INFORMATION TO USERS

This material was produced from a microfilm copy of the original document. While the most advanced technological means to photograph and reproduce this document have been used, the quality is heavily dependent upon the quality of the original submitted.

The following explanation of techniques is provided to help you understand markings or patterns which may appear on this reproduction.

1. The sign or "target" for pages apparently lacking from the document photographed is "Missing Page(s)". If it was possible to obtain the missing page(s) or section, they are spliced into the film along with adjacent pages. This may have necessitated cutting thru an image and duplicating adjacent pages to insure you complete continuity.

2. When an image on the film is obliterated with a large round black mark, it is an indication that the photographer suspected that the copy may have moved during exposure and thus cause a blurred image. You will find a good image of the page in the adjacent frame.

3. When a map, drawing or chart, etc., was part of the material being photographed the photographer followed a definite method in "sectioning" the material. It is customary to begin photoing at the upper left hand corner of a large sheet and to continue photoing from left to right in equal sections with a small overlap. If necessary, sectioning is continued again — beginning below the first row and continuing on until complete.

4. The majority of users indicate that the textual content is of greatest value, however, a somewhat higher quality reproduction could be made from "photographs" if essential to the understanding of the dissertation. Silver prints of "photographs" may be ordered at additional charge by writing the Order Department, giving the catalog number, title, author and specific pages you wish reproduced.

5. PLEASE NOTE: Some pages may have indistinct print. Filmed as received.

Xerox University Microfilms
300 North Zeeb Road
Ann Arbor, Michigan 48106

74-15,771

KALSBEEK, William Dean, 1946-
A METHOD FOR OBTAINING LOCAL POSTCENSAL ESTIMATES FOR SEVERAL TYPES OF VARIABLES.

The University of Michigan, Ph.D., 1973
Sociology, demography

University Microfilms, A XEROX Company, Ann Arbor, Michigan

THIS DISSERTATION HAS BEEN MICROFILMED EXACTLY AS RECEIVED.

A METHOD FOR OBTAINING LOCAL POSTCENSAL ESTIMATES FOR SEVERAL TYPES OF VARIABLES

by
William Dean Kalsbeek

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Biostatistics)
in The University of Michigan
1973

Doctoral Committee:

Professor Richard G. Cornell, Co-Chairman
Professor Leslie Kish, Co-Chairman
Assistant Professor John W. McGuire
Professor M. Anthony Schork

DEDICATION

To the memory of my late grandfather, Willem Kalsbeek,
whose wisdom, integrity, and insatiable quest for knowledge
have served and will continue to serve
as a great inspiration to me.

ACKNOWLEDGEMENTS

I am deeply indebted to Drs. Richard G. Cornell, Leslie Kish, M. Anthony Schork and John W. McGuire for having most ably served as members of my doctoral committee. Their many comments and suggestions were very helpful, particularly in the formative stage of my research and during the preparation of this manuscript. In addition, I wish to acknowledge Dr. Cornell for his personal involvement as director of the joint statistical project between the University of Michigan and the United States Bureau of the Census, which served as the means of financial support for my work, and also for his help in developing the analytic approach. Also, I wish to especially thank Dr. Kish for his tutelage, encouragement and personal interest in the present research as well as throughout my graduate career.

I wish to thank Dr. Benjamin Tepping of the Bureau of the Census for having served as representative for the Bureau in this project. This involved a great deal of administrative work in setting up the project as well as in procuring several different data sets. In this connection I also wish to thank Mr. Frederick Cavanaugh and Mr. Jerome Glynn of the Bureau's Population Division for arranging for me to obtain the necessary data on vital

statistics. Also thanks go to Dr. Eugene Ericksen of Temple University's Institute for Survey Research for providing me with the results of his research which made comparisons more meaningful.

I wish to acknowledge Ms. Mary Uranga for handling the many administrative details of this project. I also wish to thank Ms. Louise Dapprich for swiftly and accurately key-punching several data sets and Ms. Marija J. Norusis for her help in debugging several of the computer programs. Finally, I wish to thank Ms. Therese Birney and Ms. Glenna Petersen for their excellent job of typing this manuscript.

TABLE OF CONTENTS

Chapter		Page
I.	BACKGROUND AND MOTIVATION	1
	1.1 Introduction	1
	1.2 Description of existing methods	2
	1.3 Historical background	16
	1.4 Comparison of existing methods	19
	1.5 Motivation for proposed methods	25
	1.6 Conclusion	30
II.	THE PROPOSED METHOD	33
	2.1 Introduction	33
	2.2 Notation	36
	2.3 Grouping and classification procedures	37
	2.4 Estimation procedure	51
	2.5 Trial run	54
	2.6 Research itinerary	60
III.	AN ANALYTIC FORMULATION	61
	3.1 Introduction	61
	3.2 Preliminaries and notation	62
	3.3 General description	64
	3.4 Distribution-specific results	74
	3.5 Tabulations and summary	106
IV.	MULTIVARIATE EXTENSIONS BY MONTE CARLO SIMULATION	115
	4.1 Introduction	115
	4.2 Discussion of the models	116
	4.3 General simulation procedure	119
	4.4 Discussion of program input parameters	123
	4.5 Presentation and interpretation of results	127

Chapter Page

V. TWO ILLUSTRATIVE EXAMPLES 141

5.1 Introduction 141
5.2 General comments on procedure. 142
5.3 Population size example and results. . 149
5.4 Per capita income example and results. 156
5.5 Procedure for incidence of dilapicated housing example. 163

VI. SOME CONCLUDING REMARKS 166

6.1 Introduction 166
6.2 Procedural review of present research. 166
6.3 General conclusions. 171
6.4 Suggested further research 173

Appendix

A. APPROXIMATIONS FOR THE MOMENTS OF ORDER STATISTICS FROM THE STANDARD NORMAL DISTRIBUTION. 176

B. APPROXIMATIONS FOR THE CUMULATIVE DISTRIBUTION FUNCTION AND THE INVERSE CUMULATIVE DISTRIBUTION FUNCTION OF THE NORMAL DISTRIBUTION. 180

C. METHODS FOR OBTAINING GROUP BOUNDARIES. . . 184

D. GENERATORS OF RANDOM NORMAL DEVIATES. . . . 193

SELECTED BIBLIOGRAPHY 200

LIST OF TABLES

Table		Page
1.1	A Compendium of Research Dealing with Methods for Obtaining Postcensal Estimates	20
2.1	Several Possible Situations Describing Sources of Data for the Proposed Methods	49
2.2	Results of a Trial Run Involving 1960 Median Income in Michigan.	59
3.1	Ranges of Parameter Values Used in Computing the Analytic Results.	107
3.2	A Summary of Tabulated Analytic Results for the Normal Distribution	108
3.3	A Summary of Tabulated Analytic Results for the Logistic Type 1 Distribution.	109
3.4	A Summary of Tabulated Analytic Results for the Logistic Type 2 Distribution.	110
3.5	A Summary of Tabulated Analytic Results for the Exponential Distribution.	111
4.1	Sample Sizes and Numbers of Groups Used in Simulations.	126
4.2	Optimum Stratification Points and Branch Means for the Standard Normal Distribution.	126
4.3	Estimated Parameters of the Ericksen and Proposed Methods with High Association; n=120 ; n*=49	130
4.4	Estimated Parameters of the Ericksen and Proposed Methods with High Association; n=480 ; n*=49	131

Table		Page
4.5	Estimated Parameters of the Ericksen and Proposed Methods with High Association; n=120 ; n*=100.	132
4.6	Estimated Parameters of the Ericksen and Proposed Methods with High Association; n=480 ; n*=100.	133
4.7	Estimated Parameters of the Ericksen and Proposed Methods with Moderate Association; n=120 ; n*=49	134
4.8	Estimated Parameters of the Ericksen and Proposed Methods with Moderate Association; n=480 ; n*=49	135
4.9	Estimated Parameters of the Ericksen and Proposed Methods with Moderate Association; n=120 ; n*=100.	136
4.10	Estimated Parameters of the Ericksen and Proposed Methods with Moderate Association; n=480 ; n*=100.	137
5.1	Fitted Regression Coefficients Used by the Ericksen Method in the Illustrative Examples.	151
5.2	Relative Absolute Deviation Times 100 of the 1970 Population Size Estimates for States.	152
5.3	Mean Relative Absolute Deviation Times 100 of the 1970 Byproduct Population Size Estimates for Counties by the Actual 1970 Population Size	155
5.4	Per Cent Absolute Deviation of the 1970 Per Capita Income Estimates for States.	158
5.5	Mean Relative Absolute Deviation of 1970 Byproduct Per Capita Income Estimates for Counties by the Actual Per Capita Income. .	162
5.6	1970 State Estimates of the Proportion Times 100 of Housing Units that are Dilapidated .	165
C.1	Optimum Stratification Points for Three Distributions	188

LIST OF ILLUSTRATIONS

Figure		Page
2.1	Diagrammatic Representation of an Estimation Situation	34
2.2	Diagrammatic Representation of Rectilinear Classification	40
2.3	Diagrammatic Representation of Nonrectilinear Classification.	41
4.1	Representation of a Grid of Target Area Values on Two Symptomatic Variables. . . .	122
4.2	Boundary Points for the Four-group Simulation Model	128
4.3	Boundary Points for the Nine-group Simulation Model	128
4.4	Boundary Points for the Sixteen-group Simulation Model	129

CHAPTER I

BACKGROUND AND MOTIVATION

1.1 Introduction

We live in a society characterized by an ever growing need for self-examination in matters such as health, education, consumer affairs, political opinion, and economic well-being. For the most part this type of examination has been initiated by the governmental sector in the form of censuses, on many levels and for a variety of purposes, and a vital registration system. This trend has carried over to and become firmly established in the private and commercial sectors. The Health, Education, and Welfare administrator requires data to evaluate the effectiveness of his program while the commercial executive needs information to establish the potential market value of his product. Similarly, the local public health official utilizes vital health data to determine the segment of population in greatest need of his immunization program while a researcher in social psychology might use crime statistics to establish patterns in urban street crime for an insurance company. It becomes readily apparent that data extraction for descriptive, evaluative, or analytic purposes has become an integral part of our society.

One of the results of this quantitative awareness is the need for a variety of data for areas such as states, counties and standard metropolitan statistical areas (SMSA's) as well as for even smaller areas such as census tracts, enumerator districts and blocks. Unfortunately, periodic data for such areas do not exist except those which are obtained from sources such as the vital registration systems and the decennial and other special censuses. A need for methods useful to obtain periodic estimates for these areas clearly exists and is the rationale for our present research.

1.2 Description of existing methods

To date virtually all methods for making local postcensal estimates have been formulated for population size estimation. In many cases these techniques allow one to obtain size estimates for age-, sex-, race-, and color-specific segments of the population. However, only two methods, the Ratio-Correlation method, developed by Schmitt and Crosetti [1954], and a method due to Ericksen [1971], permit estimation of other variables. This facet of existing methods is discussed more fully in a later section.

We begin with an outline of those methods which have gained acceptance at some time in the past and those which are presently used by the United States Bureau of

the Census for making local estimates. These methods have appeared in Census Bureau and other publications although others have no doubt been contrived and used. Estimates are also made by private firms such as Rand McNally, Editor and Publisher, and Sales Management, but for the most part these sources have not published a description or evaluation of the methods used. Thus the list presented is most likely nonexhaustive and should be considered as such. In the next section is presented a brief description of the methods followed by a short historical sketch.

The following convention pertaining to subscripts in this chapter is adopted.

0--Third most recent data point (denoted by t_0).

1--Second most recent data point (denoted by t_1).

2--Most recent data point (denoted by t_2).

3--Point in time for which the estimate is made (denoted by t_3).

a--Interval between t_1 and t_2.

b--Interval between t_2 and t_3.

i--Observation on the i-th local area among <u>all</u> local areas in the total area.

j--Observation on the j-th local area among a <u>sample</u> of local areas from the total area.

Unless specified otherwise the "data points" referred to above are dates of national censuses, since it is then that the relevant data are most often available. Interval

"a" would then be the intercensal period and interval "b" the postcensal period. The variables using these subscripts are defined as they appear.

The existing methods are of two types--those which employ arithmetic extrapolation and those which use symptomatic information. The extrapolation procedures are basically curve fitting techniques in which general assumptions are made regarding population change over time. Past data are used to "fit" the assumed functional form which is then extrapolated to t_3 to obtain the estimate. Thus no population beyond t_2 is used to formulate the final result. For this reason they are more appropriately designated as population projections than population estimates. This lack of population information beyond t_2 has been a primary objection and is a major reason for largely abandoning such methods (U.S. Bureau of the Census [1966a]). They are obviously inappropriate in areas which are "volatile" in the sense that they are subject to large, unpredictable changes in population size. Since only single observations at each point in time can be used, these fluctuations are not subject to the stabilizing influence of the Central Limit Theorem which would be the case if several observations were made at each point. They also do poorly in situations where b, the interval length between t_2 and t_3, is long.

The other type of method utilizes symptomatic in-

formation. It is "symptomatic" since related auxiliary information from the local areas is used to obtain the estimate. Ideally, these variables collectively exhibit a high level of association (i.e., degree of stochastic dependence) with the variable to be estimated, which we call the criterion variable.

Extrapolation methods

(1) Arithmetic Extrapolation--This method assumes that the yearly amount of population change in the postcensal period equals the average yearly amount of change in the most recent intercensal period. Thus a presupposition is made that arithmetic growth is uniform and thus linear over the interval, t_1 to t_3. The estimate is expressed as

$$P_3 = \frac{P_2(t_3-t_1)-P_1(t_3-t_2)}{t_2-t_1} \qquad (1.2.1)$$

where P refers to population size of the area, and the t's are expressed in years.

(2) Geometric Extrapolation--This method differs from the Arithmetic Extrapolation method in that yearly rate of change instead of "raw change" is assumed uniform over the intercensal and postcensal periods and equal to the actual rate of change in the former period. The estimate is

$$P_3=P_2(1+r)^{t_3-t_2} \qquad (1.2.2)$$

where

$$r=\left(\frac{P_2}{P_1}\right)^{\frac{1}{t_2-t_1}}-1 \qquad (1.2.3)$$

and r is the intercensal yearly rate of growth.

(3) Curve-fitting methods

(a) Logistic Curve--The Logistic Curve was first named by Verhulst, a Belgian mathematician, who used it in 1838 to express population growth.

It wasn't until the 1920's that Pearl [1925] applied this theory of population growth to biological phenomena, including the number of humans in a geographic area. The estimate in this method is

$$P_3=\frac{K}{1+\exp[W+Zt]} \qquad (1.2.4)$$

where

$$W=\log\frac{K-P_0}{P} \;;\; Z=\frac{1}{m}\log\frac{P_0(K-P_1)}{P_1(K-P_0)} \qquad (1.2.5)$$

and

$$K=\frac{2P_0P_1P_2-P_1^{\,2}(P_0+P_2)}{P_0P_2-P_1^{\,2}} \;;\; m=t_1-t_0=t_2-t_1 \quad . \qquad (1.2.6)$$

The letter t represents the number of time units (e.g., decades) since t_0. Note also that the t's may be spread over time and thus do not refer to the dates of most recent censuses.

(b) Gompertz Curve--The Gompertz Curve evolves when it is assumed that the growth increments of the logarithms are

declining at a constant percentage. Its use is also discussed by Pearl and the estimate is

$$P_3=QG^{Ft} \qquad (1.2.7)$$

where

$$F^m=\frac{\log P_2-\log P_1}{\log P_1-\log P_0}\ ;$$

$$\log G=\frac{\log P_1-\log P_0}{F^m-1}\ ;\ \log Q=\log P_0-\log G\ . \qquad (1.2.8)$$

Q represents the maximum population size obtainable.

(c) Whelpton method--In the Whelpton method [1936] trends in birth, death, and migration rates are observed and projected to t_3 by age, race, sex, etc. and the estimate is

$$P_3=P_2+B_b-D_b+M_b \qquad (1.2.9)$$

where B, D, and M represent births, deaths, and net migration, respectively, which are obtained from the projections, and the subscript, b, refers to the postcensal period.

Symptomatic methods

(1) Proration method--The Proration method (also called the Apportionment method) involves the distribution of an estimated total at t_3 (e.g., for a state) among its geographic constituent parts (e.g., its counties), assuming the geographic distribution at t_2. This method is symptomatic in that population size or some other related variable is used to determine the t_2 geographic distribution. Thus

$$P_3=p_2P_3 \qquad (1.2.10)$$

where p is the estimated proportion of the total population in the local area and P is the total population.

(2) Natural Increase method--If it is reasonable to assume no net migration in the postcensal period, the estimate found by the Natural Increase method equals the population at t_2 plus the number of births and minus the number of deaths (i.e., the natural increase) in the postcensal period, assuming this information is available. Thus

$$P_3=P_2+B_b-D_b \tag{1.2.11}$$

(3) Census Component Method I--The Census Component Method I is similar to the Natural Increase method except that an attempt is made to estimate the net migration component. This and a similar Component Method II were suggested by Eldridge (U.S. Bureau of the Census [1947]). In addition, provision is made for military population in the area so that the estimate is

$$P_3=P_2-A_2+A_3+B_b-D_b+M_b-I_b \tag{1.2.12}$$

where A refers to the population of the armed forces in the area, and I refers to the net movement of civilians from the area into the armed forces. The principal component to be estimated is M_b, the net civilian migration in the postcensal period. The rate is estimated by finding the net migration rate for school children in grades two through eight. This is estimated by the difference between the percentage change in the population of school age children for the area and the United States as a whole. This rate is

multiplied by an empirically determined factor to obtain the estimate of the net migration rate for the entire civilian population. Using a provisional estimate of the population of the area at t_3 an estimate of M_b is found.

(4) Census Component Method II--Method II as suggested by Eldridge is similar to Method I except for the method used to estimate M_b. Here postcensal school enrollment, birth, and death data are used to find the difference between the actual school enrollment figures and what would be expected as a result of the postcensal birth and death data. The school age net migration rate is then found and multiplied by an empirical constant to estimate the total net migration rate. M_b is then found using a provisional population estimate as in Method I. Quite obviously, these component methods are heavily dependent on the accuracy and consistency of school enrollment data from year to year. This dependency has been partially resolved by averaging the results with those obtained by other methods. A detailed description of this method has been published by the Bureau of the Census [1966b].

(5) Vital Rates method--In the Vital Rates method (Bogue [1950]) we assume that the ratio of birth and death rate, t_2 to t_3, for the total area and the local area to be estimated are equal. In other words

$$\frac{D_3 \mathcal{D}_2}{P_3 P_2} = \frac{D_2 \mathcal{D}_3}{P_2 P_3} \qquad (1.2.13)$$

and

$$\frac{B_3 \mathcal{B}_2}{P_3 P_2} = \frac{B_2 \mathcal{B}_3}{P_2 P_3} \qquad (1.2.14)$$

where $\mathcal{B}$ and $\mathcal{D}$ are total birth and death figures, respectively. In both (1.2.13) and (1.2.14) it is assumed that only P_3 is unknown so two estimates are obtained and averaged for the final result. Thus

$$P_3 = \frac{P_2 P_3 [D_3 \mathcal{D}_2 B_2 \mathcal{B}_3 + B_3 \mathcal{B}_2 D_2 \mathcal{D}_3]}{2 D_2 \mathcal{D}_3 B_2 \mathcal{B}_3 P_2} \quad . \qquad (1.2.15)$$

(6) Housing Unit method--In the Housing Unit method (U.S. Bureau of the Census [1957]) the estimate of population size is based upon current estimates (i.e., t_3) of the number of housing units in the area and the average number of individuals per housing unit. The former is obtained from postcensal data on building, demolition, and utilities data. The estimate of average size is found by observing past trends and extrapolating to t_3. Accuracy thus depends upon (1) completeness of the reporting of data on construction and demolition, (2) the pattern of time lag between issuance of building permits and the time when the unit is ready for occupancy, (3) changes in the average size of the household, (4) changes in vacancy rate, and (5) changes in the size of the nonhousehold population. The estimate is simply

$$P_3 = H_3 S_3 \qquad (1.2.16)$$

where H represents the number of housing units and S is the average size per housing unit. More recently the method has been modified to resemble a composite method (U.S. Bureau of the Census [1969]). The over 18 population is estimated as described above and the under 18 population is estimated as the same age group is estimated in the Composite method which follows.

(7) Composite methods--The most common version of this method is attributed to Bogue and Duncan [1959] and finds estimates for four separate age groups at t_3--(1) under 5, (2) 5-17, (3) 18-45, and (4) over 45. The estimate for age group (2) is found in similar fashion as for the same age group in the Component method or

$$P_3^{(2)} = P_2^{(2)} + B_b^{(2)} - D_b^{(2)} + M_b^{(2)} \qquad (1.2.17)$$

where the superscript $(\cdot)$ indicates the age group and $M_b^{(2)}$ is estimated as in the Census Component Method II, neglecting the empirical factor. The estimates for age groups (1) and (3) are obtained from postcensal period birth records. Adding up the number of births over the postcensal period and subjecting this number to a few mortality constraints yields the age group (1) estimate, denoted as $P_3^{(1)}$. Using past fertility rates and postcensal birth records, an estimate of the number of females in age group (3), denoted by $F_3^{(3)}$, is found. We then assume that

$$S_3^{(3)} S_2^{(3)} = S_2^{(3)} S_3^{(3)}$$

or

$$S_3^{(3)} = S_2^{(3)} S_3^{(3)} / S_2^{(3)} \qquad (1.2.18)$$

where S and S are total and local area sex ratios (i.e., number of males per female), respectively. Only $S_3^{(2)}$ is assumed unknown and the estimate is

$$P_3^{(3)} = (1+S_3^{(3)}) F_3^{(3)} \quad . \qquad (1.2.19)$$

To find the estimate for age group (4) postcensal death records for this age group are used. Assuming the relationship of equation (1.2.13) we solve for $P_3^{(4)}$ as

$$P_3^{(4)} = \frac{P_2^{(4)} p_3^{(4)} D_3^{(4)} \mathcal{D}_2^{(4)}}{D_2^{(4)} \mathcal{D}_3^{(4)} p_2^{(4)}} \quad . \qquad (1.2.20)$$

Finally, the composite estimate is expressed as

$$P_3 = \sum_{u=1}^{4} P_3^{(u)} \quad . \qquad (1.2.21)$$

Other forms of this method have been suggested. Frisen [1951] suggests a similar approach where all estimates are based solely upon deaths and death rates in the total and local areas. Other approaches use the combination of indicators which proves most effective for a specific situation.

The remaining symptomatic methods considered have two major distinguishing features. For one, they involve models which assume a specific functional relationship be-

tween the criterion and symptomatic variables. The procedure consists of "fitting" the functional relationship by ordinary least squares using some available data. It is the source of these data which distinguishes the two methods discussed below. The estimates are then obtained from this fitted functional form. The second feature unique to these methods is that they may be used to estimate variables other than population size. However, the only known applications of these methods to postcensal local areas involved estimation of population size where a linear model is assumed and where births, deaths, elementary school enrollment, tax returns, etc. have been used as symptomatic information.

(8) Ratio-Correlation method--The Ratio-Correlation method (also referred to as the Regression method) is generally attributed to Schmitt and Crosetti [1954] although the idea was first introduced by Snow [1911]. It has later also been used by Goldberg et al. [1960]. Here the estimate is obtained by fitting a linear model where the criterion and symptomatic variables are the ratio of the proportion in the local area at one time to the same proportion at another time. To describe the procedure we consider the case where the variable to be estimated is population size and the symptomatic variables are births and deaths. First, we define

$$P_{ai} = \frac{P_{2i}P_1}{P_{1i}P_2} \; ; \; P_{bi} = \frac{P_{3i}P_2}{P_{2i}P_3} \; ; \; i=1,2,\ldots,N \qquad (1.2.22)$$

and

$$B_{ai} = \frac{B_{2i}B_1}{B_{1i}B_2} \; ; \; B_{bi} = \frac{B_{3i}B_2}{B_{2i}B_3} \; ; \; i=1,2,\ldots,N \qquad (1.2.23)$$

$$D_{ai} = \frac{D_{2i}D_1}{D_{1i}D_2} \; ; \; D_{bi} = \frac{D_{3i}D_2}{D_{2i}D_3} \; ; \; i=1,2,\ldots,N \qquad (1.2.24)$$

where N is the number of local areas in the total population, and a and b refer to the intercensal and postcensal periods, respectively. The functional model assumed is summarized by

$$P_{ai} = \alpha + \beta B_{ai} + \gamma D_{ai} + \varepsilon_{ai} \qquad (1.2.25)$$

and

$$P_{bi} = \alpha + \beta B_{bi} + \gamma D_{bi} + \varepsilon_{bi} \qquad (1.2.26)$$

where α, β and γ are constants which are assumed to be the same for both intercensal and postcensal periods and the ε's represent stochastic error components with zero expectation, constant variance and zero covariance. Using the P_{ai}, B_{ai}, D_{ai} and the method of ordinary least squares, we obtain estimates, $\hat{\alpha}$, $\hat{\beta}$ and $\hat{\gamma}$, for the constants. Then the estimate for P_{bi} is

$$P_{bi} = \hat{\alpha} + \hat{\beta} B_{bi} + \hat{\gamma} D_{bi} \quad . \qquad (1.2.27)$$

Finally, from equation (1.2.21)

$$P_{3i} = \frac{P_{bi}P_{2i}P_3}{P_2} \; ; \; i=1,2,\ldots,N \quad . \qquad (1.2.28)$$

For this method to produce satisfactory results (1) the assumed and actual functional relationship must be the same or at least close, (2) the level of the multiple R must be very high, and (3) the functional relationships must be the same or at least close for the intercensal and postcensal periods. Failure on any one of the points is likely to diminish the effectiveness of the method.

(9) Ericksen Regression method--The Ericksen method is a modification of the Ratio-Correlation method in that the data used to get the estimates of α, β and γ are exclusively postcensal. Furthermore, the data for P_{3j} to obtain the estimate for P_{bi} are obtained from a sample survey. This general approach, although suggested by both Woodruff [1966] and Ericksen [1973], will be assigned the latter's name since Ericksen does more to extend the approach to the multivariate case. The variables P_{bj}, B_{bj} and D_{bj} are defined in similar fashion as before for $j=1,2,\ldots,n$, where n is the number of local areas in the sample. Using these data and the least squares procedure again, as in the Ratio-Correlation procedure, the estimates of the coefficients, denoted as $\hat{\hat{\alpha}}$, $\hat{\hat{\beta}}$ and $\hat{\hat{\gamma}}$, are found. The estimate for the i-th local area becomes

$$P_{bi}=\hat{\hat{\alpha}}+\hat{\hat{\beta}}B_{bi}+\hat{\hat{\gamma}}D_{bi} \qquad (1.2.29)$$

and finally

$$P_{3i}=\frac{P_{bi}P_{2i}P_3}{P_2} \; ; \; i=1,2,\ldots,N \quad . \qquad (1.2.30)$$

Thus this approach eliminates a need for the third of the three considerations suggested at the conclusion of the discussion on the Ratio-Correlation method. However, sample survey data (most likely of the multistage variety) are used which implies the necessity that the value of P_{3j} be an estimate. This introduces an additional source of variation to the estimate--a within-local-area component. This proves to be a major consideration in limiting the optimal number of symptomatic variables used and the number of strata when such a modification is introduced. Further discussion of this method, in somewhat modified form, is taken up in Chapter III.

1.3 Historical background

The earliest local postcensal population estimates were formed using extrapolation methods. This approach was used by the Bureau of the Census until 1926. These methods worked well since most local population trends continued smoothly except in some urban centers with large numbers of immigrants. These situations first pointed out deficiencies and gave cause for experimentation on other types of methods.

Already early in this century the Bureau of the Census and others began experimenting with so-called "symptomatic" techniques. The Bureau of the Census had considered Simple Censal Ratio methods, which are similar

in principle to the Vital Rates method, using public utility comsumption data, school enrollment, etc., and Snow [1911] conceived of the idea of a Ratio-Correlation method. The failings in the study of these methods were twofold. First, allowances for geographical and year-to-year variation in the symptomatic ratios were not made, and second, the symptomatic information during that period was of questionable reliability. This led the Bureau of the Census to conclude that the mathematical techniques were better and the symptomatic methods were not used.

The first large-scale formal use of symptomatic information was in the Bureau's 1926 estimates for states and cities. An "apportionment formula" (as in the Proration method) was adopted using birth and death information in areas where this information was thought to be good. This method was used until 1934 when a form of the Component method was developed. The natural increase and net migration estimates for this method were made separately, the first from birth and death data and the second from school enrollment data. Following the experience of the traumatic depression years, which emphasized the failings of the projection-type approach, the Bureau discontinued using the arithmetic extrapolation procedures. In spite of this, recent surveys of methodology used by local government offices (U.S. Bureau of the Census [1956, 1962, 1966a]) show that the Extrapolation methods are still used.

In 1937 it was decided that states estimates would be discontinued until the 1940 census. However, with the approach of World War II the planned adoption of a modified component procedure was postponed. Estimates during the war period were made using registration data for war ration books, and by most standards were quite good (Bureau of the Census [1943]). Unfortunately, this historical diversion stymied further experimentation with other Symptomatic methods. Soon after the war Eldridge [1947] suggested two modifications of the component method now known as Census Component Methods I and II. The present working form of Method II (U.S. Bureau of the Census [1966b]) is a modified version of that originally proposed. These procedures were adopted for official use in the 1950's although new symptomatic techniques were suggested during that period. Composite methods were introduced by Frisen [1951] and Bogue and Duncan [1959], and the Ratio-Correlation method was reintroduced by Schmitt and Crosetti [1954] and Goldberg et al. [1961]. The Housing Unit method first appeared in 1957 (U.S. Bureau of the Census [1957]). These new techniques and the two census component procedures were among those most commonly used in the 1960's (U.S. Bureau of the Census [1962, 1966a]). The most recent addition to the list of methods is the Ericksen method [1971] which is similar to a method proposed by Woodruff [1966].

Thus the field has evolved to the stage where

several viable alternative methods exist, and where the propriety of each method depends on the circumstances at hand. Furthermore, the field is characterized by techniques which are used primarily for making size estimates. These points are both emphasized more fully in the next sections.

1.4 Comparison of existing methods

In this section we intend to summarize the results of recent tests of accuracy and comparisons made on the existing methods. A list, which indicates the research done in this area, is found in Table 1.1. Several general conclusions may be drawn from these results. They are as follows:

1. Symptomatic methods are clearly superior to extrapolation methods for estimates on all levels. In every case where they are considered (U.S. Bureau of the Census [1955]; Zitter and Shryock [1964]; Schmitt and Crosetti [1954]) Extrapolation methods are less accurate than even the worst Symptomatic method considered.
2. Large gains in accuracy may be obtained by averaging methods, provided the methods use symptomatic data that are uncorrelated. This conclusion has been suggested by Goldberg and Balakrishnan [1961] although improvements have been noted in instances when the results have high positive correlation (U.S. Bureau

TABLE 1.1

A COMPENDIUM OF RESEARCH DEALING WITH METHODS FOR OBTAINING POSTCENSAL ESTIMATES

Order	Investigator	1	2	3	4	5	6	7	8	9	Average of methods	Units on which estimates made	Reference year
2	Greenberg (U.S. Bureau of the Census)	X	X	X		X					3 and 5 2 and 3	States, metropolitan counties and cities in 1950	[1955]
8	Zitter and Shryock	X	X	X		X	X		X		3 and 5 3 and 8 6 and 8	States in 1960	[1964]
3	Shryock			X		X						Cities and SMSA's in 1950	[1957]
1	Schmitt and Crosetti	X			X	X			X			Washington counties in 1950	[1954]
4	Goldberg and Balakrishnan			X		X	X		X			Michigan counties in 1958	[1961]
9	Zitter and Shryock			X		X					3 and 5	Metropolitan counties and cities in 1960	[1964]

1. Extrapolation
2. Component Method I
3. Component Method II
4. Proration
5. Vital Rates
6. Bogue-Duncan Composite
7. Housing Unit
8. Ratio-Correlation
9. Ericksen

TABLE 1.1--Continued

Order	Investigator	1	2	3	4	5	6	7	8	9	Average of methods	Units on which estimates made	Reference year
11	Zitter et al.			X		X	X		X		3 and 5 3 and 6 3 and 8 3, 5 and 8	United States counties in 1960	[1968]
12	Zitter et al.			X		X	X		X		3 and 5 3 and 6	Selected states in 1960	[1968]
5	National Vital Statistics Division			X		X						Montana, Oregon and Pennsylvania counties in 1960	[1962]
6	National Vital Statistics Division			X		X	X					Ohio counties in 1960	[1962]
7	Hillery			X								Kentucky counties in 1960	[1962]
10	Goldberg et al.			X		X	X		X			Michigan counties in 1960	[1964]

1. Extrapolation
2. Component Method I
3. Component Method II
4. Proration
5. Vital Rates
6. Bogue-Duncan Composite
7. Housing Unit
8. Ratio-Correlation
9. Ericksen

TABLE 1.1--Continued

Order	Investigator	1	2	3	4	5	6	7	8	9	Average of methods	Units on which estimates made	Reference year
13	Starsinsic and Zitter							X				Cities over 50,000 which had special censuses in 1960's	[1968]
14	Ericksen			X		X	X			X		Seventy-five counties with special censuses between 1964-1967	[1971]

1. Extrapolation
2. Component Method I
3. Component Method II
4. Proration
5. Vital Rates
6. Bogue-Duncan Composite
7. Housing Unit
8. Ratio-Correlation
9. Ericksen

of the Census [1955]). This conclusion makes intuitive sense if we note that

$$\mathrm{Var}\left(\frac{P_3+P_3'}{2}\right)$$

$$= \frac{1}{4}[\mathrm{Var}(P_3)+\mathrm{Var}(P_3')+2\mathrm{Cov}(P_3,P_3')] \quad . \quad (1.4.1)$$

If $\mathrm{Var}(P_3)=\mathrm{Var}(P_3')=\sigma^2$ and $\mathrm{Cov}(P_3,P_3')=0$

$$\mathrm{Var}\left(\frac{P_3+P_3'}{2}\right)= \frac{\sigma^2}{2} \quad . \quad (1.4.2)$$

Thus the variance is halved by using the average of the two uncorrelated results. In fact, gains are achieved if

$$\mathrm{Cov}(P_3,P_3')<\sigma^2 \quad , \quad (1.4.3)$$

which will always be true in this case.

3. Most methods do not do well in areas in a continual dynamic state of population change. This conclusion is particularly noted by Zitter and Shryock [1964], Zitter et al. [1968], Shryock [1957], and Ericksen [1971]. The first authors found that most methods, particularly Census Method II, consistently underestimated the population size.
4. Estimates using the component methods become worse as the interval between t_2 and t_3 increases. This conclusion is substantiated in an evaluation of methods made by the Bureau of the Census [1969] and seems reasonable since the estimate of migration would conceivably be

less accurate as the length of time, during which it must be assessed, increases.

5. No universally superior symptomatic method emerges when each method is used individually to find the estimate (i.e., they are not averaged). Greenberg (U.S. Bureau of the Census [1955]) and Zitter and Shryock [1964] have found that Method II does best for estimates of states in both 1950 and 1960. At the same time, the latter have found that Method II does poorly on metropolitan counties and cities. Schmitt and Crosetti [1954], Goldberg and Balakrishian [1961], and Goldberg et al. [1964] have found that the Ratio-Correlation method does best on county estimates for 1950 and 1960. Ericksen [1971] found that the Ericksen method generally does better in slow-growing counties, and relatively better in 1967 than 1965 since the Current Population Survey (from which he obtained his sample data) increased the number of sample PSU's in 1967 from 357 to 449. This decreased the mean squared error of the estimate.
6. All methods generally do better on larger units. This has been found true in virtually all comparisons where it is considered. It means that the method will show better when estimating a large county than a small one. To an extent this is not the case with the Ericksen method which does better in slow-growing (and, in most

cases, smaller) counties.

