ANALYSIS OF CATEGORICAL DATA OBTAINED BY STRATIFIED RANDOM SAMPLING I.

by

E. Sobel, M. Francis and P. Imrey*

Department of Biostatistics
University of North Carolina at Chapel Hill

Institute of Statistics Mimeo Series No. 1065

March 1976

ANALYSIS OF CATEGORICAL DATA OBTAINED BY STRATIFIED RANDOM SAMPLING I.

E. Sobel, M. Francis and P. Imrey*

*E. Sobel is Research Fellow, Department of Epidemiology, and M. Francis is Assistant Professor, Department of Biostatistics, University of North Carolina, Chapel Hill, N. C. P. Imrey is Assistant Professor of Biostatistics, University of Illinois, Urbana, Ill.

ABSTRACT

A non-iterative method of analysis by linear models is extended to categorical data obtained by stratified random sampling. It is shown that, asymptotically, proportional allocation reduces the variance of estimators over that obtained by simple random sampling. The difference between the asymptotic covariance matrices of estimated parameter vectors obtained by simple random sampling and stratified random sampling with proportional allocation is shown to be positive definite under fairly non-restrictive conditions.

1. INTRODUCTION

Grizzle, Starmer and Koch (GSK, [4]) propose a weighted least-squares methodology for the analysis of a wide range of linear and non-linear models for categorical data, including testing of linear hypotheses within the underlying parameter space. Their approach is a synthesis of the work of Wald [11], Neyman [10] and Bhapkar [1,2]. The broad applicability of the method has been demonstrated in a continuing series of papers, e.g., Koch and Reinfurt [9], Forthofer and Koch [3], Koch, Johnson and Tolley [8], and Koch, Imrey and Reinfurt [7].

The research described above assumes throughout that the categorical data were obtained by independent simple random sampling within each of an arbitrary number of very large populations. Koch, Freeman and Freeman [6] discuss alternative sample designs for which this assumption may not be necessary due to the use of pseudo-replication or like methods for estimating the covariance matrix of the sample cell counts. Johnson and Koch [5], in the context of an example, show how proportions or mean scores from stratified random sampling designs may be analyzed within the original GSK context, through consideration of each stratum as a population and use of a "mediator matrix" to transform within-stratum cell counts or mean scores into the usual stratified estimates and their variances and covariances. Linear models are then applied to the statistics so derived.

In this paper we extend the suggestion of Johnson and Koch [5] to the full range of functional models described in GSK [4]. We then provide a multivariate extension of the well-known theoretic result that

stratified sampling with proportional allocation provides an unbiased estimator of a population mean with smaller variance than the estimator obtained from simple random sampling. In particular we then show that this result carries over to stratified and non-stratified estimators of the parameters in the functional models mentioned above.

2. NOTATION and the PROBLEM

As far as possible, we follow the notation used in GSK. We assume that there are s populations of interest. Within the i^{th} population we are interested in a multinomial random (response) variable with response categories $C_{i1}, \ldots, C_{ir_i}$. Notice that we allow the number of categories to vary between populations. Let

$$\pi_{ij} = \text{proportion of the } i^{th} \text{ population "belonging" to } C_{ij}$$

$$\underset{\sim}{\pi}_i' = (\pi_{i1}, \cdots, \pi_{ir_i})$$

$$\underset{\sim}{\pi}' = (\underset{\sim}{\pi}_1', \cdots, \underset{\sim}{\pi}_s')$$

The unknown π_{ij} are hypothesized to satisfy a possibly non-linear model of the form

$$\underset{u\times 1}{\underset{\sim}{F}}(\pi) = \underset{u\times v}{\underset{\sim}{X}}\ \underset{v\times 1}{\underset{\sim}{\beta}} \qquad (1)$$

where

i) $\underset{\sim}{F}(\pi)' = (f_1(\underset{\sim}{\pi}), \cdots, f_u(\underset{\sim}{\pi}))$,

ii) $u \leq \Sigma(r_i - 1)$,

iii) each $f_m(\underset{\sim}{\pi})$ has continuous second partials with respect to the π_{ij},

iv) the $f_m(\underset{\sim}{\cdot})$ are independent of one another, i.e., the rank of

$$\underset{u\times\Sigma r_i}{\underset{\sim}{H}}(\underset{\sim}{\pi}) = \left(\frac{\partial f_m(\underset{\sim}{\pi})}{\partial \pi_{ij}}\right)$$

is equal to u,

v) $\underset{\sim}{X}$ is a known u x v design matrix of rank $v \leq u$,

vi) $\underset{\sim}{\beta}$ is a v x 1 vector of unknown parameters.

Assuming simple random sampling from each population, GSK treated (1) as a null hypothesis and presented an asymptotic χ^2-test. Then, assuming (1) is valid, they gave a best asymptotically normal

(BAN) estimate of $\underset{\sim}{\beta}$. Finally they gave an asymptotic χ^2-test for the null hypothesis

$$\underset{dxv}{C}\ \underset{\sim}{\beta} = \underset{\sim}{0} \qquad , (2)$$

where

vii) $\underset{\sim}{C}$ is a known d x v matrix of rank d $\leq$ v.

