MARTIN MARIETTA ENERGY SYSTEMS LIBRARIES
3 4456 0361502 3

4
CENTRAL RESEARCH LIBRARY
DOCUMENT COLLECTION

ORNL-2957
UC-34 - Physics and Mathematics

AN EMPIRICAL STUDY OF THE SAMPLING DISTRIBUTIONS OF SOME DISEASE INCIDENCE ESTIMATES

G. J. Atta

CENTRAL RESEARCH LIBRARY
DOCUMENT COLLECTION
LIBRARY LOAN COPY
DO NOT TRANSFER TO ANOTHER PERSON
If you wish someone else to see this document, send in name with document and the library will arrange a loan.

OAK RIDGE NATIONAL LABORATORY
operated by
UNION CARBIDE CORPORATION
for the
U.S. ATOMIC ENERGY COMMISSION

Printed in USA. Price $1.75. Available from the
Office of Technical Services
Department of Commerce
Washington 25, D.C.

LEGAL NOTICE

This report was prepared as an account of Government sponsored work. Neither the United States, nor the Commission, nor any person acting on behalf of the Commission:

A. Makes any warranty or representation, expressed or implied, with respect to the accuracy, completeness, or usefulness of the information contained in this report, or that the use of any information, apparatus, method, or process disclosed in this report may not infringe privately owned rights; or

B. Assumes any liabilities with respect to the use of, or for damages resulting from the use of any information, apparatus, method, or process disclosed in this report.

As used in the above, "person acting on behalf of the Commission" includes any employee or contractor of the Commission, or employee of such contractor, to the extent that such employee or contractor of the Commission, or employee of such contractor prepares, disseminates, or provides access to, any information pursuant to his employment or contract with the Commission, or his employment with such contractor.

ORNL-2957

Contract No. W-7405-eng-26

AN EMPIRICAL STUDY OF THE SAMPLING DISTRIBUTIONS OF SOME DISEASE INCIDENCE ESTIMATES

George J. Atta

DATE ISSUED

JUL - 5 1960

Submitted as a thesis to the Faculty of the Graduate Council of the University of Tennessee in partial fulfillment of the requirements for the degree of Master of Arts in Mathematics

OAK RIDGE NATIONAL LABORATORY
Oak Ridge, Tennessee
operated by
UNION CARBIDE CORPORATION
for the
U. S. ATOMIC ENERGY COMMISSION

MARTIN MARIETTA ENERGY SYSTEMS LIBRARIES

3 4456 0361502 3

TABLE OF CONTENTS

CHAPTER		PAGE
I.	INTRODUCTION	1
	Statement of the problem	1
	Importance of the study	2
	Review of the related literature	2
	Organization of the study by chapters	3
II.	ESTIMATION OF DISEASE INCIDENCES	4
	General description	4
	Estimation of conditional probabilities	6
	Estimation of disease incidences in the unrestricted population	10
	Estimation of disease incidences in the truncated population	11
	Approximate variances of the estimates	14
III.	ASYMPTOTIC VARIANCES OF THE ESTIMATES	16
	Asymptotic variance-covariance matrix of the conditional probabilities	16
	Asymptotic variances of the incidence estimates	22
IV.	RANDOM SAMPLING EXPERIMENT	30
	Introduction	30
	Description of the study	31
	Generation of the pseudo-random numbers	34
	Generation of the sample elements	36
	Computation of the sample moments	37
	Distribution of the estimates	42
	Bias of the estimates	48

CHAPTER PAGE

V. SUMMARY AND CONCLUSIONS 64

BIBLIOGRAPHY 67

LIST OF TABLES

TABLE		PAGE
I.	The Conditional Probabilities, p_{ij}, for Case (1c)	32
II.	The Unconditional Probabilities, $p_{ij}R_j$, for Case (1c)	33
III.	Moment Statistics of the Distribution of $\hat{I}_i'$, $E(r_j) = E(r_j')$, $n = 5$	43
IV.	Moment Statistics of the Distribution of $\hat{I}_i'$, $E(r_j) = E(r_j')$, $n = 10$	44
V.	Moment Statistics of the Distribution of $\hat{I}_i'$, $E(r_j) = E(r_j')$, $n = 20$	45
VI.	Moment Statistics of the Distribution of $\hat{I}_i'$, $E(r_j) \neq E(r_j')$, $n = 5$	46
VII.	Moment Statistics of the Distribution of $\hat{I}_i'$, $E(r_j) \neq E(r_j')$, $n = 10$	47
VIII.	Bias and Standard Deviation of the Bias of $\hat{I}_i'$, $n = 5$	58
IX.	Bias and Standard Deviation of the Bias of $\hat{I}_i'$, $n = 10$	59
X.	Bias and Standard Deviation of the Bias of $\hat{I}_i'$, $n = 20$	60

LIST OF FIGURES

FIGURE PAGE

1. Flow Diagram of the Sampling Experiment 35
2. Distribution of 1,000 $\hat{I}_2'$, $E(r_j) = E(r_j')$, $n = 10$, $N = 400$ 49
3. Distribution of 1,000 $\hat{I}_3'$, $E(r_j) = E(r_j')$, $n = 10$, $N = 400$ 50
4. Distribution of 1,000 $\hat{I}_4'$, $E(r_j) = E(r_j')$, $n = 10$, $N = 400$ 51
5. Distribution of 1,000 Approximate Variances of $\hat{I}_2'$, $E(r_j) = E(r_j')$, $n = 10$, $N = 400$ 52
6. Distribution of 1,000 Approximate Variances of $\hat{I}_3'$, $E(r_j) = E(r_j')$, $n = 10$, $N = 400$ 53
7. Distribution of 1,000 Approximate Variances of $\hat{I}_4'$, $E(r_j) = E(r_j')$, $n = 10$, $N = 400$ 54
8. Distribution of 1,000 Asymptotic Variances of $\hat{I}_2'$, $E(r_j) = E(r_j')$, $n = 10$, $N = 400$ 55
9. Distribution of 1,000 Asymptotic Variances of $\hat{I}_3'$, $E(r_j) = E(r_j')$, $n = 10$, $N = 400$ 56
10. Distribution of 1,000 Asymptotic Variances of $\hat{I}_4'$, $E(r_j) = E(r_j')$, $n = 10$, $N = 400$ 57

CHAPTER I

INTRODUCTION

The estimation of disease incidences in populations subject to competing risks has been a recognized problem among statisticians for many years. For example, in studies on cancer mortality individuals may die from causes other than cancer. Disease incidence estimation has been attacked in many ways usually yielding different results. It is the writer's belief, however, that none of the previous authors have approached this problem from the maximum likelihood point of view with the aim of studying the sampling distributions of the estimates.

Statement of the problem. The general problem is illustrated most easily by reference to a particular experiment. A group of animals is exposed to a sublethal dose of radiation. The group is observed until all animals have died. At the time of death each animal is autopsied, and a cause of death is assigned. The experimenter wishes to estimate the incidence of disease A. There is, however, a complicating factor. Suppose disease B generally occurs earlier in life than does disease A. Thus, there are animals which die of disease B before they manifest disease A. The problem considered is twofold: (1) to estimate what the incidence of disease A would be if disease B were eliminated as a cause of death when the estimation is performed with a sample from a population in which all diseases are operating and (2) to study the properties of the sampling distributions of these estimates.

For the answer to the first part of the problem Kimball [7] derived a distribution free estimation procedure and gave an approximate formula for the variance of the estimate. It was felt that this

approximate variance formula was inadequate. A search for a better variance formula and an answer to the second part of the above problem, as related to Kimball's estimates, led to the results of this paper.

Importance of the study. For any estimation procedure it is most desirable to know the small sample characteristics of the estimates. Kimball's procedure is being used extensively at Oak Ridge National Laboratory. Unfortunately, in this case the complexity of the mathematics prohibited a complete analytical treatment. Thus, an empirical method was used in an endeavor to learn more about the small sample properties of the estimates.

Review of the related literature. Berkson and Gage [1] developed methods for estimating the "cure rate" and the "death rate" from cancer in human populations after treatment for cancer. Kimball and Atta [8] derived a procedure for estimating disease incidence probabilities in animal populations subject to multiple causes of death. These procedures require the assumption of specific probability distributions for the survival times in the populations considered.

Fix and Neyman [3] discussed a method for estimating "rates of risks" in patients following treatment for cancer. Their method is based on the stochastic process approach.

Harris, Meier, and Tukey [4] derived estimates of "event-rates" on the basis of a sample of observations from a distribution of such events. Kaplan and Meier [6] were concerned with the estimation of "survival rates" in individuals. In these procedures no assumption is made about the form of the distribution of survival times in the populations considered.

Sampford [12] discussed methods for estimating the response-time

distributions. These methods may be adapted to the estimation of disease incidences if certain assumptions are made about the distributions of times of death in the population. Such information, however, is usually lacking.

Organization of the study by chapters. As was indicated earlier the major part of the study reported in this paper was concerned with the small sample properties of the estimates obtained by Kimball's distribution free estimation procedure. For completeness Kimball's procedure is discussed in Chapter II.

A new asymptotic variance formula for the estimates is developed in Chapter III. Included is a derivation of the asymptotic variance-covariance matrix of the estimated conditional probabilities.

The small sample distribution and its important characteristics for Kimball's estimates of disease incidence probabilities were estimated by a random sampling (Monte Carlo) experiment performed on the ORACLE (Oak Ridge Automatic Computer and Logical Engine). The description and results of that experiment are reported in Chapter IV.

Chapter V contains general conclusions and a recommendation for further research.

CHAPTER II

ESTIMATION OF DISEASE INCIDENCES

Kimball [7] presents a method for the estimation of disease incidence probabilities in populations subject to competing risks. This chapter is virtually a copy of his paper. Many of the missing algebraic details of Kimball's derivation are supplied here.

General description. Assume a random sample of N animals from an infinite population of animals. When an animal dies, the cause of death is recorded. After all N animals in the sample have died, the observation period is divided into n intervals, not necessarily of equal length; and the number dying of each disease is calculated for each time interval.

Let a_{ij} $(i = 1, 2, \ldots, m;\ j = 1, 2, \ldots, n)$ denote the number of animals that died of cause i during the jth interval. Let the total number of animals that died of cause i be specified by

$$A_i = \sum_{j=1}^{n} a_{ij} , \tag{2.1}$$

and the total number of deaths in the jth interval by

$$r_j = \sum_{i=1}^{m} a_{ij} . \tag{2.2}$$

Then

$$N = \sum_{i=1}^{m} A_i = \sum_{j=1}^{n} r_j . \tag{2.3}$$

Define the conditional probability, p_{ij} $(i = 1, \ldots, m;$ $j = 1, \ldots, n)$, to be the probability of death in the jth interval from the ith cause, given survival through the first (j-1) intervals. Then

the probability of death in the jth interval given survival through the first (j-1) intervals takes the form

$$s_j = \sum_{i=1}^{m} p_{ij} \ . \tag{2.4}$$

The model, by the selection of the intervals, imposes the conditions

$$s_0 = 0 \quad \text{and} \quad s_n = 1 \ . \tag{2.5}$$

As a consequence of the definition of the p_{ij}, the probability of survival through the first (j-1) intervals is

$$R_j = \prod_{\ell=0}^{j-1} (1-s_\ell) \ . \tag{2.6}$$

Thus, one finds the unconditional probability of death in the jth interval from the ith cause to be $p_{ij} R_j$.

It will be shown that

$$\sum_{i=1}^{m} \sum_{j=1}^{n} p_{ij} R_j = 1 \ . \tag{2.7}$$

Write

$$\sum_{i=1}^{m} \sum_{j=1}^{n} p_{ij} R_j = \sum_{j=1}^{n} R_j \sum_{i=1}^{m} p_{ij} \ . \tag{2.8}$$

Using (2.4) and (2.6), one may write (2.8) in the form

$$\begin{aligned} \sum_{i=1}^{m} \sum_{j=1}^{n} p_{ij} R_j &= \sum_{j=1}^{n} R_j s_j \\ &= s_1(1-s_0) + s_2(1-s_0)(1-s_1) + s_3(1-s_0)(1-s_1)(1-s_2) \\ &\quad + \ldots + s_n(1-s_0)(1-s_1)(1-s_2) \ldots (1-s_{n-1}) \ . \end{aligned} \tag{2.9}$$

Since $s_0 = 0$,

$$\begin{aligned} \sum_{i=1}^{m} \sum_{j=1}^{n} p_{ij} R_j &= s_1 + s_2(1-s_1) + s_3(1-s_1)(1-s_2) \\ &\quad + \ldots + s_n(1-s_1)(1-s_2) \ldots (1-s_{n-1}) \ . \end{aligned} \tag{2.10}$$

Repeated factorization of (2.10) yields

$$(2.11)\quad \sum_{i=1}^{m}\sum_{j=1}^{n} p_{ij}\,R_j = s_1 + (1-s_1)\Big[\,s_2 + (1-s_2)\Big[\,s_3 + (1-s_3)\Big[\,\cdots \cdot \Big[\,s_{n-1} + (1-s_{n-1})\,s_n\,\Big]_{n-2}\cdots\Big]_3\Big]_2\Big]_1 .$$

Now, $s_n = 1$. Consequently, the quantity in the (n-2) brackets reduces to unity. It follows at once that the entire right-hand side of (2.11) telescopes to unity.

<u>Estimation of conditional probabilities</u>. The a_{ij} represent a random sample from a multinomial population. The likelihood function for such a sample is

$$\frac{N!}{\prod_{i=1}^{m}\prod_{j=1}^{n}(a_{ij}!)}\prod_{i=1}^{m}\prod_{j=1}^{n}\left[p_{ij}\,R_j\right]^{a_{ij}} .$$

The logarithm of the likelihood function is

$$(2.12)\quad L = \log\frac{N!}{\prod_{i=1}^{m}\prod_{j=1}^{n}(a_{ij}!)} + \sum_{i=1}^{m}\sum_{j=1}^{n} a_{ij}\log p_{ij} + \sum_{i=1}^{m}\sum_{j=1}^{n} a_{ij}\log R_j .$$

Note that

$$\sum_{i=1}^{m}\sum_{j=1}^{n} a_{ij}\log R_j = \sum_{j=1}^{n}\log R_j\sum_{i=1}^{m} a_{ij} ,$$

$$= \sum_{j=1}^{n} r_j\log R_j ,$$

$$\sum_{i=1}^{m} \sum_{j=1}^{n} a_{ij} \log R_j = r_1 \log (1-s_0) + r_2 \log (1-s_0)(1-s_1)$$

$$+ \ldots + r_n \log (1-s_0)(1-s_1) \ldots (1-s_{n-1}) ,$$

$$= r_1 \log (1-s_0) + r_2 \log (1-s_0) + r_2 \log (1-s_1)$$

$$+ \ldots + r_n \log (1-s_0) + r_n \log (1-s_1)$$

(2.13)

$$+ \ldots + r_n \log (1-s_{n-1}) ,$$

$$= \log (1-s_0) \sum_{j=1}^{n} r_j + \log (1-s_1) \sum_{j=2}^{n} r_j$$

$$+ \ldots + \log (1-s_{n-1}) r_n ,$$

$$= \sum_{j=1}^{n} S_j \log (1-s_{j-1}) ,$$

where

(2.14) $$S_j = \sum_{\ell=j}^{n} r_\ell .$$

Obviously, S_j is the number of animals that are surviving at the beginning of the jth interval.

