1

A

37582

National Library
of CanadaBibliothèque nationale
du CanadaCANADIAN THESES
ON MICROFICHETHÈSES CANADIENNES
SUR MICROFICHENAME OF AUTHOR/NOM DE L'AUTEUR IAN KEITH MACHILLANTITLE OF THESIS/TITRE DE LA THÈSE A STUDY OF CONTINGENCY TABLES
WITH GIVEN MARGINALSUNIVERSITY/UNIVERSITÉ CARLETON UNIVERSITYDEGREE FOR WHICH THESIS WAS PRESENTED/
GRADE POUR LEQUEL CETTE THÈSE FUT PRÉSENTÉE MASTERS DE SCIENCEYEAR THIS DEGREE CONFERRED/ANNÉE D'OBTENTION DE CE DEGRÉ 1978 (PROSPECTED)NAME OF SUPERVISOR/NOM DU DIRECTEUR DE THÈSE DR. J. N. K. RAO

Permission is hereby granted to the NATIONAL LIBRARY OF
CANADA to microfilm this thesis and to lend or sell copies
of the film.

L'autorisation est, par la présente, accordée à la BIBLIOTHÈ-
QUE NATIONALE DU CANADA de microfilmer cette thèse et
de prêter ou de vendre des exemplaires du film.

The author reserves other publication rights, and neither the
thesis nor extensive extracts from it may be printed or other-
wise reproduced without the author's written permission.

L'auteur se réserve les autres droits de publication; ni la
thèse ni de longs extraits de celle-ci ne doivent être imprimés
ou autrement reproduits sans l'autorisation écrite de l'auteur.

DATED/DATE JUNE 5 1977 SIGNED/SIGNÉ Ian MachillanPERMANENT ADDRESS/RÉSIDENCE FIXE 507 RIVERDALE AVE, OTTAWA ONT. K1S 1S4

National Library of Canada

Cataloguing Branch
Canadian Theses Division

Ottawa, Canada
K1A 0N4

Bibliothèque nationale du Canada

Direction du catalogage
Division des thèses canadiennes

NOTICE

The quality of this microfiche is heavily dependent upon the quality of the original thesis submitted for microfilming. Every effort has been made to ensure the highest quality of reproduction possible.

If pages are missing, contact the university which granted the degree.

Some pages may have indistinct print especially if the original pages were typed with a poor typewriter ribbon or if the university sent us a poor photocopy.

Previously copyrighted materials (journal articles, published tests, etc.) are not filmed.

Reproduction in full or in part of this film is governed by the Canadian Copyright Act, R.S.C. 1970, c. C-30. Please read the authorization forms which accompany this thesis.

**THIS DISSERTATION
HAS BEEN MICROFILMED
EXACTLY AS RECEIVED**

AVIS

La qualité de cette microfiche dépend grandement de la qualité de la thèse soumise au microfilmage. Nous avons tout fait pour assurer une qualité supérieure de reproduction.

S'il manque des pages, veuillez communiquer avec l'université qui a conféré le grade.

La qualité d'impression de certaines pages peut laisser à désirer, surtout si les pages originales ont été dactylographiées à l'aide d'un ruban usé ou si l'université nous a fait parvenir une photocopie de mauvaise qualité.

Les documents qui font déjà l'objet d'un droit d'auteur (articles de revue, examens publiés, etc.) ne sont pas microfilmés.

La reproduction, même partielle, de ce microfilm est soumise à la Loi canadienne sur le droit d'auteur, SRC 1970, c. C-30. Veuillez prendre connaissance des formules d'autorisation qui accompagnent cette thèse.

**LA THÈSE A ÉTÉ
MICROFILMÉE TELLE QUE
NOUS L'AVONS REÇUE**

A STUDY OF
CONTINGENCY TABLES
WITH GIVEN MARGINALS

by

IAN KEITH MACMILLAN

DEPARTMENT OF MATHEMATICS

MAY 1977

CARLETON UNIVERSITY

The undersigned recommend to the Faculty
of Graduate Studies acceptance of the thesis

'A Study of Contingency Tables
with Given Marginals'

submitted by Ian Keith MacMillan, B.Sc.,
in partial fulfilment of the requirements
for the degree of Master of Science.

JNKlao

.....
Thesis Supervisor

L. A. Denson

.....
Chairman, Department of Mathematics

Carleton University

ABSTRACT

The problem of estimating population cell proportions in two-way tables for which the population marginal proportions are known, is investigated. Four estimators of the population cell proportions are considered. Two of these estimators, the Raking Ratio Estimator and the Linear Correction Estimator, require iterative calculations for their solution, while the remaining two estimators, the Method of Proportional Distribution of Marginal Adjustments and the Method of Array Proportions, are calculated exactly. Various statistical properties, notably the asymptotic variances, are derived and examined for the two iterative methods at each iteration. These asymptotic variances are found to be the same for both procedures. Statistical properties of the two non-iterative estimators are also discussed. Convergence characteristics of the iterative methods are considered and situations where the procedures cannot converge are identified. Finally, properties of all the four estimators, as well as the maximum likelihood estimator of the population cell proportions, are investigated for the case when the rows and columns of the two-way table are assumed independent.

TABLE OF CONTENTS

	<u>PAGE</u>
ACKNOWLEDGEMENTS	v
INTRODUCTION	vi
CHAPTER 1	1
Review of the Literature	
CHAPTER 2	7
Raking Ratio Estimation Procedure	
CHAPTER 3	19
Linear Correction Estimation Procedure	
CHAPTER 4	34
Some Statistical Properties of RREP and LCEP	
CHAPTER 5	51
Convergence Properties of RREP and LCEP when	
Some Sample Cell Frequencies are Zero	
CHAPTER 6	63
Two Other Estimators of the Population Cell	
Proportions	
CHAPTER 7	76
Properties of the Estimators when the Row and	
Column Characteristics are Independent	
CHAPTER 8	85
Summary and Conclusions	
APPENDICES	87
BIBLIOGRAPHY	103

ACKNOWLEDGEMENTS

I wish to thank Professor J. N. K. Rao for his valuable guidance and assistance in the preparation of this paper. I also wish to thank Mr. G. J. Brackstone, Director of the Census Survey Methods Division of Statistics Canada, for introducing me to the topic of this thesis during a co-op work term. Thanks also go to Mr. M. D. Bankier for his help with the empirical results and to Susan Villeneuve for the typing.

INTRODUCTION

It is sometimes the case that sample data from one source must be adjusted to comply with established results from other sources. In effect, such an adjustment makes use of the information from these other sources and often produces 'better' estimates than had this information not been used. However, it is not always clear as to how this auxiliary data is to be incorporated into the estimation process.

For example, consider a two-way table of r rows and c columns for which the row and column proportions $p_{i.} = \sum_j p_{ij}$, $p_{.j} = \sum_i p_{ij}$ are known for the population, but the individual cell proportions p_{ij} are not. A simple random sample of size n is taken to obtain sample cell frequencies n_{ij} so that the estimates $\pi_{ij} = n_{ij}/n$ of the cell proportions p_{ij} are available. Even though π_{ij} itself is an unbiased estimator of p_{ij} it does not make use of the available marginal information. In addition, $\pi_{i.} = \sum_j \pi_{ij}$ and $\pi_{.j} = \sum_i \pi_{ij}$ will not equal the known marginals $p_{i.}$ and $p_{.j}$ except by accident.

The problem at hand, therefore, is to construct estimates p_{ij}^* of p_{ij} that satisfy the marginal constraints

$$\sum_i p_{ij}^* = p_{.j} \quad r = 1, \dots, c-1$$

$$\sum_j p_{ij}^* = p_{i.} \quad r = 1, \dots, r-1$$

$$\sum_i \sum_j p_{ij}^* = 1.$$

This is a system of $r + c - 1$ linearly independent equations in rc unknowns which has infinitely many feasible solutions p_{ij}^* such that $0 \leq p_{ij}^* \leq 1$. In view of this we would require a procedure that yields a solution which is in some way optimal. Two of the methods considered in this paper provide solutions which minimize or maximize some function with known statistical properties subject to the linear constraints. Both of these estimates are asymptotically efficient for large n .

Such optimal properties, however, are of little benefit if the estimators themselves cannot be readily calculated and the above two methods require calculation by iterative methods (albeit simple ones). Thus, two easily computable non-iterative methods which produce estimates satisfying the marginals exactly are also discussed. However, these estimates do not possess optimal properties except in special cases.

The purpose of this paper is to present the four estimation procedures described above and to examine some of their characteristics.

A review of the relevant literature for two-way tables is presented in Chapter 1. Chapter 2 gives a widely used iterative method known as the Raking Ratio Estimation Procedure (RREP) and a description of some of its properties. Another iterative method known in this paper as the Linear Correction Estimation Procedure (LCEP) is described in Chapter 3. This procedure shares several properties with the RREP which are discussed in Chapter 4 and subsequent chapters. Some of the convergence characteristics of the RREP and LCEP are examined in Chapter 5. Chapter 6 includes an examination of two non-iterative estimators and derives some of their statistical properties. Efficiency comparisons with the iterative methods in the special case where the row and column characteristics are independent are presented in Chapter 7. Chapter 8 is a summary of the results of the preceding chapters. Finally, several appendices include some of the more lengthy proofs of various propositions as well as examples of the RREP and LCEP.

CHAPTER 1

REVIEW OF THE LITERATURE

The problem of estimating cell frequencies for contingency tables with given marginals was first discussed by Deming and Stephan [6]. They reasoned that 'good' estimates q_{ij}^* of the population cell proportions p_{ij} could be generated by minimizing the modified χ^2 statistic

$$\phi = \sum_i \sum_j \frac{(n_{ij} - n q_{ij})^2}{n_{ij}} \quad (1.1)$$

with respect to q_{ij} , where the q_{ij} are subject to the marginal restrictions,

$$\sum_i q_{ij} = p_{.j} \quad j = 1, \dots, c \quad (1.2)$$

$$\sum_j q_{ij} = p_{i.} \quad i = 1, \dots, r$$

They demonstrated that the problem is easily solved by minimizing the Lagrangian function

$$\phi + 2n a_i (\sum_j q_{ij} - p_{i.}) + 2n b_j (\sum_i q_{ij} - p_{.j}) \quad (1.3)$$

with respect to q_{ij} which yields the solution

$$q_{ij}^* = \pi_{ij} (1 + a_i + b_j) \quad \begin{matrix} i = 1, \dots, r \\ j = 1, \dots, c \end{matrix} \quad (1.4)$$

where the a_i and b_j are Lagrangian multipliers. The a_i and b_j can be found exactly by substituting (1.4) into (1.2) and solving the resulting linear system. Deming and Stephan demonstrate that this technique is readily extended to tables of higher dimension.

A great deal more hand calculation was required in solving a system of linear equations when Deming and Stephan produced their results in 1940 than is the case today. With this in mind, they investigated a rapidly converging, easily computed iterative procedure which is now known as the Raking Ratio Estimation Procedure (RREP).

They incorrectly stated that this procedure yields the q_{ij}^* of (1.4).

Stephan [19] corrected this error and presented another iterative procedure for calculating the q_{ij}^* of (1.4). This procedure has two notable attributes. First, as with the RREP it is easily extended to tables of higher dimension. Second, it is applicable not only to (1.1) but to functions of the more general form

$$\sum_i \sum_j \frac{(n_{ij} - n q_{ij})^2}{c_{ij}} \quad (1.5)$$

32

where the c_{ij} are weights that do not depend on q_{ij} . Note that setting $c_{ij} = n_{ij}$ yields (1.1). In its most general form, the procedure requires starting values for the process as well as certain parameters at each iteration. If these latter parameters are judiciously selected the procedure converges rapidly. However, the procedure is more straightforward and more easily programmable on a computer if certain constants are assigned to these parameters for all iterations, and convergence is still quite rapid. Stephan's procedure with these constants is discussed in Chapter 3.

Smith [18] discussed the problem of contingency tables with given marginals at some length. He derived the maximum likelihood estimates of p_{ij} by minimizing the log likelihood function

$$\sum_i \sum_j n_{ij} \ln q_{ij} \quad (1.6)$$

with respect to q_{ij} where the q_{ij} are subject to (1.2). The method of Lagrange multipliers yields minimizing values

$$m_{ij}^* = \frac{n_{ij}}{(a_i + b_j)} \quad \begin{array}{l} i = 1, \dots, r \\ j = 1, \dots, c \end{array} \quad (1.7)$$

where the a_i and b_j are Lagrange multipliers. The substitution of (1.7) into (1.2) results in a non-linear system of equations in the

a_i and b_j whose solution requires iterative methods. Smith presented such a method, as did Saleh [20], but neither made an investigation of the convergence properties. Realizing that simplicity of calculation is a major factor in the usefulness of any estimator, Smith presented a non-iterative estimator which satisfies the marginals exactly. This estimator is discussed in Chapters 6 and 7 of this paper.

Kullback [11], using a result of Neyman, proved that the q_{ij}^* of (1.4) are asymptotically efficient (for large samples) or Best Asymptotically Normal (BAN). In the same paper, Kullback also proved that estimates produced by the RREP are also BAN.

El-Badry and Stephan [7] presented a non-iterative estimator which they showed to be asymptotically more efficient than that of Smith [18] discussed above. This estimator is examined in Chapters 6 and 7 of this paper.

In 1961, Friedlander [8] discussed an estimation method which had been used for some time at the General Registry Office in Wales, but the exact origin of which is unknown. This procedure is known as the Linear Correction Estimation Procedure (LCEP) in this paper and is discussed in Chapters 3, 4, 5 and 7.

Ireland and Kullback [11] provided a proof of convergence for the RREP, although their proof requires all n_{ij} to be positive.

They also showed that the values produced by the RREP minimize the discrimination information function

$$\sum_i \sum_j q_{ij} \ln \frac{q_{ij}}{n_{ij}} \quad (1.9)$$

with respect to q_{ij} where the q_{ij} are subject to (1.2). Details of the RREP are presented in Chapters 2, 4, 5, and 7.

Ireland [12] dealt with some of the computational aspects associated with the RREP. Among the problems discussed were computer execution times and convergence problems when some of the n_{ij} 's are zero. This latter problem is discussed in Chapter 5 of this paper.

The RREP is used at Statistics Canada in analyzing Census data. Much work has been done in establishing various statistical properties (notably the bias and variance) of the estimators produced in the first four iterations. Unpublished papers by J. N. K. Rao [15], [16], [17], Brackstone and Arora [2], [3] and Saleh [20] provide much of the material on which this thesis is based.

Procedures similar to the RREP have also been used for other problems. Bishop and Feinberg [4], and Goodman [10] used a version of the RREP in cases where the population marginal totals were unknown but where the row and column characteristics were assumed to

be independent. In such cases, the maximum likelihood estimator for the population cell value p_{ij} is the product of the sample marginals $\pi_{i.}$ and $\pi_{.j}$. However, there may be cells for which it is known a priori that $p_{ij} = 0$ in which case $\pi_{i.}\pi_{.j}$ is no longer an m.l.e. A modified version of the RREP is used to produce the maximum likelihood estimates for this constrained problem. This procedure is as follows: set

$$e_{ij}^{(0)} = \begin{cases} \frac{1}{n} & \text{if } p_{ij} \neq 0 \\ 0 & \text{otherwise} \end{cases}$$

At the $2t^{\text{th}}$ step, let

$$e_{ij}^{(2t-1)} = e_{ij}^{(2t-2)} \frac{\pi_{i.}}{\sum_j e_{ij}^{(2t-2)}}$$

and at the $(2t+1)^{\text{th}}$ step, let

$$e_{ij}^{(2t)} = e_{ij}^{(2t-1)} \frac{\pi_{.j}}{\sum_i e_{ij}^{(2t-1)}}$$

The estimates of the p_{ij} are the e_{ij}^* to which the process converges.

Notice that if $p_{ij} = 0$ for any cell, then $e_{ij}^{(t)} = 0$ for all t and therefore

$$e_{ij}^* = 0.$$

CHAPTER 2

RAKING RATIO ESTIMATION PROCEDURE

2.1 Derivation of the RREP.

Consider the problem of minimizing the discrimination information function

$$\sum_{i=1}^r \sum_{j=1}^c q_{ij} \ln \frac{q_{ij}}{\pi_{ij}} \quad (2.1)$$

with respect to q_{ij} , subject to the equality constraints

$$\sum_i q_{ij} = p_{.j} \quad j = 1, \dots, c \quad (2.2)$$

$$\sum_j q_{ij} = p_{i.} \quad i = 1, \dots, r$$

The solution can be found by minimizing the Lagrangian

$$\sum_i \sum_j q_{ij} \ln \frac{q_{ij}}{\pi_{ij}} - \lambda_i (\sum_j q_{ij} - p_{i.}) - \mu_j (\sum_i q_{ij} - p_{.j}) \quad (2.3)$$

where λ_i and μ_j are the Lagrangian multipliers of the i^{th} row and j^{th} column constraint respectively. Differentiating (2.3)

with respect to q_{ij} and setting the result to zero yields

$$\begin{aligned}
 p_{ij}^* &= e^{-1} e^{\lambda_i} e^{\mu_j} \pi_{ij} \\
 &= a_i b_j \pi_{ij}
 \end{aligned}
 \tag{2.4}$$

Since (2.4) satisfies (2.2) we have

$$p_{i.} = a_i \sum_j b_j \pi_{ij} \tag{2.5}$$

$$p_{.j} = b_j \sum_i a_i \pi_{ij} \tag{2.6}$$

which is a non-linear system in the a_i 's and b_j 's. The form of (2.5) and (2.6) is particularly simple and motivates the following iterative solution known as the Raking Ratio Estimation Procedure (RREP).

