SOME USEFUL STRATEGIES IN THE MULTIVARIATE ANALYSIS OF DATA FROM COMPLEX SURVEYS

Gary G. Koch, Daniel H. Freeman, and Jean L. Freeman, University of North Carolina

1. INTRODUCTION

For many large scale surveys like those conducted by the U.S. Bureau of the Census and the National Center for Health Statistics, data are obtained through complex designs often involving both clustering and stratification as well as multi-stage selection. Moreover, sub-population (or domain) characteristics are estimated by appropriate ratio methods. As a result, standard methods of multivariate analysis (which assume simple random samples) are not directly applicable. On the other hand, since sample sizes in such situations are usually very large, it generally can be assumed that the various estimates of sub-population characteristics do approximately have multivariate normal distributions with covariance matrices which can be consistently estimated by either direct or replication methods. Thus, a weighted least squares approach can be used to investigate various relationships among these estimates and test appropriate hypotheses. This paper is concerned with the application of this methodological strategy for analyses involving:

1. **comparisons among cross-classified sub-populations,**
2. **evaluations of the existence and nature of trends.**

2. METHODOLOGY

2.1. Definitions and Preliminaries.

Population: A set of N individuals indexed by the subscript $i = 1,2,\ldots,N$.

Sample Design: Probability random sample of size n with $\tau = 1$ trial for the measurements on each selected individual. As indicated by Cornfield [1], this sample design can be characterized by random variables U_i where

$$U_i = \begin{cases} 1 & \text{if population element i is in sample} \\ 0 & \text{otherwise} \end{cases}.$$

In this context, $\phi_i = E\{U_i\}$ represents the probability of selection for the i-th population element and $\theta_{ii'} = E\{U_i U_{i'}\}$ represents the probability for the joint selection of both the i-th and i'-th population elements. Unless stated otherwise, we shall usually assume that sampling is without replacement in which case $\theta_{ii} = E\{U_i^2\} = E\{U_i\} = \phi_i$; however, the $\theta_{ii'}$ for $i \neq i'$ must be determined by appropriate calculations which correspond directly to the specific method of selection -- although for some complex designs this can potentially involve very difficult mathematical problems.

Measurement Process:

a. Self-enumeration. In other words, each individual in the sample responds individually and independently to the survey measurement process. Eg., mailed questionnaires or personal contact situations in which the role of the enumerator or interviewer is minimized.

b. Random assignment of interviewers. There is a fixed population of B interviewers which are available for assignment to the n sampled individuals. It will be assumed that this process is undertaken at random with n_h individuals being associated with the h-th interviewer. Thus, $n = \sum_{h=1}^{B} n_h$.

The assignment of interviewers can be characterized by random variables T_{hi}, where

$$T_{hi} = \begin{cases} 1 & \text{if interviewer h is assigned to individual i given i is in sample} \\ 0 & \text{otherwise,} \end{cases}$$

and their corresponding joint probability distribution.

Basic Observational Unit: The random variable $Y_{hit}^{(\xi)}$ represents the measurement corresponding to the ξ-th attribute (where $\xi = 1,2,\ldots,m$ indexes the m attributes) of the i-th individual in the population with respect to the h-th interviewer. The subscript t indexes a conceptual sequence of replications of this overall measurement process for any specific individual in the sample. Henceforth, the ξ-superscript will usually be dropped for notational convenience.

2.2. Model. In the spirit of the approach of Wilk and Kempthorne [11], one model of interest involves assuming that the Y_{hit} can be represented as an additive function of an overall mean, a fixed main effect due to the i-th individual, a fixed main effect due to the h-th interviewer, a fixed interaction effect due to the combination of the h-th interviewer and i-th individual, and a random residual effect corresponding to the combination of the h-th interviewer, i-th individual, and t-th trial. In this regard, we have the structure

$$Y_{hit} = \overline{Y} + B_h + H_i + (BH)_{hi} + Z_{hit}$$

where $\overline{Y}$, B_h, H_i, and $(BH)_{hi}$ are determined in the following manner from $Y_{hi} = E\{Y_{hit}\}$