Substitution of (2.13) in (2.12) yields

(2.15) $$L = \log \frac{N!}{\prod_{i=1}^{m} \prod_{j=1}^{n} (a_{ij}!)} + \sum_{i=1}^{m} \sum_{j=1}^{n} a_{ij} \log p_{ij}$$

$$+ \sum_{j=1}^{n} S_j \log (1-s_{j-1}) .$$

The maximum likelihood estimates of p_{ij} are found by setting the

first partial derivatives of the function L with respect to p_{ij} equal to zero and solving the resulting equations for the estimates, $\hat{p}_{ij}$. To this end, rewrite (2.15) in the form

$$(2.16) \qquad L = \log \frac{N!}{\prod_{i=1}^{m} \prod_{j=1}^{n} (a_{ij}!)} + \sum_{i=1}^{m} \sum_{j=1}^{n} a_{ij} \log p_{ij} + \sum_{j=0}^{n-1} S_{j+1} \log (1-s_j) .$$

Examination of (2.16) shows clearly that the maximum likelihood equations for the p_{ij} of any one interval will not involve the p_{ij} of any other interval. Thus, for $k < n$, the partial derivatives of L with respect to p_{ij} for the kth interval are:

$$(2.17) \qquad \frac{\partial L}{\partial p_{ij}} = \frac{a_{ik}}{p_{ik}} - \frac{S_{k+1}}{1-s_k} \qquad (i = 1, 2, \ldots, m) .$$

Setting these partial derivatives equal to zero and using (2.4), one obtains the following system of equations for the $\hat{p}_{ik}$:

$$(2.18) \qquad \begin{aligned} (a_{1k} + S_{k+1})\, \hat{p}_{1k} + a_{1k}\, \hat{p}_{2k} + \cdots + a_{1k}\, \hat{p}_{mk} &= a_{1k} \\ a_{2k}\, \hat{p}_{1k} + (a_{2k} + S_{k+1})\, \hat{p}_{2k} + \cdots + a_{2k}\, \hat{p}_{mk} &= a_{2k} \\ \vdots \qquad\qquad \vdots \qquad\qquad \vdots \qquad & \vdots \\ a_{mk}\, \hat{p}_{1k} + a_{mk}\, \hat{p}_{2k} + \cdots + (a_{mk} + S_{k+1})\, \hat{p}_{mk} &= a_{mk} . \end{aligned}$$

Addition of the system (2.18) renders

$$(2.19) \qquad (S_{k+1} + \sum_{i=1}^{m} a_{ik})\,\hat{p}_{ik} + (S_{k+1} + \sum_{i=1}^{m} a_{ik})\,\hat{p}_{2k} + \dots + (S_{k+1} + \sum_{i=1}^{m} a_{ik})\,\hat{p}_{mk} = \sum_{i=1}^{m} a_{ik} ,$$

or

$$(2.20) \qquad (S_{k+1} + r_k)\,\hat{p}_{ik} + (S_{k+1} + r_k)\,\hat{p}_{2k} + \dots + (S_{k+1} + r_k)\,\hat{p}_{mk} = \sum_{i=1}^{m} a_{ik} .$$

Applying (2.14), one has

$$(2.21) \qquad S_k\,\hat{p}_{1k} + S_k\,\hat{p}_{2k} + \dots + S_k\,\hat{p}_{mk} = \sum_{i=1}^{m} a_{ik} .$$

Hence,

$$(2.22) \qquad \hat{p}_{1k} + \hat{p}_{2k} + \dots + \hat{p}_{mk} = \frac{\sum_{i=1}^{m} a_{ik}}{S_k} .$$

Solving (2.22) and the ith equation of (2.18) simultaneously for $\hat{p}_{ik}$, one obtains

$$(2.23) \qquad \hat{p}_{ik} = \frac{a_{ik}\,(S_k - \sum_{i=1}^{m} a_{ik})}{S_k(S_{k+1})} = \frac{a_{ik}(S_k - r_k)}{S_k(S_{k+1})} , \qquad = \frac{a_{ik}\,S_{k+1}}{S_k(S_{k+1})} = \frac{a_{ik}}{S_k} \qquad (k < n) .$$

When $k = n$, $s_n = \sum_{i=1}^{m} p_{in} = 1$. Therefore, the p_{in} are parameters of a multinomial distribution. It is well known [9, pp. 214-216] that

the maximum likelihood estimates of the p_{in} are

$$\hat{p}_{in} = \frac{a_{in}}{S_n} \quad . \tag{2.24}$$

Thus, the maximum likelihood estimates of the p_{ij} are

$$\hat{p}_{ij} = \frac{a_{ij}}{S_j} \qquad (i = 1, \ldots, m;\ j = 1, \ldots, n). \tag{2.25}$$

<u>Estimation of disease incidences in the unrestricted population.</u> In the population in which all diseases are operating, the true incidence, I_i, of the ith disease may be computed by

$$I_i = \sum_{j=1}^{n} p_{ij} R_j \quad , \tag{2.26}$$

that is, by summing the unconditional probabilities over all time intervals. An estimate of I_i is given by

$$\hat{I}_i = \sum_{j=1}^{n} \hat{p}_{ij} \hat{R}_j \quad , \tag{2.27}$$

where $\hat{R}_j$ is R_j with $\hat{p}_{ij}$ substituted for p_{ij}. Since

$$1-s_j = 1 - \sum_{i=1}^{m} p_{ij} \quad , \tag{2.28}$$

it follows that

$$1-\hat{s}_j = 1 - \frac{\sum_{i=1}^{m} a_{ij}}{S_j} = 1 - \frac{r_j}{S_j} \quad ,$$

(2.29)

$$= \frac{S_j - r_j}{S_j} = \frac{S_{j+1}}{S_j} \quad .$$

Then

$$
\begin{aligned}
R_j &= (1-\hat{s}_0)(1-\hat{s}_1) \ldots (1-\hat{s}_{j-1}) \ , \\
&= (1-\hat{s}_1)(1-\hat{s}_2) \ldots (1-\hat{s}_{j-2})(1-\hat{s}_{j-1}) \ , \\
&= \frac{S_2}{S_1} \cdot \frac{S_3}{S_2} \cdots \frac{S_{j-1}}{S_{j-2}} \cdot \frac{S_j}{S_{j-1}} \ , \\
&= \frac{S_j}{S_1} \ .
\end{aligned}
\tag{2.30}
$$

Using (2.25) and (2.30), one finds that (2.27) takes the form

$$
\hat{I}_i = \sum_{j=1}^{n} \frac{a_{ij}}{S_j} \cdot \frac{S_j}{S_1} = \frac{1}{S_1} \sum_{j=1}^{n} a_{ij} \ . \tag{2.31}
$$

Now,

$$
S_1 = \sum_{\ell=1}^{n} r_\ell = N \ . \tag{2.32}
$$

It follows immediately that

$$
\hat{I}_i = \frac{A_i}{N} \ , \tag{2.33}
$$

which is the classical result.

<u>Estimation of disease incidences in the truncated population</u>. It is desired to use a sample from a population in which all diseases considered are operating to estimate the disease incidences in a restricted population from which one particular disease has been eliminated

as a possible cause of death. This is equivalent to estimating the properties of a truncated population from an untruncated sample.

Assume that the first disease is to be eliminated. Since the indexing of the diseases may be permuted so that the ith disease becomes the first, there is no loss in generality. Define a new set of conditional probabilities by

$$p'_{ij} = \frac{p_{ij}}{1-p_{1j}} \qquad (i = 2, \ldots, m;\ j = 1, 2, \ldots, n) . \tag{2.34}$$

Let

$$q_j = \sum_{i=2}^{m} p'_{ij} , \tag{2.35}$$

$$q_0 = 0 , \tag{2.36}$$

$$R'_j = \prod_{\ell=0}^{j-1} (1 - q_\ell) . \tag{2.37}$$

Then, the incidence of the ith disease in the truncated population is

$$I'_i = \sum_{j=1}^{n} p'_{ij} R'_j \qquad (i = 2, \ldots, m) . \tag{2.38}$$

It may be shown that

$$\hat{I}'_i = \sum_{j=1}^{n} \hat{p}'_{ij} \hat{R}'_j \qquad (i = 2, \ldots, m) , \tag{2.39}$$

where

$$\hat{p}'_{ij} = \frac{\hat{p}_{ij}}{1-\hat{p}_{ij}} ,$$

(2.40)

$$\hat{R}'_j = \prod_{\ell=0}^{j-1} (1-\hat{q}_\ell) = \prod_{\ell=0}^{j-1} (1 - \sum_{i=2}^{m} \hat{p}'_{i\ell}) .$$

It will be advantageous to express $\hat{I}_i'$ in a different form. By the use of (2.6), (2.28), (2.34), (2.35), and (2.37), one has

$$
\begin{aligned}
R_j' &= \prod_{\ell=0}^{j-1} (1-q_\ell) , \\
&= (1-q_0)(1-q_1) \cdots (1-q_{j-1}) , \\
&= (1-q_1)(1-q_2) \cdots (1-q_{j-1}) , \\
&= \left[1 - \sum_{i=2}^{m} p_{i1}'\right] \left[1 - \sum_{i=2}^{m} p_{i2}'\right] \cdots \left[1 - \sum_{i=2}^{m} p_{i,\,j-1}'\right] , \\
&= \left[1 - \sum_{i=2}^{m} \frac{p_{i1}}{1-p_{11}}\right] \left[1 - \sum_{i=2}^{m} \frac{p_{i2}}{1-p_{12}}\right] \cdots \left[1 - \sum_{i=2}^{m} \frac{p_{i,\,j-1}}{1-p_{1,\,j-1}}\right] , \\
&= \left[\frac{1 - \sum_{i=1}^{m} p_{i1}}{1-p_{11}}\right] \left[\frac{1 - \sum_{i=1}^{m} p_{i2}}{1-p_{12}}\right] \cdots \left[\frac{1 - \sum_{i=1}^{m} p_{i,\,j-1}}{1-p_{1,\,j-1}}\right] , \\
&= \frac{1-s_1}{1-p_{11}} \cdot \frac{1-s_2}{1-p_{12}} \cdots \frac{1-s_{j-1}}{1-p_{1,\,j-1}} , \\
&= \frac{R_j}{\prod_{\ell=1}^{j-1} (1-p_{1\ell})} .
\end{aligned}
\tag{2.41}
$$

On substituting (2.34) and (2.41) in (2.38), one may write (2.38) in the form

$$
\begin{aligned}
I_i' &= \sum_{j=1}^{n} \frac{p_{ij}}{1-p_{1j}} \left[\frac{R_j}{\prod_{\ell=1}^{j-1} (1-p_{1\ell})}\right] , \\
&= \sum_{j=1}^{n} \frac{p_{ij}\, R_j}{\prod_{\ell=1}^{j} (1-p_{1\ell})} .
\end{aligned}
\tag{2.42}
$$

Also

$$\hat{R}_j = \frac{S_j}{S_1} = \frac{S_j}{N} \text{ and } \hat{p}_{ij} = \frac{a_{ij}}{S_j} .$$

Therefore,

$$\hat{I}_i' = \frac{1}{N} \sum_{j=1}^{n} \frac{a_{ij}}{\prod_{\ell=1}^{j} (1-\hat{p}_{1\ell})} \tag{2.43}$$

is another form for $\hat{I}_i'$.

Approximate variances of the estimates. It has been pointed out that the a_{ij} represent a sample from a multinomial population. It is well known [10, pp. 34-35] that

$$\begin{aligned} \text{Var } (a_{ij}) &= N \cdot (p_{ij} R_j)(1-p_{ij} R_j) , \\ \text{Cov } (a_{ij}, a_{i'j'}) &= - N \cdot (p_{ij} R_j)(p_{i'j'} R_{j'}) . \end{aligned} \tag{2.44}$$

As a result of (2.25), (2.30), and (2.32), these variances and covariances are estimated by

$$\begin{aligned} \text{Est Var } (a_{ij}) &= \frac{a_{ij}(N - a_{ij})}{N} , \\ \text{Est Cov } (a_{ij}, a_{i'j'}) &= - \frac{a_{ij}\, a_{i'j'}}{N} . \end{aligned} \tag{2.45}$$

Inspection of formula (2.43) shows that each term in the summation is correlated with all preceding terms. This makes the calculation of the variance of $\hat{I}_i'$ extremely laborious. In order to simplify the computation, it will be assumed that (1) the p_{ij} are small and (2) the covariances between a_{ij} and $a_{ij'}$ are also small. In other words, if the

covariance terms are negligible, and if the denominators of the successive terms in (2.43) are nearly constant, the variance may be approximated by

$$\text{(2.46)} \qquad \mathrm{Var}\,(\hat{I}_i') \cong \frac{1}{N^2} \sum_{j=1}^{n} \frac{\mathrm{Var}\,(a_{ij})}{\left[\prod_{\ell=1}^{j} (1-\hat{p}_{1\ell})\right]^2} .$$

Substitution of (2.45) in (2.46) yields

$$\mathrm{Var}\,(\hat{I}_i') \cong \frac{1}{N^3} \sum_{j=1}^{n} \frac{a_{ij}(N - a_{ij})}{\left[\prod_{\ell=1}^{j} (1-\hat{p}_{1\ell})\right]^2} .$$

Failure of the first assumption will deflate the variance estimate. Failure of the second assumption will inflate the estimate since the covariances are negative. If both assumptions fail, it is expected that these two errors will counterbalance approximately.

