Statistics Statistique Canada Canada

# SURVEY METHODOLOGY 

## DECEMBER 1975

## VOLUME 1 - SUPPLEMENTARY ISSUE

A JOURNAL PRODUCED BY
HOUSEHOLD SURVEYS DEVELOPMENT DIVISION
STATISTICAL SERVICES FIELD
STATISTICS CANADA
$\qquad$


$\square$


 . a
 0
$\begin{array}{ccc} \\ -\infty & 0\end{array}$

$\square$
(20.
$\qquad$
$\qquad$
$\qquad$
$\qquad$
$\qquad$ 4
$\qquad$
$\qquad$ \&
$\qquad$
$\qquad$
$\qquad$
$\qquad$

11

7
$\qquad$
$\qquad$
$\qquad$
$\square$
$\qquad$
$\qquad$
$\qquad$
$\qquad$



## SURVEY METHODOLOGY

December 1975 Volume 1 Supplementary Issue
ANALYTIC STUDIES OF SAMPLE SURVEY DATA
CONTENTS
Introduction ..... 1

1. Estimation of Domain Totals and Means ..... 2
2. Design of Analytical Surveys ..... 11
3. Hypothesis Testing ..... 28
4. Balanced Repeated Replication (BRR) ..... 36
5. The Jack-knife Method ..... 51
6. Taylor Expansion Method ..... 55
7. Empirical Study ..... 58
8. A General Method ..... 61
9. Estimating Relationships between Variables ..... 63
Acknowledgements ..... 76
RAN
December 1975 Volume 1 Supplementary Issue

A Journal produced by Household Surveys Development Division, Statistical Services Field, Statistics Canada.

| Editorial Board: | R. Platek - Chairman |
| :--- | :--- |
|  | M.P. Singh - Editor |
|  | P.F. Timmons |

Editorial Policy:
The objective of the Survey Methodology Journal is to provide a forum in a Canadian context for publication of articles on the practical applications of the many aspects of survey methodology. The Survey Methodology Journal will publish articles dealing with all phases of methodological development in surveys, such as, design problems in the context of practical constraints, data collection techniques and their effect on survey results, non-sampling exrors, sampling systems development and application, statistical analysis, interpretation, evaluation and interrelationships among all of these survey phases. The emphasis will be on the development strategy and evaluation of specific survey methodologies as applied to actual surveys. All papers will be refereed, however, the authors retain full responsibility for the contents of their papers and opinions expressed are not necessarily those of the Editorial Board or the Department Copies of papers in either Official Language will be made available upon request.

## Submission of Papers:

The Journal will be issued twice a year. Authors are invited to submit their papers, in either of the two Official Languages, to the Editor Dr. M.P. Singh, Household Surveys Development Division, Statistics Canada, loth Floor, Coats Building, Tunney's Pasture, Ottawa, Ontario KlA OT6. Two copies of each paper, typed space-and-a-half, are requested. Authors of articles for this journal are free to have their articles published in other statistical journals.

ANALYTIC STUDIES OF SAMPLE SURVEY DATA<br>J.N.K. Rao ${ }^{1}$


#### Abstract

Most sample surveys in the past have been 'descriptive' ir the sense that the main objective is the computation of means or totals of a number of characters of interest along with their standard errors. However, in recent years data produced from 'descriptive' surveys are also being increasingly used for 'analytical' purposes, i.e., for investigating relationships among variables. Also some sample surveys might have primary 'analytical goals' in which case the 'optimal' designing of such 'analytical surveys' become important.

These lecture notes present an account of some recent developments in the analytical studies of sample survey data. Many challenging problems remain to be solved and I hope these notes will provide stimulation for further research in this important area.


## ETUDES ANALYTIQUES DE DONNEES PROVENANT D'ENQUETES PAR SONDAGE

par J.N.K. Rao ${ }^{1}$

La plupart des enquêtes menées dans le passé ont été de type "descriptif", c'est-à-dire ayant comme principal objectif le calcul de moyennes et de totaux associé à un certain nombre de caractères ainsi que de leurs erreurs-types. Cependant, plus récemment, les données provenant d'enquêtes par sondage de type "descriptif" sont de plus en plus utilisées à des fins analytiques, c'est-à-dire pour êtudier les relations entre les variables. Il peut également se faire que des enquêtes par sondage aient principalement pour but 1""analyse"; dans un tel cas, la recherche d'un plan "optimal" pour ces enquêtes de type "analytique" devient importante.

Ces notes présentent un "bilan" sur quelques récents développements en ce qui concerne les études analytiques sur les données provenant d'enquêtes par sondage. Il reste plusieurs problèmes intéressants à résoudre et $j$ 'espère que ces notes seront un stimulant pour la recherche dans cet important secteur.

1 Professor, Department of Mathematics, Carleton University. Consultant, Statistical Services Field.

## INTRCDUCTION

Most sample surveys in the past have been "descriptive" in the sense that the main objective is the computation of means and totals of a number of characters attached to the units of the population, along with their standard errors. However, data produced from a "descriptive" survey can also be used for "analytical" purposes, i.e. for investigating relationships among variables. Such an analysis often involves the comparison of means of certain subgroups or "domains" of the population. "Domains" are usually well defined but it will not be known until after sampling which of the "domains" a particular unit belongs to. Consequently, the sample size within a "domain" is a random variable. Also, the size of the domain is usually unknown; in fact, it is often a parameter of interest.

A simple method of estimation of domain totals and means, which is applicable for any sampling design however complex it may be, will be given in Section 1. This method requires only the standard formulae pertinent to the estimation of a population total or a population ratio.

Some sample surveys might have primary "analytical" goals. "Optimal" designing of such "analytical" surveys will be considered in Section 2 employing double sampling or two-phase sampling.

Classical methods of statistical inference assume simple random sampling from infinite populations. Extensions of these methods to cover the complex designs usually employed in survey work is a formidable task. Certain tests of independence in contingency tables from stratified samples, however, have been proposed. Ingenious methods, such as the "jack-knife"
and "balanced half-sample replication", to circumvent the theoretical difficulties have also been proposed. These methods will be considered in Section 3-9.

## 1. Estimation of Domain Totals and Means

Suppose ${ }_{i} N$ elements (ultimate units) in the population and $i^{n}$ in the sample belong to domain $i\left(s a y D_{i}\right)$. Appropriate to any particular sampling design, there are three basic formulae pertinent to the estimation of population total $Y\left(Y_{j}\right)=\Sigma y_{j}$ of a character of interest " $y$ ":

$$
\begin{array}{ll}
\text { Estimator of } Y: & \hat{Y}=\hat{Y}\left(Y_{j}\right) \\
\text { Variance of } \hat{Y}: & V(\hat{Y})=V\left(Y_{j}\right) \\
\text { Estimator of } V(\hat{Y}): & V(\hat{Y})=V\left(Y_{j}\right) . \tag{1.3}
\end{array}
$$

The estimator $\hat{Y}$ usually is of the form $\hat{Y}\left(y_{j}\right)=\sum_{j \varepsilon s} w_{j} y_{j}$ where "s" denotes a sample. The operators $Y, V$ and $v$ depend on the sample design. The three basic formulae (1.1) - (1.3) are applicable for any set of characters $y_{j}$ attached to the population elements. Consequently, the same formulae could be used to estimate domain totals $i Y$ by attaching the following characters ${ }_{i} Y_{j}$ to all the elements in the population:

$$
i_{j}^{y_{j}}= \begin{cases}y_{j} & \text { if } j^{\text {th }} \text { element belongs to } D_{i}  \tag{1.4}\\ 0 & \text { otherwise }, j=1,2, \ldots, N\end{cases}
$$

Noting that ${ }_{i} Y=Y\left({ }_{i} Y_{j}\right)$, it immediately follows that

$$
\begin{align*}
\hat{Y} & =\hat{Y}\left({ }_{i} Y_{j}\right)  \tag{1.5}\\
V(\hat{Y}) & =V\left({ }_{i} Y_{j}\right)  \tag{1.6}\\
v(\hat{Y}) & =v\left({ }_{i} Y_{j}\right) \tag{1.7}
\end{align*}
$$



An alternative way of defining ${ }_{i} y_{j} \quad: \quad i_{i} y_{j}={ }_{i} a_{j} y_{j}$ where

$$
i_{j}=\left\{\begin{array}{l}
I \text { if } j^{\text {th }} \text { element belongs to } D_{i}  \tag{1.8}\\
0 \text { otherwise }
\end{array}\right.
$$

If $\hat{Y}$ and $V(\hat{Y})$ are unbiased for $Y$ and $V(\hat{Y})$ respectively, then the corresponding domain estimators $\hat{i}^{Y}$ and $v(\hat{Y})$ are also unbiased for ${ }_{i} Y$ and $V(\hat{i})$ respectively. The meaning of the formulae (1.5) - (1.7) is simply that the standard formulae be applied to the "synthetic" characters $i_{j}{ }_{j}$ defined by (1.4). Spelling out of these formulae in terms of the $y$ values of the elements in $D_{i}$ will often result in simplification and lead to useful insights. For numerical evaluation, however, it is expedient to use the standard formulae by putting $y_{j}=0$ when the element $j$ is not in $D_{i}$.

Noting that ${ }_{i} N=Y\left({ }_{i}{ }_{j}\right)$, it immediately follows that

$$
\begin{equation*}
\hat{\mathbf{N}}=\hat{Y}\left({ }_{i} a_{j}\right) \tag{1.9}
\end{equation*}
$$

Since the domain mean $\bar{Y}={ }_{i} Y /{ }_{i} N$, it follows from (1.1) and (1.9) that

$$
\begin{equation*}
\hat{\bar{Y}}=\frac{\hat{Y}\left(i y_{j}\right)}{\hat{Y}\left({ }_{i} a_{j}\right)} \tag{1.10}
\end{equation*}
$$

provided ${ }_{i} \hat{N}>0$ (i.e., at least one element in the sample belongs to $D_{i}$ ) which, in general, is a ratio estimator. Employing the classical approximate variance formulae for ratio estimators, we have:

$$
\begin{align*}
& V\left({ }_{i} \hat{\bar{Y}}\right) \doteq \frac{V\left({ }_{i} y_{j}-\bar{i}_{i} a_{j}\right)}{i^{2}}  \tag{1.11}\\
& V\left({ }_{i} \hat{\bar{Y}}\right) \equiv \frac{V\left({ }_{i} Y_{j}-\hat{\bar{Y}}_{i} a_{j}\right)}{\hat{i}^{2}} \tag{1.12}
\end{align*}
$$

For self-weighting designs, $\quad \hat{\bar{Y}}$ reduces to the simple mean $i_{i} \bar{y}=\sum_{i}^{n} y_{j} / i^{n}$. Note that ${ }_{i} y_{j}-{ }_{i} \bar{Y}_{i} a_{j}$ takes the value $y_{j}-{ }_{i} \bar{Y}$ if $j^{\text {th }}$ element is in $D_{i}$, otherwise it is zero.

For the comparison of domain means, say ${ }_{1} \bar{Y}$ and ${ }_{2} \bar{Y}$, we have:

$$
\begin{align*}
& V\left(1_{1}^{\hat{\bar{Y}}}-2^{\hat{\bar{Y}}}\right) \doteq v\left(\frac{1^{y_{j}}-{ }_{1} \bar{Y}_{1} a_{j}}{1^{N}}-\frac{2^{y_{j}}-2^{\bar{y}} 2_{j} a_{j}^{N}}{2^{2}}\right)  \tag{1.13}\\
& V\left({ }_{1} \hat{\bar{Y}}-\frac{\hat{\bar{Y}}}{2^{\prime}} \doteq v\left(\frac{1^{y_{j}}-1_{1}^{\hat{\bar{Y}}} 1_{j}}{1^{N}}-\frac{2^{Y_{j}}-2_{2}^{\hat{\bar{Y}}_{2} a_{j}}}{2^{\hat{N}}}\right)\right. \tag{1.14}
\end{align*}
$$

Since the "domains" are usually non-overlapping, the expression in brackets on r.h.s. of (1.13) reduces to $\left(Y_{j}-\bar{I}^{\bar{Y}}\right) /{ }_{1} N$ if $j^{\text {th }}$ element in $D_{1}$, to - $\left(Y_{j}-{ }_{2} \bar{Y}\right) / 2^{N}$ if $j^{\text {th }}$ element in $D_{2}$ and to zero otherwise.

If we are interested in the proportion of a measured characteristic which falls in $D_{i}$, i.e.. ${ }_{i} P={ }_{i} Y / Y$, we use:

$$
\begin{equation*}
{ }_{i} \hat{P}=\frac{\hat{Y}\left({ }_{i} y_{j}\right)}{\hat{Y}\left(y_{j}\right)} \tag{1.15}
\end{equation*}
$$

with variance:

$$
\begin{equation*}
V\left({ }_{i} \hat{P}\right) \doteq \frac{1}{\hat{Y}^{2}} V\left({ }_{i} Y_{j}-{ }_{i} P Y_{j}\right) \tag{1.16}
\end{equation*}
$$

and variance estimator:

$$
\begin{equation*}
v\left({ }_{i} \hat{P}\right) \doteq \frac{1}{\hat{Y}^{2}} v\left({ }_{i} y_{j}-\hat{i}_{i} y_{j}\right) \tag{1.17}
\end{equation*}
$$

For instance, in a consumer expenditure survey we might be interested in the proportion of total milk consumption which is attributable to families
in the income group $\$ 6,000$ to $\$ 9,000$. Note that ${ }_{i} Y_{j}-{ }_{i} P_{j}$ takes the value $y_{j}\left(1-{ }_{i}\right.$ P) if $j^{\text {th }}$ element belongs to $D_{i}$, otherwise it equals $-{ }_{i} P_{j}$. One could test the hypothesis $\bar{Y}={ }_{2} \bar{Y}$ or set confidence limits on $1^{\bar{Y}}-{ }_{2} \bar{Y}$ by assuming that $\hat{\bar{Y}}-2_{2} \hat{\bar{Y}}$ is approximately normal with mean $\bar{Y}-2_{2} \bar{Y}$ and variance $v\left({ }_{1} \hat{\bar{Y}}-2_{2} \hat{\bar{Y}}\right)$. An application of 'balanced half-sample replication' for simplifying the computation of standard errors will be given in section 1.3.

Examples:

1. Stratified Simple Random Sampling

We will now spell out the formulae for the simple case of stratified simple random sampling. We adopt the following notation for stratum $h(h=1, \ldots, L)$ and domain $i:$

$$
\begin{aligned}
i^{N}{ }_{h} & =\text { no. of elements, } \\
i^{n_{h}} & =\text { no. of elements drawn in the sample, } \\
i^{Y} Y_{h} & =\text { population total, } \\
i^{Y_{h}} & =\text { sample total, } \\
i^{S_{h}^{2}} & =\text { population mean square, } \\
i^{S_{h}^{2}} & =\text { sample mean square, } \\
i_{Y_{h}} & =\text { population mean, } \\
i_{Y_{h}} & =\text { sample mean. }
\end{aligned}
$$

For stratum $h$, we have

| $\mathrm{n}_{\mathrm{h}}$ | $=$ no. of elements drawn in the sample, |
| :--- | :--- |
| $\mathrm{N}_{\mathrm{h}}$ | $=$ no. of elements, |
| $\bar{Y}_{\mathrm{h}}$ | $=$ sample mean, |
| $\mathrm{s}_{\mathrm{h}}^{2}$ | $=$ sample mean square. |

Then letting $a_{h}=N_{h}\left(N_{h}-n_{h}\right) / n_{h}$, we get

$$
\begin{align*}
\hat{Y} & =\sum \frac{N_{h}}{n_{h}} i_{i} Y_{h}  \tag{1.18}\\
V(\hat{Y}) & =\sum \frac{a_{h}}{N_{h}-1}\left\{\left({ }_{i} N_{h}-1\right) i_{i} s_{h}^{2}+\left(\frac{1}{N_{h}}-\frac{1}{N_{h}}\right) i_{h}^{2}\right\}  \tag{1.19}\\
v(\hat{Y}) & =\sum_{h} \frac{a_{h}}{n_{h}-1}\left\{\left({ }_{i} n_{h}-1\right) i_{i} s_{h}^{2}+\left(\frac{1}{n_{h}}-\frac{1}{\Omega_{h}}\right) i_{h} Y_{h}^{2}\right\} . \tag{1.20}
\end{align*}
$$

The last terms in (1.19) and (1.20) arise due to not having the "domain"

## frame.

For the mean $i^{\bar{Y}}$ we have

$$
\begin{align*}
& i_{i}^{\hat{Y}}=\left(\sum_{h} \frac{N_{h}}{n_{h}} \quad i^{y_{h}}\right) /\left(\sum_{h}^{N_{h}} i_{h} n_{h}\right)  \tag{1.21}\\
& ={ }_{i} \bar{Y} \text { if } n_{h} \propto N_{h} \text { (ide., self weighting) } \\
& V\left({ }_{i} \hat{\bar{Y}}\right) \stackrel{i^{N}}{ } \sum_{h}^{-2} \frac{a_{h}}{N_{h}-1}\left\{\left({ }_{i} N_{h}-1\right){ }_{i} S_{h}^{2}+\left(\frac{1}{N_{h}}-\frac{1}{N_{h}}\right) i_{i} N_{h}^{2}\left(\bar{Y}_{h}-\bar{Y}^{\bar{Y}}\right)^{2}\right\} \tag{1.22}
\end{align*}
$$

Similarly,

$$
\begin{align*}
& v\left(\hat{\bar{Y}}-2_{2}^{\hat{Y}}\right) \equiv \sum_{h^{\prime}} \frac{a_{h}}{n_{h}-1}\left\{\frac{\left(n_{h}-1\right) 1_{n}^{2}}{\hat{n}^{2}}+\frac{\left(n_{n} n^{2}-1\right) 2^{s_{n}^{2}}}{2^{N^{2}}}+\frac{1^{n} h_{h}\left(n_{h}-1_{n}\right)}{n_{h}} .\right. \\
& \frac{\left(1_{h}^{\hat{Y}}-1^{\hat{Y}}\right)^{2}}{\hat{N}^{2}}+\frac{2^{n} h\left(n_{h}-2^{n}\right)}{n_{h}} \frac{\left(2^{y_{h}}-2^{\bar{y}}\right)^{2}}{\hat{N}^{2}}+2 \frac{1^{n} h^{2} 2^{n} h}{n_{h}} . \\
& \left.\frac{\left(1_{1} \bar{Y}_{h}-\frac{\hat{\bar{Y}}}{1}\right)\left(2_{2} \bar{Y}_{h}-2^{\hat{\bar{Y}}}\right)}{1^{N} 2^{N}}\right\} \tag{1.24}
\end{align*}
$$

$$
\begin{aligned}
& v(\hat{P}) \doteq \hat{Y}^{-2} \sum_{h} \frac{a_{h}}{n^{-1}}\left\{\left(1-\frac{i^{\hat{Y}}}{\hat{Y}}\right)^{2}\left(i^{2} h-1\right) i^{2} s^{2}\right. \\
& +\frac{i^{Y^{2}}}{\hat{Y}^{2}}\left[\left(n_{h}-1\right) s_{h}^{2}-\left(i_{h}-1\right) i^{S_{n}^{2}}\right]
\end{aligned}
$$

It is clear from the above formulae that the "spelling out" for domain estimation could be tedious and might lead to lengthy formulae.

In a recent Minimum Wage Survey in Manitoba, a stratified simple random sampling design was adopted: firms were first stratified by region and industry and simple random samples from each cell were selected and interviewed in three consecutive quarters. Whenever the number falling in a cell is small (say $\leq 6$ ), all the firms were enumerated, so this is a case of disproportionate allocation. Here we might be interested in comparing the large and small firms (within an industry) for the effect of minimum wage on employment.

Another interesting application is in surveys with "deaths" of elements. For instance, in the Minimum Wage Survey, some firms were not in existence at the time of interview and in addition there was non-response. Suppose for stratum $h$ (which is a cell in the two-way table) $d_{h}=$ number of deaths, $r_{h}=$ number not responding in the sample of size $m_{h}$ (say). If we assume that respondents and non-respondents are similar, the formulae (1.21)
and (1.23) are applicable, provided we use ${ }_{1} n_{h}=m_{h}-d_{h}-r_{h}, n_{h}=m_{h}-r_{h}$.

## 2. Multistage Stratified Sampling

Tin and Toe (1972) provide an application involving multistage
sampling. The following sampling design was adopted for a Consumer Expenditure Survey in Burma: $M=190$ villages were classified into $L=3$ strata on the basis of location such that $M_{1}=116, M_{2}=39$ and $M_{3}=35$. A $6 \%$ sample of villages (i.e., $m_{h} / M_{h}=0.06$ ) consisting of $m_{1}=7, m_{2}=2$ and $\mathfrak{m}_{3}=2$ villages within strata were selected by simple random sampling (srs). All households in the selected villages were substratified by main occupations into $T=5$ substrata (some substrata might be empty in some villages). Out of $N_{h j}$ households in the $j^{\text {th }}$ village of $h^{\text {th }}$ stratum, $N_{h j k}$ belong to $k^{\text {th }}$ substratum and $n_{h j k}$ out of $N_{h j k}$ were selected again by srs, where $n_{h j k} / N_{h j k}=$ 0.2. The selected households numbering $n^{\prime}=291$ were provided with account books for entering daily expenditure and were visited every month for twelve months to check and collect monthly data. Only six households did not respond $\left(n=n^{\prime}-6=291-6=285\right.$ ). Suppose we take households of size five as our domain of interest, then ${ }_{i} n=56$ and the monthly expenditure per capita for households of size five is estimated by ${ }_{i} \bar{y}={ }_{i} y /{ }_{i} n=1760 / 56=31.4$ since the design is self-weighting. For this sampling design, we have

$$
v(\bar{i} \bar{y})=A+B
$$

where $A$, the variation due to first stage sampling, is given by:

$$
A=\frac{1-a}{1^{n^{2}}} \sum_{h=1}^{L} \frac{m_{h}}{m_{h}-1}\left\{\sum_{j=1}^{m_{h}}\left({ }_{i} y_{h j}-i^{n} h_{i} \bar{y}\right)^{2}-m_{h}\left(\bar{y}_{h}-\frac{i^{n} h}{m_{h}} i^{\bar{y}}\right)^{2}\right\}
$$

and $B$, the variation due to second stage sampling, is given by:

$$
\begin{aligned}
B= & \frac{a(1-b)}{1^{n^{2}}} \sum_{h=1}^{I} \sum_{j=1}^{m} \sum_{k=1}^{T} \frac{n_{h j k}}{n_{h j k}-1}\left\{\left(i_{h j k}-1\right){ }_{i} s_{h j k}^{2}\right. \\
& \left.+\frac{i^{n} h j k}{n_{h j k}}\left(n_{h j k}-{ }_{i} n_{h j k}\right)\left\langle_{i} \bar{y}_{h j k}-{ }_{i} \bar{Y}\right)^{2}\right\}
\end{aligned}
$$

where $a=m_{h} / M_{h}, b=n_{h j k} / N_{h j k}$, and the rest of the notation is selfexplanatory.

For the Burmese data,

$$
A=0.6352, \quad B=0.0837
$$

and

$$
v\left({ }_{i} \bar{y}\right)=0.6352+0.0837=0.7189 .
$$

If sampling fraction a and second stage variation $B$ are neglected (equivalent to drawing villages with replacement), then $v\left({ }_{i} \bar{y}\right)=0.6757$ which is quite close to the correct value 0.7189 .