7. The propriety of any method depends upon the population to be estimated. This again is a conclusion which is generally made and which becomes a rather natural extension of the above conclusions.

Several investigations have dealt specifically with testing the accuracy of a particular method. Hillery [1962] has tested Component Method II on Kentucky counties using a separate school age net migration conversion factor for each economic area of the state. He concludes that this modification makes little difference in the results. Starsinsic and Zitter [1968] have investigated the Housing Unit method on estimates for cities. They conclude that (1) the Housing Unit method tends to overestimate the true size, and (2) improved results are obtained when gas and electric utility data are used and when the average housing unit size is obtained by extrapolation from t_2 to t_3 as opposed to using the t_2 figures. Finally, Rosenberg [1968] has found that stratification can significantly improve the Ratio-Correlation estimates.

1.5 Motivation for proposed method

We obtain a better perspective of postcensal estimation and what is proposed by listing the advantages and disadvantages of existing methods, using the discussion of previous sections. As a procedural note, we eliminate

Extrapolation methods from further consideration due to their deficiencies under most present day circumstances. Furthermore, they are nonsymptomatic and therefore less meaningful within the context of the present discussion. We also classify the symptomatic methods into two categories--regression (Ratio-Correlation and Ericksen procedures) and nonregression (all others).

The major advantage of the symptomatic, nonregression procedures is that they do, either singly or in combination, give good results for total area population size. (In most methods, estimates on age or racial segments of an area are less reliable.) For this reason we make no large scale attempt to improve on them although we hope that the proposed method gives at least comparable results. Instead, we intend to concentrate on making estimates on variables other than total area population. The lack of flexibility of present methods is a major weakness and dictates a need for further development among postcensal estimation techniques for local areas. We believe that methods should and can be developed which are useful in assessing attitudinal, economic, and other characteristics of local populations without the necessity of special local surveys. In addition to inflexibility, another disadvantage of methods currently used is that they depend on a set of arbitrary and often unsubstantiated assumptions. For example, the Vital Rates technique assumes that birth

and death data are equally good auxiliary indicators for population size. This may not always be true. In addition, since there is often no way to test these assumptions, the researcher is left in a quandry as to which method or combination of methods is most appropriate under the circumstances.

Symptomatic regression procedures eliminate one disadvantage of the nonregression methods in that they are applicable in theory to variables other than population size. However, they do not negate the second disadvantage since they too are limited by important and possibly restrictive assumptions. The Ratio-Correlation method assumes not only a linear model but also that the coefficients in that model remain constant over a period of time and area. The Ericksen method eliminates the necessity for the time assumption, but the limitation of linearity or known functional form remains.

The difficulties associated with model-fitting point out other disadvantages of the regression approach. First, the situation in both methods corresponds to sampling from a multivariate distribution instead of the usual textbook assumption that the "dependent" variable is random and all other "independent" variables are fixed. Thus the symptomatic or "independent" variables are not fixed but random in the proposed model. This involves difficulties (which are generally ignored) in estimating

and thus measuring reliability (Kendall [1951]), particularly when a nonlinear model is the true one.

Second, a decision must be made as to which model is used when preliminary investigations suggest a moderate multiple correlation between the variable to be estimated and the symptomatic variables. Natural symptomatic variables with multiple correlation levels upwards of 0.95 are more rare than common in nature, and even at high levels there is no assurance of the estimate's accuracy. Schmitt and Crosetti [1954] report an average error rate of 7.4 per cent using the Ratio-Correlation method despite a multiple correlation level of 0.979. What decision can be made when the level is moderate, say 0.50 to 0.80? It is usually not known whether this indicates a linear model with low level of association or some nonlinear model, possibly with a high level of association. In this situation we might be willing to take our chances and find the estimate assuming linearity. It is doubtful even then that the estimate is very precise since a linear model under low levels does poorly at estimating from the fitted equation. Thus the estimation procedure becomes more a shot in the dark and perhaps even wasteful of some nonlinear model under high level of association where we assume a linear model which shows moderate or low levels. For example, as we shall see with one of the bivariate logistic distributions considered in Chapter III, $\rho = \frac{3\alpha_0}{\Pi^2}$,

where ρ is the common linear model correlation coefficient and α_0 is the true model level of association. This means that although preliminary investigations might show $\rho=0.25$ assuming a linear model, the true level of the model would be $\alpha_0 = \frac{\pi^2}{3}(0.25)=0.82$.

Third, supposing that the researcher does not limit himself to linear models, how much has he gained? The social-psychologist or economist would be extremely naive to assume that all functional models he encounters are linear and simple. Indeed more often just the opposite is true. In terms of the regression methods, once the researcher has resigned himself to nonlinear alternatives, he must then settle on a reasonable model from preliminary work or other sources. His list of candidates is lengthy since it must include several variations in both order and degree. Once he has removed the less important variables he may then spend a good bit of time attempting to find the most reasonable model for those remaining. Finally, after some trial and error testing, what are his chances of selecting the correct or even a reasonable model? In many instances one might suspect that they would not be very great, particularly when there are several correlated symptomatic variables of comparable importance. The process in this case would undoubtedly be time-consuming and difficult.

An easy but incorrect alternative might be to

reason that if the marginal distributions of the variables resemble a normal distribution, the regression is linear or close to it. The same logistic distribution mentioned above, with its similarities to the normal and often non-linear regression, provides an adequate counter-example to the above assertion. Linearizing transformations may solve some problems but, again, in situations with several symptomatic variables this is a time-consuming, hit or miss approach with unlikely success. To avoid the preliminary work the researcher might proceed assuming linearity and use all possible symptomatic variables. In the case of the Ericksen method, this tactic would be to his disadvantage since he would be carrying "dead weight" along which adds little to the explanatory level (of multiple correlation) in his model while increasing the mean squared error of his estimate (Ericksen [1971], Chapter 2).

One further disadvantage of the regression procedures is that the symptomatic variables must be interval type (i.e., quantifiable). Ideally, we would hope that other qualitative types of variables would be usable by the method. For example, geographic location is frequently important but is difficult to quantify.

1.6 Conclusion

We must conclude that the symptomatic regression

approach with its advantages, carries with it a substantial number of disadvantages. Of course one would be mistaken to completely discount the credibility of this approach, although some severely question its usage altogether. Ehrenberg [1963], for one, sees the situation as being one where an estimate for one set of data is obtained using estimates of coefficients which have been obtained from a different set of data. Thus he reasons that the estimate will always be incorrect and should not be made by this approach.

There appears to be a need for a procedure which is useful to estimate many types of variables without being limited by stringent, unrealistic, or unverifiable assumptions. In Chapter II a method is proposed which, as the others, utilizes auxiliary information but which involves neither model-building nor model-fitting. In this way assumptions, which may impede the usefulness of the method, are avoided. The most limiting assumption that is made is that good symptomatic variables exist and are available to implement the method. Although this assumption may appear to be innocuous, it cannot be taken lightly since a method is useless unless relevant data are obtainable. To diminish the chances that this becomes a major difficulty, we rely upon the growing interest in social indicators for populations as well as existing Federal and non-Federal sources of data.

The removal of the regression methods' limitations is of particular interest when considering the suggestion that the Bureau of the Census adopt a quinquennial census to replace the current decennial census. If a method(s) is found which gives reliable estimates for census variables, perhaps a concentrated survey-type census in the interim period would be adequate in view of the great additional cost and effort required by a five-year census. Our effort here will hopefully indicate one such method.

CHAPTER II

THE PROPOSED METHOD

2.1 Introduction

Having established the background and motivation, we now present a detailed description of the proposed method as well as some indication as to how we intend to proceed in our research.

Before continuing, an important distinction is made concerning the circumstances in which the proposed method is used. In the other symptomatic methods the smallest units considered are the local area units for which the estimates are made. For example, with the Ratio-Correlation method, the population size of a county is estimated by using the symptomatic attributes (e.g., births, deaths and school enrollment) of that county as a whole. By contrast we now propose to break up the local unit into constituent geographical units or what we call base units. Continuing with the above example, these units might be enumerator districts, townships or some other geographical subunit of a county. We use the attributes for these base units to find estimates for the base units from the area of interest. These estimates are then combined to obtain the final estimate. To make later comparisons meaningful,

we consider the other regression methods in a similar light.

To further clarify this distinction consider Figure 2.1.

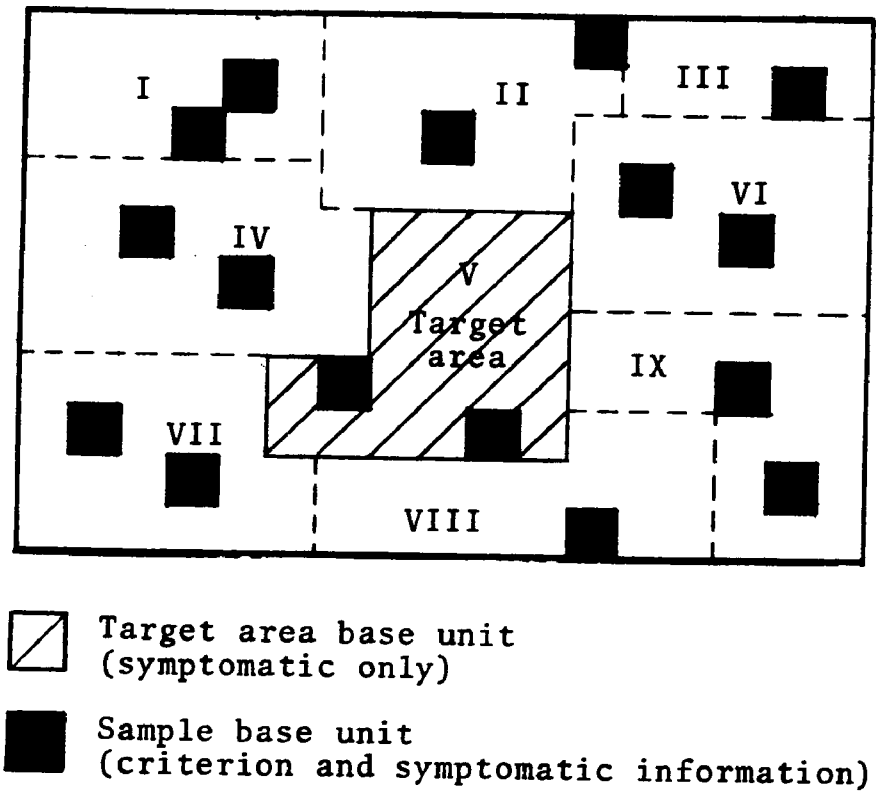


Fig. 2.1.--Diagrammatic representation of an estimation situation.

Let us allow it to be the geographical representation of a population. Suppose that this population is made up of nine local areas, designated as I through IX and that we decide to estimate some variable for area V. Area V is called the target area or the local area for which the estimate is made and the small, shaded squares in area V

are called target area base units or those base units in the target area. The point made above is that when finding the proposed estimate, we will use information from the base units which make up the areas. These base units are represented by the small squares in the figure. The meaning of the darkened squares is explained later.

As standard convention the variables (X) which provide the auxiliary information are called symptomatic variables.[1] The variable (Y), for which the estimate is to be made, is called the criterion variable.

As mentioned in Chapter I the major failing of other methods is that they are bound by rather limiting assumptions. The proposed method avoids these limitations and requires only that the symptomatic variables be "influential" on the criterion variable. This means that (1) no specific functional model (e.g., linearity) is assumed involving the criterion and symptomatic variables, (2) only a moderate level of correlation may be necessary in the nonlinear case since high association in a nonlinear model may not give high correlation, (3) estimates are possible for several types of criterion variables, and (4) estimates are possible for the base units as a byproduct of the procedure as long as relevant symptomatic information exists

---

[1]This may differ from standard usage in that symptomatic variables, as used here, include birth, death, school enrollment, tax, and other common forms of data as well as social indicators which are available for local areas.

and is available for those units. Preliminary comparison of this method with others in a moderate correlation situation has given encouraging results which are presented later (see Section 2.5).

The proposed method uses symptomatic information; however, a distinction is made as to how this symptomatic information is used. Most methods use this information directly to arrive at the estimate. The proposed method, on the other hand, uses the information to group base units which are members of a sample of base units from the total population. Referring again to Figure 2.1, the darkened blocks represent the base units in this sample, and are defined as sample base units. The symptomatic information is also used to classify target area base units into the appropriate group, which makes the technique similar in some respects to estimation from a contingency table.

2.2 Notation

Several basic symbols are used to denote various things which are brought up in this chapter. They are defined here and other conventions regarding subscripting are mentioned thereafter.

Y or y--Criterion variable (i.e., that which is to be estimated).

X or x--Symptomatic variable (i.e., that used as auxiliary information).

n--Number of units.

W--Weighting factor assigned to unit.

P--Probability that the unit is in the sample survey.

M--Measure of size.

An asterisk (*) superscripted indicates association with the target area data and no asterisk indicates association with the sample data. Referring to the criterion and symptomatic variables, $\bar{Y}$ and $\bar{X}$ refer to true means and $\bar{y}$ and $\bar{x}$ refer to estimates of the true means. We also designate s as the number of symptomatic variables and k as the number of groups formed. The subscript conventions adopted in Sections 2.3 and 2.4 are given below.

Notation for Section 2.3

The same subscript convention is adopted here as was used in Section 1.2. The subscript "a" refers to the interval between t_1 and t_2 and the subscript "b" refers to the interval between t_2 and t_3.

Notation for Section 2.4

The subscripts in Section 2.4 may be used one, two or three at a time.

g--The g-th group.

gh--The h-th base unit in the g-th group.

ghc--The c-th element in the gh-th base unit.

2.3 Grouping and classification procedures

Having dispensed with the necessary preliminaries,

we now describe the procedure followed to obtain the proposed estimate. Referring once more to Figure 2.1, the object of the method is to estimate some criterion variable in the target area. The target area base units we refer to are represented by all squares in area V and the sample base units are represented by all black squares. Furthermore, each white or shaded square is assumed to have symptomatic information only while the black squares have both symptomatic and criterion information.

Procedure

(1) Grouping--The sample base units (black squares) are broken down into k groups using either their symptomatic and criterion information or just their symptomatic information. The object of this phase is to form the groups in such a way that they are similar in some sense within groups and thus dissimilar between groups.

(2) Classification--Each target area base unit (all squares in V) is then assigned to one of the k groups. The object is to associate each target area base unit with a group of base units which are not only similar to the target area unit but among themselves also. Since we do not have criterion information in most of the target area base units, only symptomatic information can be used in this phase. Distinguishability among groups with respect to symptomatic information is quite obviously necessary here.

(3) Estimation--An estimate for each of the target area base units is obtained from the sample base units in the group to which it has been assigned. The final estimate for the target area is found by pooling the target area base unit estimates. A noteworthy sidelight of this procedure is that an estimate (and perhaps a good one if the objectives of the grouping phase are met) has been obtained for each of the target area base units. Here, similarity within groups with respect to the criterion variable is of benefit since maximally homogeneous groups produce the most reasonable estimates for the target area units and thus a more precise final estimate.

Before looking more closely at each phase, a few comments must be made regarding the definitions of multivariate data separation procedures, particularly "grouping" and "classification" as used above. We define grouping as the prior task of forming the groups from the sample observations. The problem of assigning as yet unlabeled individuals to one or another of the groups is called classification. Special care is taken with these definitions since terminology used to represent these procedures in the literature is at best haphazard. The result is a good deal of confusion since "grouping" is also called such things as clustering, classification or numerical taxonomy, and "classification" is often called discrimination, identification, allocation or assignment.

The goal of the grouping phase is to form distinct groups in s-space. They are generally defined in two ways. If the groups are formed by an algorithm which finds the optimum cut-points for each variable subject to some criterion, the result is groups defined by rectilinear boundaries. This is illustrated in Figure 2.2 where s=2. We define d_{uv} as the v-th demarcation point on X_u.

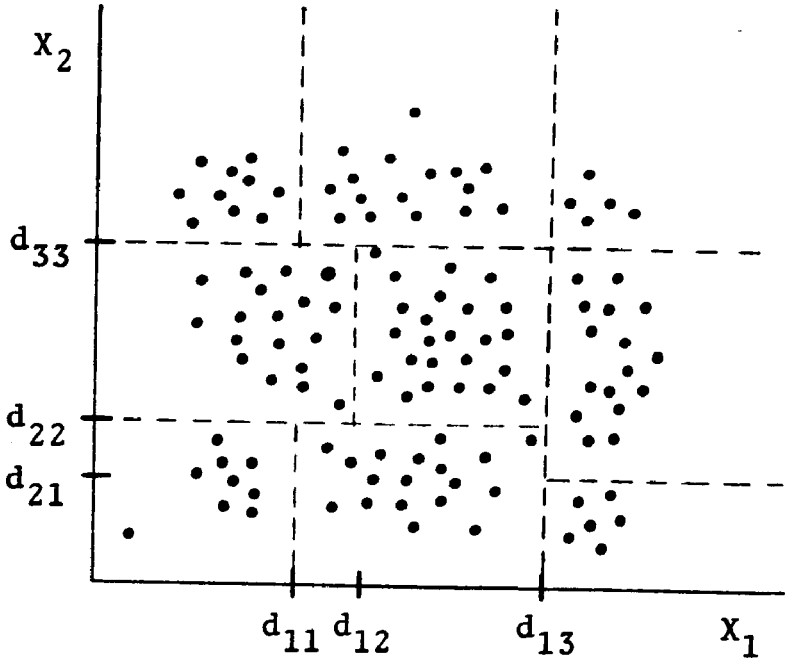


Fig. 2.2.--Diagrammatic representation of rectilinear classification.

The other definition of groups is implicit. If the algorithm calls for forming the groups such that the result is k lists of observations, the implicit boundaries might be thought of as the convex hull formed by the fringe points of each group. This is illustrated in Figure 2.3.

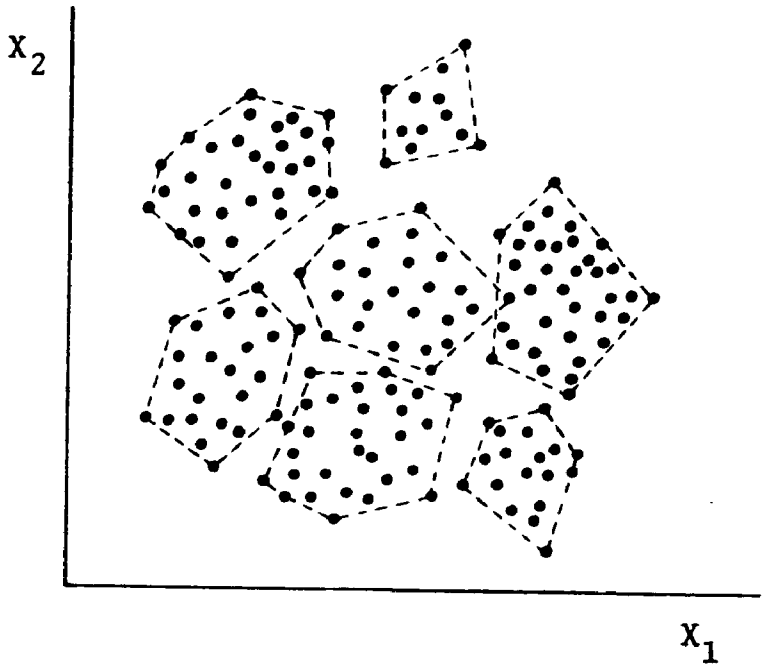


Fig. 2.3.--Diagrammatic representation of nonrectilinear classification.

The form that these boundaries take is important since it dictates which kind of classification can be used.

An appropriate question at this point is: When does one wish a grouping algorithm with output as in Figure 2.2 or as in Figure 2.3? Some light is shed on an answer when recalling that groups may be found using both symptomatic and criterion information or by using symptomatic information only. For example, when both are used, it may be disadvantageous to the classification procedure if the output is rectilinear as in Figure 2.2.

The growing interest in procedures useful for separation or summarization of data has led to a volumi-

nous bibliography of clustering or grouping procedures in the literature. Purposes for using these procedures are nearly as plentiful as the number of procedures. Unfortunately, the vast majority are of no use within the context of the proposed method since many methods deal only with clustering variables and also since most, which deal with clustering individuals, are equipped to handle only relatively small numbers.

Cormack [1971], in a recent review of grouping methods (or classification methods as he calls them), distinguishes three types of grouping algorithms as follows:

(1) Agglomerative--Based upon a series of successive fusions of the n observations.

(2) Clustering--Based upon a successive reallocation of entities between the sets of some initial partition.

(3) Divisive--Based upon the partitioning of the complete set of entities into finer partitions.

Because of the large numbers expected, we will find the divisive procedures most favorable due to their greater computing efficiency. However, some methods which employ modifications of these approaches are nearly as efficient. We postpone a detailed discussion of the procedures used until Chapter V where the major illustrative examples are presented.

Once we have formed groups which are optimally

similar in some respect, a classification rule is adopted to assign each target area observations to one of the groups. If they are formed as in Figure 2.2, a logical rule would be to "pigeon-hole" each target area base unit by simply assigning it to that group whose region includes its symptomatic values. This rule is certainly straightforward and simple to implement and might be expected to produce good results. If the groups have been formed as in Figure 2.3 the simple rule does not work very well since the regions defined in s-space are not exhaustive and perhaps not even mutually exclusive. Thus the rule breaks down when a target area observation lies outside the boundaries of all groups. This entails an extension of the rule to assign such an observation. One possibility is to use the minimum chi-square approach suggested by Tatsuoka [1971]. Although the assumption that the groups are samples of observations from separate populations would appear to disqualify this approach from consideration here, its only effect is to disallow computation of misclassification probabilities. The "distance" criterion it uses is intuitively attractive and useful. On the other hand, the rule based on the linear discriminant function first suggested by von Mises [1945] and discussed in Anderson [1958] is less useful. Under normality its rule is equivalent to the Tatsuoka rule, but the distance criterion of the latter rule is more reasonable

when normality cannot be assumed.

The minimum chi-square rule assumes that there are k groups with s-dimensional mean vectors, $\underset{\sim}{\mu}_g$, and common dispersion matrices, $\sum$. If the individual to be classified has observation vector $\underset{\sim}{X}=(X_1,X_2,\ldots,X_s)'$, where the prime refers to the transpose, the g-th deviation vector is defined as

$$\underset{\sim}{d}_g=\underset{\sim}{X}-\underset{\sim}{\mu}_g \; ; \; g=1,2,\ldots,k \quad . \tag{2.3.1}$$

The statistic used is

$$\chi_g^2=\underset{\sim}{d}_g'\sum^{-1}\underset{\sim}{d}_g \; ; \; g=1,2,\ldots,k \quad . \tag{2.3.2}$$

Since $\underset{\sim}{\mu}_g$ and $\sum$ are often unknown we substitute $\hat{\underset{\sim}{\mu}}_g$ and $\hat{\sum}$ where the latter is an estimate based on a pooling of the within sum of squares and cross products for all groups. The statistic becomes

$$\hat{\chi}_g^2=\underset{\sim}{e}_g'\hat{\sum}^{-1}\underset{\sim}{e}_g \tag{2.3.3}$$

where

$$\underset{\sim}{e}_g=\underset{\sim}{X}-\hat{\underset{\sim}{\mu}}_g \quad . \tag{2.3.4}$$

If it appears that the dispersion matrices are unequal and unknown, the modified statistic is

$$\hat{\hat{\chi}}_g^2=\underset{\sim}{e}_g'\hat{\sum}_g^{-1}\underset{\sim}{e}_g+\log|\hat{\sum}_g| \tag{2.3.5}$$

where $\hat{\sum}_g$ is the estimate for $\sum_g$, the g-th dispersion matrix and "| |" refers to the determinant. An additional modification is suggested when prior probabilities are known. That is

$$\hat{\hat{\chi}}_g^2 = \hat{\chi}_g^2 - 2 \log p_g \qquad (2.3.6)$$

where p_g is the probability that a population element is a member of the g-th group.

Truett et al. [1967] describe a procedure involving risk in a dichotomous situation which implies that, in the k-level polychotomous situation, the individual is assigned to the group, g, with maximum posterior probability

$$h(\Pi_g | \underset{\sim}{x}) = \frac{p_g f(\underset{\sim}{X} | \Pi_g)}{\sum_{q=1}^{k} p_q f(\underset{\sim}{X} | \Pi_q)} \qquad (2.3.7)$$

where Π_g refers to membership in group g and $\underset{\sim}{x}$ is an observed value of $\underset{\sim}{X}$. If $\underset{\sim}{X}$ is distributed as an s-variate normal, it can be shown that

$$h(\Pi_g | \underset{\sim}{x}) = \frac{\exp[-\hat{\hat{\chi}}_g^2/2]}{\sum_{q=1}^{k} \exp[-\hat{\hat{\chi}}_q^2/2]} \quad . \qquad (2.3.8)$$

The denominator of (2.3.8) is constant for all g and can be ignored in the classification rule. The numerator is a monotonic decreasing function of $\hat{\hat{\chi}}_g^2$, so that the modified minimum chi-square rule is thus equivalent to the maximum posterior probability rule under normality.

Having defined the "pigeon-hole" and minimum chi-square classification rules, it may be of interest to speculate when either of these rules is used. Clearly the answer depends upon which type of grouping procedure

is available. If the groups are formed so that their boundaries are (1) mutually exclusive and (2) exhaustive of the space, (e.g., a rectilinear algorithm), the natural choice is the "pigeon-hole" rule although one might conceivable use a more complicated rule. If either or both of conditions (1) and (2) do not hold, a rule like the minimum chi-square rule is a reasonable one.

As mentioned previously, there is also an option as to whether criterion and symptomatic or just symptomatic information is used. Intuitively, it is more reasonable to use both whenever possible, since we are more fully utilizing our information, but this may not always be possible. For example, if the only grouping procedure available is one which defines nonexhaustive groups, and we are wary of using the minimum chi-square rule, we might lump both sample and target area units together in the grouping phase. This allows us to skip the classification phase, but limits us to utilizing the symptomatic information only, since only symptomatic information is available in the target area.

Use of both criterion and symptomatic variables in the grouping phase necessitates an additional consideration in the classification phase. If the grouping procedure produces rectilinear-type groups defined on both criterion and symptomatic variables, we cannot simply ignore the criterion variable's demarcation points and

implement the "pigeon-hole" rule since demarcation points on the symptomatic variables might not be the same over the whole range of criterion values. Thus we prefer a grouping algorithm which defines the groups in terms of the symptomatic variables only. If the grouping procedure defines the groups as in Figure 2.3, we ignore the criterion value on each individual and the classification rule (e.g., minimum chi-square) uses the symptomatic information exclusively.

Both the Ericksen and Ratio-Correlation methods use time-ratio forms (i.e., ratios of values from one time to another) of the criterion and symptomatic variables to avoid the difficulty of inconsistencies in data collection, extreme skewness, etc. This obliges us to consider these forms also so that similar problems do not arise. Thus we consider variables measured for one point in time (e.g., X_3 or Y_3) and those measured as ratios (e.g., $X_b$or Y_b). The form of the variable used is largely determined by the availability and nature of the data.

This points to another consideration in the method--the source of data. We must assume that criterion information is available at t_3 in some sample of base units from the total population. There is little flexibility in this requirement. The symptomatic information is most often obtained from some noncensus source although in some instances census information may be used. The major

requirement of the symptomatic information is that it is available for both target area and sample base units. The several options open to us imply that the method may be carried out under many different situations. Some of the more probable possibilities are described in Table 2.1, their appropriateness depending upon which criterion and symptomatic variables are used. To avoid possible confusion it should be emphasized that in all situations described in Table 2.1 the proposed procedure is the same except for minor technical differences of implementation.

In situation 10 of Table 2.1, t_2 information is used in the grouping and classification phases. For this to be a reasonable course we assume that the results of the two phases are similar to those obtained if t_3 data are used. Regarding the grouping phase, it says that target area base units, which were similar based on past (t_2) data, are also similar based on current (t_3) data if it were available. This entails the requirement that the interval separating t_2 and t_3 be kept short and poses a potential disadvantage of the method.

The "NC" in Table 2.1 requires an explanation of what might be considered as reasonable noncensus sources. One is the vital registration system. This is the common source generally employed for making population estimates. Another source are the periodic special censuses taken by the Bureau of the Census on such things as business,

TABLE 2.1

SEVERAL POSSIBLE SITUATIONS DESCRIBING SOURCES OF DATA FOR THE PROPOSED METHOD

Situation	Source of data for symptomatic variables			Source of data for criterion variable			Form of symptomatic variable for grouping the sample units	Form of criterion variable for grouping the sample units	Form of symptomatic variable for classifying the target area
	t_1	t_2	t_3	t_1	t_2	t_3			
1	NC	NC	NC	--	C	S	X_b	Y_b	X_a^*
2	NC	NC	NC	--	C	S	X_b	--	X_a^*
3	--	NC	NC	--	C	S	X_3	Y_b	X_2^*
4	--	NC	NC	--	C	S	X_3	--	X_2^*
5	--	--	NC	--	--	S	X_3	Y_b	X_3^*
6	--	--	NC	--	--	S	X_3	--	X_3^*
7	--	NC	NC	--	C	S	X_b	Y_b	X_b^*
8	--	--	NC	--	--	S	X_3	--	X_3^*
9	--	NC	NC	--	C	S	X_b	--	X_b^*
10	--	C	--	--	--	S	X_2	--	X_2^*

C--Census NC--Noncensus
S--Sample survey "--"--Not used

housing, agriculture, manufactures and governments. There are also non-Federal sources of data. Local governments often collect and tabulate data which are common to all such governments such as local tax receipts, school millage levels, and crime statistics. Guidebooks to these sources of data (U.S. Bureau of the Census [1966c, 1970]) have been published by the Bureau of the Census and are available for public use. Also, social indicators such as a coefficient of racial mixture are or eventually will be available for local areas. Other more or less stable characteristics of local areas may be used such as mileage of paved roads or medical care capabilities measured in hospital beds available. Some data might be obtained from private sources such as local chapters of labor unions, civic organizations, or private clubs.

Any national survey whose geographical universe is the total population (i.e., usually the United States) is a useful source for the criterion information. This includes such national sample surveys as the Census Bureau's Current Population Survey and National Health Survey as well as private national surveys maintained by research installations such as the Institute for Social Research, Ann Arbor, Michigan. Of course, the type of sample survey used need not be limited to the national multistage household variety. For example, if one wishes to estimate some variables pertaining to physicians' prognostic recommenda-

tions on certain diseases for a local area, the results of a national sample of physicians, who have been asked these same questions, might be used. As another, if one wishes to estimate the learning ability of elementary school age children in a local area, the results from a national sample of elementary schools could be used.

From the above discussion we might now mention the advantages of the proposed method. First, estimates can be obtained for variables other than population size as is also the case with the Ericksen and Ratio-Correlation methods. Second, no regression model assumptions (e.g., linearity) are made and thus no model-building or model-fitting is required. Third, qualitative symptomatic variables can be used so long as the grouping algorithm permits these variables. Fourth, "byproduct" estimates of the criterion variable are found for all target area base units, which implies that these estimates can be obtained for the very small areas (e.g., Enumerator Districts) provided the necessary data are available.

2.4 Estimation procedure

We now devote a section to describe the estimation phase of the procedure. This phase will be treated as if the sample survey involved is of the multistage variety. As a reminder, the notation used in this section is described in Section 2.2.

Since the sample is multistage, it is assumed that the units designated as base units make up one of the stages in the sample design. This is necessary so that estimates for the base units can be found. The estimates for each base unit are (Kish [1967]; Section 2.8)

$$\bar{y}_{gh} = \sum_{c=1}^{n_{gh}} W_{ghc} y_{ghc} \qquad (2.4.1)$$

where

$$W_{ghc} = \left(P_{ghc} \sum_{c=1}^{n_{gh}} 1/P_{ghc} \right)^{-1} . \qquad (2.4.2)$$

Thus if within each group the P_{ghc} are equal, then

$$W_{ghc} = \frac{1}{n_{gh}} . \qquad (2.4.3)$$

The estimate for the g-th group becomes

$$\bar{y}_{g} = \sum_{h=1}^{n_{g}} W_{gh} \bar{y}_{gh} \qquad (2.4.4)$$

where, as one possibility,

$$W_{gh} = \left(P_{gh} \sum_{h=1}^{n_{g}} 1/P_{gh} \right)^{-1} . \qquad (2.4.5)$$

The final estimate is

$$\bar{y}^{*} = \sum_{g=1}^{k} W_{g}^{*} \bar{y}_{g} \qquad (2.4.6)$$

where, as one possibility, we let

$$W_{g}^{*} = M_{g}^{*} / \sum_{g=1}^{k} M_{g}^{*} . \qquad (2.4.7)$$

Changes in the type of criterion variable necessitate minor modifications of the procedure. If the criterion variable is an attribute particular to the base units (e.g., a census nonresponse level), Y_{ghc} becomes constant for all $c=1,2,\ldots,n_{gh}$ and $\bar{y}_{gh}\equiv Y_{ghc}$. If it is a count (e.g., number of influenza vaccinations), $W_{ghc}=P_{gh}/\sum_{c=1}^{n_{gh}} P_{ghc}$ and $W_{gh}=1$. If it is the proportion of individuals with some attribute, the criterion variable is defined as

$$Y_{ghc}=\begin{cases}1 \text{ if the unit has the attribute}\\ 0 \text{ otherwise}\end{cases} . \qquad (2.4.8)$$

Due to the theoretic implications of such things as the grouping and classifications procedures used as well as the many different situations regarding the source of the data, a general analytic expression for the mean squared error is not derived here. However, analytic results under less general circumstances are found in Chapter III. For the present, the relative gains can be thought of roughly in terms of the gains due to post-stratification since the grouping procedure (where the criterion information is used) is similar in principle. Assuming the sample base units to be a simple random sample, the variance of the sample mean is approximately (Kish [1967]; Section 3.4C)

$$\mathrm{Var}(\bar{y}_{p\text{-}s})=\frac{1-f}{n}\sum_{g=1}^{k} W_g S_g^2+\frac{1-f}{n}\sum_{g=1}^{k} W_g(1-W_g)S_g^2/n_g \qquad (2.4.9)$$

where $n=\sum_{g=1}^{k} n_g$, f is the sampling rate, S_g^2 is the variance of the criterion variable in group g, and the subscript, p-s, refers to post-stratification. The first component is the variance of a proportionate stratified sample. If the n_g are moderately large, the second term becomes small, and the gains over simple random sampling are roughly those of a proportionate stratified sample. The gains of proportionate to simple random sampling are approximately

$$\frac{\mathrm{Var}(\bar{y}_{SRS})-\mathrm{Var}(\bar{y}_{p-s})}{\mathrm{Var}(\bar{y}_{SRS})} = \frac{\sum_{g=1}^{k} W_g(\bar{y}_g-\bar{y})^2}{S^2} \qquad (2.4.10)$$

where S^2 and $\bar{y}$ are estimates for the population variance and mean of the criterion variable. The extension of this to the proposed method is that from this we speculate that the best results will be found when (1) the groups formed are maximally distinct among groups (i.e., maximally similar within groups) with respect to the criterion variable and (2) the n_g are large.

2.5 Trial run

As further illustration of the proposed method, we present and discuss the results of a trial run in which the median incomes of the state of Michigan and Michigan counties are estimated for 1960. The primary purpose is to give rough indication, under somewhat realistic circum-

stances, as to how results obtained by the proposed method compare with those obtained by a similar method, namely the Ericksen method. Although the Ratio-Correlation is similar also, its results are not presented. The reason is that in this situation the proposed and Ericksen methods are at an unfair advantage which is described below.