A. NOTATION - Simple Random Sampling

Within the i^{th} population, i = 1, $\cdots$, s, let

$n_{i.}$ = total sample size

n_{ij} = number of C_{ij} responses, j = 1, $\cdots$, r_i, in the sample

$p_{ij} = n_{ij}/n_{i.}$

$\underset{\sim}{p}_i' = (p_{i1}, \cdots, p_{ir_i})$

$$\underset{r_i \times r_i}{\text{var}}(\underset{\sim}{p}_i) = \underset{\sim}{V}(\underset{\sim}{\pi}_i) = \frac{1}{n_{i.}} \begin{pmatrix} \pi_{i_1}(1-\pi_{i1}) - \pi_{i1}\pi_{i2} & \cdots & -\pi_{i1}\pi_{ir_i} \\ \vdots & & \\ -\pi_{ir_i}\pi_{i1} & -\pi_{ir_i}\pi_{i2} & \cdots \pi_{ir_i}(1-\pi_{ir_i}) \end{pmatrix}.$$

Also let

$$\underset{1 \times \Sigma r_i}{\underset{\sim}{p}'} = (\underset{\sim}{p}_1', \cdots, \underset{\sim}{p}_s')$$

$$\underset{\Sigma r_i \times \Sigma r_i}{\underset{\sim}{V}(\underset{\sim}{\pi})} = \text{block diagonal matrix of the } \underset{\sim}{V}(\underset{\sim}{\pi}_i)$$

$$\underset{\sim}{V}(\underset{\sim}{p}) = \text{usual sample estimate of } \underset{\sim}{V}(\underset{\sim}{\pi})$$

$$\underset{\sim}{H} = \underset{\sim}{H}(\underset{\sim}{p})$$

$$\underset{\sim}{S} = \underset{\sim}{H}\underset{\sim}{V}(\underset{\sim}{p})\underset{\sim}{H}'$$

$$\underset{u \times u}{\underset{\sim}{S}}(\underset{\sim}{\pi}) = \underset{\sim}{H}(\underset{\sim}{\pi})\ \underset{\sim}{V}(\underset{\sim}{\pi})\ \underset{\sim}{H}(\underset{\sim}{\pi})'$$

B. NOTATION - Stratified Random Sampling

Within the i^{th} population, $i = 1, \cdots, s$, let

$n_{i.}$ = total sample size

D_i = number of strata

$n_{i(h)}$ = sample size in stratum h

$n_{i(h)j}$ = number of C_{ij} responses in the sample from stratum h, $j = 1, \cdots, r_i$

$N_{i.}$ = total population size

N_{ih} = population size of stratum h

$W_{ih} = N_{ih}/N_{i.}$

$\underset{\sim}{W}_i' = (W_{i1}, \cdots, W_{iD_i})$

$\alpha_{i(h)j}$ = proportion in stratum h "belonging" to C_{ij}

$a_{i(h)j} = n_{i(h)j}/n_{i(h)}$

$$\underset{\sim}{\alpha}'_{i(h)} = (\alpha_{i(h)1}, \cdots, \alpha_{i(h)r_i}); \quad \underset{\sim}{a}'_{i(h)} = (a_{i(h)1}, \cdots, a_{i(h)r_i})$$

$$\underset{1 \times D_i r_i}{\underset{\sim}{\alpha}'_i} = (\underset{\sim}{\alpha}'_{i(1)}, \cdots, \underset{\sim}{\alpha}'_{i(D_i)}); \quad \underset{1 \times D_i r_i}{\underset{\sim}{a}'_i} = (\underset{\sim}{a}'_{i(1)}, \cdots, \underset{\sim}{a}'_{i(D_i)})$$

$$\underset{r_i \times r_i}{\underset{\sim}{\mathrm{var}}}(\underset{\sim}{a}_{i(h)}) = \underset{\sim}{V}(\underset{\sim}{\alpha}_{i(h)})$$

$$= \frac{1}{n_{i(h)}} \begin{pmatrix} \alpha_{i(h)1}(1 - \alpha_{i(h)1}) - \alpha_{i(h)1}\,\alpha_{i(h)2} \cdots - \alpha_{i(h)1}\,\alpha_{i(h)r_i} \\ \\ - \alpha_{i(h)r_i}\,\alpha_{i(h)1} - \alpha_{i(h)r_i}\,\alpha_{i(h)2} \cdots \alpha_{i(h)r_i}(1 - \alpha_{i(h)r_i}) \end{pmatrix}$$

$\underset{D_i r_i \times D_i r_i}{\underset{\sim}{V}(\underset{\sim}{\alpha}_i)}$ = block diagonal matrix of the $\underset{\sim}{V}(\underset{\sim}{\alpha}_{i(h)})$.

Also let

$$\underset{1 \times \Sigma D_i r_i}{\underset{\sim}{\alpha}'} = (\underset{\sim}{\alpha}_1', \cdots, \underset{\sim}{\alpha}_s') \quad ; \quad \underset{1 \times \Sigma D_i r_i}{\underset{\sim}{a}'} = (\underset{\sim}{a}_1', \cdots, \underset{\sim}{a}_s')$$

$\underset{\Sigma D_i r_i \times \Sigma D_i r_i}{\underset{\sim}{V}(\underset{\sim}{\alpha})}$ = block diagonal matrix of the $\underset{\sim}{V}(\underset{\sim}{\alpha}_i)$