## 3. Interpenetrating Sub-samples

Suppose $K$ intexpenetrating equal sub-samples are selected using some sampling design and $\hat{Y}_{1}, \ldots, \hat{Y}_{K}$ denote the $K$ estimates of $Y$. Then

$$
\begin{equation*}
\hat{Y}=\frac{1}{K} \hat{K}_{1}^{K} \hat{Y}_{j} \tag{1.26}
\end{equation*}
$$

and

$$
\begin{equation*}
V(\hat{Y})=\frac{1}{K(K-1)} \sum_{1}^{K}\left(\hat{Y}_{j}-\hat{Y}\right)^{2} . \tag{1.27}
\end{equation*}
$$

(1)

We obtain the estimates of domain total from the $K$ sub-samples separately: $\hat{i}^{Y_{1}}{ }^{\prime} ._{i} \hat{Y}_{K}$ and then

$$
\begin{equation*}
\hat{i}=\frac{1}{K} \sum_{j=1}^{K} \hat{i}^{K} \tag{1.28}
\end{equation*}
$$

and

$$
\begin{equation*}
v\left(_{i} \hat{Y}\right)=\frac{1}{K(K-1)} \sum_{j=1}^{K} \zeta_{i} \hat{Y}_{j}-\hat{i}_{i} \hat{Y}^{2} \tag{1.29}
\end{equation*}
$$

For instance, in a two-stage design if primaries are drawn p.p.s. with replacement and equal work loads, say $\bar{m}$, in each selected primary, then

$$
\begin{equation*}
\hat{X}=\frac{N}{\overline{\mathrm{mK}}} i_{i}^{Y} \tag{1.30}
\end{equation*}
$$

and

$$
\begin{equation*}
v(\hat{Y})=\frac{N^{2}}{K(K-1)} \sum_{t=1}^{K}\left(\frac{i^{Y} t}{\bar{m}}-\frac{i^{Y}}{N}\right)^{2} \tag{1.31}
\end{equation*}
$$

where ${ }_{i} Y$ is sample domain total and ${ }_{i} y_{t}$ is $t^{\text {th }}$ sub-sample domain total. Similarly, for estimating domain means, if $\hat{N}_{i}, \ldots,{ }_{i} \hat{N}_{K}$ denote the estimates of ${ }_{i} N$ from the $K$ sub-samples separately, $\hat{i} N=\sum_{j=1}^{K} \hat{i}^{N}{ }_{j} / K$,

$$
\begin{equation*}
\hat{i}^{\hat{Y}}=\sum_{j=1}^{K} \hat{i}_{j}^{Y} \hat{j}_{j=1}^{K} \hat{i}_{j} \hat{N}_{j} \tag{1.32}
\end{equation*}
$$

and

$$
\begin{equation*}
v\left({ }_{i} \hat{\bar{Y}}\right)=\frac{1}{\hat{i}^{2}} \frac{1}{K(K-1)} \sum_{j=1}^{K}\left(\hat{X}_{j}-\hat{X}_{i} \hat{X}_{j}\right)^{2} \tag{1.33}
\end{equation*}
$$

REFERENCES (Section 1.1)
[1] Durbin, J. (1958) "Sampling Theory for Estimates Based on Fewer Individuals than the Number Selected". Bull. I.S.I., 36, 113-9
[2] Hartley, H.O. (1959) "Analytical Studies of Survey Data", in a volume in honour of Corrado Gini, Institute of Statistics, Rome (also reprint No. 63, Iowa State University, Ames, Iowa).
[3] Kish, L. (1965) Survey Sampling. Wiley: New York.
[4] Tin, M. and Toe, T. (1972) "Estimation for Domains in Multistage Sampling", J.A.S.A., 67, 913-6

## 2. Design of Analytical Surveys

### 2.1 Domains Identifiable in Advance

We first consider the comparison of strata, assuming that domains are strata: i.e. we assume that sampling can be done independently in each sub-group. We list several problems and their solutions. A difficulty in designing for analytical surveys is that a great variety of analytical studies are usually made.
(1) Simplest situation is stratified s.r.s. and 'objective function' is the average variance of the differences $\bar{Y}_{h}-\bar{y}_{k}$ between all $L(L-1) / 2$ pairs of strata, i.e.

$$
\begin{align*}
& \bar{v}=\{L(L-1) / 2\}^{-1} \sum_{h<k}^{\sum} V\left(\bar{y}_{h}-\bar{y}_{k}\right)  \tag{2.1}\\
& =1 \\
& =\frac{2}{L} \sum_{1}^{L} \frac{S_{h}^{2}}{n_{h}}-\frac{2}{L} \sum_{1}^{L} \frac{S_{h}^{2}}{N_{h}} \tag{2.2}
\end{align*}
$$

where, for stratum $h, n_{h}=$ no. elements in the sample, $s_{h}^{2}=$ mean square.

Suppose the cost function is $C=\Sigma c_{h}{ }^{n} h$. Minimization of $\bar{V}$ subject to fixed C leads to

$$
\begin{equation*}
n_{h}=\frac{\left(S_{h} / \sqrt{c_{h}}\right) c}{\Sigma S_{h} \sqrt{c_{h}}} \tag{2.3}
\end{equation*}
$$

which is different from the customary Neyman allocation for estimating $\bar{Y}$ :

$$
\begin{equation*}
n_{h}^{*}=\frac{\left(W_{h} S_{h} / \sqrt{c_{h}}\right) c}{\sum W_{h} S_{h} \sqrt{c_{h}}} \tag{2.4}
\end{equation*}
$$

unless $W_{h}=W$, where $W_{h}=N_{h} / N$ and $N_{h}=$ no. of elements in stratum $h$.
(2) Allocation for the criterion $\beta, V\left(\bar{y}_{l}\right)=\ldots=\beta \mathrm{V}\left(\bar{y}_{k}\right)=\phi$

Suppose $C$ is fixed and we wish to find an allocation $n^{\prime}=\left(n_{1}, \ldots, n_{L}\right)$ which makes the strata variances $V\left(\bar{y}_{h}\right)$ as nearly as possible in a given ratio $\frac{1}{\beta_{l}}=\ldots=\frac{1}{\beta_{L}}$ where $\beta_{h}$ is a prespecified constant representing the "importance" attached to stratum $h$, i.e. We wish $\phi_{h}=\beta_{h} v\left(\bar{y}_{h}\right)$ to be as nearly equal as possible, i.e.

$$
\begin{equation*}
\frac{\beta_{h} S_{h}^{2}}{n_{h}}=\phi+\frac{\beta_{h} S_{h}^{2}}{N_{h}} \tag{2.5}
\end{equation*}
$$

using $\sum c_{h} n_{h}=C$ in (2.5), we get

$$
\begin{align*}
& c=\Sigma\left\{\frac{\beta_{h} s_{h}^{2} c_{h}}{\phi+\frac{\beta_{h} s_{h}^{2}}{N_{h}}}\right\} \\
&=\Sigma\left\{\frac{N_{h} c_{h}}{N_{h} \phi}\right.  \tag{2.6}\\
& 1+\frac{B_{h} s_{h}^{2}}{}
\end{align*}
$$

The range of $\phi$ is limited by

$$
\begin{equation*}
0 \leq \phi \leq \min \left\{\beta_{h} S_{h}^{2}\left(1-\frac{1}{N_{h}}\right)\right\} \tag{2.7}
\end{equation*}
$$

From the graph of $C$ versus $\phi$, we read off $\phi=\phi^{*}$ say, given a fixed $C$. We then get the nonintegral allocation

$$
\begin{equation*}
n_{h}^{*}=\frac{N_{h}}{1+\left(\phi^{*} N_{h}\right) /\left(\beta_{h} S_{h}^{2}\right)} \tag{2.8}
\end{equation*}
$$

We round the $n_{h}^{*}$ to the nearest integer, say $\tilde{n}_{h}$, in the range $1 \leq \tilde{n}_{h} \leq N_{h}$. If $C=\sum_{C_{h}} \tilde{n}_{h}=\tilde{C}$, then ( $\tilde{n}_{1}, \ldots, \tilde{n}_{h}$ ) is a feasible allocation; if $\tilde{C} \leq C(\tilde{C}>C)$ we have to increase (decrease) some of the $n_{h}$ to obtain a feasible allocation. To accomplish this, put $\phi_{h}=\beta_{h} S_{h}^{2}\left(\frac{1}{\tilde{n}_{h}}-\frac{1}{N_{h}}\right)$ and increase (decrease) the $\tilde{n}_{h}$ by distributing (withdrawing) units, corresponding to $c-\tilde{C}$, among (from) those strata where $\phi_{h}$ is relatively high (low). This allocation often gives a $\phi$ which is quite close to the minimum attainable $\phi_{0}$. If the $n_{h}$ are small, an integer programming algorithm is needed to get the optimal solution (Chaddha et al. 1971).

Example: Suppose $\mathrm{L}=5, \mathrm{~N}_{\mathrm{h}}=4,33,20,55,56, \mathrm{~S}_{\mathrm{h}}^{2}=1$ for all h , $\beta_{h}=8,1,1,4,8, c_{h}=1$ for all $h$. The graph of $C$ versus $\phi$ for $0 \leq \phi \leq 0.40$ gives $\phi^{*}=0.27$ when $C=\Sigma \Omega_{h}=41$ is fixed. We get $\tilde{\mathrm{n}}_{1}=4, \tilde{\mathrm{n}}_{2}=4, \tilde{\mathrm{n}}_{3}=3, \tilde{\mathrm{n}}_{4}=11, \tilde{\mathrm{n}}_{5}=19$ and $\phi=0.30$, compared to optimal $\phi_{0}=0.29$.
(3) Consider again the stratified s.r.s. case but now suppose we are interested in optimal design for both strata means and overall population mean. In the Labour Force Survey we might wish to find an allocation ( $n_{1}, \ldots, n_{n}$ ) which minimises the cost subject to the following conditions:
$v\left(\bar{y}_{h}\right) \leq v_{h o}$ and $V\left(\bar{y}_{s t}\right) \leq v_{o}$ where $\bar{y}_{s t}=\sum W_{h} \bar{y}_{h}$ and $v_{h o}, v_{o}$ are prescribed gauges. The problem, therefore, is:

$$
\begin{align*}
& \text { Minimize } c=\sum c_{h} n_{h} \\
& \text { subject to } v\left(\bar{y}_{s t}\right)=\sum w_{h}^{2}\left(1-\frac{n_{h}}{N_{h}}\right) \frac{s_{h}^{2}}{n_{h}} \leq v_{o}  \tag{2.9}\\
& v\left(\bar{y}_{h}\right)=\left(1-\frac{n_{h}}{N_{h}}\right) \frac{s_{h}^{2}}{n_{h}} \leq v_{h o} \quad(h=1, \ldots, L \\
& \text { and } 1 \leq n_{h} \leq N_{h} .
\end{align*}
$$

Letting $r_{h}=\frac{1}{n_{h}}-\frac{1}{N_{h}}$, this may be reformulated as:

$$
\begin{align*}
& \text { minimise } C=\sum \frac{N_{h} C_{h}}{1+N_{h} r_{h}} \\
& \text { subject to } \sum\left(W_{h}^{2} S_{h}^{2}\right) r_{h} \leq v_{0}  \tag{2.10}\\
& \text { and } S_{h}^{2} r_{h} \leq v_{h o} \\
& \quad 0 \leq r_{h} \leq 1-\frac{1}{N_{h}}, h=1, \ldots . L
\end{align*}
$$

This is a convex programming problem because the 'objective' function C is a separate convex function of the $r_{h}$ and the inequality constraints are linear in $r_{h}$. Hartley and Hocking (1963) have given an efficient convex programming algorithm and Huddleston et al (1970) use it for sample survey problems. Due to the special nature of the inequality constraints in (2.10), an explicit solution can be obtained simply along the lines of the solution to problem (2) of section 2.2 (see p. 23).

Often times, we might be interested in more than one factor, in analogy with factorial experiements. Keyfitz (1953) investigated the effect of six factors on the family size, using a sample of 1056 Roman Catholic, French-speaking farm families of French origin, living in Quebec. Each factor has two levels: present age of wife (45-54, 55-74), age of wife at marriage (15-19, 20-24), wife's year of schooling ( $0-6,7^{+}$), farm income (high, low), distance to city (near, far), type of area (French, mixed). Keyfitz estimated for each factor the true difference between the two levels for the variable 'family size'.

Consider two factors $\alpha$ and $\tau$ each having two categories, represented by a $2 \times 2$ table with the obvious notation: $N_{i j}{ }^{\prime} n_{i j}, S_{i j}{ }^{\prime} \bar{Y}_{i j}{ }^{\prime} \bar{Y}_{i j}$ and the marginals $N_{i .}, N_{j}, n_{i .} N_{j}(i, j=1,2)$. The two levels for each factor may be compared by considering

$$
\begin{align*}
& \hat{D}_{\alpha}=\frac{N}{N}\left(\bar{y}_{11}-\bar{Y}_{21}\right)+\frac{N \cdot 2}{N}\left(\bar{y}_{12}-\bar{Y}_{22}\right)  \tag{2.11}\\
& \hat{D}_{\tau}=\frac{N_{1}}{N}\left(\bar{y}_{11}-\bar{Y}_{12}\right)+\frac{N_{2}}{N}\left(\bar{Y}_{21}-\bar{Y}_{22}\right) .
\end{align*}
$$

These estimators may be called 'proportionally weighted estimators' and provide overall comparisons.

Again, one might formulate the allocation problem as a convex programming problem:

$$
\begin{align*}
& \begin{aligned}
\text { minimise } & \sum \sum c_{i j} n_{i j} . \\
& i j
\end{aligned} \\
& \text { subject to } V\left(\hat{D}_{\alpha}\right)=\sum_{i} \sum_{j} \frac{N^{2} \cdot j}{N^{2}} S_{i j}^{2}\left(\frac{1}{n_{i j}}-\frac{1}{N_{i j}}\right) \leq V_{\alpha} \tag{2.12}
\end{align*}
$$

(1)

$$
\begin{aligned}
& V\left(\hat{D}_{\tau}\right)=\sum_{i} \sum_{j} \frac{N_{i}^{2}}{N^{2}} S_{i j}^{2}\left(\frac{1}{n_{i j}}-\frac{1}{N_{i j}}\right) \leq V_{\tau} \\
& 1 \leq n_{i j} \leq N_{i j}, i, j=1,2
\end{aligned}
$$

where $c_{i j}$ is cost/element in (i, $\left.j\right)$ th cell and $V_{a}, V_{\tau}$ are prescribed gauges. Booth and Sedransk's (1969) empirical investigation indicates that the solution to (2.12) is essentially the same as the solution to the following simpler problem:

$$
\text { minimise } \sum \sum c_{i j} n_{i j}
$$

subject to

$$
\begin{equation*}
W_{1} V\left(\hat{D}_{\alpha}\right)+\left(1-W_{1}\right) V\left(\hat{D}_{\tau}\right)=2 \frac{V_{\alpha} V_{\tau}}{V_{\alpha}+V_{\tau}}=V^{*} \text { (say) } \tag{2.13}
\end{equation*}
$$

and

$$
1 \leq n_{i j} \leq N_{i j}
$$

where $W_{1}=V_{\tau} /\left(V_{\alpha}+V_{\tau}\right)$. However, they considered only the case $V_{\alpha}=V_{\tau}$.

The solution of (2.13) is

$$
\begin{equation*}
n_{i j}^{*}=\left\{\frac{S_{i j}^{2}}{c_{i j}}\right\}^{1 / 2}\left[W_{1} N_{\cdot j}^{2}+\left(1-W_{1}\right) N_{i .}^{2}\right]^{1 / 2} \frac{\left\{\sum \sum \sigma_{i j}\left[c_{i j}\left(W_{1} N^{2} \cdot j+\left(1-W_{1}\right) N_{i}^{2}\right)\right]^{1 / 2}\right\}}{N^{2} V^{*}} \tag{2.14}
\end{equation*}
$$

provided we ignore the constraints $1 \leq n_{i j} \leq N_{i j}$. However, $n_{i j}^{*}$ usually satisfies $\quad 1 \leq n_{i j} \leq N_{i j}$.
(5) Comparison of domains with stratified s.r.s.

Suppose we have I domains and L strata represented by a $L \times I$ table
with $N_{h i}=$ No. of elements in $(h, i)-t h$ cell, $h=1, \ldots, L ; i=1, \ldots, I$. Then, if we assume the availability of frame in each cell, the estimator of $i$ th domain mean is

$$
\begin{gather*}
\hat{\bar{Y}}_{. i}=\frac{\sum_{h K_{h i}} \bar{Y}_{h i}}{N_{. i}}
\end{gather*}
$$

and

$$
\begin{equation*}
\frac{2}{I(I-1)} \sum_{\substack{I<i \\=1}}^{\sum_{n}} V\left(\frac{\hat{\bar{Y}}}{} . i-\frac{\hat{\bar{Y}}}{. i^{\prime}}\right)=\frac{2}{I} \sum_{h i N_{i}^{2}}^{N_{h i}^{2}}\left(\frac{1}{n_{h i}}-\frac{1}{N_{h i}} S_{h i}^{2}\right. \tag{2.16}
\end{equation*}
$$

which reduces to (2.2) when $I=1$ in (2.16) and $L$ in (2.2) is changed to $I$. Minimisation of (2.16) for fixed $C=\sum_{h} \sum_{i} n_{h i} c_{h i}$ leads to

$$
n_{h i}=C\left[\left(N_{h i} / N_{. i}\right) s_{h i} / \sqrt{C_{h i}}\right] /\left[\begin{array}{cc}
\sum_{h i} & \left.\sum_{h i}\left(N_{i i}\right) s_{h i} \sqrt{C_{h i}}\right] . \tag{2.17}
\end{array}\right.
$$

### 2.2 Domains not Identifiable in Advance

If domains are not identifiable in advance, two-phase or double sampling is needed. We illustrate the methods for problems (1), (4) and (5). Double sampling is appropriate only if the cost of identifying an element is small compared to the cost of obtaining $y$ - values. For example, it will be useful when sampling from files containing information about the domain to which each population element belongs. If the identification process involves visiting each person in home (say), then double sampling is impractical.
(1)

Suppose a large sample $s\left(n^{\prime}\right)$ of size $n^{\prime}$ is selected by s.r.s. and $i^{n \prime}$ of these belong to $D_{i}$, so that $i^{n}$ " is an $r . v$. We need now a sampling rule which specifies how subsampling is to be done within each domain. Suppose $i^{n}\left(s_{i} n^{\prime}\right)$ elements from $s\left(_{i} n^{\prime \prime}\right)$ are enumerated for $y$-values and ${ }_{i} \bar{y}$ denotes the mean based on these $i^{n}$ elements. Suppose the cost function is $C=c^{\prime} n^{\prime}+\Sigma_{i} C_{i} n^{\prime}$ where $c^{\prime}=$ cost of identifying per element and ${ }_{i} c^{=}$cost of getting $y$ in $D_{i}$ per element. Then

$$
\begin{aligned}
V\left({ }_{i} \bar{Y}-j_{j} \bar{Y}\right) & =E V\left({ }_{i} \bar{Y}-j^{\bar{Y}} \mid i^{n,} j^{n}\right)+V E\left({ }_{i} \bar{Y}-j^{\bar{Y}} \mid i^{n,} j^{n}\right) \\
& =E\left\{\frac{i^{S^{2}}}{i^{n}}+\frac{j^{S^{2}}}{j^{n}}\right\}=\left(\frac{i^{S^{2}}}{i^{N}}+\frac{j^{S^{2}}}{j^{N}}\right)
\end{aligned}
$$

Therefore the average variance of the differences ${ }_{i} \bar{Y}-{ }_{j} \bar{Y}$ is

$$
\bar{V}=\frac{2}{I} E\left\{\begin{array}{l}
I  \tag{2.18}\\
\sum \frac{i^{2}}{i^{n}}
\end{array}\right\}-\frac{2}{I} \frac{I}{i} \frac{i^{N}}{i^{N}}
$$

Given that $i^{\prime \prime}$ elements from $s\left(n^{\prime}\right)$ fall in $D_{i}$. Sedransk (1965) proposed to minimize the conditioned average variance $\sum_{i} s^{2} / i^{n} w . r . t$. $i^{n}$ and subject to $\sum_{i} c_{i} n=C-c^{\prime} n^{\prime}$ and $i^{n} \leq i^{n \prime}, i=1, \ldots, I$, assuming that $n^{\prime}$ is so large that $\operatorname{Pr}\left(i^{\prime}{ }^{\prime} \geq 1\right)=1$ for all $i$. The resulting sampling rule is quite complicated. especially when $I$ is not small. Sedransk proposed various approximation when $I=2$ or 3 . Since the overall average variance $\bar{V}$ is a function of $n^{\prime}$, we could select that value of $n$ ' which minimizes $\bar{V}$. The implementation of Sedransk's method for more complex situations (like problems (4) and (5)) is quite formidable, so we consider an alternative scheme, due to J.N.K. Rao (1973), which attempts to circumvent these difficulties.

Essentially, Rao's method employs $i^{n}={ }_{i} n^{\prime \prime}{ }_{i}{ }^{\prime}$, where $i^{v}$ is a constant such that $0<{ }_{i} v \leq 1$, with determination of optimal $n^{\prime}$ and $i v$. Now

$$
\begin{equation*}
\bar{V}=\frac{2}{I} \sum \frac{i^{S^{2}}}{i^{v}} E\left(\frac{1}{i^{\prime}}\right) \doteq \frac{2}{I} \sum \frac{i^{S^{2}}}{n^{\prime}\left({ }_{i} W_{i} v\right)} \tag{2.19}
\end{equation*}
$$

provided $n^{\prime}$ is sufficiently large to justify the approximation $E\left(1 / i^{\prime \prime}\right) \doteq$ $1 / E\left({ }_{i} n^{\prime}\right)$. Here ${ }_{i} W={ }_{i} N / N$. The cost $C$, however, is random for this scheme so we use the expected cost

$$
\begin{equation*}
C^{*}=E(C)=c^{\prime} n^{\prime}+n^{\prime} \Sigma_{i} W_{i} c_{i} \nu \tag{2.20}
\end{equation*}
$$

for determining the optimal $n$ ' and $i v$. This is satisfactory provided the $C . V$. of $C$ is small, which is the case when $n^{\prime}$ is large, since $C \cdot V \cdot(C) \propto 1 / \sqrt{n}$. So the problem reduces to minimising (2.19) subject to (2.20) and the inequality constraints $0<{ }_{i} \nu \leq 1, i=1, \ldots$ I. We will give the explicit solution later, employing a lemma in convex programming. This method retains its simplicity even when $I$ is large. The optimal $i^{\nu}$ (as well as $n^{\prime}$ ) depend on the weights $i W$. If the weights are not accurately known and/or if random costs are undesirable, one could use the $n$ ' obtained from the above method, draw the sample $s\left(n^{\prime}\right)$ and then minimize conditional average variance, for given $i^{n^{\prime}}$, subject to $\Sigma_{i} c_{i} n=C-c^{\prime} n^{\prime}$ and $i^{n} \leq i^{n}{ }^{\prime}$. The resulting $i^{n}$, of course, do not depend on ${ }_{i} W$. Note, however, that the determination of $n$ ' by Sedransk's method also requires a knowledge of the ${ }_{i} W$. We shall also present the explicit solution for the $i^{n}(p, 23)$.