The data are obtained from 1950 and 1960 census volumes with counties designated as the base units. To establish a set of sample base units it was decided that the 145 counties in the United States national multistage sample of the Institute for Social Research, Ann Arbor, Michigan would be used. Therefore population and symptomatic information are obtained for the ISR counties and symptomatic information only is obtained for the 83 Michigan counties. Instead of using sample estimates of the criterion variable for the county, which would be the usual procedure, the "true" value recorded in the census returns used (i.e., $\bar{y}_{gh} \equiv Y_{ghc}$), thus the advantage for the two methods mentioned previously. In spite of this, the value of the comparison remains since both are at a similar advantage.

The object is to estimate the 1960 (t_3) median income for Michigan families using three symptomatic variables: (1) median school years completed by males 25 years of age and older, (2) per cent of the population defined as urban dwellers, and (3) per cent of females 14 years of

age and older who are members of the labor force. Grouping is done using the Automatic Interaction Detector (AID) algorithm, a rectilinear-type procedure which involves both criterion and symptomatic information and which forms groups such that the symptomatic variables are maximally predictive of the criterion variable. This procedure is described in greater detail in Chapter V.

In terms of Figure 2.1, the ISR counties correspond to the black squares and the Michigan counties correspond to the squares in V. Briefly, the proposed method consists of grouping the ISR counties into 10 groups of about 15 counties each. An estimate of the median income is found for each group and each Michigan county is classified into one of these 10 groups. The final Michigan estimate is found as described in Section 2.4. The measure of size used (M_g^*) is the total 1960 population of all counties in the g-th group. The classification procedure used is the minimum chi-square procedure described in Section 2.3 under two different assumptions, namely equal dispersion and unequal dispersion among groups. The base unit weighting factors, W_{gh}, assuming unequal dispersion, are assigned three different values

$$W_{gh}^{(1)} = \left[P_{gh} \sum_{h=1}^{n_g} 1/P_{gh} \right]^{-1} \qquad (2.5.1)$$

$$W_{gh}^{(2)} = 1/n_g \qquad (2.5.2)$$

$$W_{gh}^{(3)} = P_{gh} / \sum_{h=1}^{n_g} P_{gh} \qquad (2.5.2)$$

where P_{gh} is population size. When assuming equal dispersion only $W_{gh}=1/n_g$ is used.

The Ericksen procedure is done in two ways. In the first way, designated as Ericksen (a), a linear regression model is fitted by ordinary least squares using 1960 criterion and symptomatic data for 47 states (excluding Alaska, Hawaii and Michigan). Then using the estimated coefficients, estimates are found for each Michigan county based upon its 1960 symptomatic values. The Michigan estimate is found by a weighted estimate of the county estimates where a county's "weight" is the proportion of the 1960 Michigan population living in that county. In the second way, designated as Ericksen (b), 1960 census data for the ISR counties are used to fit the linear model. The Michigan estimate is found as in Ericksen (a).

To evaluate the estimates found for all Michigan counties by the two methods, the relative absolute deviation, defined as

$$\frac{|\text{Predicted-True}|}{\text{True}}$$

is found for all counties by the methods. The tabulated results are summarized in two ways, as weighted or unweighted. The weighted result is the weighted mean of the 83 relative absolute deviations where the weight is

the proportion of 1960 Michigan population. The unweighted result is found by a simple mean of the 83 relative absolute deviations.

The results are presented in Table 2.2. The limited scope of this example makes any hard and fast conclusions difficult, although several items are noteworthy. First, all forms of the proposed method do well with the Michigan estimate and are somewhat better than the Ericksen estimates. Second, the Ericksen method does well on the county estimates although the difference is not great. Third, the drop in mean relative absolute deviation by weighting supports the conjecture that the methods do better on the larger counties than smaller ones. The proposed method records a somewhat greater reduction than the other method; however, one might speculate that this is due to the disproportionately large number of urban counties in the ISR sample. Fourth, $W_{gh}^{(3)}$ produces slightly better results here although there are no great differences except with the Michigan estimate. Fifth, the choice between equal or unequal dispersion seems unimportant here. At the very least, we conclude that the proposed method appears to compare favorably with the Ericksen method under these circumstances.

The level of multiple correlation in this case is relatively high (0.80 to 0.87) and the results appear to be quite good for the proposed method. The case of even

TABLE 2.2

RESULTS OF A TRIAL RUN INVOLVING 1960 MEDIAN INCOME IN MICHIGAN

		Relative absolute deviation for 1960 Michigan estimate of median income	Mean relative absolute deviation for 83 Michigan county estimates of median income	
			Unweighted	Weighted
Ericksen (a) (Mult. R=0.87)		0.0966	0.2135	0.1127
Ericksen (b) (Mult. R=0.80)		0.0558	0.0930	0.0787
Proposed method	$W_{gh}^{(1)}$	0.0325	0.1479	0.0643
Assuming	$W_{gh}^{(2)}$	0.0161	0.1381	0.0542
Unequal dispersion	$W_{gh}^{(3)}$	0.0006	0.1300	0.0592
Proposed method assuming equal dispersion	$W_{gh}^{(2)}$	0.0242	0.1483	0.0662

lower multiple correlation levels, considered in Chapters III and IV, may provide us with more interesting results, however.

2.6 Research itinerary

Briefly, it is our intention to study the proposed method in three ways. The first involves an analytic formulation of the estimation technique under certain prespecified conditions. Here estimates of the properties of the proposed method are obtained and compared with the Ericksen procedure. This is done in Chapter III. The second way involves a simulation of the proposed method under more general and relevant conditions and is done in Chapter IV. Finally, the proposed and Ericksen methods are used to obtain estimates on some real data, which are provided by the Bureau of the Census. This is considered in Chapter V. It is hoped that this three-pronged approach will allow us sufficient insight as to the relative feasibility and accuracy of the technique which is being proposed.

CHAPTER III

AN ANALYTIC FORMULATION

3.1 Introduction

Having supplied a brief description of the proposed method, we now wish to develop it analytically and then compare the results with other methods. These comparisons are made under several circumstances which are discussed later.

The work in this chapter is done assuming that (1) there is one criterion variable and one symptomatic variable, (2) grouping is done using symptomatic information only, (3) the grouping algorithm utilizes order statistics on observations obtained in a simple random sample from an infinite population, (4) classification is done by placing the unit into the group in which the value of the symptomatic variable falls (i.e., the "pigeon-hole" procedure described in Section 2.3), and (5) the criterion and symptomatic variables have certain bivariate distributions.

It may appear that these assumptions excessively limit the scope of the comparisons to be made. However, the general principles are retained so that comparisons, although narrow assumptionwise, are valid and of interest.

Also, an implicit generality exists since a new, single variable can be expressed as a linear combination of several symptomatic variables. The single symptomatic variable might be considered as such a linear combination. These particular assumptions also serve to make the algebraic manipulation less complicated and the computing formulas slightly more bearable. In Chapter IV some of these assumptions are lifted in favor of more relevant ones. However, there, comparisons are made using Monte Carlo methods in place of the more direct approach used here.

3.2 Preliminaries and notation

Before proceeding we expand somewhat on the assumptions made above. We then present a general description of the estimate as well as formulas for the expectation, bias, variance and mean squared error (MSE) of the proposed method. Next similar formulas are derived for the Ericksen method with which the proposed method is compared. Finally, the specific expectation, bias, variance and MSE are found for both methods and compared, assuming four different bivariate distributions: the normal, two logistics, and an exponential.

It should be noted that comparison is made with the Ericksen method alone. There are several reasons for this. First, most other methods are used exclusively to

estimate population size and would not be of general interest. Second, the Ericksen procedure is similar both in principle and implementation to the proposed method thereby simplifying the comparisons made. Third, it was felt that comparison with the Ratio-Correlation method is less meaningful since, under its assumptions, the major source of error is the change in regression relationship from intercensal to postcensal period, thus making it less comparable with the other two methods. This is not to suggest that such comparisons are not of future interest. Work in this direction may prove to be an interesting extension of the present work.

The notation used here follows the same pattern established in Chapter II with a few additions. The asterisk (*) continues to refer to the target area and $\bar{y}^*$ and $\bar{x}^*$ refer to the simple mean of the n^* criterion and symptomatic observations, respectively. The subscripts "P" and "E" refer to the proposed and Ericksen estimates, respectively. Also, the subscript "g" refers to a group and "k" to the number of groups formed. Finally, to facilitate the use of summation signs we define

$$\sum \equiv \sum_{g=1}^{k} \sum_{t=1}^{n_g} \tag{3.2.1}$$

and

$$\sum\sum' \equiv \sum_{g=1}^{k} \sum_{t=1}^{n_g} \sum_{g'=1}^{k} \sum_{t'=1}^{n_{g'}} , \tag{3.2.2}$$

$$s_g+t \neq s_{g'}+t'$$

and $\sum\sum'_{all}$ indicates summation over all s_g+t and $s_{g'}+t'$. The value n_g refers to the number of sample observations in group g, s_g is defined later as

$$s_g = \sum_{f=1}^{g-1} n_f \quad , \tag{3.2.3}$$

and the primes (') serve only to differentiate subscripts.

3.3 General description

Let us assume that we have a simple random sample

$$(x_1,y_1),(x_2,y_2),\ldots,(x_n,y_n)$$

drawn from an infinite bivariate population (thus we ignore finite population corrections) with joint cumulative distribution function (c.d.f.), F(x,y), marginal c.d.f.'s, $F_1(x)$ and $F_2(y)$, means, μ_x and μ_y, and variances, σ_x^2 and σ_y^2. We then order the symptomatic observations

$$x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$$

with corresponding criterion observations

$$y_{[1]},y_{[2]},\ldots,y_{[n]}$$

which may not be in ascending order. We also have known symptomatic observations from the target area

$$x_1^*,x_2^*,\ldots,x_{n^*}^*$$

or ordered as

$$x_{(1)}^* \leq x_{(2)}^* \leq \cdots \leq x_{(n^*)}^* \quad .$$

The purpose of the grouping procedure is to take

the sample observations and form a prespecified number of groups (k) which are similar with respect to the single symptomatic variable. This yields a set of boundaries $b_0<b_1<b_2<\ldots<b_k$. In most instances we may designate $b_0=-\infty$ and $b_k=+\infty$. The problem then is to assign values to $b_1,b_2,\ldots,b_{k-1}$. One solution is to assign those values which seem reasonable based upon one's prior knowledge of the behavior of the symptomatic variable in the population. This of course ignores the sample information and is not considered here. However, it does point to the possibility of a Bayesian approach utilizing both one's prior notions as well as the sample information. This course is not pursued any further here.

Another solution is to use the values of certain order statistics on x as the boundaries. This has some intuitive appeal since observations in close proximity are also consecutive order statistics, at least under the circumstances where there is a single symptomatic variable. The problem is to decide the group to which each order statistics most appropriately belongs.

In the simplest case where one wishes groups of equal size and n/k=e is an integer, we might let

$$b_g=x_{(ge)} \quad ; \quad g=1,2,\ldots,k-1 \quad . \tag{3.3.1}$$

In addition, we might define

$$b_g=x_{(s_{g+1})} \tag{3.3.2}$$

where

$$s_g=\{j:\min_j|\tfrac{j}{n+1}-F_1(B_{g-1})|\}\ ,\ g=2,3,\ldots,k \quad , \qquad (3.3.3)$$

and $B_0,B_1,\ldots,B_k$ are the "optimal" points of stratification on X with k strata (Note: $B_0=-\infty$ and $B_k=+\infty$). Defining s_g by (3.3.3) is reasonable since the expectation of the j-th order statistics from a standard uniform distribution (i.e., for $F_1(x)$ here) is j/n+1. The method used to determine these points is discussed in Appendix C. We also define

$$n_g=s_{g+1}-s_g \qquad (3.3.4)$$

which implies that

$$s_g=\sum_{f=1}^{g-1} n_f\ ;\ g=2,3,\ldots,k+1 \quad , \qquad (3.3.5)$$

and that the order statistics $x_{(s_g+1)},x_{(s_g+2)},\ldots,x_{(s_g+n_g)}$ occupy the g-th group or interval if we let $s_1=0$ and $s_{k+1}=n$.

If $F_1(x)$ is unknown except by speculation, the groups formed will most likely not be optimal in the sense that the groups are maximally homogeneous within and heterogeneous between. But within the context of an assumed distribution for x, defining boundaries by order statistics is quite satisfactory and will be our approach.

The form of the proposed estimate we shall use is a modified version of that suggested in Chapter II. The specific analytic context now allows us to replace the weights (see Section 2.4) by a function of what we define as

$$\Pi_{gj}=Pr(x_{(s_g)}\leq x_{(j)}{}^{*}\leq x_{(s_{g+1})}) \tag{3.3.6}$$

which is the probability that the interval between the s_g-th and s_{g+1}-th order statistics covers $x_{(j)}{}^{*}$. Our estimate then becomes

$$\bar{y}_p{}^{*}=\sum\delta_g y_{[s_g+t]} \tag{3.3.7}$$

where

$$\delta_g=\sum_{j=1}^{n^*}\frac{\Pi_{gj}}{n^* n_g} \quad . \tag{3.3.8}$$

This corresponds to

$$W_{gj}=1/n_g \tag{3.3.9}$$

and

$$W_g=\frac{\sum_{j=1}^{n^*}\Pi_{gj}}{n^*} \tag{3.3.10}$$

in terms of the terminology of Section 2.4. The expectation and variance of $\bar{y}_p{}^{*}$ follow as

$$E(\bar{y}_p{}^{*})=\sum\delta_g E(y_{[s_g+t]}) \tag{3.3.11}$$

and

$$\begin{aligned}Var(\bar{y}_p{}^{*})=&\sum\delta_g{}^{2}Var(y_{[s_g+t]})\\&+\sum\sum{}'\delta_g\delta_{g'}Cov(y_{[s_g+t]},y_{[s_{g'}+t']}) \quad .\end{aligned} \tag{3.3.12}$$

From Hansen, Hurowitz and Madow, Vol. II [1953], p. 66 we note that

$$E(y_{[j]})=E[E(y_{[j]}|x_{(j)})] \quad , \tag{3.3.13}$$

$$\mathrm{Var}(y_{[j]})=E[\mathrm{Var}(y_{[j]}|x_{(j)})] + \mathrm{Var}[E(y_{[j]}|x_{(j)})] \quad , \tag{3.3.14}$$

$$\mathrm{Cov}(y_{[j]},y_{[j']})=E[\mathrm{Cov}(y_{[j]},y_{[j']}|x_{(j)},x_{(j')})] + \mathrm{Cov}[E(y_{[j]}|x_{(j)},x_{(j')}),E(y_{[j']}|x_{(j)},x_{(j')})] \quad . \tag{3.3.15}$$

We can simplify (3.3.15) by the following theorem.

Theorem 1.--Consider a sample of size n from a (nonsingular) bivariate continuous distribution with density f(x,y), non-degenerate marginals, $f_1(x)$ and $f_2(y)$, and conditional density g(y|x). Ordering the x's yields the order statistics

$$x_{(1)}\leq x_{(2)}\leq\cdots\leq x_{(n)}$$

with corresponding

$$y_{[1]},y_{[2]},\ldots,y_{[n]} \quad .$$

The conditional covariance

$$\mathrm{Cov}[y_{[i]},y_{[j]}|x_{(i)},x_{(j)}]=0 \ , \ i\neq j \quad . \tag{3.3.16}$$

Proof.--We can represent the surface, z=f(x,y), above two-dimensional real space, R_2, by contiguous corridors defined in both directions on the real lines of x and y. Let $\nu_1^+,\nu_2^+,\nu_3^+,\ldots$, be an infinite set of such corridors where ν_i^+ is defined as the volume bounded by the surface and the planes $x=(i-1)\delta x$, $x=i\delta_x$ and z=0. Similarly, ν_i^- from $\nu_1^-,\nu_2^-,\nu_3^-,\ldots$, is defined as the volume bounded by the surface and the planes $x=(1-i)\delta_x$, $x=-i\delta_x$ and z=0. On the real line of y we may define $\zeta_1^+,\zeta_2^+,\zeta_3^+,\ldots$, and

$\zeta_1^-, \zeta_2^-, \zeta_2^-, \ldots$, where ζ_j^+ is bounded by the surface, $y=(j-1)\delta_y$, $y=j\delta_y$, $z=0$ and ζ_j^- is bounded by the surface, $y=(1-j)\delta_y$, $y=-j\delta_y$ and $z=0$. Finally, we let $\phi_{ij}(I_i, I_j)$ represent the intersection of a corridor on x and a corridor on y such that

$$I_i \equiv \begin{cases} + \text{ if } \nu_i^+ \\ - \text{ if } \nu_i^- \end{cases} ; \quad I_j \equiv \begin{cases} + \text{ if } \zeta_j^+ \\ - \text{ if } \zeta_j^- \end{cases} \tag{3.3.17}$$

which is bounded by the surface, $\nu_i^{I_i}$, $\zeta_j^{I_j}$, and $z=0$ and where δ_x and δ_y are arbitrarily small.

Since $x_{(i)}$ and $x_{(j)}$ represent no more than designated values of x along R_1 the problem reduces to showing that

$$\mathrm{Cov}[y_1, y_2 | x_1, x_2] = 0 \tag{3.3.18}$$

where x_1 and x_2 are known and y_1 and y_2 are the corresponding y values, respectively.

The mechanism for selection of a sample from a bivariate distribution can be thought of as a twofold procedure: (1) Select the x observation according to $f_1(x)$. (2) Select the y observation according to $g(y|x)$ where x is taken from (1).

The problem suggests that (1) has been completed leaving the y's according to x_1 and x_2. Knowing x_1 and x_2 involves two corridors (possibly one), say ν_u^+ and ν_v^+, which implies that the corresponding values of y are selected with probability proportional to the volume of

the slice of the corridor (e.g., ϕ_{uw} or ϕ_{vz}) in which it falls. Once the value of y_1 is found, say in ϕ_{uw}, it is clear that by this mechanism the probability that y_2 will be selected from ϕ_{vz} remains unchanged for any y_1. Thus y_1 and y_2 are uncorrelated conditioned on x_1 and x_2 and thus

$$\mathrm{Cov}[y_1, y_2 | x_1, x_2] = 0 \qquad (3.3.19)$$

which implies that

$$\mathrm{Cov}[y_{[i]}, y_{[j]} | x_{(i)}, x_{(j)}] = 0 \quad . \qquad (3.3.20)$$

Furthermore, the nature of the sampling method implies that

$$E(y_{[j]} | x_{(j)}, x_{(j')}) = E(y_{[j]} | x_{(j)}) \quad . \qquad (3.3.21)$$

As a result of Theorem 1, (3.3.15) reduces to

$$\mathrm{Cov}(y_{[j]}, y_{[j']}) = \mathrm{Cov}[E(y_{[j]} | x_{(j)}), E(y_{[j']} | x_{(j')}] \quad . \qquad (3.3.22)$$

We may now substitute the results from (3.3.13), (3.3.14) and (3.3.22) to obtain the results for (3.3.11) and (3.3.12). The distribution-specific results are dealt with later.

To obtain the bias we first consider the model

$$y = E(y|x) + \varepsilon \qquad (3.3.23)$$

where the ε's are uncorrelated error components each of which has a distribution with zero mean and constant variance. We find an estimate, $\bar{y}_o^*$, of the target area mean

from (3.3.23) if we assume that the simple mean of the n^* target area values of ε is zero. The bias is approximated by

$$\text{Bias}(\bar{y}_p^*)=E(\bar{y}_p^*)-\bar{y}_o^* \quad . \tag{3.3.24}$$

Finally, the mean squared error is found from

$$\text{MSE}(\bar{y}_p^*)=\text{Var}(\bar{y}_p^*)+\text{Bias}^2(\bar{y}_p^*) \quad . \tag{3.3.25}$$

To compute Π_{gj} we note that since $s_g<s_{g+1}$,

$$\begin{aligned}&\Pr(x_{(s_g)}\leq x_{(j)}^*\leq x_{(s_{g+1})})\\ &=\Pr(x_{(s_g)}\leq x_{(j)}^*)-\Pr(x_{(s_{g+1})}\leq x_{(j)}^*) \quad .\end{aligned} \tag{3.3.26}$$

Also since, in general,

$$\Pr(x_{(i)}\leq x_{(j)}^*)=\sum_{v=i}^{n}\binom{n}{v}F_j^{\,v}(1-F_j)^{n-v} \quad , \tag{3.3.27}$$

where $F_j\equiv F_1(x_{(j)}^*)$, (3.3.26) becomes

$$\Pi_{gj}=\sum_{v=s_g}^{s_{g+1}-1}\binom{n}{v}F_j^{\,v}(1-F_j)^{n-v} \quad , \; g=1,2,\ldots,k-1 \quad , \tag{3.3.28}$$

and

$$\Pi_{kj}=\sum_{v=s_k}^{s_{k+1}}\binom{n}{v}F_j^{\,v}(1-F_j)^{n-v} \tag{3.3.29}$$

where we let $s_1=0$. When it is feasible we use the normal approximation to the binomial so that

$$\begin{aligned}\Pi_{gj}&\doteq\Phi\left(\frac{s_{g+1}-nF_j-1/2}{\sqrt{nF_j(1-F_j)}}\right)\\ &\quad-\Phi\left(\frac{s_g-nF_j-1/2}{\sqrt{nF_j(1-F_j)}}\right) \; ; \; g=2,3,\ldots,k-1 \quad ,\end{aligned} \tag{3.3.30}$$

and

$$\Pi_{1j} \doteq \Phi\left(\frac{s_2 - nF_j - 1/2}{\sqrt{nF_j(1-F_j)}}\right) \tag{3.3.31}$$

and

$$\Pi_{kj} \doteq 1 - \Phi\left(\frac{s_k - nF_j - 1/2}{\sqrt{nF_j(1-F_j)}}\right) \quad . \tag{3.3.32}$$

Furthermore, when the distribution on x is normal, an approximation for F_j is required. A discussion of this appears in Appendix B.

To obtain the estimate by Ericksen's method we assume a special linear form of (3.3.23), namely

$$y = \alpha + \beta x + \varepsilon \quad . \tag{3.3.33}$$

The object is to estimate α and β from the sample information. Recalling that $x_1^*, x_2^*, \ldots, x_{n^*}^*$ are the symptomatic observations in the target area, our estimate is

$$\bar{y}_E^* = \frac{1}{n^*} \sum_{j=1}^{n^*} (\hat{\alpha} + \hat{\beta} x_j^*) = \hat{\alpha} + \hat{\beta}\bar{x}^* \tag{3.3.34}$$

where

$$\bar{x}^* = \frac{1}{n^*} \sum_{j=1}^{n^*} x_j^* \quad . \tag{3.3.35}$$

The ordinary least squares estimates of the coefficients are

$$\hat{\beta} = \frac{\sum_{i=1}^{n} (x_i - \bar{x}) y_i}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \tag{3.3.36}$$

and

$$\hat{\alpha}=\bar{y}-\hat{\beta}\bar{x} \tag{3.3.37}$$

where

$$\bar{y}=\frac{1}{n}\sum_{i=1}^{n} y_i \tag{3.3.38}$$

and

$$\bar{x}=\frac{1}{n}\sum_{i=1}^{n} x_i \quad . \tag{3.3.39}$$

We now obtain the expectation and variance of the estimate as

$$E(\bar{y}_E^*)=E(\hat{\alpha})+x^*E(\hat{\beta}) \tag{3.3.40}$$

and

$$Var(\bar{y}_E^*)=Var(\hat{\alpha})+\bar{x}^{*2}Var(\hat{\beta})+2\bar{x}^*Cov(\hat{\alpha},\hat{\beta}) \quad . \tag{3.3.41}$$

From (3.3.13)

$$E(\hat{\alpha})=E[E(\hat{\alpha}|\underset{\sim}{x})] \quad , \tag{3.3.42}$$

$$E(\hat{\beta})=E[E(\hat{\beta}|\underset{\sim}{x})] \quad , \tag{3.3.43}$$

$$Var(\hat{\alpha})=E[Var(\hat{\alpha}|\underset{\sim}{x})]+Var[E(\hat{\alpha}|\underset{\sim}{x})] \quad , \tag{3.3.44}$$

$$Var(\hat{\beta})=E[Var(\hat{\beta}|\underset{\sim}{x})]+Var[E(\hat{\beta}|\underset{\sim}{x})] \quad , \tag{3.3.45}$$

$$Cov(\hat{\alpha},\hat{\beta})=E[Cov(\hat{\alpha},\hat{\beta}|\underset{\sim}{x})]+Cov[E(\hat{\alpha}|\underset{\sim}{x}),E(\hat{\beta}|\underset{\sim}{x})] \quad , \tag{3.3.46}$$

where $\underset{\sim}{x}=(x_1,x_2,\ldots,x_n)$ represents the symptomatic sample information. The values in (3.3.33) and (3.3.34) are found using (3.3.35)-(3.3.39). Again the distribution-specific results are deferred until later. As before we use (3.3.23) to obtain $\bar{y}_o^*$ which yields

$$Bias(\bar{y}_E^*)=E(\bar{y}_E^*)-\bar{y}_o^* \tag{3.3.47}$$

and

$$MSE(\bar{y}_E^*)=Var(\bar{y}_E^*)+Bias^2(\bar{y}_E^*) \quad . \qquad (3.3.48)$$

Finally, we mention a theorem which will be useful to obtain approximations related to the properties of the Ericksen method. It is found in Cramér [1963] but is attributed to Slutsky [1925].

Theorem 2.--If $\theta_n,\tau_n,\ldots,\xi_n$ are random variables converging in probability to the constants u,v,...,w, respectively, any rational function $R(\theta_n,\tau_n,\ldots,\xi_n)$ converges in probability to the constant R(u,v,...,w) provided the latter is finite.

The result most useful for our purposes is a corollary of this theorem, however. In particular, as n becomes large $R(\theta_n,\tau_n,\ldots,\xi_n)$ and a function involving random variables and constants, say $R(u,\tau_n,\ldots,w)$, become "nearly equal" since they are converging in probability to the same constant, R(u,v,...,w).

3.4 Distribution-specific results

We now propose to consider in greater detail a comparison of the two methods assuming four different bivariate distributions. The four we have chosen are forms of common relevant distributions--the normal, logistic, and exponential, which display appropriately different regression equations. Certainly other bivariate forms would have been interesting to consider--the bivariates

beta, lognormal, "Student's" t, and negative binomial to name a few. Many exhibit interesting regression forms and provide interesting grounds for a comparison of the two methods. However, we leave consideration of these other distributions for future consideration.

(1) Bivariate normal--Its common place in both theory and nature as well as its unique properties compels us to include the bivariate normal among our distributions. Since its regression is linear, we expect a priori that comparison will show the Ericksen method to be both unbiased and to exhibit smaller variance here.

Consider the bivariate normal density

$$f(x,y)=(2\Pi\sigma_x\sigma_y)^{-1}(1-\rho^2)^{-1/2}$$
$$\exp\left[-\frac{1}{2(1-\rho^2)}\left\{\left(\frac{x-\mu_x}{\sigma_x}\right)^2-2\rho\left(\frac{x-\mu_x}{\sigma_x}\right)\left(\frac{y-\mu_y}{\sigma_y}\right)+\left(\frac{y-\mu_y}{\sigma_y}\right)^2\right\}\right], \quad -\infty<x,y<\infty, \tag{3.4.1}$$

which has means, μ_x and μ_y, variances, σ_x^2 and σ_y^2, measure of association, ρ, and joint c.d.f.

$$F(x,y)=\int_{-\infty}^{y}\int_{-\infty}^{x} f(x,t)\,ds\,dt \quad . \tag{3.4.2}$$

The marginal on x has normal density

$$f_1(x)=(2\Pi\sigma_x^2)^{-1/2}\exp\left[-1/2\left(\frac{x-\mu_x}{\sigma_x}\right)^2\right], \quad -\infty<x<\infty \tag{3.4.3}$$

which has mean, μ_x, variance, σ_x^2, and c.d.f.

$$F_1(x)=\int_{-\infty}^{x} f_1(s)ds \quad . \tag{3.4.4}$$

It follows, then, that the conditional distribution of y given x has normal density

$$g(y|x)=(2\Pi\sigma_y{}^2(1-\rho^2))^{-1/2}$$

$$\exp\left[-\{2\sigma_y{}^2(1-\rho^2)\}^{-1}\left\{y-\mu_y-\frac{\rho\sigma_y}{\sigma_x}\left[\frac{x-\mu_x}{\sigma_x}\right]\right\}^2\right] \tag{3.4.5}$$

which has mean, $\mu_y+(\rho\sigma_y/\sigma_x)(x-\mu_x)$, and variance, $\sigma_y{}^2(1-\rho^2)$.

It is the only distribution with a linear regression line that we consider here. We express the conditional mean as

$$E(y|x)=\alpha+\beta x \tag{3.4.6}$$

where

$$\beta=\rho\frac{\sigma_y}{\sigma_x} \tag{3.4.7}$$

and

$$\alpha=\mu_y-\beta\mu_x \quad . \tag{3.4.8}$$

Equation (3.3.23) becomes

$$y=\alpha+\beta x+\varepsilon \quad , \tag{3.4.9}$$

and assuming the mean of ε in the target area to be zero, we let

$$\bar{y}_o{}^*=\alpha+\beta\bar{x}^*=\mu_y+\rho\frac{\sigma_y}{\sigma_x}(\bar{x}^*-\mu_x) \quad . \tag{3.4.10}$$

Watterson [1959] has shown that for the bivariate normal distribution

$$E(y_{[j]})=\mu_y+\rho\sigma_y\mu_{j:n} \quad , \tag{3.4.11}$$

$$Var(y_{[j]})=(1-\rho^2)\sigma_y{}^2+\rho^2\sigma_y{}^2\sigma_{jj:n} \quad , \tag{3.4.12}$$

$$Cov(y_{[j]},y_{[j']})=\rho^2\sigma_y{}^2\sigma_{jj':n} \ , \ j\neq j' \quad , \tag{3.4.13}$$

where ρ is the correlation coefficient of x and y, $\mu_{j:n}$ is the mean of the j-th order statistic and $\sigma_{jj':n}$ is the covariance of the j-th and j'-th order statistics in a sample of size n from a standard normal population. Computation of $\mu_{j:n}$ and $\sigma_{jj':n}$ involves an approximation. A discussion of this topic appears in Appendix A. From (3.3.11) and (3.3.12) we get

$$E(\bar{y}_p{}^*)=\mu_y+\rho\sigma_y\sum\delta_g\mu_{s_g+t:n} \tag{3.4.14}$$

and

$$Var(\bar{y}_p{}^*)$$

$$=\sigma_y{}^2\left[\rho^2\sum\sum{}'\delta_g\delta_{g'}\sigma_{s_g+t;s_{g'}+t':n}+(1-\rho^2)\sum\delta_g{}^2\right] \quad . \tag{3.4.15}$$

Also, from (3.3.24), (3.3.25) and (3.4.10)

$$Bias(\bar{y}_p{}^*)=\mu_y+\rho\sigma_y\sum\delta_g\mu_{s_g+t:n}-(\alpha+\beta\bar{x}^*)$$

$$=\rho\sigma_y\left[\sum\delta_g\mu_{s_g+t:n}-\frac{(\bar{x}^*-\mu_x)}{\sigma_x}\right] \tag{3.4.16}$$

and

$$MSE(\bar{y}_p{}^*)=\sigma_y{}^2\left[\rho^2\sum_{all}\sum{}'\delta_g\delta_{g'}\sigma_{s_g+t;s_{g'}+t':n}\right.$$

$$\left.+(1-\rho^2)\sum\delta_g{}^2+\rho^2\left\{\sum\delta_g\mu_{s_g+t:n}-\frac{(\bar{x}^*-\mu_x)}{\sigma_x}\right\}^2\right] \quad . \tag{3.4.17}$$

We now turn to the distribution-specific results for the Ericksen method involving the normal distribution. Using (3.3.42) and (3.3.43),

$$E(\hat{\alpha})=E[E(\hat{\alpha}|\underset{\sim}{x})]=E(\alpha)=\alpha \tag{3.4.18}$$

and similarly

$$E(\hat{\beta})=\beta \tag{3.4.19}$$

so that

$$E(\bar{y}_E^*)=\alpha+\beta\bar{x}^* \quad . \tag{3.4.20}$$

Thus from (3.4.10)

$$\text{Bias}(\bar{y}_E^*)=0 \quad . \tag{3.4.21}$$

Using Slutsky's Theorem and since

$$E(\hat{\alpha}|\underset{\sim}{x})=\alpha \tag{3.4.22}$$

and

$$E(\hat{\beta}|\underset{\sim}{x})=\beta \quad , \tag{3.4.23}$$

(3.3.44) and (3.3.46) become, for large n,

$$\text{Var}(\hat{\alpha})=E[\text{Var}(\hat{\alpha}|\underset{\sim}{x})]=\frac{(1-\rho^2)\sigma_y^2(\mu_x^2+\sigma_x^2)}{n\sigma_x^2} \quad , \tag{3.4.24}$$

$$\text{Var}(\hat{\beta})=E[\text{Var}(\hat{\beta}|\underset{\sim}{x})]=\frac{(1-\rho^2)\sigma_y^2}{n\sigma_x^2} \quad , \tag{3.4.25}$$

$$\text{Cov}(\hat{\alpha},\hat{\beta})=E[\text{Cov}(\hat{\alpha},\hat{\beta}|\underset{\sim}{x})]=-\frac{(1-\rho^2)\sigma_y^2\mu_x}{n\sigma_x^2} \quad . \tag{3.4.26}$$

Finally, from (3.3.41) and (3.3.48)

$$\text{Var}(\bar{y}_E^*)=\text{MSE}(\bar{y}_E^*)=\frac{(1-\rho^2)\sigma_y^2}{n\sigma_x^2}[(\mu_x-\bar{x}^*)^2+\sigma_x^2] \quad . \tag{3.4.27}$$

(2) Bivariate logistic type 1--The univariate logistic distribution is similar in many respects to the normal distribution yet the bivariate forms that are considered exhibit properties which differ markedly from the bivariate normal. We now discuss the first of two different bivariate logistic distributions attributed to Gumbel [1961]. First, some of its important properties are defined. Consider the following joint c.d.f.

$$F(x,y)=[1+\exp\{-(a+bx)\}+\exp\{-(c+dy)\}]^{-1} \qquad (3.4.28)$$

with marginal c.d.f. on x

$$F_1(x)=[1+\exp\{-(a+bx)\}]^{-1} \qquad (3.4.29)$$

and on y

$$F_2(y)=[1+\exp\{-(c+dy)\}]^{-1} \qquad (3.4.30)$$

where

$$a=-\mu_x\Pi/\sigma_x\sqrt{3} \;;\; b=\Pi/\sigma_x\sqrt{3}$$

$$c=-\mu_y\Pi/\sigma_y\sqrt{3} \;;\; d=\Pi/\sigma_y\sqrt{3} \quad . \qquad (3.4.31)$$

Equation (3.4.28) corresponds to the joint density

$$f(x,y)=2bd\,\exp[-\{(a+bx)+(c+dy)\}]$$

$$\cdot[1+\exp\{-(a+bx)\}+\exp\{-(c+dy)\}]^{-3} \;,\; -\infty<x,y<\infty$$

which has means, μ_x and μ_y, variances, σ_x^2 and σ_y^2, and a level of association which corresponds mathematically to $\rho=1/2$. Although the marginals are symmetric about the mean, the bivariate function is asymmetric.

