^{y} \binom{y-1}{x-1} \frac{(-1)^{x+1}}{2^y} f_x \text{ and } t = \frac{u}{2-u}$$

(2) $\hat{d}(tn) = \sum_{x=1}^{\infty} (-1)^{x+1} \hat{f}_x t^x$ where $\hat{f}_x = f_1 \frac{\Gamma(x+\alpha)}{x!\,\Gamma(1+\alpha)} \gamma^{x-1}$

$$= f_1 t \sum_{x=1}^{\infty} (-1)^{x+1} \frac{\Gamma(x+\alpha)}{x!\Gamma(1+\alpha)} (\gamma t)^{x-1}$$

which can also be modified by Euler's transformation.

We generalize their derivations to estimate $T = \sum_{j=1}^{M} y_j$ by using $\Delta(\infty)$ where $\Delta(tn) = E\left[\sum_{j=1}^{M} Z_j^{(1)}\right] \frac{\int e^{-\lambda}(1-e^{-\lambda t})dG(\lambda)}{\int e^{-\lambda}\lambda dG(\lambda)}$, and we also derive the biases of these estimators to measure their precision.

6.2 Nonparametric Model

From Chapter 4, lemma 4.1, we know

$$E\left[\sum_{j=1}^{M} Z_j^{(x)} \middle| \lambda_j\right] \simeq \sum_{j=1}^{M} y_j \frac{e^{-\lambda_j}\lambda_j^x}{x!} .$$

Suppose that M is large and the frequency distribution of values λ_1, ... , λ_M can be approximated by a continuous distribution $G(\lambda)$.

Then,

$$E\left[\sum_{j=1}^{M} Z_j^{(x)}\right] = E_\lambda E\left[\sum_{j=1}^{M} Z_j^{(x)} \middle| \lambda_j, j=1, \ldots, M\right] = \left(\sum_{j=1}^{M} y_j\right) \int \frac{e^{-\lambda}\lambda^x}{x!} dG(\lambda).$$

Define

$$Y_{\bar{j}}(tn) = y_j \delta_{\bar{j}}(tn) = \begin{cases} y_j & \text{if the jth class shows in the second sample of size tn but does not show the basic sample} \\ 0 & \text{otherwise} \end{cases}$$

where

$$\delta_{\bar{j}}(tn) = \begin{cases} 1 & \text{if the jth class shows in the second sample of size tn but does not show in the basic sample} \\ 0 & \text{otherwise} \end{cases}$$

and

$$\Delta(t) = E_\lambda E\left[\sum_{j=1}^{M} Y_j^-(tn) \middle| \lambda_j\right] .$$

We have

$$\Delta(t) = E_\lambda\left\{\sum_{j=1}^{M} y_j\left(1 - P_j\right)^n\left[1 - \left(1 - P_j\right)^{nt}\right]\right\}$$

$$\simeq E_\lambda\left\{\sum_{j=1}^{M} y_j e^{-nP_j}\left(1 - e^{-ntP_j}\right)\right\}$$

$$= E_\lambda\left\{\sum_{j=1}^{M} y_j e^{-\lambda_j}\left(1 - e^{-\lambda_j t}\right)\right\} \text{ where } \lambda_j = nP_j$$

$$= \left(\sum_{j=1}^{M} y_j\right)\int e^{-\lambda}\left(1 - e^{-\lambda t}\right) dG(\lambda) \tag{6.1}$$

$$= E\left[\sum_{j=1}^{M} Z_j^{(1)}\right] \frac{\int e^{-\lambda}\left(1 - e^{-\lambda t}\right) dG(\lambda)}{\int e^{-\lambda}\lambda dG(\lambda)} . \tag{6.2}$$

We wish to estimate $\Delta(t)$. Substituting the expansion

$$1 - e^{-\lambda t} = \lambda t - \frac{\lambda^2 t^2}{2!} + \frac{\lambda^3 t^3}{3!} -+ \ldots$$

into (6.1), we obtain

$$\Delta(t) \simeq E\left[\sum_{j=1}^{M} Z_j^{(1)}\right] t - E\left[\sum_{j=1}^{M} Z_j^{(2)}\right] t^2 + E\left[\sum_{j=1}^{M} Z_j^{(3)}\right] t^3 -+ \ldots . \tag{6.3}$$

This result appears in Remark 3.1.(5) in Chapter 3. The right-hand side need not converge, but assuming it does, this suggests an estimator for $\Delta(t)$

$$\hat{\Delta}(t) = \left(\sum_{j=1}^{M} Z_j^{(1)}\right) t - \left(\sum_{j=1}^{M} Z_j^{(2)}\right) t^2 + \left(\sum_{j=1}^{M} Z_j^{(3)}\right) t^3 -+ \ldots . \tag{6.4}$$

The estimator $\hat{\Delta}(t)$ is a function of the data only through the statistics $\sum_{j=1}^{M} Z_j^{(x)}$. Unfortunately $\hat{\Delta}(t)$ is useless for values of t larger than 1. The geometrically increasing magnitude of t^x produces wild oscillations in $\hat{\Delta}(t)$ as the number of terms increases.

6.3 Parametric Model with a Gamma Distribution for $G(\lambda)$

The c.d.f. $G(\lambda)$ is approximated by a gamma distribution with density,

$$g(\lambda) = \frac{1}{\Gamma(\alpha)\beta^{\alpha}} \lambda^{\alpha-1} e^{-\lambda/\beta} \tag{6.5}$$

Therefore

$$E\left[\sum_{j=1}^{M} Z_j^{(x)}\right] = \left(\sum_{j=1}^{M} y_j\right) \int \frac{e^{-\lambda}\lambda^x}{x!} dG(\lambda) = \left(\sum_{j=1}^{M} y_j\right) \frac{1}{\Gamma(\alpha)\beta^{\alpha}} \int \frac{\lambda^{\alpha+x-1} e^{-\lambda(1+\frac{1}{\beta})}}{x!} d\lambda$$

$$= \left(\sum_{j=1}^{M} y_j\right) \frac{1}{\Gamma(\alpha)\beta^{\alpha}} \frac{\Gamma(x+\alpha)}{x!} \gamma^{\alpha+x} \quad \text{where } \gamma = \frac{\beta}{1+\beta}$$

$$= E\left[\sum_{j=1}^{M} Z_j^{(1)}\right] \frac{\Gamma(x+\alpha)}{x!\,\Gamma(1+\alpha)} \gamma^{x+1} \tag{6.6}$$

$E\left[\sum_{j=1}^{M} Z_j^{(x)}\right]$ is proportional to the negative binomial distribution with parameters α and γ. Integrating (6.2) we obtain

$$\Delta(t) \simeq \begin{cases} \dfrac{E\left[\sum_{j=1}^{M} Z_j^{(1)}\right]}{\alpha\gamma} \left[1 - \dfrac{1}{(1+\gamma t)^{\alpha}}\right] & \text{if } \alpha > 0 \\[2ex] \dfrac{E\left[\sum_{j=1}^{M} Z_j^{(1)}\right]}{\gamma} \log(1+\gamma t) & \text{if } \alpha = 0 . \end{cases} \tag{6.7}$$