Turning to problem (4), we distinguish two cases: (a) marginals $\mathbb{N}_{i}$.
and $N_{, j}$ unknown; (b) marginals $N_{i}$. and $N_{. j}$ known. In the case (a), the estimators are given by

$$
\begin{aligned}
& \hat{D}_{\alpha}^{\prime}=\frac{n^{\prime}}{n^{\prime}}\left(\bar{y}_{11}-\bar{y}_{21}\right)+\frac{n^{\prime} \cdot 2}{n^{\prime}}\left(\bar{y}_{12}-\bar{y}_{22}\right) \\
& \hat{D}_{\tau}^{\prime}=\frac{n^{\prime}}{n^{\prime}}\left(\bar{y}_{11}-\bar{y}_{12}\right)+\frac{n_{2}}{n^{\prime}}\left(\bar{y}_{21}-\bar{y}_{22}\right)
\end{aligned}
$$

where $n_{i}^{\prime}=\sum_{j} n_{i j}^{\prime \prime} n_{j j}^{\prime}=\sum_{i} n_{i j}^{\prime}$ and $n_{i j}^{\prime}=$ no. of elements from $s\left(n^{\prime \prime}\right)$ falling in (i $j$ )-th cell and $\bar{y}_{i j}$ is the mean based on sub-sample of size $n_{i j}$. If we take $n_{i j}=n_{i j}^{\prime} \nu_{i j} \quad 0 \leq \nu_{i j} \leq 1$, then with equal precisions for $D_{\alpha}^{\prime}$ and $D_{\tau}^{\prime}$, we minimize


$$
\begin{aligned}
\bar{v}^{\prime} & =\frac{1}{2} \quad\left\{V\left(\hat{D}_{\alpha}^{\prime}\right)+V\left(\hat{D}_{\tau}^{\prime}\right)\right\} \\
& =\frac{1}{2}\left\{\sum_{i j} \sum_{i} \frac{\tilde{g}_{i j} S_{i j}^{2} v_{i j}}{2}+\frac{A}{n^{\prime}}\right\}
\end{aligned}
$$

subject to

$$
C^{*}=n^{\prime} c^{\prime}+\sum \sum n^{\prime} c_{i j} W_{i j} \nu_{i j}
$$

and
where

$$
0<v_{i j} \leq 1 \quad, i, j=1,2
$$

$$
\begin{aligned}
w_{i j} & =N_{i j} / N_{i} \\
\tilde{g}_{i j} & =\frac{w^{2} \cdot j+w_{i}^{2}}{w_{i j}}
\end{aligned}
$$

$$
A=\Sigma W_{j}\left(\delta_{j}-\bar{\delta}\right)^{2}+\Sigma W_{i .}\left(\gamma_{i}-\bar{\gamma}\right)^{2}
$$

$$
\delta_{j}=\bar{Y}_{i j}-\bar{Y}_{2 j^{\prime}} \quad \bar{\delta}=\Sigma W_{. j} \delta_{j} \quad \gamma_{i}=\bar{Y}_{i 1}-\bar{Y}_{i 2^{\prime}} \quad \bar{\gamma}=W_{i .} Y_{i} .
$$

We will present the solution of (2.21) after considering problem (5).
(3)

Suppose a large sample $s\left(n_{h}^{\prime}\right)$ is drawn by s.r.s. from $h^{\text {th }}$ stratum and $i^{n} n_{h}^{\prime}$ of these belong to $D_{i}$. We select ${ }_{i} n_{h}={ }_{i} n_{h i} \nu_{h}$ units from $\left.s_{i} n_{h}^{\prime}\right)$ and observe $Y$-values, where $0<i \nu_{h} \leq 1$. The expected cost is

$$
\begin{equation*}
C^{*}=\sum c_{h}^{\prime} n_{h}^{\prime}+\sum \sum{ }_{i} c_{h} E\left(i_{h} n_{h}=\sum c_{h}^{\prime} n_{h}^{\prime}+\sum \sum{ }_{i} c_{h} W_{h} V_{h} v_{h}^{\prime}\right. \tag{2.22}
\end{equation*}
$$

To estimate the domain mean $\bar{Y}=\sum_{h}{ }_{i} N_{h}{ }_{i} \bar{Y}_{h} / \sum_{h}{ }_{i} N_{h}$, we eftploy the combined ratio estimator

$$
i_{i}^{\hat{Y}}=\sum_{h} \hat{N}_{h} \hat{X}_{h} / \sum_{h}{ }_{i} \hat{N}_{h}
$$

where ${ }_{i} \hat{N}_{h}=N_{h}\left({ }_{i} n_{h}^{\prime} / n_{h}^{\prime}\right)$. Noting that ${ }_{i} \hat{\bar{Y}}-{ }_{i} \bar{Y} \equiv \sum_{h}{ }_{i} \hat{N}_{h}\left({ }_{i} \bar{Y}_{h}-{ }_{i} \bar{Y}\right) /{ }_{i} N$, we find that $\bar{V}$ is of the form

$$
\begin{equation*}
\bar{V}=\frac{2}{I} \sum_{h=1}^{L} \sum_{i=1}^{I}\left\{\frac{i_{h}}{n_{h}^{\prime} i_{h}}+\frac{B_{h}}{n_{h}^{\prime}}\right\}+\text { terms not involving } n_{h}^{\prime \prime} \text { and } i_{h} \tag{2.23}
\end{equation*}
$$

where $i_{i}{ }_{h}$ and $B_{h}$ are population parameters (assumed known). Therefore, the problem is to minimize (2.23) subject to (2.22) and the constraints $0<{ }_{i} \nu_{h} \leq 1$.

Solutions: The determination of optimal $\nu_{i j}$ for problem (4) and of the optimal $U_{h}$ for problem (5) can be formulated as follows:

$$
\begin{align*}
& \text { Minimize } V=\frac{T}{I} \frac{A_{\tau}}{n^{\prime} v_{\tau}}+\frac{B}{n^{\prime}} \\
& \text { Subject to } n^{\prime} c^{\prime}+\Sigma n^{\prime} v_{\tau} c_{\tau} W_{\tau}=C^{*}  \tag{2.24}\\
& \text { and } o<v_{\tau} \leq I, \tau=I, \ldots, T
\end{align*}
$$

where $C^{*}, A_{\tau}, B, C^{\prime}, C_{\tau}$ and $W_{\tau}$ are known constants. Without going into details, we present the solution: If we minimize $V$ subject to the equality constraint by Cauchy inequality or Lagrange multiplier method, we get

$$
\sqrt{\frac{B}{n^{\prime}}}=\lambda \sqrt{n^{\prime} c^{\prime}}
$$

and

$$
\sqrt{\frac{A_{\tau}}{n^{\prime} v_{\tau}}}=\lambda \sqrt{n^{\prime} \nu_{\tau} c_{\tau} W_{\tau}}
$$

or

$$
\nu_{\tau}=\sqrt{\frac{A_{\tau}}{B}} \sqrt{\frac{c^{\prime}}{c_{\tau}^{W} W_{\tau}}}
$$

Suppose (1) denotes the index $\tau$ for which $\sqrt{\frac{A_{\tau}}{C_{\tau}^{W}}}$ is the largest, (2) the second largest, and so on. Then if $\nu_{(1)} \leq 1$, the optimal $\nu_{\tau}$ are given by (2.25). If one or more of the $\nu_{\tau}$ given by (2.25) is $>1$, we set $v_{(1)}=1$ and repeat the procedure with the remaining $\nu_{h}$. If again one or more of the remaining $v_{\tau}$ is $>1$, we set the largest equal to 1 and repeat the procedure with the remaining $\nu_{\tau}$ until all $\nu_{\tau} \leq 1$. Note that when two or more of the $\nu_{\tau}$ given by (2.25) (or in later steps) are $>1$, we should not set all of them equal to 1 , since this procedure is not optimal. The above procedure can also be written explicitly as follows:

$$
\text { Let } D_{(0)}=\infty, D_{(r)}=\frac{A_{(r)}}{W_{(r)^{C}}(r)} \quad\left(c^{\prime}+\sum_{1}^{r} c_{(i)} W_{(i)}\right)-\sum_{1}^{r} A_{(i)}, r=1, \ldots, I,
$$

$\left.{ }^{\left(D_{(1)}\right)}{ }^{<D_{(L-1)}} \cdots<D_{(0)}\right)$, then if for $r=1, \ldots, L, D_{(r)}<B \leq D_{(r-1)}$,

$$
v_{(i)}=1, i=1, \ldots, r-1(r=2, \ldots, L)
$$

and $\quad v_{(i)}=\frac{\sqrt{A}(i)}{\sqrt{C_{(i)}{ }^{W}(i)}}\left\{\frac{c^{\prime}+c_{(1)^{W}(1)}+\ldots+c_{(r-1)^{W}(r-1)}}{B+A(1)+\ldots+A_{(r-1)}^{1 / 2}}\right\}^{1}, i=r, \ldots, L$
and if $B>D_{(L-1)}$ then $\nu_{(1)}=1$. We start with the interval $\left(D_{(1)}, D_{(0)}\right)$ and stop the procedure as soon as $B$ is in the interval under consideration. In practice, it is unlikely that we need to compute D's beyond $D(3)$ or $D(4)$. For problem (5), once the fractions $\nu_{h}$ are obtained by the above procedure (which do not depend on $h_{h}^{\prime}$ ), we minimize (2.23) subject to (2.22) w.r.t. $n_{h}^{\prime}$ after substituting for $i h^{\text {. This, of course, reduces to standard Neyman allocation }}$ and we get

$$
\begin{equation*}
n_{h}^{\prime}=\frac{c^{*}}{\sum B_{h} \tilde{c}_{h}}\left(\sqrt{\frac{\tilde{B}_{h}}{c_{h}}}\right) \tag{2.26}
\end{equation*}
$$

where $\tilde{B}_{h}=B_{h}+\sum_{i} \frac{A_{i h}}{i \nu_{h}}$ and $\tilde{c}_{h}=c_{h}^{\prime}+\sum_{i} i^{c}{ }_{h} i^{W} W_{h}{ }^{\nu}{ }_{h}$.

Turning to the problem of minimizing conditional average variance subject to $i^{n} \leq i^{n^{\prime}}$ for all $i$ and $\sum_{i} c_{i} n=C-c^{\prime} n^{\prime}$, it may be formulated as follows:

$$
\begin{align*}
& \text { Minimize } \quad v=\frac{T}{\sum_{1}} \frac{A_{\tau}}{n_{\tau}} \\
& \text { subject to } \sum n_{\tau} c_{\tau}=\tilde{c}  \tag{2.27}\\
& \text { and } n_{\tau} \leq \alpha_{\tau}, \tau=1, \ldots, T
\end{align*}
$$

where $A_{\tau}, C_{\tau}, \tilde{C}$ and $\alpha_{\tau}$ are given constants. Without going into details we present the solution: If $\tilde{\mathrm{c}}>\sum \mathrm{c}_{\tau} \alpha_{\tau}$, then clearly $n_{\tau}=\alpha_{\tau}$. So we assume now that $\sum c_{\tau} \alpha_{\tau} \geq \tilde{c}$. If we minimize $V$ subject to the equality constraint, then

$$
\begin{equation*}
n_{\tau}=\frac{\tilde{c}}{\Sigma \sqrt{A_{\tau} c_{\tau}}}\left(\sqrt{\frac{A_{\tau}}{c_{\tau}}}\right), \tau=1, \ldots . T \tag{2.28}
\end{equation*}
$$

Suppose (1) denotes the index for which $\sqrt{\frac{A_{\tau}}{c_{\tau} \alpha^{2}}}$ is the largest, (2) the second largest and so on. Then if: $n(1) \leq \alpha_{(1)}$. the optimal $n_{\tau}$ are given by (2.28). If one or more of the $n_{\tau}$ given by (2.28) is $>1$, we set all of them -1 and repeat the procedure with the remaining $n_{\tau}$ until all the $n_{\tau} \leq \alpha_{\tau}$. In this problem, the above procedure and the earlier one (viz. setting only the largest $n_{\tau}$ equal to $\alpha_{\tau}$ ) both lead to the optimal solution. Again, we can wirte an explicit solution as follows:

$$
\text { Let } \tilde{c}_{(-1)}=0, \tilde{c}_{(0)}=\left(\Sigma \sqrt{A_{\tau} c_{\tau}}\right) \alpha_{(1)} \sqrt{\frac{c^{\prime}(1)}{A_{(1)}}}
$$

$$
\begin{aligned}
& \text { and } \tilde{c}_{(T-1)}=\sum c_{h} \alpha_{h} \text {. } \\
& \text { Then if } \tilde{C}_{(r-1)}<\tilde{C} \leq \tilde{C}_{(r)}, r=1,2, \ldots .(T-1) \\
& n_{(\tau)}=q_{(\tau)}, \tau=1, \ldots, r \\
& =\frac{\tilde{C}-c^{c}(1)^{\alpha}(1) \cdots-c^{c}(r)^{\alpha}(r)}{\Sigma \sqrt{A_{\tau}{ }^{c} \tau}-\sqrt{A}(1)^{c}(1) \cdots-\sqrt{A}(r)^{c}(r)} \sqrt{\frac{A^{A}(\tau)}{c^{\prime}(\tau)}}, \tau=r+1, \ldots, T \\
& \text { if } \tilde{\mathrm{C}}_{(-1)}<\tilde{\mathrm{C}} \leq \tilde{\mathrm{C}}_{(0)} \\
& { }^{n}(\tau)=\frac{\tilde{c}}{\sum \sqrt{A_{\tau}{ }^{c} \tau}}\left(\sqrt{\frac{A}{A^{C}(\tau)}}\right), \quad \tau=1, \ldots . T \text {. }
\end{aligned}
$$

We start with the interval $\left(\tilde{C}_{(-1)}, \tilde{C}_{(0)}\right)$ and stop the procedure as soon as $\tilde{C}$ is in the interval under consideration.

Finally, we turn to the solution of problem (1) which may be formulated as follows:

$$
\left.\begin{array}{l}
\text { Minimize } V=\sum_{1}^{T} \frac{A_{\tau}}{n^{\prime} \nu_{\tau} W_{\tau}}  \tag{2.29}\\
\text { subject to } C^{*}=c^{\prime} n^{\prime}+\sum c_{\tau} W_{\tau}\left(n^{\prime} \nu_{\tau}\right) \\
\text { and } 0<\nu_{\tau} \leq 1, \tau=1, \ldots, T
\end{array}\right\}
$$

where $A_{\tau}, C^{*}, C_{T}$ and $W_{\tau}$ are given constants. The optimal solution is given by J.N.K. Rao (1973). We determine the optimal $\nu_{\tau}$ for a given $n^{\prime}$ and then the optimal $n$ '. For a given $n$ ', using Cauchy's inequality, we get

$$
n^{\prime} W_{\tau} \nu_{\tau}=\sqrt{\frac{A_{\tau}}{C_{\tau}}} \frac{\left(C^{\star}-n^{\prime} c^{\prime}\right)}{\sum \sqrt{A_{\tau} c^{\prime}}} \quad(\tau=1, \ldots, L)
$$

provided $n^{\prime} W_{\tau} \nu_{\tau} \leq n^{\prime} W_{\tau}$ or

$$
n^{\prime} \geq\left\{c^{\prime}+W(1) \sqrt{\frac{A}{C^{\prime}(1)}}\left(\sum \sqrt{A_{\tau} c_{\tau}}\right)\right\}^{-1} \quad C^{*}=m_{1}^{\prime} \text { (say) }
$$

where (l) denotes the index $\tau$ with the smallest value of $W_{\tau} \sqrt{c_{\tau} / A_{\tau}}$. The minimum value of $V$ for $n^{\prime} \geq m_{1}^{\prime}$, after substituting the optimal $v_{\tau}$ into (2.29), is

$$
V_{1}\left(n^{\prime}\right)=\frac{\left(\Sigma \sqrt{A_{\tau} c_{\tau}}\right)^{2}}{C^{*}-n^{\prime} c^{\prime}}
$$

so that the minimum occurs at the value $m_{1}=m_{1}^{\prime}$. Note that $v_{(1)}=1$ when $n^{\prime}=m_{1}$. We consider next the values of $n^{\prime} \leq m_{1}$. Since $v_{(1)} \geq 1$ for these values, we set $v_{(1)}=1$ and reallocate the remaining $\nu_{\tau}$ again by Cauchy's inequality. This gives

$$
\begin{equation*}
n^{\prime} W_{\tau} \nu_{\tau}=\sqrt{\frac{A_{\tau}}{c_{\tau}}}\left\{\left(c-n^{\prime} c^{\prime}-n^{\prime} c_{(1)} W_{(1)}\right) / \sum_{(1)} \sqrt{A_{\tau} c_{\tau}}\right\}, \quad \tau \neq \tag{1}
\end{equation*}
$$

provided

$$
n^{\prime} \geq\left\{c^{\prime}+c_{(1)} W_{(1)}+\left(W_{(2)} \sqrt{\frac{c_{(2)}}{A_{(2)}}} \sum_{(1)} \sqrt{A_{\tau} c_{\tau}}\right\}^{-1} c^{*}=m_{2}^{\prime}\right. \text { (say) }
$$

where $\Sigma$ denotes summation over $\tau \neq(1)$ and (2) denotes the index $\tau$ with (1) second smallest values of $W_{\tau} \sqrt{C_{\tau} / A_{\tau}}$. Therefore, the minimum value of $V$, for $n^{\prime}$ in the range $m_{2}^{\prime} \leq n^{\prime} \leq m_{1}^{\prime}$, is given by

$$
V_{2}\left(n^{\prime}\right)=\frac{{ }^{A}(1)}{n^{\prime} W(1)}+\frac{\left(\sum \sqrt{A_{\tau} c_{\tau}}\right)^{2}}{c^{*}-n^{\prime}\left(c^{\prime}+c_{(1)} W_{(1)}\right)}
$$

We need to examine the derivation of $v_{2}\left(n^{\prime}\right)$ over the range $m_{2} \leq n^{\prime} \leq m_{1}^{\prime}$ to find the optimal $n^{\prime}$. The derivative vanishes at
and the derivative is $\lesseqgtr$ for $n^{\prime} \lesseqgtr \tilde{m}_{2}$. Consequently, if $\tilde{m}_{2}>m_{1}^{\prime}, v_{2}\left(n^{\prime}\right)$ monotonically decreases as $n^{\prime}$ increases so that the minimum occurs at $m_{2}=m_{1}$; note that $\tilde{m}_{2}$ is usually $>m_{2}^{\prime}$ since $c^{\prime} \ll c_{\tau}$. If, however, $\tilde{m}_{2}$ lies in ( $m_{2}^{\prime}, m_{1}^{\prime}$ ), the true optimum will often be given by $m_{2}=\tilde{m}_{2}$ and the procedure may be terminated here, since in practice, $V\left(n^{\prime}\right)$ will have a unique minimum.

The general procedure is now clear. If $m_{2} \neq \tilde{m}_{2}$, we set $\nu_{(1)}=1$, $\nu(2)=1$ and reallocate the remaining $\nu_{\tau}$. All in all, $T$ steps will be involved if the derivative is nonvanishing over the $T-1$ ranges $n^{\prime} \geq m_{1}^{\prime}, m_{\tau}^{\prime} \leq n^{\prime} \leq m_{\tau-1}^{\prime}$ ( $\tau=2, \ldots, T-1$ ) where

$$
\begin{equation*}
m_{\tau}^{\prime}=\left\{c^{\prime}+\sum_{1}^{\tau-1} c_{(k)} W_{(k)}+\frac{W_{(\tau)} \sqrt{c}(\tau)}{\sqrt{A}(\tau)} \underset{(\tau-1)}{\left.\sum_{j} \sqrt{A_{j} c_{j}}\right\}^{-1} c^{*}, ~ . ~}\right. \tag{2.30}
\end{equation*}
$$

(k) denoting the index $\tau$ with smallest value of $W_{\tau} \sqrt{A}{ }_{\tau} / C_{\tau}$ and $\sum_{(\tau-1)}$ denotes summation over $j$ excluding (1), ..., $(\tau-1)(j-1, \ldots, L)$. The derivative of $V_{\tau}\left(n^{\prime}\right)$ in the range $m_{\tau}^{\prime} \leq n^{\prime} \leq m_{\tau-1}^{\prime}$ vanishes at
and usually $\tilde{m}_{\tau}>m_{\tau}^{\prime}$. Denoting the optimal $n^{\prime}$ in the range $m_{\tau}^{\prime} \leq n^{\prime} \leq m_{\tau-1}^{\prime}$ by $m_{\tau}$, we compare the values $V_{\tau}\left(m_{\tau}\right)$ to find the true optimum $n^{\prime}$ and then the corresponding $\nu_{\tau}$ 's, where


$$
\begin{equation*}
V_{\tau}\left(n^{\prime}\right)=\sum_{1}^{\tau-1} \frac{A^{A}(k)}{n^{\prime} W(k)}+\frac{\left(\sum \sqrt{A_{j} C_{j}}\right)^{2}}{c^{*}-n^{\prime}\left(c^{\prime}+{ }_{\sum}^{\tau^{-1}} C_{(k)} W_{(k)}\right)} \tag{2.32}
\end{equation*}
$$

If the optimum occurs in the range $m_{r}^{\prime} \leq n^{\prime} \leq m^{\prime}(r-1)$, then optimal $v_{\tau}$ are given by

$$
\begin{aligned}
& v_{(1)}=\ldots=v_{(r-1)}=1 \\
& n^{\prime} W_{j} v_{j}=\sqrt{\frac{A_{j}}{c_{j}}} \frac{\left(C^{*}-n^{\prime} c^{\prime}-n^{\prime} \sum^{r-1} c_{(k)^{W}(k)}\right.}{\sum_{(r-1)}{\sqrt{A_{\tau} c_{\tau}}}^{1}, j \neq\{(1), \ldots(r-1)\} .}
\end{aligned}
$$

Note that $n^{\prime} \geq c^{*} /\left(c^{\prime}+\Sigma c_{\tau} W_{\tau}\right)=m_{L}^{\prime}$.

Example: Sedransk (1965) considered the comparison of farms of different sizes with the number of cattle expected to be sold, and measure of size, the number of animal units (three hogs equal one head of cattle) on the farm.

Here $T=3, C^{*}=220, C^{\prime}=1, c_{1}=c_{2}=C_{3}=4, W_{1}=W_{2}=0.25, W_{3}=0.50$, $A_{1}=S_{1}^{2}=4, A_{2}=S_{2}^{2}=1, A_{3}=S_{3}^{2}=1$. Using the above method, the following values are obtained: $m_{1}=m_{1}^{\prime}=73, \frac{3}{2} V_{1}\left(m_{1}\right)=0.436 ; m_{2}^{\prime}=55, m_{2}=\tilde{m}_{2}=64$ and $\frac{3}{2} V_{2}\left(m_{2}\right)=0.424$. One could stop here since $m_{2}=\tilde{m}_{2}$, but let us proceed to the final (third) step: $m_{3}^{\prime}=44, \tilde{m}_{3}=58 \quad m_{2}^{\prime}$ so $m_{3}=m_{2}^{\prime}=55$ and $\frac{3}{2} v_{3}\left(m_{3}\right)$ $=0.436$. Therefore, the optimal $n^{\prime}=m_{2}=64$ and $v_{(1)}=v_{1}=1, v_{(2)}=v_{2}=$ 0.72 and $v_{3}=0.36$. Sedransk's method leads to $n^{\prime}=64$ also. Here the C.V. of $C$ is about $4 \%$ only.