The marginal distribution on x is expressed as a

logistic density

$$f_1(x)=b\ \exp\{-(a+bx)\}[1+\exp\{-(a+bx)\}]^{-2}\ ,\ -\infty<x<\infty \qquad (3.4.32)$$

which has mean, μ_x, and variance, σ_x^2. From this we obtain

$$g(y|x)=\frac{f(x,y)}{f_1(x)}$$

$$=\frac{2d\ \exp\{-(c+dy)\}[1+\exp\{-(a+bx)\}]^2}{[1+\exp\{-(a+bx)\}+\exp\{-(c+dy)\}]^3} \qquad (3.4.33)$$

We now find the conditional moment generating function (m.g.f.)

$$G(t_2|x)=\int_{-\infty}^{\infty} e^{yt_2} g(y|x)\,dy$$

$$=2[1+\exp\{-(a+bx)\}]^2$$

$$\int_{-\infty}^{\infty} \frac{d\ \exp\{-(c+dy)\}e^{yt_2}dy}{[1+\exp\{-(a+bx)\}+\exp\{-(c+dy)\}]^3} \ . \qquad (3.4.34)$$

Substituting

$$z=\frac{1+\exp\{-(a+bx)\}}{1+\exp\{-(a+bx)\}+\exp\{-(c+dy)\}} \qquad (3.4.35)$$

and after a bit of algebra, we get

$$G(t_2|x)=\exp\{-\frac{ct_2}{d}\}[F_1(x)]^{t_2/d}\,\Gamma(2+t_2/d)\Gamma(1-t_2/d) \qquad (3.4.36)$$

where $\Gamma(\cdot)$ is the ordinary gamma function. The natural logarithm of $G(t_2|x)$ (i.e., conditional cumulant) is

$$\log G(t_2|x)=-\frac{ct_2}{d}+\frac{t_2}{d}\log F_1(x)$$

$$+\log(1+t_2/d)+\log\Gamma(1+t_2/d)+\log\Gamma(1-t_2/d) \ . \qquad (3.4.37)$$

Taking the first derivative in (3.4.37) with respect to t_2 and setting $t_2=0$, we get

$$E(y|x)= \frac{1}{d}[1-c+\log F_1(x)] \quad . \tag{3.4.38}$$

Thus the regression is nonlinear in x.

Similarly, after the second derivative and setting $t_2=0$ we get

$$Var(y|x)=(1-3/\Pi^2)\sigma_y{}^2 \quad . \tag{3.4.39}$$

To find $E(y_{[j]})$ we use (3.3.13) and (3.4.38) to get

$$E(y_{[j]})= \frac{1}{d}[1-c+E\{\log F_1(x_{(j)})\}] \quad . \tag{3.4.40}$$

The density of the j-th order statistic can be expressed (David [1970]) as

$$g(x_{(j)})= \frac{[F_1(x_{(j)})]^{j-1}[1-F_1(x_{(j)})]^{n-j}f_1(x_{(j)})}{B(j,n-j+1)} \tag{3.4.41}$$

where

$$B(j,n-j+1)= \frac{\Gamma(j)\Gamma(n-j+1)}{\Gamma(n+1)} \tag{3.4.42}$$

Substituting $z=F_1(x_{(j)})$ with $dz=f_1(x_{(j)})dx_{(j)}$, we can find E(log z) from the m.g.f. of log z which is

$$H(t)=E(e^{t \log z})= \frac{\Gamma(n+1)\Gamma(j+t)}{\Gamma(j)\Gamma(n+1+t)} \tag{3.4.43}$$

Taking logarithms we get

$$\log H(t)=\log \Gamma(n+1) + \log \Gamma(j+t)-\log \Gamma(j)-\log \Gamma(n+1+t) \tag{3.4.44}$$

The first derivative is

$$\frac{\delta \log H(t)}{\delta t} =\Psi(j+t)-\Psi(n+1+t)=-\sum_{q=j}^{n}(q+t)^{-1} \tag{3.4.45}$$

where $\Psi(\cdot)$ is the digamma function defined by

$$\Psi(m)=\Gamma'(m)/\Gamma(m)=\Psi(1+z)+\sum_{q=1}^{k-1}(q+z)^{-1} \qquad (3.4.46)$$

and $m=k+z$ for integral k. Setting $t=0$ in (3.4.45) we have

$$E(\log z)=\Psi(j)-\Psi(n+1)=-\sum_{q=j}^{n}q^{-1} \qquad (3.4.47)$$

which yields, from (3.4.40) and (3.4.32),

$$E(y_{[j]})=\mu_y+\frac{\sigma_y\sqrt{3}}{\pi}(1-\sum_{q=j}^{n}q^{-1}) \quad . \qquad (3.4.48)$$

The second derivative of (3.4.44) is

$$\frac{\delta^2 \log H(t)}{\delta t^2}=\sum_{q=j}^{n}(q+t)^{-2} \quad . \qquad (3.4.49)$$

Setting $t=0$ we find

$$\mathrm{Var}(\log z)=\sum_{q=j}^{n}q^{-2} \quad . \qquad (3.4.50)$$

So from (3.4.38) and (3.4.30)

$$\mathrm{Var}(E(y_{[j]}|x_{(j)}))=\frac{3\sigma_y^2}{\pi^2}\sum_{q=j}^{n}q^{-2} \quad . \qquad (3.4.51)$$

Since from (3.4.39)

$$E[\mathrm{Var}(y_{[j]}|x_{(j)})]=(1-3/\pi^2)\sigma_y^2 \quad , \qquad (3.4.52)$$

finally, from (3.3.14),

$$\mathrm{Var}(y_{[j]})=\sigma_y^2[1-\frac{3}{\pi^2}(1-\sum_{q=j}^{n}q^{-2})] \quad . \qquad (3.4.53)$$

To find the covariance we begin by noticing that

$$\mathrm{Cov}[E(y_{[j]}|x_{(j)}),E(y_{[j']}|x_{(j')})]$$

$$= \frac{1}{d^2}\,\mathrm{Cov}[\log F_1(x_{(j)}),\log F_1(x_{(j')})] \qquad (3.4.54)$$

Substituting $z=F_1(x_{(j)})$ and $z'=F_1(x_{(j')})$, we find, from David once more, that the density of z and z' is

$$h(z,z')$$

$$= \frac{\Gamma(n+1)}{\Gamma(j)\Gamma(j'-j)\Gamma(n-n'+1)}\, z^{j-1}[z'-z]^{j'-j'1}[1-z']^{n-j'}\,,$$

$$0\leq z\leq z'\leq 1 \quad . \qquad (3.4.55)$$

The joint m.g.f. of log z and log z' is

$$M(t,t')=E[\exp\{t\log z+t'\log z'\}]$$

$$= \frac{\Gamma(n+1)\Gamma(t+t'+j')\Gamma(t+j)}{\Gamma(j)\Gamma(t+t'+n+1)\Gamma(t+j')} \quad . \qquad (3.4.56)$$

Taking the logarithm,

$$\log M(t,t')=\log\Gamma(n+1)$$

$$+\log\Gamma(t+t'+j')+\log\Gamma(t+j)-\log\Gamma(j)$$

$$-\log\Gamma(t+t'+n+1)-\log\Gamma(t+j') \quad , \qquad (3.4.57)$$

$\frac{\delta^2 \log M(t,t')}{\delta t\delta t'}$, evaluated at t=0 and t'=0, yields

$$\mathrm{Cov}[\log F_1(x_{(j)}),\log F_1(x_{(j')})]=\sum_{q=j'}^{n} q^{-2} \qquad (3.4.58)$$

and thus

$$\mathrm{Cov}[E(y_{[j]}|x_{(j)}),E(y_{[j']}|x_{(j')})]= \frac{1}{d^2}\sum_{q=j'}^{n} q^{-2} \quad . \qquad (3.4.59)$$

Finally, using (3.3.22) and (3.4.30),

$$\mathrm{Cov}(y_{[j]},y_{[j']})= \frac{3\sigma_y^{\,2}}{\Pi^2}\sum_{q=j'}^{n} q^{-2} \quad . \qquad (3.4.60)$$

Using (3.4.48), (3.4.53) and (3.4.60), we substitute into (3.3.11) and (3.3.12) to obtain

$$E(\bar{y}_p^*)=\mu_y\sum\delta_g+\frac{\sigma_y\sqrt{3}}{\pi}\sum\delta_g(1-\sum_{q=s_g+t}^{n}q^{-1} \qquad (3.4.61)$$

and

$$Var(\bar{y}_p^*)=\sigma_y^2[\sum\delta_g^2\{1-\frac{3}{\pi^2}(1-\sum_{q=s_g+t}^{n}q^{-2})\}$$

$$+\sum\sum{}'\delta_g\delta_{g'}\{\frac{3}{\pi^2}\sum_{q=s_{g'}+t'}^{n}q^{-2}\}] \quad . \qquad (3.4.62)$$

From (3.3.23) and (3.4.38) we find

$$\bar{y}_o^*=\frac{[1-c+\frac{1}{n^*}\sum_{i=1}^{n^*}\log F_1(x_i^*)]}{d} \qquad (3.4.63)$$

and from (3.3.24), (3.3.25), (3.4.61) and (3.4.63)

$$Bias(\bar{y}_p^*)\doteq\mu_y\sum\delta_g$$

$$+\frac{\sigma_y\sqrt{3}}{\pi}\sum\delta_g(1-\sum_{q=s_g+t}^{n}q^{-1})$$

$$-\frac{1}{d}\{1-c+\frac{1}{n^*}\sum_{i=1}^{n^*}\log F_1(x_1^*)\} \quad . \qquad (3.4.64)$$

Finally, we substitute (3.4.62) and (3.4.64) into

$$MSE(\bar{y}_p^*)=Var(\bar{y}_p^*)+Bias^2(\bar{y}_p^*) \quad . \qquad (3.4.65)$$

We now find the properties for the Ericksen method assuming this bivariate logistic distribution. By Slutsky's Theorem if we are given $\underset{\sim}{x}$, then as n becomes large

$$\hat{\beta} \doteq \hat{\beta}' = \frac{\sum_{i=1}^{n} (x_i - \bar{x}) y_i}{n\sigma_x^2} \tag{3.4.66}$$

and

$$\hat{\alpha} \doteq \hat{\alpha}' = \bar{y} - \hat{\beta}' \bar{x} \quad . \tag{3.4.67}$$

Thus by (3.3.43)

$$E(\hat{\beta}) \doteq E(\hat{\beta}') = E[E(\hat{\beta}' | \underset{\sim}{x})]$$

$$= E\left[E\left\{\frac{\sum_{i=1}^{n} (x_i - \bar{x}) y_i}{n\sigma_x^2}\right\}\right] = \left(\frac{n-1}{n}\right)\frac{\sigma_y}{2\sigma_x} \quad . \tag{3.4.68}$$

Similarly, by (3.3.42)

$$E(\hat{\alpha}) \doteq E(\hat{\alpha}') = E[E(\hat{\alpha}' | \underset{\sim}{x})]$$

$$= E[E(\bar{y} - \hat{\beta}'\bar{x} | \underset{\sim}{x})]$$

$$= \frac{1}{n\sigma_x^2} E[\sum_{i=1}^{n} (\sigma_x^2 - x_i\bar{x} + \bar{x}^2) E(y_i | x_i)]$$

$$= \frac{1}{d}[-c - \frac{1}{n\sigma_x^2} E[\sum_{i=1}^{n} \bar{x}(x_i - \bar{x}) \log F_1(x_i)]]$$

$$= \mu_y - \frac{n-1}{n^2 d\sigma_x^2}[\sigma_x^2 + (n-2)v_3\mu_x + (n-1)\mu_x^2 + v_2] \quad , \tag{3.4.69}$$

where

$$v_2 = \frac{1}{b^2}[I_{012} - 2I_{021} - \frac{a\Pi^2}{3} - a^2 - 6] \quad , \tag{3.4.70}$$

$$v_3 = \frac{1}{b}[\frac{\Pi^2}{6} + a] \quad , \tag{3.4.71}$$

and we define

$$I_{ijk}=\int_0^1 z^i \log^j z \log^k(1-z)dz \quad . \tag{3.4.72}$$

To find the variances and covariance for α and β we use the result (again according to Slutsky) that for given $\underset{\sim}{x}$, as n becomes large

$$\hat{\beta}\doteq\hat{\beta}''=\frac{\sum_{i=1}^{n}(x_i-\mu_x)y_i}{n\sigma_x^2} \tag{3.4.73}$$

and

$$\hat{\alpha}\doteq\hat{\alpha}''=\bar{y}-\hat{\beta}''\mu_x \quad . \tag{3.4.74}$$

The random variables $\hat{\alpha}''$ and $\hat{\beta}''$ differ from $\hat{\alpha}'$ and $\hat{\beta}'$ in that $\bar{x}$ is replaced by μ_x to simplify the formulation.

Now, to find $\mathrm{Var}(\hat{\beta})$ by (3.3.45) we see, first, that

$$E[\mathrm{Var}(\hat{\beta}''|\underset{\sim}{x})]\doteq\frac{(1-3/\pi^2)\sigma_y^2}{n\sigma_x^2} \tag{3.4.75}$$

and using (3.4.73),

$$\begin{aligned}\mathrm{Var}[E(\hat{\beta}|\underset{\sim}{x})]&\doteq\mathrm{Var}[E(\hat{\beta}''|\underset{\sim}{x})]\\ &=\frac{1}{n^2d^2\sigma_x^4}\mathrm{Var}\left[\sum_{i=1}^{n}(x_i-\mu_x)(1-c+\log F_1(x_i))\right]\\ &=\frac{1}{nd^2\sigma_x^4}[(1-c)^2\sigma_x^2+(2c-1)\mu_x^2\\ &\quad-\{(6-4c)v_3+2v_4\}\mu_x+(v_1-v_3^2)+2(1-c)v_2] \end{aligned}\tag{3.4.76}$$

where

$$v_1 = \frac{1}{b^2}[I_{022} - 2I_{031} + 2aI_{021} + 2a^2 + 12a + 24] \qquad (3.4.77)$$

and

$$v_4 = \frac{1}{b}[-I_{021} - 2a - 6] \qquad (3.4.78)$$

Finally, using the results of (3.4.75) and (3.4.76), the $\mathrm{Var}(\hat{\beta})$ is obtained from (3.3.45).

From (3.4.74),

$$\mathrm{Var}[E(\hat{\alpha}|\underset{\sim}{x})] \doteq \mathrm{Var}[E(\hat{\alpha}''|\underset{\sim}{x})]$$

$$= \frac{1}{n^2 d^2 \sigma_x^4}\mathrm{Var}\left[\sum_{i=1}^{n}(\sigma_x^2 - x_i\mu_x + \mu_x^2)(1 - c + \log F_1(x_i))\right]$$

$$= \frac{1}{nd^2\sigma_x^4}[\sigma_x^4 - \{2v_3(2-c) + 2v_4\}\sigma_x^2\mu_x + (1+c^2)\sigma_x^2\mu_x^2$$

$$+ \{(v_1 - v_3^2) + 2(1-c)v_2\}\mu_x^2 - \{(6-4c)v_3 + 2v_4\}\mu_x^3$$

$$+ (2c-1)\mu_x^4] \quad . \qquad (3.4.79)$$

Also

$$E[\mathrm{Var}(\hat{\alpha}|\underset{\sim}{x})] \doteq \frac{(1 - 3/\Pi^2)\sigma_y^2(\mu_x^2 + \sigma_x^2)}{n\sigma_x^2} \qquad (3.4.80)$$

thus

$$\mathrm{Var}(\hat{\alpha}) \doteq E[\mathrm{Var}(\hat{\alpha}''|\underset{\sim}{x})] + \mathrm{Var}[E(\hat{\alpha}''|\underset{\sim}{x})] \quad , \qquad (3.4.81)$$

where the components are obtained from (3.4.79) and (3.4.80).

Finally, to find $\mathrm{Cov}(\hat{\alpha},\hat{\beta})$, from (3.4.73) and (3.4.74),

$$\mathrm{Cov}[E(\hat{\alpha}|\underset{\sim}{x}), E(\hat{\beta}|\underset{\sim}{x})] \doteq \mathrm{Cov}[E(\hat{\alpha}''|\underset{\sim}{x}), E(\hat{\beta}''|\underset{\sim}{x})]$$

$$= \mathrm{Cov}[E(\bar{y}|\underset{\sim}{x}), E(\hat{\beta}''|\underset{\sim}{x})] - \mu_x \mathrm{Var}[E(\hat{\beta}''|\underset{\sim}{x})]$$

$$=-\frac{1}{nd^2\sigma_x^4}[\{c^2-c+1\}\sigma_x^2\mu_x-\{(2-c)\nu_3+\nu_4\}\sigma_x^2$$

$$+\{2(1-c)\nu_2+(\nu_1-\nu_3^2)\}\mu_x$$

$$-\{4(1-c)\nu_3+2(\nu_4+\nu_3)\}\mu_x^2+(2c-1)\mu_x^3] \quad . \qquad (3.4.82)$$

Also, it can be shown that

$$E[Cov(\hat{\alpha},\hat{\beta}|\underset{\sim}{x})]\doteq-\frac{(1-3/\pi^2)\sigma_y^2\mu_x}{n\sigma_x^2} \quad . \qquad (3.4.83)$$

Finally,

$$Cov(\hat{\alpha},\hat{\beta})\doteq E[Cov(\hat{\alpha},\hat{\beta}|\underset{\sim}{x})]+Cov[E(\hat{\alpha}|\underset{\sim}{x}),E(\hat{\beta}|\underset{\sim}{x})] \qquad (3.4.84)$$

as given by (3.4.82) and (3.4.83).

Using the results in (3.4.63), (3.4.68), (3.4.69), (3.4.78), (3.4.81) and (3.4.84), we obtain the properties for the Ericksen method from (3.3.40)-(3.3.41) and (3.3.47)-(3.3.48).

(3) Bivariate logistic type 2--By contrast we also consider a symmetric form of the bivariate logistic distribution. It is obtained from a bivariate c.d.f. generation formula suggested by Gumbel [1960a]

$$F(x,y)=F_1(x)F_2(y)[1+\alpha_o\{1-F_1(x)\}\{1-F_2(y)\}] \ , \ -1\leq\alpha_o\leq1 \qquad (3.4.85)$$

where α_o is the parameter of association. The marginal c.d.f. on x, $F(x,\infty)$, is $F_1(x)$. Similarly, the marginal on y is $F_2(y)$. From (3.4.29) and (3.4.30)

$$F(x,y)=[1+\exp\{-(a+bx)\}]^{-1}$$

$$[1+\exp\{-(c+dy)\}]^{-1}[1+\alpha_o\exp\{-(a+bx)-(c+dy)\}$$

$$\cdot\{1+\exp\{-(a+bx)\}\}^{-1}\{1+\exp\{-(c+dy)\}\}^{-1}] \quad , \qquad (3.4.86)$$

where α_o is the association parameter which corresponds to $\rho=3\alpha_o/\pi^2$ so that

$$|\rho|\leq 0.303964 \quad . \qquad (3.4.87)$$

It follows then that

$$f(x,y)=bd\ \exp\{-(a+bx)-(c+dy)\}$$

$$[1+\exp\{-(a+bx)\}]^{-2}[1+\exp\{-(c+dy)\}]^{-2}$$

$$\cdot\left[1+\alpha_o\left\{\frac{1-\exp\{-(a+bx)\}}{1+\exp\{-(a+bx)\}}\right\}\left\{\frac{1-\exp\{-(c+dy)\}}{1+\exp\{-(c+dy)\}}\right\}\right] \quad . \qquad (3.4.88)$$

From (3.4.32) and (3.4.88) we find

$$g(y|x)=\frac{d\ \exp\{-(c+dy)\}\left[1+\alpha_o\left\{\frac{1-\exp\{-(a+bx)\}}{1+\exp\{-(a+bx)\}}\right\}\left\{\frac{1-\exp\{-(c+dy)\}}{1+\exp\{-(c+dy)\}}\right\}\right]}{[1+\exp\{-(c+dy)\}]^2} \qquad (3.4.89)$$

The conditional moment generating function is

$$G(t_2|x)=\int_{-\infty}^{\infty} e^{yt_2}g(y|x)\,dy \quad . \qquad (3.4.90)$$

Substituting $w=F_2(y)=[1+\exp\{-(c+dy)\}]^{-1}$ leads to

$$G(t_2|x)=\exp\left\{-\frac{ct_2}{d}\right\}\left[1+\frac{\alpha_o t_2}{d}\{2F_1(x)-1\}\right]$$

$$[\Gamma(1+t_2/d)\Gamma(1-t_2/d)] \quad . \qquad (3.4.91)$$

The first derivative of $\log G(t_2|x)$ with respect to t_2 is

$$\frac{\delta \log G}{\delta t_2} = -\frac{c}{d} + \frac{\frac{\alpha_o}{d}\{2F_1(x)-1\}}{1+\frac{\alpha_o t_2}{d}\{2F_1(x)-1\}}$$

$$+ \frac{1}{d}[\Psi(1+t_2/d)-\Psi(1-t_2/d)] \quad . \qquad (3.4.92)$$

Setting $t_2=0$ we obtain

$$E(y|x) = \frac{1}{d}[\alpha_o\{2F_1(x)-1\}-c]$$

$$= \frac{2\alpha_o}{d}F_1(x) - \left[\frac{\alpha_o+c}{d}\right] = \mu_y + \frac{\sigma_y\sqrt{3}\alpha_o}{\pi}\{2F_1(x)-1\} \quad . \qquad (3.4.93)$$

Thus the regression is nonlinear in x although it is linear in probability. The second derivative is

$$\frac{\delta^2 \log G}{\delta t_2^2} = \frac{-q^2}{(1+qt_2)^2} + \frac{1}{d^2}[\Psi'(1+t_2/d)+\Psi'(1-t_2/d)] \qquad (3.4.94)$$

where

$$q = \frac{\alpha_o}{d}[2F_1(x)-1] \qquad (3.4.95)$$

and Ψ' represents the derivative of Ψ with respect to t_2. Again, we set $t_2=0$ and find

$$Var(y|x) = -q^2 + \frac{2\Psi'(1)}{d^2} \quad . \qquad (3.4.96)$$

Since $\Psi'(1)=\zeta(2)=\pi^2/6$,

$$Var(y|x) = \sigma_y^2\left[1-\frac{3\alpha_o^2}{\pi^2}\{2F_1(x)-1\}^2\right] \quad . \qquad (3.4.97)$$

Using (3.3.13) we note that

$$E(y_{[j]}) = E[E(y_{[j]}|x_{(j)}]$$

$$= \frac{1}{d}[2\alpha_o E[F_1(x_{(j)})]-(\alpha_o+c)] \quad . \qquad (3.4.98)$$

As with the first bivariate logistic distribution, the density of $z=F_1(x_{(j)})$ is

$$h(z)= \frac{z^{j-1}(1-z)^{n-j}}{B(j,n+1-j)} \quad , \quad 0\leq z\leq 1 \quad . \qquad (3.4.99)$$

We now find

$$E[F_1(x_{(j)})]=E(z)= \frac{B(j+1,n+1-j)}{B(j,n+1-j)} = \frac{j}{n+1} = p_j \quad , \qquad (3.4.100)$$

and thus from (3.4.98)

$$E(y_{[j]})= \frac{1}{d}[2\alpha_o p_j - (\alpha_o+c)]$$

$$=\mu_y + \frac{\sigma_y\sqrt{3}\alpha_o}{\pi}(2p_j-1) \quad . \qquad (3.4.101)$$

Using (3.4.93) we now wish to find

$$Var[E(y_{[j]}|x_{(j)})]= \frac{4\alpha_o^2}{d^2}Var[F_1(x_{(j)})]$$

$$= \frac{4\alpha_o^2}{d^2}Var\ (z) \quad . \qquad (3.4.102)$$

But

$$Var(z)=E(z^2)-[E(z)]^2 \qquad (3.4.103)$$

and from (3.4.99)

$$E(z^2)= \frac{j(j+1)}{(n+1)(n+2)} \quad . \qquad (3.4.104)$$

So

$$Var(z)- \frac{j(j+1)}{(n+1)(n+2)} - \left[\frac{j}{n+1}\right]^2 = \frac{p_j(1-p_j)}{n+2} \qquad (3.4.105)$$

and thus

$$\mathrm{Var}[E(y_{[j]}|x_{(j)})] = \frac{4\alpha_o^2 p_j(1-p_j)}{d^2(n+2)}$$

$$= \sigma_y^2\left[\frac{12\alpha_o^2 p_j(1-p_j)}{\Pi^2(n+2)}\right] \quad . \qquad (3.4.106)$$

Next we find from (3.4.97) that

$$E[\mathrm{Var}(y_{[j]}|x_{(j)})]$$

$$= \sigma_y^2\left[1 - \frac{3\alpha_o^2}{\Pi^2}E\{[2F_1(x_{(j)})-1]\}\right] \quad . \qquad (3.4.107)$$

But

$$E\{[2F_1(x_{(j)})-1]^2\} = E\{(2z-1)^2\} = 1-4\left(\frac{n+1}{n+2}\right)p_j(1-p_j) \qquad (3.4.108)$$

so

$$E[\mathrm{Var}(y_{[j]}|x_{(j)})]$$

$$= \sigma_y\left[1 - \frac{3\alpha_o^2}{\Pi^2}\left\{1-4\left(\frac{n+1}{n+2}\right)p_j(1-p_j)\right\}\right] \quad . \qquad (3.4.109)$$

Finally, using (3.3.14)

$$\mathrm{Var}(y_{[j]}) = \sigma_y^2\left[1 - \frac{3\alpha_o^2}{\Pi^2}\{1-4p_j(1-p_j)\}\right] \quad . \qquad (3.4.110)$$

Using the joint density of $z=F_1(x_{(j)})$ and $z'=F_1(x_{(j')})$, which is found for the first bivariate logistic, and (3.4.93) we find that

$$\mathrm{Cov}[E(y_{[j]}|x_{(j)}), E(y_{[j']}|x_{(j')})]$$

$$= \frac{4\alpha_o^2}{d^2}\mathrm{Cov}[F_1(x_{(j)}, F_1(x_{(j')})]$$

$$= \frac{4\alpha_o^2}{d^2}\mathrm{Cov}(z,z') \quad . \tag{3.4.111}$$

But

$$\mathrm{Cov}(z,z')=E(zz')-E(z)E(z') \tag{3.4.112}$$

and

$$E(zz')= \frac{\Gamma(n+1)\Gamma(j+1)\Gamma(j'+2)}{\Gamma(n+3)\Gamma(j)\Gamma(j'+1)} = \frac{j(j'+1)}{(n+1)(n+2)} \quad . \tag{3.4.113}$$

Thus

$$\mathrm{Cov}(z,z')= \frac{j(j'+1)}{(n+1)(n+2)} - \frac{jj'}{(n+1)^2} = \frac{p_j(1-p_{j'})}{n+2} \tag{3.4.114}$$

and

$$\mathrm{Cov}[E(y_{[j]}|x_{(j)}),E(y_{[j']}|x_{(j')})]$$

$$= \frac{4\alpha_o^2 p_j(1-p_{j'})}{d^2(n+2)}$$

$$=\sigma_y^2\left(\frac{12\alpha_o^2 p_j(1-p_{j'})}{\Pi^2(n+2)}\right)$$

$$=\mathrm{Cov}[y_{[j]},y_{[j']}] \tag{3.4.115}$$

Since

$$E[\mathrm{Cov}(y_{[j]},y_{[j']}|x_{(j)},x_{(j')})]=0 \quad . \tag{3.4.116}$$

For the proposed method we use (3.4.101) to write

$$E(\bar{y}_p^*)=\sum\delta_g E[y_{[s_g+t]}]$$

$$=\mu_y+ \frac{\sigma_y\sqrt{3}\alpha_o}{\Pi}\sum\delta_g(2p_{s_g+t}-1) \quad . \tag{3.4.117}$$

We obtain Π_{gj} and thus δ_g as with the bivariate logistic type 1. Furthermore, from (3.4.110) and (3.4.115) it

follows that

$$\mathrm{Var}(\bar{y}_p^*)=\sigma_y^2\left[\sum\delta_g^2\left\{1-\frac{3\alpha_o^2}{\Pi^2}[1-4p_{s_g+t}(1-p_{s_g+t})]\right\}\right.$$

$$\left.+\frac{12\alpha_o^2}{\Pi^2(n+2)}\sum\sum{}'\delta_g\delta_{g'}p_{s_g+t}(1-p_{s_{g'}+t'})\right] \quad . \qquad (3.4.118)$$

Also from (3.3.23) and (3.4.93) we obtain

$$\bar{y}_o^*\doteq\mu_y+\frac{\sigma_y\sqrt{3}\alpha_o}{\Pi}\left[\frac{2}{n^*}\sum_{i=1}^{n^*}F_1(x_i^*)-1\right] \quad . \qquad (3.4.119)$$

From (3.3.24), using our results from (3.4.117) and (3.4.119), we find that

$$\mathrm{Bias}(\bar{y}_p^*)\doteq\frac{\sigma_y\sqrt{3}\alpha_o}{\Pi}\left[\sum\delta_g(2p_{s_g+t}-1)-\left\{\frac{2}{n^*}\sum_{i=1}^{n^*}F_1(x_i)-1\right\}\right] \qquad (3.4.120)$$

Then (3.4.118) and (3.4.120) can be substituted into (3.3.25) to calculate the $\mathrm{MSE}(\bar{y}_p^*)$.

We now find the properties for the Ericksen method assuming a sample of size n from the bivariate logistic type 2. As before by Slutsky's Theorem we use the results of (3.4.66) and (3.4.67). Thus

$$E(\hat{\beta})\doteq E(\hat{\beta}')=E[E(\hat{\beta}'|\underset{\sim}{x})]$$

$$=\frac{1}{n\sigma_x^2}E\left[\sum_{i=1}^{n}(x_i-\bar{x})E(y_i|x_i)\right]$$

$$=\left(\frac{n-1}{n}\right)\frac{3\alpha_o}{\Pi^2}\frac{\sigma_y}{\sigma_x} \quad . \qquad (3.4.121)$$

Similarly,

$$E(\hat{\alpha}) \doteq E(\hat{\alpha}') = E[E(\hat{\alpha}'|\underset{\sim}{x})]$$

$$= \frac{1}{n\sigma_x^2} E\left[\sum_{i=1}^{n} (\sigma_x^2 - x_i\bar{x} + \bar{x}^2) E(y_i|x_i)\right]$$

$$= \mu_y + \frac{\alpha_o}{nd\sigma_x^2}\left[\frac{(n-1)^2}{n}\mu_x^2 + \left\{\frac{n-1}{n}\right\}\sigma_x^2 - \left\{\frac{2v_5(n-1)(n-2)}{n}\right\}\mu_x - 2v_6\right] \qquad (3.4.122)$$

where

$$v_5 = \frac{1-a}{2b} = \frac{\sqrt{3}\sigma_x}{2\pi} + \frac{\mu_x}{2} \qquad (3.4.123)$$

and

$$v_6 = \frac{1}{b^2}\left[\frac{(2-a)^2}{2} - 2I_{111}\right] \qquad (3.4.124)$$

and I_{111} is defined by (3.4.72).

To find $\mathrm{Var}(\hat{\alpha})$, $\mathrm{Var}(\hat{\beta})$ and $\mathrm{Cov}(\hat{\alpha},\hat{\beta})$ we also use Slutsky's Theorem as defined by the results (3.4.73) and (3.4.74). First, we note that

$$E[\mathrm{Var}(\hat{\beta}|\underset{\sim}{x})] \doteq E[\mathrm{Var}(\hat{\beta}''|\underset{\sim}{x})]$$

$$= \frac{1}{n^2\sigma_x^4} E\left[\sum_{i=1}^{n} (x_i^2 - 2x_i\mu_x + \mu_x^2)\mathrm{Var}(y_i|x_i)\right]$$

$$= \frac{\sigma_y^2}{n\sigma_x^4}\left[(1-s)\sigma_x^2 + \frac{2}{3}s\mu_x^2 + 8s(v_7 - v_5)\mu_x + 4s(v_6 - v_8)\right] \qquad (3.4.125)$$

where

$$v_7 = \frac{1}{b}\left(\frac{1}{2} - \frac{a}{3}\right) \quad , \qquad (3.4.126)$$

$$v_8 = \frac{1}{b^2}[89/54+a^2/3-a-2I_{211}] \quad , \tag{3.4.127}$$

and

$$s=3\alpha_o{}^2/\Pi^2 \quad . \tag{3.4.128}$$

Also

$$\begin{aligned}
\mathrm{Var}[E(\hat{\beta}|\underset{\sim}{x})] &\doteq \mathrm{Var}[E(\hat{\beta}''|\underset{\sim}{x})] \\
&= \frac{1}{n^2\sigma_x{}^4}\mathrm{Var}[\sum_{i=1}^{n}(x_i-\mu_x)E(y_i|x_i)] \\
&= \frac{1}{nd^2\sigma_x{}^4}[(\alpha_o+c)^2\sigma_x{}^2-\alpha_o(\tfrac{5}{3}\alpha_o+2c)\mu_x{}^2 \\
&\quad +8\alpha_o\{(\tfrac{3}{2}\alpha_o+c)v_5-\alpha_o v_7\}\mu_x \\
&\quad +4\alpha_o\{\alpha_o(v_8-v_5{}^2)-(\alpha_o+c)v_6\}] \quad .
\end{aligned} \tag{3.4.129}$$

Thus the $\mathrm{Var}(\hat{\beta})$ is given by (3.3.45) using (3.4.125) and (3.4.129).

Similarly,

$$\begin{aligned}
E[\mathrm{Var}(\hat{\alpha}|\underset{\sim}{x})] &\doteq E[\mathrm{Var}(\hat{\alpha}''|\underset{\sim}{x})] \\
&= \frac{1}{n^2\sigma_x{}^4}E[\sum_{i=1}^{n}(\sigma_x{}^2+\mu_x{}^2-\mu_x x_i)^2\mathrm{Var}(y_i|x_i)] \\
&= \frac{\sigma_y{}^2}{n\sigma_x{}^4}[(1-\tfrac{s}{3})\sigma_x{}^4+(1+\tfrac{s}{3})\sigma_x{}^2\mu_x{}^2+8s(v_7-v_5)\sigma_x{}^2\mu_x \\
&\quad +4s(v_6-v_8)\mu_x{}^2+8s(v_7-v_5)\mu_x{}^3+\tfrac{2}{3}s\mu_x{}^4]
\end{aligned} \tag{3.4.130}$$

and

$$\mathrm{Var}[E(\hat{\alpha}|\underset{\sim}{x})] \doteq \mathrm{Var}[E(\hat{\alpha}''|\underset{\sim}{x})]$$

$$= \frac{1}{n^2\sigma_x^4}\mathrm{Var}\Big[\sum_{i=1}^{n}(\sigma_x^2+\mu_x^2-\mu_x x_i)E(y_i|x_i)\Big]$$

$$= \frac{1}{nd^2\sigma_x^4}\Big[\frac{\alpha_o^2}{3}\sigma_x^4+\Big\{(\alpha_o+c)^2-2\alpha_o(\alpha_o+c)+\frac{2}{3}\alpha_o^2\Big\}\sigma_x^2\mu_x^2$$

$$+\Big\{4\alpha_o(\alpha_o+c)v_5-8\alpha_o^2(v_7-\frac{v_5}{2})\Big\}\sigma_x^2\mu_x$$

$$+\Big\{4\alpha_o^2(v_8-v_5^2)-4\alpha_o(\alpha_o+c)v_6\Big\}\mu_x^2$$

$$+\Big\{4\alpha_o(\alpha_o+c)v_5-8\alpha_o^2(v_7-\frac{v_5}{2})+4\alpha_o(\alpha_o+c)v_5\Big\}\mu_x^3$$

$$+\Big\{\frac{\alpha_o^2}{3}-2\alpha_o(\alpha_o+c)\Big\}\mu_x^4\Big] \quad . \qquad (3.4.131)$$

Substitution of these expressions in (3.3.44) gives $\mathrm{Var}(\hat{\alpha})$. To find $\mathrm{Cov}(\hat{\alpha},\hat{\beta})$ we calculate

$$E[\mathrm{Cov}(\hat{\alpha},\hat{\beta}|\underset{\sim}{x})] \doteq E[\mathrm{Cov}(\hat{\alpha}'',\hat{\beta}''|\underset{\sim}{x})]$$

$$= \mathrm{Cov}(\bar{y},\hat{\beta}''|\underset{\sim}{x})-\mu_x\mathrm{Var}(\hat{\beta}''|\underset{\sim}{x})$$

$$= \frac{1}{n^2\sigma_x^4}E\Big[\sum_{i=1}^{n}\{\sigma_x^2x_i-\sigma_x^2\mu_x-\mu_x(x_i-\mu_x)^2\}\mathrm{Var}(y_i|x_i)\Big]$$

$$= \frac{\sigma_y^2}{n\sigma_x^4}[4s(v_5-v_7)\sigma_x^2-(1-\frac{s}{3})\sigma_x^2\mu_x+4s(v_6-v_8)\mu_x$$

$$+8s(v_5-v_7)\mu_x^2-\frac{2}{3}s\mu_x^3] \qquad (3.4.132)$$

and

$$\mathrm{Cov}[E(\hat{\alpha}|\underset{\sim}{x}),E(\hat{\beta}|\underset{\sim}{x})]\doteq\mathrm{Cov}[E(\hat{\alpha}''|\underset{\sim}{x}),E(\hat{\beta}''|\underset{\sim}{x})]$$

$$= \frac{1}{n^2\sigma_x^4}\mathrm{Cov}[\sum_{i=1}^{n}(x_i-\mu_x)E(y_i|x_i),$$

$$\sum_{j=1}^{n}(\sigma_x^2+\mu_x^2-\mu_x x_j)E(y_j|x_j)]$$

$$= \frac{1}{nd^2\sigma_x^4}\left[\left\{4\alpha_o^2(v_7-\frac{v_5}{2})-2\alpha_o(\alpha_o+c)v_5\right\}\sigma_x^2\right.$$

$$-\left\{c(\alpha_o+c)+\frac{\alpha_o^2}{3}\right\}\sigma_x^2\mu_x$$

$$+\left\{4\alpha_o(\alpha_o+c)v_6-4\alpha_o^2(v_8-v_5^2)\right\}\mu_x$$

$$+\left\{8\alpha_o^2(v_7-\frac{v_5}{2})-8\alpha_o(\alpha_o+c)v_5\right\}\mu_x^2$$

$$\left.+\left\{2\alpha_o(\alpha_o+c)-\frac{\alpha_o^2}{3}\right\}\mu_x^3\right] \qquad (3.4.133)$$

and then substitute in (3.3.46).