Hence

$$\hat{\Delta}(t) = \begin{cases} \dfrac{\sum_{j=1}^{M} Z_j^{(1)}}{\hat{\alpha}\hat{\gamma}} \left[1 - \dfrac{1}{(1+\hat{\gamma}t)^{\hat{\alpha}}} \right] & \text{if } \hat{\alpha} > 0 \\ \dfrac{\sum_{j=1}^{M} Z_j^{(1)}}{\hat{\gamma}} \log(1+\hat{\gamma}t) & \text{if } \hat{\alpha} = 0 \end{cases}$$

6.3.1 Example

From Section 5.5 we obtained

$\hat{\alpha} = 8.78268266064$ $\hat{\beta} = .01083547363$ $\hat{\gamma} = .01071932467$

so $\hat{\Delta}(t) = 142{,}982.4414 \left[1 - \dfrac{1}{(1+.01071932467t)^{8.78268266064}} \right]$

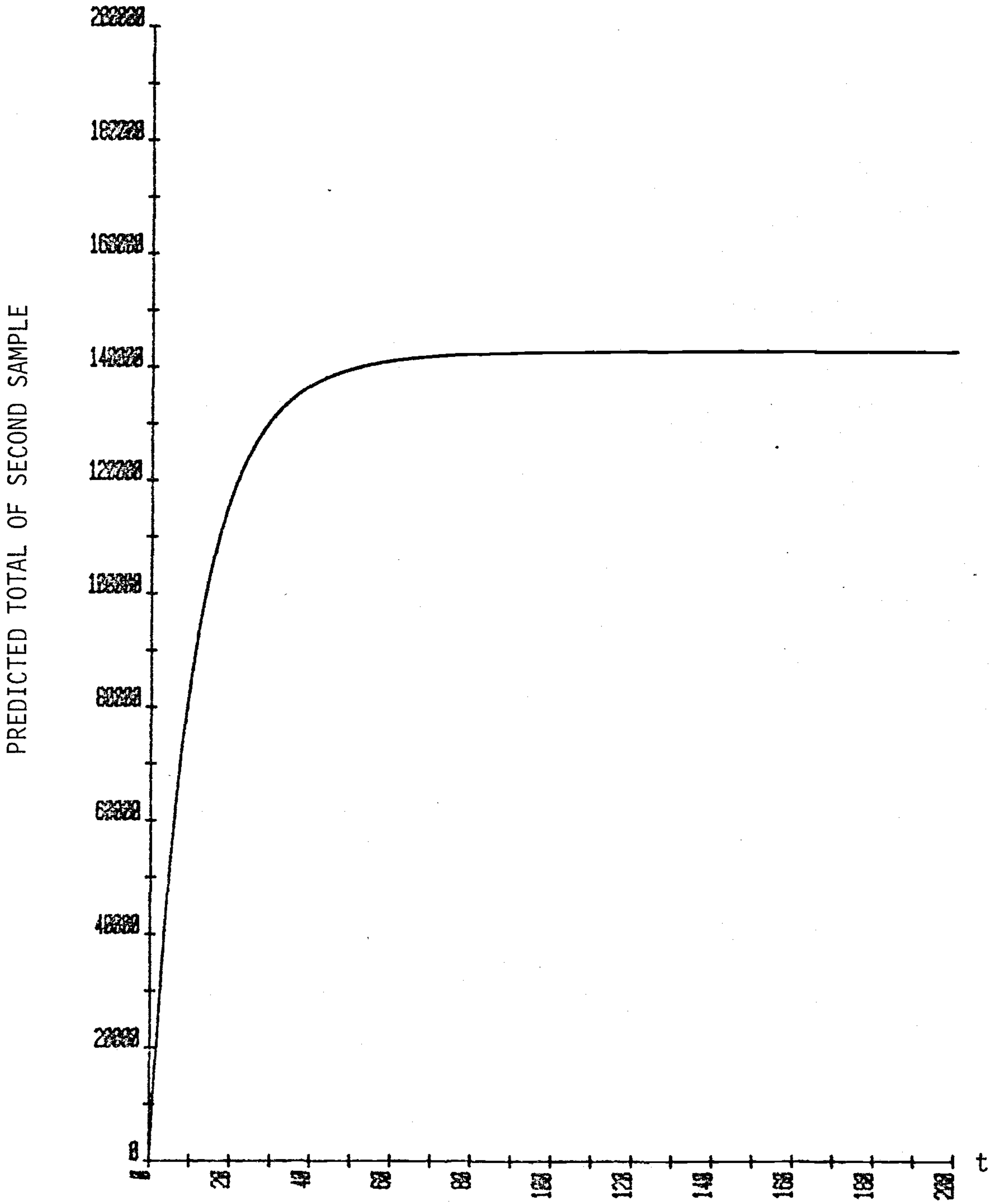
(see Figure 6.1). Hence we can claim $\hat{T} = \hat{\Delta}(\infty) = 142{,}982$. Using the same approach as that of the last chapter, we can find the asymptotic variance of $\hat{\Delta}(t)$

$$\hat{\text{Var}}\left(\hat{\Delta}(\infty)\right) \simeq \begin{cases} 4.29237317 \times 10^9 & \text{without Euler's transformation} \\ 4.258910831 \times 10^9 & \text{with Euler's transformation} \end{cases}$$

The relative standard error is

$$\begin{cases} .46 & \text{without Euler's transformation} \\ .46 & \text{with Euler's transformation} \end{cases} .$$

Figure 6.1



$$\hat{\Delta}(t) = \frac{\sum_{j=1}^{M} Z_j^{(1)}}{\hat{\alpha}\,\hat{\gamma}}\left[1 - \frac{1}{(1+\hat{\gamma}t)^{\hat{\alpha}}}\right] \text{ where } \hat{\alpha} = 8.78268266064,$$

$$\hat{\gamma} = .01071930467 \text{ and } \sum_{j=1}^{M} Z_j^{(1)} = 13,461$$

6.4 Euler's Transformation

Euler's transformation is a method of forcing oscillating series like $\Delta(t) = \sum_{x=1}^{\infty} (-1)^{x+1} \eta_x t^x$, where $\eta_x = E\left[\sum_{j=1}^{M} Z_j^{(x)}\right]$, to converge rapidly.

Efron and Thisted showed

$$\Delta(t) = \sum_{x=1}^{\infty} (-1)^{x+1} \eta_x t^x = \sum_{y=1}^{\infty} \xi_y u^y \text{ where } t = \frac{u}{2-u}, \ 0 \leq u \leq 2,$$

and $\xi_y = \sum_{x=1}^{y} \binom{y-1}{x-1} \frac{(-1)^{x+1}}{2^y} \eta_x$.

6.4.1 Nonparametric Estimator for $\Delta(t)$

Define

$$\Delta_E(u) = \sum_{y=1}^{\infty} \xi_y u^y$$

$$\Delta^{x_0}(t) = \sum_{x=1}^{x_0} (-1)^{x+1} \eta_x t^x$$

$$\Delta_E^{x_0}(u) = \sum_{y=1}^{x_0} \xi_y u^y \ .$$

Good and Toulmin suggest estimating $\Delta(t)$ by

$$\hat{\Delta}^{x_0}(u) = \sum_{y=1}^{x_0} \hat{\xi}_y u^y \quad \text{where} \quad u = \frac{2t}{1+t} \quad \text{and}$$

$\hat{\xi}_y = \sum_{x=1}^{y} \binom{y-1}{x-1} \frac{(-1)^{x+1}}{2^y} \hat{\eta}_x$. The $\hat{\eta}_x$ is taken to be the nonparametric estimator $\sum_{j=1}^{M} Z_j^{(x)}$.