$$\overline{Y} = \frac{1}{NB} \sum_{h=1}^{B} \sum_{i=1}^{N} Y_{hi}$$

$$B_h = \frac{1}{N} \sum_{i=1}^{N} (Y_{hi} - \overline{Y}) = (\overline{Y}_h - \overline{Y})$$

$$H_i = \frac{1}{B} \sum_{h=1}^{B} (Y_{hi} - \overline{Y}) \equiv (Y_i - \overline{Y})$$

$$(BH)_{hi} = (Y_{hi} - B_h - H_i - \overline{Y}) = (Y_{hi} - \overline{Y}_h - Y_i + \overline{Y})$$

$$Z_{hit} = (Y_{hit} - Y_{hi}).$$

With respect to measurement error models like those developed by the U.S. Bureau of the Census (see [4]), the $\{Z_{hit}\}$ reflect trial to trial variation in the context of all potentially observable measurements of a particular attribute for any specific individual and interviewer combination. Thus, this source of variation represents intrinsic response error which is due to factors which are not under control with respect to the sampling selection and measurement data collection processes. Such response errors may be due to structural weaknesses or vagueness in the definition of the phenomena being measured for an individual (eg., attitudes related to political opinions, consumer taste preferences, etc.) or a consequence of some underlying stochastic process (occurrences of motor vehicle accidents and subsequent injuries, outcomes pertaining to judgments in court cases, survival subsequent to diagnosis and treatment for disease, pregnancy outcomes like birthweight, etc.). In view of these considerations, we shall henceforth assume that the $\{Z_{hit}\}$ are mutually uncorrelated and $\text{Var}\{Z_{hit}\} = E\{Z_{hit}^2\} = \eta_{hi}^2$.

The $\{T_{hi}\}$ give rise to external response error with respect to the fact that there is controlled variability in the basic measurement process to the extent that different interviewers can potentially tend to report different observations for a particular attribute of a specific individual. The nature of this source of error is characterized by the $\{B_h\}$ and $\{(BH)_{hi}\}$. Since these quantities are rather difficult to manipulate in general terms, we shall henceforth assume that they are unimportant and can be neglected; i.e., we shall assume $B_h=0$ for all h and $(BH)_{hi}=0$ for all h,i. For other discussion, see Koch [7].

Finally, the $\{U_i\}$ reflect sampling error since their joint probability distribution characterizes the selection aspects of the survey design in the sense of which individuals are in the sample and which are not.

a. Special Case for Only Sampling Errors. In this situation, $\eta_{hi}^2 = 0$ for all h,i and thus $Z_{hit} \equiv 0$. As a result, the model becomes

$$Y_{hit} = \bar{Y} + H_i.$$

Examples might include determinations of whether a product was defective or not in certain types of acceptance (inspection) sampling or the cost of purchases of stock items in inventory samples.

b. Special Case for Only Response Errors. In this situation $H_{hi} = 0$ for all h,i. As a result, the model becomes

$$Y_{hit} = \bar{Y} + Z_{hit}.$$

Examples might include determinations of the distribution of the ratio of the harmonic vs. geometric mean for the observations corresponding to the faces of four twelve-sided dice, determinations based on repeated simulations of a given stochastic process which are all based on the same computer random number generator (although with different starting points), repeated experimental observations on different samples from basically the same bacteria culture, repeated observations on the preferences of a specific individual with respect to (blind) paired comparisons of particular food or beverage products. However, as will be argued later, perhaps the most important situations of this type involve single observations on distinct individuals who belong to a matched set based on twin relationships, other family relationships, or equivalence with respect to several demographic or other characteristics.

2.3. Other Assumptions.

a. No interaction in the measurement process in the sense that $E_t\{Y_{hit} | \text{any specification of } \{U_i\} \text{ and } \{T_{hi}\}\} = Y_{hi}$ and hence conditional expected value notation of the type $E_t\{\ |\ \}$ will not be used in any of the remaining discussion.

b. The measurement process is unbiased in the sense that the population mean $\bar{Y}$ (or the population total $Y = N\bar{Y}$) is the population parameter of interest.