CHAPTER III

ASYMPTOTIC VARIANCES OF THE ESTIMATES

The results of the sampling experiment to be described in Chapter IV indicate that the assumptions made in deriving the approximate variance formula (2.47) of the last chapter may not be valid and that the two errors introduced there often do not counterbalance each other. Attention will be given next to one possibility for obtaining better estimates for the variances of $\hat{I}_i'$ that do not depend upon the questionable assumptions. An asymptotic variance formula, based on the Taylor series expansion of $\hat{I}_i'$, will be derived in this chapter.

<u>Asymptotic variance-covariance matrix of the conditional probabilities</u>. The asymptotic variance-covariance matrix of the p_{ij} may be found by inverting the symmetric positive definite matrix

$$(3.1) \qquad \Psi = \left(-E \frac{\partial^2 L}{\partial p_{ij} \, \partial p_{i'j'}}\right) \qquad (j, j' < n) ,$$

where E is the operator of mathematical expectation.

Because of the particular nature of the likelihood function, Ψ takes the form

$$(3.2) \qquad \Psi = \begin{pmatrix} \Psi_1 & 0 & \cdots & 0 \\ 0 & \Psi_2 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & \Psi_{n-1} \end{pmatrix} ,$$

where the Ψ_k $(k < n)$ are $m \times m$ block matrices with elements $- E \dfrac{\partial^2 L}{\partial p_{ik} \, \partial p_{i'k}}$.

The elements of Ψ_k may be computed with the aid of

Theorem 3.1. Let r_k, R_k, and S_k be the quantities defined by (2.2), (2.6), and (2.14) respectively. If $E(a_{ij}) = Np_{ij} R_j$, then

$$E(S_k) = NR_k \quad . \tag{3.3}$$

Proof. By well-known rules of mathematical expectation

$$\begin{aligned} E(r_k) &= E \sum_{i=1}^{m} a_{ik} = \sum_{i=1}^{m} E(a_{ik}) \quad , \\ &= NR_k \sum_{i=1}^{m} p_{ik} \quad . \end{aligned} \tag{3.4}$$

Now, for k = 1, the proof of (3.3) is trivial. Write

$$E(S_1) = E \sum_{j=1}^{n} r_j = E(N) = N \quad . \tag{3.5}$$

Since $R_1 = (1-s_0) = 1$, equation (3.5) will not be altered if the right-hand side is multiplied by R_1. Therefore,

$$E(S_1) = NR_1 \quad . \tag{3.6}$$

To prove the theorem by induction, assume it to be valid for k, and consider

$$S_{k+1} = S_k - r_k \quad . \tag{3.7}$$

Taking expected values of (3.7) and using (3.3) and (3.4), one has

$$E(S_{k+1}) = E(S_k) - E(r_k) = NR_k - NR_k \sum_{i=1}^{m} p_{ik} \quad ,$$

$$E(S_{k+1}) = NR_k \left(1 - \sum_{i=1}^{m} p_{ik}\right) = NR_k(1-s_k) \quad ,$$

(3.8)

$$= NR_{k+1} \quad .$$

Thus, the theorem is proved for all k.

It was seen in (2.17) that for the kth period

$$\frac{\partial L}{\partial p_{ik}} = \frac{a_{ik}}{p_{ik}} - \frac{S_{k+1}}{1-s_k} \qquad (i = 1, \ldots, m; \quad k < n). \tag{3.9}$$

Consequently,

$$\frac{\partial^2 L}{\partial p_{ik}^2} = - \frac{a_{ik}}{p_{ik}^2} - \frac{S_{k+1}}{(1-s_k)^2} \quad ,$$

(3.10)

$$\frac{\partial^2 L}{\partial p_{ik}\, \partial p_{i'k}} = - \frac{S_{k+1}}{(1-s_k)^2} \qquad (i \neq i') \quad .$$

Now, use equation (2.6) and theorem 3.1 to evaluate the negative of the expected values of (3.10). To this end write

$$- E \frac{\partial^2 L}{\partial p_{ik}^2} = E \frac{a_{ik}}{p_{ik}^2} + E \frac{S_{k+1}}{(1-s_k)^2} \quad ,$$

$$= N \frac{p_{ik} R_k}{p_{ik}^2} + \frac{NR_{k+1}}{(1-s_k)^2} \quad ,$$

(3.11)

$$= \frac{N p_{ik} R_k}{p_{ik}^2} + \frac{NR_k(1-s_k)}{(1-s_k)^2} \quad ,$$

$$= NR_k \left[\frac{1}{p_{ik}} + \frac{1}{1-s_k}\right] \quad ,$$

$$(3.11)\qquad \begin{aligned} -E\,\frac{\partial^2 L}{\partial p_{ik}\,\partial p_{i'k}} &= E\,\frac{S_{k+1}}{(1-s_k)^2}\ , \\ &= \frac{NR_k}{1-s_k}\ . \end{aligned}$$

It follows that

$$(3.12)\quad \Psi_k = NR_k \begin{pmatrix} \frac{1}{p_{1k}} + \frac{1}{1-s_k} & \frac{1}{1-s_k} & \cdots & \frac{1}{1-s_k} \\ \frac{1}{1-s_k} & \frac{1}{p_{2k}} + \frac{1}{1-s_k} & \cdots & \frac{1}{1-s_k} \\ \vdots & \vdots & & \vdots \\ \frac{1}{1-s_k} & \frac{1}{1-s_k} & \cdots & \frac{1}{p_{mk}} + \frac{1}{1-s_k} \end{pmatrix}.$$

Rewrite (3.12) as

$$(3.13)\quad N^{-1}\,R_k^{-1}\,\Psi_k = \begin{pmatrix} \frac{1}{p_{1k}} & 0 & \cdots & 0 \\ 0 & \frac{1}{p_{2k}} & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & \frac{1}{p_{mk}} \end{pmatrix} + \frac{1}{1-s_k}\,\Phi\ ,$$

where Φ is a $k \times k$ matrix whose elements are all unity.

Roy and Sarhan [11] have proved

<u>Theorem 3.2</u>. Let A be any $m \times m$ nonsingular square matrix and A^{-1} its inverse. Let D be an $m \times m$ diagonal matrix with elements $(1/d_1, \ldots, 1/d_m)$. Let (u/d), (v/d), (u), and (v) be $m \times 1$ vectors

with elements u_i/d_i, v_i/d_i, u_i, and v_i respectively. Finally, let λ denote a scalar. If

$$(3.14) \qquad A = D - \frac{\lambda}{1 + \lambda \sum_{i=1}^{m} u_i \frac{v_i}{d_i}} (u/d)(v/d)' ,$$

then

$$(3.15) \qquad A^{-1} = D^{-1} + \lambda (u)(v)' .$$

Setting $\lambda = -1$, $(u) = (v) = (d)$, and $A = N^{-1} R_k^{-1} \Psi_k$, one may write (3.14) in the form

$$(3.16) \qquad N^{-1} R_k^{-1} \Psi_k = D + \frac{1}{1 - \sum_{i=1}^{m} d_i} \Phi ,$$

where Φ is defined as above. Now if D is the diagonal matrix of (3.13) and $(d)' = (p_{1k}, \ldots, p_{mk})$, the result of theorem 3.2 enables one to write

$$(3.17) \qquad NR_k \Psi_k^{-1} = \begin{pmatrix} p_{1k} & 0 & \cdots & 0 \\ 0 & p_{2k} & & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & & p_{mk} \end{pmatrix} - \begin{pmatrix} p_{1k}^2 & p_{1k}p_{2k} & \cdots & p_{1k}p_{mk} \\ p_{1k}p_{2k} & p_{2k}^2 & \cdots & p_{2k}p_{mk} \\ \vdots & \vdots & & \vdots \\ p_{1k}p_{mk} & p_{2k}p_{mk} & \cdots & p_{mk}^2 \end{pmatrix} .$$

Therefore,

$$\Psi_k^{-1} = \frac{1}{NR_k} \begin{pmatrix} p_{1k}-p_{1k}^2 & -p_{1k}p_{2k} & \cdots & -p_{1k}p_{mk} \\ -p_{1k}p_{2k} & p_{2k}-p_{2k}^2 & \cdots & -p_{2k}p_{mk} \\ \vdots & \vdots & & \vdots \\ -p_{1k}p_{mk} & -p_{2k}p_{mk} & \cdots & p_{mk}-p_{mk}^2 \end{pmatrix},$$

(3.18)

$$= \frac{1}{NR_k} \begin{pmatrix} p_{1k}(1-p_{1k}) & -p_{1k}p_{2k} & \cdots & -p_{1k}p_{mk} \\ -p_{1k}p_{2k} & p_{2k}(1-p_{2k}) & \cdots & -p_{2k}p_{mk} \\ \vdots & \vdots & & \vdots \\ -p_{1k}p_{mk} & -p_{2k}p_{mk} & \cdots & p_{mk}(1-p_{mk}) \end{pmatrix}.$$

Recall that when $k = n$, the p_{in} are parameters of a multinomial distribution. It is well known [9, pp. 214-216] that the $\hat{p}_{in}$ have an asymptotic variance-covariance matrix that is equal to Ψ_n^{-1}. Thus, the variance-covariance matrix of the $\hat{p}_{ij}$ is given by

$$(3.19) \qquad \Omega = \frac{1}{N} \begin{pmatrix} \Omega_1 & 0 & \cdots & 0 \\ 0 & \Omega_2 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & \Omega_n \end{pmatrix},$$

where

$$(3.20)\quad \Omega_k = \frac{1}{R_k}\begin{pmatrix} p_{1k}(1-p_{1k}) & -p_{1k}p_{2k} & \cdots & -p_{1k}p_{mk} \\ -p_{1k}p_{2k} & p_{2k}(1-p_{2k}) & \cdots & -p_{2k}p_{mk} \\ \vdots & \vdots & & \vdots \\ -p_{1k}p_{mk} & -p_{2k}p_{mk} & \cdots & p_{mk}(1-p_{mk}) \end{pmatrix}.$$

<u>Asymptotic variances of the incidence estimates</u>. It will be convenient to let $h = i$ and $t = j$ in (2.39). Then

$$(3.21)\qquad \hat{I}_h' = \sum_{t=1}^{n} \hat{p}_{ht}' \hat{R}_t' \qquad (h = 2, \ldots, m) .$$

Expand (3.21) into its Taylor series about the p_{ij} ($i = 1, \ldots, m$; $j = 1, \ldots, n$) retaining only the linear terms. Then,

$$(3.22)\qquad \hat{I}_h' \cong I_h' + \sum_{i=1}^{m} \sum_{j=1}^{n} (\hat{p}_{ij} - p_{ij})\theta_{hij} ,$$

where

$$(3.23)\qquad \theta_{hij} = \sum_{t=1}^{n} p_{ht}' \frac{\partial R_t'}{\partial p_{ij}} + \sum_{t=1}^{n} R_t' \frac{\partial p_{ht}'}{\partial p_{ij}} .$$

Note that

$$(3.24)\qquad \frac{\partial p_{ht}'}{\partial p_{ij}} = 0 \qquad (t \neq j) .$$

Therefore,

$$\theta_{hij} = R_j' \frac{\partial p_{hj}'}{\partial p_{ij}} + \sum_{t=1}^{n} p_{ht}' \frac{\partial R_t'}{\partial p_{ij}} ,$$

(3.25)

$$= R_j' \frac{\partial p_{hj}'}{\partial p_{ij}} + \sum_{t=1}^{j} p_{ht}' \frac{\partial R_t'}{\partial p_{ij}} + p_{h,\ j+1}' \frac{\partial R_{j+1}'}{\partial p_{ij}} + \sum_{t=j+2}^{n} p_{ht}' \frac{\partial R_t'}{\partial p_{ij}} .$$

To calculate the first term on the right-hand side of (3.25), one observes that

$$\frac{\partial p_{hj}'}{\partial p_{1j}} = \frac{p_{hj}}{(1-p_{1j})^2} ,$$

(3.26)

$$\frac{\partial p_{hj}'}{\partial p_{ij}} = \frac{1}{1-p_{1j}} \qquad (i = h) ,$$

$$\frac{\partial p_{hj}'}{\partial p_{ij}} = 0 \qquad (i \neq h) .$$

To calculate the last three terms of (3.25), note that

$$(1-q_j) = 1 - \sum_{i=2}^{m} p_{ij}' = 1 - \frac{1}{1-p_{1j}} \sum_{i=2}^{m} p_{ij} ,$$

(3.27)

$$= \frac{1 - \sum_{i=1}^{m} p_{ij}}{1-p_{1j}} = \frac{1-s_j}{1-p_{1j}} .$$

In calculating $\frac{\partial R_t'}{\partial p_{ij}}$ three cases may be distinguished:

<u>case (i)</u>. If $t < j + 1$,

$$R_t' = \prod_{\ell=0}^{t-1} (1-q_\ell) \quad . \tag{3.28}$$

Obviously, all R_t' $(t < j + 1)$ do not involve the j subscript; and hence, are not functions of the p_{ij}. Thus,

$$\frac{\partial R_t'}{\partial p_{ij}} = 0 \qquad (i = 1, 2, \ldots, m). \tag{3.29}$$

<u>case (ii)</u>. If $t = j + 1$,

$$\begin{aligned} R_{j+1}' &= \prod_{\ell=0}^{j} (1-q_\ell) \ , \\ &= R_j' \ (1-q_j) \ . \end{aligned} \tag{3.30}$$

Substitution of (3.27) in (3.30) yields

$$R_{j+1}' = R_j' \left(\frac{1-s_j}{1-p_{1j}}\right) \quad . \tag{3.31}$$

Differentiation of (3.31) with respect to P_{1j} gives

$$\begin{aligned} \frac{\partial R_{j+1}}{\partial p_{1j}} &= \frac{-(1-p_{1j})+(1-s_j)}{(1-p_{1j})^2} \ R_j' = \frac{p_{1j}-s_j}{(1-p_{1j})^2} \ R_j' \ , \\ &= - \frac{s_j-p_{1j}}{(1-p_{1j})^2} \ R_j' \ . \end{aligned} \tag{3.32}$$

Differentiation of (3.31) with respect to p_{ij} $(1 < j \leq n)$ gives

$$\frac{\partial R'_{j+1}}{\partial p_{ij}} = - \frac{R'_j}{(1-p_{1j})} \quad . \tag{3.33}$$