Using several of the preceding results, we can obtain the properties for the Ericksen method from (3.3.40), (3.3.41), (3.3.47) and (3.3.48).

(4) Bivariate exponential--The exponential distribution is a particularly relevant distribution to economists, demographers, and biologists. It is often associated with decay processes pertaining to such diverse things as atomic particles or human life. In addition, it is particularly simple to work with algebraically and has prop-

erties of interest to the present work. We begin by finding some of these properties. Much of the following parallels that which is found by Gumbel [1960b].

As with the bivariate logistic type 2 the c.d.f. of this bivariate distribution is obtained from (3.4.85) and written as

$$F(x,y)=[1-\exp(-ax)][1-\exp(-by)]$$
$$[1+\alpha_1\exp\{-(ax+by)\}] \ , \ 0\leq x,y\leq\infty \qquad (3.4.134)$$

where α_1 is the association parameter which corresponds to $\rho=\frac{\alpha_1}{4}$ so that

$$|\rho|\leq 0.25 \quad . \qquad (3.4.135)$$

We define

$$a=1/\mu_x=1/\sigma_x \qquad (3.4.136)$$

and

$$b=1/\mu_y=1/\sigma_y \quad . \qquad (3.4.137)$$

We note that $F(x,\infty)$ yields the appropriate marginal on x, namely,

$$F_1(x)=[1-\exp[-ax)] \ , \ 0\leq x<\infty \quad . \qquad (3.4.138)$$

Also (3.4.134) corresponds to the joint density

$$f(x,y)=ab\ \exp\{-(ax+by)\}$$
$$[1+\alpha_1\{2\ \exp(-ax)-1\}\{2\ \exp(-by)-1\}] \qquad (3.4.139)$$

and thus

$$f_1(x)=a\ \exp(-ax) \quad . \qquad (3.4.140)$$

From (3.4.139) and (3.4.140)

$$g(y|x)=b\ \exp(-by)[1+\alpha_1\{1-2\ \exp(-ax)\}]$$

$$-2\alpha_1 b\ \exp(-2by)\{1-2\ \exp(-ax)\} \quad . \qquad (3.4.141)$$

Now

$$E(y|x)=\int_0^\infty yg(y|x)dy=\frac{1}{b}[1+\frac{\alpha_1}{2}\{1-2\ \exp(-ax)\}] \quad . \qquad (3.4.142)$$

Similarly,

$$E(y^2|x)=\frac{1}{b^2}[2+\frac{3\alpha_1}{2}\{1-2\ \exp(-ax)\}] \quad . \qquad (3.4.143)$$

So

$$Var(y|x)=\frac{1}{b^2}[1+\frac{\alpha_1}{2}\{1-2\ \exp(-ax)\}$$

$$-\frac{\alpha_1^2}{4}\{1-2\ \exp(-ax)\}^2] \quad . \qquad (3.4.144)$$

Using the density for $z=F_1(x_{(j)})$, as found previously, and (3.3.13) we find that

$$E(y_{[j]})=E[E(y_{[j]}|x_{(j)})]$$

$$=\frac{1}{b}[(1+\frac{\alpha_1}{2})-\alpha_1 E\{\exp(-ax_{(j)})\}] \quad . \qquad (3.4.145)$$

Since

$$E\{\exp(-ax)\}=E\{1-z\}=1-p_j \quad , \qquad (3.4.146)$$

it follows that

$$E(y_{[j]})=\frac{1}{b}[1+\frac{\alpha_1}{2}(2p_j-1)]=\mu_y[1+\frac{\alpha_1}{2}(2p_j-1)] \quad . \qquad (3.4.147)$$

Using (3.4.142)

$$\text{Var}[E(y_{[j]}|x_{(j)})] = \frac{\alpha_1^2}{b^2}\text{Var}\{\exp(-ax_{(j)})\} \quad . \qquad (3.4.148)$$

Now

$$\text{Var}\{\exp(-ax_{(j)})\} = \text{Var}\{1-z\} = \text{Var}(z) = \frac{p_j(1-p_j)}{n+2} \quad . \qquad (3.4.149)$$

Thus

$$\text{Var}[E(y_{[j]}|x_{(j)})] = \alpha_1^2\sigma_y^2\left[\frac{p_j(1-p_j)}{n+2}\right] \quad . \qquad (3.4.150)$$

Also from (3.4.144), (3.4.146) and (3.4.149)

$$E[\text{Var}(y_{[j]}|x_{(j)})] = \sigma_y^2 + \frac{1}{b^2}\left[\frac{\alpha_1}{2}(1-\frac{\alpha_1}{2})\right.$$

$$\left.-\alpha_1(1-\alpha_1)(1-p_j)-\alpha_1^2(1-p_j)(1-\frac{j}{n+2})\right] \qquad (3.4.151)$$

and finally from (3.3.14)

$$\text{Var}(y_{[j]}) = \sigma_y^2\left[1+\frac{(2p_j-1)\alpha_1}{2}+\frac{\{4p_j(1-p_j)-1\}\alpha_1^2}{4}\right] . \qquad (3.4.152)$$

Using (3.4.142) once more,

$$\text{Cov}[E(y_{[j]}|x_{(j)}),E(y_{[j']}|x_{(j')})]$$

$$= \frac{\alpha_1^2}{b}\text{Cov}[\exp(-ax_{(j)}),\exp(-ax_{(j')})] \quad . \qquad (3.4.153)$$

Again, using the joint density of $z=F_1(x_{(j)})$ and $z'=F_1(x_{(j')})$ found previously, and since $\exp(-ax_{(j)})=1-z$ and $\exp(-ax_{(j')})=1-z'$,

$$\text{Cov}[\exp(-ax_{(j)}),\exp(-ax_{(j')})]$$

$$=\text{Cov}(1-z,1-z')=\text{Cov}(z,z')$$

$$= \frac{p_j(1-p_{j'})}{n+2} \quad . \tag{3.4.154}$$

Finally, from Theorem 1

$$\mathrm{Cov}(y_{[j]},y_{[j']}) = \alpha_1^2\sigma_y^2\left[\frac{p_j(1-p_{j'})}{n+2}\right] \quad . \tag{3.4.155}$$

Thus for the proposed method from (3.3.11) and (3.3.12) as well as (3.4.147), (3.4.152) and (3.4.155), we find that

$$E(\bar{y}_p^*) = \mu_y[1+ \frac{\alpha_1}{2}\sum\delta_g(2p_{s_g+t}-1)] \tag{3.4.156}$$

and

$$\mathrm{Var}(\bar{y}_p^*) = \sigma_y^2\left[\sum\delta_g^2\left\{1+ \frac{(2p_{s_g+t}-1)\alpha_1}{2} + \frac{(4p_{s_g+t}(1-p_{s_g+t})-1)\alpha_1^2}{4}\right\} + \sum\sum{}'\delta_g\delta_{g'}\alpha_1^2\left\{\frac{p_{s_g+t}(1-p_{s_{g'}+t'})}{n+2}\right\}\right] \quad . \tag{3.4.157}$$

From (3.3.23) and (3.4.142) we obtain

$$\bar{y}_o^* \doteq \mu_y[1+ \frac{\alpha_1}{2} - \frac{\alpha_1}{n^*}\sum_{i=1}^{n^*}\exp(-ax_i^*)] \quad . \tag{3.4.158}$$

Using the results in (3.4.156)-(3.4.158), we find that

$$\mathrm{Bias}(\bar{y}_p^*) \doteq \mu_y[\frac{\alpha_1}{2}\{\sum\delta_g(2p_{s_g+t}-1)-1\} + \frac{\alpha_1}{n^*}\sum_{i=1}^{n^*}\exp(-ax_i^*)] \quad . \tag{3.4.159}$$

Finally, substituting (3.4.157) and (3.4.159) into (3.3.25) yields the $\mathrm{MSE}(\bar{y}_p^*)$.

For the Ericksen estimate we use Slutsky's Theorem and the results of (3.4.66) and (3.4.67) so that

$$E(\hat{\beta}) \doteq E(\hat{\beta}') = \frac{1}{n\sigma_x^2} E\left[\sum_{i=1}^{n} (x_i - \bar{x}) E(y_i | x_i)\right]$$

$$= \left(\frac{n-1}{n}\right) \frac{\alpha_1}{4} \frac{\sigma_y}{\sigma_x} \quad . \tag{3.4.160}$$

Similarly,

$$E(\hat{\alpha}) \doteq E(\hat{\alpha}') = E[E(\bar{y} - \hat{\beta}'\bar{x} | \underset{\sim}{x})]$$

$$= \frac{1}{n\sigma_x^2} E\left[\sum_{i=1}^{n} (\sigma_x^2 - x_i\bar{x} + \bar{x}^2) E(y_i | x_i)\right]$$

$$= \mu_y \left[1 - \frac{\alpha_1}{4}\left(\frac{n^2-1}{n^2}\right)\right] \quad . \tag{3.4.161}$$

Thus

$$E(\bar{y}_E^*) = E(\hat{\alpha}) + \bar{x}^* E(\hat{\beta})$$

$$= \mu_y + \frac{\alpha_1}{4}\left[\left(\frac{n-1}{n}\right)\frac{\sigma_y}{\sigma_x}\bar{x}^{*2} - \left(\frac{n^2-1}{n^2}\right)\mu_y\right] \quad . \tag{3.4.162}$$

Using (3.3.47) and (3.4.158),

$$\text{Bias}(\bar{y}_E^*) \doteq \frac{\alpha_1}{4}\left[\left(\frac{n-1}{n}\right)\frac{\sigma_y}{\sigma_x}\bar{x}^* - \left(\frac{n^2-1}{n^2}\right)\mu_y\right]$$

$$- \mu_y\left[\frac{\alpha_1}{2} - \frac{\alpha_1}{n^*}\sum_{i=1}^{n^*} \exp(-ax_i^*)\right] \quad . \tag{3.4.163}$$

Once again, using the results of (3.4.73) and (3.4.74),

$$E[\text{Var}(\hat{\beta} | \underset{\sim}{x})] \doteq E[\text{Var}(\hat{\beta}'' | \underset{\sim}{x})]$$

$$= \frac{1}{n^2\sigma_x^2} E\left[\sum_{i=1}^{n} (x_i - \mu_x)^2 \text{Var}(y_i | x_i)\right]$$

$$= \frac{\sigma_y^2}{n\sigma_x^2}\left[1+ \frac{\alpha_1}{4} - \frac{5}{27}\alpha_1^2\right] \quad . \qquad (3.4.164)$$

Also

$$\mathrm{Var}[E(\hat{\beta}|\underset{\sim}{x})] \doteq \mathrm{Var}[E(\hat{\beta}''|\underset{\sim}{x})]$$

$$= \frac{1}{n^2\sigma_x^4}\mathrm{Var}\left[\sum_{i=1}^{n}(x_i-\mu_x)E(y_i|x_i)\right]$$

$$= \frac{\sigma_y^2}{n\sigma_x^2}\left[2+ \frac{3}{4}\alpha_1 - \frac{1}{16}\alpha_1^2\right] \quad . \qquad (3.4.165)$$

Thus

$$\mathrm{Var}(\hat{\beta}) \doteq \frac{\sigma_y^2}{n\sigma_x^2}\left[2+ \frac{3}{4}\alpha_1 - \frac{1}{16}\alpha_1^2\right] \quad . \qquad (3.4.166)$$

Similarly,

$$E[\mathrm{Var}(\hat{\alpha}|\underset{\sim}{x})] \doteq E[\mathrm{Var}(\hat{\alpha}''|\underset{\sim}{x})]$$

$$= \frac{1}{n^2\sigma_x^4}E\left[\sum_{i=1}^{n}(\sigma_x^2+\mu_x^2-\mu_x x_i)^2\mathrm{Var}(y_i|x_i)\right]$$

$$= \frac{\sigma_y^2}{n}\left[2- \frac{1}{4}\alpha_1 - \frac{23}{108}\alpha_1^2\right] \qquad (3.4.167)$$

and

$$\mathrm{Var}[E(\hat{\alpha}|\underset{\sim}{x})] \doteq \mathrm{Var}[E(\hat{\alpha}''|\underset{\sim}{x})]$$

$$= \frac{1}{n^2\sigma_x^4}\mathrm{Var}\left[\sum_{i=1}^{n}(2\mu_x^2-\mu_x x_i)E(y_i|x_i)\right]$$

$$= \frac{\sigma_y^2}{n}\left[1+ \frac{65}{432}\alpha_1^2\right] \quad . \qquad (3.4.168)$$

Thus

$$\mathrm{Var}(\hat{\alpha}) \doteq \frac{\sigma_y^2}{n}[3 - \frac{1}{4}\alpha_1 + \frac{1}{16}\alpha_1^2] \qquad (3.4.169)$$

Also

$$\mathrm{Cov}[E(\hat{\alpha}|\underset{\sim}{x}), E(\hat{\beta}|\underset{\sim}{x})] \doteq \mathrm{Cov}[E(\hat{\alpha}''|\underset{\sim}{x}), E(\hat{\beta}''|\underset{\sim}{x})]$$

$$= \frac{1}{n^2\sigma_x^4}\mathrm{Cov}[\sum_{i=1}^{n}(2\mu_x^2 - \mu_x x_i)E(y_i|x_i),$$

$$\sum_{j=1}^{n}(x_j - \mu_x)E(y_j|x_j)]$$

$$= -\frac{\sigma_y^2}{n\sigma_x}[1 + \frac{1}{4}\alpha_1 + \frac{41}{432}\alpha_1^2] \qquad (3.4.170)$$

and

$$E[\mathrm{Cov}(\hat{\alpha}, \hat{\beta}|\underset{\sim}{x})] \doteq E[\mathrm{Cov}(\hat{\alpha}'', \hat{\beta}''|\underset{\sim}{x})]$$

$$= E[\mathrm{Cov}(\bar{y}, \hat{\beta}''|\underset{\sim}{x}) - \mu_x \mathrm{Var}(\hat{\beta}''|\underset{\sim}{x})]$$

$$= \frac{1}{n^2\mu_x^3}E[\sum_{i=1}^{n}(3\mu_x x_i - 2\mu_x^2 - x_i^2)\mathrm{Var}(y_i|x_i)]$$

$$= -\frac{\sigma_y^2}{n\sigma_x}[1 - \frac{68}{432}\alpha_1^2] \quad . \qquad (3.4.171)$$

Finally,

$$\mathrm{Cov}(\hat{\alpha}, \hat{\beta}) \doteq -\frac{\sigma_y^2}{n\sigma_x}[2 + \frac{1}{4}\alpha_1 - \frac{1}{16}\alpha_1^2] \quad , \qquad (3.4.172)$$

and from (3.3.41), (3.4.166), (3.4.169) and (3.4.172) we obtain

$$\mathrm{Var}(\bar{y}_E^*) \doteq \frac{\sigma_y^2}{n}\left[\left\{3+\frac{2\bar{x}^{*2}}{\mu_x^2}-\frac{4\bar{x}^*}{\mu_x}\right\}\right.$$

$$\left.-\left\{1-\frac{3\bar{x}^{*2}}{\mu_x^2}+\frac{2\bar{x}^*}{\mu_x}\right\}\frac{\alpha_1}{4}-\left\{1+\frac{\bar{x}^{*2}}{\mu_x^2}-\frac{2\bar{x}^*}{\mu_x}\right\}\frac{\alpha_1^2}{16}\right] \quad . \qquad (3.4.173)$$

The $\mathrm{MSE}(\bar{y}_E^*)$ is found by substituting (3.4.163) and (3.4.173) into (3.3.48).

3.5 Tabulation of analytic results

Results based upon the formulation in Section 3.4 are presented in Tables 3.2-3.5. Each table deals with one of the bivariate distributions discussed earlier. The input set of parameter values (i.e., n, k, n^*, etc.) is the same for each of the tables.

The results are a small portion of those which could be obtained if all combinations of the parameters, presented in Table 3.1, were used. For clarity and simplicity we present only enough cases to give some indication as to how the two methods compare in general and, more specifically, when the proposed method gives better results. Also it should be noted that, with two exceptions, n and k are small. This is due to the expense involved in programming these results and must be kept in mind when considering that approximations are involved. Finally, the input values for the values of X in the target area are found by dividing the range into n^* points

TABLE 3.1

RANGES OF PARAMETER VALUES USED IN COMPUTING THE ANALYTIC RESULTS

Parameter	Low	High
n	30	60
k	3	6
n*	15	30
Percentile X range in target area	(50,80)	(20,80)
Association level	0.75	0.95
σ_y	5	10

TABLE 3.2

A SUMMARY OF TABULATED ANALYTIC RESULTS
FOR THE NORMAL DISTRIBUTION

Entry	n	k	n*	σ_y [a]	Target area percentile range on X	ρ	Variance		Bias		Mean squared error	
							Proposed	Ericksen	Proposed	Ericksen	Proposed	Ericksen
(1)	120	20	100	5	(5,95)	0.95	0.21	0.02	0.08	0.00	0.22	0.02
(2)	120	20	100	5	(50,95)	0.95	0.08	0.02	0.10	0.00	0.09	0.02
(3)	30	3	30	5	(50,80)	0.95	0.97	0.09	0.31	0.00	1.06	0.09
(4)	60	3	30	5	(50,80)	0.95	0.48	0.05	0.15	0.00	0.51	0.05
(5)	60	6	30	5	(50,80)	0.95	0.57	0.05	0.18	0.00	0.60	0.05
(6)	30	3	15	5	(50,80)	0.95	0.97	0.09	0.32	0.00	1.07	0.09
(7)	30	3	15	10	(50,80)	0.95	3.89	0.38	0.64	0.00	4.29	0.38
(8)	60	6	30	5	(50,80)	0.75	0.71	0.21	0.14	0.00	0.73	0.21
(9)	30	3	15	5	(20,80)	0.75	0.91	0.36	0.21	0.00	0.95	0.36
(10)	30	3	15	5	(50,80)	0.75	1.05	0.42	0.25	0.00	1.12	0.42

[a] $\sigma_x=10$; $\mu_x=50$; $\mu_y=60$.

TABLE 3.3

A SUMMARY OF TABULATED ANALYTIC RESULTS
FOR THE LOGISTIC TYPE 1 DISTRIBUTION

Entry	n	k	n*	σ_y [a]	Target area percentile range on X	Variance		Bias		Mean squared error	
						Proposed	Ericksen	Proposed	Ericksen	Proposed	Ericksen
(1)	120	20	100	5	(5,95)	0.32	0.21	-0.01	-0.28	0.32	0.28
(2)	120	20	100	5	(50,95)	0.19	12.12	0.01	-0.30	0.19	12.21
(3)	30	3	30	5	(50,80)	1.51	18.46	-0.19	-0.62	1.54	18.85
(4)	60	3	30	5	(50,80)	0.77	9.23	-0.22	-0.64	0.81	9.64
(5)	60	6	30	5	(50,80)	1.16	9.23	-0.03	-0.64	1.16	9.64
(6)	30	3	15	5	(50,80)	1.50	18.56	-0.19	-0.61	1.54	18.94
(7)	30	3	15	10	(50,80)	6.02	23.47	-0.38	-1.23	6.16	24.98
(8)	60	6	30	5	(50,80)	1.16	9.23	-0.03	-0.64	1.16	9.64
(9)	30	3	15	5	(20,80)	1.34	0.83	-0.25	-0.56	1.41	1.15
(10)	30	3	15	5	(50,80)	1.50	18.56	-0.19	-0.61	1.54	18.94

[a] $\sigma_x=10$; $\mu_x=50$; $\mu_y=60$.

TABLE 3.4

A SUMMARY OF TABULATED ANALYTIC RESULTS
FOR THE LOGISTIC TYPE 2 DISTRIBUTION

Entry	n	k	n*	σ_y [a]	Target area percentile range on X	α_o	Variance		Bias		Mean squared error	
							Proposed	Ericksen	Proposed	Ericksen	Proposed	Ericksen
(1)	120	20	100	5	(5,95)	0.95	0.23	0.21	0.02	-0.01	0.23	0.21
(2)	120	20	100	5	(50,95)	0.95	0.18	12.25	0.01	-0.31	0.18	12.35
(3)	30	3	30	5	(50,80)	0.95	1.21	18.12	-0.06	-0.38	1.21	18.27
(4)	60	3	30	5	(50,80)	0.95	0.62	9.06	-0.11	-0.30	0.63	9.15
(5)	60	6	30	5	(50,80)	0.95	0.96	9.06	0.02	-0.30	0.96	9.15
(6)	30	3	15	5	(50,80)	0.95	1.21	18.22	-0.05	-0.38	1.21	18.36
(7)	30	3	15	10	(50,80)	0.95	4.83	23.06	-0.11	-0.76	4.84	23.64
(8)	60	6	30	5	(50,80)	0.75	0.95	8.85	0.02	-0.24	0.95	8.90
(9)	30	3	15	5	(20,80)	0.75	1.01	0.83	0.07	-0.07	1.01	0.84
(10)	30	3	15	5	(50,80)	0.75	1.20	17.78	-0.04	-0.30	1.20	17.88

[a] σ_x=10 ; μ_x= 50 ; μ_y=60.

TABLE 3.5

A SUMMARY OF TABULATED ANALYTIC RESULTS
FOR THE EXPONENTIAL DISTRIBUTION

Entry	n	k	n*	σ_y[a]	Target area percentile range on X	α_1	Variance		Bias		Mean squared error	
							Proposed	Ericksen	Proposed	Ericksen	Proposed	Ericksen
(1)	120	20	100	5	(5,95)	0.95	0.23	0.19	0.02	-0.13	0.23	0.21
(2)	120	20	100	5	(50,95)	0.95	0.28	0.40	-0.01	-0.56	0.27	0.72
(3)	30	3	30	5	(50,80)	0.95	2.01	0.92	-0.03	-0.65	2.01	1.34
(4)	60	3	30	5	(50,80)	0.95	1.14	0.46	-0.03	-0.63	1.15	0.86
(5)	60	6	30	5	(50,80)	0.95	1.16	0.46	-0.01	-0.63	1.16	0.86
(6)	30	3	15	5	(50,80)	0.95	1.97	0.92	-0.03	-0.65	1.98	1.34
(7)	30	3	15	10	(50,80)	0.95	7.90	3.68	-0.07	-1.30	7.90	5.37
(8)	60	6	30	5	(50,80)	0.75	1.12	0.45	-0.00	-0.50	1.12	0.70
(9)	30	3	15	5	(20,80)	0.75	0.96	0.80	-0.06	-0.24	0.97	0.86
(10)[b]	30	3	15	5	(50,80)	0.75	1.91	0.90	-0.03	-0.51	1.91	1.17
(11)[b]	30	3	15	5	(50,80)	0.75	1.32	0.90	-0.03	-0.51	1.32	1.17

[a] $\mu_x=\mu_y=\sigma_x=5$.

[b] Optimum stratification points for the normal distribution were used to obtain the group boundaries.

which are equally spaced in probability. The estimates of the target area mean is found for each distribution as in Section 2.4.

Our format will be to state some general conclusions as well as to point out more specific tendencies of the methods relative to the distributions. First of all, it is generally true that the proposed and Ericksen methods give their best results when n is large. Second, an increase in σ_y appears to increase the variance of both estimates while the bias increases also and approximately in direct proportion to σ_y. Third, the Ericksen method gives a smaller variance for all distributions when the range on X is symmetric about the 50-th percentile. The difference of the variances is smaller for the non-normal distributions, however. Fourth, for all but the normal distribution, the bias of the Ericksen estimate is larger than the bias of the proposed method. Fifth, it appears that n/k exerts a stronger influence on the variance than does k alone, since, based upon entries (4) and (5) in each table, the variance decreases from (4) to (5) while k has increased and n/k has decreased. This conclusion must be viewed with caution since both n and k are small. Sixth, for both methods the effect of a change in n^* appears to be small when all else is held constant. This is based upon a comparison of entries (3) and (6) in each table.

Seventh, the effect of a lower level of association usually results in a corresponding increase in the variance. This is not always true due perhaps to the approximations involved. Eighth, with the logistic distributions the variance of the Ericksen method becomes very large in comparison to the proposed method when the values of X in the target area are extreme. The bias of the Ericksen estimate also is greater under these circumstances. One final observation involves a comparison of entries (10) and (11) in Table 3.5. Contrary to expectation the result where the group boundaries have been found using the optimum stratification points on the exponential distribution gives a higher variance than the one which uses the stratification points for the normal distribution.

We conclude by stating that, for these distributions and parameters, the proposed estimate gives better results than the Ericksen estimate when (1) regression is nonlinear, (2) association is either high or moderate, and (3) the range of values on X in the target area is not symmetric about the mean. In other instances involving nonlinearity the Ericksen estimates are less variable provided the target area X values are symmetric about the mean. This would imply that for distributions such as these, the Ericksen method would be best used if the target area base unit is "typical" with respect to all local

(or target) areas. Finally, as expected the Ericksen method does very well with the bivariate normal distribution due to the linear regression which it assumes.

CHAPTER IV

MULTIVARIATE EXTENSIONS BY MONTE CARLO SIMULATIONS

4.1 Introduction

In this chapter we compare the Ericksen and proposed methods by Monte Carlo simulation in a multivariate setting. Specifically, we consider the situation where we have two symptomatic variables and a criterion variable with a prespecified distribution and corresponding regression figure. The Monte Carlo simulation enables us to investigate the comparative behavior of the methods under somewhat more relevant conditions than were studied analytically in Chapter III.

Our approach is to consider two basic models. The first is what we call the linear model, wherein the regression figure of the criterion variable on the symptomatic variables assumes a linear functional form (i.e., a two-dimensional plane in the case of two symptomatic variables). The other is a nonlinear model in which the regression figure assumes some nonlinear functional form. In both cases we carefully specify the distribution and parameters in advance. Then by simulating a large number of samples from these distributions and performing the

method on each sample, we obtain estimates of the parameters of both methods. This approach allows us the opportunity to explore several interesting and illustrative hypothetical situations. Since we speculate that the proposed method will do best in nonlinear settings, these simulations allow us to create such settings and to observe the results.

4.2 Discussion of the models

Notationally, we designate Y as the criterion variable and X_1 and X_2 as the symptomatic variables. Furthermore, define

$$\underset{\sim}{Z}=\begin{pmatrix}\underset{\sim}{X}\\ Y\end{pmatrix} \tag{4.2.1}$$

with specific value

$$z=\begin{pmatrix}\underset{\sim}{x}\\ y\end{pmatrix} \tag{4.2.2}$$

where

$$\underset{\sim}{X}=\begin{pmatrix}X_1\\ X_2\end{pmatrix} \tag{4.2.3}$$

has specific value

$$x=\begin{pmatrix}x_1\\ x_2\end{pmatrix} \quad . \tag{4.2.4}$$

We represent the mean vector of Z as

$$\underset{\sim}{U}=\begin{pmatrix}\underset{\sim}{U}_x\\ U_y\end{pmatrix} \tag{4.2.5}$$

and the covariance matrix as

$$V=\begin{pmatrix} V_{xx} & V_{xy} \\ V_{xy}' & V_{yy} \end{pmatrix} \tag{4.2.6}$$

where

$$\underset{\sim}{U}_x=\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} \; ; \; U_y=\mu_y \tag{4.2.7}$$

and

$$V_{xx}=\begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{pmatrix} \; ; \; V_{xy}=\begin{pmatrix} \sigma_{1y} \\ \sigma_{2y} \end{pmatrix} \; ; \; V_{yy}=\sigma_{yy} \tag{4.2.8}$$

and the subscripts 1 and 2 refer to the X's. We also denote the conditional expectation as $E(Y|x_1,x_2)$ and the conditional variance as $\sigma_{y.12}^2 \equiv Var(Y|x_1,x_2)$. In general, we express both models as

$$Y=E(Y|x_1,x_2)+\varepsilon \tag{4.2.9}$$

where it is assumed that $E(\varepsilon)=0$, $Var(\varepsilon)=\sigma_{y.12}^2$, and the ε's are uncorrelated.

As in Chapter III the most obvious multivariate distribution to use for the linear model is the normal, specifically the trivariate normal distribution here. With the parameters as defined above we have

$$E(Y|x_1,x_2)=\beta_0+\beta_1X_1+\beta_2X_2 \tag{4.2.10}$$

with

$$\beta_0=\mu_y-V_{xy}'V_{xx}^{-1}\underset{\sim}{U}_x \; ; \; (\beta_1,\beta_2)=V_{xy}'V_{xx}^{-1} \tag{4.2.11}$$

and

$$\sigma_{y.12}^{2}=V_{yy}-V_{xy}'V_{xx}^{-1}V_{xy} \quad . \tag{4.2.12}$$

The magnitude of the multiple correlation coefficient, R, can be expressed as

$$R=\left(\frac{V_{xy}'V_{xx}^{-1}V_{xy}}{V_{yy}}\right)^{1/2} \quad . \tag{4.2.13}$$

For our nonlinear model we again let $\underset{\sim}{X}$ be distributed as a bivariate normal with mean, $\underset{\sim}{U}_x$, and variance, V_{xx}, but then we specify that the distribution of Y given $\underset{\sim}{x}$ be a logistic such that

$$F(Y|x_1,x_2)=[1+\exp\{-(a+by)\}]^{-1} \quad , \quad -\infty<y<\infty \quad , \tag{4.2.14}$$

and thus

$$E(Y|x_1,x_2)=a/b$$

with

$$a=\frac{-[60-c\{0.5-[1+\exp\{-(\alpha+\beta X_1+\gamma X_2)\}]\Pi}{\sigma_{y.12}\sqrt{3}} \tag{4.2.15}$$

and

$$b=\frac{\Pi}{\sigma_{y.12}\sqrt{3}} \tag{4.2.16}$$

where c, α, β and γ are arbitrary constants which determine the shape of the regression figure. Geometrically, with $\beta=\gamma$, the figure can be pictured as a sheet of paper, which, when viewed from the side about the line $X_1=X_2$, assumes the shape of an ordinary logistic curve. Allowing $\beta\neq\gamma$

serves only to distort the figure somewhat.

Selection of the nonlinear model is done with the idea that it be somewhat realistic. There is no doubt that an even "more nonlinear" model might be selected such as some parabolic function or one involving absolute values in which curve-fitting would most assuredly do poorly. However, with little solid evidence as to the relevance of such situations, we select the logistic model which is a more common one in nature (e.g., in demography; Pearl [1925]). This is not to deny the existence of other more complex models, but we delay consideration of other, possibly more nonlinear, models to the examples of Chapter V where real data are used. Finally, this model is relevant since it describes a situation where the Ericksen method might be used erroneously, since the distributions of the variables are similar to the normal distribution but the regression figure is nonlinear.

4.3 General simulation procedure

Without going into great detail, the principle of research by simulation, where we wish to investigate some properties of a statistic, involves (1) the generation of a sample from a prespecified distribution or under prespecified circumstances, (2) the computation of the statistic from that sample, and (3) the repetition of (1) and (2) a large number of times. The number of repe-

titions involved is determined by such things as the complexity of the statistic, the nature of the underlying distribution, the size of the sample, and the amount of available computing funds. In general, sample observations are obtained by the generation and transformation of pseudo-random uniformly distributed numbers. A more detailed discussion of this phenomenon as well as a specific description of the procedure we use is taken up in Appendix D.

In terms of the discussion carried on in Chapter II the three phases in the proposed procedure are done as follows:

(1) Grouping--Since simulations involving even relatively simple statistics can be expensive, simplicity is of utmost importance. To meet this condition we define boundaries on each variable in a hierarchical scheme (see Appendix C), which involves the optimum stratification points (OSP's) of the underlying distributions. Thus the group boundaries are rectilinear and determined in advance using the parameters of the distributions involved. On the surface this may appear to put the proposed procedure at an advantage. However, this is most likely not the case since, as a disadvantage, the grouping is done in all instances using symptomatic information only. Use of an algorithm like AID or some other clustering procedure which uses criterion information would often produce better results.

(2) Classification--With rectilinear group boundaries the natural preference in the classification phase is "pigeonholing," since, in most real life instances, it would be used in this setting. It also carries with it a simplicity of implementation which increases computing time efficiency.

(3) Estimation--Since the sampling is done by simple random selection from a hypothetically infinite population, our base units are the observations and thus $W_{ghc}=1$, $W_{gh}=1/n_g$, and $W_g=n_g*/n*$, where the asterisk (*) superscript refers to the target area as before.

The Ericksen estimate in all settings is obtained, assuming a linear model, by the method of ordinary least squares. Estimates of the coefficients obtained from the sample are then used to obtain an estimate for the target area observations which are fixed and known in advance.

The target area units are designated in the following manner: One of the input parameters to the program is the range of X_1* and X_2* values (in terms of the underlying marginal distributions of X_1 and X_2). The range is represented by lower and upper percentiles on the two distributions. Then, depending on the value of n* which is the even square of an integer, a grid of values in X_1 and X_2 are determined by evenly spaced percentiles over the designated range of the two variables. This process is illustrated in Figure 4.1 for the case where $n*=5^2=25$

and the lower and upper percentiles on both variables are, say, 20 and 80. This means that the lower left most dot is at the 20-th percentile on both X_1 and X_2, the middle leftmost dot is the 50-th percentile on X_1 and the 20-th percentile on X_2, and so forth. Once these points are established the corresponding values of Y for the target area are found from $E(Y|x_1,x_2)$ particular to the model and are completely determined by our procedure for specifying n* and the grid in Figure 4.1. Thus the true mean of the criterion variable for the target area (which we are attempting all along to estimate) remains fixed. This allows us to estimate, for both the Ericksen and proposed methods, the biases which are obtained in both cases as the estimate over all repetitions. The estimates of the mean squared error are found using the variance and bias in each case.