2.4. Linear Sample Statistics. Let us consider the model described in (2.2) with respect to the statistics

$$y_t = \frac{1}{B} \sum_{i=1}^{N} \sum_{h=1}^{B} W_{hi} U_i T_{hi} Y_{hit}$$

which is a linear combination of the observed elements in the sample with the W_{hi} being known specified coefficients. On the basis of previous assumptions, we have

$$E\{y_t\} = \frac{1}{B} \sum_{i=1}^{N} \sum_{h=1}^{B} W_{hi} \phi_i \lambda_{hi} Y_{hi}$$

where $\phi_i = E\{U_i\}$ and $\lambda_{hi} = E\{T_{hi} | U_i = 1\}$. In the remainder of this paper, we shall assume that the weights are the reciprocals of the probabilities associated with each (h,i) combination; i.e., $W_{hi} = (1/\phi_i \lambda_{hi})$ so that y_t represents a generalized Horvitz and Thompson [5] estimator which accounts for non-uniform assignments of interviewer effects. We shall now assume that the $\{\lambda_{hi}\}$ are uniform in the sense that $E\{T_{hi}|U_i=1\} = (1/B)$ in which case we can write y_t in the usual Horvitz-Thompson form

$$y_t = \sum_{i=1}^{N} W_i U_i \{\sum_{h=1}^{B} T_{hi} Y_{hit}\} = \sum_{i=1}^{N} \left(\frac{1}{\phi_i}\right) U_i Y_{it}$$

where $W_i = (1/\phi_i)$ and $Y_{it} = \{\sum_{h=1}^{B} T_{hi} Y_{hit}\}$ so that

$$E\{y_t\} = \sum_{i=1}^{N} Y_i = Y = (N\bar{Y})$$

since $E_{t,T}\{Y_{it}\} = Y_i$ with $E_{t,T}\{\ \}$ being interpreted as expected value with respect to both conceptual repeated trials as well as interviewer assignments.

As indicated in Koch [7], the variance of y_t under the model (2.2) in conjunction with the assumptions (2.3) can be written as

$$\text{Var}\{y_t\} = \frac{1}{n}\{(SRV) + (n-1)(CRV)\} + \{(SV)\}$$

with

$$(SRV) = \text{Simple Response Variance} = N^2[\frac{1}{N}\sum_{i=1}^{N} (\bar{\phi}/\phi_i)\ E_{t,T}\{(Y_{it}-Y_i)^2\}];$$

$$(CRV) = \text{Correlated Response Variance} = \frac{N^2}{N(N-1)} \sum_{i \neq i'}^{N} \Psi_{ii'} E_{t,T}\{(Y_{it}-Y_i)(Y_{i't}-Y_{i'})\}$$

(SV) = Sampling Variance

$$= \sum_{i=1}^{N} \sum_{i'=1}^{N} \{\frac{\theta_{ii'}}{\phi_i \phi_{i'}} - 1\} Y_i Y_{i'} \text{ ; where}$$

$$\overline{\phi} = \frac{1}{N} \sum_{i=1}^{N} \phi_i = (n/N),\ \overline{\theta} = \frac{1}{N(N-1)} \sum_{i \neq i'}^{N} \theta_{ii'} =$$

$$n(n-1)/N(N-1), \text{ and } \Psi_{ii'} = \frac{\theta_{ii'} \overline{\phi}^2}{\overline{\theta} \phi_i \phi_{i'}}.$$

In accordance with the assumptions regarding <u>intrinsic response error</u> (due to uncontrolled observational factors) and <u>external response error</u> (due to interviewer effects, etc.) given in 2.2 and the assumptions in 2.3 it follows that there is no <u>correlated response variance</u> component since CRV = 0 under these conditions and $Var\{y_t\}$ simplifies to $Var\{y_t\} = \frac{1}{n}\{(SRV)\} + \{(SV)\}$ where $(SRV) = N^2 \frac{1}{N} \sum_{i=1}^{N} (\overline{\phi}/\phi_i) \eta_i^2$ with η_i^2 being defined by $\eta_i^2 = \frac{1}{B} \sum_{h=1}^{B} \eta_{hit}^2$.