The RREP will now be described. Let $a_i^{(t)}$ and $b_j^{(t)}$ denote the t^{th} iterated estimates of a_i and b_j and let $p_{ij}^{(2t-1)} = a_i^{(t)} b_j^{(t)} \pi_{ij}$ and $p_{ij}^{(2t)} = a_i^{(t)} b_j^{(t+1)} \pi_{ij}$ denote the $(2t-1)^{\text{th}}$ and $2t^{\text{th}}$ iterated estimates of p_{ij}^* . Set

$$a_i^{(0)} = b_j^{(1)} = 1 \tag{2.7}$$

which yields

$$p_{ij}^{(0)} = a_i^{(0)} b_j^{(1)} \pi_{ij} = \pi_{ij} \quad (2.8)$$

Succeeding $a_i^{(t)}$ and $b_j^{(t)}$ are obtained by rewriting (2.5) and (2.6) in terms of the estimators of a_i and b_j :

$$a_i^{(t)} = p_{i.} / \sum_j b_j^{(t)} \pi_{ij} \quad (2.9)$$

$$b_j^{(t)} = p_{.j} / \sum_i a_i^{(t-1)} \pi_{ij} \quad (2.10)$$

Therefore,

$$\begin{aligned} p_{ij}^{(2t-1)} &= a_i^{(t)} b_j^{(t)} \pi_{ij} \\ &= \frac{p_{i.} b_j^{(t)} \pi_{ij}}{\sum_j b_j^{(t)} \pi_{ij}} \\ &= \frac{p_{i.} a_i^{(t-1)} b_j^{(t)} \pi_{ij}}{\sum_j a_i^{(t-1)} b_j^{(t)} \pi_{ij}} \\ &= \frac{p_{i.}}{p_{i.}^{(2t-2)}} p_{ij}^{(2t-2)} \quad (2.11) \end{aligned}$$

Similarly, from (2.10) we get

$$p_{ij}^{(2t)} = \frac{p_{.j}}{p_{.j}^{(2t-1)}} p_{ij}^{(2t-1)} \quad (2.12)$$

2.2. Properties of the RREP.

For the first three iterations, we have (assuming $\pi_{i.} > 0$ and $\pi_{.j} > 0$ for all i and j)

$$p_{ij}^{(0)} = \pi_{ij}$$

$$p_{ij}^{(1)} = \frac{p_{i.}}{\pi_{i.}} \pi_{ij}$$

$$p_{ij}^{(2)} = \frac{p_{.j}}{\sum_i \frac{p_{i.}}{\pi_{i.}} \pi_{ij}} \frac{p_{i.}}{\pi_{i.}} \pi_{ij}$$

so that at the zeroth iteration, no marginal information is used; at the first iteration the row marginals are used, whereas at the second iteration both row and column marginals are used. The form of (2.11) and (2.12) shows that the row and column information is used in alternate iterations.

By this procedure, one set of marginals is satisfied at each iteration. For example,

$$\begin{aligned}
 \sum_j p_{ij}^{(2t-1)} &= \sum_j \frac{p_{i.}}{p_{i.}^{(2t-2)}} p_{ij}^{(2t-2)} \\
 &= \frac{p_{i.}}{p_{i.}^{(2t-2)}} p_{i.}^{(2t-2)} \\
 &= p_{i.} \quad .
 \end{aligned}
 \tag{2.13}$$

Similarly,

$$\sum_i p_{ij}^{(2t)} = p_{.j} \quad .$$

This provides a simple way of checking calculations, especially when they are done by hand.

The choice of starting values was purely arbitrary. We could as well set $a_i^{(1)} = 1$, $b_j^{(0)} = 1$ and define $p_{ij}^{(2t-1)} = a_i^{(t)} b_j^{(t)} \pi_{ij}$ and $p_{ij}^{(2t)} = a_i^{(t+1)} b_j^{(t)} \pi_{ij}$. Use of recursion relations similar to (2.9) and (2.10) yield

$$p_{ij}^{(2t-1)} = \frac{p_{.j}}{p_{.j}^{(2t-2)}} p_{ij}^{(2t-2)} \tag{2.16}$$

$$p_{ij}^{(2t)} = \frac{p_{i.}}{p_{i.}^{(2t-1)}} p_{ij}^{(2t-1)} \tag{2.17}$$

Note that (2.16) and (2.17) are similar to (2.11) and (2.12) except that where the row constraints are satisfied by one, the

column constraints are satisfied by the other.

Ireland and Kullback [11] used the properties of the minimum discrimination function (2.1) to show that $p_{ij}^{(t)}$ converges to p_{ij}^* if all $n_{ij} > 0$. This proof holds irrespective of the constraints satisfied at the first iteration.

Note that the RREP cannot converge to a negative value.

This property follows from equations (2.11) and (2.12) since $p_{ij}^{(0)} = \pi_{ij} \geq 0$. It is obviously a necessary property for an estimator of cell proportions and it is mentioned here to contrast the RREP with the Linear Correction Estimation Procedure (LCEP) discussed in the next chapter.

We now consider an example of a 2×2 table to illustrate the convergence properties and an undesirable feature of the RREP.

Consider minimizing

$$\sum_i \sum_j q_{ij} \ln \frac{q_{ij}}{\pi_{ij}} \quad (2.19)$$

with respect to q_{ij} , subject to the three linearly independent constraints

$$q_{11} + q_{12} = p_1.$$

$$q_{21} + q_{22} = p_2.$$

(2.20)

$$q_{11} + q_{21} = p_{.1}$$

Using (2.20) to express the remaining q_{ij} in terms of q_{11} , substituting into (2.19), differentiating with respect to q_{11} and setting the result to zero yields

$$\ln \left[\frac{p_{.1} (p_{2.} - p_{.1} + p_{11}^*) \pi_{12} \pi_{21}}{(p_{.1} - p_{11}^*) (p_{1.} - p_{11}^*) \pi_{11} \pi_{22}} \right] = 0. \quad (2.21)$$

This leads to the solution

$$p_{11}^* = \frac{-B \pm (B^2 + 4(K-1)p_{1.}p_{.1})^{\frac{1}{2}}}{2(K-1)} \quad (2.22)$$

where

$$K = \frac{\pi_{12} \pi_{21}}{\pi_{11} \pi_{22}}$$

$$B = Kp_{2.} - Kp_{.1} + p_{1.} + p_{.1}$$

From (2.21) it is seen that p_{11}^* depends on the π_{ij} 's only through the 'cross ratio' $(\pi_{12}\pi_{21})/(\pi_{11}\pi_{22})$. Thus, for a given set of marginals we might take two samples of the same size yielding very different π_{ij} 's and yet get the same p_{ij}^* in both cases. For example, suppose the marginals are $p_{1.} = .25$, $p_{2.} = .75$, $p_{.1} = .35$, $p_{.2} = .65$ and suppose that sample 1 yields $\pi_{11} = .1$, $\pi_{12} = .2$, $\pi_{21} = .2$, $\pi_{22} = .5$ and sample 2 yields $\pi_{11} = .5$, $\pi_{12} = .2$, $\pi_{21} = .2$, $\pi_{22} = .1$. The results of the RREP for these two samples are as follows:

Sample 1

Iteration 0

$p_{11}^{(0)} = .1$	$p_{12}^{(0)} = .2$	$p_{1.}^{(0)} = .3$
$p_{21}^{(0)} = .2$	$p_{22}^{(0)} = .5$	$p_{2.}^{(0)} = .7$
$p_{.1}^{(0)} = .3$	$p_{.2}^{(0)} = .7$	

Iteration 1

$$\begin{array}{lll}
 p_{11}^{(1)} = .0833 & p_{12}^{(1)} = .1667 & p_{1\cdot}^{(1)} = .2500 \\
 p_{21}^{(1)} = .2143 & p_{22}^{(1)} = .5357 & p_{2\cdot}^{(1)} = .7500 \\
 p_{\cdot 1}^{(1)} = .2976 & p_{\cdot 2}^{(1)} = .7024 &
 \end{array}$$

Iteration 2

$$\begin{array}{lll}
 p_{11}^{(2)} = .0980 & p_{12}^{(2)} = .1542 & p_{1\cdot}^{(2)} = .2522 \\
 p_{21}^{(2)} = .2520 & p_{22}^{(2)} = .4958 & p_{2\cdot}^{(2)} = .7478 \\
 p_{\cdot 1}^{(2)} = .3500 & p_{\cdot 2}^{(2)} = .6500 &
 \end{array}$$

Iteration 3

$$\begin{array}{lll}
 p_{11}^{(3)} = .0971 & p_{12}^{(3)} = .1529 & p_{1\cdot}^{(3)} = .2500 \\
 p_{21}^{(3)} = .2528 & p_{22}^{(3)} = .4972 & p_{2\cdot}^{(3)} = .7500 \\
 p_{\cdot 1}^{(3)} = .3499 & p_{\cdot 2}^{(3)} = .6501 &
 \end{array}$$

Iteration 4

$$\begin{array}{lll}
 p_{11}^{(4)} = .0972 & p_{12}^{(4)} = .1528 & p_{1\cdot}^{(4)} = .2500 \\
 p_{21}^{(4)} = .2528 & p_{22}^{(4)} = .4972 & p_{2\cdot}^{(4)} = .7500 \\
 p_{\cdot 1}^{(4)} = .3500 & p_{\cdot 2}^{(4)} = .6500 &
 \end{array}$$

Sample 2

Iteration 0

$$p_{11}^{(0)} = .5$$

$$p_{12}^{(0)} = .2$$

$$p_{1\cdot}^{(0)} = .7$$

$$p_{21}^{(0)} = .2$$

$$p_{22}^{(0)} = .1$$

$$p_{2\cdot}^{(0)} = .3$$

$$p_{\cdot 1}^{(0)} = .7$$

$$p_{\cdot 2}^{(0)} = .3$$

Iteration 1

$$p_{11}^{(1)} = .1785$$

$$p_{12}^{(1)} = .0715$$

$$p_{1\cdot}^{(1)} = .2500$$

$$p_{21}^{(1)} = .5000$$

$$p_{22}^{(1)} = .2500$$

$$p_{2\cdot}^{(1)} = .7500$$

$$p_{\cdot 1}^{(1)} = .6785$$

$$p_{\cdot 2}^{(1)} = .3215$$

Iteration 2

$$p_{11}^{(2)} = .0921$$

$$p_{12}^{(2)} = .1445$$

$$p_{1\cdot}^{(2)} = .2366$$

$$p_{21}^{(2)} = .2579$$

$$p_{22}^{(2)} = .5055$$

$$p_{2\cdot}^{(2)} = .7634$$

$$p_{\cdot 1}^{(2)} = .3500$$

$$p_{\cdot 2}^{(2)} = .6500$$

Iteration 3

$$p_{11}^{(3)} = .0973$$

$$p_{12}^{(3)} = .1527$$

$$p_{1\cdot}^{(3)} = .2500$$

$$p_{21}^{(3)} = .2534$$

$$p_{22}^{(3)} = .4967$$

$$p_{2\cdot}^{(3)} = .7500$$

$$p_{\cdot 1}^{(3)} = .3567$$

$$p_{\cdot 2}^{(3)} = .6493$$

Iteration 4

$$\begin{array}{lll}
 p_{11}^{(4)} = .0972 & p_{12}^{(4)} = .1528 & p_{1.}^{(4)} = .2500 \\
 p_{21}^{(4)} = .2528 & p_{22}^{(4)} = .4972 & p_{2.}^{(4)} = .7500 \\
 p_{.1}^{(4)} = .3500 & p_{.2}^{(4)} = .6500 &
 \end{array}$$

Note that convergence is quite rapid in both cases, yielding four-digit accuracy in the marginals after four iterations. Note also that one set of marginal constraints is satisfied at each iteration after the zeroth.

For both samples, (2.22) yields the solution $p_{11}^* = .0972$ even though the π_{ij} -pattern for sample 2 is very different from that of sample 1. This undesirable feature persists for larger tables involving more than two rows or columns, but to a lesser extent. For instance, consider a 3×2 table. If we express the unknowns in terms of p_{11}^* and p_{22}^* , make substitutions in (2.1) and differentiate with respect to p_{11}^* and p_{22}^* , we get

$$\ln \left[\frac{\pi_{12}\pi_{31}p_{11}^*(p_{3.}-p_{.1}+p_{2.}+p_{22}^*+p_{11}^*)}{\pi_{11}\pi_{32}(p_{1.}-p_{11}^*)(p_{.1}-p_{2.}+p_{22}^*-p_{11}^*)} \right] = 0 \quad (2.23)$$

and

$$\ln \left[\frac{\pi_{32}\pi_{21}}{\pi_{22}\pi_{31}} \frac{p_{22}^* (p_{.1} - p_{2.} + p_{22}^* - p_{11}^*)}{(p_{2.} - p_{22}^*)(p_{3.} - p_{.1} + p_{2.} - p_{22}^* + p_{11}^*)} \right] = 0 \quad (2.24)$$

A sufficient condition for the p_{ij}^* to remain the same for different π_{ij} -patterns is that the ratios $(\pi_{12}\pi_{31})/(\pi_{11}\pi_{32})$ and $(\pi_{32}\pi_{21})/(\pi_{22}\pi_{31})$ remain the same. This may be impossible to achieve for samples of the same size unlike the 2×2 case but for samples of different sizes it is clearly feasible.

CHAPTER 3

LINEAR CORRECTION ESTIMATION PROCEDURE

3.1. Derivation of the LCEP.

Consider minimizing the modified χ^2 statistic

$$\sum_{i=1}^r \sum_{j=1}^c (n_{ij} - nq_{ij})^2 / n_{ij} \quad (3.1)$$

with respect to q_{ij} , subject to the equality constraints

$$\begin{aligned} \sum_i q_{ij} &= p_{.j} & j = 1, \dots, c \\ \sum_j q_{ij} &= p_{i.} & i = 1, \dots, r \end{aligned} \quad (3.2)$$

As in Chapter 2, the solution to (3.1) subject to (3.2) denoted by q_{ij}^* , can be found by minimizing the Lagrangian function

$$\sum_{i=1}^r \sum_{j=1}^c (n_{ij} - nq_{ij})^2 / n_{ij} + \lambda_i (\sum_j q_{ij} - p_{i.}) + \mu_j (\sum_i q_{ij} - p_{.j}) \quad (3.3)$$

where λ_i and μ_j are the Lagrangian multipliers of the i^{th} row and j^{th} column constraint respectively. Differentiating (3.3) with respect to q_{ij} and setting the result to zero yields

$$2nq_{ij}^* = n_{ij}(2n - \lambda_i - \mu_j)$$

or

$$q_{ij}^* = \pi_{ij} (1 + a_i + b_j) \quad (3.4)$$

Substituting (3.5) into (3.2) gives

$$p_{i.} = \pi_{i.} + a_i \pi_{i.} + \sum_j b_j \pi_{ij} \quad (3.5)$$

$$p_{.j} = \pi_{.j} + b_j \pi_{.j} + \sum_i a_i \pi_{ij} \quad (3.6)$$

Stephan [19] presented an iterative procedure for the solution

to (3.4) - (3.6), a slightly modified version of which is as follows.

Let $a_i^{(t)}$ and $b_j^{(t)}$ be the t^{th} iterated estimates of a_i and b_j .

Similarly, let $q_{ij}^{(2t-1)} = \pi_{ij} (1 + a_i^{(t)} + b_j^{(t)})$ and $q_{ij}^{(2t)} = \pi_{ij} (1 + a_i^{(t)} + b_j^{(t+1)})$

be the $(2t-1)^{\text{th}}$ and $2t^{\text{th}}$ iterated estimates of q_{ij}^* and begin the

procedure by setting $a_i^{(0)}, b_j^{(1)} = 0$ so that $q_{ij}^{(0)} = \pi_{ij}$. The

$a_i^{(t)}$ and $b_j^{(t)}$ are calculated as follows:

$$a_i^{(t)} = a_i^{(t-1)} + \frac{\pi_{i.}}{\pi_{i.}} (p_{i.} - q_{i.}^{(2t-2)}) \quad (3.7)$$

$$b_j^{(t)} = b_j^{(t-1)} + \frac{\pi_{.j}}{\pi_{.j}} (p_{.j} - q_{.j}^{(2t-3)}) \quad (3.8)$$

where $\Theta_{i.}^{(t)}$ and $\Theta_{.j}^{(t)}$ are parameters selected at the t^{th} iteration to accelerate the speed of convergence. The process is more straightforward and more easily implemented on a computer if we let $\Theta_{i.}^{(t)} = \Theta_{.j}^{(t)} = 1$ for all t ; the resulting procedure still converges quite rapidly.

Now, consider the difference between two successive estimates of q_{ij}^* , taking $\Theta_{i.}^{(t)} = \Theta_{.j}^{(t)} = 1$ for all t :

$$q_{ij}^{(2t)} - q_{ij}^{(2t-1)} = \pi_{ij} (b_j^{(t+1)} - b_j^{(t)}) \quad (3.9)$$

Substituting (3.8) gives

$$q_{ij}^{(2t)} = q_{ij}^{(2t-1)} + \frac{\pi_{ij}}{\pi_{.j}} (p_{.j} - q_{.j}^{(2t-1)}) \quad (3.10)$$

Similarly,

$$q_{ij}^{(2t+1)} = q_{ij}^{(2t)} + \frac{\pi_{ij}}{\pi_{i.}} (p_{i.} - q_{i.}^{(2t)}) \quad (3.11)$$

Note that (3.10) and (3.11) provide a direct solution by avoiding the calculation of the Lagrange multiplier estimates at each iteration.