$\underset{\sim}{V}(\underset{\sim}{a})$ = usual sample estimate of $\underset{\sim}{V}(\underset{\sim}{\alpha})$

$$\underset{\Sigma r_i \times \Sigma D_i r_i}{\underset{\sim}{Q}} = \begin{pmatrix} \underset{\sim}{W}_1' \otimes \underset{r_1 \times r_1}{\underset{\sim}{I}} & & \underset{\sim}{0} \\ & \ddots & \\ \underset{\sim}{0} & & \underset{\sim}{W}_s' \otimes \underset{r_s \times r_s}{\underset{\sim}{I}} \end{pmatrix}$$

where $\otimes$ is the Kronecker product and $\underset{r_i \times r_i}{\underset{\sim}{I}}$ is the $r_i \times r_i$ identity matrix. Finally let

$$\underset{\sim}{S}_{st} = \underset{\sim}{H}_{st} \underset{\sim}{V}(\underset{\sim}{a}) \underset{\sim}{H}_{st}'$$

$$\underset{\sim}{S}_{st}(\underset{\sim}{\alpha}) = \underset{\sim}{H}_{st}(\underset{\sim}{\alpha}) \underset{\sim}{V}(\underset{\sim}{\alpha}) \underset{\sim}{H}_{st}(\underset{\sim}{\alpha})'$$

where $\underset{\sim}{H}_{st}$ and $\underset{\sim}{H}_{st}(\underset{\sim}{\alpha})$ are defined in (3) in Section 4.

We remark here that even though $\underset{\sim}{V}(\underset{\sim}{\pi})$ and $\underset{\sim}{V}(\underset{\sim}{\alpha})$ are singular and thus only positive semi-definite, both $\underset{\sim}{S}(\underset{\sim}{\pi})$ and $\underset{\sim}{S}_{st}(\underset{\sim}{\alpha})$ are non-singular and thus positive definite as a consequence of assumption iv) following equation (1). We shall use the positive definiteness in the proof of the theorem in Section 6.

3. UNDER SIMPLE RANDOM SAMPLING

GSK pointed out that if (1) is valid, then

$$\underset{\sim}{SS}[\underset{\sim}{F}(\underset{\sim}{\pi}) = \underset{\sim}{X}\underset{\sim}{\beta}] = (\underset{\sim}{F}(\underset{\sim}{p}) - \underset{\sim}{X}\underset{\sim}{b})'\underset{\sim}{S}^{-1}(\underset{\sim}{F}(\underset{\sim}{p}) - \underset{\sim}{X}\underset{\sim}{b})$$

is asymptotically a $\chi^2(u-v)$ where

$$\underset{\sim}{b} = (\underset{\sim}{X}'\underset{\sim}{S}^{-1}\underset{\sim}{X})^{-1}\underset{\sim}{X}'\underset{\sim}{S}^{-1}\underset{\sim}{F}(\underset{\sim}{p}).$$

This provides a test of (1). Note that if u = v then (1) must trivially be true. If (1) is true the b is a BAN estimation of $\underset{\sim}{\beta}$. GSK suggested that to test (2) one use

$$\underset{\sim}{SS}[\underset{\sim}{C}\underset{\sim}{\beta} = \underset{\sim}{0}] = \underset{\sim}{b}'\underset{\sim}{C}'(\underset{\sim}{C}(\underset{\sim}{X}'\underset{\sim}{S}^{-1}\underset{\sim}{X})^{-1}\underset{\sim}{C}')^{-1}\underset{\sim}{C}\underset{\sim}{b}$$

which, if (2) is true, is asymptotically a $\chi^2(d)$. Note that if (1) is true and $\underset{\sim}{F}(\underset{\sim}{\pi})$ is linear, $\underset{\sim}{S}(\underset{\sim}{\pi})$ is the covariance matrix of $\underset{\sim}{F}(\underset{\sim}{p})$ and that if $\underset{\sim}{F}(\underset{\sim}{\pi})$ is non-linear

$$\underset{\sim}{\text{var}}\ (\underset{\sim}{F}(\underset{\sim}{p}))\underset{\sim}{S}^{-1}(\underset{\sim}{\pi}) \to \underset{\sim}{I}$$

as the $n_i \to \infty$ so that $n_{i(h)}/n_i$ is constant. Similarly,

$$\underset{\sim}{S}\underset{\sim}{S}^{-1}(\underset{\sim}{\pi}) \to \underset{\sim}{I},$$

and $(\underset{\sim}{X}'\underset{\sim}{S}^{-1}\underset{\sim}{X})^{-1}$ is asymptotically equal to the covariance matrix of $\underset{\sim}{b}$.

4. UNDER STRATIFIED RANDOM SAMPLING

In order to test (1), obtain a BAN estimator of $\underset{\sim}{\beta}$ and test (2) under stratified random sampling we basically treat this situation from the point of view of simple random sampling from $\sum_1^s D_i$ populations. Without loss of generality, we assume

$$\alpha_{i(h)j} > 0, \text{ for each } h, j, i.$$

Since

$$\pi_{ij} = \sum_{h=1}^{D_i} W_{ih}\alpha_{i(h)j}$$

we have

$$\underset{\sim}{Q}\,\underset{\sim}{\alpha} = \underset{\sim}{\pi}$$

and thus if (1) is true

$$\underset{\sim}{T}(\underset{\sim}{\alpha}) = \underset{\sim}{F}(\underset{\sim}{Q}\underset{\sim}{\alpha}) = \underset{\sim}{X}\,\underset{\sim}{\beta}. \qquad (1')$$

Note that this differs from Johnson and Koch [5], who calculate functions within each stratum before pooling using stratum weights. Such an approach is inappropriate if $\underset{\sim}{F}(\cdot)$ is non-linear, while it produces results identical to ours in the linear case.