6.4.2 Parametric Estimator for $\Delta(t)$

From (6.3) and (6.6) we know

$$\Delta(t) \approx \eta_1 t - \eta_2 t^2 + \eta_3 t^3 -+ \ldots$$

$$\eta_x = \eta_1 \frac{\Gamma(x+\alpha)}{x!\Gamma(1+\alpha)} \gamma^{x-1} .$$

We obtain $\Delta(t) \approx \eta_1 t \sum_{x=1}^{\infty} (-1)^{x+1} \frac{\Gamma(x+\alpha)}{x!\Gamma(1+\alpha)} (\gamma t)^{x-1}$

which diverges for $\gamma t > 1$. If we estimate η_1, α, and γ, we obtain an estimator of $\Delta(t)$. According to Efron and Thisted, for $-1 < \alpha \leqq 1$, the series $\sum_{y=1}^{\infty} \xi_y u^y$ converges in the nicest possible way, having $\xi_y \geqq 0$ for all y. Using Euler's transformation we obtain the estimator

$$\hat{\Delta}_E^{x_0}(u) = \sum_{y=1}^{x_0} \hat{\xi}_y u^y \quad \text{where } u = \frac{2t}{1+t}$$

and $$\hat{\xi}_y = \sum_{x=1}^{y} \binom{y-1}{x-1} \frac{(-1)^{x+1}}{2^y} \hat{\eta}_1 \frac{\Gamma(x+\hat{\alpha})}{x!\Gamma(1+\hat{\alpha})} \hat{\gamma}^{x-1} .$$

6.4.3 Example

Initially let us consider the parametric estimator $\hat{\Delta}_E^{x_0}(u)$ with Euler's transformation. The values of $\hat{\xi}_y$ are in Table 6.1. One way to choose x_0 is to require $\hat{\Delta}^{x_0}(1) \approx \sum_{j=1}^{M} Y_j = 14{,}165$. This gives $x_0 = 38$, and so we do not consider $\hat{\xi}_y$, $y \geqq 39$. Since $\sum_{y=29}^{38} \xi_y = .00000522259$, we decide to choose $x_0 = 29$. Let us choose $t = 100$. From Figure

y	$\hat{\xi}_y$	y	$\hat{\xi}_y$
1	6730.5	26	.00003380035
2	3188.80362999268	27	.00001514092
3	1509.57766569919	28	.00000673407
4	714.02261275796	29	.00000296968
5	337.42502726722	30	.00000129620
6	159.30509997155	31	.00000055862
7	75.13553619057	32	.00000023690
8	35.39960803598	33	.00000009839
9	16.65943914926	34	.00000003968
10	7.83068586624	35	.00000001535
11	3.67605589976	36	.00000000556
12	1.72333189187	37	.00000000178
13	.80671026984	38	.00000000042
14	.37703393043	39	-.00000000001
15	.17591546659	40	-.00000000011
16	.08192720133	41	-.00000000010
17	.03807890877	42	-.00000000007
18	.01766019281	43	-.00000000005
19	.00817093799	44	-.00000000003
20	.00377060640	45	-.00000000002
21	.00173497792	46	-.00000000001
22	.00079575682	47	-.00000000001
23	.00036366811	48	-0
24	.00016552792	49 and more	
25	.00007499638		

Table 6.1

$$\xi_y = \sum_{x=1}^{y} \binom{y-1}{x-1} \frac{(-1)^{x+1}}{2^y} \hat{\eta}_1 \frac{\Gamma(x+\hat{\alpha})}{x!\Gamma(1+\hat{\alpha})} \hat{\gamma}^{x-1}$$ where $\hat{\eta}_1 = 13{,}461$, $\hat{\eta}_2 = 8.78268266$

and $\hat{\gamma} = .01071932467$

6.1 this seems large enough and if we suppose that $\lambda_j = {}^{1000}/_{14,115}$, the expected fraction of distinct units observed in the second sample is

$$1 - e^{-100\lambda_j} = .9991621419 .$$

We calculate

$$\sum_{j=1}^{M} \hat{y}_j = \hat{\Delta}_E^{29}\left({}^{200}/_{101}\right) = 167,493$$

$$\text{and } \hat{\Delta}_E^{38}\left({}^{200}/_{101}\right) = 172,129 .$$

(see Figure 6.2).

If we consider the nonparametric estimator $\hat{\Delta}(t)$ without Euler's transformation

$$\hat{\Delta}(t) = \hat{\eta}_1 t - \hat{\eta}_2 t^2 + \hat{\eta}_3 t^3 = 13461t - 671t^2 + 33t^3$$

$$= 149,118 \qquad \text{when } t = 14.115$$

The reasons we consider $t = 14.115$ are that $t = {}^N/_n$ and, if there do not exist duplicated cases, then $\sum_{j=1}^{M} \hat{y}_j = \frac{N}{n} \sum_{i=1}^{n} Y_i$ where

$$\sum_{i=1}^{n} Y_i = \sum_{j=1}^{M} Z_j^{(1)} .$$

If we consider the nonparametric estimate of $\hat{\Delta}_E^{x_0}(u)$ with Euler's transformation, we get

$$\hat{\xi}_y = {}^{13,461}/_{2^y} - 671(y-1)/_{2^y} + 33(y-1)(y-2)/_{2^{y+1}}$$

and the table of values of $\hat{\xi}_y$ is in Table 6.2. From this table we

y	ξ_y	y	ξ_y
1	6730.5	27	0.00005021691
2	3197.5	28	0.00002580509
3	1519.0	29	0.00001331232
4	721.6875	30	0.00000689179
5	342.96875	31	0.00000357907
6	163.0625	32	0.00000186381
7	77.578125	33	0.0000097288
8	36.94140625	34	0.00000050885
9	17.611328125	35	0.00000026659
10	8.408203125	36	0.00000013986
11	4.021484375	37	0.00000007345
12	1.92749023438	38	0.00000003861
13	0.92614746094	39	0.00000002030
14	0.4462890625	40	0.00000001068
15	0.21575927734	41	0.00000000562
16	0.10469055176	42	0.00000000296
17	0.05100250244	43	0.00000000156
18	0.02495574951	44	0.00000000082
19	0.01226806641	45	0.00000000043
20	0.00606060028	46	0.00000000023
21	0.00300931931	47	0.00000000012
22	0.00150203705	48	0.00000000006
23	0.00075364113	49	0.00000000003
24	0.00038009882	50	0.00000000002
25	0.00019267201	51	0.00000000001
26	0.00009813905	52 and more	0