## REFERENCES

[1] CHADDHA, R.L. et al (1971). "Allocation of total sample size when only the stratum means are of interest", Technometrics, Vol. 13, 817-31.
[2] HARTLEY, H.O. \& HOCKING, R. (1963). "Convex programming by tangential approximation", Management Science, Vol. 9, 600-12.

[3] HUDDLESTON, H.F.. CLAYPOOL, P.L. and HOCKING, R.R. (1970). "Optimal sample allocation to strata using convex programming", Applied Statistics. Vol. 19, 273 - 8.
[4] RAO, J.N.K. (1973). "Double sampling for stratification and analytical surveys", Biometrika, Vol. 60, 125-33.
[5] SEDRANSK (1965). "A double sampling scheme for analytical surveys". J.A.S.A., Vol. 60, 985-1004.
[6] KEYFITZ (1953). "A factorial arrangement of comparison of family size". Amer. J. Sociology, Vol. 58, 470-80.
[7] BOOTH, G. \& SEDRANSK, J. (1969). "Planning some two-factor comparative surveys", J.A.S.A., Vol. 64, 560-73.
[8] SEDRANSK, J. (1973). "Design and analysis of analytical surveys", Paper presented at the I.A.S.S. meetings, Vienna.

## 3. Hypothesis Testing

Classical statistical tests of significance or tests of goodness of fit are based on the assumption of simple random sampling from an infinite population. However, survey populations are finite and most survey designs are complex involving clustering, unequal probabilities, etcetra. Only recently, some attention has been given to hypothesis testing from complex survey designs.

### 3.1 Tests of independence in contingency tables from stratified simple random samples

Nathan (1969, 73) considered stratified s.r.s. and developed appropriate tests of independence in contingency tables. Stratification is used only for reducing the variance of estimated total and/or for administrative convenience and is considered to have no intrinsic interest as an additional variable when testing independence between two qualitative variables. That is, the hypothesis of general independence in the population between the two qualitative variables is of interest, irrespective of the stratification employed for sampling.

## Suppose the population is divided into $t$ strata and the elements

 within each stratum classified according to an $r \times s$ table. Let $\mathrm{P}_{\mathrm{ijk}}=\mathrm{N}_{\mathrm{ijj}} / \mathrm{N}$ where $\mathrm{N}_{i j k}=$ no. elements in $k^{\text {th }}$ stratum falling in (i, j)-th cell. Denote the marginal totals as $P_{i j} \ldots P_{i} \ldots P_{j} .{ }_{j} P_{._{k}}$ etcetra where $P_{._{k}}=N_{k} / N$ is the known stratum weight ( $i=1, \ldots, r, j=1, \ldots, s, k=1, \ldots, t$ ). The hypothesis of interest is$$
\begin{equation*}
H_{0}: P_{i j .}=P_{i \ldots} P_{. j .}, \quad i=1, \ldots, r ; j=1, \ldots, s \tag{3.1}
\end{equation*}
$$

The hypothesis $H_{o}$ in general is not equivalent to testing independence within strata:

$$
\begin{equation*}
H_{0}^{\prime}: \frac{P_{i j k}}{P_{\cdots} \cdot}\left(=\frac{N_{i j k}}{N_{k}}\right)=\frac{P_{i \cdot k}}{P_{\ldots k}} \frac{P_{j k}}{P_{\cdot j}}\left(=\frac{N_{i . k} N_{j k}}{N_{k}^{2}}\right) \tag{3.2}
\end{equation*}
$$

$H_{0}=H_{0}^{\prime}$ if and only of

$$
\begin{equation*}
P_{i j k}=\frac{P_{i j .} P_{i . k} P_{i j k}}{P_{i \ldots} P_{j .} P_{\cdot k}} \tag{3.3}
\end{equation*}
$$

The meaning of $(3.3)$ is not clear. If one of the classifications, say that indered by $i$, corresponds to domains, $H_{o}$ is formulated as

$$
\begin{equation*}
H_{0}: \frac{P_{1 j}}{P_{1 \ldots}}=\frac{P_{2 j}}{P_{2 \ldots}}=\ldots=\frac{P_{r j}}{P_{r \ldots}}\left(=P_{. j}\right)(j=1, \ldots, s) \tag{3.4}
\end{equation*}
$$

i.e. the conditional probability of having character $j$ given the domain is constant over all domains of study, i.e. a homogeneity test.

Nathan assumed infinite $N$ or that $n / N$ is negligible so that the likelihood function, based on a stratified s.r.s. ( $n_{1}, \ldots, n_{\tau}$ ) is given by

$$
\begin{equation*}
L\left(P_{i j k}\right)=\frac{t}{\pi}\left\{\frac{n_{k!}}{\Pi n_{i j k}!} \quad \frac{11}{i, j}\left(\frac{P_{i j k}}{P_{i, k}}\right)^{n_{i j k}}\right\} \tag{3.5}
\end{equation*}
$$

where $n_{i j k}=$ no. of elements in the sample from $k$-th stratum falling in (i, j)-th cell. The theory for finite $N$, when $n / N$ is substantial, is not available.

## Likelihood ratio test statistic

Maximizing (3.5) subject to $\sum_{i, j} P_{i j k}=P_{\ldots k}$ and $P_{i j k} \geq 0$ leads to the maximum likelihood estimators (M)

$$
\begin{equation*}
\hat{p}_{i j k}=n_{i j k} \frac{p_{k}}{n_{k}} \tag{3.6}
\end{equation*}
$$

ML estimators under $H_{0}$ and the constraints $\sum_{i, j} P_{i j k}=P_{H_{k}}$ and $P_{i j k} \geq 0$ can be obtained by using Lagrange multipliers $\lambda_{k}$ and $\mu_{i j}$ and the Newton-Raphson method. Starting values of $P_{i j k}, \lambda_{k}$ and $\mu_{i j}$ in the iteration may be taken as (3.6), $\lambda_{k}=0$ and $\mu_{i j}=0$. Four or five iterations might suffice when $r, s$ and $t$ are small. Denote the restricted $M L$ estimators as $\hat{P}_{i j k}$. The likelihood ratio statistic, therefore, is given by

$$
\begin{equation*}
\lambda=\frac{L\left(\hat{\hat{P}}_{i j k}\right)}{L\left(\hat{P}_{i j k}\right)} \tag{3.7}
\end{equation*}
$$

or $G=-2 \ln \lambda=2 \sum_{i} \sum_{j k} \sum_{i j k} \ln \left(\hat{P}_{i j k} / \hat{P}_{i j k}\right)$
is the log-likelihood ratio. Under $H_{O^{\prime}} G$ is asymptotically (as $n \rightarrow \infty$ with r, s, t fixed) distributed as chi-square with ( $r-1$ ) (s-l) d.f. Therefore, $H_{o}$ is rejected if $G>x_{\alpha}^{2}$ where $X_{\alpha}^{2}$ is the upper $\alpha$-point of chi-square distribution with ( $x-1$ ) ( $s-1$ ) d.f.

Other test statistics based on the ML estimators $\hat{P}_{i j k}$ and restricted MU estimators $\hat{\hat{P}}_{i j k}$ are the following:
chi-square statistic:

$$
\begin{equation*}
x^{2}=\sum_{k}\left(\frac{n_{k}}{P_{\ldots k}}\right) \sum_{i, j} \frac{\left[\hat{P}_{i j k}-\hat{\hat{P}}_{i j k}\right]^{2}}{\hat{\hat{r}}_{i j k}} \tag{3.8}
\end{equation*}
$$

chi-one square statistic:

$$
\begin{equation*}
x_{1}^{2}=\sum_{k}\left(\frac{n_{k}}{P_{\ldots k}}\right) \sum_{i, j} \frac{\left[\hat{P}_{i j k}-\hat{\hat{P}}_{i j k}\right]^{2}}{P_{i j k}} \tag{3.9}
\end{equation*}
$$

All three statistics $G, X^{2}$ and $X_{1}^{2}$ have the same asymptotic distribution under $H_{0}$ and also same asymptotic non-null distribution. These statistics, however, are computationally combersome as $\hat{\mathrm{P}}_{\mathrm{ijk}}$ involve iterative calculations. Two other statistics, which are computationally simpler, are given below.

First, one could consider the restricted ML estimators after the first iteration only, say $\tilde{P}_{i j k}$, and replace ${\underset{\mathrm{P}}{i j k}}$ by $\tilde{\mathrm{P}}_{\mathrm{ij} k}$ in (3.9):

$$
\begin{equation*}
\tilde{x}_{1}^{2}=\sum_{k}\left(\frac{n_{k}}{P_{0} \cdot k}\right) \sum_{i, j} \frac{\left[\hat{P}_{i j k}-\tilde{P}_{i j k}\right]^{2}}{\hat{P}_{i j k}} \tag{3.10}
\end{equation*}
$$

The estimators $\tilde{P}_{i j k}$ are equivalent to those obtained by minimising $a x_{1}^{2}$ statistic:

$$
\begin{align*}
x_{1}^{2} & =\sum_{i j k} \sum_{j} \frac{\left(n_{i j k}-n_{k} \frac{P_{i j k}}{P_{\ldots j}}\right)^{2}}{n_{i j k}} \\
& =\sum_{k} \frac{n_{k}}{P_{k}} \sum_{i, j} \frac{\left(P_{i j k}-\hat{P}_{i j k}\right)^{2}}{P_{i j k}} \tag{3.11}
\end{align*}
$$

subject to a linearised $H_{0}$ and $\sum_{k} P_{i j k}=P \ldots k$.

A large sample statistic based on asymptotic normality of estimators of $H_{0}$ has also been considered. The hypothesis $H_{0}$ (given by (3.4)) may be written as

$$
\begin{equation*}
\mathrm{H}_{0}: \underset{\sim}{q}{ }^{\prime} \underset{\sim}{r}=\underset{\sim}{n} \tag{3.12}
\end{equation*}
$$

where

$$
\begin{aligned}
& {\underset{\sim}{q}}^{\prime}=\left(q_{11}, \ldots, q_{r, s-1}\right), q_{i j}-p_{i j} / p_{i \ldots} \text { and } \\
& \Gamma \quad=\left({ }_{\sim}^{Y} 11, \ldots,{\underset{\sim}{Y}}^{Y}, 1, s-1\right) \text { is } r(s-1) \times(r-1)(s-1) \text { matrix } \\
& \gamma_{i j}^{\prime}=\left(y_{11}^{(i j)}, \ldots, \gamma_{r, s-1}^{(i, j)}\right), i=1, \ldots, r-1 ; j=1, \ldots, s-1 \\
& \gamma_{i^{\prime} j^{\prime}}^{(i j)}= \begin{cases}1 & \text { if } j^{\prime}=j, \quad i^{\prime}=i \\
-1 & \text { if } j^{\prime}=j, \\
i^{\prime}=i+1, & i^{\prime}=1, \ldots, r \\
0 & \text { otherwise }\end{cases}
\end{aligned}
$$

where

The test statistic is

$$
\begin{equation*}
G_{1}=\left(\hat{q}^{\prime} \underset{\sim}{\Gamma}\right)\left(\Gamma_{\sim}^{\prime} \underset{\sim}{\underset{\sim}{\Gamma}} \underset{\sim}{r}\right)^{-1}\left(\hat{q}^{\prime} \Gamma\right)^{\prime} \tag{3.13}
\end{equation*}
$$

where

$$
\hat{q}_{i}{ }^{\prime j}=\frac{\hat{P}_{i^{\prime} j}}{\hat{p}_{i \ldots}}, \hat{\Sigma} \text { is a consistent estimator of the variance- }
$$

covariance matrix of $\underset{\sim}{q}$. Again, the statistics (3.10) and (3.13) have the same asymptotic null and non-null distributions as $G, X^{2}$ and $X_{1}^{2}$. One advantage of (3.13) is it can incorporate finite population correction factor in comoutina $\underset{\sim}{\underset{\sim}{r}}$. A Monte Carlo study has shown that. for finite $n$. the differences between the powers of these five statistics are auite small.

Example: Consider the data in Table 1 obtained from a (stratified) sample survey undertaken in a Canadian Maritime Province in 1952. Responses are classified by groups of 'occupational disadvantage' $(x=2)$, by incidence of

psychiatric disorders ( $r=3$ ) within $t=3$ geographic - and social-area strata: $n_{1}=92, n_{2}=112, n_{3}=78, n=282$.

$$
\text { Table 1: } n_{i j k} \text { - values }
$$

|  | 1 | 2 | P. ${ }_{k}$ |
| :---: | :---: | :---: | :---: |
| $k=1:\left\{\begin{array}{l}1 \\ 2 \\ 3\end{array}\right.$ | 10 | 18 |  |
|  | 6 | 46 | 0.17 |
|  | 2 | 10 |  |
| $\mathrm{k}=2: \quad$ | 17 | 18 |  |
|  | 29 | 31 | 0.35 |
|  | 8 | 9 |  |
| $k=3:\left\{\begin{array}{l}2 \\ 3\end{array}\right.$ | 4 | 18 |  |
|  | 9 | 30 | 0.48 |
|  | 8 | 9 |  |

If we test $H_{0}^{\prime}$ (within strata) using the $G-$ statistic, we get $G_{1}=6.449$, $G_{2}=0.011, G_{3}=4.363$. Comparing these values with $X_{0.05}^{2}(2 \mathrm{~d} . f)=5.991$, we see that the hypothesis of independence within strata is rejected for stratum 1 but not in strata 2 and 3. Pooled statistic $G_{1}+G_{2}+G_{3}=10.823$ is also not significant compared to $X_{0.05}^{2}$ (6d.f). Turning to $H_{0}$, we have the following:
(1)

Statistic
(3.7)
(3.10)
(3.13)


Value of Test
Statistic
1.929
where $\hat{\hat{q}}^{\prime}=\hat{\sim}^{\prime}-\hat{\sim}^{\prime} \underline{\sim}^{\prime} \underset{\sim}{\left(\Gamma^{\prime}\right.} \underset{\sim}{\sum} \tilde{\sim}^{-1} \Gamma_{\sim}^{\prime} \underset{\sim}{\underset{\sim}{\Sigma}}$. It is seen that $H_{0}$ is not rejected by any of the three methods considered above, since $X_{0.05}^{2}(2 \mathrm{d.f})=$.5.991 .

In the case of proportional allocation (i.e. $n_{k}=n$ P ..k ,

$$
\hat{P}_{i j}=\frac{n_{i j}}{n}=\sum_{k} p \ldots k\left(\frac{n_{i j k}}{n_{k}}\right)
$$

are unbiased estimates of $\mathrm{P}_{\mathrm{ij}}$. . Therefore, classical methods applied to overall frequencies $n_{i j}$. might be appropriate. The classical log-likelihood ratio will be

$$
\begin{align*}
G^{*}= & 2\left[\sum_{i j} \sum_{i j} n_{i} \ln \left(n_{i j}\right)=\sum_{i} n_{i .} \ln \left(n_{i .}\right)\right. \\
& \left.-\sum_{j} n_{\cdot j \cdot} \ln \left(n_{. j}\right)+n \ln (n)\right] \tag{3.14}
\end{align*}
$$

or chi-square statistic

$$
\begin{equation*}
x^{\star^{2}}=\sum_{i} \sum_{j}\left(n_{i j}-\frac{n_{i} .^{n} j_{0}}{n}\right)^{2} /\left(\frac{n_{i} n_{i}{ }^{n}}{n}\right) \tag{3.15}
\end{equation*}
$$

In fact, the asymptotic distribution of $G *$ or $X^{*}{ }^{2}$ under $H_{0}$ is chi-square with (r-1) (s-1) d.f.. i.e. same as before. However, their asymptotic power is not greater than that of the detailed test statistics. A Monte Carlo study has indicated that for finite samples or large samples, differences in power
(
are quite small so that over-all test statistics might be preferred under proportional sampling.

The test statistics we have considezed aro bused on the assurrtion of s.r.s. within each stratum. No theoretical results are available when cluster sampling and/or unequal probability sampling is used within strata: The method of balanced repeated replication (BRR) has been used to construct test statistics under stratified cluster sampling (Nathan, 1973). We will give this application after describing the method of $B R R$.

Brackstone and Gosselin (1973) consider the problem of testing that undercoverage rate in the 1971 Census is evenly distributed with respect to categories of interest, i.e., a test of homogeneity. If $M_{i}=$ no. missed persons in category $i, E_{i}=$ no. persons enumerated in category $i$, then
$H_{0}: \frac{M_{1}}{M_{1}+E}=\ldots=\frac{M_{r}}{M_{r}+E_{r}}=\frac{M}{M+E}$ or equivalently $H_{0}: M_{1}=M \frac{E_{1}}{E}, \ldots, M_{r}=M \frac{E_{r}}{E}$ where $M=\Sigma M_{i}, E=\Sigma E_{i}$ and the quantities $M, E$ and $E_{i}$ are known.

Based on a stratified s.r.s. of missed persons, we estimate $P_{i .}=\frac{M_{i}}{M}$ by $\hat{P}_{i .}$ and the large-sample statistic $\left(\underset{\sim}{P}-{\underset{\sim}{P}}_{0}\right)^{\prime} \hat{\Sigma}_{\sim}^{-1} \hat{\sim}\left(\underset{\sim}{p}-P_{0}\right)$ is assymptotically $X^{2}$ with $k-1 d f$ under $H_{0}$, where $\hat{P}_{\sim}$, is the vector of $\hat{P}_{i}$. 's, ${ }_{\sim}^{P}$ is the vector of $P_{10}=E_{i} / E$ and ${\underset{\sim}{N}}^{\hat{N}}$ is the estimated variance-covariance matrix of $\hat{P}_{i}$ 's.

## REFERENCES

[1] NATHAN, G. (1969) "Tests of independence in contingency tables from stratified samples", in New Developments in Survey Sampling, Wiley, 578-600.
[2] NATHAN, G. (1973) "Tests of independence in contingency tables from complex samples", Pater presented at the IASS Meeting, Vienna.
[3] GARZA-HERNANDEZ, T., and MCCARTHY, P.J. (1962) "A test of homogeneity for a stratified sample", Proc. Soc. Statist. Sec., A.S.A., 200-2.
[4] BRACKSTONE, G.J., and GOSSELIN, J.F. (1973) "1971 Census Evaluation Porgramme, Reverse Record Check; Analysis of missed persons: Method of Analysis and preliminary results", Statistics Canada Report.
4. Balanced Repeated Replication (BRR)

Since the efficiency of an estimator generally increases with the number of strata for the same sample size, quite often the primary units are stratified to the extent that two primary units are selected from each stratum. One could, of course, stratify to the extent that one primary per stratum is selected, but difficulties with variance estimation arise. If the two primaries within a stratum are selected with replacement, there will be only two independent replicates available for the estimation of variance and confidence intervals for the population parameter (based on the two replicates) will then be much wider than they need to be. One cannot always estimate the variance by first estimating within strata and then pooling over strata. For instance, to estimate the ratio $T=Y / X$ where $Y$ and $X$ are unknown, one cannot write $T$ as $T=\sum_{1}^{L}\left(X_{h} / X\right)\left(Y_{h} / X_{h}\right)$ and then estimate $Y_{h} / X_{h}$ separately from the data in stratum $h$, since $X_{h} / X$ is unknown. Moreover, even if $X_{h} / X$ is known, 'separate ratio estimate' of $T$ might lead to large bias when $L$ is large, unlike a 'combined ratio estimate' of $T$. It may also be noted that the variance estimate based on the two independent replicates refers to the average of replicate estimates rather than to the estimate prepared for the entire sample. The two estimates will not be the same in general except in the linear case. BRR is a method to overcome these difficulties and it often leads to stable variance estimates and facilitates statistical inference from complex survey data.

To fix the ideas, consider the estimation of $\bar{Y}$ from a stratified s.r.s. design with two elements per stratum. Assume first that the elements are selected with replacement. Suppose $Y_{h 1}$ and $y_{h 2}$ denote the observaticns from stratum $h(h=1 \ldots, L)$, then $\bar{Y}$ is estimated by $\bar{Y}_{s t}=\Sigma W_{h} \bar{Y}_{h}$ where $W_{h}=N_{h} / N$ and $\bar{y}_{h}=\left(y_{h 1}+y_{h 2}\right) / 2$. The customary variance estimator, of course, is given by

$$
\begin{equation*}
v\left(\bar{y}_{s t}\right)=(1 / 2) \sum W_{h}^{2} s_{h}^{2}=(1 / 4) \sum W_{h}^{2} d_{h}^{2} \tag{4.1}
\end{equation*}
$$

where $d_{h}=Y_{h 1}-Y_{h 2}$. The variance estimate based on the independent replicates ( $y_{I 1}, \ldots, y_{L 1}$ ) and $y_{12}, \ldots, y_{L 2}$ ) is

$$
\begin{equation*}
v_{R}\left(\bar{y}_{s t}\right)=(1 / 4)\left(\bar{y}_{s t 1}-\bar{y}_{s t 2}\right)^{2} \tag{4.2}
\end{equation*}
$$

where $\bar{y}_{s t 1}=\Sigma W_{h} Y_{h 1}, \bar{y}_{s t 2}=\Sigma W_{h} y_{h 2}$ and $\bar{y}_{s t}=\left(\bar{y}_{s t 1}+y_{s t 2}\right) / 2$. Of course, (4.2) is in general not equal to (4.1) and its stability relative to (4.1) will be very poor as it is based only on two replicates.

Suppose now we form a half-sample replicate by selecting one of $y_{11}, y_{12}$, one of $y_{21}, y_{22}, \ldots$ and one of $y_{L 1}, y_{L 2}$, then the half-sample estimate of $\bar{Y}$ is

$$
\begin{equation*}
\bar{y}_{s}=\sum w_{h}\left(\delta_{h 1} y_{h 1}+\delta_{h 2} y_{h 2}\right) \tag{4.3}
\end{equation*}
$$

where $\delta_{h l}=1$ if element $(h 1)$ is selected for the half-sample, o otherwise and $\delta_{h 2}=1-\delta_{h 1}$. There are in all $2^{L}$ possible half samples for a given sample and if $\bar{Y}_{s, i}$ denotes the estimate for $i{ }^{\text {th }}$ half-sample ( $i=1, \ldots, 2^{L}$ ) then

$$
\begin{equation*}
\frac{1}{2^{L}} \sum_{i=1}^{2^{L}} \bar{y}_{s, i}=\bar{y}_{s t} \tag{4.4}
\end{equation*}
$$

Also

$$
\begin{equation*}
\bar{v}_{s, i}-\bar{v}_{s t}=(1 / 2) \sum_{1}^{L} W_{r} \delta_{h}^{(i)} d_{h} \tag{4.5}
\end{equation*}
$$

where $\delta_{h}^{(i)}=2 \delta_{h 1}-1$ and equals +1 if (hl) is in the $i$ th half sample and is -1 if (h2) is in the $i$ th half sample so that $\frac{1}{2^{L}} \Sigma \delta_{h}^{(i)}=0$. Similarly

$$
\begin{equation*}
\left(\bar{y}_{s, i}-\bar{y}_{s t}\right)^{2}=(1 / 4) \sum w_{h}^{2} d_{h}+(1 / 2) \sum_{h<h} \sum_{h} \delta_{h}^{(i)} \delta_{h}^{(i)} w_{h} w_{h}, d_{h} d_{h} \tag{4.6}
\end{equation*}
$$

since $\left[\delta_{h}^{(i)}\right]^{2}=1$. Since $\frac{l}{L} \sum_{2} \delta_{h}^{(i)} \delta_{h}^{(i)}=0$, it immediately follows that

$$
\begin{equation*}
\frac{1}{2^{L}} \sum_{i}\left(\bar{y}_{s i}-\bar{y}_{s t}\right)^{2}=v\left(\bar{y}_{s t}\right) \tag{4.7}
\end{equation*}
$$

i.e. there is no loss of information if we use all the $2^{L}$ half-sample replicates.