It is clear that, with $\underset{\sim}{T}(\cdot) \equiv \underset{\sim}{F}(\underset{\sim}{Q}\cdot)$, (1) and (1′) are equivalent. Next we show that the conditions corresponding to iii) and iv) of section 2 are valid for

$$\underset{\sim}{T}(\underset{\sim}{\alpha}) = (t_1(\underset{\sim}{\alpha}), \ldots, t_u(\underset{\sim}{\alpha}))' .$$

We have

$$\frac{\partial t_m(\underset{\sim}{\alpha})}{\partial \alpha_{i(h)j}} = \frac{\partial f_m(\underset{\sim}{G}\underset{\sim}{\alpha})}{\partial \alpha_{i(h)j}} = \frac{\partial f_m(\underset{\sim}{\pi})}{\partial \pi_{ij}} \frac{\partial \pi_{ij}}{\partial \alpha_{i(h)j}} = W_{ih} \frac{\partial f_m(\underset{\sim}{\pi})}{\partial \pi_{ij}}$$

and

$$\frac{\partial^2 t_m(\underset{\sim}{\alpha})}{\partial \alpha_{i(h)j} \partial \alpha_{i'(h')j'}} = \frac{\partial [W_{ih} \partial f_m(\underset{\sim}{G}\underset{\sim}{\alpha})/\partial \pi_{ij}]}{\partial \alpha_{i'(h')j'}}$$

$$= W_{ih} W_{i'h'} \frac{\partial^2 f_m(\underset{\sim}{\pi})}{\partial \pi_{ij} \partial \pi_{i'j'}} .$$

Thus each $t_m(\alpha)$ has continuous second partials. Let

$$\underset{u \times \Sigma D_i r_i}{\underset{\sim}{H}_{st}(\underset{\sim}{\alpha})} = \left(\frac{\partial t_m(\underset{\sim}{\alpha})}{\partial \alpha_{i(h)j}} \right)$$

and

$$\underset{\sim}{H}_{st} = \underset{\sim}{H}_{st}(\underset{\sim}{a}) = \left(\frac{\partial t_m(\underset{\sim}{\alpha})}{\partial \alpha_{i(h)j}} \,\middle|\, \underset{\sim}{\alpha} = \underset{\sim}{a} \right)$$

Then

$$\underset{\sim}{H}_{st}(\underset{\sim}{\alpha}) = \begin{pmatrix} \left(\frac{\partial f_1(\underset{\sim}{\pi})}{\partial \underset{\sim}{\pi}_1}\right)' W_{11} & \cdots & \left(\frac{\partial f_1(\underset{\sim}{\pi})}{\partial \underset{\sim}{\pi}_1}\right)' W_{1D_1} & \cdots & \left(\frac{\partial f_1(\underset{\sim}{\pi})}{\partial \underset{\sim}{\pi}_s}\right)' W_{s1} & \cdots & \left(\frac{\partial f_1(\underset{\sim}{\pi})}{\partial \underset{\sim}{\pi}_s}\right)' W'_{sD_s} \\ \vdots & & & & & & \vdots \\ \left(\frac{\partial f_u(\underset{\sim}{\pi})}{\partial \underset{\sim}{\pi}_1}\right)' W_{11} & & \cdots & & & & \left(\frac{\partial f_u(\underset{\sim}{\pi})}{\partial \underset{\sim}{\pi}_s}\right)' W'_{sD_s} \end{pmatrix}$$

$$= \begin{pmatrix} \underset{\sim}{W}'_1 \otimes \left(\frac{\partial f_1(\underset{\sim}{\pi})}{\partial \underset{\sim}{\pi}_1}\right)' & \cdots & \underset{\sim}{W}'_s \otimes \left(\frac{\partial f_1(\underset{\sim}{\pi})}{\partial \underset{\sim}{\pi}_s}\right)' \\ \vdots & & \\ \underset{\sim}{W}'_1 \otimes \left(\frac{\partial f_u(\underset{\sim}{\pi})}{\partial \underset{\sim}{\pi}_1}\right)' & \cdots & \underset{\sim}{W}'_s \otimes \left(\frac{f_u(\underset{\sim}{\pi})}{\partial \underset{\sim}{\pi}_s}\right)' \end{pmatrix}$$

$$= \underset{\sim}{H}(\underset{\sim}{\pi})\underset{\sim}{Q}.$$

Since $\underset{\sim}{Q}$ has only one non-zero entry in each column and is $\Sigma r_i \times \Sigma D_i r_i$,

$$\mathrm{rk}\ \underset{\sim}{Q} = \Sigma r_i .$$

Thus

$$\mathrm{rk}\ \underset{\sim st}{H}(\underset{\sim}{\alpha}) = u .$$

Consequently

$$\underset{\sim}{SS}[\underset{\sim}{T}(\underset{\sim}{\alpha}) = \underset{\sim}{X}\underset{\sim}{\beta}] = (\underset{\sim}{T}(\underset{\sim}{a}) - \underset{\sim}{X}\underset{\sim st}{b})\underset{\sim st}{S^{-1}}(\underset{\sim}{T}(\underset{\sim}{a}) - \underset{\sim}{X}\underset{\sim st}{b})$$

is asymptotically $\chi^2(u-v)$ where

$$\underset{\sim st}{b} = (\underset{\sim}{X}'\underset{\sim st}{S^{-1}}\underset{\sim}{X})^{-1}\underset{\sim}{X}'\underset{\sim st}{S^{-1}}\underset{\sim}{T}(\underset{\sim}{a}).$$