<u>case (iii)</u>. If $t > j + 1$,

$$\begin{aligned} R'_t &= (1-q_j) \prod_{\ell=0}^{j-1} (1-q_\ell) \prod_{\ell=j+1}^{t-1} (1-q_\ell) \quad , \\ &= R'_j \, (1-q_j) \prod_{\ell=j+1}^{t-1} (1-q_\ell) \quad . \end{aligned} \tag{3.34}$$

Substitution of (3.27) in (3.34) yields

$$R'_t = R'_j \frac{1-s_j}{1-p_{1j}} \prod_{\ell=j+1}^{t-1} (1-q_\ell) \quad . \tag{3.35}$$

Differentiating (3.35) with respect to p_{1j} gives

$$\frac{\partial R'_t}{\partial p_{1j}} = - \frac{R'_j(s_j-p_{1j})}{(1-p_{1j})^2} \prod_{\ell=j+1}^{t-1} (1-q_\ell) \quad . \tag{3.36}$$

On multiplying both the numerator and denominator by the quantity $(1-q_j)$, equation (3.36) becomes

$$\begin{aligned} \frac{\partial R'_t}{\partial p_{1j}} &= - \frac{R'_j(s_j-p_{1j})(1-q_j)}{(1-p_{1j})^2(1-q_j)} \prod_{\ell=j+1}^{t-1} (1-q_\ell) \quad , \\ &= - \frac{R'_t(s_j-p_{1j})}{(1-q_j)(1-p_{1j})^2} \quad . \end{aligned} \tag{3.37}$$

Differentiation of (3.35) with respect to p_{ij} $(1 < i \leq j)$ and multiplication of the numerator and denominator of the resulting differential equation by the quantity $(1-q_j)$ gives

$$\frac{\partial R_t'}{\partial p_{ij}} = -\frac{R_t'}{(1-p_{1j})(1-q_j)} \quad . \tag{3.38}$$

Thus, the calculation of θ_{hij} must also be divided into cases:

<u>case (i)</u>. If $i = 1$ and $j < n$,

$$\theta_{h1j} = \frac{R_j' \, p_{hj}}{(1-p_{1j})^2} - \frac{p_{h,\,j+1}' \, R_j'(s_j-p_{1j})}{(1-p_{1j})^2}$$

$$- \sum_{t=j+2}^{n} \frac{p_{ht}' \, R_t'(s_j-p_{1j})}{(1-q_j)(1-p_{1j})^2} \quad . \tag{3.39}$$

Multiplying the numerator and denominator of the second term on the right-hand side of (3.39) by the quantity $(1-s_j)$ and using (2.34), (2.41), and (3.27); one has

$$\theta_{h1j} = \frac{R_j' \, p_{hj}}{(1-p_{1j})^2} - \frac{p_{h,\,j+1}' \, R_j'(1-s_j)(s_j-p_{1j})}{(1-s_j)(1-p_{1j})^2}$$

$$- \sum_{t=j+2}^{n} \frac{p_{ht}' \, R_t'(s_j-p_{1j})(1-p_{1j})}{(1-s_j)(1-p_{1j})^2} \quad ,$$

$$= \frac{1}{1-p_{1j}} \left[\frac{R_j \, p_{hj}}{\prod_{\ell=1}^{j} (1-p_{1\ell})} - \frac{s_j-p_{1j}}{(1-s_j)} \sum_{\substack{t=j+1 \\ j<n}}^{n} \frac{p_{ht} \, R_t}{\prod_{\ell=1}^{t} (1-p_{1\ell})} \right] \quad . \tag{3.40}$$

<u>case (ii)</u>. If $i = 1$ and $j = n$,

$$\theta_{h1n} = \frac{R_n \, p_{hn}}{(1-p_{1n}) \prod_{\ell=1}^{n} (1-p_{1\ell})} \tag{3.41}$$

is almost immediate. In this case $t < n + 1$ and by (3.29), $\frac{\partial R_t'}{\partial p_{ij}} = 0$. Indeed, the last three terms on the right-hand side of (3.25) vanish; and (3.41) follows at once.

<u>case (iii)</u>. If $i = h$ and $j < n$,

$$\theta_{hij} = \frac{R_j'}{1-p_{1j}} - \frac{p_{h,\,j+1}' \, R_j'}{(1-p_{1j})} - \sum_{t=j+2}^{n} p_{ht}' \frac{R_t'}{(1-p_{1j})(1-q_j)} \; . \tag{3.42}$$

Devices similar to those used in case (i) reduce this to the form

$$\theta_{hij} = \frac{R_j}{\prod_{\ell=1}^{j} (1-p_{1\ell})} - \frac{1}{1-s_j} \sum_{\substack{t=j+1 \\ j<n}}^{n} \frac{p_{ht} \, R_t}{\prod_{\ell=1}^{t} (1-p_{1\ell})} \; . \tag{3.43}$$

<u>case (iv)</u>. If $i = h$ and $j = n$,

$$\theta_{hin} = \frac{R_n}{\prod_{\ell=1}^{n} (1-p_{1\ell})} \tag{3.44}$$

is immediate. This case is analogous to case (ii).

<u>case (v)</u>. If $i \neq h$ and $j < n$,

$$\theta_{hij} = - \frac{p_{h,\,j+1}' \, R_j'}{(1-p_{1j})} - \sum_{\substack{t=j+2 \\ j<n}}^{n} p_{ht}' \frac{R_t'}{(1-p_{1j})(1-q_j)} \; . \tag{3.45}$$

Devices used in case (i) reduce this form to

$$\theta_{hij} = -\frac{1}{1-s_j} \sum_{\substack{t=j+1 \\ j<n}}^{n} \frac{p_{ht} R_t}{\prod_{\ell=1}^{t} (1-p_{1\ell})} . \tag{3.46}$$

<u>case (vi)</u>. If $i \neq h$ and $j = n$,

$$\theta_{hin} = 0 \tag{3.47}$$

is trivial. This follows at once from (3.25), (3.26), and (3.29).

Consider (3.22). Transposing I_h' to the left-hand side and squaring the resulting equation, one obtains

$$\begin{aligned}(\hat{I}_h' - I_h')^2 = & \sum_{i=1}^{m} \sum_{j=1}^{n} (\hat{p}_{ij} - p_{ij})^2 \theta_{hij}^2 \\ & + \sum_{i=1}^{m} \sum_{j=1}^{n} (\hat{p}_{ij} - p_{ij})(\hat{p}_{i'j} - p_{i'j}) \theta_{hij} \theta_{hi'j} \\ & + \sum_{i=1}^{m} \sum_{j=1}^{n} (\hat{p}_{ij} - p_{ij})(\hat{p}_{ij'} - p_{ij'}) \theta_{hij} \theta_{hij'} \\ & + \sum_{i=1}^{m} \sum_{j=1}^{n} (\hat{p}_{ij} - p_{ij})(\hat{p}_{i'j'} - p_{i'j'}) \theta_{hij} \theta_{hi'j'} .\end{aligned} \tag{3.48}$$

Now, the variance of $\hat{I}_h'$ is defined as

$$\text{Var}(\hat{I}_h') = E(\hat{I}_h' - I_h')^2 . \tag{3.49}$$

Then

$$\text{Var}\,(I_h') \cong \sum_{i=1}^{m}\sum_{j=1}^{n} E(\hat{p}_{ij} - p_{ij})^2\,\theta_{hij}^2$$

$$+ \sum_{i=1}^{m}\sum_{j=1}^{n} E(\hat{p}_{ij} - p_{ij})(\hat{p}_{i'j} - p_{i'j})\,\theta_{hij}\,\theta_{hi'j}$$

(3.50)

$$+ \sum_{i=1}^{m}\sum_{j=1}^{n} E(\hat{p}_{ij} - p_{ij})(\hat{p}_{ij'} - p_{ij'})\,\theta_{hij}\,\theta_{hij'}$$

$$+ \sum_{i=1}^{m}\sum_{j=1}^{n} E(\hat{p}_{ij} - p_{ij})(\hat{p}_{i'j'} - p_{i'j'})\,\theta_{hij}\,\theta_{hi'j'}\,.$$

When $j \neq j'$, the variance-covariance matrix (3.19) shows that

$$(3.51)\quad E(\hat{p}_{ij} - p_{ij})(\hat{p}_{ij'} - p_{ij'}) = E(\hat{p}_{ij} - p_{ij})(\hat{p}_{i'j'} - p_{i'j'}) = 0\quad .$$

Thus, the last two terms on the right-hand side of equation (3.50) vanish. It follows that

$$\text{Var}\,(\hat{I}_h') \cong \sum_{i=1}^{m}\sum_{j=1}^{n} \theta_{hij}^2\,\text{Var}\,(\hat{p}_{ij})$$

(3.52)

$$+ \sum_{i=1}^{m}\sum_{j=1}^{n} \theta_{hij}\,\theta_{hi'j}\,\text{Cov}(\hat{p}_{ij}, \hat{p}_{i'j})\,.$$

It was thought that the asymptotic variance formula (3.52) would be an improvement upon the approximate variance formula (2.47). In Chapter IV a comparison of these two variance formulas with the population variances is made for a wide range of parameters.

CHAPTER IV

RANDOM SAMPLING EXPERIMENT

Introduction. When statisticians are confronted with a difficult problem in distribution theory, they sometimes resort to the Monte Carlo method. The process consists of drawings from random sequences of numbers, or at least pseudo-random sequences, in accordance with an artificial model, whereby the statistic whose distribution is sought can be observed over and over again and the distribution estimated empirically.

A large scale random sampling (Monte Carlo) experiment was conducted in order to study the small sample distributions of Kimball's estimates of disease incidences as well as the small sample distributions of the estimates for the variances of the estimated disease incidences, $\hat{I}_i'$.

The general idea of the experiment was as follows. A set of parameters N, m, and p_{ij} were assigned, and with a synthetic model sample approximations to the distribution of a_{ij} were generated on the ORACLE. Large sample approximations to the distributions of $\hat{I}_i'$ were then computed; that is, large but finite populations of values of $\hat{I}_i'$ were generated which approximate the corresponding true life infinite populations. Since the generated populations of values of $\hat{I}_i'$ were finite, their variances could be computed by means of statistical formulas. Values for variance formulas (2.47) and (3.52) were computed from the sample of a_{ij}. These values were compared with the basic finite population variances. The assumption that the synthetic model is a good approximation to the infinite populations of samples then allows one to judge

the accuracy of both variance formulas by how closely they approximate the true (pseudo-true) population variances.

Description of the study. The experiment was divided into the following cases:

1. $E(r_j) = E(r_j')$
   a) $n = 5$
   b) $n = 10$
   c) $n = 20$
2. $E(r_j) \neq E(r_j')$
   a) $n = 5$
   b) $n = 10$

The cases were chosen to provide some information about the effects of variations in the number of time intervals and the number of deaths per time interval on the estimation procedure. For each case and for sample sizes of 50, 100, 200, 400, and 800, a thousand random samples were generated from a random distribution with a predetermined set of conditional probabilities, p_{ij}. Each of the samples was generated with $m = 4$. An attempt to simulate an actual experiment influenced the selection of the p_{ij}. The p_{ij} were specified for case (1c). These are displayed in Table I. The unconditional probabilities, $p_{ij}R_j$, are displayed in Table II. The p_{ij} for case (1b) were derived from case (1c) by pooling the unconditional probabilities in the adjacent time intervals of case (1c) by pairs. Similarly, the p_{ij} for case (1a) were derived from case (1b) by pooling the unconditional probabilities in the adjacent time intervals of case (1b) by pairs. The p_{ij} for case (2a) and case (2b) were derived by pooling the unconditional probabilities in the adjacent time intervals of case (1c) such that the

TABLE I

THE CONDITIONAL PROBABILITIES, p_{ij}, FOR CASE (1c)

Time Interval	Disease			
	1	2	3	4
1	0.0200	0.0245	0.0050	0.0005
2	.0216	.0253	.0053	.0005
3	.0233	.0261	.0056	.0006
4	.0253	.0271	.0059	.0006
5	.0300	.0250	.0062	.0012
6	.0327	.0260	.0067	.0013
7	.0286	.0271	.0143	.0014
8	.0315	.0285	.0154	.0015
9	.0350	.0292	.0167	.0025
10	.0391	.0309	.0182	.0027
11	.0470	.0300	.0200	.0030
12	.0533	.0322	.0222	.0033
13	.0638	.0250	.0312	.0050
14	.0743	.0271	.0357	.0057
15	.0800	.0300	.0500	.0067
16	.0980	.0340	.0600	.0080
17	.1175	.0325	.0875	.0125
18	.1600	.0400	.1167	.0167
19	.2450	.0550	.1750	.0250
20	0.5000	0.1000	0.3500	0.0500

TABLE II

THE UNCONDITIONAL PROBABILITIES, $p_{ij}R_j$, FOR CASE (1c)

Time Interval	Disease			
	1	2	3	4
1	0.0200	0.0245	0.0050	0.0005
2	.0205	.0240	.0050	.0005
3	.0210	.0235	.0050	.0005
4	.0215	.0230	.0050	.0005
5	.0240	.0200	.0050	.0010
6	.0245	.0195	.0050	.0010
7	.0200	.0190	.0100	.0010
8	.0205	.0185	.0100	.0010
9	.0210	.0175	.0100	.0015
10	.0215	.0170	.0100	.0015
11	.0235	.0150	.0100	.0015
12	.0240	.0145	.0100	.0015
13	.0255	.0100	.0125	.0020
14	.0260	.0095	.0125	.0020
15	.0240	.0090	.0150	.0020
16	.0245	.0085	.0150	.0020
17	.0235	.0065	.0175	.0025
18	.0240	.0060	.0175	.0025
19	.0245	.0055	.0175	.0025
20	0.0250	0.0050	0.0175	0.0025

number of time intervals that were combined formed the distributions (1, 3, 7, 7, 2) and (1, 1, 1, 2, 2, 2, 5, 3, 2, 1) respectively.

An examination of (2.43) shows clearly that the estimation procedure is undefined if in the nth interval either no deaths occur or all deaths are due to the first disease. In practice this difficulty may be overcome by judicious choice of the time intervals. In the empirical study, however, these situations were unavoidable in some samples since the time intervals were specified in advance of the experiment. When either anomaly developed in a sample, that sample was rejected, and a new sample was generated to replace it. If the rejection rate exceeded 10 per cent (100 in the present experiment) of the total number of samples drawn, the corresponding sample size was eliminated from the study. This expedient was used because of the excessive amount of machine time required to generate 1000 admissible samples under these conditions.