2.5. <u>Estimators for the Variance of a Linear Sample Statistic</u>. Here, we shall restrict attention to the usual Horvitz-Thompson statistic $y_t = \sum_{i=1}^{N} (\frac{1}{\phi_i}) U_i Y_{it}$ under the assumptions given in (2.2)-(2.4) with respect to the uncorrelated nature of <u>intrinsic response errors</u> and the nonexistence of <u>external response errors</u>.

a. <u>Direct Methods</u>. A lower bound estimator for the variance of y_t is the Horvitz-Thompson quadratic statistic

$$(\hat{SV}) = \sum_{i=1}^{N} \sum_{i'=1}^{N} \{\frac{1}{\phi_i \phi_{i'}} - \frac{1}{\theta_{ii'}}\} U_i U_{i'} Y_{it} Y_{i't}$$

for which $E\{(\hat{SV})\} = \frac{N^2}{n} \{\frac{1}{N} \sum_{i=1}^{N} (\frac{\overline{\phi}}{\phi_i} - \frac{n}{N}) \eta_i^2\} + (SV)$.

Similarly, an upper bound estimator for $Var\ t_y$ is

$$(\hat{\hat{SV}}) = \sum_{i=1}^{N} \sum_{i'=1}^{N} \{\frac{1}{(1-\phi_i)(1-\phi_{i'})}\}^{1/2} \cdot \{\frac{1}{\phi_i \phi_{i'}} - \frac{1}{\theta_{ii'}}\} U_i U_{i'} Y_{it} Y_{i't}$$

for which

$$E\{(\hat{\hat{SV}})\} = \frac{1}{n}\{(SRV)\} + \sum_{i=1}^{N} \sum_{i'=1}^{N} \{\frac{1}{(1-\phi_i)(1-\phi_{i'})}\}^{1/2} \cdot \{\frac{\theta_{ii'}}{\phi_i \phi_{i'}} - 1\} Y_i Y_{i'}.$$

Although $(\hat{SV})$ and $(\hat{\hat{SV}})$ appear to be quite different, it follows that if n and N are large and if n << N so that the terms $(1-\phi_i)$ which tend to behave like (1-n/N) can be replaced by 1's, then $(\hat{SV}) \approx (\hat{\hat{SV}})$. Finally, these considerations can be simplified if all the ϕ_i are equal to (n/N) in which case

$$E\{(\hat{SV})\} = Var\{y_t\} - (\frac{1}{N}) \text{ (SRV)}$$

$$E\{(\hat{\hat{SV}})\} = Var\{y_t\} + \frac{n}{(N-n)} \text{ (SV)}$$

Moreover, if all the $\theta_{ii'}$ are equal to $n(n-1)/N(N-1)$, then the expressions for $(\hat{SV})$ and $(\hat{\hat{SV}})$ also simplify to the familiar forms

$$(\hat{SV}) = N^2 (\frac{1}{n}) (1-\frac{n}{N}) \{\frac{1}{(N-1)} \sum_{i=1}^{N} U_i (Y_{it} - \overline{y}_t)^2\}$$

$$(\hat{\hat{SV}}) = N^2 (\frac{1}{n}) \frac{1}{(N-1)} \sum_{i=1}^{N} U_i (Y_{it} - \overline{y}_t)^2$$

where $\overline{y}_t = (y_t/N)$ is the estimator for the population mean (in this case, the ordinary sample mean) and $s^2 = \{\frac{1}{(N-1)} \sum_{i=1}^{N} U_i (Y_{it} - \overline{y}_t)^2\}$ is the sample variance estimator in the usual sense.