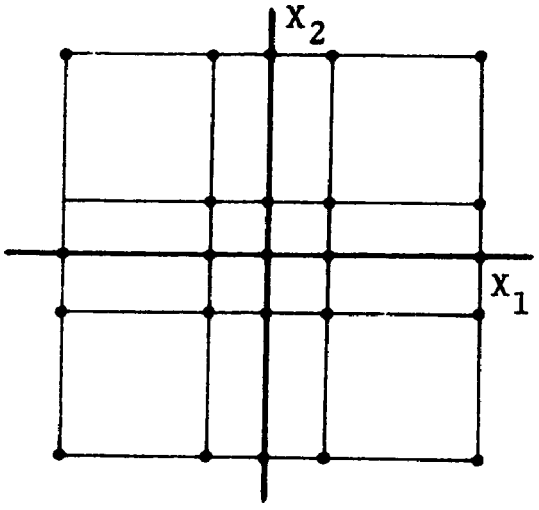


Fig. 4.1.--Representation of a grid of target area values on two symptomatic variables.

4.4 Discussion of program input parameters

It is our intention to deal with two situations for each model, one with high association and the other with moderate association. Variation in the value of V allows us to vary the association level in the linear model in which we interpret R=0.95 as being high association and R=0.82 as being moderate. Since the multivariate distribution generated for the nonlinear model is determined implicitly from the joint distribution on $\underset{\sim}{X}$ and the conditional distribution of Y given $\underset{\sim}{x}$, the interpretable level of association (i.e., R) is altered by changing $E(Y|x_1,x_2)$ and $\sigma_{y.12}{}^2$. In terms of the parameters we use, high association in the nonlinear model corresponds to a multiple correlation value of about 0.80 and moderate association corresponds to a level of about 0.50.

Specific values used for the situations described above are now given. To express all parameters in the same units we use $\sigma_{y.12}$ instead of $\sigma_{y.12}{}^2$.

(1) Linear model--high association

$$\underset{\sim}{U}=\begin{bmatrix}50\\50\\60\end{bmatrix} \quad ; \quad V=\begin{bmatrix}25.0 & 15.0 & 42.5\\15.0 & 25.0 & 42.5\\42.5 & 42.5 & 100.0\end{bmatrix} \qquad (4.4.1)$$

which implies that

$$R=0.95 \; ; \; \sigma_{y.12}=3.16 \; ; \; \beta_0=-46.25 \; ; \; \beta_1=\beta_2=1.06 \quad . \qquad (4.4.2)$$

(2) Linear model--moderate association

$$\underset{\sim}{U}=\begin{bmatrix}50\\50\\60\end{bmatrix} \; ; \; V=\begin{bmatrix}25.0 & 15.0 & 36.5\\15.0 & 25.0 & 36.5\\36.5 & 36.5 & 100.0\end{bmatrix} \qquad (4.4.3)$$

which implies that

$$R=0.82 \; ; \; \sigma_{y.12}=5.74 \; ; \; \beta_0=-31.25 \; ; \; \beta_1=\beta_2=0.91 \quad . \qquad (4.4.4)$$

(3) Nonlinear model--high association

$$\underset{\sim}{U}=\begin{bmatrix}50\\50\\60^{o}\end{bmatrix} \; ; \; V=\begin{bmatrix}25 & 15 & 25^{o}\\15 & 25 & 25^{o}\\25^{o} & 25^{o} & 45^{o}\end{bmatrix} \; ; \; \sigma_{y.12}=0.32 \qquad (4.4.5)$$

with

$$c=14 \; ; \; \alpha=-200 \; ; \; \beta=3.5 \; ; \; \gamma=0.5 \qquad (4.4.6)$$

which implies a level of association that corresponds to $R\doteq 0.82$. The superscript "o" on elements in $\underset{\sim}{U}$ and V indicates that those values have been obtained empirically and are not input parameters.

(4) Nonlinear model--moderate association

$$\underset{\sim}{U}=\begin{bmatrix}50\\50\\60^{o}\end{bmatrix} \; ; \; V=\begin{bmatrix}25 & 15 & 25^{o}\\15 & 25 & 25^{o}\\25^{o} & 25^{o} & 90^{o}\end{bmatrix} \; ; \; \sigma_{y.12}=7.00 \qquad (4.4.7)$$

with

$$c=14 \; ; \; \alpha=-200 \; ; \; \beta=3.5 \; ; \; \gamma=0.5 \quad . \qquad (4.4.8)$$

which implies a level of association that corresponds to $R\doteq 0.50$.

For all runs the range of values assumed by X_1^*

and X_2^* have a lower percentile of 5 and upper percentile of 65 while two different values for n* are used. One is with n*=49 (or 7 per variable) and the other is n*=100 (or 10 per variable).

Sample sizes assume two values: n=120 and n=480. Relatively large samples are used since it is felt that in most instances where the procedure is applicable the sample size, based on a major survey, will be large. Monetary constraints prevent a large number of repetitions. However, to best utilize available resources, we allow more repetitions with n=120 than with n=480, since intuition dictates that the estimate is less variable with a larger sample and therefore the parameter is estimable at the same level of accuracy with few repetitions. As a result we use 400 repetitions when n=120 and 100 repetitions when n=480.

As mentioned previously grouping is done by a priori definition of boundaries using a hierarchical scheme involving optimum stratification points. Using the notation of Appendix C, where (S_1,S_2) defines a grouping scheme with S_1 branches on X_1 and S_2 branches on X_2, we indicate in Table 4.1 the values of $k=S_1 \times S_2$ that are used in combination with the sample sizes. By considering these values we allow ourselves a variety on values of n, k, and n/k, all of which may be influential in the quality of the proposed estimate.

TABLE 4.1

SAMPLE SIZES AND NUMBER OF GROUPS
USED IN SIMULATIONS

S_1	S_2	k	n	n/k
2	2	4	120	30.0
3	3	9	120	13.3
4	4	16	120	7.5
2	2	4	480	120.0
3	3	9	480	53.3
4	4	16	480	30.0

TABLE 4.2

OPTIMUM STRATIFICATION POINTS AND BRANCH MEANS
FOR THE STANDARD NORMAL DISTRIBUTION

S_i	Branch Mean 1	OSP 1	Branch Mean 2	OSP 2	Branch Mean 3	OSP 3	Branch Mean 2
2	-0.798	0.0	0.798				
3	-1.178	-0.550	0.0	0.550	1.178		
4	-1.435	-0.885	-0.415	0.0	0.415	0.885	1.435

Although a general discussion of the grouping procedure is found in Appendix C, the results of the procedure for obtaining the specific boundary points are given here. Using the results from Table C.1 and formula (C.12), the OSP's and branch means for the standard normal distribution are given in Table 4.2. These values are used to obtain the final boundaries which are given in Figures 4.2-4.4.

4.5 Presentation and interpretation of results

The results of the simulation run described in the preceding sections are given in Tables 4.3 through 4.10. Before proceeding with a discussion of those results it behooves us to interject a note of caution as to the generality of any conclusions that might be drawn from these data. We must remember that the models, population distributions, and target area have been produced synthetically and that some of the results may be attributed to our particular simulation methodology. This is brought out further as our discussion progresses, but it does not take away the importance of this portion of our research. The situations with which we deal retain, in principle, the essence of both the Ericksen and proposed methods so that meaningful comparisons can be made. It is the magnitude of the results which may not be the same in all situations.

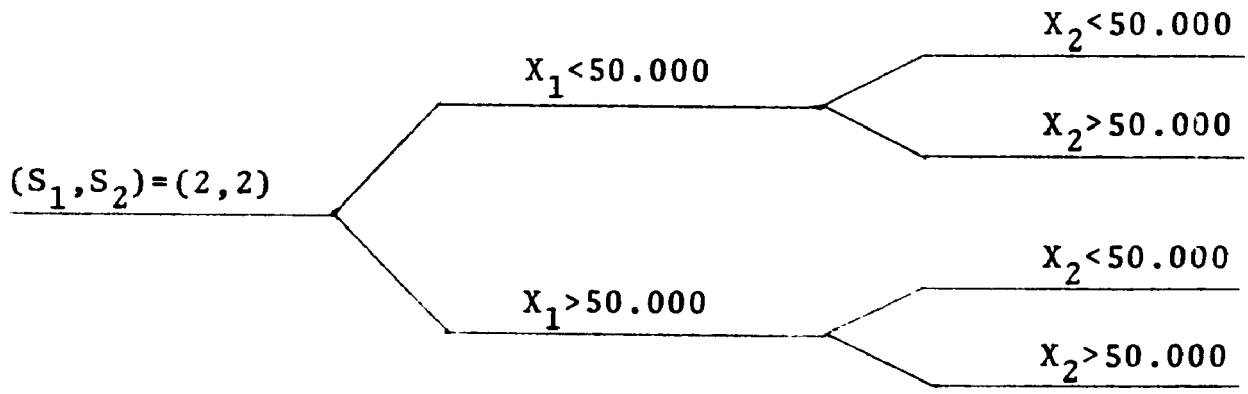


Fig. 4.2.--Boundary points for the four-group simulation model.

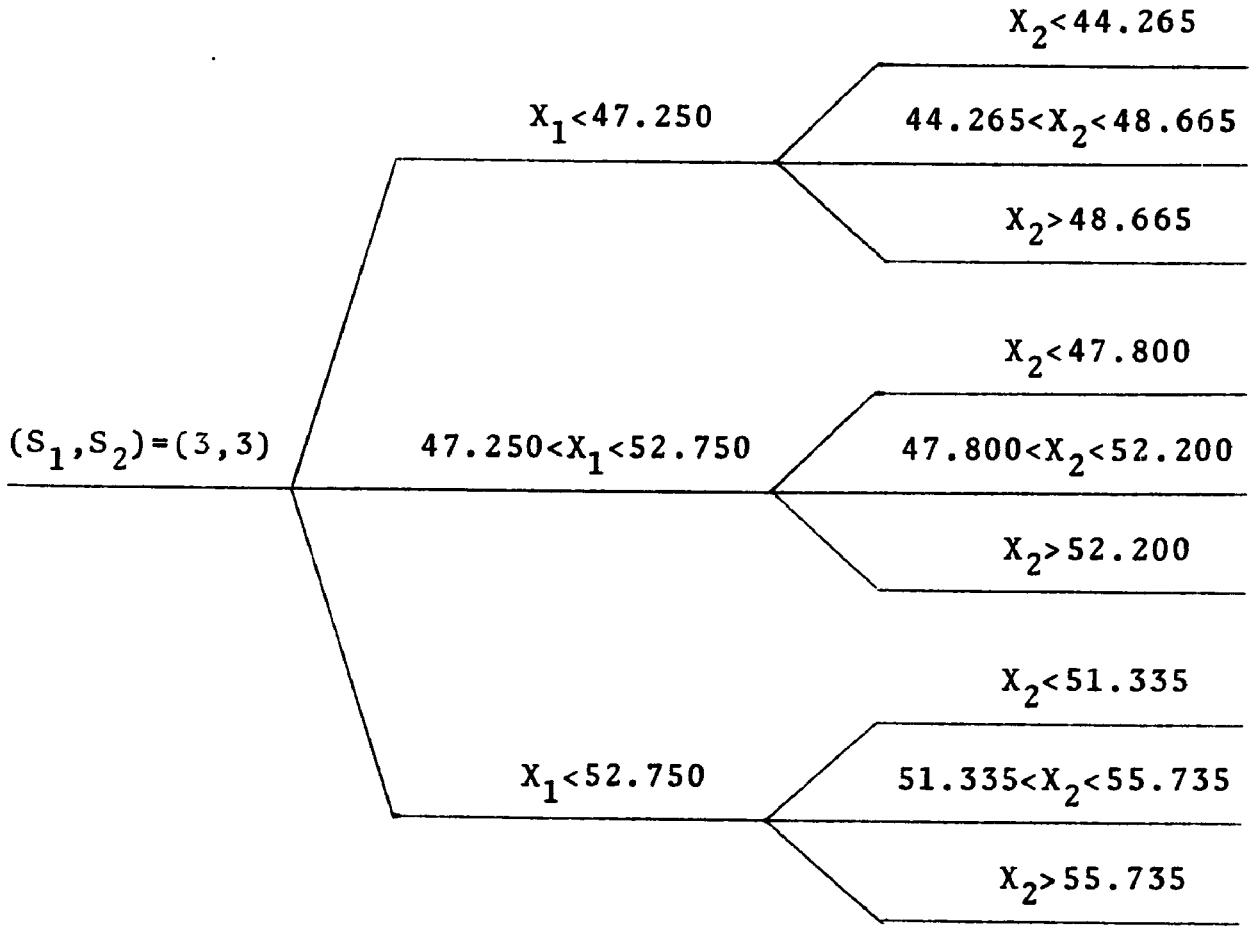


Fig. 4.3.--Boundary points for the nine-group simulation model.

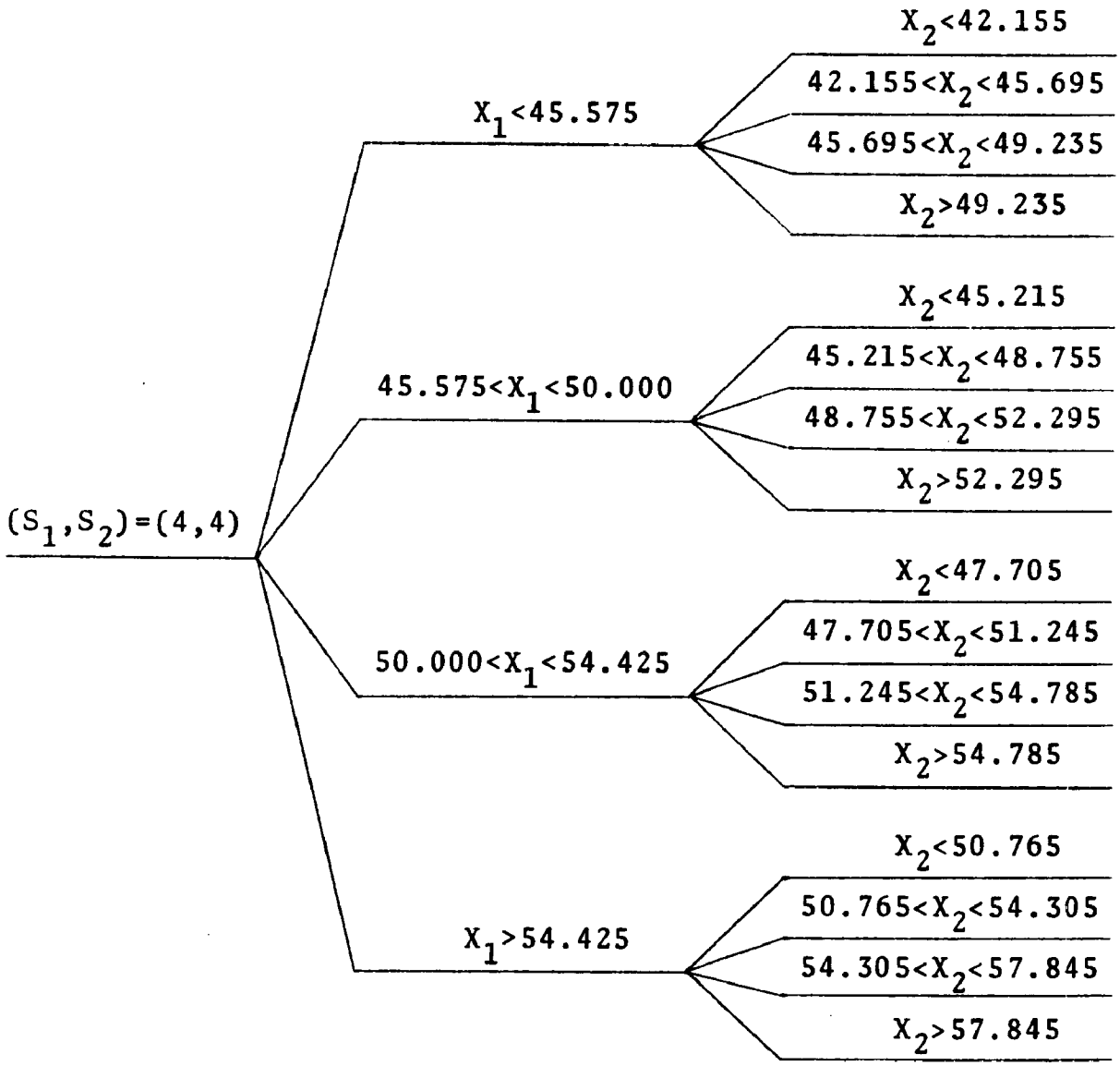


Fig. 4.4.--Boundary points for the sixteen-group simulation model.

TABLE 4.3

ESTIMATED PARAMETERS OF THE ERICKSEN AND PROPOSED METHODS WITH HIGH ASSOCIATION
n=120 ; n*=49

	S_1	S_2	k	n/k	Expectation	Variance	Bias[a]	Mean squared error
Linear model								
Ericksen	-	-	-	-	54.8826	0.0963	-0.0179	0.0966
Proposed	2	2	4	30.0	55.8781	0.3437	0.9776	1.2993
Proposed	3	3	9	13.3	55.1200	0.3011	0.2195	0.3493
Proposed	4	4	16	7.5	54.9032	0.8910	0.0027	0.8910
Nonlinear model								
Ericksen	-	-	-	-	57.1789	0.1626	0.7225	0.6847
Proposed	2	2	4	30.0	57.0621	0.0596	0.6058	0.4266
Proposed	3	3	9	13.3	56.2383	0.3690	-0.2180	0.4165
Proposed	4	4	16	7.5	56.5233	0.7946	0.0670	0.7991

[a]True target area values for Y: Linear model--54.901; Nonlinear model--56.456.

TABLE 4.4

ESTIMATED PARAMETERS OF THE ERICKSEN AND PROPOSED METHODS WITH HIGH ASSOCIATION
n=480 ; n*=49

	S_1	S_2	k	n/k	Expectation	Variance	Bias[a]	Mean squared error
Linear model								
Ericksen	-	-	-	-	54.9214	0.0261	0.0209	0.0265
Proposed	2	2	4	120.0	55.9000	0.0781	0.9995	1.0771
Proposed	3	3	9	53.3	55.1665	0.0638	0.2660	0.1345
Proposed	4	4	16	30.0	55.0926	0.0656	0.1921	0.1025
Nonlinear model								
Ericksen	-	-	-	-	57.1961	0.0325	0.7397	0.5797
Proposed	2	2	4	120.0	57.0527	0.0119	0.5964	0.3676
Proposed	3	3	9	53.3	56.2435	0.0609	-1.2168	0.1062
Proposed	4	4	16	30.0	56.6869	0.0128	0.2306	0.0659

[a]True target area values for Y: Linear model--54.901; Nonlinear model--56.456.

TABLE 4.5

ESTIMATED PARAMETERS OF THE ERICKSEN AND PROPOSED METHODS WITH HIGH ASSOCIATION
n=120 ; n*=100

	S_1	S_2	k	n/k	Expectation	Variance	Bias[a]	Mean squared error
Linear model								
Ericksen	-	-	-	-	55.0293	0.1175	0.0015	0.1175
Proposed	2	2	4	30.0	56.1619	0.3591	1.1340	1.6451
Proposed	3	3	9	13.3	54.7880	0.2913	-0.2399	0.3488
Proposed	4	4	16	7.5	54.9017	0.9623	-0.1261	0.9782
Nonlinear model								
Ericksen	-	-	-	-	57.0293	0.1526	0.9679	1.0894
Proposed	2	2	4	30.0	57.2380	0.0635	0.9414	0.9497
Proposed	3	3	9	13.3	56.2400	0.3661	-0.0566	0.3693
Proposed	4	4	16	7.5	56.6353	0.7966	0.3388	0.9113

[a]True target area values for Y: Linear model--55.028; Nonlinear model--56.297.

TABLE 4.6

ESTIMATED PARAMETERS OF THE ERICKSEN AND PROPOSED METHODS WITH HIGH ASSOCIATION
n=480 ; n*=100

	S_1	S_2	k	n/k	Expectation	Variance	Bias[a]	Mean squared error
Linear model								
Ericksen	-	-	-	-	55.0427	0.0254	0.0149	0.0256
Proposed	2	2	4	120.0	56.1902	0.0781	1.1624	1.4292
Proposed	3	3	9	53.3	54.8209	0.0606	-0.2069	0.1035
Proposed	4	4	16	30.0	55.0594	0.0554	0.0316	0.0564
Nonlinear model								
Ericksen	-	-	-	-	57.2623	0.0367	0.9658	0.9694
Proposed	2	2	4	120.0	57.2522	0.0156	0.9556	0.9287
Proposed	3	3	9	53.3	56.2134	0.0733	-0.0831	0.0802
Proposed	4	4	16	30.0	56.7654	0.0176	0.4688	0.2373

[a]True target area values for Y: Linear model--55.028; Nonlinear model--56.297.

TABLE 4.7

ESTIMATED PARAMETERS OF THE ERICKSEN AND PROPOSED METHODS WITH MODERATE ASSOCIATION
n=120 ; n*=49

	S_1	S_2	k	n/k	Expectation	Variance	Bias[a]	Mean squared error
Linear model								
Ericksen	-	-	-	-	55.6250	0.3256	0.0046	0.3256
Proposed	2	2	4	30.0	56.4692	0.5121	0.8488	1.2325
Proposed	3	3	9	13.3	55.8548	0.6232	0.2344	0.6781
Proposed	4	4	16	7.5	55.7177	0.9944	0.0973	1.0039
Nonlinear model								
Ericksen	-	-	-	-	57.1183	0.6581	0.6620	1.0963
Proposed	2	2	4	30.0	56.9859	0.5599	0.5296	0.8404
Proposed	3	3	9	13.3	56.2555	1.1735	-0.2008	1.2138
Proposed	4	4	16	7.5	56.5368	1.2633	0.0805	1.2697

[a]True target area values for Y: Linear model--55.620; Nonlinear model--56.456.

TABLE 4.8

ESTIMATED PARAMETERS OF THE ERICKSEN AND PROPOSED METHODS WITH MODERATE ASSOCIATION
n=480 ; n*=49

	S_1	S_2	k	n/k	Expectation	Variance	Bias[a]	Mean squared error
Linear model								
Ericksen	-	-	-	-	55.5908	0.1073	-0.0296	0.1082
Proposed	2	2	4	120.0	56.4315	0.1574	0.8111	0.8152
Proposed	3	3	9	53.3	55.7209	0.2072	0.1004	0.2173
Proposed	4	4	16	30.0	55.6709	0.1941	0.0505	0.1967
Nonlinear model								
Ericksen	-	-	-	-	57.1667	0.1664	0.7104	0.6710
Proposed	2	2	4	120.0	57.0355	0.1454	0.5792	0.4809
Proposed	3	3	9	53.3	56.1987	0.2655	-0.2577	0.3319
Proposed	4	4	16	30.0	56.6265	0.1793	0.1701	0.2082

[a]True target area values for Y: Linear model--55.620; Nonlinear model--46.456.

TABLE 4.9

ESTIMATED PARAMETERS OF THE ERICKSEN AND PROPOSED METHODS WITH MODERATE ASSOCIATION
n=120 ; n*=100

	S_1	S_2	k	n/k	Expectation	Variance	Bias[a]	Mean squared error
Linear model								
Ericksen	-	-	-	-	55.7268	0.3574	-0.0030	0.3574
Proposed	2	2	4	30.0	56.6775	0.5940	0.9477	1.4922
Proposed	3	3	9	13.3	55.6026	0.6736	-0.1272	0.6898
Proposed	4	4	16	7.5	55.6816	1.2866	-0.0482	1.2889
Nonlinear model								
Ericksen	-	-	-	-	57.2720	0.7384	0.9754	1.6899
Proposed	2	2	4	30.0	57.2524	0.6183	0.9559	1.5320
Proposed	3	3	9	13.3	56.2488	1.1841	-0.0478	1.1863
Proposed	4	4	16	7.5	56.6761	1.5503	0.3795	1.6944

[a]True target area values for Y: Linear model--55.730; Nonlinear model--56.297.

TABLE 4.10

ESTIMATED PARAMETERS OF THE ERICKSEN AND PROPOSED METHODS WITH MODERATE ASSOCIATION
n=480 ; n*=100

	S_1	S_2	k	n/k	Expectation	Variance	Bias[a]	Mean squared error
Linear model								
Ericksen	-	-	-	-	55.7018	0.1008	-0.0280	0.1016
Proposed	2	2	4	120.0	56.6781	0.1906	0.9482	1.0898
Proposed	3	3	9	53.3	55.5665	0.1832	-0.1634	0.2099
Proposed	4	4	16	30.0	55.7440	0.1933	0.0142	0.1935
Nonlinear model								
Ericksen	-	-	-	-	57.2288	0.1780	0.9322	1.0470
Proposed	2	2	4	120.0	57.2396	0.1456	0.9430	1.0348
Proposed	3	3	9	53.3	56.2591	0.2391	-0.0374	0.2405
Proposed	4	4	16	30.0	56.7539	0.1736	0.4574	0.3828

[a]True target area values for Y: Linear model--55.730; Nonlinear model--56.297.

We might also point out that the results in each of the tables constitute what we have been referring to as a "run." Thus the target area values of the variables are the same for each of the methods considered in a table are the same. Also the sample observations on X_1 and X_2 are the same for each method in a table. The sample criterion values, of course, differ for the linear and nonlinear models.

One intuitively reasonable pattern established in these tables is that the results in high association situations are consistently better than the results in moderate association situations. Another intuitively reasonable result is that when everything else is held constant an increase in sample size from n=120 to n=480 reduces the variance of the estimates for both methods in all situations. The improvement varies, however. For the Ericksen method reduction is greater with the nonlinear model and generally increases as k, the number of groups formed, increases when n=480. Improvements in mean squared error due to sample size increase follow about the same pattern as with variances although the magnitude of improvement is somewhat less. There seems to be little effect with respect to variance or mean squared error which is attributable to change in n^*, the size of the target area.

Surprisingly, no general pattern emerges which would support the hypothesis that the larger k is the

smaller the variance. In fact, there are instances where the largest k yields the largest variance when n=120. This is no doubt due to a small average group size of 7.5 in that case. It begins to appear that for any given sample size increase in k yields improvement in the variance and mean squared error up to a point beyond which there is a trend in the direction of increased variance and mean squared error. This point of diminishing returns on k appears to vary directly with the sample size.

With the estimation procedure as conceived here, the bias value will be due in large part to how we define the values on X_1^* and X_2^* and how we select our group boundaries. If the range had been between the fifth and 95-th percentile for both variables in these simulation runs, the bias values for both methods would likely have been small with both models since here the target areas of observation effectively cover the ranges of X_1 and X_2. This should be taken into account before considering generalizations.

A comparison of methods leads us to several general conclusions. First, in linear model situations the Ericksen procedure does somewhat better although with the larger sample the results are much closer. In terms of the population size estimation presented in Chapter V, we might expect a priori that the proposed estimate will do comparably well although probably not better than the

Ericksen method. With the nonlinear model and high association the proposed method as expected generally does better particularly with the large sample size. This is the case when we used both the variance and mean squared error as criteria. With moderate association the two methods yield comparable results with respect to the variance, but with respect to mean squared error the proposed method does somewhat better particularly with the larger sample size. Had we defined moderate association in the nonlinear model as corresponding to a multiple correlation higher than the 0.50 we used, the superiority of the proposed method possibly would have been more clear cut.

CHAPTER V

TWO ILLUSTRATIVE EXAMPLES

5.1 Introduction

As a final facet of our research we observe the behavior of the proposed method when it is applied to real data. In this chapter we concentrate less upon the various properties of the estimators and more upon the comparative accuracy of the methods. By this approach we hope to gain insight as to the circumstances under which the proposed method performs well. It may also indicate difficulties of the procedure and necessary modifications. As in Chapters III and IV, the proposed method is compared with the Ericksen [1971] method since the methods are applicable under essentially the same circumstances.

The discussion in this chapter focuses on two illustrative examples while the methodology associated with a third is mentioned. The criterion variable for the first example is population size. The availability of three ideal symptomatic variables (i.e., births, deaths and school enrollment) with a high level of multiple correlation (see Ericksen [1971]) and a linear functional model makes it quite unlikely a priori that the proposed method will give better results than the Ericksen method.

It will be interesting, however, to observe whether they are at least comparable.

The second example uses per capita income as its criterion variable. The primary usefulness of this variable is the availability of symptomatic variables which may exhibit a nonlinear relationship to the criterion variable. Based upon previous results it is in this situation that the proposed method should do best. These symptomatic variables are obtained from several data sources collected by the Bureau of the Census. Another reason for considering the income variable is the growing need for estimates of this type of variable. The recent change in governmental fiscal policy from one of central distribution to one of local revenue sharing involves a need for local estimates of different variables such as per capita income whose values determine a local area's level of Federal monetary allotment.

The third example involves estimation of the incidence of dilapidated housing in local areas. Here accuracy comparisons cannot be made since the actual 1970 figures are not available from the Bureau. Thus we go only so far as to describe the implementation of the proposed procedure and to present the estimates. This is considered in Section 5.5.

5.2 General comments on procedure

In each of the examples discussed below the cri-

terion variable estimates are made for states. Thus in terms of the definitions of Chapter II, each target area is a state and the base units in the classification and estimation phases are counties.

In both instances the Current Population Survey (CPS), which is conducted monthly by the Bureau of the Census, is used as the source of sample information. Briefly, the CPS is a national multistage probability sample of the population. The first stage consists of Primary Sampling Units (PSU's) which are essentially counties or groups of contiguous counties. They are of two types--self-representing or non-self-representing. A self-representing PSU most generally is a Standard Metropolitan Statistical Area with a population of no less than 250,000. A non-self-representing PSU consists of a less densely populated county or group of contiguous counties. All PSU's are associated with a stratum, and the number of strata is equal to the number of PSU's included in the sample. Each of the self-representing PSU's are considered a stratum and are automatically a part of the sample since the design calls for one PSU selection per stratum. In the remaining strata non-self-representing PSU's are selected with probabilities proportional to their size. Each of the 449 sample PSU's are broken down into Enumerator Districts (ED's), which make up the second stage. The selected ED's are further divided into blocks constituting

a third stage. Finally, the selected blocks are divided into segments which contain approximately six dwelling units and which are called Ultimate Sampling Units. Selection probabilities at each stage are arranged so that the overall sampling rate for all dwellings is 1/2244.61.

Referring once more to the discussion of Chapter II, the major contribution of the sample is to provide, in this case, estimates of the criterion variable for the sample base units. However, we note that it is difficult to obtain such estimates since counties are not a stage in the sample design. Instead we let the PSU's be the sample base units of the grouping phase. The effect of this modification is most likely minimal since both counties and PSU's are arbitrary geographic entities so that one might easily envision a county as being a "small PSU" without losing the principle of the method. However, it is worth noting that the counties which make up a non-self-representing PSU are chosen so that they are different as possible subject to the geographic constraint of contiguity.

The estimates in all examples are made for 1970 since, for the first two, known values are available from the Bureau of the Census (U.S. Bureau of the Census [1972]). Therefore, comparisons of accuracy for the state estimates as well as for the county estimates can be made. The criterion for measuring the accuracy of an estimate is the

relative absolute deviation as defined in Section 2.5.

The general procedure followed in the examples is as follows:

(1) Grouping--The grouping of the PSU's is done using the Automatic Interaction Detector, version II (AID II) which is attributed to Morgan and Sonquist [1963]. Input to the program consists of estimates of the criterion variable and selected symptomatic variables. The latter are obtained from the 1967 County-City Data Book (U.S. Bureau of the Census [1967]) and the vital registry of births, deaths and school enrollment.

The objective of the AID II algorithm is to form groups so that the predictability of the criterion variable by the symptomatic variables is maximized. In other words, the criterion variable is centrally incorporated in the group formation. To implement this objective the procedure consists of a series of dichotomous splits on the recoded (i.e., categorized) symptomatic variables in such a way as to maximize, for each split,

$$\frac{n_1\bar{y}_1^{\,2}+n_2\bar{y}_2^{\,2}}{n\bar{y}^2}$$

where n_1 and n_2 are the numbers in the split groups and $\bar{y}_1$ and $\bar{y}_2$ are the criterion means in these groups. The values n and $\bar{y}$ are for the unsplit group. These splits continue until an additional prespecified proportion of the total variation is no longer explained by another

split or until the designated minimal group size is attained.

Satisfying the grouping rule for AID II implies that the variation on the criterion variable is maximally reduced by the split. Putting it another way, knowing the group mean allows reduction of criterion variation from $\sum y^2$ to $\sum y^2 - n\bar{y}^2$, but knowing the two means reduces it further to $\sum y^2 - n_1\bar{y}_1^2 - n_2\bar{y}_2^2$. The split is made so that this reduction (i.e., $n_1\bar{y}_1^2 + n_2\bar{y}_2^2 - n\bar{y}^2$) is maximal and thus a larger portion of the criterion variation is explained.

Output from AID II consists of a nonsymmetric branch network on the symptomatic variables. Each branch is defined by the categories of the symptomatic variables on it as well as the number and criterion mean of the observations on that branch. The size of the branch network (and implicitly the number of groups, k) is determined by the reducibility level (i.e., the amount of reduction in the unexplained sum of squares necessary for a further split) and the minimum group size. The branching implies that the group boundaries are rectilinear and exhaustive so long as the categories for each symptomatic variable are mutually exclusive and exhaustive over the range of values assumed by that variable. This mirrors a similar characteristic of the "preferred grouping procedure" described in Chapter II since both criterion and

symptomatic information are used in selecting groups but the groups are defined in terms of the symptomatic variables only.

An important advantage of this procedure is that additivity and linearity, relating to the criterion and symptomatic variables, are not required. This is especially appropriate for the present situation since no specific assumptions are made regarding a functional model. Also its inherent simplicity makes it practical for large data sets as is the case here.

The most recent version of this algorithm, called AID III (Sonquist et.al. [1971]), has several useful options. One is a covariance option which is found most useful in situations where one symptomatic variable prevails as income often does in economic studies. Assuming some type of linear relationship between the criterion and this variable, the covariate, the groups are divided using that variable directly. The output indicates an estimate of the linear relationship for each group which is useful in the estimation phase. Other options allow for such things as prespecification of a sequence of splits, compulsory symmetry of splits, and assignment of priorities to the symptomatic variables.

(2) Classification--Since grouping by AID II yields rectilinear groups, the natural classification procedure and the one we use is the "pigeon-hole" rule whereby an ob-

servation is assigned to that group whose boundaries include the observation's symptomatic values. As mentioned previously counties are the geographic units being assigned although the geographic units in the grouping phase are PSU's.

(3) Estimation--Since the CPS is a multistage sample, the description of the estimation phase parallels the discussion in Section 2.4. The value for $\bar{y}_{gh}$ is found from (2.4.1) where P_{ghc} is the known probability that a particular household was selected into the sample. The estimate for the group mean, $\bar{y}_g$, is found using (2.4.4) where W_{gh} is proportional to the expected number of persons who would be sampled within that PSU if sampling were done by simple random selection. This is found by first noting that the multistage design of the CPS implies that

$$P_{gh} \cdot P_w = 1/2244.61 \qquad (5.2.1)$$

where P_w is defined as the within-PSU sampling rate. Thus

$$P_w = \{[2244.61] \cdot [P_{gh}]\}^{-1} \qquad (5.2.2)$$

and

$$N_{gh} = [\text{Actual 1960 population}] \cdot [P_w] \qquad (5.2.3)$$

where N_{gh} is the expected number of persons sampled in the gh-th PSU. Finally,

$$W_{gh} = N_{gh} / \sum_{h=1}^{n_g} N_{gh} \quad . \qquad (5.2.4)$$

The final estimate, $\bar{y}^*$, is found as in (2.4.6) where M_g^* in

(2.4.7) is the 1960 population of all counties in group g.