A general description of the computations is given in the form of a flow diagram, Figure 1. The arrow "→" inside a box means "is replaced by", and C_1, C_2, and C_3 are merely counters. The computational details in Figure 1 are discussed in the following sections.

<u>Generation of the pseudo-random numbers</u>. The procedure employed to generate the pseudo-random numbers, ρ_ℓ, was one based on methods described by Taussky and Todd [13]. If ρ_0 denotes any odd number and $\rho = 5^{15}$, then the recurrence relation

$$\rho_{\ell+1} = \rho \cdot \rho_\ell \pmod{2^{39}} \tag{4.1}$$

forms a pseudo-random sequence of numbers. The sequence is periodic

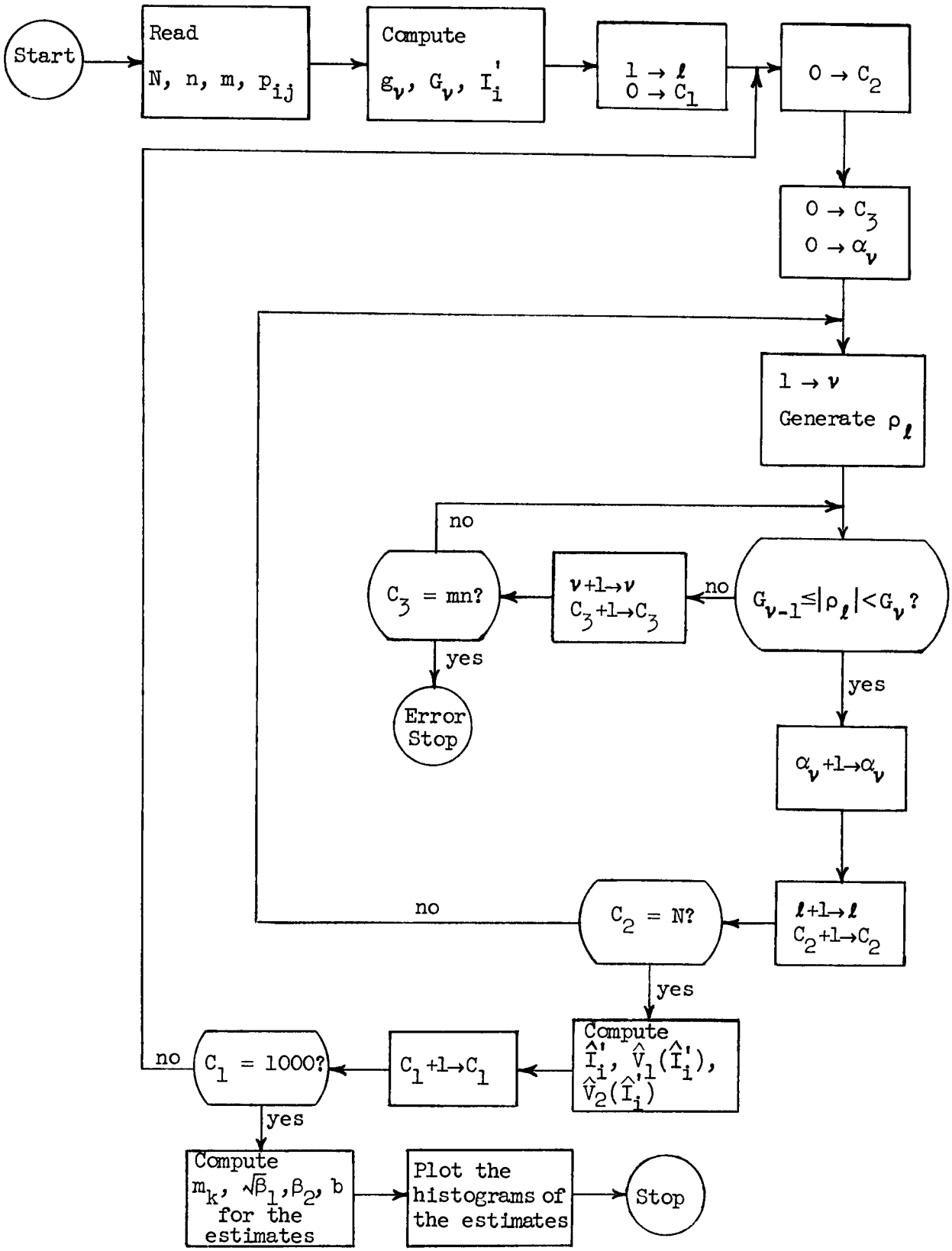

Figure 1. Flow diagram of the sampling experiment

with period 2^{37}. These numbers have been submitted to various statistical tests, and they satisfy the usual tests for randomness.

On a high-speed computing machine such as the ORACLE, this method breaks down to the generation of sequences of random numbers uniformly distributed on the interval (-1, 1), the end points of the interval as well as zero being excluded. Thus, the random numbers as represented in the ORACLE satisfy the inequalities

$$0 < |\rho_\ell| < 1 \ . \tag{4.2}$$

<u>Generation of the sample elements</u>. Let

$$\nu = n(i-1) + j \qquad (i = 1, \ldots, m;\ j = 1, \ldots, n) \ . \tag{4.3}$$

It will be convenient to introduce the quantities

$$\alpha_\nu = a_{ij} \quad \text{and} \quad g_\nu = p_{ij}R_j \qquad (\nu = 1, 2, \ldots, mn) \ . \tag{4.4}$$

The probability that an element belongs to one of the first ν categories of the population is given by

$$G_\nu = \sum_{t=1}^{\nu} g_t \ . \tag{4.5}$$

The generation of a sample element may be described as follows: Define $G_0 \equiv 0$. Let the sample element, z, be characterized by the set of variables $(z_1, z_2, \ldots, z_\nu)$. If an element belongs to the νth category of the population, $z = z_\nu$. Generate a random number by using

(4.1) with $\ell = 0$. If

$$G_{\nu-1} \leq \rho_{\ell+1} < G_{\nu} \quad , \tag{4.6}$$

set $z = z_{\nu}$. For a sample of size N repeat the process with $\ell = 1, 2, \ldots, N-1$ successively. The number of sample elements, α_{ν}, in the νth category of the population is the number of times that z assumes the value z_{ν}. Thus,

$$N \equiv \sum_{\nu=1}^{mn} \alpha_{\nu} \quad . \tag{4.7}$$

<u>Computation of the sample moments</u>. If $x_1, x_2, \ldots, x_N$ is an independent random sample from a population with probability density f(x), the kth sample moment is defined to be

$$m_k' \equiv \frac{1}{N} \sum_{t=1}^{N} x_t^k \quad . \tag{4.8}$$

Define $m_1' \equiv \bar{x}$. Then the kth moment about the sample mean (central moment) is denoted by

$$m_k \equiv \frac{1}{N} \sum_{t=1}^{N} (x_t - \bar{x})^k \quad . \tag{4.9}$$

Using the binomial expansion, one may write (4.9) in the form

$$m_k = \frac{1}{N} \sum_{t=1}^{N} \sum_{r=0}^{k} (-1)^r \binom{k}{r} x_t^{k-r} \bar{x}^r ,$$

(4.10)

$$= \sum_{r=0}^{k} (-1)^r \binom{k}{r} \bar{x}^r \sum_{t=1}^{N} \frac{x_t^{k-r}}{N} .$$

The use of definition (4.8) reduces this to the form

(4.11) $$m_k = \sum_{r=0}^{k} (-1)^r \binom{k}{r} \bar{x}^r m'_{k-r} ,$$

where $\binom{k}{r} \equiv \frac{k!}{r!(k-r)!}$.

Many authors of statistics textbooks advocate using (4.11) to compute the central moments of a sampling distribution. This method should, indeed, be recommended for a desk calculator since it is less tedious than computing individual deviations from the mean and summing their powers. On the other hand, when a high-speed digital computer is available, serious difficulties may arise, especially for large N.

The central moments of the distributions of the incidence and variance estimates were first computed by utilizing (4.11). Some of the fourth moments turned out to be negative and hence inadmissible.

Householder [5] distinguished three types of error in the result of a computation: (1) propagated error, (2) generated error, and (3) residual error. Propagated errors are those due to the original errors in data representation. Generated errors are those that arise as a result of round-off. Finally, residual errors are defined as those due to truncation after the computation is terminated.

It was the accumulation of the generated errors which rendered

the moment results incorrect. Hence, it was important to know how such errors affected the computations.

The moment computations were performed by using floating-point arithmetic. The number, ξ, in most floating-point systems is represented by

$$\xi^* = \pm\ (\xi_1\beta^{-1} + \xi_2\beta^{-2} + \ldots + \xi_\lambda\beta^{-\lambda})\beta^\eta \quad , \tag{4.12}$$

where β is the base, usually 2 or 10; λ is a positive integer; and η is any integer, possibly zero. The quantity in parentheses is usually called the mantissa of ξ^*. The floating-point routine chosen to perform the moment computations was designed to operate with numbers of the form (4.12) where

$$\begin{aligned} &\tfrac{1}{2} \le |\xi_1\beta^{-1} + \xi_2\beta^{-2} + \ldots + \xi_\lambda\beta^{-\lambda}| < 1 \quad , \\ &|\eta| \le 2^7 \quad , \\ &\beta = 2 \quad , \\ &\lambda = 31 \quad . \end{aligned} \tag{4.13}$$

The floating-point operation of addition may be described as follows. Assume $N_1 \equiv d_1 2^{e_1}$, $N_2 \equiv d_2 2^{e_2}$, $N_3 \equiv d_3 2^{e_3}$, and $e_2 \ge e_1$. Clearly, the d_i are the mantissas of the N_i. Consider $N_3 = N_1 \pm N_2$. The plus sign will be used consistently, but the results obviously hold for subtraction. Thus,

$$d_3 = \left(d_1 2^{e_1 - e_2}\right)^* + d_2,\ e_3 = e_2 \qquad (|d_3| < 1) \quad ,$$

(4.14)

$$d_3' = \left(d_3 2^{-1}\right)^*, \; e_3 = e_2 + 1 \qquad (1 \le |d_3| < 2) \; ,$$

where the asterisk superscript indicates truncation of the number of magnitude digits at λ. Thus, digits are dropped, and an error in the result is generated. Given N numbers to be added together, these errors can build up and render the result useless.

Suppose one is interested in summing a set of N numbers. Let n_1 of these numbers be nonnegative, and $N = n_1 + n_2$. There are sequences

(4.15)
$$\left\{x_k\right\}_1^{n_1} \quad \text{and} \quad \left\{x_k'\right\}_1^{n_2}$$

of nonnegative and negative elements respectively for which

(4.16)
$$|x_k| \le |x_{k'}| \quad \text{and} \quad |x_k'| \le |x_{k'}'| \qquad (k < k') \; .$$

To keep the error in the sum as small as possible, the sequences (4.15) should be added separately and their respective sums combined. Indeed, this ordering process was employed prior to computing the sample moments. Moreover, the computation of the moments was carried out by using (4.9) rather than (4.11) since fewer summations were involved. Thus, greater accuracy was achieved in the results.

Error bounds may be placed on the results of (4.9). Let

(4.17)
$$y_t = y_t^* + \epsilon_t \qquad (t = 1, 2, \ldots, N) \; ,$$

where y_t^* is the machine representation of y_t. Then

$$\sum_{t=1}^{N} y_t = \sum_{t=1}^{N} (y_t^* + \epsilon_t) \quad ,$$

(4.18)

$$\sum_{t=1}^{N} y_t - \sum_{t=1}^{N} y_t^* = \sum_{t=1}^{N} \epsilon_t \quad ,$$

so that

(4.19) $$\left| \sum_{t=1}^{N} y_t - \sum_{t=1}^{N} y_t^* \right| \leq \sum_{t=1}^{N} |\epsilon_t| \quad .$$

If $|\epsilon_t| \leq \epsilon$, then

(4.20) $$\left| \sum_{t=1}^{N} y_t - \sum_{t=1}^{N} y_t^* \right| \leq N\epsilon \quad .$$

Thus given N numbers, each of which can be in error by as much as ϵ, their sum can be in error by as much as $N\epsilon$.