Thus if (1′) is valid, $\underset{\sim st}{b}$ is a BAN estimator of $\underset{\sim}{\beta}$. The test suggested by GSK for (2) then uses

$$\underset{\sim}{SS}[\underset{\sim}{C}\underset{\sim}{\beta} = \underset{\sim}{0}] = \underset{\sim st}{b}'\underset{\sim}{C}'(\underset{\sim}{C}(\underset{\sim}{X}'\underset{\sim st}{S^{-1}}\underset{\sim}{X})^{-1}\underset{\sim}{C}')^{-1}\underset{\sim st}{Cb}$$

which is asymptotically $\chi^2(d)$ if (2) is true.

5. COMPARISON OF $\underset{\sim}{b}$ and $\underset{\sim st}{b}$ with PROPORTIONAL ALLOCATION

We now assume that

$$n_{i(h)} = W_{ih}n_{i\cdot},\ h = 1, \cdots, D_i,\ i = 1, \cdots, s.$$

Then the usual estimate of π_{ij} is

$$\hat{\pi}_{ij} = \sum_{h=1}^{D_i} W_{ih}\, a_{i(h)j}$$

Put $\underset{\sim}{\hat{\pi}}_i' = (\hat{\pi}_{i1}, \cdots, \hat{\pi}_{ir_i})$ and $\underset{1 \times \Sigma r_i}{\underset{\sim}{\hat{\pi}}'} = (\underset{\sim}{\hat{\pi}}_1', \cdots, \underset{\sim}{\hat{\pi}}_s')$. Then

$$\underset{\sim}{p}_{st} = \underset{\sim}{\hat{\pi}} = \underset{\sim}{Q}\underset{\sim}{a}$$

and

$$\underset{\sim}{\text{var}}(\underset{\sim}{\hat{\pi}}) = \underset{\sim}{\text{var}}\,(\underset{\sim}{p}_{st}) = (\text{cov}(\hat{\pi}_{ij}, \hat{\pi}_{i'j'}))$$

$$= \underset{\sim}{Q}\underset{\sim}{V}(\alpha)\underset{\sim}{Q}',$$

Also, in this case, since

$$\underset{\sim}{H}_{st}(\underset{\sim}{\alpha}) = \underset{\sim}{H}(\underset{\sim}{\pi})\underset{\sim}{Q},$$

$$\underset{\sim}{S}_{st}(\underset{\sim}{\alpha}) = \underset{\sim}{H}(\underset{\sim}{\pi})\underset{\sim}{\text{var}}(\underset{\sim}{\hat{\pi}})\underset{\sim}{H}(\underset{\sim}{\pi})'$$

and

$$\underset{\sim}{S}_{st} = \underset{\sim}{H}\,\underset{\sim}{V}(\underset{\sim}{p}_{st})\,\underset{\sim}{H}'$$

where

$$\underset{\sim}{V}(\underset{\sim}{p}_{st}) = \text{usual estimate of } \underset{\sim}{\text{var}}(\underset{\sim}{p}_{st}).$$

As a notational convenience, $\underset{\sim}{A} \leq \underset{\sim}{B}$ and $\underset{\sim}{A} < \underset{\sim}{B}$, where $\underset{\sim}{A}$ and $\underset{\sim}{B}$ are square matrices, means that $\underset{\sim}{B}-\underset{\sim}{A}$ is positive semi-definite and positive definite, respectively. $\geq$ and $>$ have equivalent meanings.

Lemma

$$\underset{\sim}{\text{var}}\ (\hat{\underset{\sim}{\pi}}) \leq \underset{\sim}{V}(\underset{\sim}{\pi}).$$

Proof:

First note that both $\underset{\sim}{\text{var}}\ (\hat{\underset{\sim}{\pi}})$ and $\underset{\sim}{V}(\underset{\sim}{\pi})$ are composed of (square) block diagonal matrices:

$$\underset{\sim}{V}\ (\underset{\sim}{\pi}) = \begin{pmatrix} \underset{\sim}{V}\ (\underset{\sim}{\pi}_1) & \cdots & \underset{\sim}{0} \\ \vdots & \ddots & \vdots \\ \underset{\sim}{0} & \cdots & \underset{\sim}{V}\ (\underset{\sim}{\pi}_s) \end{pmatrix}$$

and

$$\underset{\sim}{\text{var}}(\hat{\underset{\sim}{\pi}}) = \begin{pmatrix} \underset{\sim}{\text{var}}(\hat{\underset{\sim}{\pi}}_1) & \cdots & \underset{\sim}{0} \\ \vdots & \ddots & \vdots \\ \underset{\sim}{0} & \cdots & \underset{\sim}{\text{var}}(\hat{\underset{\sim}{\pi}}_s) \end{pmatrix}$$