Table 6.2

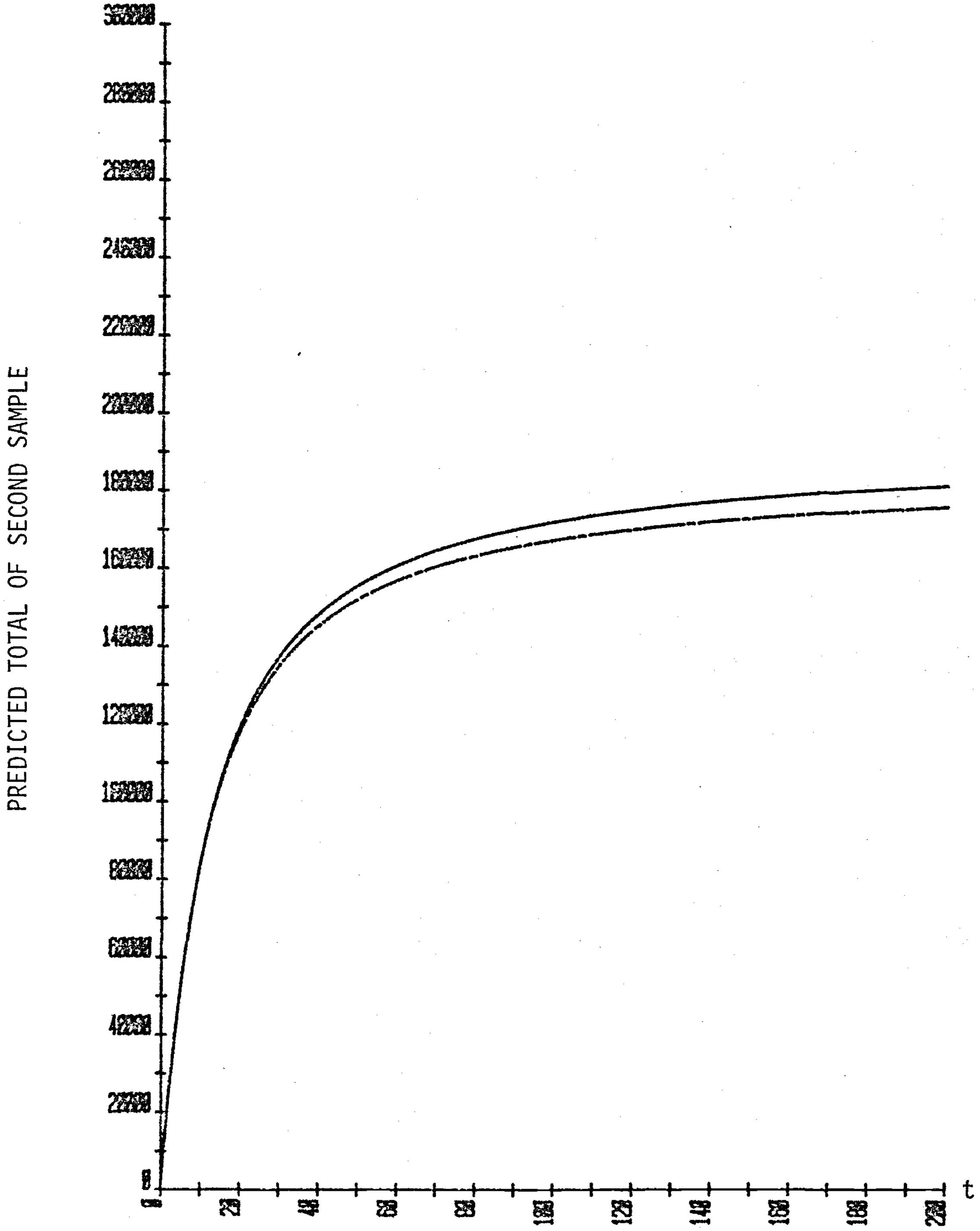
$$\xi_y = \frac{1}{2^y}\hat{n}_1 - \binom{y-1}{1}\frac{1}{2^y}\hat{n}_2 + \binom{y-1}{2}\frac{1}{2^y}\hat{n}_3 \quad \text{where } \hat{n}_x = \sum_{j=1}^{M} Z_j^{(x)} \text{ and}$$

$\hat{n}_1 = 13.461$, $\hat{n}_2 = 671$, $\hat{n}_3 = 33$

y	Accumulative $\hat{\xi}_y$	y	Accumulative $\hat{\xi}_y$
1	12823.0	27	0.00010371208
2	6092.5	28	0.00005349517
3	2895.0	29	0.00002769008
4	1376.0	30	0.00001437776
5	654.3125	31	0.00000748597
6	311.34375	32	0.00000390690
7	143.28125	33	0.00000204309
8	70.703125	34	0.00000107021
9	33.76171875	35	0.00000056135
10	16.150390625	36	0.00000029476
11	7.7421875	37	0.00000015491
12	3.720703125	38	0.00000008145
13	1.79321289064	39	0.00000004285
14	0.86706542968	40	0.00000002254
15	0.42077636719	41	0.00000001186
16	0.20501708984	42	0.00000000624
17	0.10032653809	43	0.00000000328
18	0.04932403564	44	0.00000000173
19	0.02436828613	45	0.00000000091
20	0.01210021973	46	0.00000000048
21	0.00603961945	47	0.00000000025
22	0.00303030014	48	0.00000000013
23	0.00152826309	49	0.00000000007
24	0.00077462196	50	0.00000000004
25	0.00039452314	51	0.00000000002
26	0.00020185113	52	0.00000000001
		53 and more	0

Table 6.3: Accumulative $\hat{\xi}_y$ from Table 6.2.

Figure 6.2



$\hat{\Delta}_E^{x_0}(u)$, where $u = \frac{2t}{1+t}$, in Section 6.4.2

——— means $\hat{\Delta}_E^{38}(u)$.

– – – means $\hat{\Delta}_E^{29}(u)$

know we can choose $x_0 = 31$ since $\sum_{x=31}^{\infty} \hat{\xi}_y < .00001$ (see Table 6.3). We calculate $\hat{\Delta}_E^{31}\left(^{200}/_{101}\right) = 221{,}314$. This is the value we claim for the estimate of $\sum_{j=1}^{M} y_j$ (see Figure 6.3). Note $\hat{\Delta}_E^{29}\left(^{200}/_{101}\right) = 210{,}177$.

6.5 The Bias of $\hat{\Delta}(t)$

From the expressions for $\hat{\Delta}(t)$ and $\hat{\Delta}^{x_0}(t)$ in Section 6.4, we see that it would be difficult to find their variances. In this section we try to find their biases. Using Euler's transformation and substituting $u = \frac{2t}{1+t}$, we have

$$\hat{\Delta}^{x_0}(t) = \sum_{x=1}^{x_0} (-1)^{x+1} \hat{\eta}_x t^x \sum_{y=x}^{x_0} \binom{y-1}{x-1} \left(\frac{1}{1+t}\right)^x \left(\frac{t}{1+t}\right)^{y-x}$$

Define

$$h_x^{x_0} = (-1)^{x+1} t^x \sum_{y=x}^{x_0} \binom{y-1}{x-1} \left(\frac{1}{1+t}\right)^x \left(\frac{t}{1+t}\right)^{y-x}, \text{ and}$$

$$h_x = (-1)^{x+1} t^x \sum_{y=x}^{\infty} \binom{y-1}{x-1} \left(\frac{1}{1+t}\right)^x \left(\frac{t}{1+t}\right)^{y-x},$$