Now,

$$y_t y_{t'} = (y_t^* + \epsilon_t)(y_{t'}^* + \epsilon_{t'}) \qquad (t \neq t') \quad ,$$

(4.21)

$$= y_t^* y_{t'}^* + \epsilon_t y_{t'}^* + \epsilon_{t'} y_t^* + \epsilon_t \epsilon_{t'} \quad ,$$

so that

(4.22) $$|y_t y_{t'} - y_t^* y_{t'}^*| \leq |\epsilon_t y_{t'}^*| + |\epsilon_{t'} y_t^*| + |\epsilon_t \epsilon_{t'}| \quad .$$

If $|y_t^*| \leq 1-\epsilon$, one may write

(4.23) $$|y_t y_{t'} - y_t^* y_{t'}^*| \leq (1-\epsilon)(|\epsilon_t| + |\epsilon_{t'}|) + |\epsilon_t \epsilon_{t'}| \quad .$$

Since $|\epsilon_t| \leq \epsilon$ $(t = 1, 2, \ldots, N)$, it is easy to verify by induction that

(4.24) $$\left| y_t^k - y_t^{*k} \right| \leq k\epsilon \quad .$$

Define $y_t \equiv x_t - \bar{x}$ and $\epsilon = 2^{-31}$. Using (4.9), (4.20), and (4.24), one finds that

$$|Nm_k - (Nm_k)^*| \leq Nk\epsilon = Nk\, 2^{-31} \quad . \tag{4.25}$$

Distribution of the estimates. Altogether twenty-one sets of parameters were used in the computation. It has already been pointed out that one thousand samples were generated for each set of parameters. When the first disease was eliminated as a cause of the death in the population, $\hat{I}_i'$ and $\mathrm{Var}(\hat{I}_i')$ were computed for each sample. The variances were computed by using both the approximate and asymptotic variance formulas. The moment statistics of the distributions of $\hat{I}_i'$ $(i = 2, 3, 4)$ are shown in Tables III - VII. The sample variance, the measure of skewness, and the measure of kurtosis are denoted by

$$\begin{aligned} \hat{\sigma}^2 &\equiv m_2 \quad , \\ \sqrt{\beta_1} &\equiv \frac{m_3}{m_2^{3/2}} \quad , \\ \beta_2 &\equiv \frac{m_4}{m_2^2} \quad , \end{aligned} \tag{4.26}$$

respectively. For the normal distribution it is well known that

$$\sqrt{\beta_1} = 0 \quad \text{and} \quad \beta_2 = 3 \quad . \tag{4.27}$$

To further depict these distributions, the one thousand values of $\hat{I}_i'$ for each set of parameters were arranged in histograms. It would not be feasible to present each of these histograms here, but those arising from case (1b) with a sample size of four hundred are exhibited in

TABLE III

MOMENT STATISTICS OF THE DISTRIBUTION OF $\hat{I}_i'$, $E(r_j) = E(R_j')$, $n = 5$

N	i	I_i'	$E(\hat{I}_i')$	$\hat{\sigma}^2$	$\sqrt{\beta_1}$	β_2
	2	0.44764	0.44930	0.011194	0.1468	2.889
50	3	.48384	.47995	.012677	- .1121	3.027
	4	.06852	.07075	.003937	.0112	4.873
	2	.44764	.44880	.005468	.0516	2.898
100	3	.48384	.48281	.006256	- .0106	3.123
	4	.06852	.06839	.001828	.7365	3.718
	2	.44764	.44800	.002539	- .0093	3.092
200	3	.48384	.48345	.002916	- .0525	2.977
	4	.06852	.06855	.000914	.4353	3.233
	2	.44764	.44631	.001263	- .0286	3.215
400	3	.48384	.48498	.001274	- .0284	2.997
	4	.06852	.06871	.000419	.2021	2.858
	2	.44764	.44805	.000624	.0452	3.116
800	3	.48384	.48432	.000645	- .0991	3.495
	4	0.06852	0.06763	0.000216	0.3846	3.385

TABLE IV

MOMENT STATISTICS OF THE DISTRIBUTION OF $\hat{I}_i'$, $E(r_j) = E(r_j')$, $n = 10$

N	i	I_i'	$E(\hat{I}_i')$	$\hat{\sigma}^2$	$\sqrt{\beta_1}$	β_2
	2	0.43899	0.44031	0.006360	0.2097	3.160
100	3	.49155	.49095	.007144	- .0561	3.020
	4	.06946	.06864	.002320	.1045	4.393
	2	.43899	.44046	.002826	- .0233	2.964
200	3	.49155	.49061	.003080	- .1429	2.852
	4	.06946	.06893	.001075	.6729	3.837
	2	.43899	.44026	.001310	.1006	2.892
400	3	.49155	.49061	.001491	- .1157	2.805
	4	.06946	.06913	.000542	.5491	3.671
	2	.43899	.43962	.000688	- .0314	3.048
800	3	.49155	.49180	.000743	.1110	2.860
	4	0.06946	0.06858	0.000222	0.1865	2.821

TABLE V

MOMENT STATISTICS OF THE DISTRIBUTION OF $\hat{I}_i'$, $E(r_j) = E(r_j')$, $n = 20$

N	i	I_i'	$E(\hat{I}_i')$	$\hat{\sigma}^2$	$\sqrt{\beta_1}$	β_2
	2	0.43937	0.43640	0.003204	0.2981	3.421
200	3	.49303	.49282	.003788	- .1858	3.123
	4	.06760	.07078	.001187	.6749	3.337
	2	.43937	.43663	.001613	.2444	3.233
400	3	.49303	.49442	.001822	- .0861	3.257
	4	.06760	.06895	.000589	.4860	3.271
	2	.43937	.43521	.000748	.1115	2.746
800	3	.49303	.49563	.000879	- .2299	3.048
	4	0.06760	0.06916	0.000299	0.2413	2.940

TABLE VI

MOMENT STATISTICS OF THE DISTRIBUTION OF $\hat{I}_i'$, $E(r_j) \neq E(r_j')$, $n = 5$

N	i	I_i'	$E(\hat{I}_i')$	$\hat{\sigma}^2$	$\sqrt{\beta_1}$	β_2
	2	0.45924	0.45749	0.010416	0.2879	3.094
50	3	.47365	.47271	.011338	-0.1954	3.102
	4	.06711	.06980	.003450	1.1114	4.903
	2	.45924	.45866	.005509	0.2946	3.376
100	3	.47365	.47609	.005837	-0.1816	3.085
	4	.06711	.06525	.001780	0.8727	3.940
	2	.45924	.45896	.002686	0.1538	2.924
200	3	.47365	.47199	.002927	-0.1031	2.887
	4	.06711	.06905	.000910	0.5819	3.562
	2	.45924	.46049	.001384	0.2008	2.880
400	3	.47365	.47189	.001527	-0.2220	3.028
	4	.06711	.06762	.000427	0.3088	3.159
	2	.45924	.45915	.000627	0.0803	3.170
800	3	.47365	.47454	.000665	0.0271	3.091
	4	0.06711	0.06631	0.000200	0.2271	3.004

TABLE VII

MOMENT STATISTICS OF THE DISTRIBUTION OF $\hat{I}_i'$, $E(r_j) \neq E(r_j')$, n = 10

N	i	I_i'	$E(\hat{I}_i')$	$\hat{\sigma}^2$	$\sqrt{\beta_1}$	β_2
	2	0.44462	0.44614	0.005896	0.3345	3.192
100	3	.48664	.48671	.006362	-0.1887	2.822
	4	.06874	.06715	.002216	1.1813	5.279
	2	.44462	.44163	.002849	0.2527	3.089
200	3	.48664	.48955	.003310	-0.2099	3.136
	4	.06874	.06882	.001163	1.1002	6.548
	2	.44462	.44367	.001389	0.1178	2.992
400	3	.48664	.48786	.001511	-0.0076	2.857
	4	.06874	.06847	.000512	0.4299	3.440
	2	.44462	.44474	.000721	0.0823	3.212
800	3	.48664	.48730	.000830	-0.1261	2.992
	4	0.06874	0.06796	0.000254	0.4571	3.658

Figures 2 - 4 to illustrate the nature of the distributions. The histograms were plotted on the ORACLE's photographic curve plotter. Examination of Tables III - VII and Figures 2 - 4 indicates that the sampling distributions of $\hat{I}_2'$ and $\hat{I}_3'$ approximate the normal distribution. The sampling distributions of $\hat{I}_4'$ appear to be positively skewed although the amount of skewness is small.

Similar tables and histograms were constructed for the variance estimates obtained from the two variance formulas. These distributions resembled the chi-square distribution. Figures 5 - 10 represent typical histograms of the variance estimates. Again, those histograms arising from case (1b) with a sample size of four hundred were chosen for an illustration of the distributions.

<u>Bias of the estimates</u>. Consider a sample of $x_1, x_2, \ldots, x_N$ independent random variables with a probability density function $f(x; \theta_1, \theta_2, \ldots, \theta_m)$, where the θ_i are the parameters in the distribution. If $\hat{\theta}_i(x_1, x_2, \ldots, x_m)$ is an estimator of θ_i, the bias of $\hat{\theta}_i$ is defined by

$$b_i \equiv E(\hat{\theta}_i) - \theta_i \quad . \tag{4.28}$$

If $b_i = 0$, then $\hat{\theta}_i$ is said to be unbiased. If $b_i > 0$, $\hat{\theta}_i$ is said to be positively biased; while if $b_i < 0$, $\hat{\theta}_i$ is said to be negatively biased. Unbiased estimates enjoy great popularity, but use of them is not a crucial requirement. If an estimate differs but little from the parameter value relative to the standard deviation of the estimate, the estimate may be quite satisfactory.

The biases in $\hat{I}_i'$ together with their standard deviations, s.d.(b_i), are presented in Tables VIII - X. There is no evidence of

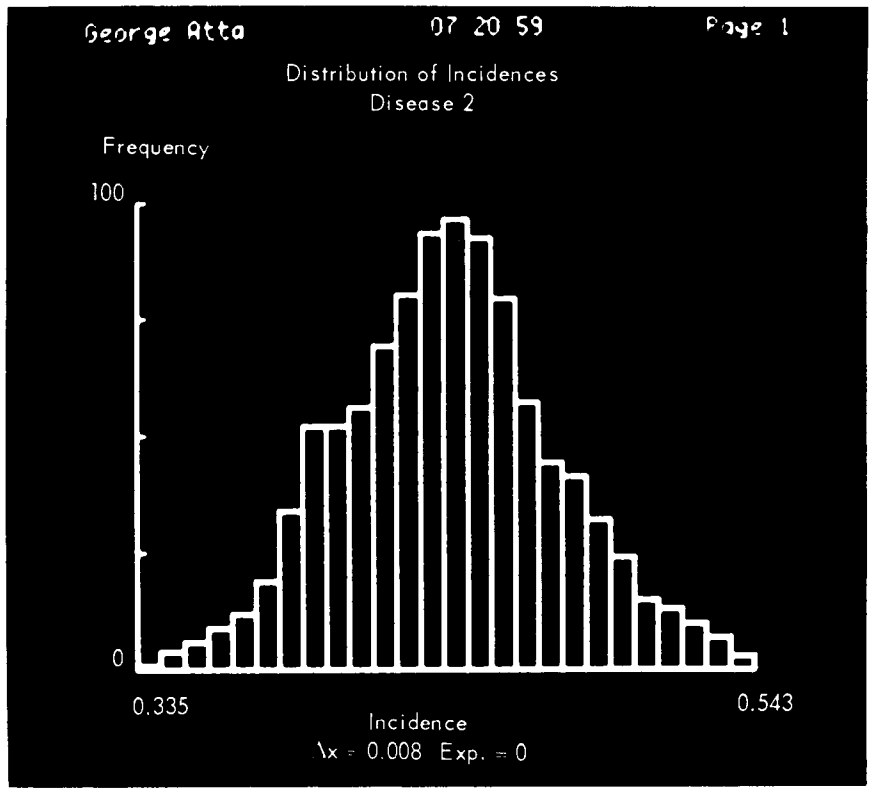

Figure 2. Distribution of 1,000 $\hat{I}_2'$, $E(r_j) = E(r_j')$, $n = 10$, $N = 400$.

Note: The interval length is denoted by Δx while "exp." specifies the power of ten by which the abscissas were multiplied.

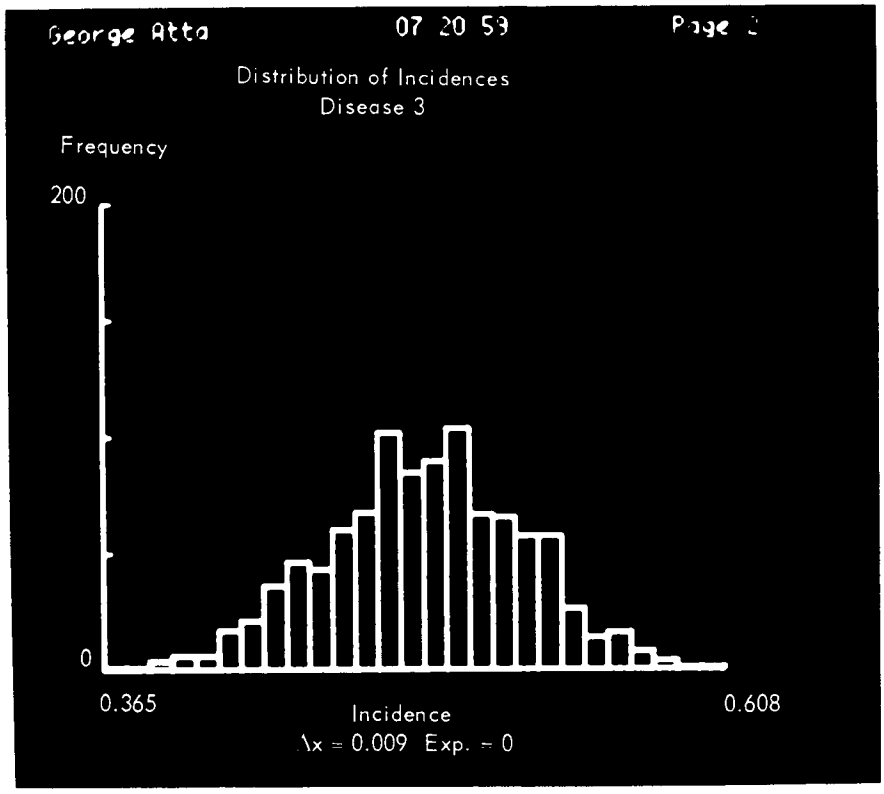

Figure 3. Distribution of 1,000 $\hat{I}_3'$, $E(r_j) = E(r_j')$, n = 10 N = 400.

Note: The interval length is denoted by Δx while "exp." specifies the power of ten by which the abscissas were multiplied.

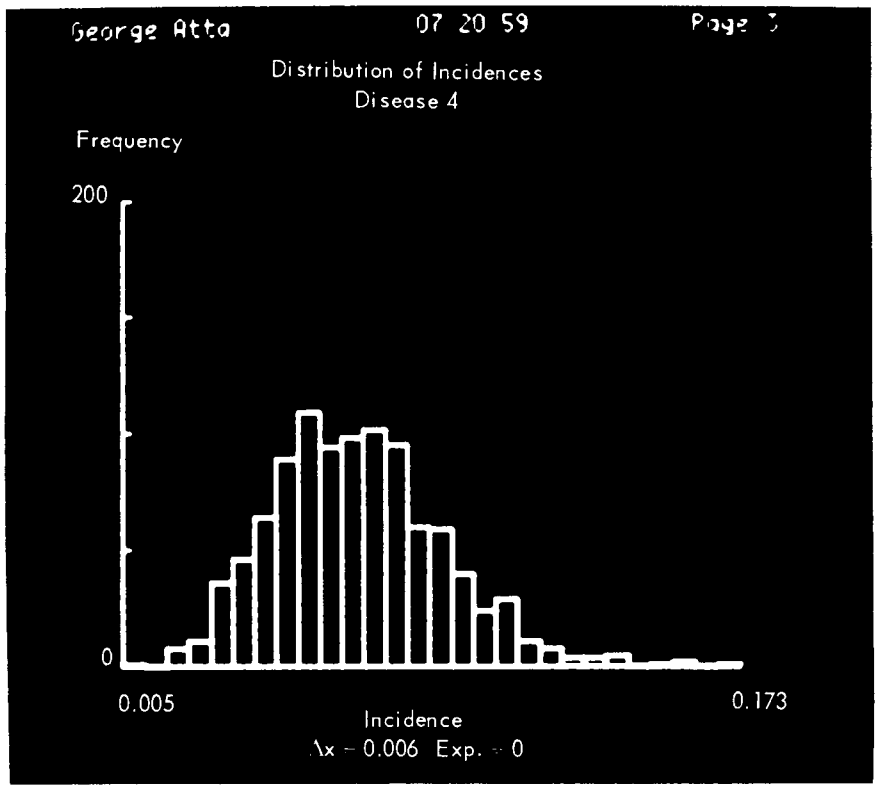

Figure 4. Distribution of 1,000 $\hat{I}_4'$, $E(r_j) = E(r_j')$, $n = 10$, $N = 400$.