If $2^{L}$ is large (say $L=30$ ), one would naturally wish to select only a fraction of these half-samples. Suppose we select $k$ half-samples from $2^{\text {L }}$ by s.r.s. with replacement and use the variance estimate

$$
\begin{equation*}
v_{k}\left(\bar{y}_{s t}\right)=\sum_{i}^{k}\left(\bar{y}_{s, i}-\bar{y}_{s t}\right)^{2} / k \tag{4.8}
\end{equation*}
$$

Since the conditional expectation, $E_{2}\left\{v_{k}\left(\bar{y}_{s t}\right)\right\}$ over the $2^{\text {Lh }}$ half-samples (for a given sample) is $v\left(y_{s t}\right)$, it immediately follows that

$$
\begin{align*}
v\left[v_{k}\left(\bar{y}_{s t}\right]\right. & =v_{1} E_{2}\left\{v_{k}\left(\bar{y}_{s t}\right)\right\}+E_{1} v_{2}\left\{v_{k}\left(\bar{y}_{s t}\right)\right\} \\
& =v_{1}\left[v\left(\bar{y}_{s t}\right)\right]+E_{1} v_{2}\left\{v_{k}\left(y_{s t}\right)\right\} \\
& \geq v_{1}\left[v\left(\bar{y}_{s t}\right)\right] \tag{4.9}
\end{align*}
$$

where $E_{1}$ and $V_{1}$ respectively, denote the expectation and variance over the main sample and $V_{2}$ is the conditional variance. Consequently, the variance estimator $(4.8)$ is less stable than $v\left(\bar{y}_{s t}\right)$.

The question that naturally arises now is whether it is possible to select a set of $k$ half-samples such that (4.8) is equal to $v\left(\bar{y}_{s t}\right)$, i.e. is it possible to select a set of $k$ half-samples such that

$$
\begin{equation*}
\sum_{i=1}^{k} \delta_{h}^{(i)} \delta_{h^{\prime}}^{(i)}=0 \forall h, h^{\prime}=1, \ldots, I\left(h<h^{\prime}\right), \tag{4.10}
\end{equation*}
$$

in view of (4.6)? This leads us into the method of $B R R$ which provides a set of half-samples satisfying (4.8).

If the two elements within a stratum are selected by s.r.s. without replacement, then we modify (4.3) to

$$
\begin{equation*}
\bar{y}_{s}=\sum W_{h} \sqrt{1-\frac{2}{N_{h}}}\left\{\left(\delta_{h 1} y_{h 1}+\delta_{h 2} y_{h 2}\right)-\bar{y}_{h}\right\}+\bar{y}_{s t} \tag{4.11}
\end{equation*}
$$

because then

$$
\begin{equation*}
\left(\bar{y}_{s, i}-\bar{y}_{s t}\right)^{2}=v\left(\bar{y}_{s t}\right)+(1 / 4) \sum_{h<h} \sum_{h}\left(1-\frac{2}{N_{h}}\right) \delta_{h}^{(i)} \delta_{h}^{(i)} W_{h} W_{h}, d_{h} d_{h} . \tag{4.12}
\end{equation*}
$$

Plackett and Burman (1946) have given a method of constructing $\mathrm{k} \times \mathrm{k}$ orthogonal matrices when k is a multiple of 4 . Suppose $L=5,6,7$ or 8 , then the smallest orthogonal matrix that can be used for this case is $8 \times 8$ because of the multiple-of-4 constraint. We give the $8 \times 8$ matrix below where rows identify half-sample and the columns correspond to strata:

| $\delta_{h}^{(1)}:$ | +1 | -1 | -1 | +1 | -1 | +1 | +1 | -1 |
| :--- | :--- | :--- | :--- | :--- | :--- | :--- | :--- | :--- |
| $\delta_{h}^{(2)}:$ | +1 | +1 | -1 | -1 | +1 | +1 | +1 | -1 |
| $\delta_{h}^{(3)}:$ | +1 | +1 | +1 | -1 | -1 | +1 | -1 | -1 |
| $\delta_{h}^{(4)}:$ | -1 | +1 | +1 | +1 | -1 | -1 | +1 | -1 |
| $\delta_{h}^{(5)}:$ | +1 | -1 | +1 | +1 | +1 | -1 | -1 | -1 |
| $\delta_{h}^{(6)}:$ | -1 | +1 | -1 | +1 | +1 | +1 | -1 | -1 |
| $\delta_{h}^{(7)}:$ | -1 | -1 | +1 | -1 | +1 | +1 | +1 | -1 |
| $\delta_{h}^{(8)}:$ | -1 | -1 | -1 | -1 | -1 | -1 | -1 | -1 |

Any set of 5 columns for $L=5$ case, 6 columns for $L=6$ case, 7 columns for $L=7$ case or all the 8 columns for $L=8$ case defines a set of $k=8$ balanced half-samples with the property (4.10). The average of $k$ values $\bar{Y}_{s, i}$ will be equal to $\bar{y}_{\text {st }}$ when

$$
\begin{equation*}
\frac{1}{k} \sum_{i=1}^{k} \delta_{h}^{(i)}=0, \quad \forall h=1, \ldots . L \tag{4.13}
\end{equation*}
$$

This property is satisfied except for the last column of the orthogonal matrix given above, i.e. except in the case of $L=8$ strata. However, the property (4.10) is always satisfied. Note that the number of half-samples, $k$, equals the smallest integral multiple of 4 which is greater than $L$ if both (4.10) and (4.13) are to be satisfied. We call this 'full orthogonal balance'.

These results easily generalise to multi-stage sampling in which primaries are selected p.p.s. with replacement within strata. The estimator of $Y$ corresponding to (4.3) is

$$
\begin{equation*}
\hat{Y}_{s}=\sum_{1}^{L}\left(\delta_{h 1} \frac{\hat{Y}_{h 1}}{p_{h 1}}+\delta_{h 2} \frac{\hat{Y}_{h 2}}{p_{h 2}}\right) \tag{4.14}
\end{equation*}
$$

where $\hat{Y}_{h i}$ is the unbiased estimator of (hi)-th primary total $Y_{\text {ni }}$ wased on the elements selected from that primary ( $i=1,2$ ) and $P_{h i}$ is the selection probability for (hi)-th primary. If a set of $k$ balanced half-samples are selected, then

$$
\begin{equation*}
\mathrm{v}_{\mathrm{k}}(\hat{Y})=\sum_{\mathrm{l}}^{\mathrm{K}}\left(\hat{Y}_{\mathrm{S}, \mathrm{i}}-\hat{Y}\right)^{2} / \mathrm{k} \tag{4.15}
\end{equation*}
$$

is taken as the estimator of $V(\hat{Y})$ where $\hat{Y}$ is the estimator of $Y$ based on the entire sample:

$$
\begin{equation*}
\hat{Y}=\sum_{1}^{\mathrm{L}}(1 / 2)\left(\frac{\hat{Y}_{h 1}}{\mathrm{p}_{\mathrm{h} 1}}+\frac{\hat{Y}_{\mathrm{h} 2}}{\mathrm{p}_{\mathrm{h} 2}}\right) \tag{4.16}
\end{equation*}
$$

Due to the property (4.10), it follows that (4.3) reduces to

$$
\begin{equation*}
v_{k}(\hat{Y})=1 / 4 \quad \sum_{1}^{L}\left(\hat{Y}_{h 1} / p_{h 1}-\hat{Y}_{h 2} / P_{h 2}\right)^{2}=v(\hat{Y}) \tag{4.17}
\end{equation*}
$$

which is the customary unbiased estimator of $V(\hat{Y})$.

## Application to ratio estimation

Suppose we are interested in estimating $R=Y / X$ from a multistage p.p.s. sample (within strata) where the primaries are selected with replacement. Then the estimator of $R$ based on a half-sample is

$$
\hat{R}_{S}=\frac{\hat{Y}_{S}}{\hat{X}_{S}}
$$

where $\hat{X}_{s}$ is of the form (4.14) with $y$ replaced by $x$. The variance estimator based on $k$ balanced half-samples is taken as

$$
\begin{equation*}
v_{k}\left(\hat{R}_{s}\right)=\frac{1}{k} \sum_{l}^{k}\left(\hat{R}_{s i}-\hat{R}^{k}\right)^{2} \tag{4.19}
\end{equation*}
$$

where $\hat{R}=\hat{Y} / \hat{X}$. A similar procedure is adopted for ratio estimators of $Y$ when $X$ is known. If $L$ is large, we could approximate

$$
\begin{equation*}
\hat{R}_{s i}-\hat{R} \doteq \frac{\hat{Y}_{s i}-\hat{R}^{\hat{X}_{s i}}}{\hat{X}} \tag{4.20}
\end{equation*}
$$

since $\hat{X}_{s i}$ and $\hat{X}$ both converge to $X$. Therefore,

$$
\begin{align*}
v_{k}\left(\hat{R}_{s}\right) & \doteq \frac{1}{\hat{X^{2}}}\left\{\frac{1}{K} \hat{\sum}\left[\left(\hat{Y}_{s i}-\hat{Y}\right)-\hat{R}^{\hat{R}}\left(\hat{X}_{s i}-\hat{X}\right)\right]\right. \\
& =\frac{1}{\hat{X^{2}}}\left\{v(\hat{Y})-2 \hat{R} \operatorname{Cov}(\hat{Y}, \hat{X})+\hat{R}^{2} v(\hat{X})\right\} \tag{4.21}
\end{align*}
$$

in view of (4.10), where

$$
\begin{equation*}
\operatorname{Cov}(\hat{Y}, \hat{X})=(1 / 4) \sum_{1}^{L}\left(\frac{\hat{Y}_{h 1}}{p_{h 1}}-\frac{\hat{Y}_{h 2}}{p_{h 2}}\right)\left(\frac{\hat{X}_{h 1}}{p_{h 1}}-\frac{\hat{X}_{h 2}}{p_{h 2}}\right) \tag{4.22}
\end{equation*}
$$

Consequently, for large $L, v_{k}\left(\hat{R}_{S}\right)$ is approximately equal to the customary estimator, $V(\hat{R})$, of $V(\hat{R})$. If the property ( 4.13 ) also holds, then $\hat{\bar{R}}=\frac{1}{\mathrm{k}} \sum \hat{\mathrm{R}}_{\mathrm{S}, \mathrm{i}} \doteq \hat{\mathrm{R}}$ in view of (4.20) Corresponding to the half-sample estimate $\hat{R}_{S, i}$ we have the estimate $\tilde{R}_{S_{, i}}$ from the complimentary set of data and we get the variance estimate

$$
\begin{equation*}
\tilde{v}_{k}\left(\hat{R}_{s}\right)=\frac{1}{k} \sum_{1}^{k}\left(\tilde{R}_{s, i}-\hat{R}^{2}\right. \tag{4.23}
\end{equation*}
$$

(1)

The complimentary estimators $\tilde{R}_{s, 1}, \ldots, \tilde{R}_{s, k}$ also form a balanced set of half-samples. We could also take average of (4.19) and (4.23):

$$
\begin{equation*}
\bar{v}_{\mathrm{k}}\left(\hat{R}_{\mathrm{s}}\right)=\frac{\hat{v}_{\mathrm{k}}\left(\hat{R}_{\mathrm{S}}\right)+\tilde{v}_{\mathrm{k}}\left(\hat{R}_{\mathrm{S}}\right)}{2} \tag{4.24}
\end{equation*}
$$

Another variance estimator is

$$
v_{k}^{*}\left(\hat{R}_{S}\right)=(1 / 4) \quad \begin{align*}
& \mathrm{k}  \tag{4.25}\\
& 1
\end{align*}\left(\hat{R}_{S, i}-\tilde{R}_{S, i}\right)^{2}
$$

A motivation for $(4.24)$ is that $\hat{\bar{R}}_{S, i}=(1 / 2)\left[\hat{R}_{S, i}+\tilde{R}_{S, i}\right] \doteq \hat{R}$ (equality exact in the linear case) and since $\hat{R}_{S i}$ and $\tilde{R}_{s i}$ are independent, $V\left(\hat{R}_{S, i}\right)$ is estimated unbiasedly by $\left(\hat{R}_{s i}-\tilde{R}_{s i}\right)^{2} / 4$.

No theoretical results are available (even for the case of
stratified s.r.s.) on the finite sample performances of the variance estimators $v_{k}\left(\hat{R}_{S}\right), \tilde{v}_{k}\left(\hat{R}_{S}\right), \bar{v}_{k}\left(\hat{R}_{S}\right), \hat{v}_{k}^{*}\left(\hat{R}_{S}\right)$ and the customary one $v(\hat{R})$. We will, however, briefly describe later the results of some empirical studies.

If primaries are selected with unequal probabilities without replacement and sampling fraction is large, no simple correction similar to (4.11) is available.

Example: Koch and Thompson (1972) give an interesting application of BRR in the comparison of domain means employing two charactexs: height and weight. In a U.S. Health Examination Survey, employing a complex design with two primaries per stratum, the heights, $y_{1}$, and weights, $y_{2}$, of $n=7119$ boys are recorded. The two domains are: $D_{1}=$ negro, six years old, male; $D_{2}=$ white, six years old, male. The estimator of ( $\bar{Y}_{1}{ }_{1}{ }_{i} \bar{Y}_{2}$ ), $i=1,2$ is


$$
\left[\hat{i}_{i \hat{\bar{Y}}_{2}}^{\hat{\bar{Y}}_{2}}\right]=\hat{i \stackrel{\hat{\bar{Y}}}{2}}
$$

where

$$
\begin{equation*}
\hat{\overline{\bar{Y}}}_{j}=\frac{\sum_{i}^{n} w_{t i}{ }^{Y} j t}{\sum_{j}^{n} w_{t i^{a}}}, j=1,2 ; i=1,2 \tag{4.26}
\end{equation*}
$$

where ${ }_{i} y_{j t}=y_{j t}$ if $t^{\text {th }}$ child is in $D_{i}$, $=0$ otherwise and $i_{j t}=1$ if $t^{\text {th }}$ child in $D_{i},=0$ otherwise and $w_{t}$ is the weight attached to $t^{\text {th }}$ child, i.e. the estimator of $Y$ is of the form $\hat{Y}=\sum w_{t} y_{t}$. With $L=20$ strata, we get $k=20$ balanced half-sample estimates

$$
\begin{equation*}
\hat{\bar{Y}}_{j}^{(r)}=\frac{\sum_{i}^{n} w_{t} \delta_{t}^{(r)} i^{y} j t}{\sum_{i} w_{t} \delta_{t}^{(r)} i^{a} t}, r=1, \ldots, k \tag{4.27}
\end{equation*}
$$

where $\delta_{t}^{(r)}=1$ if $t^{\text {th }}$ child is in $r^{\text {th }}$ half-sample; $=0$ otherwise. Denote

$$
\frac{\hat{\bar{Y}}}{\underline{\sim}}(x)=\left[\begin{array}{c}
\hat{\bar{Y}}^{(x)} \\
i^{1}(x) \\
\hat{\bar{Y}}_{2}(x)
\end{array}\right] .
$$

We may therefore estimate the variance-covariance matrix of $\underset{\sim}{\hat{\bar{Y}}}=\left[\begin{array}{c}\underset{\sim}{\hat{\bar{Y}}} \\ \underset{\sim}{\hat{Y}}\end{array}\right]$ by

$$
\begin{equation*}
{\underset{\sim}{D}}(\underset{\sim}{\hat{\bar{Y}}})=\frac{1}{20} \sum_{r=1}^{20}[\underset{\sim}{\hat{\bar{Y}}}(x)-\underset{\sim}{\hat{Y}}]\left[\hat{\bar{Y}}_{\sim}^{\hat{Y}}(x)-\underset{\sim}{\hat{\bar{Y}}}\right] \tag{4.28}
\end{equation*}
$$


Since $n$ is large we could assume multivariate normality of $\underset{\sim}{\hat{Y}} \underset{\sim}{\hat{Y}}$ and any contrast in $i_{j} \bar{Y}_{j}$, say $\underset{\sim}{C}{ }^{\prime}{ }_{\underset{\sim}{Y}}^{\hat{Y}}$, may be estimated by $\underset{\sim}{C}{ }^{\prime} \underset{\sim}{\hat{Y}}$. For example, if $\underset{\sim}{C}{ }^{\prime}=(1,0,-1,0)$ we will have a comparison of height of negro vs. height
of white. We could use the statistic

$$
\begin{equation*}
x^{2}=\frac{\left(\underset{\sim}{C}{ }^{\prime} \underset{\sim}{\underset{\sim}{Y}}\right)^{2}}{{\underset{\sim}{C}}^{\prime}{\underset{\sim}{\sim}}_{k}^{C}} \tag{4.29}
\end{equation*}
$$

as a $X^{2}$ with 1 d.f. If $\underset{\sim}{c}{ }^{\prime}=(1,0,-1,0)$ we get $\underset{\sim}{C^{\prime}}{ }^{\prime} \underset{\sim}{\underset{Y}{y}}=0.58$, $C^{C}{ }^{1}{\underset{\sim}{k}}^{X} \underset{\sim}{C}=0.8168$ and $X^{2}=0.40$. Similarly for the comparison of weights, $x^{2}=0.40$. Both $X^{\prime}$ 's are not significant suggesting that the two domains are similar w.r.t. average height and weight.

One could make a joint test of height and weight by considering

$$
\underset{\sim}{c}=\left[\begin{array}{rrrr}
1 & 0 & -1 & 0 \\
0 & 1 & 0 & -1
\end{array}\right]
$$

and the statistic

$$
\begin{equation*}
X^{2}={\underset{\sim}{Y}}_{\underset{\sim}{\prime}}^{C_{\sim}^{\prime}}\left[{\underset{\sim}{C}}_{\underset{\sim}{C}}^{D_{k}} C^{\prime}\right] \underset{\sim}{C} \underset{\sim}{\hat{Y}} \tag{4.30}
\end{equation*}
$$

which is asymptotically $X^{2}$ with 2 d.f. Here we get $X^{2}=8.70$ which is significant, i.e. differences cannot be explained in terms of any variable separately. Here is an illustration that multivariate tests are more sensitive than separate univariate tests. One reason for this is the data exhibit a 'cross over effect' which arises when direction of differences between the two domains w.r.t. two positively correlated characters is reversed. Here $\operatorname{corr}_{\mathbb{N}}\left(y_{1}, y_{2}\right)=0.89, \operatorname{corr}_{W}\left(Y_{1}, y_{2}\right)=0.77$, but $\hat{\bar{Y}}_{1}-\hat{2}_{2} \hat{\bar{Y}}_{1}=$ 0.58 cm and $1_{2}^{\hat{Y}_{2}}=2_{2}^{\hat{Y}_{2}}=-0.28 \mathrm{~kg}$.

Application to testing in contingency tables from a stratified cluster sample
Suppose two primaries are selected within each stratum by s.r.s. with replacement, and subsampling is done such that we have two independent unbiased estimates of the probabilities $P_{\text {ijh }}$ from the two primaries within stratum $h(L=1, \ldots, L)$, where $P_{i j h}$ is the probability that an element is
in stratum $h$ and is classified into (i, j) ${ }^{\text {th }}$ cell. Denote the two estimates of $P_{i j h}$ in stratum $h$ by $\hat{P}_{i j l}$ and $\hat{P}_{i j 2}$. If we select a set of $k$ balanced half-samples, then the estimate of $P_{i j}$. from $u^{\text {th }}$ half-sample is

$$
\begin{equation*}
\left.\hat{P}_{i j,, u}=\sum_{h}^{\{ } \delta_{h 1} \hat{P}_{i j 1}+\delta_{h 2} \hat{P}_{i j 2}\right\}, u=1, \ldots, k \tag{4.31}
\end{equation*}
$$

where $\delta_{h l}$ is as defined before. Similarly, from the complimentary set corresponding to $u^{\text {th }}$ half-sample we get the independent estimate of $P_{i j}$. given by

$$
\begin{equation*}
\tilde{\mathrm{P}}_{i j ., u}=\sum_{h}\left\{\delta_{h 2} \hat{P}_{i j 1}+\delta_{h 1} \hat{P}_{i j 2}\right\}, u=1, \ldots, k \tag{4.32}
\end{equation*}
$$

Under the null hypothesis $H_{0}: P_{i j}=P_{i .} P_{. j .}$, the following statistics have zero expectation:

$$
\begin{align*}
& \hat{Q}_{i j, u}=\hat{P}_{i j \ldots u}+\tilde{P}_{i j \ldots u}-\hat{P}_{i \ldots, u} \tilde{P}_{. j \ldots u}-\tilde{P}_{i \ldots, u} \hat{P}_{. j \ldots u}  \tag{4.33}\\
& \hat{T}_{i j, u}=\hat{P}_{i j \ldots u} \tilde{P}_{r s \ldots, u}-\hat{P}_{i s \ldots, u} \tilde{P}_{r j \ldots, u} \quad \begin{array}{l}
i=1, \ldots, r-1 \\
j=1, \ldots, s-1 .
\end{array} \tag{4.34}
\end{align*}
$$

Let

$$
\begin{align*}
& \hat{\underline{Q}}_{u}=\left(\hat{Q}_{11, u}, \ldots, \hat{Q}_{r-1, s-1, u}\right)^{\prime}, \hat{\bar{Q}}_{\sim}^{\hat{Q}}=\left(\hat{\bar{Q}}_{11}, \ldots, \bar{Q}_{r-1, s-1}\right)^{\prime} \\
& \hat{T}_{\sim}=\left(\hat{T}_{11, u}, \ldots, \hat{T}_{r-1, s-1, u}\right)^{\prime}, \hat{\bar{T}}_{\sim}^{\hat{T}}=\left(\hat{\bar{T}}_{11} \ldots \hat{\bar{T}}_{r-1, s-1}\right)^{\prime} \tag{4.35}
\end{align*}
$$

where

$$
\begin{equation*}
\hat{\bar{Q}}_{i j}=\frac{1}{k} \sum_{u=1}^{k} \hat{Q}_{i j, u}, \hat{\bar{T}}_{i j}=\frac{1}{k} \sum_{u=1}^{k} \hat{T}_{i j, u} \tag{4.36}
\end{equation*}
$$

Then the statistics

$$
\begin{align*}
& x_{Q}^{2}=\hat{\bar{Q}}^{\prime} \hat{\sim}_{Q}^{-1} \underset{\sim}{\hat{Q}}  \tag{4.37}\\
& x_{T}^{2}=\hat{\bar{T}}_{\sim}^{\prime} \hat{\sim}_{\sim}^{D} \tag{4.38}
\end{align*}
$$

can be used to test $H_{0}$ by comparing them with $X_{0.05}^{2}[(r-1)(s-1)]$, where $\hat{D}_{Q}$ and $\hat{\sim}_{T}$ are the estimated variance-covariance matrices of $\underset{\sim}{\hat{Q}}$ and $\underset{\sim}{\hat{T}}$ respectively; expressions for $\hat{\sim}_{Q}$ and $\hat{D}_{T}$ are given in Nathan (1973a).