Stephan's procedure has been expressed in different forms in the literature. For example, Friedlander [8] considered the following alternative expression to (3.4):

$$q_{ij}^* = \pi_{ij}(a_i + b_j) \quad (3.12)$$

Proceeding as before, let $a_i^{(t)}$ and $b_j^{(t)}$ be the t^{th} iterated approximations to a_i and b_j and let $q_{ij}^{(2t-1)} = \pi_{ij}(a_i^{(t)} + b_j^{(t)})$ and $q_{ij}^{(2t)} = \pi_{ij}(a_i^{(t)} + b_j^{(t+1)})$ be the $(2t-1)^{\text{th}}$ and $2t^{\text{th}}$ iterated estimates of q_{ij}^* . Calculate $a_i^{(t)}$ and $b_j^{(t)}$ as follows:

$$a_i^{(t)} = \frac{p_{i\cdot}}{\pi_{i\cdot}} - \sum_j b_j^{(t)} \frac{\pi_{ij}}{\pi_{i\cdot}} \quad (3.13)$$

$$b_j^{(t)} = \frac{p_{\cdot j}}{\pi_{\cdot j}} - \sum_i a_i^{(t-1)} \frac{\pi_{ij}}{\pi_{\cdot j}} \quad (3.14)$$

Begin the procedure by selecting

$$a_i^{(0)} = \frac{1}{2} \frac{p_{i\cdot}}{\pi_{i\cdot}} \quad (3.15)$$

which, after substitution into (3.14) gives

$$q_{ij}^{(0)} = \pi_{ij} \left(\frac{1}{2} \frac{p_{i.}}{\pi_{i.}} + \frac{p_{.j}}{\pi_{.j}} - \frac{1}{2} \sum_a \frac{p_{a.}}{\pi_{a.}} \frac{\pi_{aj}}{\pi_{.j}} \right) \quad (3.16)$$

Note that this starting value is different from that of Stephan's procedure given earlier in this chapter.

An examination of two successive estimates of $q_{ij}^{(t)}$ gives the same result as (3.9) and (3.10), i.e.,

$$\begin{aligned} q_{ij}^{(2t)} - q_{ij}^{(2t-1)} &= \pi_{ij} (b_j^{(t+1)} - b_j^{(t)}) \\ &= \frac{\pi_{ij}}{\pi_{.j}} (p_{.j} - \sum_i (a_i^{(t)} + b_j^{(t)}) \pi_{ij}) \\ &= \frac{\pi_{ij}}{\pi_{.j}} (p_{.j} - q_{.j}^{(2t-1)}) \end{aligned} \quad (3.17)$$

or

$$q_{ij}^{(2t)} = q_{ij}^{(2t-1)} + \frac{\pi_{ij}}{\pi_{.j}} (p_{.j} - q_{.j}^{(2t-1)}) \quad (3.18)$$

Similarly,

$$q_{ij}^{(2t+1)} = q_{ij}^{(2t)} + \frac{\pi_{ij}}{\pi_{i.}} (p_{i.} - q_{i.}^{(2t)}) \quad (3.19)$$

Saleh [20] followed the method used in deriving the RREP given in Chapter 2 to generate yet another iterative procedure. As before, let $a_i^{(t)}$ and $b_j^{(t)}$ be iterated estimates of the a_i and b_j of (3.4)

and let $q_{ij}^{(2t-1)} = \pi_{ij}(1 + a_i^{(t)} + b_j^{(t)})$ and $q_{ij}^{(2t)} = \pi_{ij}(1 + a_i^{(t)} + b_j^{(t+1)})$

denote iterated estimates of q_{ij}^* . Begin by setting $a_i^{(0)} = 1$ for all

i and $b_j^{(1)} = -1$ for all j , yielding $q_{ij}^{(0)} = \pi_{ij}$ for all i and j .

Succeeding $a_i^{(t)}$ and $b_j^{(t)}$ are obtained by replacing a_i and b_j

in (3.5) and (3.6) by their iterated estimates. This yields

$$a_i^{(t)} = (p_{i\cdot} - \pi_{i\cdot} - \sum_j b_j^{(t)} \pi_{ij}) / \pi_{i\cdot} \quad (3.20)$$

$$b_j^{(t)} = (p_{\cdot j} - \pi_{\cdot j} - \sum_i a_i^{(t-1)} \pi_{ij}) / \pi_{\cdot j} \quad (3.21)$$

Note that (3.20) and (3.21) are identical to (3.13) and (3.14) except

for the constant term -1 . Again, consider the difference between

the successive $q_{ij}^{(t)}$:

$$q_{ij}^{(2t)} - q_{ij}^{(2t-1)} = \pi_{ij}(b_j^{(t+1)} - b_j^{(t)})$$

Substitution of (3.21) yields as before

$$q_{ij}^{(2t)} = q_{ij}^{(2t-1)} + \frac{\pi_{ij}}{\pi_{\cdot j}} (p_{\cdot j} - q_{\cdot j}^{(2t-1)}) \quad (3.22)$$

Similarly,

$$q_{ij}^{(2t+1)} = q_{ij}^{(2t)} + \frac{\pi_{ij}}{\pi_{i.}} (p_{i.} - q_{i.}^{(2t)}) . \quad (3.23)$$

The three methods discussed above, then, calculate estimates of the q_{ij}^* at the t^{th} iteration from those at the $(t-1)^{\text{th}}$ iteration in identical ways. The differences between the methods are in their starting values: the procedures of Stephan and Saleh use the sample cell estimates π_{ij} at the zeroth iteration while Friedlander's procedure defines the zeroth iteration estimates by (3.16). Use of Stephan's and Saleh's starting values yields the estimates

$$q_{ij}^{(0)} = \pi_{ij}$$

$$q_{ij}^{(1)} = \frac{p_{i.}}{\pi_{i.}} \pi_{ij}$$

$$q_{ij}^{(2)} = \frac{p_{i.}}{\pi_{i.}} \pi_{ij} + \frac{\pi_{ij}}{\pi_{.j}} (p_{.j} - \sum_i \frac{p_{i.}}{\pi_{i.}} \pi_{ij})$$

$$q_{ij}^{(2t)} = q_{ij}^{(2t-1)} + \frac{\pi_{ij}}{\pi_{.j}} (p_{.j} - q_{.j}^{(2t-1)})$$

$$q_{ij}^{(2t+1)} = q_{ij}^{(2t)} + \frac{\pi_{ij}}{\pi_{i.}} (p_{i.} - q_{i.}^{(2t)}) . \quad (3.24)$$

This procedure, described by (3.24), will be known as the Linear Correction Estimation Procedure (LCEP) for the remainder of this thesis. Note that at both the zeroth and first iterations, the estimates $q_{ij}^{(0)}$ and $q_{ij}^{(1)}$ are the same as those for the RREP.

3.2. Properties of the LCEP.

The LCEP shares many properties with the RREP. The introduction of marginal information occurs in the same fashion as seen from (3.24), i.e., the LCEP makes use of row and column information at alternate iterations. The LCEP satisfies one set of marginals at each iteration as does the RREP. For example,

$$\begin{aligned} \sum_i q_{ij}^{(2t)} &= \sum_i q_{ij}^{(2t-1)} + \sum_i \frac{\pi_{ij}}{\pi_{.j}} (p_{.j} - q_{.j}^{(2t-1)}) \\ &= q_{.j}^{(2t-1)} + p_{.j} - q_{.j}^{(2t-1)} \\ &= p_{.j} \end{aligned} \tag{3.25}$$

Thus, the self-checking feature of the RREP is also present in the LCEP. Also, as with the RREP, the LCEP can be started by satisfying either the row or the column constraints at the first iteration.

The RREP and LCEP can be compared in other ways. The correction at each iteration of the LCEP involves the addition of the difference between the actual and estimated marginal values, whereas the RREP involves

the multiplication of the ratio of these two quantities. However, the RREP can be written in terms of the difference between the actual and estimated marginal totals as follows :

$$\begin{aligned}
 p_{ij}^{(2t)} &= \frac{p_{.j}}{p_{.j}^{(2t-1)}} p_{ij}^{(2t-1)} \\
 &= p_{ij}^{(2t-1)} + \frac{p_{ij}^{(2t-1)}}{p_{.j}^{(2t-1)}} (p_{.j} - p_{.j}^{(2t-1)}) .
 \end{aligned} \tag{3.26}$$

Thus, it is the multiplier of the difference that distinguishes the RREP from the LCEP. The RREP uses the most recent estimates for this multiplier while the LCEP uses zeroth iteration estimates at all iterations. One therefore might expect the RREP to converge more quickly than the LCEP... partially for...plies..

One characteristic of the LCEP which the RREP does not share is that the q_{ij}^* may be found exactly by solving a system of $c-1$ or $r-1$ linear equations as shown by Deming and Stephan [6]. This is obtained by making use of the fact that at most $r + c - 1$ of the $r + c$ marginal constraints are linearly independent. Therefore, ignoring say, the last column constraint and setting the corresponding Lagrange multiplier, b_c , to zero, we obtain from (3.5)

$$a_i = (p_{i.} - \sum_j b_j \pi_{ij} - \pi_{i.}) / \pi_{i.} . \tag{3.27}$$

Substitution of (3.27) into (3.6) yields

$$p_j = p_{.j} + b_j \pi_{.j} + \sum_i \pi_{ij} (p_i - \sum_j b_j \pi_{ij}) / \pi_{.i} \quad (3.28)$$

$$j = 1, \dots, c-1$$

which is clearly linear in the $c-1$ b_j 's. However, the solution of a linear system requires a substantial amount of calculation, even for moderately large tables. The LCEP clearly requires little calculation at each iteration and, as will be illustrated shortly, converges very rapidly. Hence, despite the fact that the calculation of the exact solution is possible, practical considerations favour the use of the LCEP.

Friedlander [8] considered the LCEP in the form of (3.7).

We showed that the LCEP as well as Stephan's procedure (as expressed in (3.13) - (3.15)) converge provided that (3.5) and (3.6) constitute a consistent linear system (i.e., at least one solution exists) in the a_i and b_j . A notable feature of Friedlander's proof is that it does not require $\pi_{ij} > 0$ for all i and j , unlike Kullback's proof of convergence for the RREP. Thus, the conditions for convergence of the RREP are more stringent than those for the LCEP. In fact, it is shown in Chapter 5 that the LCEP converges in all cases in which the RREP converges but the RREP does not always converge when the LCEP converges,

although these latter cases are of little practical interest.

The two 2×2 examples given in Chapter 2 will be considered in order to demonstrate the properties of the LCEP already discussed, as well as an undesirable property of the method. The marginals for these examples are $p_{1.} = .25$, $p_{2.} = .75$, $p_{.1} = .35$, $p_{.2} = .65$. Sample 1 gives $\pi_{11} = .1$, $\pi_{21} = .2$, $\pi_{12} = .2$, $\pi_{22} = .5$ whereas sample 2 has $\pi_{11} = .5$, $\pi_{21} = .2$, $\pi_{12} = .2$, $\pi_{22} = .1$.

The solution to the first example will be found by solving the linear system in the Lagrange multipliers (3.5) and (3.6). The system (3.28) with $b_2 = 0$ becomes

$$.35 = .3b_1 + (.1(.25 - .1b_1)/.3 + .2(.75 - .2b_1)/.7)$$

yielding

$$b_1 = .25$$

Substitution into (3.27) gives $a_1 = -.25$, $a_2 = 0$ so that $q_{11}^* = .1$ is the exact solution.

The solution to the second example will be found by the same method as that used in Chapter 2. Using (3.2) to express the q_{1j} in terms of q_{11} , substituting into (3.1), differentiating with respect to q_{11} and setting the result to zero gives

$$q_{11}^* = \left[\frac{p_{1.}}{\pi_{12}} + \frac{p_{.1}}{\pi_{21}} + \frac{p_{.1}}{\pi_{22}} - \frac{p_{2.}}{\pi_{22}} \right] / \left[\frac{1}{\pi_{11}} + \frac{1}{\pi_{12}} + \frac{1}{\pi_{21}} + \frac{1}{\pi_{22}} \right] \quad (3.29)$$

Note that if $p_{2.}/\pi_{22}$ is sufficiently large, q_{11}^* will be negative.

This indeed occurs in the second example where the solution is $q_{11}^* = -.0454$.

The results of applying the LCEP to these examples are as follows:

Sample 1

Iteration 0

$$p_{11}^{(0)} = .1$$

$$p_{12}^{(0)} = .2$$

$$p_{1.}^{(0)} = .3$$

$$p_{21}^{(0)} = .2$$

$$p_{22}^{(0)} = .5$$

$$p_{2.}^{(0)} = .7$$

$$p_{.1}^{(0)} = .3$$

$$p_{.2}^{(0)} = .7$$

Iteration 1

$$p_{11}^{(1)} = .8333$$

$$p_{12}^{(1)} = .1667$$

$$p_{1.}^{(1)} = .250$$

$$p_{21}^{(1)} = .2143$$

$$p_{22}^{(1)} = .5357$$

$$p_{2.}^{(1)} = .750$$

$$p_{.1}^{(1)} = .2976$$

$$p_{.2}^{(1)} = .7024$$

Iteration 2

$$p_{11}^{(2)} = .1008$$

$$p_{12}^{(2)} = .1517$$

$$p_{1.}^{(2)} = .2525$$

$$p_{21}^{(2)} = .2492$$

$$p_{22}^{(2)} = .4983$$

$$p_{2.}^{(2)} = .7475$$

$$p_{.1}^{(2)} = .35$$

$$p_{.2}^{(2)} = .65$$

Iteration 3

$$p_{11}^{(3)} = .1000$$

$$p_{12}^{(3)} = .1500$$

$$p_{1.}^{(3)} = .25$$

$$p_{21}^{(3)} = .2499$$

$$p_{22}^{(3)} = .5001$$

$$p_{2.}^{(3)} = .75$$

$$p_{.1}^{(3)} = .3499$$

$$p_{.2}^{(3)} = .6501$$

Iteration 4

$$p_{11}^{(4)} = .1000$$

$$p_{12}^{(4)} = .1500$$

$$p_{1.}^{(4)} = .25$$

$$p_{21}^{(4)} = .2500$$

$$p_{22}^{(4)} = .5000$$

$$p_{2.}^{(4)} = .75$$

$$p_{.1}^{(4)} = .35$$

$$p_{.2}^{(4)} = .65$$

Sample 2

Iteration 0

$$p_{11}^{(0)} = .5$$

$$p_{12}^{(0)} = .2$$

$$p_{1.}^{(0)} = .7$$

$$p_{21}^{(0)} = .2$$

$$p_{22}^{(0)} = .1$$

$$p_{2.}^{(0)} = .3$$

$$p_{.1}^{(0)} = .7$$

$$p_{.2}^{(0)} = .3$$

Iteration 1

$$p_{11}^{(1)} = .1885$$

$$p_{12}^{(1)} = .0715$$

$$p_{1.}^{(1)} = .25$$

$$p_{21}^{(1)} = .5$$

$$p_{22}^{(1)} = .25$$

$$p_{2.}^{(1)} = .75$$

$$p_{.1}^{(1)} = .6785$$

$$p_{.2}^{(1)} = .3215$$

Iteration 2

$$p_{11}^{(2)} = -.056143$$

$$p_{12}^{(2)} = .2905$$

$$p_{1.}^{(2)} = .2344$$

$$p_{21}^{(2)} = .40613$$

$$p_{22}^{(2)} = .3595$$

$$p_{2.}^{(2)} = .7656$$

$$p_{.1}^{(2)} = .35$$

$$p_{.2}^{(2)} = .65$$

Iteration 3

$$p_{11}^{(3)} = -.0449$$

$$p_{12}^{(3)} = .2949$$

$$p_{1.}^{(3)} = .2500$$

$$p_{21}^{(3)} = .3957$$

$$p_{22}^{(3)} = .3543$$

$$p_{2.}^{(3)} = .7500$$

$$p_{.1}^{(3)} = .3507$$

$$p_{.2}^{(3)} = .6493$$

Iteration 4

$$p_{11}^{(4)} = -.0455$$

$$p_{12}^{(4)} = .2955$$

$$p_{1.}^{(4)} = .2500$$

$$p_{21}^{(4)} = .3955$$

$$p_{22}^{(4)} = .3545$$

$$p_{2.}^{(4)} = .7500$$

$$p_{.1}^{(4)} = .3500$$

$$p_{.2}^{(4)} = .6500$$

These examples illustrate that the LCEP converges at approximately the same rate as the RREP, contrary to our predictions made from equation (3.26).

CHAPTER 4

SOME STATISTICAL PROPERTIES OF

RREP AND LCEP

4.1. Consistency and Unbiasedness of the Estimators.

Both the Raking Ratio Estimators (RRE) and the Linear Correction Estimators (LCE) are consistent (i.e., they reduce to population proportions when the sample size n equals the population size N) at each iteration. Clearly, when $n = N$, the sample cell frequencies n_{ij} equal the population cell frequencies N_{ij} and so $\hat{\pi}_{ij} = p_{ij}$ for all i and j . Thus,

$$p_{ij}^{(1)} = \frac{p_{i.}}{p_{i.}} p_{ij} = p_{ij} \quad (4.1)$$

If $p_{ij}^{(2t-1)} = p_{ij}$, then

$$p_{ij}^{(2t)} = \frac{p_{.j}}{p_{.j}^{(2t-1)}} p_{ij}^{(2t-1)} = p_{ij} \quad (4.2)$$

and so

$$p_{ij}^{(2t+1)} = p_{ij} \quad (4.3)$$

as well. Thus, by induction, $p_{ij}^{(t)}$ is consistent for all t .

Similarly, if $q_{ij}^{(2t-1)} = p_{ij}$, then

$$\begin{aligned}
 q_{ij}^{(2t)} &= q_{ij}^{(2t-1)} + \frac{p_{ij}}{p_{\cdot j}} (p_{\cdot j} - q_{\cdot j}^{(2t-1)}) \\
 &= p_{ij} + \frac{p_{ij}}{p_{\cdot j}} (p_{\cdot j} - p_{\cdot j}) \\
 &= p_{ij}
 \end{aligned} \tag{4.4}$$

and

$$q_{ij}^{(2t+1)} = p_{ij} \tag{4.5}$$

Since $q_{ij}^{(1)} = p_{ij}^{(1)}$ and $p_{ij}^{(1)}$ is consistent, it follows by induction that $q_{ij}^{(t)}$ is consistent for all t .