Note: The interval length is denoted by Δx while "exp." specifies the power of ten by which the abscissas were multiplied.

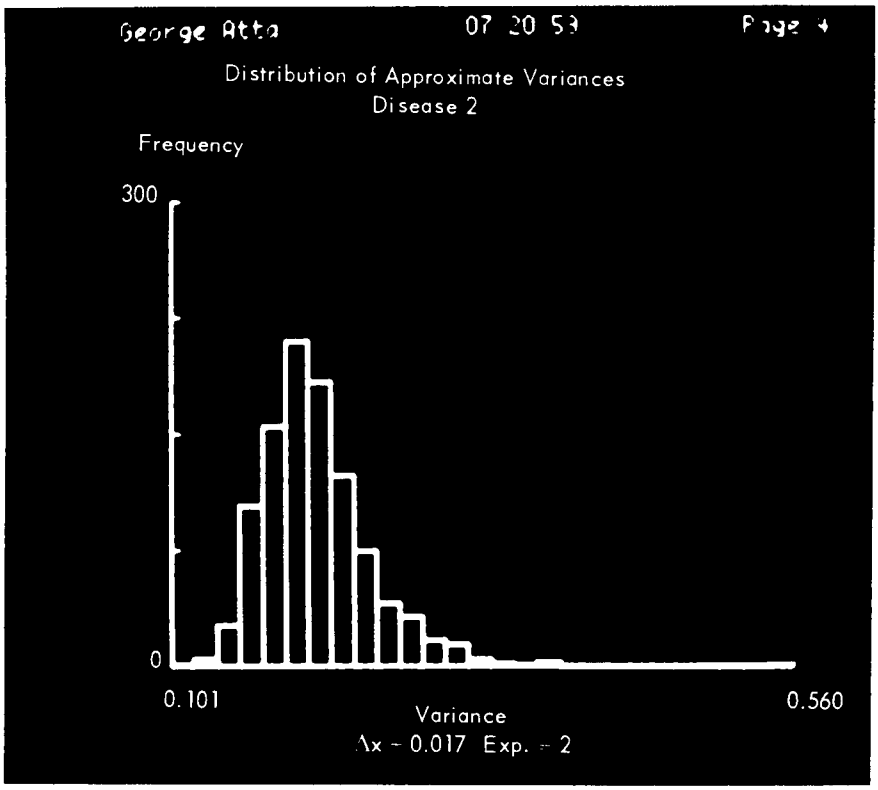

Figure 5. Distribution of 1,000 approximate variances of $\hat{I}_2'$, $E(r_j) = E(r_j')$, $n = 10$, $N = 400$.

Note: The interval length is denoted by Δx while "exp." specifies the power of ten by which the abscissas were multiplied.

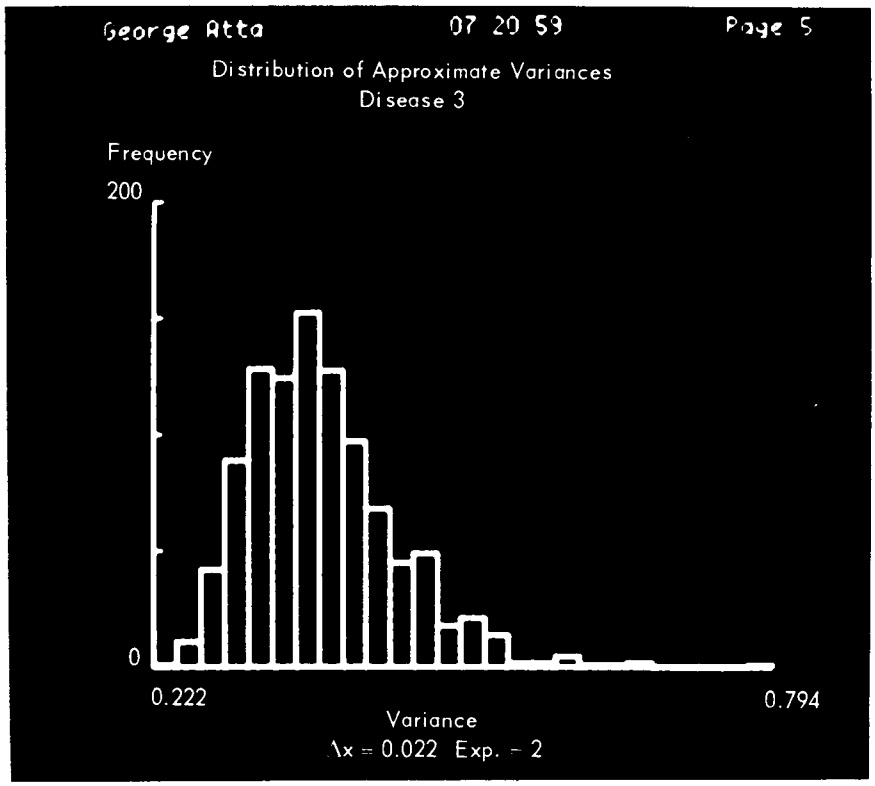

Figure 6. Distribution of 1,000 approximate variances of $\hat{I}_3'$, $E(r_j) = E(r_j')$, $n = 10$, $N = 400$.

Note: The interval length is denoted by Δx while "exp." specifies the power of ten by which the abscissas were multiplied.

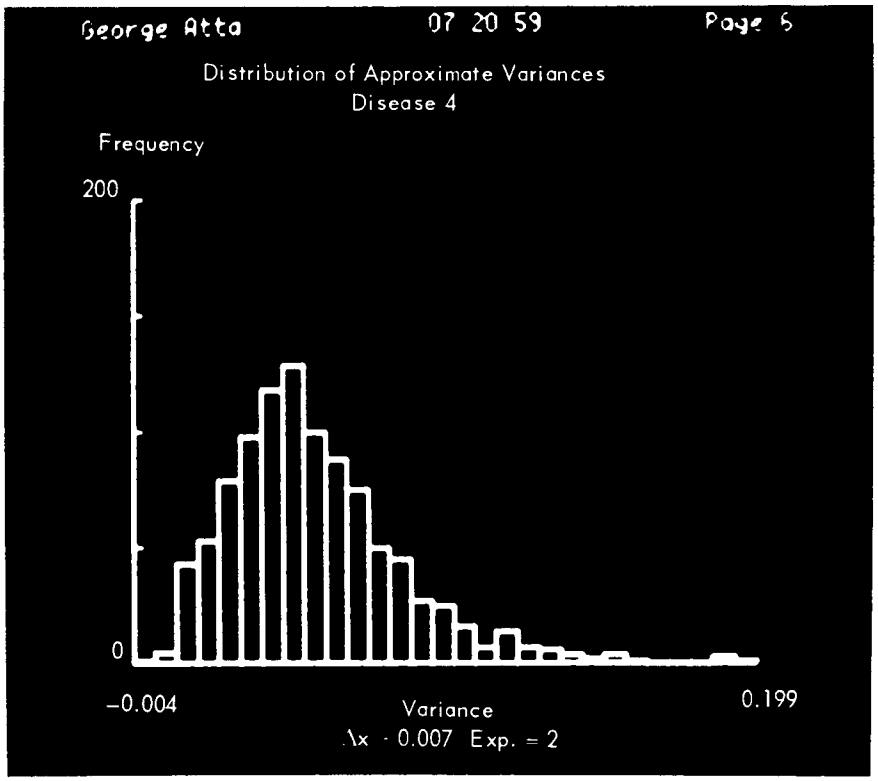

Figure 7. Distribution of 1,000 approximate variances of $\hat{I}_4'$, $E(r_j) = E(r_j')$, $n = 10$, $N = 400$.

Note: The interval length is denoted by Δx while "exp." specifies the power of ten by which the abscissas were multiplied.

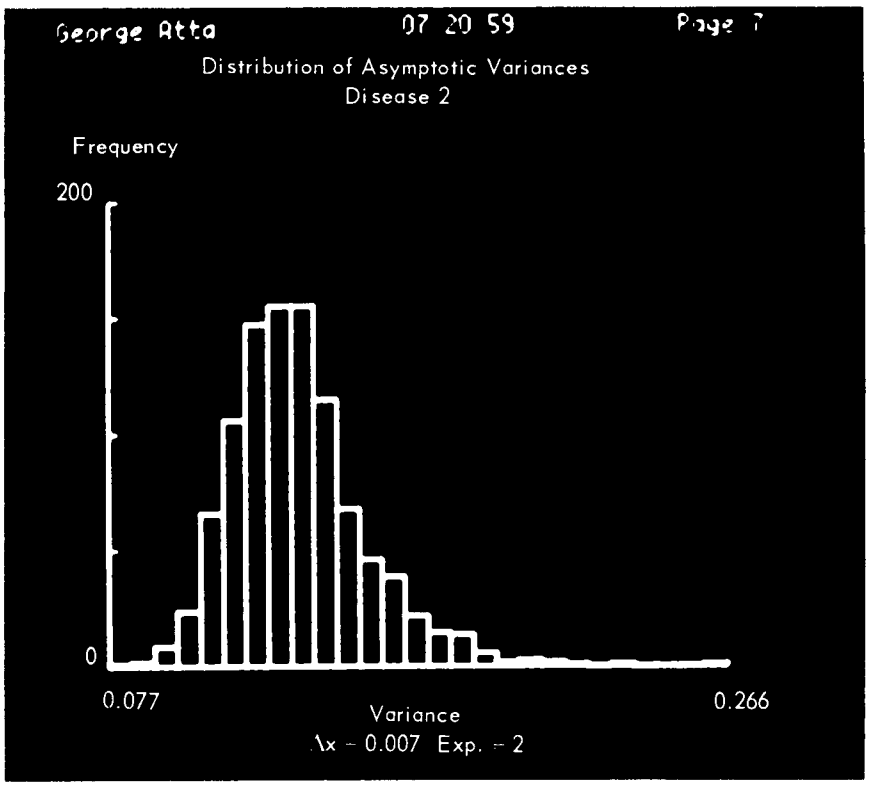

Figure 8. Distribution of 1,000 asymptotic variances of $\hat{I}_2'$, $E(r_j) = E(r_j')$, $n = 10$, $N = 400$.

Note: The interval length is denoted by Δx while "exp." specifies the power of ten by which the abscissas were multiplied.

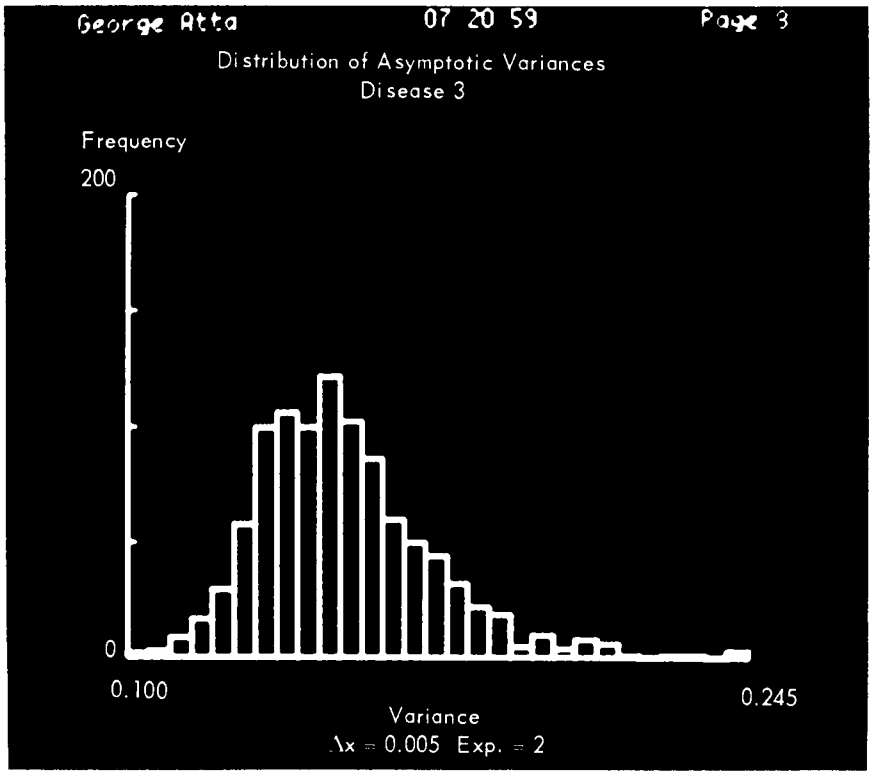

Figure 9. Distribution of 1,000 asymptotic variances of $\hat{I}_3'$, $E(r_j) = E(r_j')$, $n = 10$, $N = 400$.

Note: The interval length is denoted by Δx while "exp." specifies the power of ten by which the abscissas were multiplied.