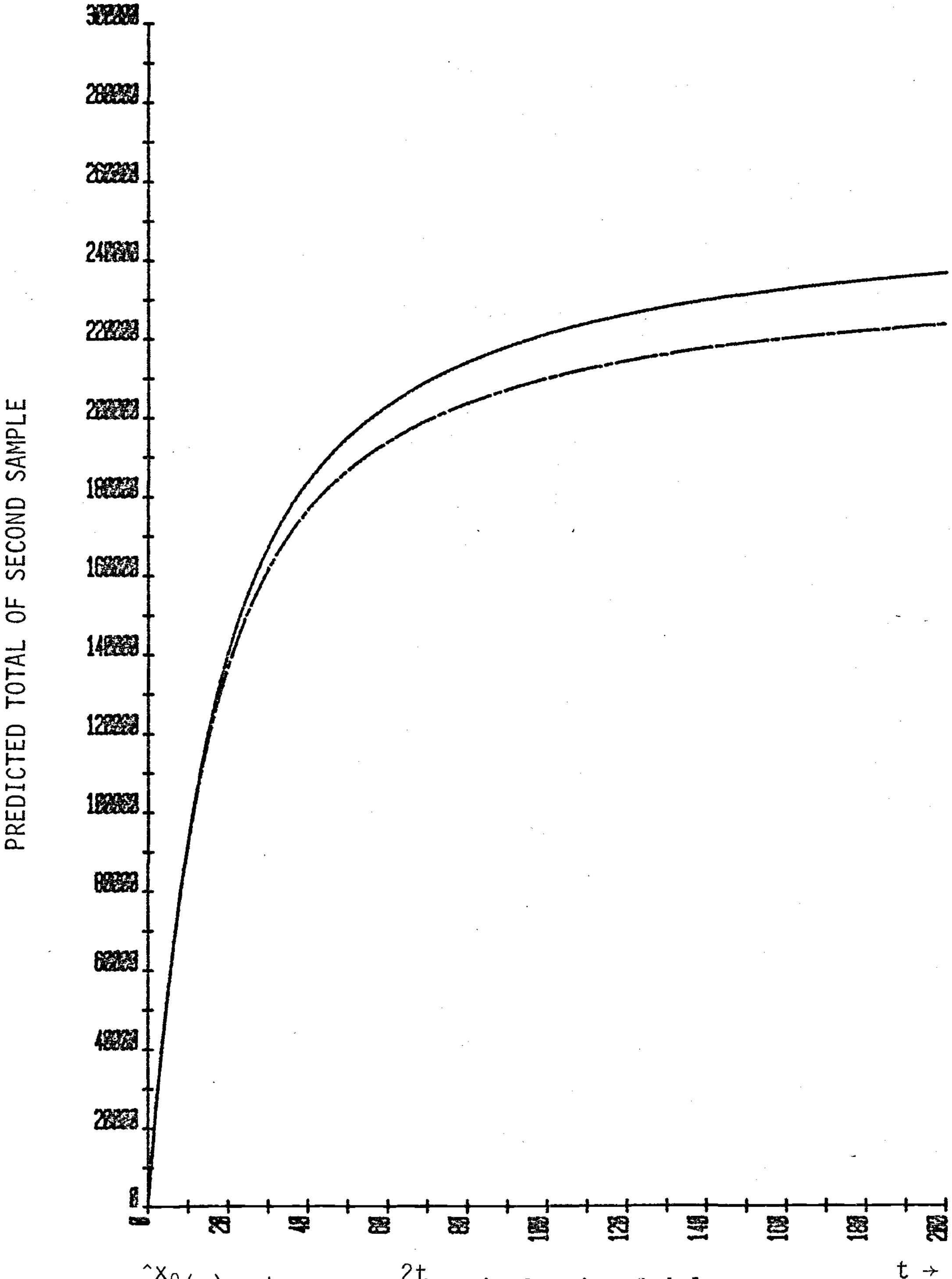
so that

$$\hat{\Delta}^{x_0}(t) = \sum_{x=1}^{x_0} h_x^{x_0} \hat{\eta}_x$$

and

$$\hat{\Delta}(t) = \sum_{x=1}^{\infty} h_x \hat{\eta}_x \quad \text{where } \hat{\eta}_x = \sum_{j=1}^{M} Z_j^{(x)} .$$

Figure 6.3



$\hat{\Delta}_E^{x_0}(u)$, where $u = \frac{2t}{1+t}$, in Section 6.4.1

——— means $\hat{\Delta}_E^{31}(u)$

—-·— means $\hat{\Delta}_E^{29}(u)$

Define $H(\lambda) = \sum_{x=1}^{\infty} h_x \lambda^x / x!$ where $0 < \lambda < \infty$

and $H^{x_0}(\lambda) = \sum_{x=1}^{x_0} h_x^{x_0} \lambda^x / x!$

Then

$$E[\hat{\Delta}(t)] = \sum_{x=1}^{\infty} h_x \eta_x = \sum_{x=1}^{\infty} h_x \left(\sum_{j=1}^{M} y_j\right) \int_0^{\infty} \frac{e^{-\lambda}\lambda^x}{x!} \, dG(\lambda)$$

$$= \left(\sum_{j=1}^{M} y_j\right) \int_0^{\infty} e^{-\lambda} H(\lambda) dG(\lambda)$$

$$E\left\{\hat{\Delta}(t) - \Delta(t)\right\} = \left(\sum_{j=1}^{M} y_j\right) \int_0^{\infty} e^{-\lambda}\left[H(\lambda) - \left(1 - e^{-\lambda t}\right)\right] dG(\lambda)$$

which, for $t = \infty$, becomes

$$E\left\{\hat{\Delta}(\infty) - \Delta(\infty)\right\} = \left(\sum_{j=1}^{M} y_j\right) \int_0^{\infty} e^{-\lambda}\left[H(\lambda) - 1\right] dG(\lambda) .$$

It is convenient to rewrite this in a form which depends on

$\eta_+ = \sum_{x=1}^{\infty} \eta_x$ rather than $\sum_{j=1}^{M} y_j$. Define

$$P = \int_0^{\infty} \left(1 - e^{-\lambda}\right) dG(\lambda)$$

$$d\tilde{G}(\lambda) = \frac{1 - e^{-\lambda}}{P} \, dG(\lambda) .$$

Since $\eta_+ = \sum_{x=1}^{\infty} \eta_x = \sum_{x=1}^{\infty} \left(\sum_{j=1}^{M} y_j\right) \int_0^{\infty} \frac{e^{-\lambda}\lambda^x}{x!} \, dG(\lambda) = \left(\sum_{j=1}^{M} y_j\right) \int_0^{\infty} \left(1 - e^{-\lambda}\right) dG(\lambda)$

$$= \left(\Sigma y_j\right) P ,$$

$$E\left\{\hat{\Delta}(t)-\Delta(t)\right\} = \left(\sum_{j=1}^{M} y_j\right)\int_0^\infty e^{-\lambda}\left[H(\lambda) - \left(1 - e^{-\lambda t}\right)\right] dG(\lambda)$$

$$= \frac{\frac{1 - e^{-\lambda}}{P}}{\frac{1 - e^{-\lambda}}{P}}\left(\sum_{j=1}^{M} y_j\right)\int_0^\infty e^{-\lambda}\left[H(\lambda) - \left(1 - e^{-\lambda t}\right)\right] dG(\lambda)$$

$$= n_+\int_0^\infty \frac{e^{-\lambda}}{1 - e^{-\lambda}}\left[H(\lambda) - \left(1 - e^{-\lambda t}\right)\right] d\tilde{G}(\lambda)$$

and

$$E\{\hat{\Delta}(\infty)-\Delta(\infty)\} = n_+\int_0^\infty \frac{e^{-\lambda}}{1 - e^{-\lambda}}\left[H(\lambda) - 1\right] d\tilde{G}(\lambda) .$$