Both the RRE and the LCE are unbiased at the zeroth and first iterations. We have

$$p_{ij}^{(0)} = q_{ij}^{(0)} = \pi_{ij} \tag{4.6}$$

and it is well known [5] that $E(\pi_{ij}) = p_{ij}$. At the first iteration we have

$$\begin{aligned}
 p_{ij}^{(1)} &= q_{ij}^{(1)} = \frac{p_{i\cdot}}{\pi_{i\cdot}} \pi_{ij} \\
 &= \frac{N_{i\cdot}}{N} \frac{n_{ij}}{n_{i\cdot}}
 \end{aligned} \tag{4.7}$$

Denote E_2 as the expectation given $n_{i.}$ and E_1 as the expectation over $n_{i.}$. Noting that for a given $n_{i.}$ the sample from the i^{th} row is equivalent to a simple random sample of size $n_{i.}$ from the $N_{i.}$ population units in that row, we set $E_2 \frac{n_{ij}}{n_{i.}} = \frac{N_{ij}}{N_{i.}}$ and hence

$$\begin{aligned} E(p_{ij}^{(1)}) &= E(q_{ij}^{(1)}) \\ &= \frac{N_{i.}}{N} E_1 E_2 \frac{n_{ij}}{n_{i.}} = \frac{N_{i.}}{N} E_1 \frac{N_{ij}}{N_{i.}} = p_{ij} \end{aligned} \quad (4.8)$$

4.2. Variances of the Estimators.

Rao [15] derived asymptotic variance formulae for $p_{ij}^{(2)}$ and $p_{ij}^{(4)}$ by means of Taylor expansions. Consider the RRE and LCE to be functions of the sample values, i.e., let

$$p_{ij}^{(2t-1)} = f_{(2t-1)ij}(\pi) = \frac{p_{ij}}{\sum_b f_{(2t-2)ib}(\pi)} f_{(2t-2)ij}(\pi) \quad (4.9)$$

$$p_{ij}^{(2t)} = f_{(2t)ij}(\pi) = \frac{p_{i.}}{\sum_a f_{(2t-1)aj}(\pi)} f_{(2t-1)ij}(\pi)$$

$$q_{ij}^{(2t-1)} = g_{(2t-1)ij}(\pi) = g_{(2t-2)ij}(\pi) +$$

$$\frac{n_{ij}}{n_{i.}} (p_{i.} - \sum_b g_{(2t-2)ib}(\pi)) \quad (4.10)$$

$$q_{ij}^{(2t)} = g_{(2t)ij}(\underline{\pi}) = g_{(2t-1)ij}(\underline{\pi}) +$$

$$\frac{\pi_{ij}}{\pi_{ij}} (p_{ij} - \sum_a g_{(2t-2)aj}(\underline{\pi}))$$

where $\underline{\pi}$ is the vector of the π_{ij} 's. Note that by (4.2) - (4.5)

we have, for all t ,

$$f_{(t)ij}(\underline{p}) = g_{(t)ij}(\underline{p}) = p_{ij} \quad (4.11)$$

where $\underline{p}$ is the vector of p_{ij} 's.

Let

$$f_{(t)ij}^{(uv)}(\underline{\pi}) = \frac{\partial f_{(t)ij}(\underline{\pi})}{\partial \pi_{uv}}$$

$$f_{(t)ij}^{(uv)}(\underline{p}) = \frac{\partial f_{(t)ij}(\underline{p})}{\partial p_{uv}} \Big|_{\underline{\pi} = \underline{p}} \quad (4.12)$$

and similarly for $g_{(t)ij}(\underline{\pi})$. Consider first order Taylor expansions of $f_{(t)ij}$ and $g_{(t)ij}$ about $\underline{p}$. These are, using (4.11)

$$\begin{aligned} f_{(t)ij}(\underline{\pi}) &\approx f_{(t)ij}(\underline{p}) + \sum_u \sum_v (\pi_{uv} - p_{uv}) f_{(t)ij}^{(uv)}(\underline{p}) \\ &= p_{ij} + \sum_u \sum_v (\pi_{uv} - p_{uv}) f_{(t)ij}^{(uv)}(\underline{p}) \end{aligned} \quad (4.13)$$

$$\begin{aligned}
 g_{(t)ij}^{(p)} &\approx g_{(t)ij}^{(p)} + \sum_u \sum_v (\pi_{uv} - p_{uv}) g_{(t)ij}^{(uv)}(p) \\
 &= p_{ij} + \sum_u \sum_v (\pi_{uv} - p_{uv}) g_{(t)ij}^{(uv)}(p) \quad (4.14)
 \end{aligned}$$

Thus, the variance of $p_{ij}^{(t)}$ can be approximated by

$$\begin{aligned}
 V[p_{ij}^{(t)}] &\approx V[p_{ij} + \sum_u \sum_v (\pi_{uv} - p_{uv}) f_{(t)ij}^{(uv)}(p)] \\
 &= \sum_u \sum_v [f_{(t)ij}^{(uv)}(p)]^2 V(\pi_{uv}) \\
 &\quad + \sum_{(a,b)} \sum_{(u,v)} \sum_{(a,b)} \sum_{(u,v)} (f_{(t)ij}^{(ab)}(p)) (f_{(t)ij}^{(uv)}(p)) \text{cov}(\pi_{ab}, \pi_{uv}) \quad (4.15)
 \end{aligned}$$

and similarly for $q_{ij}^{(t)}$. Using the well known results [5],

$$V(\pi_{uv}) = (1/n - 1/N) p_{uv} (1 - p_{uv}) \quad (4.16)$$

and

$$\text{cov}(\pi_{ab}, \pi_{uv}) = (1/n - 1/N) p_{ab} p_{uv} \delta_{(ab)(uv)}$$

(4.15) becomes

$$\begin{aligned}
V[p_{ij}^{(t)}] &\approx (1/n - 1/N) \sum_u \sum_v [f_{(t)ij}^{(uv)}(p)]^2 p_{uv} (1-p_{uv}) \\
&- (1/n - 1/N) \sum_{(ab)} \sum_{(uv)} \sum_u \sum_v [f_{(t)ij}^{(ab)}(p)] [f_{(t)ij}^{(uv)}(p)] p_{ab} p_{uv} \\
&= (1/n - 1/N) \left\{ \sum_u \sum_v [f_{(t)ij}^{(uv)}(p)]^2 p_{uv} - \left[\sum_u \sum_v f_{(t)ij}^{(uv)}(p) p_{uv} \right]^2 \right\}
\end{aligned}
\tag{4.17}$$

and similarly for $q_{ij}^{(t)}$.

Consider now the derivatives of $f_{(2t-1)ij}^{(uv)}(\pi)$ with respect

to π_{uv} :

$$\begin{aligned}
f_{(2t-1)ij}^{(uv)}(\pi) &= p_{i.} \left\{ f_{(2t-2)ij}^{(uv)}(\pi) \sum_b f_{(2t-2)1b}(\pi) \right. \\
&\quad \left. - f_{(2t-2)1j}(\pi) \sum_b f_{(2t-2)1b}^{(uv)}(\pi) \right\} / \left[\sum_b f_{(2t-1)1b}(\pi) \right]^2
\end{aligned}$$

By (4.11) we get

$$f_{(2t-1)ij}^{(uv)}(p) = \frac{p_{i.} f_{(2t-2)ij}^{(uv)}(p) - p_{ij} \sum_b f_{(2t-2)1b}^{(uv)}(p)}{p_{i.}} \tag{4.18}$$

Similarly,

$$f_{(2t)ij}^{(uv)}(\underline{p}) = \frac{p_{\cdot j} f_{(2t-1)ij}^{(uv)}(\underline{p}) - p_{ij} \sum_a f_{(2t-1)ajj}^{(uv)}(\underline{p})}{p_{\cdot j}} \quad (4.19)$$

Turning to the derivative of $g_{(2t-1)ij}(\pi)$ with respect to π_{uv} , we see that

$$\begin{aligned} g_{(2t-1)ij}^{(uv)}(\pi) &= g_{(2t-2)ij}^{(uv)}(\pi) \\ &+ (p_{i\cdot} - \sum_b g_{(2t-2)ib}(\pi)) ((\pi_{ij}^{(uv)} \pi_{i\cdot} - \pi_{ij} \pi_{i\cdot}^{(uv)}) / (\pi_{i\cdot})^2) \\ &- \pi_{ij} (\sum_b g_{(2t-2)ib}^{(uv)}(\pi)) / \pi_{i\cdot} \end{aligned}$$

where

$$\pi_{ij}^{(uv)} = \frac{\partial \pi_{ij}}{\partial \pi_{uv}} = \begin{cases} 1 & \text{if } (i,j) = (u,v) \\ 0 & \text{otherwise} \end{cases}$$

and

$$\pi_{i\cdot}^{(uv)} = \frac{\partial \pi_{i\cdot}}{\partial \pi_{uv}} = \begin{cases} 1 & \text{if } i = u \\ 0 & \text{otherwise} \end{cases}$$

Again, by (4.11) we get

$$\begin{aligned}
g_{(2t-1)ij}^{(uv)}(\underline{p}) &= g_{(2t-2)ij}^{(uv)}(\underline{p}) \\
&\quad + (p_{i.} - p_{i.}) (p_{ij}^{(uv)} p_{i.} - p_{ij} p_{i.}^{(uv)}) / (p_{i.})^2 \\
&\quad - p_{ij}^{(rs)} \left(\sum_b g_{(2t-2)ib}^{(uv)}(\underline{p}) \right) / p_{i.} \\
&= \frac{p_{i.} g_{(2t-2)ij}^{(uv)}(\underline{p}) - p_{ij} \left(\sum_b g_{(2t-2)ib}^{(uv)}(\underline{p}) \right)}{p_{i.}} \quad (4.19)
\end{aligned}$$

where

$$p_{ij}^{(uv)} = \pi_{ij}^{(uv)} \mid \underline{\pi} = \underline{p} = \begin{cases} 1 & \text{if } (i,j) = (u,v) \\ 0 & \text{otherwise} \end{cases}$$

and

$$p_{i.}^{(rs)} = \pi_{i.}^{(rs)} \mid \underline{\pi} = \underline{p} = \begin{cases} 1 & \text{if } i = u \\ 0 & \text{otherwise} \end{cases}$$

Similarly,

$$g_{(2t)ij}^{(uv)}(\underline{p}) = \frac{p_{.j} g_{(2t-1)ij}^{(uv)}(\underline{p}) - p_{ij} \left[\sum_a g_{(2t-1)aj}^{(uv)}(\underline{p}) \right]}{p_{.j}} \quad (4.20)$$

Since $g_{(1)ij}(\underline{\pi}) = f_{(1)ij}(\underline{\pi})$, we have by induction and the previous four equations that

$$f_{(t)ij}^{(uv)}(p) = g_{(t)ij}^{(uv)}(p) \quad (4.21)$$

for all t so that

$$p_{ij}^{(t)} = f_{(t)ij}(\pi) \simeq g_{(t)ij}(\pi) = q_{ij}^{(t)}$$

for all t . Moreover, by (4.21) the variances of the RRE and the LCE are equal at all iterations to this degree of approximation.

Properties of the variance for the first four iterations of the RREP were investigated by Brackstone [3]. Since the values to which the estimators converge are Best Asymptotically Normal (BAN), one would expect the variance to decrease at each iteration. While this is usually the case, it will be shown that there are exceptions.

The expression for $V[p_{ij}^{(t)}]$ given by (4.17) involves a sum of squares and the square of a sum. The square of the sum is zero for all iterations except the zeroth. The proof is as follows.

$$\text{Since } p_{ij}^{(0)} = \pi_{ij},$$

$$f_{(0)ij}^{(uv)}(\pi) = \alpha_{iu} \alpha_{jv} = f_{(0)ij}^{(uv)}(p) \quad (4.22)$$

where $\alpha_{iu} = \begin{cases} 1 & \text{if } i = u \\ 0 & \text{otherwise} \end{cases}$

Clearly,

$$\sum_u \sum_v p_{uv} f^{(uv)}_{(0)ij}(\underline{p}) = p_{ij} \quad (4.23)$$

Thus, by the recursion relations (4.18) and (4.19) we get

$$\begin{aligned} \sum_u \sum_v p_{uv} f^{(uv)}_{(1)ij}(\underline{p}) &= \sum_u \sum_v \left[f^{(uv)}_{(0)ij}(\underline{p}) - \frac{p_{ij}}{p_{1i}} \sum_{ab} f^{(uv)}_{(0)ab}(\underline{p}) \right] p_{uv} \\ &= p_{ij} - \frac{p_{ij}}{p_{1i}} \sum_v p_{1v} \\ &= 0 \end{aligned} \quad (4.24)$$

Assume that

$$\sum_u \sum_v p_{uv} f^{(uv)}_{(2t-2)ij}(\underline{p}) = 0$$

Then by (4.18)

$$\sum_u \sum_v p_{uv} f^{(uv)}_{(2t-1)ij}(\underline{p}) = 0 \quad (4.25)$$

Similarly,

$$\sum_u \sum_v p_{uv} f_{(2h)ij}^{(uv)}(\underline{p}) = 0 \quad (4.26)$$

Thus, the square of the sum in (4.17) is zero for all $h \geq 1$. Therefore,

(4.17) becomes

$$V(p_{ij}^{(h)}) \simeq (1/n - 1/N) \left[\sum_u \sum_v [f_{(h)ij}^{(uv)}(\underline{p})]^2 p_{uv} \right] \quad (4.27)$$

for $h \geq 1$.

For $h = 0$ we have

$$V(p_{ij}^{(0)}) \simeq (1/n - 1/N) \left[\sum_u \sum_v [f_{(0)ij}^{(uv)}(\underline{p})]^2 p_{uv} - p_{ij}^2 \right] \quad (4.28)$$

We have already seen from (4.22) that

$$f_{(0)ij}^{(uv)}(\underline{p}) = \alpha_{iu} \alpha_{jv}$$

Using the recursion relations (4.18) and (4.19) we get

$$f_{(1)ij}^{(uv)}(\underline{p}) = \alpha_{iu} \left(\alpha_{jv} - \frac{p_{uj}}{p_u} \right) \quad (4.29)$$

and

$$f_{(2)ij}^{(uv)}(p) = (\alpha_{in} - \frac{p_{ij}}{p_{.j}})(\alpha_{jv} - \frac{p_{ij}}{p_{.i}}) \quad (4.30)$$

Substituting these into (4.27) yields

$$v(p_{ij}^{(0)}) \approx (1/n - 1/N) p_{ij} (1 - p_{ij}) \quad (4.31)$$

$$v(p_{ij}^{(1)}) \approx (1/n - 1/N) p_{ij} (1 - \frac{p_{ij}}{p_{.i}}) \quad (4.32)$$

and, after some algebra,

$$v(p_{ij}^{(2)}) \approx (1/n - 1/N) p_{ij} \left[1 - \frac{p_{ij}}{p_{.j}} - \frac{p_{ij}}{p_{.i}} + \frac{2p_{ij}^2}{p_{.i} p_{.j}} - \frac{p_{ij}^2 p_{.j}^2}{p_{.i}^2 p_{.j}^2 p_{.w}} \right] \quad (4.33)$$

Thus, when comparing variances for the first two iterations, we get

$$\begin{aligned} v(p_{ij}^{(0)}) - v(p_{ij}^{(1)}) \\ \approx (1/n - 1/N) p_{ij} (\frac{1}{p_{.i}} - 1) > 0, \quad (4.34) \end{aligned}$$

$$V(p_{1j}^{(0)}) - V(p_{1j}^{(2)})$$

$$\approx (1/n - 1/N) \left[p_{1j}^2 \left(\frac{1}{p_{.j}} - 1 \right) + \frac{p_{1j}^2}{p_{1.}} \left(1 - \frac{p_{1j}}{p_{.j}} \right)^2 + \frac{p_{1j}^2}{p_{.j}^2} \sum_a \frac{p_{aj}^2}{p_{a.}} \right]$$

> 0

(4.35)

and finally

$$V(p_{1j}^{(1)}) - V(p_{1j}^{(2)}) \approx (1/n - 1/N) \left[\frac{p_{1j}^2}{p_{.j}} \left(1 - 2 \frac{p_{1j}}{p_{1.}} \right) + \frac{p_{1j}^2}{p_{.j}^2} \sum_a \frac{p_{aj}^2}{p_{a.}} \right]$$

$$> 0 \text{ if } \frac{p_{1j}}{p_{1.}} \leq \frac{1}{2}$$

(4.36)

Note that the condition $\frac{p_{1j}}{p_{1.}} \leq \frac{1}{2}$ is only a sufficient condition.

However, as demonstrated in the following example, it is possible that

$V(p_{1j}^{(2)})$ could exceed $V(p_{1j}^{(1)})$. Suppose $p_{11} = .3$, $p_{12} = .05$,

$p_{21} = .1$, $p_{22} = .55$. Then

$$V(p_{11}^{(1)}) - V(p_{11}^{(2)}) \approx (1/n - 1/N) (-.0074).$$

Brackstone's results [3] for iterations 3 and 4 are summarized

below:

$$1) \quad V(p_{ij}^{(t)}) < V(p_{ij}^{(0)})$$

i) for $t = 1, 2$

ii) for $t = 3$ if

$$\frac{p_{ij}}{p_{i.}} \leq \frac{1}{2} \quad \text{or} \quad \frac{p_{ij}}{p_{.j}} \leq \frac{1}{2}$$

iii) for $t = 4$ if $\frac{p_{ij}}{p_{.j}} \leq \frac{1}{2}$

$$2) \quad V(p_{ij}^{(t)}) < V(p_{ij}^{(1)}) \quad \text{for } t = 3 \quad \text{if} \quad \frac{p_{ij}}{p_{i.}} \leq \frac{1}{2}$$

Note that in case 1) the sufficient conditions for the inequality curiously become more stringent as t increases.

4.3. An Empirical Comparison of the RREP and LCEP.

The RREP is used at Statistics Canada in analyzing Census data, although the procedure is not used to find estimates of p_{ij} . To appreciate how the procedure is used, one requires an understanding of the sampling techniques adopted in the Canadian Census.

For the 1974 test Census, two questionnaires were used.