In the case of a self-weighting design, $P_{i j}=n_{i j} . / n$ is unbiased for $P_{i j}$. . so one could use the classical statistic given by (3.14). A linked simulation study by Nathan (1973) indicates that the differences in performance of $(4.37)$ and (3.14) are small, so that the use of the simpler overall statistic (3.14) may be justified in the case of self-weighting designs.

One could also use the large-sample statistic (3.13) with an
estimate of $\sum_{\sim}$ based on balanced half-sample estimates $\hat{P}_{i j}, u$ :
est. $\operatorname{cov}\left(\hat{q}_{i^{\prime} j} \prime \prime \hat{q}_{i j}\right)=\frac{1}{k} \sum_{u=1}^{k}\left\{\hat{q}_{i^{\prime} j \prime, u}-\hat{q}_{i \prime j}\right\}\left\{\hat{q}_{i j, u}-\hat{q}_{i j}\right\}$

$$
\begin{aligned}
& i, i^{\prime}=1, \ldots, r \\
& j, j^{\prime}=1, \ldots, s-1
\end{aligned}
$$

where $\hat{q}_{i j}=\frac{P_{i j}}{\hat{P}_{i .}}$ is the estimate of $\frac{P_{i j}}{P_{i .}}$ based on the entire sample and
$\hat{q}_{i j, u}=\frac{\hat{P}_{i j \ldots, u}}{\hat{p}_{i \ldots, u}}$. The use of (3.13) of course, involves the inversion of a large matrix $\underset{\sim}{\Gamma} \hat{\Sigma}_{\sim}^{\Gamma}$.

A computationally simpler statistic is the following:

$$
\begin{equation*}
x_{M}^{2}=\frac{4 \Sigma \sum \sum\left(\hat{\bar{Z}}_{i j}-\hat{\bar{Z}}\right)^{2}}{1-\frac{2}{\bar{\pi}} \sin ^{-1}\left(\frac{1}{1-r s}\right)} \tag{4.40}
\end{equation*}
$$

where $\hat{\bar{Z}}_{i j}=\frac{1}{k} \sum_{u=1}^{k} \operatorname{sg}\left[\hat{P}_{i j \ldots u}-\tilde{\mathcal{P}}_{i \ldots \ldots} \tilde{\mathrm{P}}_{\text {ij..... }}\right], \hat{\bar{Z}}=\frac{1}{r s} \sum_{i j} \hat{\bar{Z}}_{i j}$
and $\operatorname{sg}[]=1$ if [ ] is positive and = 0 if [ ] is negative. If rs is large $l /(l-r s)$ is small and $\sin ^{-1}\left[l /(1-r s] \doteq l /(1-r s) \cdot x_{M}^{2}\right.$ is asymptotically $\mathrm{X}_{(\mathrm{rs}-1)}^{2}$ under some stringent assumptions (McCarthy, 1969). Nathan's (1973a) empirical investigation indicates that (4.37) performs better than (4.40).

## Partially balanced half-samples

If the number of strata, $L$, is large (say $\geq 50$ ), full orthogonal balancing would require $k \geq$ half-samples and the processing of results for such a large number of half-samples might be quite expensive. To reduce the number of half-samples, we could generate a smaller number of partially balanced half-samples as follows: We group the I strata into G groups, each of $L / G$ strata (assume $L / G$ is an integer) and then use a set of $k \geq \frac{L}{G}$ balanced half-samples in each of the groups. Then $\sum_{i} \delta_{h}^{(i)} \delta_{h}^{(i)}=0$ when $L$ and $L$ ' belong to the same group or belong to different groups but do not correspond to the same column in the two balanced sets. Therefore, the variance estimator is

$$
\begin{equation*}
v_{k}^{p}\left(\bar{y}_{s t}\right)=\frac{1}{k} \sum_{1}^{k}\left(\bar{y}_{s, i}-\bar{y}_{s t}\right)^{2}=(1 / 4) \sum w_{h}^{2} d_{h}^{2}+(1 / 2) \sum^{*} w_{h} w_{h}, d_{h} d_{h} \tag{4.42}
\end{equation*}
$$

where $\Sigma^{*}$ denotes summation over all pairs ( $h, h^{\prime}$ ) such that $h<h^{\prime}$, $h$ is from one group of strata and $h^{\prime}$ is from another group of strata and $h$ and $h^{\prime}$ represent corresponding columns from the $L / G$ orthogonal columns that make up a balanced set. The number of terms in $\Sigma^{*}$ is ( $L / G$ ) $\frac{G(G-1)}{2}=\frac{L(G-1)}{2}$ which increases with $G$.

Example: $L=6, k=4$. Here we need $k=8$ to achieve full orthogonal balancing. Since $k=4$, we can divide the six strata into $G=2$ groups each of $L / G=3$ strata and use the first three columns of a $4 \times 4$ orthogonal matrix for each of the 2 groups:

|  | 1 | 2 | 3 | 4 | 5 | 6 |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| $\delta_{h}^{(1)}$ | -1 | -1 | -1 | -1 | -1 | -1 |
| $\delta_{h}^{(2)}$ | -1 | +1 | +1 | -1 | +1 | +1 |
| $\delta_{h}^{(3)}$ | +1 | -1 | +1 | +1 | -1 | +1 |
| $\delta_{h}^{(4)}$ | +1 | +1 | -1 | +1 | +1 | -1 |

Efficiency of partially balanced half-sample variance estimator will depend on $G$ as well as on the arrangement of strata into groups, for a given $G$. In the linear case, if we assume that the $L$ strata are randomly split into $G$ groups each of size $L / G$, then

$$
\begin{equation*}
v\left[v_{k}^{P}\left(\bar{y}_{s t}\right)\right]=v\left[v_{k}\left(\bar{y}_{s t}\right)\right]+\frac{G-1}{L-1} \sum_{\substack{L<h \\ \\=1}}^{\sum} w_{h}^{2} W_{h}^{2}, \sigma_{h y}^{2} \sigma_{h \prime y}^{2} \tag{4.43}
\end{equation*}
$$

which shows that (4.43) increases with $G$. The loss in using $v_{k}^{P}\left(\vec{y}_{s t}\right)$ is minimised when $G-2$. For the special case of $W_{I}^{2} \sigma_{I y}^{2}=\ldots=W_{L}^{2} \sigma_{L y}^{2}$ and $\beta_{1}=\ldots=\beta_{L}=3$ where $\beta_{h}$ is the kurtosis measure in stratum $h$, we have

$$
\begin{equation*}
\frac{v\left[v_{k}^{P}\left(\bar{y}_{s t}\right)\right]}{v\left[v_{k}\left(\bar{y}_{s t}\right)\right]}=G \tag{4.44}
\end{equation*}
$$

so that the loss in precision by using a G-order partially balanced set of samples instead of a balanced set of half-samples might be very large. Of course, the choice of $G$ will depend on a trade-off between precision and computational cost.

Lee $(1972,73)$ has investigated several strata arrangements into G groups and his investigations suggest the following as a 'good' pattern: A A A (alternatate ascending order arrangement)
(1) Arrange the L strata first in ascending order of the magnitude of $a_{h}=$ $W_{h}^{2} \sigma_{h y}^{2}$; (2) Divide the L strata arranged in this order into $G$ groups, each of size L/G; (3) Reverse the order of the L/G strata in each of the second, fourth, sixth, .... groups.

In practice, to implement $A A A$ we will need estimates of $a_{h}$ from past data.

Additional applications of BRR are given in a paper by G.C. Koch, D.H. Freeman and J.L. Freeman "Some useful strategies in the multivariate analysis of data from complex surveys", Proc. Soc. Statist. Sec. (A.S.A.), 1973, 8 - 17. A weighted least squares approach is used to investigate various relationships among domain means and to test relevant hypotheses. Examples include: (1) comparisons among cross-classified domains; (2) evaluation of the existence and nature of trends.

## REFERENCES

[1] Koch, G.C. and Johnson, W.D. (1972) "Application of multivariate analysis to complex sample survey data", JASA, Vol. 67, 780-2.
[2] Lee, K.H. (1972) "The use of partially balanced designs for half replication method of variance estimation", JASA, Vol. 67, 324-34.
[3] Lee, K.H. (1973) "Using partially balanced designs for the half-sample replication method of variance estimation", JASA, Vol. 68, 612-4.
[4] McCarthy, P.J. (1966) Replication, an Approach to the Analysis of Data from Complex Surveys, N.C.H.S. Vital and Health Statistics, Serits 2, No. 14, Washington, D.C.
[5] McCarthy, P.J. (1969) "Pseudo-replication: half samples", Rev. of I.S.I.., Vol. 33, 239-64.
[6] Plackett, R.L. and Burman, P.J. (1946) "The design of optimum multifactorial experiments", Biometrika, 33, 305-25.
[7] Nathan, G. (1973a) Approximate Tests of Independence in Contingency Tables from Complex Stratified Samples. N.C.H.S. Vital and Health Statistics, Series 2, No. 53, Washington, D.C.
[8] Kish, L. and Frankel, M.R. (1970) "Balanced repeated replication for standard errors", JASA, Vol. 65, 1071-94.

## 5. The Jack-knife Method

To fix ideas, consider the case of simple random sampling and ratio estimation. Suppose $\left(y_{i}, x_{i}\right)(i=1, \ldots, n)$ denotes the sample and $\bar{y}$ and $\bar{x}$ denote the sample means. The customary ratio estimator $r=\bar{y} / \bar{x}$ of the population ratio $R=\bar{Y} / \bar{X}$ has bias of order $1 / n$. Suppose we split the sample at random into $g$ groups, each of size $p=n / g$ and compute the customary estimators, $r_{j}^{\prime}$, by omitting $j^{\text {th }}$ group from the sample, $j=1, \ldots$, g, i.e. $r_{j}^{\prime}=\left(n \bar{y}-p \bar{y}_{j}\right) /\left(n \bar{x}-p \bar{x}_{j}\right)$ where $\bar{y}_{j}$ and $\bar{x}_{j}$ are the $j^{\text {th }}$ group means. Let $\bar{r}_{g}^{\prime}=\Sigma_{\underline{r}_{j}^{\prime}} / g$, then the estimator

$$
\begin{equation*}
r_{Q}=\Sigma \tilde{r}_{j} / g=g r-(g-1) \bar{r}_{g}^{\prime} \tag{5.1}
\end{equation*}
$$

will have bias of order $n^{-2}$ at most (ignoring f.p.c.), where

$$
\begin{equation*}
\tilde{r}_{j}=g r-(g-1) r_{j}^{\prime} \tag{5.2}
\end{equation*}
$$

is called a 'pseudo-value'. If f.p.c. is not negligible, we modify (5.1) as

$$
\begin{equation*}
r_{Q}^{*}=w r-(w-1) \bar{r}_{g}^{\prime} \tag{5.3}
\end{equation*}
$$

where $w=g[1-(n-p) / N]$. Bias of $r_{Q}^{*}$ will not contain terms of order $n^{-1}$ as well as of order $\mathrm{N}^{-1}$. Quenouille (1956) proposed this method in the context of bias reduction. Investigation by kao and Rao (1971) incicates that for ratio estimation, $g=\Omega$ is an optimum choice in that both bias and variance of $r_{Q}$ are decreasing functions of $g$ under certain reasonable models:

$$
\begin{gathered}
y_{i}=\alpha+\beta x_{i}+e_{i}, \quad i=1, \ldots, n \\
E\left(e_{i} \mid x_{i}\right)=0, V\left(e_{i} \mid x_{i}\right) \propto x_{i}^{t}, t \geq 0, E\left(e_{i} e_{j} \mid x_{i}, x_{j}\right)=0, i \neq j
\end{gathered}
$$

x is distributed as a gamma $\mathrm{r} . \mathrm{v}$.
Extensive empirical and semi-empirical work (Rao, 1969; Rao and Kuzik, 1974) support this choice.

Tukey proposed that in many instances the pseudo-values $\tilde{r}_{1}, \ldots, \tilde{r}_{n}$ (using $g=n$ ) can be treated as approximately i.i.d. (independent, identically distributed) so that $V\left(r_{Q}\right)$ (or m.s.e.r.) may be estimated by

$$
\begin{align*}
v\left(r_{Q}\right) & =\frac{1}{n(n-1)} \sum_{1}^{n}\left(\tilde{r}_{j}-\tilde{r}_{Q}\right)^{2}  \tag{5.5}\\
& =\frac{(n-1)}{n} \sum_{1}^{n}\left(r_{j}^{\prime}-\bar{r}_{g}^{\prime \prime}\right)^{2} \tag{5.6}
\end{align*}
$$

If f.p.c. is not negligible, we multiply (5.5) or (5.6) by
(1 $-n / N$ ). One could also use

$$
\begin{equation*}
v_{Q}(r)=\frac{1}{n(n-1)} \sum_{1}^{n}\left(\tilde{r}_{j}-r\right)^{2} \tag{5,7}
\end{equation*}
$$

For large $n, v\left(r_{Q}\right)$ is approximately equal to the customary variance estimator

$$
\begin{equation*}
v(r)=\frac{1}{n(n-1) x^{-2}}\left\{\sum_{1}^{n} y_{i}^{2}-2 r \sum_{1}^{n} y_{i} x_{i}+r^{2} \sum_{1}^{n} x_{i}^{2}\right\} \tag{5.8}
\end{equation*}
$$

noting that

$$
\begin{equation*}
r_{j}^{\prime}-\bar{x}_{n}^{\prime} \doteq\left\{\left(y_{j}-\bar{y}\right)-x\left(x_{j}-\bar{x}\right)\right\} /\{(n-1) \bar{x}\}=\frac{y_{j}-r x_{j}}{\bar{x}(n-1)} . \tag{5.9}
\end{equation*}
$$

Investigation by Rao and Rao (1971) using the model (5.4) indicates that $v(r)$ underestimates m.s.e. (r), whereas $v\left(r_{Q}\right)$ overestimates it, although the absolute bias of $v(r)$ is likely to be smaller than that of $v\left(r_{Q}\right)$.

If p.s.u."s are selected with probabilities $p_{j}$ and with replacement, formulae (5.5) - (5.7) hold good, provided $y_{j}$ and $x_{j}$ are replaced by $\hat{Y}_{j} / p_{j}$ and $\hat{X}_{j} / p_{j}$ respectively, where $\hat{Y}_{j}$ and $\hat{X}_{j}$ are unbiased estimates of $j$ th primary totals $Y_{j}$ and $X_{j}$ based on the elements selected from that primary.

## Stratified sampling

H.L. Jones (1974) has obtained 'jack-knife' estimators for the case of stratified simple random sampling without replacement, employing Taylor expansions. Suppose $n_{h}$ units are selected from the $N_{h}$ units in stratum $h(h=l, \ldots, L)$ by srswor independently in each stratum. Then the combined ratio estimator of $R$ is

$$
\begin{equation*}
\hat{R}=\frac{\sum W_{h} \bar{Y}_{h}}{\sum W_{h} \bar{x}_{h}} \tag{5.10}
\end{equation*}
$$

where $W_{h}=N_{h} / N$ and $\bar{Y}_{h}$ and $\bar{x}_{h}$ are $h$-th stratum sample means. Let $\hat{R}_{\text {(hi) }}$ denote the estimator of $R$, of the same form as $R$, obtained by omitting $i^{\text {th }}$ unit in stratum $h$ and let $\hat{R}(h)=\sum_{i} \hat{R}_{(h i)} / n_{h}$. Then Jones' jack-knife estimator of $R$ is

$$
\begin{equation*}
\hat{R}_{J}=\left[1+\sum_{1}^{L}\left(n_{h}-1\right)\left(1-\frac{n_{h}}{N_{h}}\right)\right] \hat{R}-\sum_{1}^{L}\left(n_{h}-1\right)\left(1-\frac{n_{h}}{N_{h}} \hat{R}_{(h)}\right. \tag{5.11}
\end{equation*}
$$

whose bias (for large L) does not involve second order population moments. If $1-\frac{h_{h}}{N_{h}} \doteq 1$ and $n_{h}=2$, we get

$$
\begin{equation*}
\hat{R}_{J}=2 \hat{R}-\hat{Z}_{l}^{L} \hat{R}_{(h)} \tag{5.12}
\end{equation*}
$$

which was proposed by McCartiny (1966). Jones' estimator of mse $\hat{R}$ (or of $\hat{R}_{J}$ ) is

$$
\begin{equation*}
v_{J}(\hat{R})=\sum_{1}^{L}\left(1-\frac{n_{h}}{N_{h}}\right) \frac{1}{n_{h}} s_{(h)}^{2} \tag{5.13}
\end{equation*}
$$

where $\left.s_{(h)}^{2}=\sum_{1}^{n} \hat{\left(R_{(h i)}\right.}-\hat{R}_{(h)}\right)^{2} / n_{h}$. For $n_{h}=2$ and $1-n_{h} / N_{h} \doteq 1$,
(5.13) reduces to

$$
\left.\left.v_{J}(\hat{R})=\begin{array}{l}
\mathrm{L}  \tag{5.14}\\
1
\end{array} \frac{1}{2} \sum_{1}^{2} \hat{R}_{(h i)}-\hat{R}_{(h)}\right)^{2}=\sum_{1}^{L} \frac{1}{4} \hat{R}_{(h l)}-\hat{R}_{(h 2)}\right)^{2}
$$

McCarthey (1966) proposed

$$
v_{M}(\hat{R})=\begin{array}{lll}
L & \frac{1}{2} & \sum_{1}  \tag{5.15}\\
1 & \hat{R}_{(h i)}-\frac{1}{L} & \left.\sum \hat{R}_{(h)}\right)^{2} .
\end{array}
$$

Lee (1973) employed

$$
v_{L}(\hat{R})=\begin{array}{lll}
L & 2  \tag{5.16}\\
\sum & \frac{1}{2} & \hat{R}^{(h i)} \\
1
\end{array} \hat{R}^{2} .
$$

Kish and Frankel (1970) suggested, for $n_{h}=2$, deleting one unit from stratum $h$ but including the other unit twice and then computing the estimator from (5.10): in all 2 L estimators. For the case of $f_{h}=2 / N_{h}=f$, i.e. equal $N_{h}$. one of their estimators of m.s.e. (R) reduces to Jones' (5.13). Other estimators of m.s.e. (R) are similar to (4.21) and (4.22) in the case of BRR:


$$
\begin{align*}
& v_{1}(\hat{R})=(1-f) \hat{\sum} \begin{array}{l}
\mathrm{L}\left(\hat{R}_{(h i)}-\hat{R}\right)^{2} \\
v_{2}(\hat{R})=(1-f) \sum_{1}^{L}\left(\hat{R}_{(h j)}-\hat{R}^{2}\right. \text { where (hj) is the } \\
\text { unit omitted }
\end{array}  \tag{5.17}\\
& v_{3}(\hat{R})=\frac{v_{1}(\hat{R})+v_{2}(\hat{R})}{2} . \tag{5.18}
\end{align*}
$$

For large L, all the variance estimators are approximately equal to the customary variance estimator $\left(n_{h}=2\right)$

$$
\begin{equation*}
\hat{v(\hat{R})}=\frac{1}{\left(\Sigma W_{h} \bar{x}_{h}\right)^{2}}\left[\Sigma \frac{W_{h}^{2}}{2}\left(s_{y h}^{2}-2 \hat{R} s_{y x h}+\hat{R}^{2} s_{x h}^{2}\right)\right] \tag{5.20}
\end{equation*}
$$

where $s_{y h}^{2}=\left(y_{h 1}-y_{h 2}\right)^{2} / 2, s_{y x h}=\left(y_{h 1}-y_{h 2}\right)\left(x_{h 1}-x_{h 2}\right) / 2$,

$$
x_{x h}^{2}=\left(x_{h 1}-x_{h 2}\right)^{2} / 2
$$

An advantage of the 'jack-knife' method is that variance estimators are available for general $n_{n} \geq 2$ unlike the $B R R$ which is applicable only for $n_{h}=2$.

If P.S.U."s are selected with probabilities $p_{h i}$ with replacement from each stratum, then the above formulae hold good, provided $y_{h i}$ is replaced by $\hat{Y}_{h i} / p_{h i}$ and $x_{h i}$ by $\hat{X}_{h i} / p_{h i}$.

## 6. Taylor Expansion Method

Suppose $\underset{\sim}{Y}=\left(Y_{1}, Y_{2}, \ldots, Y_{k}\right)$ is a vector of population parameters and $\hat{\sim}$

* Extension of $B R R$ to $n_{n}>2$ has recently been given by M. Gurney (Technical Memorandum, U.S. Bureau of Census) and J.I. Borack (1971, Ph. D. Thesis, Cornell University).
parameter of interest is $g(\underset{\sim}{Y})$ which is estimated by $g(\underset{\sim}{Y})$. Then by Taylor series expansion,

$$
\begin{equation*}
g(\hat{Y}) \doteq g(\underset{\sim}{Y})+\left.\sum_{1}^{k}\left(\hat{Y}_{i}-Y_{i}\right) \frac{\partial g(\hat{Y})}{\partial Y_{i}}\right|_{\hat{Y}_{i}=Y_{i}} \tag{6.1}
\end{equation*}
$$

to first approximation. Therefore $E g(\underset{\sim}{Y}) \doteq g(\underset{\sim}{Y})$ and

$$
\left.\operatorname{mse}(g(\hat{Y})) \stackrel{\sum}{\underline{V}}\left[\begin{array}{c}
k \\
1
\end{array} \hat{Y}_{i}-Y_{i}\right) \frac{\partial g(\underset{\sim}{Y})}{\partial Y_{i}}\right]=\operatorname{Var}\left[\sum_{1}^{k} \hat{Y}_{i} \frac{\partial g(\underset{\sim}{Y})}{\partial Y_{i}}\right]
$$

Now $\hat{Y}_{i}=\sum_{1}^{n} w_{j} Y_{i j}=\sum_{1}^{n} y_{i j}^{\prime}$ (say) where $w_{j}$ are the weights attached to the elements in the sample, and $\left\{\partial g(\hat{Y}) / \partial \hat{Y}_{i} \mid \hat{Y}_{i}=Y_{i}\right\}=\left\{\partial g\left(Y_{i j}^{\prime}\right) / \partial Y_{i j}^{\prime} \mid Y_{i j}^{\prime}=E\left(Y_{i j}^{\prime}\right)\right\}$ for all j. Therefore

$$
\operatorname{mse}(g(\hat{Y})) \equiv \operatorname{Var}\left[\begin{array}{cc}
k & k \\
1 & \sum \\
1 & \partial g_{\left(y_{i j}^{\prime}\right)}
\end{array} y_{i j}^{\prime}\right]
$$

$$
\left.\left.=\operatorname{Var} \begin{array}{rr}
n & k \\
1 & \left(\sum_{1}^{n}\right. \\
& \frac{\partial g}{\partial E\left(y_{i j}^{\prime}\right)}
\end{array} y_{i j}^{\prime}\right)\right]=\operatorname{Var}\left(\begin{array}{c}
n \\
1
\end{array} z_{j}\right) \operatorname{say}
$$

where $z_{j}=\sum_{i}^{k} \frac{\partial g}{\partial E\left(Y_{i j}^{\prime}\right)} y_{i j}^{\prime}$. This method avoids the computation of $M$ variances of $\hat{Y}_{i}$ and $M(M-1) / 2$ covariances of $\hat{Y}_{i}$ and $\hat{Y}_{i}$, from (6.2) ; we only need to compute $Z_{j}$ 's and apply the usual formula for single variate. For variance estimation, we substitute the estimates of partial derivatives $\partial g / \partial E\left(y_{i j}^{\prime}\right)$ from the sample and then apply the usual variance estimator formula for a single variate. The Taylor expansion method is general and applicable to any sample design, provided the variance estimator formula for estimated total is known. The above method of simplification is due to Woodruff (1971).