Similarly

$$E\{\hat{\Delta}^{x_0}(t)-\Delta(t)\} = n_+\int_0^\infty \frac{e^{-\lambda}}{1 - e^{-\lambda}}\left[H^{x_0}(\lambda) - \left(1 - e^{-\lambda t}\right)\right] d\tilde{G}(\lambda) ,$$

and for $t = \infty$

$$E\{\hat{\Delta}^{x_0}(\infty)-\Delta(\infty)\} = n_+\int_0^\infty \frac{e^{-\lambda}}{1 - e^{-\lambda}}\left[H^{x_0}(\lambda) - 1\right] d\tilde{G}(\lambda) .$$

We use the integrands

$$B_t(\lambda) = \frac{e^{-\lambda}}{1 - e^{-\lambda}}\left[H(\lambda) - \left(1 - e^{-\lambda t}\right)\right]$$

$$B_t^{x_0}(\lambda) = \frac{e^{-\lambda}}{1 - e^{-\lambda}}\left[H^{x_0}(\lambda) - \left(1 - e^{-\lambda t}\right)\right]$$

to measure the bias of $\hat{\Delta}$ for any $G(\lambda)$.

6.5.1 Example

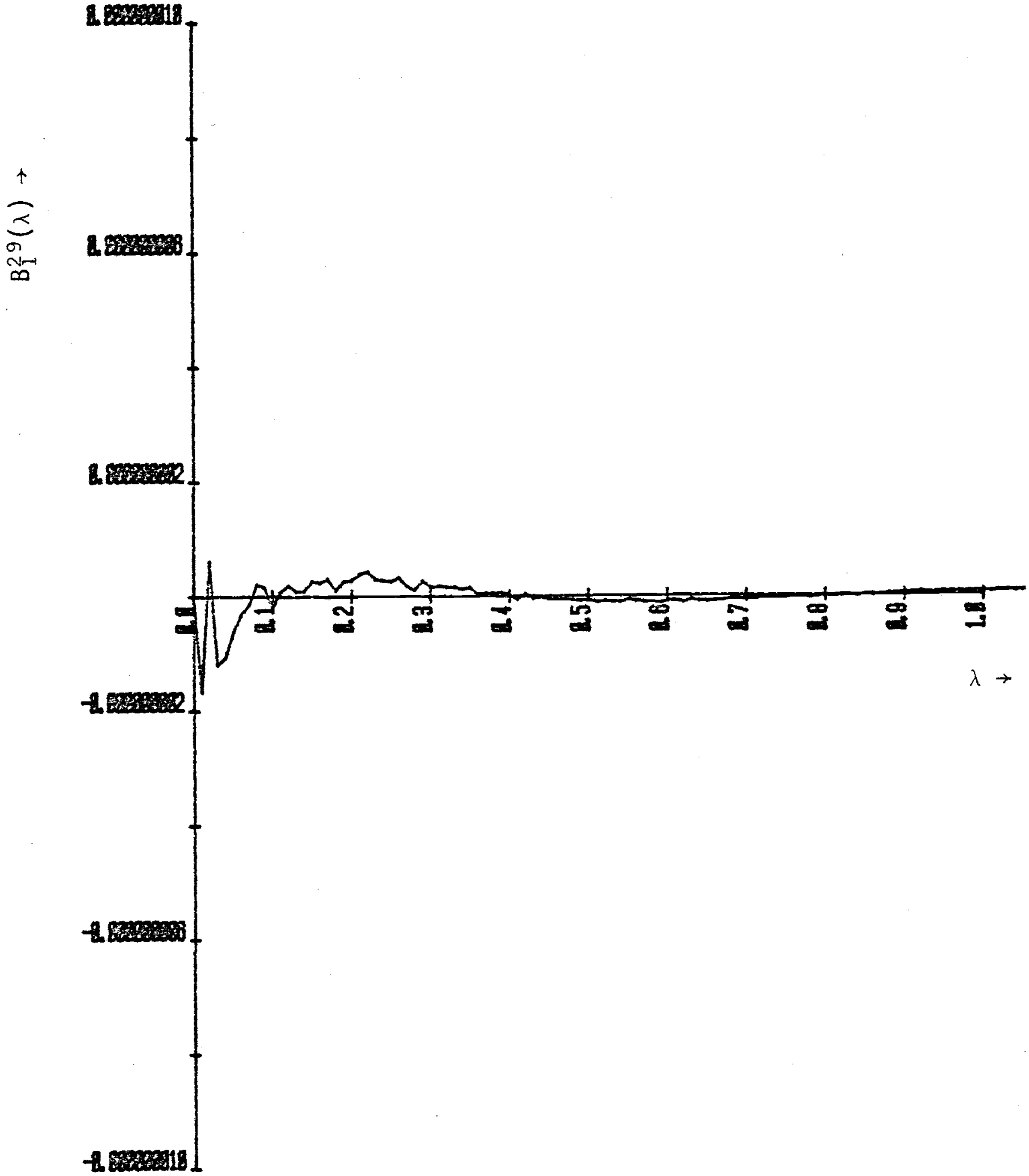
We compute $B_t^{x_0}(\lambda)$ in Table 6.4 and Figures 6.4, 6.5 and 6.6. The maximum bias of $\hat{\Delta}_E^{x_0}\left(= n_+\left\{\max_\lambda B_t^{x_0}(\lambda)\right\}\right)$ is .00000694085 for $x_0 = 29$, t=1; .00000198310 for $x_0 = 31$, t=1; 1,062,375 for $x_0 = 29$, t=100; 1,034,045 for $x_0 = 31$, t=100; and the relative bias $\left(= \text{Bias}/\hat{\Delta}^{x_0}(t)\right)$ is: $.54 \times 10^{-9}$ for $x_0 = 29$, t=1 and the parametric model with the gamma distribution; $.15 \times 10^{-9}$ for $x_0 = 31$, t=1, and the nonparametric model; 6.34 for $x_0 = 29$, t=100, and the parametric model with the gamma distribution; 4.67 for $x_0 = 31$, t=100, and the nonparametric model.

$B_t^{x_0}(\lambda)$	1×10^{-11}	$\frac{1000}{14115}$	$\frac{2000}{14115}$	$\frac{3000}{14115}$	$\frac{4000}{14115}$	$\frac{5000}{14115}$	$\frac{6000}{14115}$	$\frac{7000}{14115}$	$\frac{8000}{14115}$	$\frac{9000}{14115}$
$B_1^{29}(\lambda)$	0	$-.49\times10^{-9}$	1×10^{-11}	$.3\times10^{-9}$	$.23\times10^{-9}$	$.7\times10^{-10}$	$-.4\times10^{-10}$	$-.8\times10^{-10}$	$-.11\times10^{-9}$	$-.9\times10^{-10}$
$B_1^{31}(\lambda)$	0	$-.8\times10^{-10}$	$-.5\times10^{-10}$	$-.14\times10^{-9}$	$.8\times10^{-10}$	$.3\times10^{-10}$	$-.2\times10^{-10}$	0	$-.2\times10^{-10}$	$-.1\times10^{-10}$
$B_{100}^{29}(\lambda)$	-74.99999999930	1.29000387654	2.08045834322	0.7447105037	-.13649066111	-.49577148252	-.52078000274	-.38115206215	-.19140356295	-.01734218806
$B_{100}^{31}(\lambda)$	-72.99999999930	1.65384757369	1.99291149494	0.55014339244	-.28611353371	-.55553788581	-.49916822146	-.30762042606	-.09824700846	.06921067045