Thus if we show that

$$\underset{\sim}{\text{var}}(\hat{\underset{\sim}{\pi}}_i) \leq \underset{\sim}{V}(\underset{\sim}{\pi}_i),\ i = 1, \cdots, s$$

we are done. But

$$n_{i\cdot}\left(\underset{\sim}{V}(\underset{\sim}{\pi}_i) - \underset{\sim}{\mathrm{var}}(\underset{\sim}{\hat{\pi}}_i)\right) =$$

$$\begin{pmatrix} \sum_h W_{ih}(\alpha_{i(h)1} - \pi_{i1})^2 & \cdots & \sum_h W_{ih}(\alpha_{i(h)1} - \pi_{i1})(\alpha_{i(h)r_i} - \pi_{ir_i}) \\ \vdots & & \vdots \\ \sum_h W_{ih}(\alpha_{i(h)r_i} - \pi_{ir_i})(\alpha_{i(h)1} - \pi_{i1}) & \cdots & \sum_h W_{ih}(\alpha_{i(h)r_i} - \pi_{ir_i})^2 \end{pmatrix}$$

Now let

$$\underset{\sim}{X}_i' = (X_{i1}, \ldots, X_{ir_i})$$

$$X_{ij} = \alpha_{i(h)j}, \quad j = 1, \ldots, r_i$$

where h is a randomly chosen integer, $1 \le h \le D_i$, with

$$P(h = h_o) = W_{ih_o}.$$

Then

$$E(X_{ij}) = \sum_h W_{ih}\alpha_{i(h)j} = \pi_{ij}$$

$$\mathrm{var}(X_{ij}) = \sum_h W_{ih}(\alpha_{i(h)j} - \pi_{ij})^2$$

and

$$\mathrm{cov}(X_{ij}, X_{ij'}) = \sum_h W_{ih}(\alpha_{i(h)j} - \pi_{ij})(\alpha_{i(h)j'} - \pi_{ij'})$$

Thus the covariance matrix of $\underset{\sim}{X}_i$ is $n_{i\cdot}\left(\underset{\sim}{V}(\underset{\sim}{\pi}_i) - \underset{\sim}{\mathrm{var}}(\underset{\sim}{\hat{\pi}}_i)\right)$, which is therefore positive semi-definite.

Theorem

Asymptotically, with proportional allocation,

$$\text{var } \underset{\sim}{c}'\underset{\sim}{b}_{st} \leq \text{var } \underset{\sim}{c}'\underset{\sim}{b}$$

for all vectors $\underset{\sim}{c} \neq \underset{\sim}{0}$. This is equivalent to

$$(\underset{\sim}{X}'\underset{\sim}{S}_{st}^{-1}\underset{\sim}{X})^{-1} \leq (\underset{\sim}{X}'\underset{\sim}{S}^{-1}\underset{\sim}{X})^{-1} .$$

Proof:

Since

$$\text{var } \underset{\sim}{c}'\underset{\sim}{b}_{st} = \underset{\sim}{c}'(\underset{\sim}{X}'\underset{\sim}{S}_{st}^{-1}\underset{\sim}{X})^{-1}\underset{\sim}{c}$$

$$\text{var } \underset{\sim}{c}'\underset{\sim}{b} = \underset{\sim}{c}'(\underset{\sim}{X}'\underset{\sim}{S}^{-1}\underset{\sim}{X})^{-1}\underset{\sim}{c}$$

the equivalence is clear. Now from the lemma and the fact that $\underset{\sim}{H}(\pi)$ has linearly independent rows it follows that

$$\underset{\sim}{S}(\pi) - \underset{\sim}{S}_{st}(\alpha) = \underset{\sim}{H}(\pi)[\underset{\sim}{V}(\pi)-\underset{\sim}{\text{var}}(\hat{\pi})]\underset{\sim}{H}'(\pi) \geq \underset{\sim}{0}.$$

Since $\underset{\sim}{S}_{st}(\alpha) > \underset{\sim}{0}$, its inverse possesses a (symmetric) square root $\underset{\sim}{P}_1(\alpha) = \underset{\sim}{P}_1$, say for convenience. Then $\underset{\sim}{P}_1\underset{\sim}{S}(\pi)\underset{\sim}{P}_1 > \underset{\sim}{0}$ and by the Principal Axis Theorem there exists an orthogonal matrix $\underset{\sim}{P}_2(\alpha) = \underset{\sim}{P}_2$, say, such that

$$\underset{\sim}{P}_2(\underset{\sim}{P}_1\underset{\sim}{S}(\pi)\underset{\sim}{P}_1)\underset{\sim}{P}_2' = \underset{\sim}{D},$$

where $\underset{\sim}{D}$ is diagonal of full rank. Thus, with $\underset{\sim}{P}(\alpha) = \underset{\sim}{P} = \underset{\sim}{P}_2\underset{\sim}{P}_1$,

$$\underset{\sim}{P}\underset{\sim}{S}(\pi)\underset{\sim}{P} = \underset{\sim}{D}$$

$$\underset{\sim}{P}\underset{\sim}{S}_{st}(\alpha)\underset{\sim}{P} = \underset{\sim}{P}_2\underset{\sim}{P}_2 = \underset{\sim}{I}.$$