A short questionnaire (form 2A) requested information on age, sex, marital status, mother tongue, tenure and dwelling type. A longer questionnaire (form 2B) requested not only the information of the 2A form

but also information on education, labour force status and migration. A one in three systematic sample of households received the 2B form while the rest received the 2A form.

A two-way table was constructed from 2A data pertaining to household characteristics (which included characteristics of the head of the household) for each suitably sized sub-population known as a weighting area. Thus, the population cell counts, N_{ij} , and the marginal population totals, $N_{i.} = \sum_j N_{ij}$, $N_{.j} = \sum_i N_{ij}$ are known for all i and j . However, the population cell counts for a 2B item, denoted by Y_{ij} , are unknown although sample cell counts, y_{ij} , are available from the 2B tabulations. The RREP is used on the table described above to calculate weights for each cell. These weights are then applied to the y_{ij} and these estimates are summed over the rows and columns to obtain an estimate of the population total, $Y_R^{(t)} = \sum_i \sum_j w_{ij}^{(t)} y_{ij}$ for that 2B characteristic. The weight at the t^{th} iteration, $w_{ij}^{(t)}$, is calculated as follows:

$$w_{ij}^{(0)} = \frac{N}{n}$$

$$w_{ij}^{(1)} = \frac{N_{i.}}{n_{i.}}$$

$$w_{ij}^{(2)} = \frac{N_{.j}}{\sum_a w_{aj}^{(1)} n_{aj}} w_{ij}^{(1)}$$

$$w_{ij}^{(2t-1)} = \frac{N_{i\cdot}}{\sum_b w_{ib}^{(2t-2)} n_{ib}} w_{ij}^{(2t-2)}$$

$$w_{ij}^{(2t)} = \frac{N_{\cdot j}}{\sum_a w_{aj}^{(2t-1)} n_{aj}} w_{ij}^{(2t-1)} \quad (4.37)$$

Note that (4.37) can be derived from (2.11) and (2.12) by the relation

$$N_{ij}^{(t)} = w_{ij}^{(t)} n_{ij}$$

Estimates of the bias and variance of $Y_R^{(t)}$ for $t = 0, \dots, 4$ were derived by Brackstone and Arora [2]. These estimates were programmed on Statistics Canada's IBM 370 and applied to data of 15 weighting areas of the 1974 Census for several 2B items. A comprehensive account of the results of the study can be found in Bankier [1].

The bias and variance formulae mentioned above make extensive use of the Taylor expansion technique similar to that used earlier in this chapter. Since we have seen that the LCE and RRE are identical to the first order approximations these formulae are also valid for the LCEP estimates of 2B items, i.e. $Y_L^{(t)} = \sum_i \sum_j u_{ij}^{(t)} y_{ij}$. The $u_{ij}^{(t)}$ are the with iterated weights calculated by the LCEP as follows:

$$u_{1j}^{(0)} = \frac{N}{n} = w_{1j}^{(0)}$$

$$u_{1j}^{(1)} = \frac{N_{1.}}{n_{1.}} = w_{1j}^{(1)}$$

$$u_{1j}^{(2)} = \frac{N_{1.}}{n_{1.}} + \frac{1}{n_{.j}} \left(N_{.j} - \sum_i \frac{N_{i.}}{n_{i.}} n_{ij} \right)$$

$$u_{1j}^{(2t-1)} = u_{1j}^{(2t-2)} + \frac{1}{n_{1.}} \left(N_{1.} - \sum_j u_{1j}^{(2t-2)} n_{1j} \right)$$

$$u_{1j}^{(2t)} = u_{1j}^{(2t-1)} + \frac{1}{n_{.j}} \left(N_{.j} - \sum_i u_{ij}^{(2t-1)} n_{ij} \right)$$

(4.38)

The bias and variance program for 2B forms was run for seven weighting areas using the LCEP weight estimates in place of those of the RREP.

The resulting figures for the population estimates agreed with those of the RREP almost exactly, with the worst agreement being two significant digits.

CHAPTER 5

CONVERGENCE PROPERTIES OF RREP

AND LCEP WHEN SOME SAMPLE

CELL FREQUENCIES ARE ZERO

5.1. Restrictions on the Marginal Constraints.

It was mentioned in Chapters 2 and 3 that proofs exist which guarantee the convergence of the RREP and LCEP if certain conditions are fulfilled. Friedlander's proof of convergence of the LCEP [8] requires that the system of marginal constraints expressed in terms of the Lagrangian multipliers, (3.5) and (3.6), be consistent. Kullback's proof of convergence for the RREP requires $\pi_{ij} > 0$ for all i and j . However, empirical results from the Statistics Canada study [1] indicate that the RREP can converge even when some $\pi_{ij} = 0$ and that convergence is still quite rapid.

An important characteristic of both the RREP and LCEP is that they both produce estimates of zero for cells in which $\pi_{ij} = 0$. This mapping of zero cells to zero cells can easily be seen for the RREP as follows. We have,

$$p_{ij}^{(0)} = \pi_{ij}$$

$$p_{ij}^{(2t-1)} = \frac{p_{i \cdot}}{p_{i \cdot}^{(2t-2)}} p_{ij}^{(2t-2)}$$

$$p_{ij}^{(2t)} = \frac{p_{\cdot j}}{p_{\cdot j}^{(2t-1)}} p_{ij}^{(2t-1)} \quad (5.1)$$

so that it easily follows by induction that $p_{ij}^{(t)} = 0$ for all t if

$\pi_{ij} = 0$. The LCEP can be stated as

$$q_{ij}^{(0)} = \pi_{ij}$$

$$q_{ij}^{(2t-1)} = q_{ij}^{(2t-2)} + \frac{\pi_{ij}}{\pi_{i \cdot}} (p_{i \cdot} - q_{i \cdot}^{(2t-2)})$$

$$q_{ij}^{(2t)} = q_{ij}^{(2t-1)} + \frac{\pi_{ij}}{\pi_{\cdot j}} (p_{\cdot j} - q_{\cdot j}^{(2t-1)}) \quad (5.2)$$

so that we have $q_{ij}^{(t)} = 0$ for all t if $\pi_{ij} = 0$.

These estimates, $p_{ij}^{(t)} = q_{ij}^{(t)} = 0$, of the solutions p_{ij}^* and q_{ij}^* for cells in which $\pi_{ij} = 0$ are reasonable since, by (2.4) and (3.4)

$$p_{ij}^* = a_i b_j \pi_{ij} \quad (5.3)$$

$$q_{ij}^* = (1 + c_i + d_j) \pi_{ij} \quad (5.4)$$

so that if there is a finite solution to the a_i, b_j, c_i, d_j , then $p_{ij}^* = q_{ij}^* = 0$. Note also that the solutions $p_{ij}^* = q_{ij}^* = 0$ are the only solutions which avoid making the functions to be minimized, namely (2.1) and (3.1), infinite. Thus, the fact that $\pi_{ij} = 0$ for some cells places additional constraints on the linear system of marginal equations, i.e., $q_{ij} = 0$ for all (i,j) in which $\pi_{ij} = 0$. This augmented system of constraints may be written

$$\begin{aligned} \sum_j q_{ij} &= p_i. & i &= 1, \dots, r \\ \sum_i q_{ij} &= p_j. & j &= 1, \dots, c \end{aligned} \quad (5.5)$$

$$q_{ij} = 0 \text{ for all } (i,j) \text{ where } \pi_{ij} = 0.$$

Under the constraints (5.5), three situations of interest can occur: 1) The system (5.5) is inconsistent. 2) The system is consistent but $q_{ij} < 0$ for at least one (i,j) in all solutions. 3) The system is consistent and there is at least one solution for which $q_{ij} \geq 0$ for all (i,j) .

5.2. Marginal Constraints which are Inconsistent.

An example of the first situation is the following 3×3 table where the zeros denote cells for which $\pi_{ij} = 0$:

	0		
0		0	$p_{2.}$
	0		

$p_{.2}$

Since the RREP and LCEP satisfy the row and column constraints at successive iterations, we get for this example

$$\begin{aligned}
 p_{22}^{(2t-1)} &= q_{22}^{(2t-1)} = p_{2.} \\
 p_{22}^{(2t)} &= q_{22}^{(2t)} = p_{.2}
 \end{aligned}
 \tag{5.6}$$

Thus, if $p_{.2} \neq p_{2.}$ then the estimates of p_{22} for both procedures oscillate between $p_{2.}$ and $p_{.2}$ and hence do not converge. This is easily explained if we examine the augmented system of constraints (5.5) for this situation where we have

$$\begin{aligned}
 q_{22} &= p_{2.} \\
 q_{22} &= p_{.2}
 \end{aligned}
 \tag{5.7}$$

which is clearly inconsistent if $p_{2.} \neq p_{.2}$.

It is of practical importance to be able to easily recognize whether the system of marginal constraints (5.5) is inconsistent since

neither the RREP nor the LCEP will converge if this is the case. To this end, the concept of a separable table plays a major role. A table is separable if its rows and columns can be arranged to produce a table of the following block diagonal form:

$$\begin{array}{ccccccc}
 B_1 & 0 & 0 & . & . & . & 0 \\
 0 & B_2 & 0 & . & . & . & 0 \\
 0 & 0 & B_3 & & & & . \\
 . & . & & . & & & . \\
 . & . & & & . & & . \\
 . & . & & & & . & 0 \\
 0 & 0 & . & . & . & 0 & B_k
 \end{array}$$

where the B_i are sub-tables in which the π_{ij} may be zero or nonzero and the zeros denote sub-tables in which all the π_{ij} are zero.

The importance of 'separability' is apparent in the following proposition: The system of constraints (5.5) is inconsistent if and only if the table is separable unless the sum of the row population proportions equals the sum of the column population proportions for all sub-tables. It is assumed that $\pi_{i1} > 0$, $i = 1, \dots, r$, and $\pi_{j1} > 0$, $j = 1, \dots, c$.

The proof of the above proposition is as follows. Suppose

a solution to (5.5), denoted by d_{ij} , exists for some separable table.

Then, denoting the summation over the rows and the summation over the columns of the sub-table B_t as $\sum_i^r$ and $\sum_j^c$ respectively, we have

$$\begin{aligned} \sum_i^r d_{ij} &= p_{.j} && \text{for all } j \text{ in } B_t \\ \sum_j^c d_{ij} &= p_{i.} && \text{for all } i \text{ in } B_t \end{aligned} \quad (5.7)$$

and so,

$$\sum_i^r p_{i.} = \sum_i^r \sum_j^c d_{ij} = \sum_j^c \sum_i^r d_{ij} = \sum_j^c p_{.j} \quad (5.8)$$

This result is more easily appreciated when expressed in the converse; i.e., if the table is separable and if for any sub-table the sum of the row population proportions does not equal the sum of the column population proportions, the system (5.5) is inconsistent. This proves the sufficiency. A proof of the necessity is considerably longer and is given in Appendix A.

This result is particularly helpful in determining whether the LCEP will converge in a given situation. Friedlander's proof of convergence of the LCEP [8] requires that the marginal constraints (3.5) and (3.6) be consistent in the a_i and b_j . It is shown in Appendix

B that (3.5) and (3.6) are a consistent system in the $r + c$ a_i 's and b_j 's if and only if (5.5) is consistent in the rc q_{ij} 's. This result, together with the previous proposition, shows that the LCEP will not converge if and only if the table is separable unless the sum of the row population proportions equals the sum of the column population proportions for all sub-tables.

The behaviour of the RREP and LCEP in a situation in which (5.5) is an inconsistent system is illustrated in an example given in Appendix C. The results include not only the cell estimates $p_{ij}^{(h)}$ and $q_{ij}^{(h)}$ and their respective marginal estimates but also a measure of the differences between the true and estimated marginal values.

These statistics, $(\sum_i (p_{i.}^{(2h)} - p_{i.})^2 / p_{i.}^2)^{\frac{1}{2}}$, $(\sum_j (p_{.j}^{(2h-1)} - p_{.j})^2 / p_{.j}^2)^{\frac{1}{2}}$, $(\sum_i (q_{i.}^{(2h)} - p_{i.})^2 / p_{i.}^2)^{\frac{1}{2}}$ and $(\sum_j (q_{.j}^{(2h-1)} - p_{.j})^2 / p_{.j}^2)^{\frac{1}{2}}$ are estimates

of how close the RREP and LCEP are to complete convergence since if these statistics are zero at a particular iteration, convergence has occurred. In the example given in Appendix C, these convergence measures decrease rapidly through the first and second iterations and then stabilize. Thus, it appears that in this example the procedures are producing estimates which are as close as possible in some sense to a solution which satisfies the marginals. Notice also that in this example the cell estimates oscillate in the higher iterations.

Clearly, it is advisable to set up contingency tables and their corresponding collapsing strategies to avoid the problem of inconsistency.

5.3. Marginal Constraints with Infeasible Solutions.

An example of the second situation under consideration, in which (5.5) is consistent but at least one $q_{ij} < 0$ for all possible solutions, is the following 2×2 case. Suppose $p_1 = .25$, $p_2 = .75$, $p_{.1} = .35$, $p_{.2} = .65$ with $\pi_{22} = 0$ and the other $\pi_{ij} > 0$. The system (5.5) reduces to a system of three linearly independent equations in three unknowns, the solution to which is $q_{11}^* = -.4$. The results of applying the RREP and LCEP to this problem, in which $\pi_{11} = .5$, $\pi_{12} = .3$, $\pi_{21} = .2$ are as follows:

RREP		LCEP	
Iteration 1			
.1786	.0714	.1786	.0714
.7500	.0000	.7500	.0000
Iteration 2			
.0673	.6500	-.1830	.6500
.2827	.0000	.5300	.0000
Iteration 3			
.0235	.2265	-.3380	.5880
.7500	.0000	.7500	.0000

RREP

LCEP

Iteration 4

.0106	.6500	-.3768	.6500
.3394	.0000	.7268	.0000

Note that the LCEP converges towards the solution while the RREP does not. This is to be expected since the RREP cannot converge to an estimate which is less than zero. In this example, the RREP does its best, i.e., $p_{11}(t) \rightarrow 0$ as t increases and zero is the closest non-negative estimate of the solution $q_{11}^* = -.04$. The other $p_{ij}(t)$ behave as they would in the inconsistent system with $q_{11} = q_{22} = 0$, i.e., they oscillate between the true row and column marginal values.

This situation may occur when the solution to the constraints (5.5) is not unique. The following example illustrates this property, as well as the fact that only one π_{ij} need be zero to give rise to this case and thus prevent convergence of the RREP. Suppose $\pi_{11} = 0$ so that $q_{11} = 0$ and suppose all other $\pi_{ij} > 0$. If $p_{1.}$ and $p_{.1}$ are 'large' then, since

$$\sum_{b=2}^B q_{1b} = p_{1.} \quad (5.9)$$

at least one of the q_{1b} is 'large'. However, we must also have

$$\sum_a q_{aj} = p_{.j} \quad (5.10)$$

so that if all $q_{1j} > 0$ we must have

$$q_{1j} < p_{.j} \quad j = 2, \dots, c$$

i.e., an upper bound for the q_{1j} is the population marginal proportions.

Now, if the $p_{.j}$, $j = 1, \dots, c$ are 'small' so that

$$\sum_{b=2}^c p_{.b} < p_{.1} \quad (5.11)$$

then there must be some $q_{1j} < 0$ for any solution since otherwise all the q_{1j} will be 'small' and lead to a violation of (5.9).

The results of applying the RREP and LCEP to a hypothetical example of the above type are given in Appendix D. Again, the LCEP converges whereas the RREP does not due to the fact that at least one $q_{1j} < 0$ in any solution. Note that for cells in which $q_{1j}^{(t)} < 0$, $t > 1$, $p_{1j}^{(t)} \rightarrow 0$ as t increases, indicating that, as in the previous 2×2 example, the RREP is converging as best it can to a negative estimate. Notice also that at higher iterations the RREP behaves as it would if (5.5) were the inconsistent system in which $\pi_{1j} = 0$ for those cells in which $p_{1j}^{(t)} \rightarrow 0$, i.e., the nonzero cell estimates oscillate.

5.4. Marginal Constraints with a Feasible Solution.

The third situation under consideration is that in which (5.5) is a consistent system with at least one solution where q_{1j} is non-negative for all i and j . Little will be said about this case here since, as shown in the Statistics Canada study [1], the RREP is likely

to converge in this situation; indeed, there was no case of an apparent lack of convergence of the RREP in this study. In the light of this evidence, it is conjectured that if there is at least one solution to (5.5) for which all q_{ij} are non-negative, then the RREP will converge.

There is a significant implication to this conjecture. Ireland [12] briefly discusses possible methods of detecting a lack of convergence in the RREP. These methods involve making checks on the iterated values as the procedure progresses. Unfortunately, it is by no means certain that these checks detect all possible cases of a lack of convergence. Also, there may be cases where several iterations are required before these methods ascertain whether the RREP is converging or not.

The conjecture cited above holds that the RREP will converge if the following region is not empty:

$$\begin{aligned} \sum_j q_{ij} &= p_i & i &= 1, \dots, r \\ \sum_i q_{ij} &= p_j & j &= 1, \dots, c \\ q_{ij} &= 0 & \text{if } \pi_{ij} &= 0 \\ q_{ij} &\geq 0 & \text{for all } i \text{ and } j. \end{aligned} \quad (5.12)$$

Note that (5.12) may be seen as the feasibility region to a linear

programming problem and so the Artificial Basis Technique [9] may be used to determine whether this region is empty or not. The advantages of using this technique, as opposed to those of Ireland, to ascertain convergence of the RREP are twofold: 1) This technique virtually guarantees whether the RREP will converge or not and 2) because it does not depend on the behaviour of the RREP in a given problem, it avoids the difficulty of possibly requiring several iterations before detecting a lack of convergence.