## Example

For stratified srswor,

$$
\hat{R}=\begin{array}{llll}
L & n_{h} & L & n_{h} \\
\Sigma & \sum^{h} & y_{h i}^{\prime} / \Sigma & \sum^{h} \\
l & 1 & x_{h i}^{\prime}=\frac{y}{x} \\
l
\end{array}
$$

where

$$
y_{h i}^{\prime}=\frac{N_{h}}{n_{h}} y_{h i}
$$

and

$$
x_{h i}^{\prime}=\frac{N_{h}}{n_{h}} x_{h i} .
$$

Also

$$
\frac{\partial \hat{R}}{\partial y_{h i}^{\prime}}=\frac{1}{X}, \quad \frac{\partial g}{\partial E\left(x_{h i}^{\prime}\right)}=-\frac{Y}{X^{2}}
$$

so

$$
\begin{aligned}
& \frac{\partial g}{\partial E\left(y_{h i}^{\prime}\right)}=\frac{1}{x} \\
& \frac{\partial g}{\partial E\left(x_{h i}^{\prime}\right)}=-\frac{Y}{x^{2}}
\end{aligned}
$$

and

$$
\begin{aligned}
& \sum_{1}^{n}\left\{\frac{\partial g}{\partial E\left(y_{h i}^{\prime}\right)} y_{h i}^{\prime}+\frac{\partial g}{\partial E\left(x_{h i}^{\prime}\right)} x_{h i}^{\prime}\right\}=\sum_{1}^{L} \sum_{1}^{n}\left\{\frac{1}{x} y_{h i}^{\prime}-\frac{y}{x^{2}} x_{h i}^{\prime}\right\} \\
& =\begin{array}{ll}
I & n_{h} \\
I & \sum_{1}^{h}
\end{array}\left\{\frac{N_{h}}{n_{h}}\left(\frac{y_{h i}}{X}-\frac{Y}{x^{2}} x_{h i}\right)\right\} .
\end{aligned}
$$

Therefore, letting $z_{h i}=\frac{y_{h i}}{\hat{x}}-\frac{\hat{y}}{\hat{x}^{2}} \quad x_{h i}$, we get

$$
\begin{aligned}
& v(\hat{R})=\sum v\left(\sum_{h}^{L} z_{h i}\right) \\
& 1 \quad 1 \\
&=\sum_{1}^{L} n_{h}^{2}\left(\frac{1}{n_{h}}-\frac{1}{N_{h}}\right) \frac{\sum^{n_{h}}\left(z_{h i}-\bar{z}_{h}\right)^{2}}{n_{h}-1}
\end{aligned}
$$

$$
\begin{equation*}
=\frac{1}{\hat{x}^{2}} \sum_{1}^{L} \frac{N_{h}^{2}}{n_{h}}\left(1-\frac{n_{h}}{N_{h}}\right) \sum_{1}^{n_{h}}\left(y_{h i}-\hat{R} x_{h i}-\bar{y}_{h}+\hat{R} \bar{x}_{h}\right)^{2} /\left(n_{h}-1\right) \tag{6.3}
\end{equation*}
$$

which is identical to (5.20) when $n_{h}=2$ and $1-\frac{n_{h}}{N_{h}} \doteq 1$

## 7. Empirical Study

Frankel (1972) empirically compared and evaluated BRR, 'jack-knife' and Taylor series method using data from the Current Population Survey of the U.S. Bureau of Census. Treating the sample as population, three clustered and stratified sample designs are constructed (approximately 14 elements/ cluster) with $n_{h}=2$ clusters/stratum ( $N_{h}$ 's are equal) by simple random sampling: (1) 6 strata $(n \doteq 170)$; (2) 12 strata ( $n \doteq 340$ ); (3) 30 strata ( $\mathrm{n} \rightleftharpoons 847$ ). For each design, $M=200$ or 300 independent samples are drawn and $g(\hat{Y}), V_{i}[g(\hat{\sim})]$ for each of the methods $i$ are computed. Empirical values of bias and m.s.e. of $v_{i}$ are obtained:

$$
B\left(v_{i}\right)=\frac{1}{M} \sum_{j=1}^{M}\left\{v_{i}[g(\underset{\sim}{Y})]\right\}-V[g(\underset{\sim}{Y})]
$$

where

$$
V[g(\hat{Y})]=\frac{1}{M} \sum_{j=1}^{M}\left\{[g(\hat{Y})]_{j}-\frac{1}{M} \sum[g(\hat{Y})]_{j}\right\}^{2}
$$

and

$$
\begin{aligned}
& V\left[v_{i}\right]=\frac{1}{M} \sum_{j=1}^{M}\left[\left\{v_{i}[g(\underset{\sim}{Y})]\right\}_{j}-\left\{\frac{1}{M} \sum_{1}^{M} v_{i}[g(\hat{Y})]\right\}_{j}\right]^{2} \\
& \text { m.s.e. }\left[v_{i}\right]=V\left(v_{i}\right)+B^{2}\left(v_{i}\right) .
\end{aligned}
$$

The variance estimators chosen are: for $B R R$ eqs. (4.24) and (4.25): for the "jack-knife" eqs. (5.13) and (5.19); (6.3) for Taylor series method. Results on the relative merits of the three methods with regard to bias and m.s.e. are not clear-cut. Therefore, a different criterion is chosen:
degree to which a method makes the approximation $t_{i}=\frac{g(\underset{\sim}{Y})-\operatorname{Eg}(\underset{\sim}{\underset{Y}{\sim}})}{\sqrt{v_{i}}[g(\underset{\sim}{\hat{Y}})]} \quad t$ (L d.f.) most valid. For each sample $t_{i}$ is computed from each of the 5 variance estimators and then the proportion of times each t-ratio falls in the limits specified by t-distribution with I d.f. is computed. These proportions are given in Table $1-3$ for ratio $g(\underset{\sim}{Y})=\hat{Y} / \hat{X}$ and difference of ratios $g(\hat{Y})=\hat{Y}_{1} / \hat{X}-\hat{Y}_{2} / \hat{X}=\left(\hat{Y}_{1}-\hat{Y}_{2}\right) / \hat{X}=\hat{Z} / \hat{X}$ where $z_{h i}=Y_{1 h i}-Y_{2 h i}$ and $X_{h i}=$ number of elements in (hi) ${ }^{\text {th }}$ cluster; values given are averages of proportions for 6 ratios and 12 differences of ratios.

Table 1: 6 strata design $(M=300)$

## Difference of Ratios

|  |  | 2.576 | 1.960 | 1.645 | 1.282 | 1.000 |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| BRR | [ eq. (4.22) | 0.9500 | 0.8997 | 0.8497 | 0.7578 | 0.6483 |
|  | [ eq. (4.22a) | 0.9481 | 0.8950 | 0.8450 | 0.7503 | 0.6436 |
| Jack-knife | eq. (5.19) | 0.9464 | 0.8939 | 0.8397 | 0.7428 | 0.6367 |
|  | eq. (5.13) | 0.9458 | 0.8889 | 0.8389 | 0.7400 | 0.6353 |
| Taylor | $\{\mathrm{eq}$. (6.3) | 0.9450 | 0.8842 | 0.8372 | 0.7381 | 0.6306 |
| Theoretical | rop. | 0.9580 | 0.9023 | 0.8489 | 0.7529 | 0.6441 |

Sable 2: 12 strata design $(M=200)$

## Ratio



## Difference of ratios

BRR $\left\{\begin{array}{lllllll}\text { eq. (4.22) } & 0.9658 & 0.9117 & 0.8617 & 0.7614 & 0.6458 \\ \text { eq. (4.22a) } & 0.9656 & 0.9097 & 0.8594 & 0.7586 & 0.6422\end{array}\right.$
Jack-knife $\left\{\begin{array}{llllll}\text { eq. (5.19) } & 0.9653 & 0.9083 & 0.8558 & 0.7561 & 0.6375 \\ \text { eq. (5.13) } & 0.9653 & 0.9083 & 0.8544 & 0.7558 & 0.6369\end{array}\right.$
Theoretical Prop.

Table 3: 30 strata design $(M=200)$

Ratio
$\pm 2.576 \pm 1.960 \pm 1.645 \pm 1.282 \pm 1.000$
BRR $\left\{\begin{array}{llllll}\text { eq. (4.22) } & 0.9825 & 0.9444 & 0.8906 & 0.7894 & 0.6569 \\ \text { eq. (4.22a) } & 0.9819 & 0.9437 & 0.8894 & 0.7881 & 0.6569\end{array}\right.$
Jack-knife $\left\{\begin{array}{llllll}\text { eq. (5.19) } & 0.9819 & 0.9431 & 0.8887 & 0.7856 & 0.6537 \\ \text { eq. (5.13) } & 0.9819 & 0.9431 & 0.8881 & 0.7850 & 0.6537\end{array}\right.$
$\begin{array}{lllllll}\text { Taylor }\left\{\begin{array}{llll}\text { eq. (6.3) } & 0.9819 & 0.9431 & 0.8881\end{array}\right. & 0.7844 & 0.6537 \\ \text { oretical Prop. } & 0.9848 & 0.9407 & 0.8896 & 0.7903 & 0.6747\end{array}$

## Difference of ratios

BRR $\left\{\begin{array}{lllllll}\text { eq. (4.22) } & 0.9829 & 0.9462 & 0.8875 & 0.7783 & 0.6475 \\ \text { eq. (4.22a) } & 0.9825 & 0.9454 & 0.8867 & 0.7779 & 0.6462\end{array}\right.$
Jack-knife $\left\{\begin{array}{llllll}\text { eq. }(5.19) & 0.9821 & 0.9433 & 0.8842 & 0.7742 & 0.6433 \\ \text { eq. (5.13) } & 0.9821 & 0.9433 & 0.8842 & 0.7742 & 0.6433\end{array}\right.$
Taylor $\left\{\begin{array}{llllll}\text { eq. (6.3) } & 0.9821 & 0.9433 & 0.8842 & 0.7742 & 0.6429\end{array}\right.$
Theoretical Prop.

Table 1: 6 strata design $(M=300)$
$\pm 2.576 \pm 1.960 \pm 1.645 \pm 1.282 \pm 1.000$
BRR $\left\{\begin{array}{llllll}\text { eq. (4.22) } & 0.9558 & 0.9042 & 0.8450 & 0.7562 & 0.6450 \\ \text { eq. (4.22a) } & 0.9533 & 0.8996 & 0.8404 & 0.7487 & 0.6379\end{array}\right.$
Jack-knife $\left\{\begin{array}{lllllll}\text { eq. (5.19) } & 0.9508 & 0.8942 & 0.8362 & 0.7421 & 0.6329 \\ \text { eq. (5.13) } & 0.9500 & 0.8912 & 0.8337 & 0.7396 & 0.6329\end{array}\right.$

Taylor $\left\{\begin{array}{llllll}\text { eq. (6.3) } & 0.9483 & 0.8879 & 0.8329 & 0.7379 & 0.6279\end{array}\right.$
Theoretical Prop.
0.9580
0.9023
0.8489
0.7529
0.6441

Results in Tables $1-3$ clearly show that the average proportions produced by $B R R$ (eq. 4.24) agree better with theoretical proportions than the average proportions produced by the other methods, and the agreement is good. However, the differences between the methods are small. Since (6.3) is computationally simpler than $B R R$ and 'jack-knife' variance estimators, Frankel (1972) recommends Taylor series method for ratios and differences of ratios.

## 8. A General Method

Lee (1973) proposed a general method which yields BRR or 'jack-knife'
as special case. Suppose the L strata are divided into g groups ( $g$ L) with $r_{t}$ strata in $t^{\text {th }}$ group ( $t=1, \ldots g$ ). In each group $B R R$ is used to select half samples (assuming 2 units per stratum selected with replacement). Suppose $q_{t} \leq 2^{r} t$ half samples are selected by BRR from $t^{\text {th }}$ group. Then the variance estimator for ratio estimation is

$$
\begin{equation*}
\left.v_{g}(\hat{R})=\sum_{t=1}^{g} \sum_{\alpha=1}^{q} \hat{(R}_{(t \alpha)}-\hat{R}\right)^{2} / q_{t}, \quad q_{t} \geq 2 \tag{8.1}
\end{equation*}
$$

where

$$
\begin{equation*}
\tilde{R}_{(t d)}=\frac{\sum^{(t)} W_{h}\left(y_{h 1}+y_{h 2}\right) / 2+\sum_{h}^{t} W_{h}\left(\delta_{h 1}^{(\alpha)} y_{h 1}+\delta_{h 2}^{(\alpha)} y_{h 2}\right)}{\sum_{h}^{(t)} W_{h}\left(x_{h 1}+x_{h 2}\right) / 2+\sum_{h}^{t} W_{h}\left(\delta_{h 1}^{(\alpha)} x_{h 1}+\delta_{h 2}^{(\alpha)} x_{h 2}\right)} \tag{8.2}
\end{equation*}
$$

where $\delta_{h l}^{(\alpha)}=1$ if (hl) is selected for $\alpha^{\text {th }}$ half sample, $=0$ otherwise and $\delta_{h 2}^{(\alpha)}=1-\delta_{h 1}^{(\alpha)} ; \Sigma^{t}$ is summation over the $r_{t}$ strata in $t^{\text {th }}$ group and $\Sigma^{(t)}$ is summation over remaining strata. In the linear case, $v_{g}(R)$ reduces to $v\left(\bar{y}_{s t}\right)$ as in the case of $B R R$ and the 'jack-knife' variance estimator (5.16) and $g=1$ gives the $B R R$ variance estimator (4.15). Lee (1973) made a small empirical study which indicates that the absolute bias of $v_{g}(R)$ increases with $g$, whereas m.s.e. of $v_{g}(R)$ decreases as $g$ increases, i.e.' 'jack-knife' method leads to smallest mean square error and 'BRR' to smallest absolute bias. Frankel (1972) did not use (5.16), so Lee's result is readily not comparable to Frankel's. Lee has not considered the behaviour of the t-statistic.

Lee's estimator (8.1) is applicable to any sample design provided 2 p.s.u.'s are selected from each stratum with replacement and we replace $Y_{h i}$ by $\hat{Y}_{h i} / P_{h i}$ as before
9. Estimating Relationships between Variables

Consider the finite population of pairs $\left(y_{1}, x_{1}\right), \ldots,\left(y_{N}, x_{N}\right)$. Several survey practitioners estimate population values like the finite population regression coefficient (of $y$ on $x$ )

$$
\begin{equation*}
B=\frac{S_{x y}}{S_{x}^{2}}=\frac{\sum\left(x_{i}-\bar{x}\right)\left(y_{i}-\bar{Y}\right)}{\sum\left(x_{i}=\bar{X}\right)^{2}}=\frac{N \sum y_{i} x_{i}-Y X}{N \sum x_{i}^{2}-x^{2}} \tag{9.1}
\end{equation*}
$$

from the sample selected by a specified sample design. For instance, with unequal probabilities $p_{i}$ with replacement, $S_{x y}$ and $S_{x}^{2}$ are estimated term by term and their ratio taken as an estimate of $B$ :

$$
b_{p p s}=\frac{\left(\sum_{1}^{n} p_{i}^{-1}\right)\left(\sum_{1}^{n} p_{i}^{-1} x_{i} y_{i}\right)-\left(\sum_{1}^{n} p_{i}^{-1} y_{i}\right)\left(\sum_{1}^{n} p_{i}^{-1} x_{i}\right)}{\left(\sum_{1}^{n} p_{i}^{-1}\right)\left(\sum_{1}^{n} p_{i}^{-1} x_{i}\right)-\left(\sum_{1}^{n} p_{i}^{-1} x_{i}\right)^{2}}
$$

One could also estimate $S_{x y}$ and $S_{x}^{2}$ unbiasedly, noting that

$$
\begin{align*}
& s_{x}^{2}=\frac{N}{N(N-1)} \sum_{i<j} \sum\left(x_{i}-x_{j}\right)^{2} \text { and } s_{x y}=\frac{2}{N(N-1)} \sum_{i<j} \sum\left(x_{i}-x_{j}\right)\left(y_{i}-y_{j}\right): \\
& \tilde{b}_{p p s}= \frac{\sum \sum \sum\left\{t_{i} t_{j} / E\left(t_{i} t_{j}\right)\right\}\left(x_{i}-x_{j}\right)\left(y_{i}-y_{j}\right)}{V} \begin{aligned}
& \sum \sum\left\{t_{i} t_{j} / E\left(t_{i} t_{j}\right)\right\}\left(x_{i}-x_{j}\right)^{2}
\end{aligned} \tag{9.3}
\end{align*}
$$

where $v=$ no. of district units in the sample, $E\left(t_{i} t_{j}\right)=n(n-1) p_{i} p_{j}$. Unlike in the case of total $Y$, mean $\bar{Y}$ or ratio $Y / X$, it is not clear what meaning one can attach to $B$ without an underlying model. If we assume the model

$$
\begin{equation*}
y_{i}=\alpha+\beta x_{j}+\varepsilon_{j}, \quad j=1, \ldots, N \tag{9.4}
\end{equation*}
$$

with the $E_{j}$ assumed as i.i.d. random variables with mean 0 and variance $\sigma^{2}$, then $B$ is the least square estimate of $\beta$.

Some people advocating the estimation of B (e.g., Kish and Frankel, 1974) argue that some researchers could profit by knowing how much of a relationship between $y$ and $x$ could be explained or accounted for by a linear model (9.4) or some other model. They emphasize that it is not necessary to assume a model like (9.4), since we are only asking how much of the variability in $y$ can be explained by the particular model we choose to try. However, others argue that researchers' primary aim is to discover some (at least approximate) relationship between $y$ and $x$ and then the amount of variation that cannot be accounted for by such a relationship would have some natural relevance to the evaluation of the results obtained. That is, most users are concerned with estimating parameters of an appropriate model rather than estimating descriptive expressions like B (see Brewer and Mellor, 1974 for an illuminating discussion and also the discussions on Kish and Frankel's 1974 paper by Konijn and T.M.F. Smith).

We consider both approaches here and present some empirical results on descriptive measures like $B$. However, one should not forget that the problem of estimating relationships among variables is deeper than simply estimating descriptive quantities and that it demands close collaboration between the statistician and the subject matter specialist.

### 9.1 Descriptive Measures

Frankel (1972) employed estimates of measures like $B$ which are strictly valid only under simple random sampling. However, since the sampling designs used in his study are self-weighting, these estimates are consistent for large samples. Wakimoto (1971), derived an unbiased
estimate of the correlation coefficient $\rho=S_{x y} /\left(S_{x}^{2} S_{y}^{2}\right)^{1 / 2}$ for stratified simple random sampling of elements: estimate of $s_{y}^{2}$ is
and similar expressions for $\hat{S}_{X Y}, \hat{S}_{X}^{2}$ and then

$$
\begin{equation*}
\hat{\rho}=\hat{S}_{x y} /\left(\hat{S}_{x}^{2} \hat{S}_{Y}^{2}\right) 1 / 2 \quad \text { (also Koop, 1970) } \tag{9.6}
\end{equation*}
$$

For measuring relationship between $y$ and $\underset{\sim}{x}=\left(x_{i}, \ldots, x_{p}\right)$, Kish and Frankel (1974) choose to define parameters as $B_{j}$ such that

$$
\begin{equation*}
\sum_{i=1}^{N}\left(y_{i}-\sum_{j=0}^{p} B_{j} x_{j i}\right)^{2} \quad \text { is minimum, }\left(x_{0 i}=1\right) \tag{9.7}
\end{equation*}
$$

i.e., $B_{j}$ are the ordinary least squares regression coefficients. The estimates $b_{j}$ are obtained by treating the sample ( $y_{i}, x_{1 i}, \ldots, x_{p i}$ ), $i=1, \ldots, n$ as if it is a simple random sample and, therefore, by minimizing

$$
\begin{equation*}
\sum_{i=1}^{n}\left(y_{i}-\sum_{j=0}^{p} b_{j} x_{j i}\right)^{2} \tag{9.8}
\end{equation*}
$$

w.r.t. $b_{j}$ which leads to the customary normal equations:

$$
\begin{equation*}
\sum_{j=0}^{p} u_{k j} b_{j}=v_{k}, k=0,1, \ldots, p \tag{9.9}
\end{equation*}
$$

where $u_{k j}=\sum_{i=1}^{n} x_{j i} x_{k i}, v_{k}=\sum_{i=1}^{n} Y_{i} x_{k i}$. Half sample estimates and 'Jack-knife' estimates of $B_{j}$ are similarly obtained.

We need the partial derivates $\partial b_{j} / \partial u_{k s}$ and $\partial b_{j} / \partial v_{k}(s=0,1, \ldots, p)$
to find the variable estimates by Taylor expansion method. Tepping (1968) has given a systematic method for this: Differentiating (9.9) w.r.t. us $(\ell=0,1, \ldots, p)$, we get

$$
\begin{array}{r}
\sum_{j=0}^{p} u_{k j} \frac{\partial b_{j}}{\partial u_{s \ell}}=-\left(1-\delta_{s \ell}\right) \delta_{k s} b_{\ell}-\delta_{k \ell} b_{s}  \tag{9.10}\\
k, s, \ell=0,1, \ldots, p
\end{array}
$$

and differentiation w.r.t. $v_{s}$ leads to

$$
\begin{equation*}
\sum_{j=0}^{p} u_{k j} \frac{\partial b_{j}}{\partial v_{s}}=\delta_{k s}, \quad k, s=0,1, \ldots, p \tag{9.11}
\end{equation*}
$$

where $\delta_{k s}=1$ if $k=s,=0$ otherwise. The system of equations (9.10) can be divided into $(p+1)^{2}$ subsystems $(s, 2=0,1, \ldots, p)$, each of $p+1$ linear equations in $p+1$ variables $\partial b_{j} / \partial u_{s l}, j=0,1, \ldots, p$. However., since $u_{s \ell}=u_{\ell S}$ only (l/2) (p + 2) (p + 2) of the subsystems are distinct. Similarly, the system (9.11) can be subdivided into $p+1$ subsystems each of $p+1$ linear equations in $p+1$ variables $\partial b_{j} / \partial v_{s}, j=0,1, \ldots, p$. In total, we need to solve $(1 / 2)(p+1)(p+4)$ sets, $p+1$ linear equations in each set. For BRR, we have to solve $L$ sets ( $L=$ no. of strata) each of $p+1$ linear equations. So the computations involved depend on the relative magnitudes of $L$ and $(1 / 2(p+1)(p+4)$. For the jack-knife, one could use short-cut methods for deletion of an observation ( $y_{i}, x_{0 i}, \ldots, x_{p i}$ ) from regression equations.