But $\underset{\sim}{D}-\underset{\sim}{I} = \underset{\sim}{P}[\underset{\sim}{S}(\pi) - \underset{\sim}{S}_{st}(\alpha)]\underset{\sim}{P}' \geq \underset{\sim}{0}$ and therefore every diagonal element of $\underset{\sim}{D}$ is greater than or equal to 1. Let

$$\underset{\sim}{E} = \underset{\sim}{I} - \underset{\sim}{D}^{-1} :$$

Then

$$(\underset{\sim}{X}'\underset{\sim}{S}_{st}^{-1}(\alpha)\underset{\sim}{X})^{-1} = (\underset{\sim}{X}'\underset{\sim}{P}'\underset{\sim}{P}\underset{\sim}{X})^{-1} = [(\underset{\sim}{P}\underset{\sim}{X})'(\underset{\sim}{P}\underset{\sim}{X})]^{-1}$$

$$(\underset{\sim}{X}'\underset{\sim}{S}^{-1}(\pi)\underset{\sim}{X})^{-1} = (\underset{\sim}{X}'\underset{\sim}{P}'\underset{\sim}{D}^{-1}\underset{\sim}{P}\underset{\sim}{X})^{-1} = [(\underset{\sim}{P}\underset{\sim}{X})'(\underset{\sim}{I}-\underset{\sim}{E})(\underset{\sim}{P}\underset{\sim}{X})]^{-1}.$$

Put

$$\underset{\sim}{U}(\underset{\sim}{\delta}) = [(\underset{\sim}{P}\underset{\sim}{X})'(\underset{\sim}{I}-\underset{\sim}{E}(\underset{\sim}{\delta}))(\underset{\sim}{P}\underset{\sim}{X})]^{-1}$$

where

$$\underset{\sim}{\delta} \in \underset{\sim}{\sum} = \left\{\underset{\sim}{\delta}' = (\delta_1, \ldots, \delta_u) \mid 0 \le \delta_k < 1,\ k = 1, \ldots, u\right\}$$

$$\underset{\sim}{E}(\underset{\sim}{\delta}) = \text{diagonal matrix with } \underset{\sim}{\delta} \text{ as the diagonal vector}$$

Thus

$$\underset{\sim}{U}(\underset{\sim}{0}) = (\underset{\sim}{X}'\underset{\sim}{S}_{st}^{-1}\underset{\sim}{X})^{-1}$$

$$\underset{\sim}{U}(\underset{\sim}{\delta}_{\pi}) = (\underset{\sim}{X}'\underset{\sim}{S}^{-1}(\pi)\underset{\sim}{X})^{-1}$$

for some $\underset{\sim}{\delta}_{\pi} \in \underset{\sim}{\sum}$. We are then finished once we prove that $\underset{\sim}{x}'\underset{\sim}{U}(\underset{\sim}{\delta})\underset{\sim}{x}$, $\underset{\sim}{x} \neq \underset{\sim}{0}$, is a non-decreasing function of each δ_k for $\underset{\sim}{\delta} \in \underset{\sim}{\sum}$. But

$$\frac{\partial[\underset{\sim}{x}'\underset{\sim}{U}(\underset{\sim}{\delta})\underset{\sim}{x}]}{\partial\delta_k} = \underset{\sim}{x}'\left[\frac{\partial\underset{\sim}{U}(\underset{\sim}{\delta})}{\partial\delta_k}\right]\underset{\sim}{x}$$

$$= -\underset{\sim}{x}'\underset{\sim}{U}(\underset{\sim}{\delta})(\underset{\sim}{P}\underset{\sim}{X})'\left[\frac{\partial(\underset{\sim}{I}-\underset{\sim}{E}(\underset{\sim}{\delta}))}{\partial\delta_k}\right](\underset{\sim}{P}\underset{\sim}{X})\underset{\sim}{U}(\underset{\sim}{\delta})\underset{\sim}{x}$$

$$= [\underset{\sim}{U}(\delta)\underset{\sim}{x}]'\underset{\sim}{\theta}_k\underset{\sim}{\theta}_k' [\underset{\sim}{U}(\delta)\underset{\sim}{x}]$$

$$= [\underset{\sim}{\theta}_k'\underset{\sim}{U}(\delta)\underset{\sim}{x}]'[\underset{\sim}{\theta}_k'\underset{\sim}{U}(\delta)\underset{\sim}{x}]$$

$$\geq 0,$$

where $\underset{\sim}{\theta}_k$ is the k^{th} column of $\underset{\sim}{P}\underset{\sim}{X}$. This completes the proof.

Corollary

Let $\underset{\sim}{X}' = (\underset{\sim}{X}_1', \ldots, \underset{\sim}{X}_s')$ where $\underset{\sim}{X}_i$ is defined in the proof of the lemma. Let $\underset{\sim}{X}$ be a matrix whose $\prod_{i=1}^{s} D_i$ rows are the realizations of $\underset{\sim}{X}$. Then $\text{var } \underset{\sim}{c}'\underset{\sim}{b}_{st} < \text{var } \underset{\sim}{c}'\underset{\sim}{b}$ for all $\underset{\sim}{c} \neq \underset{\sim}{0}$ if and only if

$$\text{rank } [\underset{\sim}{\ell}, \underset{\sim}{X}\underset{\sim}{H}'] = u + 1,$$

where $\underset{\sim}{\ell}$ is a column vector of 1's.