CHAPTER 6

TWO OTHER ESTIMATORS OF THE POPULATION CELL FREQUENCIES

6.1. The Estimators and Some of Their Properties.

As mentioned in the introduction, the system of marginal constraints consists of rc unknowns and $r + c - 1$ linearly independent equations so that the solution is not unique. Indeed, in the previous chapters we discussed two estimation procedures, the RREP and LCEP, which are similar in many respects but which yield different estimates. In this chapter, two estimators of the population cell proportions p_{ij} which satisfy the marginals exactly are discussed. These estimators are non-iterative, very simple to calculate, and because they satisfy the marginals exactly have a self-checking property. However, unlike the RRE and LCE they are not derived by minimizing a function with known statistical properties.

Smith [18] proposed the following estimator of the p_{ij} :

$$u_{ij} = \pi_{ij} + p_{.j}(p_{i.} - \pi_{i.}) + p_{i.}(p_{.j} - \pi_{.j}) \quad (6.1)$$

This estimation technique is known as the Method of Proportional Distribution of Marginal Adjustments. Essentially, this estimator involves two linear corrections to π_{ij} , one for the row constraints and one for the column constraints. It is easy to check that u_{ij} satisfies the marginals exactly.

Another non-iterative estimator was proposed by El-Badry and Stephan [7]. They examined the estimates

$$v_{ij} = \pi_{ij} \frac{p_{1.}}{\pi_{1.}} - p_{1.} \left[\sum_a \frac{p_{a.}}{\pi_{a.}} \pi_{aj} - p_{.j} \right] \quad (6.2)$$

and alternatively

$$v_{ij}^* = \pi_{ij} \frac{p_{.j}}{\pi_{.j}} - p_{.j} \left[\sum_b \frac{p_{.b}}{\pi_{.b}} \pi_{ib} - p_{1.} \right]$$

This estimation technique is known as the Method of Array Proportions.

Note that both v_{ij} and v_{ij}^* are combinations of ratio estimates and a linear correction term. It is easy to see that v_{ij} and v_{ij}^* satisfy the marginals exactly. For instance,

$$\begin{aligned} \sum_i v_{ij} &= \sum_i \pi_{ij} \frac{p_{1.}}{\pi_{1.}} - \sum_i p_{1.} \left[\sum_a \frac{p_{a.}}{\pi_{a.}} \pi_{aj} - p_{.j} \right] \\ &= \sum_i \pi_{ij} \frac{p_{1.}}{\pi_{1.}} - \sum_a \frac{p_{a.}}{\pi_{a.}} \pi_{aj} + p_{.j} \\ &= p_{.j} \end{aligned}$$

and

$$\sum_j v_{ij} = \frac{p_{i.}}{\pi_{i.}} \sum_j \pi_{ij} - p_{i.} \sum_a \frac{p_{a.}}{\pi_{a.}} \sum_j \pi_{aj} + p_{i.} \sum_j p_{ij}$$

$$= \frac{p_{i.}}{\pi_{i.}} \pi_{i.} - p_{i.} \sum_a p_{a.} + p_{i.}$$

$$= p_{i.}$$

Both u_{ij} and v_{ij} are consistent estimators, i.e., they reduce to the population cell proportions when $\pi_{ij} = p_{ij}$.

This is obvious for u_{ij} from (6.1). For v_{ij} , consider (6.2) with

$\pi_{ij} = p_{ij}$, i.e.,

$$v_{ij} = p_{ij} \frac{p_{i.}}{p_{i.}} - p_{i.} \left[\sum_a \frac{p_{a.}}{p_{a.}} p_{aj} - p_{.j} \right]$$

$$= p_{ij} - p_{i.} [p_{.j} - p_{.j}]$$

$$= p_{ij}$$

(6.3)

As well as being consistent, both u_{ij} and v_{ij} are unbiased, unlike the RRE and LCE for iterations greater than one.

For the estimator u_{ij} we get

$$\begin{aligned}
 E(u_{ij}) &= E(\pi_{ij}) + p_{.j}p_{i.} - p_{.j} E(\pi_{i.}) + p_{i.}p_{.j} - p_{i.}E(\pi_{.j}) \\
 &= p_{ij} \quad (6.4)
 \end{aligned}$$

As in Chapter 4, denote E as the expectation given $\pi_{i.}$ and E_1 as the expectation over $\pi_{i.}$ and consider $E(v_{ij})$:

$$\begin{aligned}
 E(v_{ij}) &= E_1 E_2 (v_{ij}) \\
 &= E_1 E_2 \left(\pi_{ij} \frac{p_{i.}}{p_{i.}} \right) - p_{i.} E_1 E_2 \left(\sum_a \frac{p_{a.}}{p_{a.}} \pi_{aj} \right) + p_{i.}p_{.j} \\
 &= E_1 \left(p_{ij} \frac{p_{i.}}{p_{i.}} \right) - p_{i.} E_1 \left(\sum_a \frac{p_{a.}}{p_{a.}} p_{aj} \right) + p_{i.}p_{.j} \\
 &= p_{ij} \quad (6.5)
 \end{aligned}$$

6.2. Variances of the Estimators.

The variance of u_{ij} can be derived using variances and covariances of the estimator π_{ij} . By a well known result [5] we have

$$V(\pi_{ij}) = \left(\frac{1}{n} - \frac{1}{N} \right) p_{ij} (1 - p_{ij}) \quad (6.6)$$

and

$$\text{cov}(\pi_{ij}, \pi_{uv}) = - \left(\frac{1}{n} - \frac{1}{N} \right) p_{ij} p_{uv} \quad (i, j) \neq (u, v)$$

From elementary properties of covariances it is easily seen that

$$\text{cov}(\pi_{ij}, \pi_{i.}) = - \left(\frac{1}{n} - \frac{1}{N} \right) p_{ij} (1 - p_{i.})$$

$$\text{cov}(\pi_{ij}, \pi_{.j}) = - \left(\frac{1}{n} - \frac{1}{N} \right) p_{ij} (1 - p_{.j})$$

$$\text{cov}(\pi_{i.}, \pi_{.j}) = - \left(\frac{1}{n} - \frac{1}{N} \right) [p_{i.} p_{.j} - p_{ij}] \quad (6.7)$$

and so, following El-Badry and Stephan [7] we get

$$\begin{aligned} v(u_{ij}) = & \left(\frac{1}{n} - \frac{1}{N} \right) [v(\pi_{ij}) + p_{.j}^2 v(\pi_{i.}) + p_{i.}^2 v(\pi_{.j}) \\ & - 2 p_{.j} \text{cov}(\pi_{ij}, \pi_{i.}) \\ & - 2 p_{i.} \text{cov}(\pi_{ij}, \pi_{.j}) \\ & + 2 p_{i.} p_{.j} \text{cov}(\pi_{i.}, \pi_{.j})] \end{aligned}$$

$$\begin{aligned}
v_{ij} = & \left(\frac{1}{n} - \frac{1}{N} \right) [p_{ij}(1 - p_{ij}) + 2p_{ij}(3p_{i.}p_{.j} - p_{i.} - p_{.j}) \\
& + p_{i.}p_{.j}(p_{i.} + p_{.j} - 4p_{i.}p_{.j})] .
\end{aligned}
\tag{6.8}$$

El-Badry and Stephan estimate $v(v_{ij})$ in a complicated fashion requiring the calculation of expectations using summations of probabilities. The technique developed by Rao [15] used in Chapter 4 to estimate the variances of the RRE and LCE yields the same result more quickly. Consider v_{ij} as a function of the π_{ij} and employ a first order Taylor's expansion about the population cell proportions p_{ij} . Using the notation of Chapter 4, this becomes

$$v_{ij} = f_{ij}(\pi) \tag{6.9}$$

$$\begin{aligned}
f_{ij}^{(uv)}(\pi) &= \frac{\partial f_{ij}(\pi)}{\partial \pi_{uv}} \\
&= p_{i.} \left[\frac{\pi_{i.}\pi_{ij}^{(uv)} - \pi_{ij}\pi_{i.}^{(uv)}}{(\pi_{i.})^2} \right] \\
&= p_{i.} \left[\sum_a \frac{(\pi_{a.}\pi_{aj}^{(uv)} - \pi_{aj}\pi_{a.}^{(uv)})}{(\pi_{a.})^2} p_{a.} \right]
\end{aligned}
\tag{6.10}$$

which at $\pi = p$ becomes

$$f_{ij}^{(uv)}(p) = \alpha_{iu} \left(\frac{p_{i \cdot} \alpha_{jv} - p_{ij}}{p_{i \cdot}} \right) - p_{i \cdot} \left(\alpha_{jv} - \frac{p_{uj}}{p_{\cdot}} \right) \quad (6.11)$$

From (4.17) we have

$$v(v_{ij}) = v(f_{ij}(\pi)) \approx \left(\frac{1}{n} - \frac{1}{N} \right) \left[\sum_u \sum_v (f_{ij}^{(uv)}(p))^2 p_{uv} - \left(\sum_u \sum_v f_{ij}^{(uv)}(p) p_{uv} \right)^2 \right] \quad (6.12)$$

Now,

$$\sum_u \sum_v f_{ij}^{(uv)}(p) p_{uv} = \sum_u \sum_v p_{uv} \left[\alpha_{iu} \left(\frac{p_{i \cdot} \alpha_{jv} - p_{ij}}{p_{i \cdot}} \right) - p_{i \cdot} \left(\alpha_{jv} - \frac{p_{uj}}{p_{\cdot}} \right) \right] \quad (6.13)$$

Let

$$\sum_u \sum_v (f_{ij}^{(uv)}(p))^2 p_{uv} = \sum_u \sum_v T_1^2 p_{uv} - 2 \sum_u \sum_v T_1 T_2 p_{uv} + \sum_u \sum_v T_2^2 p_{uv} \quad (6.14)$$

where

$$T_1 = \frac{\alpha_{iu} (p_{i \cdot} \alpha_{jv} - p_{ij})}{p_{i \cdot}}, \quad T_2 = p_{i \cdot} \left(\alpha_{jv} - \frac{p_{uj}}{p_{\cdot}} \right)$$

After some algebra, we get,

$$\begin{aligned} \sum_u \sum_v T_{11}^2 p_{uv} &= \frac{p_{1j}}{p_{1.}} (p_{1.} - p_{1j}) \\ \sum_u \sum_v T_{12}^2 p_{uv} &= p_{1j} (p_{1.} - p_{1j}) \\ \sum_u \sum_v T_{22}^2 p_{uv} &= p_{1.}^2 (p_{.j} - \sum_u \frac{p_{uj}^2}{p_{u.}}) \end{aligned} \quad (6.15)$$

Substitution of (6.13) and (6.15) in (6.12) yields

$$v(v_{1j}) = \left(\frac{1}{n} - \frac{1}{N}\right) \left[\frac{p_{1j}}{p_{1.}} (p_{1.} - p_{1j}) (1 - 2p_{1.}) + p_{1.}^2 \sum_u \frac{p_{uj} (p_{u.} - p_{uj})}{p_{u.}} \right] \quad (6.16)$$

A similar derivation gives,

$$v(v_{.j}^*) = \left(\frac{1}{n} - \frac{1}{N}\right) \left[\frac{p_{1j}}{p_{.j}} (p_{.j} - p_{1j}) (1 - 2p_{.j}) + p_{.j}^2 \sum_v \frac{p_{1v} (p_{.v} - p_{1v})}{p_{.v}} \right] \quad (6.17)$$

It is of interest to see how the estimators u_{1j} , v_{1j} and π_{1j} compare with respect to their variances. Consider

$$V(u_{ij}) - V(\pi_{ij}) = \left(\frac{1}{n} - \frac{1}{N}\right) \{2p_{ij}(3p_{i.}p_{.j} - p_{i.} - p_{.j}) + p_{i.}p_{.j}(p_{i.} + p_{.j} - 4p_{ij})\}$$

which is greater than zero if

$$\left(4 - \frac{1}{p_{i.}} - \frac{1}{p_{.j}}\right) \left(1 - \frac{p_{i.}p_{.j}}{2p_{ij}}\right) > 1 \quad (6.18)$$

This inequality is fulfilled when $p_{i.} = p_{.j} = .4$ and $p_{ij} < .04$.

A comparison of $V(v_{ij})$ and $V(\pi_{ij})$ results in much more tedious algebra than the case above. El-Badry and Stephan [7] show that $V(v_{ij}) - V(\pi_{ij})$ can be greater than zero when

$$p_{i.} + p_{.j} < 1$$

and

$$p_{ij} < \frac{p_{i.}}{(2p_{i.}-1)^2} \left\{ (p_{i.}(1-p_{i.})(1-p_{.j})(p_{i.}(1-p_{i.}) + p_{.j}(1-3p_{i.}+p_{i.}^2)) \right. \\ \left. - p_{i.}(1-p_{i.}-p_{i.}p_{.j})) \right\} \quad (6.19)$$

These inequalities are satisfied, for example, when $p_{1.} = p_{.j} = .4$ and $p_{1j} < .035$.

The results of (6.18) and (6.19) may be compared with $v(p_{1j}^{(2)}) - v(\pi_{1j})$ since the second iteration of the RREP and LCEP is the first iteration to make use of both row and column information as do v_{1j} and u_{1j} . It was shown in (4.35) that $v(p_{1j}^{(2)}) > v(\pi_{1j})$ unconditionally so that in this respect $p_{1j}^{(2)}$ and $q_{1j}^{(2)}$ are superior to u_{1j} and v_{1j} .

El-Badry and Stephan showed that

$$v(u_{1j}) \geq v(v_{1j}) \quad (6.20)$$

with equality when $p_{1j} = p_{1.} p_{.j}$ for all i . Their algebra involved

maximizing $\sum_{u \neq 1} \frac{p_{uj}(p_{u.} - p_{uj})}{p_{u.}}$ subject to the constraint $\sum_{u \neq 1} p_{uj} = p_{.j} - p_{1j}$,

substituting the result into $v(v_{1j})$ and comparing this with $v(u_{1j})$.

It is interesting to compare $v(p_{1j}^{(2)})$ with $v(v_{1j})$. It was shown in (4.33) that

$$v(p_{1j}^{(2)}) = \left(\frac{1}{n} - \frac{1}{N}\right)(p_{1j} - \frac{p_{1j}^2}{p_{1.}} - \frac{p_{1j}^2}{p_{.j}} + \frac{2p_{1j}^3}{p_{1.}p_{.j}} - \frac{p_{1j}^2}{p_{.j}^2} \sum \frac{p_{uj}^2}{p_{u.}}) \quad (6.21)$$

Now, (6.16) may be written

$$v(v_{ij}) = \left(\frac{1}{n} - \frac{1}{N}\right) \left(p_{ij} - \frac{p_{i.}^2}{p_{i.}} - 2p_{ij}p_{.j} + 2p_{ij}^2 + p_{i.}^2 p_{.j} - p_{i.}^2 \sum_{k=1}^K \frac{p_{ik}^2}{p_{i.}}\right) \quad (6.22)$$

So that, after some algebra, we get

$$v(v_{ij}) - v(p_{ij}^{(2)}) = \left(\frac{1}{n} - \frac{1}{N}\right) (p_{ij} - p_{i.}p_{.j}) \left[\frac{1}{p_{.j}} (p_{ij} - p_{i.}p_{.j}) - 2 \frac{p_{ij}^2}{p_{i.}p_{.j}} + \frac{1}{2} (p_{ij} + p_{i.}p_{.j}) \sum_{k=1}^K \frac{p_{ik}^2}{p_{i.}}\right] \quad (6.23)$$

Since $v(p_{ij}^{(2)}) < v(\pi_{ij})$ unconditionally but $v(v_{ij})$ may be greater than $v(\pi_{ij})$ there are cases where (6.23) is positive. However, it is possible for (6.23) to be negative and this need not occur only in hypothetical cases. For example, Appendix E contains the results of applying (6.23) to a table of data from the Statistics Canada Study [1]. The results show that (6.23) is negative in 11 of the 21 cells.

There are drawbacks to using the estimator u_{ij} and v_{ij} in practical situations. For example, both can yield negative estimates of the population cell proportions even when all $\pi_{ij} > 0$. A hypothetical example of this is the 2×2 table where $p_{.1} = .3$.

$p_{2.} = .7$, $p_{.1} = .2$, $p_{.2} = .8$ and $\pi_{11} = \pi_{22} = .1$, $\pi_{21} = \pi_{12} = .4$.
 In this case, $u_{11} = -.030$, $v_{11} = -.066$ and $v_{11}^+ = -.036$.

A second undesirable characteristic of the estimates u_{ij} and v_{ij} is that, unlike the HEP and LCEP, they do not yield zero estimates of the population cell proportions when $\pi_{ij} = 0$. This in itself is not a drawback, but when the methods are used to estimate cell weights, as at Statistics Canada (see Chapter 4), it can lead to problems. Denote the cell weight estimates corresponding to the cell proportion estimates u_{ij} and v_{ij} by w_{ij} and v_{ij} respectively and let N_{ij} and N_{ij} be the corresponding population cell frequency estimates. Expressions for w_{ij} and v_{ij} are derived by equating two different expressions for each of N_{ij} and N_{ij} , namely

$$N_{ij} = u_{ij} N_{i.} = N_{ij}$$

and

$$N_{ij} = v_{ij} N_{.j} = N_{ij}$$

These yield

$$u_{ij} = N_{ij} / N_{i.}$$

$$v_{ij} = N_{ij} / N_{.j}$$

If n_{ij} is zero but neither u_{ij} nor v_{ij} is zero, then $u_{ij}^{w_{ij}}$ and $v_{ij}^{w_{ij}}$ are infinite. However, this poses a problem when these weights are to be applied to sample cell frequencies of 2B items y_{ij} . The only possible convention to adopt is to let $u_{ij}^{w_{ij}} y_{ij} = v_{ij}^{w_{ij}} y_{ij} = 0$ for these cells but this implies that we must also let $u_{ij}^{w_{ij} n_{ij}} = v_{ij}^{w_{ij} n_{ij}} = 0$ which conflicts with u_{ij} and v_{ij} not being zero. Since the cell estimates u_{ij} and v_{ij} satisfy the marginals exactly, the adopted convention requiring $u_{ij}^{w_{ij} n_{ij}} = v_{ij}^{w_{ij} n_{ij}} = 0$ results in the marginals not being satisfied.