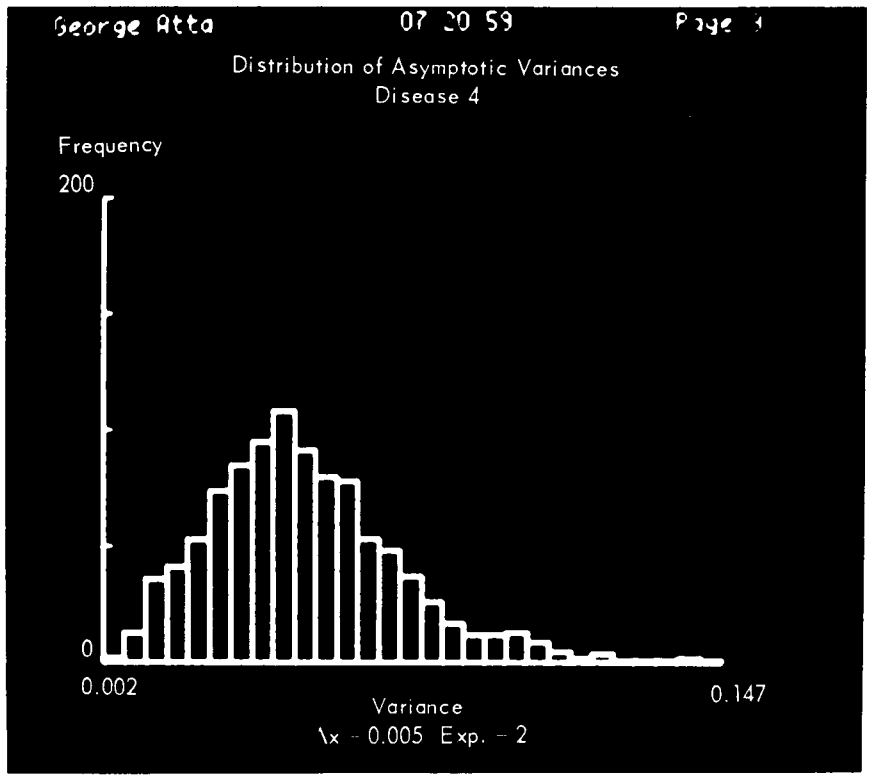

Figure 10. Distribution of 1,000 asymptotic variances of $\hat{I}_4'$, $E(r_j) = E(r_j')$, $n = 10$, $N = 400$.

Note: The interval length is denoted by Δx while "exp." specifies the power of ten by which the abscissas were multiplied.

TABLE VIII

BIAS AND STANDARD DEVIATION OF THE BIAS OF $\hat{I}_i'$, n = 5

		$E(r_j) = E(r_j')$		$E(r_j) \neq E(r_j')$	
N	i	b_i	s.d.(b_i)	b_i	s.d.(b_i)
50	2	0.00166	0.00335	-0.00175	0.00323
	3	- .00389	.00356	- .00094	.00337
	4	.00223	.00198	.00268	.00186
100	2	.00116	.00234	- .00057	.00235
	3	- .00102	.00250	.00244	.00242
	4	- .00013	.00135	- .00186	.00133
200	2	.00036	.00159	- .00028	.00164
	3	- .00038	.00171	.00165	.00171
	4	.00002	.00096	.00193	.00095
400	2	- .00133	.00112	.00125	.00118
	3	.00115	.00113	- .00176	.00124
	4	.00018	.00065	.00050	.00065
800	2	.00041	.00079	- .00009	.00079
	3	.00049	.00080	.00089	.00082
	4	-0.00089	0.00046	-0.00080	0.00045

TABLE IX

BIAS AND STANDARD DEVIATION OF THE BIAS OF $\hat{I}_i'$, n = 10

		$E(r_j) = E(r_j')$		$E(r_j) \neq E(r_j')$	
N	i	b_i	s.d.(b_i)	b_i	s.d.(b_i)
	2	0.00133	0.00252	0.00152	0.00243
100	3	- .00061	.00267	.00007	.00252
	4	- .00072	.00152	- .00159	.00149
	2	.00148	.00168	- .00299	.00169
200	3	- .00094	.00176	.00291	.00182
	4	- .00053	.00104	.00008	.00108
	2	.00128	.00114	- .00094	.00118
400	3	- .00094	.00122	.00121	.00123
	4	- .00033	.00074	- .00027	.00072
	2	.00063	.00083	.00012	.00085
800	3	.00025	.00086	.00066	.00091
	4	-0.00088	0.00047	-0.00077	0.00050

TABLE X

BIAS AND STANDARD DEVIATION OF THE BIAS OF $\hat{I}_i'$, n = 20

N	i	b_i	s.d.(b_i)
200	2	-0.00297	0.00179
	3	- .00021	.00195
	4	.00318	.00109
400	2	- .00274	.00127
	3	.00139	.00135
	4	.00135	.00077
800	2	- .00416	.00086
	3	.00260	.00094
	4	0.00156	0.00055

systematic behavior in the bias. No pronounced relationship could be discerned between the bias and either the sample size or the number of time intervals in the sample. The variations in the number of deaths per time interval appear to have no noticeable effect on the bias. Actually the indicated biases are all so small relative to the distribution standard deviations that they may actually be zero in the true infinite populations.

Similar tables were constructed for the bias in the estimate of $\mathrm{Var}(\hat{I}_i')$. Here the population variance was obtained by setting $\mathrm{Var}(\hat{I}_i') \equiv \hat{\sigma}_i^2$, where $\hat{\sigma}_i^2$ is the sample variance of $\hat{I}_i'$. These tables indicate that the variances computed from the approximate variance formula were positively biased while those obtained from the asymptotic formula were predominantly negatively biased. In other words, the asymptotic formula tended to overcorrect for the bias in the approximate formula.

To compensate for the bias in estimating $\mathrm{Var}(\hat{I}_i')$, consider the problem of deriving a correction factor for the variance estimator. Let $\hat{V}_1(\hat{I}_i')$ and $\hat{V}_2(\hat{I}_i')$ be the expected values of the sample variances based on the N variances computed from the approximate variance formula and the asymptotic variance formula respectively. In any case, the bias is given by

$$b_i \equiv \mathrm{Var}(\hat{I}_i') - \hat{V}(\hat{I}_i') \quad . \tag{4.29}$$

For each I_i' these biases and the reciprocals of their corresponding sample sizes were plotted as points in a plane. These scatter diagrams indicated a rather strong linear trend. Since $E(b_i) = 0$ as $N \to \infty$, the

function

$$b = \frac{\gamma}{N} \tag{4.30}$$

was fitted to the data by the method of least squares. A scatter diagram for the γ's and the incidences indicated that

$$\gamma = \delta_1 I' + \delta_2 I'^2 \tag{4.31}$$

would yield a satisfactory fit. Substituting (4.31) in (4.30), one finds that

$$b = \frac{\delta_1 I' + \delta_2 I'^2}{N} \quad . \tag{4.32}$$

This expresses the bias as a function of the incidence and sample size. The function (4.32) may be fitted to the data by least squares.

On setting $b = \mathrm{Var}(\hat{I}') - \hat{V}(\hat{I}')$ in (4.32), this function takes the form

$$\mathrm{Var}(\hat{I}') - \hat{V}(\hat{I}') = \frac{\delta_1 I' + \delta_2 I'^2}{N} \quad , \tag{4.33}$$

so that

$$\mathrm{Var}(\hat{I}') = \hat{V}(\hat{I}') + \frac{\delta_1 I' + \delta_2 I'^2}{N} \quad . \tag{4.34}$$

The above technique was applied to the data obtained from both variance formulas. As a result of these computations, the variances of the incidence estimates were found to be

$$\mathrm{Var}(\hat{I}_i') \cong \hat{V}_1(\hat{I}_i') + \frac{1.715 I_i' - 6.207 I_i'^2}{N} \tag{4.35}$$

if the approximate variance formula is used, and

$$\text{(4.36)} \qquad \mathrm{Var}(\hat{I}_i') \cong \hat{V}_2(\hat{I}_i') + \frac{0.3748 I_i' - 0.4414 I_i'^2}{N}$$

if the asymptotic variance formula is used. Thus, it appears that one can remove most of the bias in estimating the sample variances by adding a factor, which is a function of both the incidence and sample size, to the variance estimate computed from either variance formula.

The least squares fit of (4.32) to the approximate variance data was better than the corresponding fit to the asymptotic variance data. Hence, one would prefer formula (4.35), both because it is more reliable and because the variance estimate is easier to compute.

CHAPTER V

SUMMARY AND CONCLUSIONS

A random sample of N animals from an infinite population is observed until all animals in the sample have died. At the time of death, the cause of death is recorded. The observation period is divided into n intervals, not necessarily of equal length, and the number of animals dying from each cause of death is calculated for each time interval. It is desired to estimate the disease incidences in a population in which one of the diseases has been eliminated as a cause of death. The estimation is to be performed with a sample from a population in which all causes of death are operating.

An estimation procedure developed by Kimball [7] is discussed in Chapter II. This method is based on the maximum likelihood estimates of the parameters of a distribution very similar to the multinomial distribution. In an attempt to improve upon Kimball's approximate formula for the variances of the estimates, a true asymptotic variance formula is derived in Chapter III.

To provide some information on the small sample properties of the estimates, a large-scale random sampling experiment was designed and carried out on the ORACLE. This empirical study reflected favorably upon the estimation procedure. The sampling distributions of $\hat{I}_2'$ and $\hat{I}_3'$ were approximately normal while the distributions of $\hat{I}_4'$ were positively skewed. This is not surprising, for I_2' and I_3' are nearly equal to one-half while I_4' is less than one-tenth. The biases in the incidence estimates were negligible.

The distributions of the variance estimates approximated the

chi-square distribution, as would be expected. One cannot infer from this study that the asymptotic variance formula is an improvement on the approximate formula. It turned out that the latter always over-estimated the actual variances whereas the former tended to under-estimate the actual variances. A correction factor for both variance formulas was derived that, in practice, should provide sample variances that have very little bias.

Unless a high-speed machine is available, it is impracticable to use the asymptotic variance formula. Since the computation procedure is elementary, the approximate variance formula has greater utility. Furthermore, its correction factor seems to produce more accurate solutions.

In the derivation of the correction factors for both variance formulas, it was necessary to fit a curve of the form

$$b = N^{-1}(\delta_1 I' + \delta_2 I'^2) \tag{5.1}$$

to three clusters of points. This restriction was imposed by the design of the sampling experiment. Improved correction factors may be obtained by designing an experiment such that greater variation in the true incidences results.

BIBLIOGRAPHY

[1] Joseph Berkson and Robert P. Gage, "Survival curve for cancer patients following treatment," Journ. Amer. Stat. Assoc., Vol. 47 (1952), pp. 501-515.

[2] Harold Cramér, Mathematical Methods of Statistics, Princeton University Press, 1946.

[3] Evelyn Fix and Jerzy Neyman, "A simple stochastic model of recovery relapse, death and loss of patients," Human Biology, Vol. 23 (1951), pp. 205-241.

[4] T. E. Harris, Paul Meier, and John W. Tukey, "Timing of the distribution of events between observations: a contribution to the theory of follow-up studies," Human Biology, Vol. 22 (1950), pp. 249-270.

[5] Alston S. Householder, Principles of Numerical Analysis, McGraw-Hill Book Company, 1953.

[6] E. L. Kaplan and Paul Meier, "Non-parametric estimation of survivorship," unpublished paper, 1954.

[7] A. W. Kimball, "Disease incidence estimation in populations subject to multiple causes of death," Bull. of Internat. Stat. Inst., Vol. 36 (1958), pp. 193-204.

[8] A. W. Kimball and G. J. Atta, "A model for the multiple-cause-of-death problem," Math. Semiann. Prog. Rep., ORNL 2134 (1956), pp. 13-15.

[9] Alexander M. Mood, Introduction to the Theory of Statistics, McGraw-Hill Book Company, 1950.

[10] C. Radhakrishna Rao, Advanced Statistical Methods in Biometric Research, John Wiley and Sons, 1952.

[11] S. N. Roy and A. E. Sarhan, "On inverting a class of patterned matrices," Biometrika, Vol. 43 (1956), pp. 227-231.

[12] M. R. Sampford, "The estimation of response-time distribution II. Multi-stimulus distributions," Biometrics, Vol. 8 (1952), pp. 307-369.

[13] Olga Taussky and John Todd, "Generation and testing of pseudo-random numbers," Symposium on Monte Carlo Methods, John Wiley and Sons, 1956, pp. 15-28.

ORNL-2957
UC-34 Physics and Mathematics
TID-4500 (15th ed.)

INTERNAL DISTRIBUTION

1. C. E. Center
2. Biology Library
3. Health Physics Library
4-5. Central Research Library
6. Reactor Experimental Engineering Library
7-26. Laboratory Records Department
27. Laboratory Records, ORNL R.C.
28. A. M. Weinberg
29. L. B. Emlet (K-25)
30. J. P. Murray (Y-12)
31. J. A. Swartout
32. R. S. Cockreham
33. E. D. Shipley
34. M. L. Nelson
35. A. S. Householder
36. A. H. Snell
37. E. H. Taylor
38. A. Hollaender
39. W. H. Jordan
40. H. E. Seagren
41. M. T. Kelley
42. K. Z. Morgan
43. D. Phillips
44. C. E. Winters
45. E. O. Wollan
46. E. P. Blizard
47. R. S. Livingston
48. C. P. Keim
49. J. W. Givens
50. C. D. Susano
51. R. A. Charpie
52. J. Harrison
53. N. M. Dismuke
54. A. H. Culkowski
55. A. W. Kimball
56. M. R. Arnette
57. P. M. Reyling
58. E. Leach
59. A. C. Downing, Jr.
60. G. J. Atta
61. J. A. Lane
62. R. W. Johnson
63. M. J. Skinner
64. J. L. Gabbard
65. E. C. Long
66. M. A. Kastenbaum
67. J. P. Kelly (K-25)
68. S. E. Atta
69. D. A. Gardiner
70. C. A. Preskitt
71. F. L. Culler
72. W. S. Snyder
73. J. L. Fowler
74. C. J. Borkowski
75. G. E. Boyd
76. W. R. Grimes
77. D. S. Billington
78. J. H. Frye, Jr.
79. R. B. Briggs
80. A. Simon
81. H. Bottenbruch
82. E. L. Cooper
83. M. Feliciano
84. W. Gautschi
85. A. A. Grau
86. N. B. Alexander
87. D. E. Arnurius
88. L. L. Bumgarner
89. M. M. Callahan
90. H. P. Carter
91. C. T. Fike
92. J. E. Parham
93. D. J. Wehe
94. A. C. Upton
95. ORNL - Y-12 Technical Library, Document Reference Section

EXTERNAL DISTRIBUTION

96. Division of Research and Development, AEC, ORO
97. A. E. S. Green, Florida State University
98-99. Department of Mathematics, University of Tennessee (1 copy each to G. E. Albert and J. A. Cooley)
100. J. Furth, Roswell Park Memorial Institute, Buffalo, N.Y.
101. G. Sacher, Argonne National Laboratory
102-711. Given distribution as shown in TID-4500 (15th ed.) under Physics and Mathematics category (75 copies - OTS)