Table 6.4

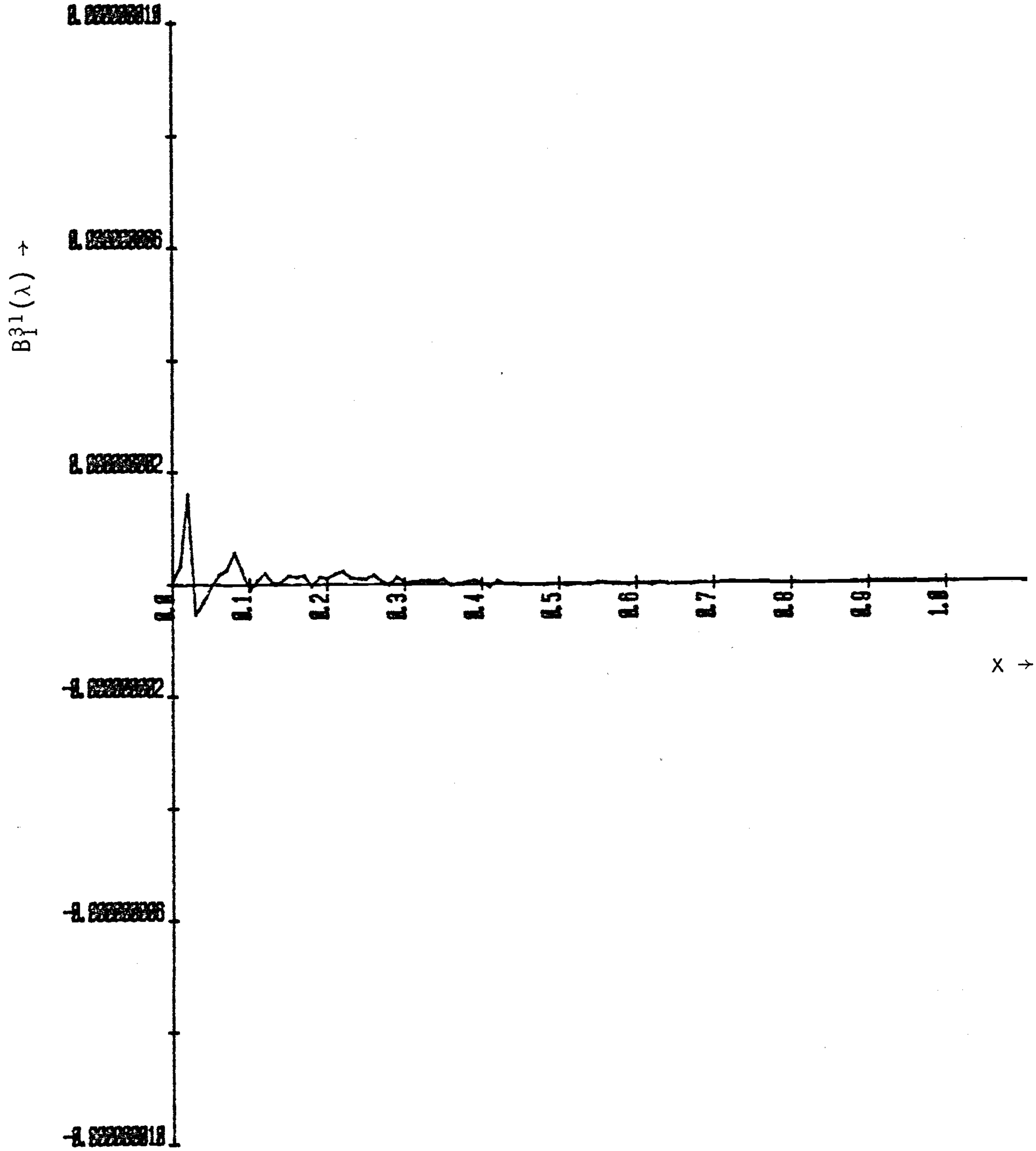
The Bias Function $B_t^{x_0}(\lambda)$; in Section 6.5, for $\hat{\Delta}^{x_0}(t)$, at x_0 = 29 or x_0 = 31 and t = 1 or t = 100

Figure 6.4



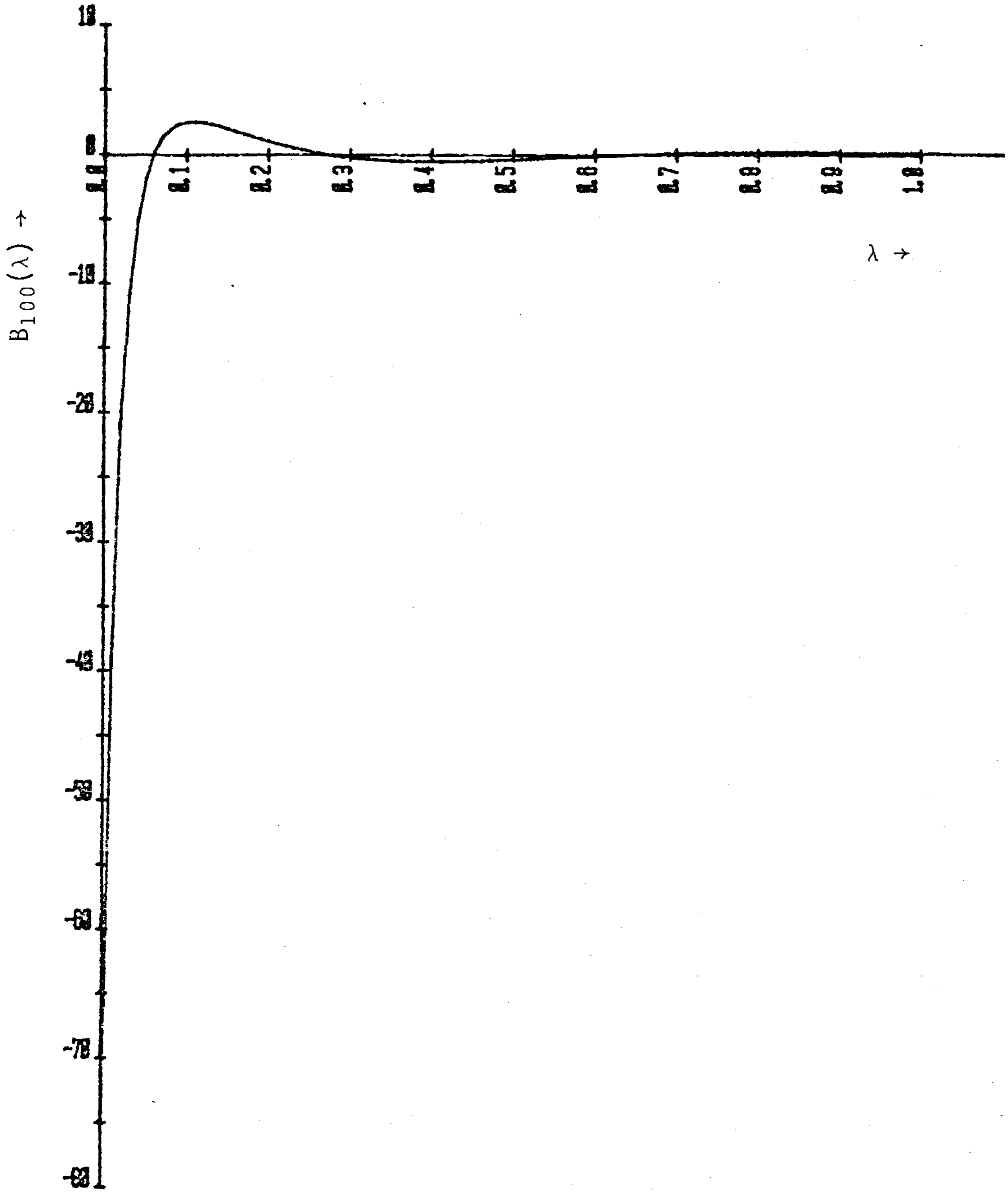
The Bias Function $B_I^{29}(\lambda)$, in Section 6.5, for $\hat{\Delta}^{29}(1)$.