CHAPTER 7

PROPERTIES OF THE ESTIMATORS WHEN

THE ROW AND COLUMN CHARACTERISTICS ARE INDEPENDENT

7.1. Variances of the Estimators under Independence.

The estimators of the population cell proportions p_{ij} discussed thus far, namely u_{ij} (6.1), v_{ij} (6.2), $p_{ij}^{(t)}$ (2.11) and $q_{ij}^{(t)}$ (3.24) share a notable property when the row and column characteristics are independent, i.e., when $p_{ij} = p_{i.} p_{.j}$ for all i and j . Now, if it is known that $p_{ij} = p_{i.} p_{.j}$ for all i and j then obviously there is not estimation problem since $p_{i.}$ and $p_{.j}$ are known. However, the following results are relevant in cases where independence is not known but where approximate independence is assumed.

If we set $p_{ij} = p_{i.} p_{.j}$, then the equations (6.8), (6.12) and (4.33) yield the following:

$$\begin{aligned}
 V(u_{ij}) &= \left(\frac{1}{n} + \frac{1}{m} \right) [p_{i.} p_{.j} (1 - p_{i.} p_{.j}) + 2 p_{i.} p_{.j} (p_{i.} p_{.j} - p_{i.} - p_{.j}) \\
 &\quad + p_{i.} p_{.j} (p_{i.} + p_{.j} - 2 p_{i.} p_{.j})] \\
 &= \frac{1}{n} + \frac{1}{m} - p_{i.} p_{.j} \quad (7.1)
 \end{aligned}$$

$$\begin{aligned}
 v(v_{1j}) &= \left(\frac{1}{n} - \frac{1}{N}\right) \left[\frac{p_{1.} p_{.j}}{p_{1.}} (p_{1.} - p_{1.} p_{.j}) (1 - 2p_{1.}) \right. \\
 &\quad \left. + p_{1.}^2 \sum_a \frac{p_{a.} p_{.j} (p_{a.} - p_{a.} p_{.j})}{p_{a.}} \right] \\
 &= \left(\frac{1}{n} - \frac{1}{N}\right) p_{1.} p_{.j} (1 - p_{1.}) (1 - p_{.j}) \\
 &= v(u_{1j}) \quad (7.2)
 \end{aligned}$$

$$\begin{aligned}
 v(q_{1j}^{(2)}) &= v(p_{1j}^{(2)}) = \left(\frac{1}{n} - \frac{1}{N}\right) p_{1.} p_{.j} \left[1 - \frac{p_{1.} p_{.j}}{p_{.j}} - \frac{p_{1.} p_{.j}}{p_{1.}} \right. \\
 &\quad \left. + 2 \frac{p_{1.}^2 p_{.j}^2}{p_{1.} p_{.j}} - \frac{p_{1.} p_{.j}}{p_{.j}^2} \sum_a \frac{p_{a.}^2 p_{.j}^2}{p_{a.}} \right] \\
 &= \left(\frac{1}{n} - \frac{1}{N}\right) p_{1.} p_{.j} (1 - p_{1.}) (1 - p_{.j}) \\
 &= v(v_{1j}) = v(u_{1j}) \quad (7.3)
 \end{aligned}$$

It is seen from (2.11) and (3.24) that $p_{1j}^{(2)}$ and $q_{1j}^{(2)}$ are the first estimates of the RREP and LREP to make use of both row and column marginal information. Thus, if the rows and columns are independent, one might intuitively expect that $p_{1j}^{(2)}$ and $q_{1j}^{(2)}$

contain all the information that the marginals can provide and the variance of the cell estimates at higher iterations is no better than at the second iteration. This is in fact the case as is shown in the next few lines.

It was shown in (4.21) that

$$q_{ij}^{(t)} \approx p_{ij}^{(t)} \approx p_{ij} + \sum_u \sum_v (\pi_{uv} - p_{uv}) f_{ij}^{(uv)}(p) \quad (7.4)$$

where the notation is as defined in (4.12). Recursion formulae for the $f_{ij}^{(uv)}(p)$ were developed in (4.18) and (4.19), i.e.,

$$\begin{aligned} f_{(2h+1)ij}^{(uv)}(p) &= f_{(2h-2)ij}^{(uv)}(p) - \frac{p_{1j}}{p_{1.} b} f_{(2h-2)ib}^{(uv)}(p) \\ f_{(2h)ij}^{(uv)}(p) &= f_{(2h-1)ij}^{(uv)}(p) - \frac{p_{1j}}{p_{.j} a} f_{(2h-1)aj}^{(uv)}(p) \end{aligned} \quad (7.5)$$

If $p_{1j} = p_{1.} p_{.j}$ for all i and j , then using (4.30) we have

$$\begin{aligned} \sum_b f_{(2)ib}^{(uv)}(p) &= \sum_b (\pi_{ib} - p_{i.}) (\pi_{bv} - p_{.v}) \\ &= (\pi_{i.} - p_{i.}) \sum_b (\pi_{bv} - p_{.v}) \\ &= 0 \end{aligned} \quad (7.6)$$

Similarly,

$$\sum_a f^{(uw)}(2) a_j(p) = 0 \quad (7.7)$$

Substitution of (7.6) and (7.7) into (7.5) and (7.4) yields

$$u_{1j}^{(1)} \approx p_{1j}^{(1)} \approx p_{1j}^{(2)} \approx q_{1j}^{(2)} \quad (7.8)$$

to first order approximations. Therefore,

$$v(q_{1j}^{(1)}) \approx v(p_{1j}^{(1)}) \approx v(p_{1j}^{(2)}) \approx v(q_{1j}^{(2)}) \approx v(u_{1j}) \approx v(v_{1j}) \quad (7.9)$$

There are important implications of the results (7.8) and (7.9):

- 1) If the RREP or LREP is used for a table with independent (or approximately independent) rows and columns, the procedure may be stopped, from the efficiency point of view, after the second iteration; 2) If exact agreement of the marginals is necessary, then the easily computed estimators u_{1j} and v_{1j} should be considered since both satisfy the marginals exactly and have the same variance as $p_{1j}^{(1)}$ and $q_{1j}^{(1)}$ for

all $t \geq 2$; 3) By (7.8) it is seen that corrections to $p_{ij}^{(t)}$ and $q_{ij}^{(t)}$ after the second iteration are 'small' so that the second iteration estimates should satisfy the marginals quite well.

7.2. Variance Comparisons with the Maximum Likelihood Estimator of p_{ij}

Finally, we explore the relationship of the four estimators discussed to this point with maximum likelihood estimators. It is well known that maximum likelihood estimators are Best Asymptotically Normal under sampling with replacement. Since the estimates generated by the RREP and LCEP are also BAN under with replacement sampling, it follows from (7.9) that the m.l.e.'s, denoted by m_{ij}^* , of the population cell proportions $p_{ij} = p_{i.} p_{.j}$ for all i and j should have the following variance under sampling with replacement:

$$V(m_{ij}^*) \sim \frac{1}{n} p_{i.} p_{.j} (1 - p_{i.}) (1 - p_{.j}). \quad (7.10)$$

This will now be verified by reducing the general asymptotic variance formula for m_{ij}^* given by Rao [17] under the condition $p_{ij} = p_{i.} p_{.j}$ for all i and j .

Rao [17], using a result of Silvey [21] showed that the variance covariance matrix of the maximum likelihood estimator $\hat{\theta}_{ij}^*$ is given by $\frac{1}{n} D(p)$ where

$$D(p) = B^{-1}(p) - A(p) F^{-1}(p) A^T(p) \quad (7.12)$$

for which

$$B^{-1}(p) = \text{diag}(p_{11}, \dots, p_{1,s-1}; \dots; p_{r-1,1}, \dots, p_{r-1,s-1};$$

$$p_{1,s}, \dots, p_{r-1,s}; p_{r,1}, \dots, p_{r,s-1}, p_{rs})$$

$A(p) =$

$p_{1,1}$	0_{s-1}	$\dots$	0_{s-1}	$\text{diag}(p_{11}, \dots, p_{1,s-1})$	$p_{1,s}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
0_{s-1}	0_{s-1}	$\dots$	$p_{r-1,1}$	$\text{diag}(p_{r-1,1}, \dots, p_{r-1,s-1})$	$p_{r-1,s}$
$\text{diag}(p_{1,s}, \dots, p_{r-1,s})$	$0_{r-1,s-1}$				p_{rs}
$0_{s-1,r-1}$			$\text{diag}(p_{r1}, \dots, p_{r,s-1})$		$p_{r,s}$
0_{r-1}^T			0_{r-1}^T		p_{rs}

$$F(p) = \begin{bmatrix} \text{diag}(p_1, \dots, p_{r-1}, \cdot) & \begin{matrix} p_{11} & \dots & p_{1,c-1} \\ \vdots & & \vdots \\ p_{r-1,1} & \dots & p_{r-1,c-1} \end{matrix} & \begin{matrix} p_{1\cdot} \\ \vdots \\ p_{r-1,\cdot} \end{matrix} \\ \begin{matrix} p_{11} & \dots & p_{r-1,1} \\ \vdots & & \vdots \\ p_{1,c-1} & \dots & p_{r-1,c-1} \end{matrix} & \text{diag}(p_{\cdot 1}, \dots, p_{\cdot, c-1}) & \begin{matrix} p_{\cdot 1} \\ \vdots \\ p_{\cdot, c-1} \end{matrix} \\ \begin{matrix} p_{1\cdot} & \dots & p_{r-1,\cdot} \end{matrix} & \begin{matrix} p_{\cdot 1} & \dots & p_{\cdot, c-1} \end{matrix} & 1 \end{bmatrix}$$

where 0_{ab} is a null matrix of order $a \times b$, 0_a is a vector of a zeros, and

$$p_{i\cdot}^T = (p_{i,1}, \dots, p_{i,c-1}) \quad i = 1, \dots, r$$

$$p_{\cdot j}^T = (p_{1,j}, \dots, p_{r-1,j}) \quad j = 1, \dots, c$$

It can be shown (by direct verification) that if $p_{ij} = p_{i\cdot} p_{\cdot j}$

for all i and j , then

$$F^{-1}(p) = \frac{1}{p_{r\cdot} p_{\cdot r}} \begin{bmatrix} E_1 & 0_{r-1, c-1} & 0_{r-1, 1} \\ 0_{c-1, r-1} & E_2 & 0_{c-1, 1} \\ 0_{r-1, c-1}^T & 0_{r-1, 1}^T & p_{r\cdot} p_{\cdot r} \end{bmatrix}$$

where

$$E_1 = \begin{bmatrix} \frac{p_{.e}(p_{1.} + p_{r.})}{p_{1.}} & p_{.e} & \dots & p_{.e} \\ p_{.e} & \frac{p_{.e}(p_{2.} + p_{r.})}{p_{2.}} & \dots & p_{.e} \\ \vdots & \vdots & \ddots & \vdots \\ p_{.e} & \dots & p_{.e} & \frac{p_{.e}(p_{r-1.} + p_{r.})}{p_{r-1.}} \end{bmatrix}$$

$$E_2 = \begin{bmatrix} \frac{p_{r.}(p_{1.} + p_{.e})}{p_{1.}} & p_{r.} & \dots & p_{r.} \\ p_{r.} & \frac{p_{r.}(p_{2.} + p_{.e})}{p_{2.}} & \dots & p_{r.} \\ \vdots & \vdots & \ddots & \vdots \\ p_{r.} & \dots & p_{r.} & \frac{p_{r.}(p_{r-1.} + p_{.e})}{p_{r-1.}} \end{bmatrix}$$

$$e_p^T = (-p_{1.}, \dots, -p_{.e}), \quad e_r^T = (-p_{r.}, \dots, -p_{r.})$$

$$\text{and } p_{ee} = p_{r.} + p_{.e} - p_{r.} p_{.e}$$

Thus, we find, from (7.11) and (7.12) after some algebra, that the diagonal elements of $D(p)$ are $p_{1.} p_{.j} (1 - p_{1.}) (1 - p_{.j})$ so that

$$V(m_{ij}^*) = \frac{1}{n} p_{1.} p_{.j} (1 - p_{1.}) (1 - p_{.j})$$

which agrees with (7.10).

CHAPTER 8

SUMMARY AND CONCLUSIONS

We have examined the problem of estimating from a simple random sample the population cell proportions p_{ij} of a two-way contingency table for which the population marginal proportions $p_{i.}$ and $p_{.j}$ are known. Four estimation methods were considered: the Raking Ratio Estimation Procedure ($p_{ij}^{(t)}$), the Linear Correction Estimation Procedure ($q_{ij}^{(t)}$), the Method of Proportional Distribution of Marginal Adjustments (u_{ij}) and the Method of Array Proportions (v_{ij}).

The Raking Ratio Estimation Procedure (RREP) and the Linear Correction Estimation Procedure (LOEP) are easily computed iterative procedures which converge rapidly to estimates of the p_{ij} which are Best Asymptotically Normal. The estimates are consistent at all iterations and are unbiased at the zeroth and first iterations. Also,

$p_{ij}^{(t)} \approx q_{ij}^{(t)}$ for all t to first order approximations. If the rows and columns of the table are independent, then $v(p_{ij}^{(t)}) = v(q_{ij}^{(t)}) =$

$v(m_{ij}^*)$ for all $t \geq 2$, where m_{ij}^* is the maximum likelihood estimator

of p_{ij} . The LOEP differs from the RREP in two respects: 1) the LOEP

has the unfortunate characteristic of being able to produce negative

estimates of the p_{ij} and 2) the values to which the LOEP converges

can be found by solving a system of $c-1$ linear equations. The RREP has the undesirable property of possibly yielding identical estimates of p_{ij} for tables having very different π_{ij} -patterns.

The two non-iterative estimators u_{ij} and v_{ij} are easily calculated and satisfy the marginals exactly but they are not asymptotically optimal unlike the RREP and LCEP. Both u_{ij} and v_{ij} are consistent and unbiased. The estimator v_{ij} is superior to u_{ij} since its asymptotic variance is less than or equal to that of u_{ij} . In cases where the rows and columns are independent, $V(u_{ij}) = V(v_{ij}) \approx V(p_{ij}^{(2)})$. As with the LCEP, both u_{ij} and v_{ij} may yield negative estimates of p_{ij} .

Due to the variety of properties of the estimates discussed, the choice of estimator for a given situation depends on the requirements of the problem at hand.

APPENDIX A

It was stated in chapter 5 that the system of constraints (5.5) is inconsistent if and only if the table under investigation is separable unless the sum of the row population proportions equals the sum of the column population proportions in all sub-tables. The proof of the sufficiency of this result was given in chapter 5. The necessity is proven here.

Consider the following matrix formulation of (5.5):

$$A Q = P \quad (A.1)$$

where A is the $(r+c) \times (rc)$ matrix

$$A^T = \begin{bmatrix} \delta_{11} & 0 & \dots & \delta_{1c} & \delta_1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & \delta_{rc} & \delta_r \end{bmatrix}$$

in which 0_c is a column vector of c zeros.

$$\delta_1^T = (\delta_{11}, \delta_{12}, \dots, \delta_{1c})$$

$$\delta_r = (\delta_{r1}, \delta_{r2}, \dots, \delta_{rc})$$

$$\delta_{ij} = \begin{cases} 1 & \text{if } n_{ij} > 0 \\ 0 & \text{if } n_{ij} = 0 \end{cases}$$

Iteration 3

				Row Totals	
	.0000	.3269	.3312	.3300	.9880
	.0000	.2876	.2873	.4931	.9880
	.0040	.0000	.0000	.0000	.0040
	.2656	-.2079	-.0288	-.0249	.0040
	.0040	.0000	.0000	.0000	.0040
	.2951	-.0744	-.0458	-.1709	.0040
	.0040	.0000	.0000	.0000	.0040
	.2994	-.0701	-.0441	-.1812	.0040
Column Totals	.0120	.3269	.3312	.3300	
	.8601	-.0648	.0886	.1161	

Iteration 4

				Row Totals	
	.0000	.0040	.0040	.0040	.0120
	.0000	.3105	.1509	.4370	.8985
	.3293	.0000	.0000	.0000	.3293
	.3082	-.1849	-.0345	-.0277	.0611
	.3293	.0000	.0000	.0000	.3293
	.3377	-.0629	-.0571	-.1961	.0216
	.3294	.0000	.0000	.0000	.3294
	.3421	-.0586	-.0554	-.2093	.0188
Column Totals	.9880	.0040	.0040	.0040	
Totals	.9880	.0040	.0040	.0040	

22

OF/DE

and where

$$Q^T = (q_{11}, q_{12}, \dots, q_{1c}; q_{21}, q_{22}, \dots, q_{2c}; \dots; q_{r1}, q_{r2}, \dots, q_{rc})$$

$$P^T = (p_1, p_2, \dots, p_r, p_{.1}, p_{.2}, \dots, p_{.c})$$

Note that the first r rows and the remaining c rows of A form the matrices of coefficients of the row and column constraints respectively of (5.5).

The rank of A is at most $r+c-1$ since the sum of the first r rows equals the sum of the last c rows. Thus, we may write

$$\begin{aligned} 0_{rc}^T &= \alpha_1 a^{(1)} + \alpha_2 a^{(2)} + \dots + \alpha_r a^{(r)} + \alpha_{r+1} a^{(r+1)} \\ &\quad + \dots + \alpha_{r+c} a^{(r+c)} \end{aligned} \quad (A.2)$$

where the $a^{(i)}$, $i=1, \dots, r$ are the rows of A corresponding to the row constraints, the $a^{(r+j)}$, $j=1, \dots, c$ are the rows of A corresponding to the column constraints and the α_i , $i=1, \dots, r$ and α_{r+j} , $j=1, \dots, c$ are scalars not all equal to zero. Note that for any (i, j) the corresponding δ_{ij} occurs exactly twice in A since each cell occurs in one row and one column constraint and so (A.2) may be expressed component-wise as

$$0 = \delta_{ij} (\alpha_i + \alpha_{r+j}) \quad (A.3)$$

Thus, if $\delta_{ij} = 1$ then α_i and α_{r+j} are both zero or both nonzero. If $\delta_{ij} = 0$ this condition need not hold.