In general, the Ericksen procedure is done by first fitting the assumed linear model by weighted least squares, the weights assigned to the PSU being of two types. With the first type each non-self-representing PSU is assigned a weight of 1 and each self-representing PSU a weight equal to its 1960 population size divided by 250,000. This approach is followed by Ericksen [1972] and will be referred to as the original Ericksen method. With the second type each PSU is assigned a weight proportional to N_{gh} in (5.2.3) and will be called the revised Ericksen method. In both cases it is clear that the large self-representing PSU's are given the greatest weight since $P_{gh}=1$ and the 1960 population is large in these PSU's.

The following two sections contain an explicit description of the two examples along with tabulation and discussion of the results.

5.3 Population size example and results

The variables involved in this example are defined as follows:

Y_1 ∝ ratio of the total 1970 population to the total 1960 population.

X_{11} ∝ ratio of the total 1970 live births by place of residence to the total 1960 live births by place of residence.

X_{12} ∝ ratio of the total 1970 deaths by place of residence to the total 1960 deaths by place of residence.

X_{13} ∝ ratio of the total 1970 school enrollment to the 1960 total school enrollment.

We use the ratio form as a means of standardization as well as to reduce the effect of nonuniformity among states regarding the collection of vital statistics. Therefore, the final estimate is a ratio, but since the denominator is known the estimate can be obtained directly.

To implement both methods a CPS estimate of Y_1 is used, and to increase the accuracy of the estimate the results of five quarterly CPS surveys, centering on April 1, 1970, are combined. In addition, several of the PSU's in New England are combined since some vital statistics are available only for these combined areas. Thus instead of the original 449 PSU's there are 440 after merging.

The fitted regression coefficients for both this and later examples, as used by the Ericksen method, are found in Table 5.1, and the state population estimates for different versions of the Ericksen and proposed methods are given in Table 5.2. Several state estimates cannot be found since they do not participate in the national vital registry.

One general conclusion to be drawn from these results is that, as expected, the Ericksen method generally does better, although the proposed method gives a more

TABLE 5.1

FITTED REGRESSION COEFFICIENTS USED BY THE ERICKSEN METHOD IN THE ILLUSTRATIVE EXAMPLES

Criterion variable	Constant	Symptomatic variables						Multiple correlation
		(Coefficient)	Variable	(Coefficient)	Variable	(Coefficient)	Variable	
Y_1[a]	16.828	(0.146)	X_{11}	(0.156)	X_{12}	(0.567)	X_{13}	0.64
Y_1[b]	15.800	(0.218)	X_{11}	(0.142)	X_{12}	(0.520)	X_{13}	0.65
Y_2[a]	1144.952	(-23.289)	X_{21}	(11.768)	X_{22}	(-34.334)	X_{23}	0.81
Y_2[a]	1592.533	(2.781)	X_{24}	(5.846)	X_{25}	(66.734)	X_{26}	0.77
Y_3[a]	28.787	(0.056)	X_{31}	(-0.119)	X_{32}	(-0.048)	X_{33}	0.56

[a]Revised Ericksen method.

[b]Original Ericksen method.

TABLE 5.2

RELATIVE ABSOLUTE DEVIATION TIMES 100 OF THE
1970 POPULATION SIZE ESTIMATES FOR STATES

State	Number of counties	Proposed method		Ericksen method	
		k=23	k=64	Original	Revised
Alabama	67	1.4	0.6	0.9	0.6
Arizona	14	4.8	13.7	1.0	0.8
Arkansas	75	5.4	5.9	1.3	0.2
California	58	2.9	0.6	3.5	2.1
Colorado	63	6.4	5.3	2.5	0.8
Delaware	3	4.2	13.3	1.0	1.0
Florida	67	2.2	1.9	2.0	0.0
Georgia	158	2.6	5.0	1.0	0.1
Hawaii	4	6.0	11.9	2.4	1.1
Idaho	44	2.1	0.7	1.2	0.3
Illinois	102	1.3	0.3	1.0	2.7
Indiana	92	3.0	1.7	1.1	0.3
Iowa	99	2.9	2.7	0.2	1.8
Kansas	105	0.9	0.6	2.8	0.9
Kentucky	120	0.7	2.3	1.8	0.5
Louisiana	64	1.9	5.8	0.3	1.5
Maine	16	5.4	4.1	1.0	2.7
Maryland	24	6.1	1.9	3.9	1.9
Massachusetts	14	0.4	3.2	2.5	0.7
Michigan	83	2.2	1.8	0.2	2.3
Minnesota	87	2.6	0.9	0.2	2.3
Mississippi	82	2.7	2.7	1.0	2.5
Missouri	115	7.0	5.7	---	11.9
Montana	56	3.5	3.5	1.4	0.8

TABLE 5.2--Continued

State	Number of counties	Proposed method		Ericksen method	
		k=23	k=64	Original	Revised
Nebraska	93	2.0	4.1	0.9	1.3
Nevada	17	17.8	11.8	4.7	3.2
New Jersey	21	0.1	0.9	2.5	1.0
New Mexico	32	1.5	2.2	1.3	1.2
New York	58	2.6	2.5	0.7	2.1
North Carolina	100	0.0	1.3	4.5	3.5
Ohio	88	1.9	1.5	1.1	0.3
Oklahoma	77	7.0	7.3	0.0	1.4
Pennsylvania	67	3.1	7.6	1.7	3.6
Rhode Island	5	2.6	3.4	0.3	1.6
South Carolina	46	4.7	1.0	2.7	1.5
South Dakota	67	5.7	5.4	0.4	2.9
Tennessee	95	2.4	3.3	2.3	1.2
Utah	29	8.8	4.7	0.5	0.5
Vermont	14	2.0	4.0	2.2	1.0
Washington	39	1.7	0.4	3.3	3.1
West Virginia	55	1.2	2.7	4.4	0.6
Wisconsin	71	1.2	4.0	1.7	0.0
SUMMARY: Total number of times gives most accurate results		3	8	13	18
		20	22	--	--
		--	--	14	28

accurate estimate in about a quarter of the states. This was not expected a priori. Generally, the more accurate results come in states with a large number of counties and the poorest come in states with few counties. Also, the number of groups formed (k=23 or 64) does not appear to affect the results much here since k=64 gives better results in 22 of 42 states' estimates. One interesting feature of the two versions of the Ericksen method is that the revised form does as good as or better in 28 of the states.

We might conclude that although the Ericksen method generally does better, the proposed method does quite well in view of the relative accuracy of the county byproduct estimates found in Table 5.3. Here the proposed method gives somewhat less accurate results in all categories although for both methods the results tend to improve as the population size increases. The proposed method with 23 groups gives consistently better results than with 64 groups. One peculiar result for both methods is that one version does well on the state estimates while doing poorly on the county estimates. For example, with the Ericksen method the revised version does better than the original in 28 of 42 counties although its overall mean deviation (5.66) is larger than the mean for the original version (4.84).

TABLE 5.3

MEAN RELATIVE ABSOLUTE DEVIATION TIMES 100 OF THE 1970 BYPRODUCT POPULATION SIZE ESTIMATES FOR COUNTIES BY THE ACTUAL 1970 POPULATION SIZE

Actual 1970 population	Number of counties	Proposed method		Revised Ericksen method
		k=23	k=64	
<5000	220	11.29	15.20	8.72
5000- 10000	429	8.92	10.97	5.88
10000- 25000	851	7.74	10.00	6.70
25000- 50000	494	7.53	9.84	4.41
50000-100000	292	7.45	9.51	4.56
100000-250000	166	6.77	8.75	3.28
250000-500000	69	6.65	8.92	3.06
>500000	65	5.96	8.19	3.86
TOTAL	2586	8.03	10.36	5.66[a]

[a]The original Ericksen method overall mean is 4.84 (Ericksen [1972]).

5.4 Per capita income example and results

The variables involved in this example are defined as follows:

Y_2 ∝ 1970 per capita income.

X_{21} ∝ per cent natural increase in population from 1960 to 1965.

X_{22} ∝ 1960 per capita aggregate income.

X_{23} ∝ 1964 per cent of the population on public assistance.

X_{24} ∝ 1960 per cent of the population living in an urban area.

X_{25} ∝ 1960 per cent of the population with a high school education or more.

X_{26} ∝ 1960 per cent of the population with an income of $10,000 or more.

As mentioned previously, per capita income is used as a criterion variable since it was felt that among the available symptomatic variables some could be found which exhibit a nonlinear relationship with the criterion variable. The thought was that this would demonstrate the potential usefulness of the proposed method. However, preliminary investigations of the candidate symptomatic variables indicates that there does not appear to be any such relationship among the candidate variables. This conclusion is based upon the results of fitting several nonlinear models to the sample data. For each set of variables there is little statistical evidence to indicate that

the model tested fits the data better than a linear model. Thus it appears that we are comparing the methods within the context of a linear model as in the first example.

Two sets of symptomatic variables are used in this example: (A) X_{21}, X_{22} and X_{23}; and (B) X_{24}, X_{25} and X_{26}. The multiple correlation of these two sets of variables with Y_2 is given in Table 5.1. These two sets of variables were selected according to two criteria. The first group consists of the first three variables added to the regression equation in a linear stepwise regression algorithm. For the second group the selection rule was that the variable be the same or highly correlated with the symptomatic variables used in the trial run (see Section 2.5).

As in the first example the form of the criterion variable is a ratio although here the ratio is of total income to the total population in 1970. In addition, the estimate for each PSU is obtained from a single round of the CPS taken in March 1970. Estimates obtained by the methods appear in Table 5.4.

Several inferences can be made from these results. First of all with the exception of the states with very low per capita income, most estimates are less than seven per cent in error. Second, with variable set A the proposed method does better than the Ericksen method (for either k=20 or k=59) in 29 of 47 states, but with vari-

TABLE 5.4

PER CENT ABSOLUTE DEVIATION OF THE 1970 PER CAPITA INCOME ESTIMATES FOR STATES

State	Number of counties	Actual 1970 per capita income	Proposed method			Ericksen method	
			Variable set A		Variable set B	Variable set A	Variable set B
			k=20	k=59	k=21		
Alabama	67	2317	6.9	7.2	6.3	4.4	5.8
Arizona	14	2937	1.5	7.8	8.3	2.6	2.7
Arkansas	75	2142	3.6	3.5	4.9	1.1	4.9
California	58	36.4	5.0	3.1	2.7	2.5	0.7
Colorado	63	3106	1.8	0.2	1.0	3.0	1.0
Connecticut	8	3885	8.9	7.1	9.4	5.5	9.0
Delaware	3	3265	1.5	6.9	5.1	1.5	2.2
Florida	67	3058	6.6	6.4	5.1	4.7	8.8
Georgia	159	2640	9.6	9.3	3.4	9.2	3.9
Idaho	44	2644	3.5	3.7	5.3	4.7	2.6
Illinois	102	3495	4.0	2.0	3.9	1.5	2.0
Indiana	92	3070	0.3	0.6	0.7	0.9	3.5
Iowa	99	2884	0.1	1.3	2.0	0.9	5.4
Kansas	105	2929	0.7	0.6	0.2	2.3	2.5
Kentucky	120	2425	3.6	1.7	2.0	2.1	0.8
Louisiana	64	2330	0.7	1.8	16.1	1.6	12.0
Maine	16	2548	4.6	4.4	0.1	4.8	1.9
Maryland	24	35.2	11.3	11.5	7.6	8.7	4.6
Massachusetts	14	3408	2.3	3.5	3.0	3.1	5.0
Michigan	83	3357	4.0	3.6	1.6	6.0	4.8

TABLE 5.4--Continued

State	Number of counties	Actual 1970 per capita income	Proposed method			Ericksen method	
			Variable set A		Variable set B	Variable set A	Variable set B
			k=20	k=59	k=21		
Minnesota	87	3038	4.5	4.6	1.9	4.1	4.9
Mississippi	82	1925	7.3	8.3	15.0	6.2	15.0
Missouri	115	2952	3.0	4.1	5.6	1.9	5.6
Montana	57	2696	5.9	4.1	6.6	6.7	3.1
Nebraska	93	2797	2.2	3.6	1.1	1.3	3.1
Nevada	17	3554	3.9	1.0	1.0	1.2	0.1
New Yampshire	10	2985	0.5	3.7	2.1	2.5	7.4
New Jersey	21	3674	4.6	6.0	5.2	2.1	3.5
New Mexico	32	2437	9.0	1.6	28.4	6.8	22.5
New York	62	3608	11.7	13.5	6.8	6.6	5.9
North Carolina	100	2474	5.3	6.5	2.7	6.8	5.1
North Dakota	53	2469	0.5	0.5	4.6	0.0	0.5
Ohio	88	3199	1.2	2.6	1.0	0.0	2.3
Oklahoma	77	2694	0.2	0.5	2.7	2.0	0.5
Oregon	36	3148	2.0	0.8	2.9	2.1	5.5
Pennsylvania	67	3066	0.2	1.2	1.5	0.7	3.9
Rhode Island	5	3121	0.1	2.3	5.5	3.1	9.7
South Carolina	46	2303	3.6	4.1	2.9	4.2	0.7
South Dakota	67	2387	1.5	2.0	4.8	2.1	2.9
Tennessee	95	2464	8.8	4.7	13.1	6.1	10.3
Texas	254	2792	1.6	1.0	5.0	1.6	1.0
Utah	29	2697	3.0	4.5	16.0	3.0	13.0
Vermont	14	2996	5.5	3.2	2.1	3.5	8.2

TABLE 5.4--Continued

State	Number of counties	Actual 1970 per capita income	Proposed method			Ericksen method	
			Variable set A		Variable set B	Variable set A	Variable set B
			k=20	k=59	k=21		
Washington	39	3357	4.6	3.7	4.4	3.4	4.9
West Virginia	55	2333	1.4	3.2	6.2	0.8	4.2
Wisconsin	72	3032	0.4	0.8	3.4	1.5	2.1
Wyoming	23	2895	5.8	7.5	7.0	7.1	4.5
SUMMARY: Total			12	5	8	13	9
number of times			--	--	--	30	17
gives most ac-			28	19	--	--	--
curate results			17	12	--	18	--
			--	--	22	--	25

able set B it does better in only 22 states. For both methods variable set A gives somewhat better results, although the multiple correlations associated with the groups are about equal. Also, the proposed method does best in the midwestern states with moderate per capita income. The Ericksen method generally provides better estimates in the southern rural states with low per capita income and in the eastern states with high per capita income. The proposed method does somewhat better with k=20 than with k=59 although there is no detectable cause for this. In general, we might conclude that the two methods do about equally well overall but that the proposed method seems to produce somewhat better results with moderate per capita income states while the Ericksen method appears to do better at the extremes.

As with the population size example the byproduct county estimates offer some interesting results. These are found in Table 5.5. The mean deviations for the proposed method tend to be somewhat higher although the magnitude of the difference is less than with the first example. Both methods tend to give their best results in the counties with moderate per capita income and the worst results at the extremes, although this is more noticable with the proposed method. The proposed method consistently does better with k=20 than with k=59, and with both methods the variable set A tends to give slightly better results.

TABLE 5.5

MEAN RELATIVE ABSOLUTE DEVIATION OF 1970 BYPRODUCT PER CAPITA INCOME ESTIMATES FOR COUNTIES BY THE ACTUAL 1970 PER CAPITA INCOME

Actual 1970 per capita income	Number of counties	Proposed method			Ericksen method	
		Variable set A		Variable set B	Variable set A	Variable set B
		k=20	k=59	k=21		
<1500	105	16.13	18.24	44.05	11.31	45.65
1500-1900	432	11.98	14.74	16.22	8.94	16.54
1900-2300	680	11.08	14.54	11.92	8.27	7.74
2300-2600	741	8.68	12.20	9.94	7.12	7.32
2600-2900	503	8.01	8.84	8.17	6.38	8.01
2900-3100	230	6.87	8.04	7.56	5.40	9.25
3100-3300	112	6.80	8.65	8.71	5.53	9.20
3300-3600	92	5.75	8.37	8.82	4.89	6.84
3600-3900	38	6.80	6.45	6.09	6.84	7.45
>3900	27	19.56	19.88	16.54	9.25	8.86
TOTAL	2960	9.63	12.17	11.97	7.42	10.46

5.5 Procedure for incidence of dilapidated housing example

The variables involved in this example are defined as follows:

Y_3 ∝ 1970 proportion of all housing units which are defined as dilapidated.[1]

X_{31} ∝ 1960 per cent of the population with an income of under $3000.

X_{32} ∝ 1960 per cent of all housing units with sound plumbing.

X_{33} ∝ 1960 index of home equipment.[2]

Since no census figures are available to determine the accuracy of state estimates found for this criterion variable, we present only a brief discussion of the procedure used to obtain the estimates and a notion as to the comparative magnitude of the estimates obtained.

The procedure used to obtain these estimates is much the same as with the previous examples. The major difference is that the sample estimates for Y_3 are obtained from the 1970 Components of Change Survey (CINCH).

---

[1] A housing unit is defined to be dilapidated if it does not provide adequate shelter and in its present condition endangers the health and safety of its inhabitants. This might be due to such defects as holes, open cracks, etc. over a large area of the foundation or walls, substantial sagging of floors and roof, or extensive damage by storm, fire or flood.

[2] The index of home equipment is found by adding together the per cents of occupied housing units with washing machine, home food freezer, air conditioning, television set, telephone and automobile. The maximum value is 600.

This survey was conducted by the Housing Division in the Bureau of the Census in order to produce estimates concerning the condition of housing, particularly in urban areas. In sample design this survey parallels the current CPS multistage design except that 357 instead of 449 PSU's were involved. The grouping phase in this example involved k=17 groups obtained from the AID II algorithm. The state estimates for both methods are given in Table 5.6.

In terms of relative magnitude both methods yield results which vary from zero to 13.1. However, in general, the proposed method gives consistently lower estimates than the Ericksen method for the southern states and about the same or slightly higher estimates for the northern states. No other major difference is immediately evident.

TABLE 5.6

1970 STATE ESTIMATES OF THE PROPORTION TIMES 100 OF HOUSING UNITS THAT ARE DILAPIDATED

State	Method		State	Method	
	Proposed	Ericksen		Proposed	Ericksen
Alabama	5.9	9.8	Nebraska	3.6	2.7
Arizona	3.8	4.5	Nevada	4.5	4.4
Arkansas	7.3	11.4	New Hampshire	3.5	4.2
California	2.8	2.4	New Jersey	2.5	1.9
Colorado	3.5	3.4	New Mexico	4.2	6.3
Connecticut	2.5	1.7	New York	2.9	4.4
Delaware	2.3	2.0	North Carolina	6.9	9.0
Florida	4.5	5.7	North Dakota	4.4	4.8
Georgia	6.3	8.9	Ohio	3.0	2.7
Idaho	3.5	3.0	Oklahoma	4.8	5.7
Illinois	3.1	3.3	Oregon	3.0	2.8
Indiana	3.4	3.5	Pennsylvania	2.9	2.9
Iowa	4.1	3.2	Rhode Island	2.9	4.6
Kansas	3.8	2.6	South Carolina	6.8	10.2
Kentucky	5.6	9.4	South Dakota	4.1	5.4
Louisiana	6.2	7.2	Tennessee	5.8	7.8
Maine	4.8	6.3	Texas	5.1	5.9
Maryland	2.8	2.7	Utah	2.2	1.6
Massachusetts	3.1	3.7	Vermont	4.4	4.8
Michigan	2.8	2.3	Washington	3.1	2.7
Minnesota	3.5	3.2	West Virginia	6.5	8.0
Mississippi	6.0	13.1	Wisconsin	3.2	2.5
Missouri	4.9	5.3	Wyoming	3.5	3.8
Montana	4.5	4.3			

CHAPTER VI

SOME CONCLUDING REMARKS

6.1 Introduction

To conclude we present a brief outline of the work we have done, some general conclusions, and some extensions. This is intended to give the reader an overall perspective of our work and to tie together the results obtained by the various avenues we have followed.

6.2 Procedural review of present research

The investigation of any method for estimating a population parameter can follow either of two extremes--one is theoretical or analytical and the other is empirical or situation-specific. Neither alone is unequivocably advantageous. The theoretical approach allows one to obtain results expressed in terms of formulas from which conclusions can be drawn for wide ranges of parameter values, but these conclusions are limited by the assumptions set forth. The purely empirical approach sacrifices this control in favor of what might be called "relevance." An important middle ground of these two extremes is an approach such as Monte Carlo simulation which allows one some of the best of both extremes. The work we have done

follows a three-pronged approach.

The introductory material of Chapter I establishes the need for methodology which circumvents the assumptions which limit several present methods. This need is shown by a description of the present methods as well as a brief procedural history of postcensal estimation in general. A method which is believed to meet this need is proposed in Chapter II.

The proposed method, like the method attributed to Ericksen [1971], involves using sample data on the criterion variable (i.e., the variable to be estimated) as well as data on related or symptomatic variables to arrive at the estimate. However, no assumed model is fitted as in the Ericksen method. Instead the sample observations are grouped on the basis of some criterion such as Euclidean distance or criterion variable predictability by the symptomatic variables. Estimates of the criterion variable are then found for all groups, and each target area observation is assigned a value equal to the estimate for the group into which it is classified. Finally, the criterion estimates for the target area observations are combined to obtain an estimate for the target area. Thus it is hypothesized that the proposed method transcends any functional assumptions and that the only requirement for satisfactory results is that the level of association between the criterion and symp-

tomatic variables be sufficiently high.

The initial research phase discussed in Chapter III involves the development of an analytic approach to the problem of estimating various statistical characteristics of the proposed and Ericksen methods. The latter can be expressed in terms of the least squares estimates inherent in a linear model so that these characteristics can be found directly from the model. To describe the characteristics of the proposed method involves developing a mathematically relevant conceptualization of the grouping and classification phases of the method. Since only bivariate distributions (i.e., one symptomatic variable) are dealt with, the group boundaries which determine the groups are defined in terms of order statistics on the sample values of the symptomatic variable. Classification is defined by what is called "pigeon-holing" whereby an observation is assigned to that group whose range of symptomatic variable values include its value.

In Chapter IV several of the restricting assumptions (e.g., a single symptomatic variable) are removed and estimation of the statistical characteristics is pursued by means of Monte Carlo simulation. The work here involves two symptomatic variables and two general models, one linear, involving a trivariate normal distribution, and one nonlinear, involving (1) a bivariate normal distribution on the symptomatic variables, (2) a univariate

logistic distribution on the criterion variable given the symptomatic variables, and (3) a regression figure which is a logistic function based upon a linear combination of the symptomatic variables. Several simulation runs are made while varying such things as n, the sample size; k, the number of groups; n*, the number of observations in the target area; and the level of association.

The final phase of the research involves making estimates by both methods using some data provided by the United States Bureau of the Census. Estimates are made for the variables: total population size, per capita income, and incidence of dilapidated housing units. In each case the target areas are states and the sources of sample data are the Current Population Survey for the first two and the 1970 Components of Change (CINCH) survey for the third. The sources of symptomatic data are the vital registry for the first example's symptomatic variables and the 1967 County-City Data Book for the remaining examples, which contains an assortment of data by county and SMSA obtained from 1960 census information and other data collected between 1960 and 1967. Observations are measured by Primary Sampling Units for the sample and by counties for the target areas. The grouping algorithm used is AID II, which performs a series of dichotomous splits on the symptomatic variables according to the rule that prediction of the criterion variable by the symptomatic

variables must be maximized by a split. Classification involves "pigeon-holing" observations as mentioned previously. In each case three symptomatic variables are used. Since 1970 census figures are available for the first two examples, the state estimates and accompanying byproduct county estimates are made and then compared for accuracy with the actual values.

Among the reasons for the selection of per capita income as a criterion variable in this portion of the research is that it was thought that there would be nonlinear relationships between this criterion variable and several appropriate symptomatic variables. Such is not the case, however. After a series of model-fitting calculations in an attempt to establish such a nonlinear relationship, it is clear that no such relationship exists among the available symptomatic variables. Thus the example involving per capita income, like the population size example, demonstrates the ability of the proposed method to perform in a linear setting.

It is felt that these approaches allow a thorough examination of the proposed method as well as relevant comparisons with the Ericksen method, which exemplifies the principle of model-fitting using symptomatic information. It is these comparisons which indicate the circumstances under which the proposed method is advantageous.

6.3 General conclusions

We now summarize the major findings of our research with the intention of describing the strengths and shortcomings of the proposed method. As an extension of this discussion we present in the next section some of the questions that have been raised and some suggestions for further research.

The major findings of this study are as follows:

1. The hypothesis that the proposed method eliminates the need for a model assumption is essentially correct. This conclusion is based upon the fact that in the analytic work and simulations it is evident that in instances where target areas are not "typical" (i.e., the target area range of values on X is somewhat extreme toward one end of the total range on X) and the level of association is quite high, the method performs comparatively better than the Ericksen method with respect to both variance and bias.
2. The method performs quite well in some apparently linear settings. This conclusion is based upon the results of the per capita income example where variable set A (see Section 5.4) is used. In that instance better state estimates are obtained in about three-fifths of the states.
3. In the linear setting when AID II is the grouping algorithm, the proposed method's byproduct estimates

are not as accurate as those of the Ericksen method. This is a general trend found in the example results. Trends in the methods' accuracy levels by the true value of the criterion variable generally parallel each other although the Ericksen method gives consistently more accurate results. Judging from the analytic findings, we speculate that in the nonlinear case with high association the proposed byproduct estimates would generally be more accurate.

4. The value of n/k appears to be more important than the values of n and k individually. This is reflected in the results and conclusions of the analytic work, simulations and examples.
5. Moderate changes in the value of n* appear to have little influence on the quality of the proposed method's results. This is borne out mainly in the findings of the simulations and somewhat inconclusively in the analytic work. One possible qualification is that the more accurate state population size estimates are in states with a relatively large number of counties (i.e., n*).
6. The range of target area values on X relative to the total range on X is important but only within the context of the variables, type of regression model and level of association involved. This can be seen by noting that in the analytic and simulation work the

proposed method does relatively better than the Ericksen method, but in the per capita income example the proposed method generally does better in the "typical" states (e.g., Michigan, Minnesota).

7. The bias and variance appear to vary directly as the variance of the criterion variable in the population. This is most noticable in the analytic work where σ_y is allowed to vary and particularly since the effect is nearly directly proportion to σ_y.

We end this section with a short description of the circumstances under which the proposed method is the recommended one in view of the above conclusions. Essentially, this situation is the one where there is evidence to believe that there exists good (in terms of association) symptomatic variables which appear to exhibit a nonlinear relationship to the criterion variable. If the range of symptomatic variables at least partially covers portions of the regression figure which deviate from linearity, the proposed method produces better results than the regression-fitting method. On the other hand, when there is indication that the model is linear with high association, the proposed method most definitely should not be used, meaning that the Ericksen method is recommended under these circumstances.

6.4 Suggested further research

During the course of research in most areas, one

encounters points at which several alternative courses may be followed. Yet in most such instances only one course can be pursued leaving many courses of study which might be of future interest. Such has been the case in the present research. Therefore, we mention several of the alternate paths which might have been followed and which may prove to be interesting extensions of the present work.

Certainly one major extension is to find some data for an important application for which nonlinearity among the criterion and symptomatic variables is evident. This might imply the necessity of a change in the type of sample data from that of the multistage national variety to that of a sample of individuals (e.g., physicians, teachers). In this case the symptomatic variables used are not limited to those which are available by Primary Sampling Units, Census Tracts, etc. In this nonlinearity setting it would be interesting to observe the accuracy of not only the target area estimates but also the by-product estimates of the observations which make up the target area. It may also be of interest to find instances where several qualitative symptomatic variables are used.

Another extension involves a modification of the estimation phase of the proposed method. Instead of a simple or weighted mean, a regression or ratio estimate might be used to improve the group estimate of the cri-

terion variable. As one way to implement this extension, AID III, with its covariance option, could be used. In the output the criterion variable is defined, for each group, as an estimated linear function of the covariate.

Finally, it may be interesting to try a clustering algorithm which uses some clustering criterion other than the one used by AID II. There is reason to believe that the poor showing of the byproduct estimates in the example may have been due to the fact that the groups were poorly defined. This implies that the chances were high that a target area observation would be assigned to a "wrong" group. This could be verified by using grouping algorithms with differing criteria on the same body of data and then comparing the results. Other clustering methods were not used in this work because none were accessible at this time which would handle the magnitude of data involved.

APPENDIX A

APPROXIMATIONS FOR THE MOMENTS OF ORDER STATISTICS FROM THE STANDARD NORMAL DISTRIBUTION

In Section 3.4(1) we are required to find $\mu_{j:n}$, the expectation of the j-th ordered observation in a sample of size n from a standard normal distribution; $\sigma_{jj:n}$, the variance of the j-th observation from the same population; and $\sigma_{jj':n}$, the covariance of the j-th and j'-th observations. While exact values are theoretically obtainable for any sample size (n), the most extensive tabulation of $\mu_{j:n}$, $\sigma_{jj:n}$, and $\sigma_{jj':n}$ has been for $n \leq 20$ (Teichroew [1956], and Sarhan and Greenberg [1956]). The complexity involved with large samples prevents a tractable general solution although attempts have been made at simplification for small samples. For example, Bose and Gupta [1959] have shown that the moments and product moments of order statistics can be expressed in terms of elementary functions for $n \leq 5$. Similarly, Godwin [1949] and Jones [1948] have developed elementary formulation with small n. Introduction of various recurrence relations by Govindarajulu [1963] and Srikantan [1962] have reduced the number of calculations needed for evaluation of the moments.

Despite these many efforts tabulation for large n

remains infeasible. Since the proposed estimate most often involves large samples, it is necessary to explore other avenues. One obvious alternative is some kind of approximation. Thus we undertake a brief discussion of existing approximations and a description of the approximation which is used.

If we designate $\Phi(\cdot)$ as the distribution function of the standard normal and $\Psi(\cdot)$ as its inverse function, then most intuitive approximation for $\mu_{j:n}$ is to consider

$$\mu_{j:n} \doteq \Psi\left(\frac{j}{n+1}\right) \tag{A.1}$$

since $p_j = 1-q_j = \frac{j}{n+1}$ is the expectation of the j-th order statistics from a standard uniform distribution. This is an asymptotic solution so that for moderate n the approximation will not be very good. Van Zwet [1964] has suggested several inequalities from which one can find bounds for $\mu_{j:n}$. Blom [1958] proposes a modification to (A.1). That is

$$\mu_{j:n} \doteq \Psi\left(\frac{j-\alpha}{n+1-2\alpha}\right) \tag{A.2}$$

where he suggests using $\alpha = 3/8$ for $n \leq 20$. Harter [1961] has found empirically reasonable values of for other sample sizes.

The approach taken for our approximations was first suggested by Pearson and Pearson [1931] but has since been systematically developed by David and Johnson [1954]. The approximations are based upon a Taylor series expansion of

the j-th order statistics from a standard uniform, denoted by $U_{(j)}$, about $E(U_{(j)})=p_j$. Since $U_{(j)}=\Phi(X_{(j)})$ and thus $X_{(j)}=\Psi(U_{(j)})$, the expansion yields

$$X_{(j)}=\Psi_j+(U_{(j)}-p_j)\Psi_j'+1/2(U_{(j)}-p_j)^2\Psi_j''$$
$$+1/6(U_{(j)}-p_j)^3\Psi_j'''+\ldots \quad . \qquad (A.3)$$

where $\Psi_j\equiv\Psi(p_j)$ and the primes refer to derivatives with respect to p_j. Thus the estimated moments to order $(n+2)^{-2}$ for $1\leq j<j'\leq n$ are

$$\mu_{j:n}\doteq\Psi_j+\frac{p_jq_j}{2(n+2)}\Psi_j''$$
$$+\frac{p_jq_j}{(n+2)^2}[1/3(q_j-p_j)\Psi_j'''+1/8p_jq_j\Psi_j''''] \quad , \qquad (A.4)$$

$$\sigma_{jj:n}\doteq\frac{p_jq_j}{n+2}\Psi_j'^2$$
$$+\frac{p_jq_j}{(n+2)^2}[2(q_j-p_j)\Psi_j'\Psi_j''$$
$$+p_jq_j(\Psi_j'\Psi_j'''+1/2\Psi_j''^2)] \quad , \qquad (A.5)$$

$$\sigma_{jj':n}\doteq\frac{p_jq_j'}{n+2}\Psi_j'\Psi_{j'}'$$
$$+\frac{p_jq_j'}{(n+2)^2}[(q_j-p_j)\Psi_j''\Psi_{j'}'$$
$$+(q_{j'}-p_{j'})\Psi_{j'}''\Psi_j'+1/2p_jq_j\Psi_j'''\Psi_{j'}'$$
$$+1/2p_{j'}q_{j'}\Psi_j'\Psi_{j'}'''+1/2p_jq_j\Psi_j''\Psi_{j'}''] \quad . \qquad (A.6)$$

To find Ψ_j we use the approximation given in Appendix B. Also, for the standard normal distribution we find,

by noting that $p_j=\Phi(\Psi_j)$,

$$\Psi_j'=1/\phi(\Psi_j) \tag{A.7}$$

$$\Psi_j''=\Psi_j/\phi^2(\Psi_j) \tag{A.8}$$

$$\Psi_j'''=(1+2\Psi_j^2)/\phi^3(\Psi_j) \tag{A.9}$$

$$\Psi_j''''=\Psi_j(7+6\Psi_j^2)/\phi^4(\Psi_j) \tag{A.10}$$

where $\phi(x)=\Phi'(x)$ is the standard normal density. The approximations of (A.4)-(A.6) can be expanded to order $(n+2)^{-3}$, however, since n is large, we neglect the modification.

As a word of caution, we must note that for values of j such that j/n is near 0 or 1, the rate of convergence is relatively slow. Despite this, we resist the temptation to use the Blom estimates for $\mu_{j:n}$ to maintain uniformity of method throughout.

APPENDIX B

APPROXIMATIONS FOR THE CUMULATIVE DISTRIBUTION FUNCTION AND THE INVERSE CUMULATIVE DISTRIBUTION FUNCTION OF THE NORMAL DISTRIBUTION

There are several straightforward approximations for the cumulative distribution function (c.d.f.) of the standard normal distribution, which is defined as

$$\Phi(x)=p=(2\Pi)^{-1/2}\int_{-\infty}^{x}\exp(-t^2/2)dt \;;\; -\infty<x<\infty \quad . \qquad (B.1)$$

Hastings [1955] suggests several empirically developed approximations which are intended specifically for high speed computers. Accuracy of these approximations, expressed as the maximum absolute error, varies. As a general rule greater accuracy necessitates a corresponding loss in computer time efficiency, which is measured as the expected time required per unit generated. Burr [1967] develops an approximation from the c.d.f. of a distribution which closely resembles the normal. His results are comparable although slightly less accurate than the Hastings results. This approach does have the advantage of a simple algebraic expression from which a reasonable approximation for the inverse c.d.f., $\Psi(p)$, can be found.

The approximation we use is obtained from one of Hastings' approximations. The formula is found in the Handbook of Mathematical Functions, p. 932 [U.S. Department of Commerce, 1964]. It is expressed as

$$\Phi(x)=1-\phi(x)[r_1u+r_2u^2+r_3u^3] \; ; \; 0\leq x\leq\infty \quad , \tag{B.2}$$

where

$$\phi(x)=d\Phi(x)/dx \tag{B.3}$$

and

$$u=(1+wx)^{-1} \tag{B.4}$$

with

$$\left.\begin{array}{ll} w=0.3326700 & \Gamma_2=-0.1201676 \\ r_1=0.4361836 & \Gamma_3=\ 0.9372980 \end{array}\right\} \quad . \tag{B.5}$$

This approximation has a maximum absolute error of less than 1×10^{-5}. Since it is good only for positive x, a small modification is added to the algorithm so that approximations for negative x may be found.