Tables 4-6 give empirical results (taken from Frankel's 1972 paper) on t-statistic for simple correlation coefficient, multiple regression coefficients and multiple correlation coefficient; values given are averages of proportions for 12,8 and 2 coefficients respectively (Taylor expansion method not used for multiple correlation coefficient).

Table 4: 6 strata design $(M=300)$

## Simple Correlation Coefficient



Multiple Regression Coefficient
BRR $\left\{\begin{array}{llllll}\text { eq. (4.22) } & 0.9662 & 0.9150 & 0.8600 & 0.7683 & 0.6642 \\ \text { eq. (4.22a) } & 0.9587 & 0.8996 & 0.8433 & 0.7446 & 0.6446\end{array}\right.$
Jack-knife $\left\{\begin{array}{llllll}\text { eq. }(5.19) & 0.9521 & 0.8833 & 0.8304 & 0.7312 & 0.6200 \\ \text { eq. (5.13) } & 0.9454 & 0.8796 & 0.8258 & 0.7262 & 0.6142\end{array}\right.$
Taylor $\left\{\begin{array}{llllll}\text { eq. (6.3) } & 0.9421 & 0.8733 & 0.8146 & 0.7167 & 0.6029\end{array}\right.$
Theoretical Prop.

Multiple Correlation Coefficient
BRR $\left\{\begin{array}{llllll}\text { eq. (4.24) } & 0.9350 & 0.8950 & 0.8233 & 0.7383 & 0.6133 \\ \text { eq. (4.22a) } & 0.9033 & 0.8217 & 0.7583 & 0.6417 & 0.5467\end{array}\right.$
Jack-knife $\left\{\begin{array}{lllll}\text { eq. (5.19) } & 0.9117 & 0.8400 & 0.7800 & 0.6600 \\ \text { eq. (5.13) } & 0.8850 & 0.8033 & 0.7350 & 0.6133\end{array} 0.000\right.$

Table 5: 12 strata design $(M=200)$

Simple Correlation Coefficient


Multiple Regression Coefficient
BRR $\left\{\begin{array}{llllll}\text { eq. (4.22) } & 0.9733 & 0.9337 & 0.8746 & 0.7733 & 0.6529 \\ \text { eq. (4.22a) } & 0.9700 & 0.9250 & 0.8654 & 0.7646 & 0.6412\end{array}\right.$
Jack-knife $\left\{\begin{array}{llllll}\text { eq. (5.19) } & 0.9675 & 0.9162 & 0.8542 & 0.7496 & 0.6283 \\ \text { eq. (5.13) } & 0.9671 & 0.9142 & 0.8508 & 0.7471 & 0.6250\end{array}\right.$
Theoretical Prop.

## Multiple Correlation Coefficient

BRR $\left\{\begin{array}{llllll}\text { eq. (4.22) } & 0.9200 & 0.8500 & 0.7900 & 0.6767 & 0.5500 \\ \text { eq. (4.22a) } & 0.9067 & 0.8150 & 0.7400 & 0.6067 & 0.5067\end{array}\right.$
Jack-knife $\left\{\begin{array}{lllll}\text { eq. (5.19) } & 0.8950 & 0.8133 & 0.7383 & 0.6333 \\ \text { eq. (5.13) } & 0.8850 & 0.7933 & 0.7067 & 0.5933\end{array} 0.467\right.$
Theoretical Prop.

Table 6: 30 strata design ( $M-200$ )

Simple Correlation Coefficient
BRR $\left\{\begin{array}{lllllll}\text { eq. (4.22) } & \pm 2.576 \pm 1.960 \pm 1.645 & 0.9725 & 0.9108 & 0.8617 & 0.7533 & 0.6325 \\ \text { eq. (4.22a) } & 0.9696 & 0.9083 & 0.8550 & 0.7467 & 0.6212\end{array}\right.$
Jack-knife $\left\{\begin{array}{llllll}\text { eq. (5.19) } & 0.9658 & 0.9021 & 0.8471 & 0.7346 & 0.6137 \\ \text { eq. (5.13) } & 0.9658 & 0.9008 & 0.8442 & 0.7333 & 0.6112\end{array}\right.$
Theoretical Prop.
eq. (6.3)

Multiple Regression Coefficient
BRR $\left\{\begin{array}{lllllll}\text { eq. (4.22) } & 0.9825 & 0.9381 & 0.8900 & 0.7887 & 0.6706 \\ \text { eq. (4.22a) } & 0.9812 & 0.9369 & 0.8881 & 0.7831 & 0.6687\end{array}\right.$
Jack-knife $\left\{\begin{array}{llllll}\text { eq. (5.19) } & 0.9800 & 0.9325 & 0.8844 & 0.7787 & 0.6631 \\ \text { eq. (5.13) } & 0.9794 & 0.9319 & 0.8844 & 0.7787 & 0.6619\end{array}\right.$
Theoretical Prop.

Multiple Correlation Coefficient
BRR $\left\{\begin{array}{llllll}\text { eq. (4.22) } & 0.9125 & 0.8250 & 0.7350 & 0.6375 & 0.5272 \\ \text { eq. (4.22a) } & 0.8975 & 0.8100 & 0.7175 & 0.6125 & 0.4975\end{array}\right.$
Jack-knife $\left\{\begin{array}{llllll}\text { eq. (5.19) } & 0.8950 & 0.7925 & 0.7025 & 0.5950 & 0.4950 \\ \text { eq. (5.13) } & 0.8875 & 0.7925 & 0.6975 & 0.5825 & 0.4700\end{array}\right.$
Theoretical Prop.

Results in Tables $4-6$ again clearly show that the average proportions produced by $B R R$ (eq. 4.22) agree better with theoretical proportions than those produced by the other methods. Also agreement is good except in the case of multiple correlation coefficient -- note that the observed proportion is decreasing as the number of strata increases, unlike in the estimation of other parameters.

### 9.2 Estimation of models

Konijn (1962) formulated a regression model appropriate to a two-stage sampling design. Suppose the $h^{\text {th }}$ cluster ( $h=1$, ... L) has $N_{h}$ elements and the elements in the $h^{\text {th }}$ cluster are assumed to have been selected from an infinite super-populatior satisfying the linear model

$$
\begin{align*}
& Y_{h i}=\alpha_{h}+\beta_{h} x_{h i}, \quad i=1, \ldots, N_{h},  \tag{9.12}\\
& h=1, \ldots, L \\
& E\left(e_{h i} \mid x_{h i}\right)=0, E\left(e_{h i}^{2} \mid x_{h i}\right)=\sigma_{i}^{2}, e_{h i} \text { and } e_{h j} \text { uncorrelated. }
\end{align*}
$$

The parameters of interest are taken as

$$
\begin{equation*}
\alpha=\frac{\sum N_{h} \alpha_{h}}{N}, \quad \beta=\frac{\sum N_{h} \beta_{h}}{N} \text { while } N=\Sigma N_{h} \tag{9.13}
\end{equation*}
$$

A justification for choosing these parameters is that an individual selected at random from the population will satisfy the model $y=\alpha+\beta x+e, E(e \mid x)=0$.

Suppose \& clusters are selected by some sampling method (without replacement) with inclusion probabilities $\pi_{h h \prime}$ " and if $h^{\text {th }}$ cluster is selected $n_{h}$ (predetermined) elements are selected by simple random sampling $\left(\sum_{h}^{l} n_{h}\right)$. Given that $h^{\text {th }}$ cluster is in the sample, we have $E\left(b_{h}\right)=\beta_{h}$ where $b_{h}^{l}$ is the ordinary least squares estimator: $b_{h}=\sum_{i} y_{h i}\left(x_{h i}-\bar{x}_{h}\right) / \sum_{i}\left(x_{h i}-\bar{x}_{h}\right)^{2}$. Therefore, an unbiased estimator of $\beta$ is

$$
\begin{equation*}
\beta=\frac{\ell N_{h} b_{h}}{\sum} \frac{\ell}{N \pi_{h}}=\sum_{1} b_{h} / \ell \quad \text { if } \quad \pi_{h}=\ell \frac{N_{h}}{N} . \tag{9.14}
\end{equation*}
$$

In the case of stratified simple random sampling, $\pi_{h}=1$ and

$$
\hat{B}=\begin{align*}
& L  \tag{9.15}\\
& I
\end{align*} N_{h} \mathrm{~b}_{\mathrm{h}} / \mathbb{N} .
$$

Similarly, one obtains unbiased estimatcr of $\alpha$, noting that $\alpha_{h}$, given $h^{\text {th }}$ cluster in sample, is estimated by $a_{h}=\bar{y}_{h}-b_{h} \bar{x}_{h}$. One could, of course, estimate any specified linear combination of $\beta_{h}{ }^{\prime}$ s or $\alpha_{h}{ }^{\prime}$ s. The usual estimate

$$
\begin{equation*}
b=\frac{\sum \sum\left(x_{h i}-\bar{x}\right) y_{h i}}{\sum \sum\left(x_{h i}-\bar{x}\right)^{2}}, \bar{x}=\sum_{n_{h}} \bar{x}_{h} / n \tag{9.16}
\end{equation*}
$$

is not unbiased unless $\alpha_{h}=\alpha, \beta_{h}=\beta$. The variance of $\hat{\beta}$ is estimated unbiasedly by

$$
\begin{align*}
v(\hat{\beta})= & \sum_{1}^{\ell} w_{h}^{2}{\hat{\sigma_{h}}}_{h}^{2}\left[\pi_{h} \sum_{i}\left(x_{h i}-\bar{x}_{h i}\right)^{2}\right] \\
& +\sum_{h h^{\prime}}^{\ell} \frac{\left(\pi_{h} \pi_{h} \cdot i \pi_{h h^{\prime}}\right)}{\pi_{h h^{\prime}}}\left(\frac{b_{h} W_{h}}{\pi_{h}}-\frac{b_{h^{\prime}} W_{h}^{\prime}}{\pi_{h}}\right)^{2} \tag{9.17}
\end{align*}
$$

where $\hat{\sigma}_{h}^{2}=\sum_{i}\left(y_{h i}-a_{h}-b_{h} x_{h i}\right)^{2} /\left(n_{h}-2\right)$, assuming $n_{h}>2$ for each $h$. Similarly $v(\alpha)$ and $\operatorname{cov}(\alpha, \beta)$ are obtained. Konijn also treats the case of sampling clusters with replacement. Note that the variance estimate (9.17) is the sum of two components: First component is due to within cluster variability and the second is due to the fact that a sample of clusters is selected.

An alternative formulation for two stage sampling is due to Fuller (1972) which reflects the 'cluster effect':

$$
\begin{align*}
y_{h i}=\alpha+\beta x_{h i}+u_{h}+e_{h i} \quad & i=1, \ldots, N_{h}  \tag{9.18}\\
h & =1, \ldots, L
\end{align*}
$$

$$
=\alpha+\beta x_{h i}+\tilde{e}_{h i}
$$

where $u_{1}, \ldots, u_{L}$ is a random sample from an infinite population with mean 0 and variance $\sigma_{u}^{2}$, $e_{h i}$ have mean zerc and variance $\sigma_{e}^{2}$ and distributed independently of $\mathrm{v}_{\mathrm{h}}$. This model induces correlations between elements in the same cluster which takes into account the fact that olements within a cluster are often positively correlated:

$$
\operatorname{cov}\left(\tilde{e}_{h i^{\prime}}, \tilde{e}_{h^{\prime} i^{\prime}}\right)= \begin{cases}\sigma_{e}^{2} \div \sigma_{u}^{2} & \text { if } h=h^{\prime}, i=i^{\prime}  \tag{9.19}\\ \sigma_{e}^{2} & \text { if } h=h^{\prime}, i=/ i^{\prime} \\ 0 & \text { if } h \neq h^{\prime}\end{cases}
$$

If we assume that a two-stage random sample of $l$ clusters with $n_{h}$ elements form $h^{\text {th }}$ selected cluster, then the estimates of $\alpha$ and $\beta$ are simply obtained by applying ordinary least squares to 'transformed' observations

$$
\begin{aligned}
& \tilde{y}_{h i}=y_{h i}-\eta_{h} \bar{y}_{h}, \tilde{x}_{h i}=x_{h i}-\eta_{h} \bar{x}_{h} \\
& \eta_{h}=1-\left\{\frac{\sigma_{e}^{2}}{\sigma_{e}^{2}+\eta_{h} \sigma_{u}^{2}}\right\}^{1 / 2}
\end{aligned}
$$

i.e., $\hat{\beta}=\sum \sum \tilde{y}_{h i}\left(\tilde{x}_{h i}-\frac{\tilde{x}}{x} / \sum \sum\left(\tilde{x}_{h i}-\tilde{\bar{x}}\right)^{2}, \tilde{x}=\sum \sum x_{h i} / n\right.$ and $\hat{x}=\frac{\tilde{y}}{}-\hat{\beta} \tilde{\bar{x}}$. Since $\eta_{h}$ are unknown, one could obtain estimates of $\sigma_{e}^{2}$ and $\sigma_{u}^{2}$ and, hence, of $\eta_{h}$ by the method of fitting constants. The resulting estimates of $\alpha$ and $\beta$ are also unbiased provided the $\tilde{e}_{\text {hi }}$ have symmetric distribution. The model (9.18) is the simplest model for two-stage sampling which suggests itself various extensions suitable for different practical situations. For instance, if we believe that the slopes vary from cluster to cluster, we could change (9.18) to

$$
\begin{equation*}
y_{h i}=\alpha+\beta x_{h i}+\delta_{h}\left(x_{h i}-\bar{x}_{h}\right)+\tilde{e}_{h i} \tag{9.21}
\end{equation*}
$$


where $\delta_{h}$ 's are regarded as random slopes taken from a distribution with mean 0 and variance $\sigma_{\delta}^{2}$ and $\varepsilon_{h}$ 's assumed independent of $u_{h}$ 's (see Fuller (1972) for details of estimation of $\alpha$ and $\beta$ ). Fuller also gives procedures for testing mocels of the type (9.18) anci is.21).

Porter (1973) has providea an extension of Konijn's model to time series sample survey data. For each element in the finite population, it is assumed that $T$ observations ( $T=$ no. of time periods) are observable and the 'economic' relationship

$$
\begin{equation*}
{\underset{\sim}{y}}_{i}=\underset{\sim}{x} \underset{\sim}{\beta}{ }_{i}+\underset{\sim}{e}, \quad i=1, \ldots, N \tag{9.22}
\end{equation*}
$$

is assumed, where $\underset{\sim}{y}$ is $T X I$ vector of observations on the dependent variable, $X_{\sim}$ is a $\operatorname{Txp}$ matrix of observations (with rank $p$ ) on $p$ independent variables, $\underset{\sim}{\beta}$ is a pxl vector of parameters for $i^{\text {th }}$ element and $\underset{\sim}{e}{ }_{i}$ is the TXI vector of errors with zero mean for each i. Suppose now that $n$ elements are selected by using a specified sampling design, without replacement and the same $n$ elements observed in $T$ successive periods. Therefore, the observations on the $n$ elements for $T$ periods obey

$$
\begin{equation*}
\underset{\sim}{y}{ }_{i}=\underset{\sim}{X_{i}}{\underset{\sim}{i}}^{i}+\underset{\sim}{e}, \quad i=1, \ldots, n . \tag{9.23}
\end{equation*}
$$

Porter chooses $\quad \beta=\frac{1}{N} \sum_{1}^{N} \beta_{i}$ as the parameter vector of interest (or weighted average $\left.\Sigma w_{i} \beta_{i}\right)$.

The following assumptions are made: (1) $n>p, T>p$; (2) independent
 independently distributed and variance-covariance matrix of $\underset{\sim}{e}=\sigma_{i}^{2} I_{\sim}^{T}$. Let $b_{\sim}=\left(\underset{\sim}{x}{\underset{\sim}{i}}_{x_{i}}^{x_{i}}\right)^{-1} \underset{\sim}{x}{\underset{\sim}{\prime}}_{\sim}^{y}$ be the least-squares estimator of $\underset{\sim}{\beta} i^{\prime}$ so that given
the sample $s, E\left(b_{i} \mid s\right)={\underset{\sim}{i}}_{\beta}$. Unbiased estimator of $\underset{\sim}{\beta}$ chosen is the HorvitzThompson estimator

$$
\begin{equation*}
\hat{\hat{b}}=\frac{2}{\hat{N}} \frac{n}{I} \frac{{\underset{\sim}{i}}^{T_{i}}}{\pi_{i}} \tag{9.24}
\end{equation*}
$$

and an unbiased estimator of its variance-covariance matrix is
 (9.25) is of the Horvitz-Thompson type* and a 'better variance-matrix estimator of the Yates-Grundy form is obtained by replacing the first two terms in brackets of (9.25) by

$$
\begin{equation*}
\sum_{i<j} \sum_{i}^{\left(\pi_{i} \pi_{j}-\pi_{i j}\right)} \pi_{i j}\left(\frac{{\underset{\sim}{i}}^{\pi_{i}}}{\pi_{i}}-\frac{{\underset{\sim}{b}}_{j}}{\pi_{j}}\right)\left(\frac{{\underset{\sim}{b}}_{i}}{\pi_{i}}-\frac{{\underset{\sim}{b}}_{i}}{\pi_{j}}\right) \tag{9.26}
\end{equation*}
$$

An approximate Aitken-type estimator of $\beta$ which uses more of the model than (9.24) is also given (second part of assumption (2) is needed for this). Assumption (3) is also relaxed by letting

$$
E\left(e_{\sim}^{e} e_{\sim}^{\prime}\right)=\sigma_{i j} I_{\sim}^{T}, \quad i \neq j .
$$

## REFERENCES

[1] QUENOUILLE, M.H. (1956). Notes on bias in estimation. Biometrika, 43, 353-60.

[^0][2] RAO, P.S.R.S. \& RAO, J.N.K. (1971). Small sample results for ratio estimators. Biometrika, 58, 625-30.
[3] RAO, J.N.K. (1969). Ratio and regression estimators. In New Developments in Survey Sampling, 213 - 34. Wiley-Inter Science.
[4] RAO, J.N.K. \& KUZIK, R. (1974). Sampling errors in ratio estimation. Sankhya, Series C. 1.
[5] JONES, H.L. (1974). Jackknife estimation of functions of strata means. Biometrika, 61, 343-8.
[6] McCARTHY, P.J. (1966). Replication: an approach to the analysis of data from complex surveys. NCHS report, Series 2, No. 14, U.S. Dept. of Health, Education and Welfare.
[7] LEE, K.H. (1973). Variance estimation in stratified sampling. JASA, 68, 336-42.
[8] KISH, L. \& FRANKEL, M.R. (1970). Balanced repeated replication for standard errors. JASA, 65, 1071-94.
[9] WOODRUFF, R.S. (1971). A simple method for approximating the variance of a complicated estimate. JASA, 66, 411-4.
[10] FRANKEL, M.R. (1972). Inference from complex survey samples. Proc. Soc. Statist. Sec., ASA, 20 - 6.
[11] KISH, L. \& FRANKEL, M.R. (1974). Inference from complex samples (with discussion) . JRSS, Series B, 36, 1 - 37.
[12] BREWER, K.R.W. \& MELIOR, R.W. (1972). The effect of sample structure on analytical surveys. Paper presented at IASS meetings, Vienna (appeared in Australian J. Statist., 1974).
[13] WAKIMOTO, K. (1971). Stratified random sampling III: estimation of the correlation coefficient. Ann. Inst. Statist. Math. (Tokyo), 23, 339 - 53.
[14] KOOP, J.C. (1970). Estimation of correlation for a finite universe. Metrika, 15, 105 - 9.
[15] COCHRAN, W.G. (1973). Sampling Techniques. Wiley.
[16] TEPPING, B. (1968). Variance estimation in complex surveys. Proc. Soc. Statist. Sec., ASA, 11 - 8
[17] KONIJN, H.S. (1962). Regression analysis in sample surveys. JASA, 57, 590-606.
[18] FULLER, W. (1972). Regression analysis for sample surveys. Paper presented at the IASS meetings, Vienna (also to appear in Sankhya, Series C).
[19] PORTER, R.D. (1973). On the use of survey sample weights in the linear model. Annals of Economic and Social Measurement, 2, 141 - 58 (Porter is an economist working for the Board of Governors, Federal Reserve System, Washington, D.C.).

## ACKNOWLEDGMEN

I wish to thank: A. Ashraf and M. Hidiroglou for their help in the preparation of these notes; $R$. Platek and M.P. Singh for their interest in this project.

DATE DUE DATE DE RETOUR

I.OWI-MARTIN No. 1137
June 1975 Volume 1 Number 1
A Journal produced by Household Surveys Development Division, Statistical Services Field, Statistics Canada.
CONTENTS
Reinterview Programs and Response Errors
R. PLATEK and P.F. TIMMONS ..... 1
A Strategy for Up-dating Continuous Surveys
R. PLATEK and M.P. SINGH ..... 16
Components of Variance Model in Multi-Stage Stratified Samples
G.B. GRAY ..... 27
Non-Interview Patterns in the Canadian Labour Force Survey
R. SUGAVANAM ..... 44
A Comparison of Some Binomial Factors for the Labour Force Survey
M. LAWES ..... 59
Some Estimators for Domain Totals
M.P. SINGH and R. TESSIER ..... 74
Sample Design of the Family Expenditure Survey (1974)
M. LAWES and G.B. GRAY ..... 87
A Computer Algorithm for Joint Probabilities of Selection
M.A. HIDIROGLOU and G.B. GRAY ..... 99
The Development of an Automated Estimation System
A. SATIN and A. HARLEY ..... 109
A Journal produced by Household Surveys Development Division, Statistical Services Field, Statistics Canada.
CONTENTS
Controlled Random Rounding
I.P. FELLEGI ..... 123
On A Ratio Estimate With Post-Stratified Weighting G.B. GRAY and P.D. GHANGURDE ..... 134
Measurement of Response Errors in Censuses and Sample Surveys G.J. BRACKSTONE, J.F. GOSSELIN and B.E. GARTON. ..... 144
The Telephone Experiment in the Canadian Labour Force Survey R.C. MUIRHEAD, A.R. GOWER and F.T. NEWTON ..... 158
On the Improvement of Sample Survey Estimates $V$. TREMBLAY ..... 181
Some Variance Estimators for Multistage Sampling G.B. GRAY, M.A. HIDIROGLOU and M. CAIRNS ..... 197
The Methodology of the Canadian Travel Survey, 1971 A. ASHRAF ..... 208
Methods Test Panel Phase II - Data Analysis R. TESSIER ..... 228
Estimation of Process Average in Attribute Sampling Plans P.D. GHANGURDE ..... 244
December 1975 Volume 1 Supplementary Issue
CONTENTS
Analytic Studies of Sample Survey Data

J.N.K. RAO


[^0]:    * This variance estimator takes negative values 'often' unlike the Yates-Grundy variance estimator.