If (A.1) is an inconsistent system then the rank of A is at most $r+c-2$. Thus, we can write (A.2) with at least one α_i or α_{r+j} equal to zero and not all other α_i and α_{r+j} equal to zero. Under these conditions, consider a rearrangement of the rows and columns of the table with the cells for which both α_i and α_{r+j} not zero in the upper left corner and denote this sub-table as A_1 . Denote the remaining blocks in the table as in the following diagram:

A_1	B_2
B_1	A_2

Thus, the sub-tables B_1 and B_2 contain cells for which one and only one of α_i and α_{r+j} is zero while A_2 contains cells for which both α_i and α_{r+j} are zero.

There must be at least one cell in A_1 . Assuming the contrary implies $\alpha_i = 0$, $i=1, \dots, r$ or $\alpha_{r+j} = 0$, $j=1, \dots, c$. The former condition reduces (A.3) to $\delta_{ij} \alpha_{r+j} = 0$. By hypothesis, at least one α_{r+j} , say α_{r+b} , is not zero. Thus, by (A.3) we have $\delta_{ib} = 0$ for all i so that $\pi_{ib} = 0$ for all i . Hence, $\pi_{ib} = 0$ which is contrary to the conditions

assumed. The condition $\alpha_{r+j} = 0, j=1, \dots, c$ leads to a comparable result.

Similar reasoning shows that A_2 must contain at least one cell. Assuming the contrary implies that either $\alpha_i \neq 0, i=1, \dots, r$ or $\alpha_{r+j} \neq 0, j=1, \dots, c$. The former condition implies that, by hypothesis, some α_{r+j} , say α_{r+b} , equals zero. Thus, (A.3) reduces to $\delta_{ib} \alpha_i = 0, i=1, \dots, r$ from which we get $\delta_{ib} = 0, i=1, \dots, r$ and so $\pi_{ib} = 0$ for all i . Hence, $\pi_{.b} = 0$ which is contrary to the assumptions. A similar result holds if the condition $\alpha_{r+j} = 0; j=1, \dots, c$ is assumed.

Consider now the cells in B_1 and B_2 . Since one and only one of α_i and α_{r+j} is zero for these cells, it follows from (A.3) that $\delta_{ij} = 0$ so that $\pi_{ij} = 0$. Thus, B_1 and B_2 contain only cells for which $\pi_{ij} = 0$ and so the table is separable. This proves the necessity.

APPENDIX B

It was shown in chapter 3 that the solution to the problem of minimizing (3.1) subject to (3.2) could be expressed as

$$q_{ij}^* = (1 + c_i + d_j) \pi_{ij} = (a_i + b_j) \pi_{ij} \quad (\text{B.1})$$

where a_i, b_j, c_i, d_j are Lagrange multipliers. Consider (5.5) expressed in terms of the above a_i and b_j , i.e.

$$\begin{aligned} \sum_j \pi_{ij} (a_i + b_j) &= p_i, & i = 1, \dots, r \\ \sum_i \pi_{ij} (a_i + b_j) &= p_j, & j = 1, \dots, c \end{aligned} \quad (\text{B.2})$$

Clearly, if there is a solution to (B.2) then there is a solution to (5.5) namely (B.1). On the other hand, if there is a solution to (5.5) in the q_{ij} it is not clear whether there is necessarily a solution to (B.2) in the a_i and b_j . The following proof, however, shows that (5.5) is consistent if and only if (B.2) is consistent.

It was shown in appendix A that (5.5) may be written

$$A Q = P \quad (\text{B.3})$$

where the notation is as defined in (A.1). Now, (B.2) can be written

$$A \Pi A^T L = P \quad (B.4)$$

where

$$\Pi = \text{diag} (\pi_{11}, \dots, \pi_{1c}; \pi_{21}, \dots, \pi_{2c}; \dots; \pi_{r1}, \dots, \pi_{rc})$$

$$L^T = (a_1, a_2, \dots, a_r, b_1, b_2, \dots, b_c)$$

Since $A \Pi A^T$ is symmetric we can write

$$A \Pi A^T = (\Pi^{\frac{1}{2}} A^T)^T \Pi^{\frac{1}{2}} A^T \quad (B.5)$$

where

$$\Pi^{\frac{1}{2}} = \text{diag} (\pi_{11}^{\frac{1}{2}}, \dots, \pi_{1c}^{\frac{1}{2}}; \pi_{21}^{\frac{1}{2}}, \dots, \pi_{2c}^{\frac{1}{2}}; \dots; \pi_{r1}^{\frac{1}{2}}, \dots, \pi_{rc}^{\frac{1}{2}})$$

Now, $\Pi^{\frac{1}{2}} A^T$ is the matrix A^T with the nonzero rows multiplied by nonzero constants. Thus, the rank of $\Pi^{\frac{1}{2}} A^T$ is equal to the rank of A^T . Since the rank of a matrix M is equal to the rank of $M^T M$ [14], the rank of $A \Pi A^T$ is equal to the rank of A^T .

It can be shown that if M is an $m \times n$ matrix and N is an $n \times p$ matrix, then the column space of $M N$ is contained in the column space of M [14]. Using this result, it is seen that since the rank of $A \Pi A^T$ is equal to the rank of A , the column space of $A \Pi A^T$ is the same as the column space of A . Thus, if P belongs to this column space

then both (B.3) and (B.4) have a solution. If P does not belong to this column space, then (B.3) and (B.4) are both inconsistent. This proves the result.

APPENDIX C

Consider applying the RREP and LCEP to the following hypothetical separable table of sample cell values and population marginals:

.04	.00	.15	.00	$\pi_{1.} = .19$ $p_{1.} = .225$
.06	.00	.15	.00	$\pi_{2.} = .21$ $p_{2.} = .18$
.00	.10	.00	.20	$\pi_{3.} = .30$ $p_{3.} = .35$
.00	.05	.00	.25	$\pi_{4.} = .30$ $p_{4.} = .245$
$\pi_{.1} = .10$ $p_{.1} = .15$	$\pi_{.2} = .15$ $p_{.2} = .10$	$\pi_{.3} = .30$ $p_{.3} = .21$	$\pi_{.4} = .45$ $p_{.4} = .54$	

Interchanging columns 2 and 3 shows that the above table is separable. Note that the sum of the population row marginals does not equal the sum of the population column marginals for either sub-table (for example, $p_{.1} + p_{.3} \neq p_{1.} + p_{2.}$).

In the following tables of iterated values, the upper value in each cell is the RREP estimate while the lower value is the LCEP estimate. The phrase 'RREP SSQM' refers to either $[\sum_i (p_{i.}^{(2t)} - p_{i.})^2 / p_{i.}^2]^{\frac{1}{2}}$ or $[\sum_j (p_{.j}^{(2t-1)} - p_{.j})^2 / p_{.j}^2]^{\frac{1}{2}}$. Similarly, 'LCEP SSQM' denotes either $[\sum_i (q_{i.}^{(2t)} - q_{i.})^2 / q_{i.}^2]^{\frac{1}{2}}$ or $[\sum_j (q_{.j}^{(2t-1)} - q_{.j})^2 / q_{.j}^2]^{\frac{1}{2}}$.

Iteration 1

				Row Totals	
	.0474	.0000	.1776	.0000	.2250
	.0474	.0000	.1776	.0000	.2250
	.0514	.0000	.1286	.0000	.1800
	.0514	.0000	.1286	.0000	.1800
	.0000	.1167	.0000	.2333	.3500
	.0000	.1167	.0000	.2333	.3500
	.0000	.0408	.0000	.2042	.2450
	.0000	.0408	.0000	.2042	.2450
Column	.0988	.1575	.3062	.4375	
Totals	.0988	.1575	.3062	.4375	

RREP SSQM = .555 LCEP SSQM = .555

Iteration 2

				Row Totals	
	.0719	.0000	.1218	.0000	.1937
	.0678	.0000	.1295	.0000	.1974
	.0781	.0000	.0882	.0000	.1663
	.0822	.0000	.0805	.0000	.1626
	.0000	.0741	.0000	.2880	.3621
	.0000	.0217	.0000	.2789	.3572
	.0000	.0259	.0000	.2520	.2779
	.0000	.0217	.0000	.2611	.2828
Column	.1500	.1000	.2100	.5400	
Totals	.1500	.1000	.2100	.5400	

RREP SSQM = .048 LCEP SSQM = .048

Iteration 3

					Row Totals
	.0835	.0000	.1415	.0000	.2250
	.0737	.0000	.1513	.0000	.2250
	.0845	.0000	.0955	.0000	.1800
	.0871	.0000	.0929	.0000	.1800
	.0000	.0716	.0000	.2784	.3500
	.0000	.0759	.0000	.2741	.3500
	.0000	.0229	.0000	.2221	.2450
	.0000	.0154	.0000	.2296	.2450
Column	.1681	.0945	.2784	.5005	
Totals	.1608	.0913	.2741	.5037	

RREP SSQM = .034 LCEP SSQM = .038

Iteration 4

				Row Totals	
	.0745	.0000	.1254	.0000	.1999
	.0694	.0000	.1342	.0000	.2036
	.0755	.0000	.0846	.0000	.1601
	.0806	.0000	.0758	.0000	.1564
	.0000	.0758	.0000	.3003	.3761
	.0000	.0817	.0000	.2902	.3719
	.0000	.0242	.0000	.2397	.2639
	.0000	.0183	.0000	.2498	.2681
Column	.1500	.1000	.2100	.5400	
Totals	.1500	.1000	.2100	.5400	

RREP SSQM = .041 LCEP SSQM = .044

Iteration 5

				Row Totals	
	.0839	.0000	.1411	.0000	.2250
	.0739	.0000	.1511	.0000	.2250
	.0849	.0000	.0951	.0000	.1800
	.0874	.0000	.0926	.0000	.1800
	.0000	.0705	.0000	.2795	.3500
	.0000	.0744	.0000	.2756	.3500
	.0000	.0225	.0000	.2225	.2450
	.0000	.0144	.0000	.2306	.2450
Column	.1687	.0930	.2363	.5020	
Totals	.1612	.0888	.2438	.5062	

RREP SSQM = .036 LCEP SSQM = .044

Iteration 6

				Row Totals	
	.0746	.0000	.1254	.0000	.2000
	.0694	.0000	.1343	.0000	.2036
	.0754	.0000	.0846	.0000	.1600
	.0806	.0000	.0757	.0000	.1564
	.0000	.0758	.0000	.3006	.3765
	.0000	.0819	.0000	.2906	.3725
	.0000	.0242	.0000	.2394	.2635
	.0000	.0181	.0000	.2494	.2675
Column	.1500	.1000	.2100	.5400	
Totals	.1500	.1000	.2100	.5400	

RREP SSQM = .041 LCEP SSQM = .044

APPENDIX D

This appendix includes the results of an application of the RREP and LCEP to a 4×4 table in which only one $\pi_{ij} = 0$ but where all solutions to the system (5.5) have at least one $q_{ij} < 0$. The table of sample cell values and population marginals is as follows:

.00	.10	.10	.20	$\pi_1 = .40$ $p_1 = .988$
.05	.10	.01	.01	$\pi_2 = .17$ $p_2 = .004$
.05	.05	.02	.09	$\pi_3 = .21$ $p_3 = .004$
.05	.05	.02	.1010	$\pi_4 = .22$ $p_4 = .004$
$\pi_1 = .15$ $p_1 = .988$	$\pi_2 = .30$ $p_2 = .004$	$\pi_3 = .15$ $p_3 = .004$	$\pi_4 = .40$ $p_4 = .004$	

In the following tables of iterated estimates, the upper value in each cell is the RREP estimate and the lower value in each cell is the LCEP estimate.

Iteration 1

				Row Totals	
	.0000	.2470	.2470	.4940	.9880
	.0000	.2470	.2470	.4940	.9880
	.0012	.0024	.0002	.0002	.0040
	.0012	.0024	.0002	.0002	.0040
	.0009	.0010	.0004	.0017	.0040
	.0009	.0010	.0004	.0017	.0040
	.0009	.0009	.0004	.0018	.0040
	.0009	.0009	.0004	.0018	.0040
Column	.0030	.2513	.2480	.4977	
Totals	.0030	.2513	.2480	.4977	

Iteration 2

				Row Totals	
	.0000	.0039	.0040	.0040	.0120
	.0000	.1646	.0843	.2471	.4961
	.3826	.0001	.0000	.0000	.3827
	.3295	-.0801	-.0160	-.0121	.2213
	.3097	.0000	.0000	.0000	.3098
	.3293	-.0402	-.0321	-.1094	.1475
	.2957	.0000	.0000	.0000	.2957
	.3292	-.0403	-.0322	-.1216	.1351
Column	.9880	.0040	.0040	.0040	
Totals	.9880	.0040	.0040	.0040	

Iteration 5

					Row Totals
	.0000	.3294	.3293	.3293	.9880
	.0000	.3329	.1733	.4818	.9880
	.0040	.0000	.0000	.0000	.0040
	.2914	-.2185	-.0378	-.0311	.0040
	.0040	.0000	.0000	.0000	.0040
	.3335	-.0671	-.0588	-.2036	.0040
	.0040	.0000	.0000	.0000	.0040
	.3387	-.0620	-.0567	-.2160	.0040
Column Totals	.0120	.3294	.3293	.3293	
	.9836	-.0148	.0200	.0311	

Iteration 6

					Row Totals
	.0000	.0040	.0040	.0040	.0120
	.0000	.3392	.1626	.4682	.9700
	.3293	.0000	.0000	.0000	.3293
	.2995	-.2123	-.0389	-.0317	.0166
	.3293	.0000	.0000	.0000	.3293
	.3417	-.0640	-.0609	-.2097	.0070
	.3294	.0000	.0000	.0000	.3294
	.3468	-.0589	-.0588	-.2228	.0063
Column Totals	.9880	.0040	.0040	.0040	
	.9880	.0040	.0040	.0040	

APPENDIX E

The following table shows the variance comparisons of $p_{ij}^{(2)}$ and v_{ij} as calculated from (6.23). The data used is from the household table of weighting area number 12 of the Statistics Canada study [1]. The results of (6.23) multiplied by 10^3 are:

-.043	.036	.007
-.161	.000	.025
.141	-.018	-.239
-.092	-.057	.065
.076	-.028	-.031
-.071	-.025	.033
-.042	.003	.018

The results clearly indicate that we may have $V(v_{ij}) < V(p_{ij}^{(2)})$.

BIBLIOGRAPHY

- [1] Bankier, M. D. (1977). 'Properties of Raking Ratio Estimators'.
Masters Thesis: Carleton University, Department of Mathematics.
- [2] Brackstone, G. J. and Arora, H. R. (1975). 'The Bias and Variance
of the Raking Ratio Estimator'. Technical Report, Statistics
Canada.
- [3] Brackstone, G. J. (1974). 'Raking Ratio Estimation'. Technical
Report, Statistics Canada.
- [4] Bishop, Y. M. M. and Feinberg, S. E. (1969). 'Incomplete Two-
Dimensional Contingency Tables'. Biometrics 25, 119-128.
- [5] Cochran, W. G. (1963). Sampling Techniques. John Wiley and Sons,
New York.
- [6] Deming, W. E. and Stephan, F. F. (1940). 'On a Least Squares
Adjustment of a Sampled Frequency Table when the Expected
Marginal Totals are Known'. Ann. Math. Statist. 34, 911-934.
- [7] El-Badry, M. A. and Stephan, F. F. (1955). 'On Adjusting Sample
Tabulations to Census Counts'. J. Amer. Statist. Assoc.,
738-762.
- [8] Friedlander, D. (1961). 'A Technique for Estimating a Contingency
Table, Given the Marginal Totals and some Supplementary Data'.
J. Roy. Statist. Soc., Ser A 124, 412-420.

- [9] Gass, S. I. (1964). Linear Programming. McGraw-Hill, New York.
- [10] Goodman, L. A. (1968). 'The Analysis of Cross-Classified Data: Independence, Quasi-Independence, and Interactions in Contingency Tables with or without Missing Entries'. J. Amer. Statist. Assoc. 63, 1091-1131.
- [11] Ireland, C. T. and Kullback, S. (1968). 'Contingency Tables with Given Marginals'. Biometrika, 179-188.
- [12] Ireland, C. T. (1974). 'The Rake's Progress'. Unpublished paper presented at the August 1974 American Statistical Association Meetings.
- [13] Neyman, J. (1949). 'Contribution to the Theory of the χ^2 Test'. Proc. Berkeley Symp., 239-273.
- [14] Noble, B. (1969). Applied Linear Algebra. Prentice-Hall, New Jersey.
- [15] Rao, J. N. K. (1974). 'Raking Ratio Estimators'. Technical Report, Statistics Canada.
- [16] Rao, J. N. K. (1974). 'Raking Ratio Estimators II'. Technical Report, Statistics Canada.
- [17] Rao, J. N. K. (1974). 'Raking Ratio Estimators III'. Technical Report, Statistics Canada.
- [18] Smith, J. H. (1947). 'Estimation of Linear Functions'. Ann. Math. Statist., 18, No. 2, 231-255.

- [19] Stephan, F. F. (1942). 'An Iterative Method of Adjusting Sample Frequency Tables when Expected Marginal Totals are Known'.
Ann. Math. Statist. 13, No. 2, 166-178.
- [20] Saleh, A. K. M. E. (1975). 'Asymptotic Equivalence of Contingency Probabilities using Maximum Likelihood Method, Minimum χ^2 Method and Minimum Discrimination Information'. Technical Report, Statistics Canada.
- [21] Silvey, S. D. (1959). 'The Lagrangian Multiplier Test'. Ann. Math. Statist., 389-407.

1
END

RECEIVED

FIN