There are also several approximations for the inverse function, $\Psi(p)$. One of those in current use is the one suggested by Burr, to which we have already referred. Another, suggested by Marsaglia [1958] is relatively fast and accurate, but has the disadvantages of a complicated algorithm and excessive computer storage requirements. The approximation suggested by Muller [1958] also has large storage requirements with accurate results. Hastings [1955] presents an empirically derived approximation which,

despite its simplicity, is relatively slow. Tarter [1968] has suggested an approximation which is obtained from a polynomial expansion in terms of a more workable function. For example, with the half-normal c.d.f. a useful function is -log(1-p). The most accurate results are found by transforming estimates of the half-normal distribution function and using a separate approximation in the tail area.

The algorithm we have chosen is attributed to Wetherill [1965]. It is simple to implement, requires little storage, and is sufficiently fast and accurate for our purposes. The approximation is found from a series solution to a differential equation which is obtained from an approximate series expansion of the inverse function of (B.1). The estimate is

$$\Psi(p)=\sum_{i=1}^{5} a_{2i+1} w^{2i+1} \qquad \text{(B.6)}$$

where

$$w=2p-1 \qquad \text{(B.7)}$$

with

$$\left.\begin{array}{ll} a_1=1.253314 & a_7=\ 0.3630 \\ a_3=0.328117 & a_9=-0.8559 \\ a_5=0.180392 & a_{11}=\ 1.0480 \end{array}\right\} \quad . \qquad \text{(B.8)}$$

In the analytic results of Chapter III and the simulations of Chapter IV the percentiles of the target area observations on X are arbitrarily restricted as

$$F^{-1}(0.05) \leq X^* \leq F^{-1}(0.95)$$

so that over most of this interval the maximum absolute error is less than 0.001 except at the extremes where the error is less than 0.01. Wetherill does provide a special tail area approximation, but it is not used since it was felt that, by using it, the gains in accuracy would not outweigh the loss in computer time efficiency. Moreover, he suggests that, as a rule of thumb, the tail area approximation is not really necessary for $|w|<0.90$.

APPENDIX C

METHODS FOR OBTAINING GROUP BOUNDARIES

In Chapters III and IV we assume specific distributions on the criterion and symptomatic variables. In each case a grouping algorithm must be devised in such a way as to satisfy the criteria suggested in Chapter II. We can simplify the solution to this problem by noticing the parallel between group formation and stratification. In both instances we wish to form groups or strata which are homogeneous within and thus distinguishable or heterogeneous among themselves. This suggests that by finding the distribution-specific "optimum" stratification boundaries we obtain what a reasonable rectilinear-type grouping algorithm might have produced.

We find this approach particularly useful for the simulation done in Chapter IV. To simulate results obtained from relatively large samples requires simplicity wherever possible. Most ordinary grouping algorithms are expensive to run which makes a worthwhile simulation study monetarily prohibitive.

Optimum stratification points, that is, points found subject to the optimization of some characteristic for a single variable (e.g., minimizing the variance of

the final estimate), can be found exactly by an often difficult hit-or-miss iteration scheme or more easily by approximation and without appreciable loss in precision of the final estimate. Sethi [1963] has already obtained the exact optimum stratification points (OSP's) as suggested by Dalenius [1957] for the normal, gamma and beta distributions. Since OSP's vary with method of sample allocation, we have selected those found by assuming Neyman (optimum) allocation which produces the most precise stratified estimate of the population mean. With Neyman allocation OSP's are found so as to minimize the variance of the stratified estimate. Thus if we define $(B_0,B_1,\ldots,B_{k-1},B_k)$ as the OSP's for Neyman allocation where B_0 and B_k are endpoints, $B_1,\ldots,B_{k-1}$ are found such that

$$\frac{(B_g-\mu_g)^2+\sigma_g^2}{\sigma_g} = \frac{(B_g-\mu_{g+1})^2+\sigma_{g+1}^2}{\sigma_{g+1}} \; ; \; g=1,2,\ldots,k-1 \quad . \quad (C.1)$$

To avoid cumbersome and expensive iterations we obtain OSP's for the logistic and exponential distributions by approximation. Two methods which are commonly used are the so-called cumulative $\sqrt{f}$ method attributed to Dalenius and Hodges [1959] and the equal $W_g(b_g-b_{g-1})$ method due to Ekman [1959]. Both are simplified extensions of the criterion used to find the exact points. Cochran [1961], Sethi [1963], Kpedekpo [1970] and Hess et al. [1966] compare the above two approximations with others. The general conclusion is that these two methods are the "best" (as

measured by gains of the stratified over the simple random estimate), and of the two methods, the cum $\sqrt{f}$ is preferred and therefore used here.

To implement the cum $\sqrt{f}$ method on the continuous logistic and exponential distributions we proceed as follows. First, we designate our theoretically infinite populations as being finite and of size 10,000. We then form 200 equally spaced intervals on the ranges of the distributions. For both distributions the ranges are infinite so we arbitrarily select the 0.5-th and 99.5-th percentiles as endpoints for the logistic and the 99-th percentile as the upper endpoint for the exponential distribution (zero is the lower endpoint). Without loss of generality we use the standard forms of both distributions. For the logistic

$$f(x)=\exp(-x)[1+\exp(-x)]^{-2} \; ; \; -\infty<x<\infty \tag{C.2}$$

and thus

$$F(x)=[1+\exp(-x)]^{-1} \tag{C.3}$$

which has

$$\mu=0 \; ; \; \sigma^2=1 \quad . \tag{C.4}$$

For the exponential

$$f(x)=\exp(-x) \; ; \; 0\leq x<\infty \tag{C.5}$$

and thus

$$F(x)=[1-\exp(-x)] \tag{C.6}$$

which has

$$\mu=\sigma=1 \quad . \tag{C.7}$$

The endpoints in this instance are -5.29 and 5.29 for the

logistic and 4.61 for the exponential.

We then find the theoretical frequency in each interval for our hypothetical population. Each interval frequency is weighted up to account for the truncation so that the sum of the intervals equals 10,000. In each interval the cumulated sum of the square root of interval frequencies is recorded. The rule is to choose the B_g so that they create k equal intervals on the cum $\sqrt{f}$ scale. The OSP's for the three distributions are found in Table C.1.

In Chapter IV, when there is more than one variable the units are first grouped into primary strata with respect to one variable, then stratified on a second varable within each primary stratum, and so forth until all variables have been used in the stratification process. This is known as multiple or deep stratification.

A number of procedures have been suggested which provide multivariate OSP's. They differ only in which characteristic of the estimate is being optimized. Dalenius [1957] suggests as a criterion the minimization of a weighted loss function

$$L = \sum_{i=1}^{s} a_i L_i \qquad (C.8)$$

where

$$L_i = \frac{V_i^{(m)} - V_i^{(o)}}{V_i^{(o)}} \quad . \qquad (C.9)$$

TABLE C.1

OPTIMUM STRATIFICATION POINTS FOR THREE DISTRIBUTIONS

Number of Strata (i.e., Groups)	Quantile (i.e., $F(B_g)$) Times 1000 for OSP's Approximated by the Cum f Method									
g=	1	2	3	4	5	6	7	8	9	10
Normal										
2[a]	500									
3[a]	291	709								
4[a]	188	500	812							
10[a]	44	131	242	367	500	633	758	869	956	
20	24	52	90	140	191	243	303	369	439	500
g+10=	561	631	697	757	809	860	910	948	976	
Logistic										
10	44	124	228	358	500	642	772	876	956	
20	21	44	81	124	170	228	289	358	434	500
g+10=	566	642	711	772	830	876	919	956	979	
Exponential										
10	172	328	467	589	698	788	863	921	964	
20	88	172	250	328	399	467	531	589	646	698
g+10=	744	788	827	863	894	921	944	964	979	

[a]Exact OSP's [Sethi, 1963].

The a_i represent measures of importance for the i-th stratification variable, $V_i^{(m)}$ refers to the variance of the estimate by the multivariate OSP's on the i-th variable, and $V_i^{(o)}$ refers to the variance of the estimate by the univariate OSP's on the i-th variable. Murthy [1967] proposes maximization of

$$E = \sum_{i=1}^{s} E_i \qquad (C.10)$$

where

$$E_i = V_i^{(o)} / V_i^{(m)} \quad . \qquad (C.11)$$

Finally, Chakravarthy [1954] and Ghosh [1963] suggest minimization of the generalized variance which is the determinant of the dispersion matrix of the s stratified means. These approaches imply iterative solutions which are theoretically possible but practically infeasible.

To find approximate multivariate OSP's we suggest a procedure in which optimum allocation is done at successive stages of stratification. The OSP's are found for the first variable, then the second variable conditioned on the stratum means of the first variable, then the third given results from the first and second variables, and so on until all variables have been considered. This hierarchical approach is, in principle, not unlike other grouping algorithms (e.g., AID, see Section 5.2).

The stratum means, which involve only the normal distribution, are found by obtaining a formula for the mean

of a truncated normal distribution. If X is distributed as $N(\mu,\sigma^2)$, and W is distributed as a doubly truncated $N(\mu,\sigma^2)$, which is truncated to the left at w_1 and to the right at w_2, then it can be shown (Cohen [1957]) that

$$E(W)=(Q_1-Q_2)\sigma+\mu \tag{C.12}$$

where

$$Q_1=\frac{\phi(z_1)}{\Phi(z_2)-\Phi(z_1)} \quad , \tag{C.13}$$

$$Q_2=\frac{\phi(z_2)}{\Phi(z_2)-\Phi(z_1)} \quad , \tag{C.14}$$

and

$$z_1=\frac{w_1-\mu}{\sigma} \; ; \; z_2=\frac{w_2-\mu}{\sigma} \quad . \tag{C.15}$$

Using a hierarchical scheme, in which the symptomatic variable boundary points are found first, implies that these boundary points will be the same whether or not the criterion variable is included. Thus, although technically we can use the criterion information, we gain little by considering it since the classification boundaries do not change.

If our hierarchical scheme considers the symptomatic variables first, we obtain the boundary points from the bivariate normal distribution. We define the random variables, X_1 and X_2, with specific values, x_1 and x_2, means, μ_1 and μ_2, and covariance matrix

$$V=\begin{pmatrix}\sigma_{11} & \sigma_{12}\\ \sigma_{21} & \sigma_{22}\end{pmatrix} \quad . \tag{C.16}$$

Thus

$$X_1 \sim N(\mu_1,\sigma_{11}) \tag{C.17}$$

and

$$X_2|x_1 \sim N(\mu_2+\frac{\sigma_{12}}{\sigma_{11}}(x_1-\mu_1),\sigma_{22}-\frac{\sigma_{12}^2}{\sigma_{11}}) \quad . \tag{C.18}$$

In the above $X_2|x_1$ reads "random variable X_2 given $X_1=x_1$."

Since X_1 and X_2 are jointly distributed as a bivariate normal in both models of Chapter IV, the boundary points are identical if the population parameters are the same. To define these boundaries we need a few definitions. We define a hierarchical scheme by (S_1,S_2), where S_i is the number of branches formed on variable X_i (i.e., $k=S_1\times S_2$). We also designate B_v as the v-th "optimum" quantile point. For example, from Table C.1 with $S_1=3$, $B_2=\mu_1+\sigma_{11}^{1/2}\Phi^{-1}(0.709)$ $=\mu_1+0.55\sigma_{11}^{1/2}$ for the normal and also from Table C.1, $B_2=\mu_1+\sigma_{11}^{1/2}[\log(0.732)-\log(0.268)]=\mu_1+1.00\sigma_{11}^{1/2}$ for the logistic. Finally, let $\mu_{2.1}^{(v)}$ and $\Gamma_{2.1}^{(v)}$ represent the mean and standard deviation for $X_2|\mu_{(1)}^{(v)}$, where $\mu_{(1)}^{(v)}$ is the mean of the v-th branch on X_1 defined by (C.12). With the scheme (S_1,S_2) the boundaries of the v-th branch of X_1 and the w-th branch of X_2 for all $v=1,2,\ldots,S_1$ and $w=1,2,\ldots,S_2$ are

(1) On X_1 $(\mu_1+B_{v-1}\sigma_{11}^{1/2},\mu_1+B_v\sigma_{11}^{1/2})$ (C.19)

and

$$(2)\ \text{On } X_2\ (\mu_{2.1}^{(v)}+B_{w-1}\sigma_{2.1}^{(v)},\mu_{2.1}^{(v)}+B_w\sigma_{2.1}^{(v)}) \quad . \quad (C.20)$$

APPENDIX D

GENERATORS OF RANDOM NORMAL DEVIATES

Chapter IV involves the generation of samples from two trivariate distributions described by two different models. The linear model involves a trivariate normal, and the nonlinear model involves a bivariate normal distribution on the two symptomatic variables with the condition distribution of the criterion variable given the symptomatic variables as a logistic. A more detailed account of the two models is given in Section 4.2. We now present a general discussion of sample simulation from prespecified multivariate distributions. We also indicate the method used in the Monte Carlo simulations.

The usual procedure for generating an observation from an s-variate distribution with cumulative distribution function (c.d.f.)

$$F(X_1,X_2,\ldots,X_s)\equiv F(\underset{\sim}{X}) \qquad (D.1)$$

is an s-step process. First, x_1 is simulated from the marginal c.d.f. of X_1. Then x_2 is simulated from the conditional c.d.f. of $X_2|x_1$ which reads "random variable X_2 given $X_1=x_1$." In general, x_j is simulated from the conditional c.d.f. of $X_j|x_1,x_2,\ldots,x_{j-1}$. The resulting $\underset{\sim}{x}=(x_1,$

$x_2,\ldots,x_s$) is a simulation of a unit in a population with a random variable X whose c.d.f. is $F(\underset{\sim}{X})$. We note that since

$$F(\underset{\sim}{X})=F_1(X_1)\cdot F_2(X_2|x_1)\cdot F_3(X_3|x_1,x_2)$$
$$\cdot\ldots\cdot F_s(X_s|x_1,x_2,\ldots,x_{s-1}) \quad . \qquad (D.2)$$

The desired simulation is obtained from s independent observations on the basis of the above conditional distributions.

The linear model involves the random vector

$$\underset{\sim}{Z}=\begin{pmatrix}\underset{\sim}{X}\\ Y\end{pmatrix} \qquad (D.3)$$

with specific value

$$\underset{\sim}{z}=\begin{pmatrix}\underset{\sim}{x}\\ y\end{pmatrix} \qquad (D.4)$$

where

$$\underset{\sim}{X}=\begin{pmatrix}X_1\\ X_2\end{pmatrix} \qquad (D.5)$$

with specific value

$$\underset{\sim}{x}=\begin{pmatrix}x_1\\ x_2\end{pmatrix} \qquad (D.6)$$

We assume $\underset{\sim}{Z}$ to be normally distributed with mean vector

$$\underset{\sim}{U}=\begin{pmatrix}\mu_1\\ \mu_2\\ \mu_y\end{pmatrix}=\begin{pmatrix}\underset{\sim}{U}_1\\ U_2\end{pmatrix} \qquad (D.7)$$

and covariance matrix

$$V=\begin{bmatrix} V_{11} & V_{12} \\ V_{12} & V_{22} \end{bmatrix} \quad . \tag{D.8}$$

From this notation the relevant distributions are

$$X_1 \sim N(\mu_1, \sigma_{11}) \tag{D.9}$$

$$X_2|x_1 \sim N(\mu_2 + \frac{\sigma_{12}}{\sigma_{11}}(x_1-\mu_1), \sigma_{22} - \frac{\sigma_{12}^2}{\sigma_{11}}) \tag{D.10}$$

$$Y|x_1,x_2 \sim N(\mu_y + V_{12}'V_{11}^{-1}(x-U_1), \sigma_{yy} - V_{12}'V_{11}^{-1}V_{12}) \quad . \tag{D.11}$$

This implies that an observation from a trivariate normal is obtained from three random univariate normals via the generation of three standard, independent uniforms.

In the nonlinear case the relevant distributions are the same for X_1 and $X_2|x_1$. However, for the third we now have

$$Y|x_1,x_2 \sim \text{Logistic}(a,b) \tag{D.12}$$

or

$$F_3(Y|x_1,x_2) = [1+\exp\{-(a+by)\}]^{-1} \; ; \; -\infty<y<\infty \tag{D.13}$$

where

$$a = \frac{-[60-c\{0.5-[1+\exp\{-(\alpha+\beta x_1+\gamma x_2)\}]^{-1}\}]\Pi}{\sqrt{3}\sigma_{y.12}} \tag{D.14}$$

and

$$b = \frac{\Pi}{\sqrt{3}\sigma_{y.12}} \tag{D.15}$$

where $\sigma_{y.12}$ is the condition standard deviation of $Y|x_1,x_2$

and c, α, β and γ are arbitrary constants.

The two most common methods for obtaining independent standard uniforms are based upon the generation of random sequences of numbers by either of two recurrence relationships. One, due to Lehmer [1951], is called a multiplicative congruential in which

$$r_i = r_{i-1} \cdot \lambda \pmod{M} \qquad (D.16)$$

where λ is some constant. The second is called a mixed congruential since

$$r_i = r_{k-1} \cdot \lambda + \eta \pmod{M} \qquad (D.17)$$

where η is some other constant. This generator is generally attributed to Rotenberg [1960] although he was not the first to consider the idea. The value of the uniform, p_i, is obtained as

$$p_i = r_i/M \; ; \; i=1,2,3,\ldots \quad . \qquad (D.18)$$

Comparisons of mixed and multiplicative generators by Allard et al. [1963], Gannon and Schmittroth [1963] and Kronmal [1964] indicate that, in general, the multiplicative congruential, although slightly slower, demonstrates consistently better statistical properties (e.g., the serial correlation among observations) than the mixed congruential.

Once satisfactory uniformly distributed observations have been generated there are several algorithms available which yield normal deviates. The most obvious

one involves a direct transformation using the standard uniforms and the inverse c.d.f.'s. For the logistic distribution we have the advantage of an exact expression for the inverse c.d.f. of F_3, namely

$$Q_3(p|x_1,x_2)=\frac{\log(p)-\log(1-p)-a}{b} \quad ; \quad 0\leq p\leq 1 \quad . \qquad (D.19)$$

For the standard normal we could use the approximation suggested in Appendix B. Once the inverse of the conditional c.d.f.'s, $F_j(X_j|x_1,x_2,\ldots,x_{j-1})$ in (D.9), are computed, say as $Q_j(p_j|x_1,x_2,\ldots,x_{j-1})$, we apply the s random uniforms, $\underset{\sim}{p}=(p_1,p_2,\ldots,p_s)$, to the Q_j's to yield the simulated observation, $\underset{\sim}{x}$.

Another algorithm uses a linear combination of several uniforms and the Central Limit Theorem to obtain an approximately normally distributed random observation. That is,

$$x=\sum_{i=1}^{12} p_i - 6 \dot{\sim} N(0,1) \quad . \qquad (D.20)$$

Box and Muller [1958] suggest an exact two-for-two inverse method by which two independent uniform random variables, p_1 and p_2, are transformed into two independent standard normals, x_1 and x_2. The transformations are

$$x_1=(-2 \log p_1)^{1/2}\cdot\mathrm{Cos}(2\pi p_2) \qquad (D.21)$$

and

$$x_2=(-2 \log p_1)^{1/2}\cdot\mathrm{Sin}(2\pi p_2) \quad . \qquad (D.22)$$

Sibuya [1961] and Marsaglia and Bray [1964] have similarly modified this transformation to speed up the computing time. The modified transformations are

$$x_1=w_1\cdot\{-2\ \log(w_1^2+w_2^2)/(w_1^2+w_2^2)\}^{1/2} \qquad (D.23)$$

and

$$x_2=x_1\cdot(w_2/z_1) \quad . \qquad (D.24)$$

This requires that $w_1^2+w_2^2<1$ where

$$w_1=2p_1-1 \qquad (D.25)$$

and

$$w_2=2p_2-1 \quad . \qquad (D.26)$$

Finally, we mention an algorithm due to Marsaglia [1961] in which the x axis is divided into equal intervals and the normal deviates are selected according to different formulations which are chosen with varying probabilities. This method is generally fastest but its weakness is its large computer memory requirement.

The random normal generator we use in the simulation studies is due to Chen [1971]. It uses a multiplicative congruential uniform generator which is modified so that two generators are used simultaneously. That is

$$\Gamma_{1i}=\Gamma_{1,i-1}(2^{14}+3)(\text{mod } 2^{31}) \qquad (D.27)$$

and

$$\Gamma_{2i}=[\Gamma_{2,i-1}(2^{18}+3)]\cdot[2^{18}+3](\text{mod } 2^{31}) \qquad (D.28)$$

are used. This is done to reduce serial correlation and is procedurally convenient since the Box and Muller two-for-two transformation is used to obtain pairs of independent normals. Several tests of randomness and normality show the method to be statistically acceptable.

SELECTED BIBLIOGRAPHY

Allard, J. L.; Dobell, A. R.; and Hull, T. E. [1963]. "Mixed Congruential Random Number Generators for Decimal Machines." Journal of the Association of Computing Machines, Vol. 10, pp. 131-141.

Anderson, T. W. [1958]. An Introduction to Multivariate Statistical Analysis. Wiley and Sons, New York.

Blom, G. [1958]. Statistical Estimates and Transformed Beta-Variables. Almqvist and Wiksell, Uppsala, Sweden.

Bogue, D. J. [1950]. "A Technique for Making Extensive Population Estimates." Journal of the American Statistical Association, Vol. 45, pp. 149-163.

Bogue, D. J., and Duncan, B. D. [1959]. "A Composite Method for Estimating Postcensal Population of Small Areas by Age, Sex and Color." National Office of Vital Statistics, Vital Statistics--Special Reports, Vol. 47, No. 6.

Bose, R. C., and Gupta, S. S. [1959]. "Moments of Order Statistics from a Normal Population." Biometrika, Vol. 46, pp. 433-446.

Box, G. F. P., and Muller, M. E. [1958]. "A Note on the Generation of Random Normal Deviates." Annals of Mathematical Statistics, Vol. 29, pp. 231-237.

Burr, I. W. [1967]. "A Useful Approximation to the Normal Distribution Function, with Application to Simulation." Technometrics, Vol. 9, pp. 1-15.

Chakravarty, I. M. [1954]. "On the Problem of Planning a Multistage Survey for Multiple Correlated Characters." Sankhyä, Vol. 14, pp. 211-216.

Chen, E. H. [1971]. "A Random Normal Number Generator for 32-Bit-Word Computers." Journal of the American Statistical Association, Vol. 66, pp. 400-403.

Cochran, W. G. [1961]. "Comparison of Methods for Determining Stratum Boundaries." Bulletin of the International Statistical Institute (Tokyo), Vol. 39, pp. 345-358.

Cohen, A. C. [1957]. "On the Solution of Estimating Equations for Truncated and Censored Samples from Normal Populations." Biometrika, Vol. 44, pp. 225-236.

Cormack, R. M. [1971]. "A Review of Classification." Journal of the Royal Statistical Society, Series A, Vol. 134, pp. 321-353.

Cramér, H. [1963]. Mathematical Methods of Statistics. Princeton University Press, Princeton, N.J.

Dalenius, T. [1957]. Sampling in Sweden. Almqvist and Wiksell, Stockholm, Sweden.

Dalenius, T., and Hodges, J. L., Jr. [1959]. "Minimum Variance Stratification." Journal of the American Statistical Association, Vol. 54, pp. 88-101.

David, F. A. [1970]. Order Statistics. Wiley and Sons, New York.

David, F. N., and Johnson, N. L. [1954]. "Statistical Treatment of Censored Data; I: Fundamental Formulae." Biometrika, Vol. 41, pp. 228-240.

Ehrenberg, A. S. C. [1963]. "Bivariate Regression is Useless." Applied Statistics, Vol. 12, pp. 161-179.

Ekman, G. [1959]. "An Approximation Useful in Univariate Stratification." Annals of Mathematical Statistics, Vol. 30, pp. 219-229.

Ericksen, E. P. [1971]. "A Method for Combining Sample Survey Data and Symptomatic Indicators to Obtain Estimates for Local Units." Unpublished doctoral dissertation, University of Michigan, Ann Arbor.

Ericksen, E. P. [1972]. "Test of a Statistical Procedure for Computing Estimates for Local Areas." Final Report: Joint Statistical Project, Institute for Survey Research, Temple University and United States Bureau of the Census.

Ericksen, E. P. [1973]. "A Method for Combining Sample Survey Data and Symptomatic Indicators to Obtain Population Estimates for Local Areas." Demography, Vol. 10, pp. 137-160.

Frisen, C. M. [1951]. "A Technique for Estimating Population from Morbidity Data." Unpublished doctoral dissertation, Stanford University, Palo Alto, Cal.

Gannon, L. J., and Schmittroth, L. A. [1963]. Computer Generation and Testing of Random Numbers. U.S. Department of Defense Contract AT(10-1).

Ghosh, S. P. [1963]. "Optimum Stratification with Two Characters." Annals of Mathematical Statistics, Vol. 34, pp. 866-872.

Godwin, H. J. [1949]. "Some Low Moments of Order Statistics." Annals of Mathematical Statistics, Vol. 20, pp. 279-285.

Goldberg, D. J.; Feldt, A.; and Smit, J. W. [1960]. "Estimates of Population Change in Michigan: 1950-1960." Michigan Population Studies, No. 1, University of Michigan, Ann Arbor.

Goldberg, D. J., and Balakrishnan, T. R. [1961]. "A Partial Evaluation of Four Estimating Techniques." Michigan Population Studies, No. 2, University of Michigan, Ann Arbor.

Goldberg, D. J.; Rao, V. R.; and Namboodiri, N. K. [1964]. "A Test of Accuracy of the Ratio-Correlation Population Estimates." Land Economics, Vol. 40, pp. 100-103.

Govindarajulu, Z. [1963]. "On Moments of Order Statistics and Quasi-ranges from Normal Populations." Annals of Mathematical Statistics, Vol. 34, pp. 633-651.

Gumbel, E. J. [1960a]. "Multivariate Distributions with Given Margins and Analytic Examples." Bulletin de l'Institut Internationale de Statistique, Vol. 37.

Gumbel, E. J. [1960b]. "Bivariate Exponential Distributions." Journal of the American Statistical Association, Vol. 55, pp. 698-707.

Gumbel, E. J. [1961]. "Bivariate Logistic Distributions." Journal of the American Statistical Association, Vol. 56, pp. 335-349.

Hansen, M. H.; Hurwitz, W. N.; and Madow, W. G. [1953]. Sample Survey Methods and Theory, Vol. 2. Wiley and Sons, New York.

Harter, H. L. [1961]. "Expected Values of Normal Order Statistics." Biometrika, Vol. 31, pp. 1122-1147.

Hastings, C. [1955]. Approximations for Digital Computers. Princeton University Press, Princeton, N.J.

Hess, I.; Sethi, V. K.; and Balakrishnan, T. R. [1966]. "Stratification: A Practical Investigation." Journal of the American Statistical Association, Vol. 61, pp. 74-90.

Hillery, G. A. [1962]. Postcensal Population Estimates for Small Areas: An Evaluation of Selected Techniques. University of Kentucky Agricultural Experiment Station, Lexington.

Jones, H. L. [1948]. "Exact Lower Moments of Order Statistics in Small Samples from a Normal Distribution." Annals of Mathematical Statistics, Vol. 19, pp. 270-273.

Kendall, M. G. [1951]. "Regression, Structure and Functional Relationship--Part I." Biometrika, Vol. 38, pp. 11-25.

Kish, L. [1967]. Survey Sampling. Wiley and Sons, New York.

Kpedekpo, G. M. K. [1970]. "The Efficiency of Optimum Stratification on Frequency Distributions with and without Discontinuities on Zero Classes." Sankhyä, Vol. 32, pp. 1-12.

Kronmall, R. [1964]. "Evaluation of a Pseudo-random Normal Number Generator." Journal of the Association for Computing Machines, Vol. 11, pp. 357-363.

Lehmer, D. H. [1951]. "Mathematical Methods in Large-scale Computing Units." Annals of the Computing Laboratory Harvard University, Vol. 26, pp. 141-146.

Marsaglia, G. [1958]. "Expressing a Random Variable in Terms of Uniform Random Variables." Annals of Mathematical Statistics, Vol. 32, pp. 894-898.

Marsaglia, G. [1961]. Procedures for Generating Normal Random Variables II. Boeing Scientific Research Laboratory, Mathematical Note No. 243.

Marsaglia, G., and Bray, T. A. [1964]. "A Convenient Method for Generating Normal Variables." SIAM Review, Vol. 6, pp. 260-264.

Mises, R. von [1945]. "On the Classification of Observational Data into Distinct Groups." Annals of Mathematical Statistics, Vol. 16, pp. 68-73.

Morgan, J. N., and Sonquist, J. A. [1963]. "Problems in the Analysis of Survey Data, and a Proposal." Journal of the American Statistical Association, Vol. 58, pp. 415-435.

Muller, M. E. [1958]. "An Inverse Method for the Generation of Random Normal Deviates on Large Scale Computers." Mathematical Tables and Other Aids to Computing, Vol. 12, pp. 167-173.

Murthy, M. M. [1967]. Sampling Theory and Methods. Statistical Publishing Society.

National Vital Statstics Division [1962]. Preliminary Report of the Study Group on Postcensal Population Estimates. The Public Health Conference on Records and Statistics, Washington, D.C.

Pearl, R. [1925]. The Biology of Population Growth. Alfred A. Knopf, New York.

Pearson, K., and Pearson, M. V. [1931]. "On the Mean Character and Variance of a Ranked Individual, and on the Mean and Variance of the Intervals between Ranked Individuals; I: Symmetrical Distributions (Normal and Rectangular)." Biometrika, Vol. 23, pp. 364-397.

Rosenberg, H. [1968]. "Improving Current Population Estimates through Stratification." Land Economics, Vol. 44, pp. 331-338.

Rotenberg, A. [1960]. "A New Pseudo-random Number Generator." Journal of the Association of Computing Machines, Vol. 7, pp. 75-77.

Sarhan, A. E., and Greenberg, B. G. [1956]. "Estimation of Location and Scale Parameters by Order Statistics from Singly and Doubly Censored Samples; Part I: The Normal Distribution up to Samples of Size 10." Annals of Mathematical Statistics, Vol. 27, pp. 427-451.

Schmitt, R. C., and Crosetti, A. H. [1954]. "Accuracy of the Ratio-Correlation Method of Estimating Postcensal Population." Land Economics, Vol. 30, No. 3, pp. 279-280.

Sethi, V. K. [1963]. "A Note on Optimum Stratification of Populations for Estimating the Population Means." Australian Journal of Statistics, Vol. 5, pp. 20-33.

Shryock, H. S. [1957]. "Development of Postcensal Population Estimates for Local Areas." In National Bureau of Economic Research, "Studies in Income and Wealth," Vol. 21, Regional Income, Princeton University Press, Princeton, N.J., pp. 377-399.

Slutsky, E. [1925]. Über Stochastische Asymptoten und Grenzwerte, M, 5, No. 3, p. 3.

Snow, E. C. [1911]. "The Application of the Method of Multiple Correlation to the Estimation of Postcensal Populations." Journal of the Royal Statistical Society, Vol. 74, pp. 575-620.

Sonquist, J. A.; Baker, E. L.; and Morgan, J. N. [1971]. Searching for Structure (Alias, AID III), Institute for Social Research, University of Michigan, Ann Arbor.

Srikantan, K. S. [1962]. "Recurrence Relations between the PDF's of Order Statistics, and Some Applications." Annals of Mathematical Statistics, Vol. 33, pp. 169-177.

Starsinsic, D. E., and Zitter, M. [1968]. "Accuracy of the Housing-Unit Method in Preparing Population Estimates for Cities." Demography, Vol. 5, pp. 475-489.

Tarter, M. E. [1968]. "Inverse Cumulative Approximation and Applications." Biometrika, Vol. 55, pp. 29-41.

Tatsuoka, M. M. [1971]. Multivariate Analysis. Wiley and Sons, New York.

Teichroew, D. [1956]. "Tables of Expected Values of Order Statistics and Products of Order Statistics for Samples of Size Twenty or Less from the Normal Distribution." Annals of Mathematical Statistics, Vol. 27, pp. 410-426.

Truett, J.; Cornfield, J.; and Kannel, W. [1967]. "A Mathematical Analysis of the Risk of Coronary Heart Disease in Framingham." Journal of Chronic Diseases, Vol. 20, pp. 511-524.

U.S. Bureau of the Census [1943]. Population, Series P-3, No. 33. U.S. Government Printing Office, Washington, D.C.

U.S. Bureau of the Census [1947]. Population--Special Reports, Series P-47, No. 4. U.S. Government Pringing Office, Washington, D.C.

U.S. Bureau of the Census [1955]. Current Population Reports, Series P-25, No. 116. "Current Status of Population Estimates Prepared by State Agencies." U.S. Government Printing Office, Washington, D.C.

U.S. Bureau of the Census [1957]. Current Population Reports, Series P-25, No. 156. "Estimates of the Population of the New Orleans Standard Metropolitan Area: July 1, 1956." U.S. Government Printing Office, Washington, D.C.

U.S. Bureau of the Census [1962]. Current Population Reports, Series P-25, No. 244. "Local Population Estimates Prepared by State and City Agencies: Mail Survey of 1960." U.S. Government Printing Office, Washington, D.C.

U.S. Bureau of the Census [1966a]. Current Population Reports, Series P-25, No. 328. "Inventory of State and Local Agencies Preparing Population Estimates: Survey of 1965." U.S. Government Printing Office, Washington, D.C.

U.S. Bureau of the Census [1966b]. Current Population Reports, Series P-25, No. 339. "Methods of Population Estimation: Part 1." U.S. Government Printing Office, Washington, D.C.

U.S. Bureau of the Census [1966c]. Directory of Federal Statistics for Local Areas: 1966. U.S. Government Printing Office, Washington, D.C.

U.S. Bureau of the Census [1967]. 1967 County and City Data Book. U.S. Government Printing Office, Washington, D.C.

U.S. Bureau of the Census [1969]. Current Population Reports, Series P-25, No. 427. "Estimates of the Population of Counties and Metropolitan Areas, July 1, 1966: A Summary Report." U.S. Government Printing Office, Washington, D.C.

U.S. Bureau of the Census [1970]. Directory of Non-Federal Statistics for States and Local Areas: 1969. U.S. Government Printing Office, Washington, D.C.

U.S. Bureau of the Census [1972]. 1972 County and City Data Book. U.S. Government Printing Office, Washington, D.C. (in press)

U.S. Department of Commerce [1964]. Handbook of Mathematical Functions. Edited by M. Abramowitz and I. A. Stegum. National Bureau of Standards, Applied Mathematics Series-55.

Watterson, G. A. [1959]. "Linear Estimation in Censored Samples for Multivariate Normal Populations." Annals of Mathematical Statistics, Vol. 30, pp. 814-824.

Wetherill, G. B. [1965]. "An Approximation to the Inverse Normal Function Suitable for the Generation of Random Normal Deviates on Electronic Computers." Applied Statistics, Vol. 14, pp. 201-205.

Whelpton, P. K. [1936]. "An Empirical Method of Calculating Future Population." Journal of the American Statistical Association, Vol. 31, pp. 457-473.

Woodruff, R. S. [1966]. "Use of a Regression Technique to Produce Area Breakdowns of the Monthly National Estimates of Retail Trade." Journal of the American Statistical Association, Vol. 61, pp. 496-504.

Zitter, M., and Shryock, H. S. [1964]. "Accuracy of Methods for Preparing Postcensal Population Estimates for States and Local Areas." Demography, Vol. 1, pp. 227-241.

Zitter, M.; Starsinsic, D.; and Wood, D. [1968]. "Accuracy of Methods of Preparing Population Estimates for Counties: A Summary Compilation of Recent Studies." Paper presented at the 1968 meetings of the Population Association of America.

Zwet, W. R. van [1964]. Convex Transformations of Random Variables. Mathematical Center Tracts, Vol. 7, Mathematisch Centrum, Amsterdam.