Proof:

Since $\underset{\sim}{S}(\underset{\sim}{\pi}) - \underset{\sim}{S}_{st}(\underset{\sim}{\alpha})$ is the covariance matrix of $\underset{\sim}{H}\underset{\sim}{X}$ under the probability mass function

$$P(\underset{\sim}{X}' = (\underset{\sim}{\alpha}_{1(h_1)}', \ldots, \underset{\sim}{\alpha}_{s(h_s)}')) = \prod_{1}^{s} P(\underset{\sim}{X}_i = \underset{\sim}{\alpha}_{i(h_i)}) = \prod_{1}^{s} W_{i(h_i)}$$

the condition given is equivalent to $\underset{\sim}{S}(\underset{\sim}{\pi}) - \underset{\sim}{S}_{st}(\underset{\sim}{\alpha}) > \underset{\sim}{0}$. The corollary then follows directly from the proof of the theorem.

6. DISCUSSION

Under the model $\underset{\sim}{F}(\underset{\sim}{\pi}) = \underset{\sim}{X}\underset{\sim}{\beta}$, where $\underset{\sim}{F}(\cdot)$ satisfies the conditions given in Section 2, the estimator of $\underset{\sim}{\beta}$ under the least-squares algorithm behaves asymptotically as a linear function of the estimators of the cell probabilities (see [10]). Consequently if $\underset{\sim}{F}(\cdot)$ is such that its components, $f_k(\cdot)$, are derived from non-overlapping groups of the s populations, then the results in Section 5 are rather obvious. We have shown that this extends to rather arbitrary linear models.

The approach taken above ignores the finite population correction factor. However, as Johnson and Koch [5] remark, multiplication of $\underset{\sim}{V}(\underset{\sim}{p})$ by (1-f), where f is the sampling fraction, results in valid large-sample statistics based on the work of Wald [11] when the population is considered finite. If the fpc were considered here the basic results would clearly remain valid when all covariance matrices are multiplied by the appropriate $(1-f_i)$. This is easily seen by consideration of sequences of problems with stratum sizes N_{ih} and sample sizes $n_{i(h)}$, with both N_{ih} and $n_{i(h)}$ increasing such that their ratio approaches $f_i = n_{i.}/N_{i.}$. The resulting hypergeometric distributions obey the same central limit theory as their multinomial counterparts.

The weighted least-squares procedure is vulnerable to small sample problems due to observed zero cell counts. In the unstratified case if this happens S will not be of full rank because the corresponding π_{ij} will be estimated by 0. GSK [4] suggest using an "estimate" $1/r_i n_i$. However, in the stratified case the invertibility of $\underset{\sim}{S}_{st}$ depends upon the covariance matrix of the stratified estimates of the π_{ij}, not upon

the covariance matrix from the individual strata, i.e., the covariance matrix of the estimates of the $\alpha_{i(h)j}$ for fixed i, h. Thus, problems will be encountered when all strata within a given population have the same cell empty. Following GSK, each corresponding $\alpha_{i(h)j}$ can be estimated by $1/r_i n_{i(h)}$. Such an occurrence presumably would arise no more often than having a zero cell count using simple random sampling with identical total sample size.

Finally, the formulation and results in Sections 4 and 5 allow for a general study of optimal allocation as a function of costs associated with sampling from each stratum, the model $\underset{\sim}{F}(\underset{\sim}{\pi}) = \underset{\sim}{X}\underset{\sim}{\beta}$ and the variance-covariance matrix of the estimators. Results along these lines will be presented in a sequel.

REFERENCES

[1] Bhapkar, V.P., "Notes on Analysis of Categorical Data," Series No. 477, Institute of Statistics, University of North Carolina, 1966 (mimeographed).

[2] Bhapkar, V.P., "A Note on the Equivalence of Two Test Criteria for Hypotheses in Categorical Data," Journal of the American Statistical Association, 61 (March 1966) 228-35.

[3] Forthofer, Ronald N. and Koch, Gary G., "An Analysis for Compounded Functions of Categorical Data," Biometrics, 29 (March 1973), 143-57.

[4] Grizzle, James, E., Starmer, C. Frank and Koch, Gary G., "Analysis of Categorical Data by Linear Models," Biometrics, 25 (September 1969), 489-503.

[5] Johnson, William D. and Koch, Gary G., "Analysis of Qualitative Data: Linear Models," Health Sciences Research, (Winter 1970), 358-69.

[6] Koch, Gary G., Freeman, Daniel H. Jr., and Freeman, Jean L., "Strategies in the Multivariate Analysis of Data from Complex Surveys," International Statistical Review, 43 (No. 1, 1975), 59-78.

[7] Koch, Gary G., Imrey, Peter B., and Reinfurt, Donald W., "Linear Model Analysis of Categorical Data with Incomplete Response Vectors," Biometrics, 28 (September 1972), 663-92.

[8] Koch, Gary G., Johnson, William D., and Tolley, H. Dennis, "A Linear Models Approach to the Analysis of Survival and Extent of Disease in Multidimensional Contingency Tables," Journal of the American Statistical Association, 67 (December 1972), 783-96.

[9] Koch, Gary G. and Reinfurt, Donald W., "The Analysis of Categorical Data from Mixed Models," Biometrics, 27 (March 1971), 157-73.

[10] Neyman, Jenzy, "Contribution to the Theory of the χ^2 Test," in J. Neyman, Ed., Proceedings of the Berkeley Symposium on Mathematical Statistics and Probability, University of California Press, Berkeley and Los Angeles, 1949, 239-73.

[11] Wald, Abraham, "Tests of Statistical Hypotheses Concerning Several Parameters where the Number of Observations Is Large," Transactions of the American Mathematics Society, 54 (November 1943), 426-482.