Figure 6.5 for $B_I^{31}(\lambda)$



The Bias Function $B_I^{31}(\lambda)$, in Section 6.5, for $\hat{\Delta}^{31}(1)$.

Figure 6.6



The Bias Function $B^{29}_{100}(\lambda)$ and $B^{31}_{100}(\lambda)$, in Section 6.5, for

$B^{29}_{100}(\lambda)$ and $B^{31}_{100}(\lambda)$.

CHAPTER 7

SUMMARY

In the literature there are five methods for estimating the population size when sampling from a list that contains duplication and when the extent of duplication cannot be determined. In this thesis these methods are generalized to estimate population totals when a measurement is associated with each member of the population. Also, the variances of those estimates are estimated.

The five estimators are illustrated and compared for a population of size N = 14,115 with M = 12,000 distinct classes, 9,885 of them having 1 unit and 2,115 of them having 2 units. The measurements y_j, j=1, 2, ... , 12,000, are assumed to be Poisson distributed with mean 15. In other words, the expected population total is 180,000. We simulate two samples of size n = 1,000, the first sampling without replacement (Goodman's method) and the second sampling with replacement for the other methods. The five sampling methods compared as follows:

(1) By Goodman's method we have an unbiased estimate

$$\widehat{\sum_{j=1}^{M} y_j} = \Sigma_{r=1}^{n} A_r \Sigma_{j=1}^{M} Z_j^{(r)} = 163{,}652, \text{ where } A_r$$

$$= 1 - (-1)^r \frac{[N-n+r-1]^{(r)}}{n^{(r)}}, \text{ with relative standard}$$

error .058.

(2) By Good and Toulmin's method we have

$$\sum_{j=1}^{M} \hat{y}_j = \Sigma_{j=1}^{M} \tilde{Y}_j(N) = \Sigma_{r=1}^{n} \Sigma_{j=1}^{M} Z_j^{(r)} + \left(\frac{N}{n} - 1\right) \Sigma_{j=1}^{M} Z_j^{(1)} - \frac{\left(\frac{N}{n} - 1\right)^2}{\frac{N}{n}} \Sigma_{j=1}^{M} Z_j^{(2)} = 182{,}529$$

with relative standard error .036.

(3) By Harris' method for obtaining the upper and lower bounds of a population total we have

$$\sup \sum_{j=1}^{M} Y_j(N) = \sum_{j=1}^{M} Y_j + (t-1) \sum_{j=1}^{M} Z_j^{(1)} = 190{,}706$$

$$\inf \sum_{j=1}^{M} Y_j(N) = \sum_{j=1}^{M} Y_j = 14{,}165.$$

(4) By Good and Rao's method we have

$$\sum_{j=1}^{M} \hat{y}_j = \frac{\sum_{j=1}^{M} Z_j^{(1)}}{\frac{\Gamma(\hat{\alpha}+1)}{\Gamma(\hat{\alpha})} \frac{\hat{\beta}}{(1+\hat{\beta})^{\hat{\alpha}+1}}} = 157{,}177$$ with relative standard error .36.

(5) By Efron and Thisted's method we have

$$\sum_{j=1}^{M} \hat{y}_j = \frac{\sum_{j=1}^{M} Z_j^{(1)}}{\hat{\alpha}\hat{\gamma}} \left[1 - \frac{1}{(1+\hat{\gamma}t)}\right] = 142{,}982$$ with relative standard error .45.

$\sum_{j=1}^{M} \hat{y}_j = \hat{\Delta}_E^{29}(u) = 167{,}493$ in Section 6.4.2, with relative bias 6.34.

$\sum_{j=1}^{M} \hat{y}_j = \hat{\Delta}_E^{31}(u) = 221{,}314$, in Section 6.4.1, with relative bias 4.67.

Goodman's method does not involve any approximation. Good and Toulmin's method is based on some approximation but less than the other methods. Furthermore the relative standard deviations of these two estimators are small. Since Good and Toulmin's method and Efron and Thisted's method are to find the prediction of population total, they can be applied for the growing population. Since the precision of Good and Rao's method is low and Efron and Thisted's method even lower, extreme care should be exercised if either of these methods is employed.

BIBLIOGRAPHY

[1] Bromwich, T. (1926) An Introduction to the Theory of Infinite Series. 2nd edition. Cambridge University Press.

[2] Efron, B. and R. Thisted. (1975) Estimating the number of unseen species (How many words did Shakespeare know?). Technical Report No. 70, Department of Statistics, Stanford University.

[3] Engene, S. (1974) "On species frequency models," Biometrika 61: 201-208.

[4] Engene, S. (1977) "Comments on Two Different Approaches to the Analysis of Species Frequency Data," Biometrics 33:205-213.

[5] Fisher, R.A., A.S. Corbet and C.B. Williams. (1943). "The Relation Between the Number of Species and the Number of Individuals in a Random Sample of an Animal Population," The Journal of Animal Ecology 12:42-58.

[6] Good, I.J. (1953) "The Population Frequencies of Species and the Estimation of Population Parameters," Biometrika 40:237-264.

[7] Good, I.J. and G.H. Toulmin. (1956) "The Number of New Species, and the Increase in Population Coverage, When a Sample is Increased," Biometrika 43:45-63.

[8] Goodman, L.A. (1949) "On the Estimation of the Number of Classes in a Population," The Annals of Mathematical Statistics 20: 521-548.

[9] Hardy, G.H. (1949) Divergent Series. Oxford Clarendon Press.

[10] Harris, B. (1959) "Determining Bounds on Integrals with Application to Cataloguing Problems," The Annals of Mathematical Statistics 30:521-548.

[11] Knott, M. (1967) "Models for Cataloguing Problems," The Annals of Mathematical Statistics 38:1255-1260.

[12] Rao, C.R. (1965) Linear Statistical Inference and its Applications. John Wiley and Sons, New York.

[13] Rao, C.R. (1971) "Some Comments on the Logarithmic Series Distribution in the Analysis of Insect Trap Data," Statistical Ecology, Vol. 1, Ed. G.P. Patil, E.C. Pielou & W.E. Waters, Pennsylvania State University Press:131-142.

[14] Rao, J.N.K. (1968) "Some Nonresponse Sampling Theory When The Frame Contains An Unknown Amount of Duplication," Journal of the American Statistical Association 63:87-90.

[15] Schuster, J.J. (1973) The Validation of a Sample Survey. Technical Report No. 65, Department of Statistics, University of Florida.