In summary, if <u>sampling variance</u> is the most important source of error, then $(\hat{SV})$ is the most appropriate estimator since $(\hat{\hat{SV}})$ is needlessly conservative. However, if <u>intrinsic response variance</u> is the most important source of error, then $(\hat{\hat{SV}})$ is the most appropriate estimator since $(\hat{SV})$ will tend to underestimate the actual variance in a potentially misleading manner.

b. <u>Replication Methods</u>. For many surveys involving complex multistage selection procedures, the numerical calculations associated with estimators like $(\hat{SV})$ or $(\hat{\hat{SV}})$ can require considerable effort with respect to programming as well as substantial computer time costs. The main reasons for this is that these expressions involve n^2 terms and the subscript i may be a vector subscript. Thus, in recent years, there has been considerable interest in the development of alternative estimation procedures for the variances of sample statistics. In particular, one such method which has been already used extensively by the National Center for Health Statistics as well as other institutions or organizations engaged in survey research is the method of balanced repeated replication (BRR) as discussed for example by Kish and Frankel [6], Koch and Lemeshow [8] and McCarthy [9]. The principal concept which governs the use of BRR is that variability of a statistic based on a total sample can be estimated in terms of the variability of that statistic across subsamples (called replications) which reproduce (except for size) the complex design of the entire sample. Hence, BRR has considerable appeal in those situations where clustering causes the underlying distribution theory for determining the $\theta_{ii'}$ as well as the computational effort for calculating $(\hat{SV})$ or $(\hat{\hat{SV}})$ to become impractical. One specific version of BRR is the method of balanced half samples. This procedure is characterized by a matrix H with elements h_{ik} defined by

$$h_{ik} = \begin{cases} 1 & \text{individual i is in k-th half sample} \\ 0 & \text{otherwise} \end{cases}.$$

For each half sample, we form the estimator

$y_{tk} = 2\sum_{i=1}^{N}(\frac{1}{\phi_i})U_i h_{ik} Y_{it}$ which is directly analogous to $\bar{y}_t$ with respect to estimating the population total Y. The resulting estimator V for the variance of y_t is

$V = \frac{1}{L}\sum_{k=1}^{L}(y_{tk} - y_t)^2$ where L denotes the number of half sample partitions. In this context, it should be noted that the appropriate choice of the matrix $\underset{\sim}{H}$ represents a very important feature of this method for determining the estimator V. Some efficient strategies for this purpose are described in [9]. Finally, it should be recognized that this method of estimating variance primarily pertains to those situations where there are no important sources of external response errors (eg., interviewer effects) as assumed in most parts of this paper. However, appropriate modifications with respect to the definition of $\underset{\sim}{H}$ are certainly within the scope of the general BRR approach for constructing replication estimators of variance which reasonably reflect this source of error as well as intrinsic response error and sampling error.

2.6. Estimators for the Covariance of Two Linear Sample Statistics. Suppose $y_t^{(\xi)}$ and $y_t^{(\xi')}$ correspond to the Horvitz-Thompson statistics for estimating the population totals corresponding to the ξ-th and ξ'-th attributes respectively. Then the methods described in (2.5) can be used to determine estimators for $\mathrm{Var}\{y_t^{(\xi)}\}$, $\mathrm{Var}\{y_t^{(\xi')}\}$, and $\mathrm{Var}\{y_t^{(\xi)} + y_t^{(\xi')}\}$ where

$$y_t^{(\xi)} + y_t^{(\xi')} = \sum_{i=1}^{N}(\frac{1}{\phi_i})U_i(Y_{it}^{(\xi)} + Y_{it}^{(\xi')})$$

is the Horvitz-Thompson statistic for estimating the sum of the population totals associated with the ξ-th and ξ'-th attributes in terms of the respective sums for individuals who are selected in the sample. However, this means that an estimator for Cov $y_t^{(\xi)}$, $y_t^{(\xi')}$ can be directly obtained from the identity relationship

$$\mathrm{Cov}\{y_t^{(\xi)}, y_t^{(\xi')}\} = \frac{1}{2}[\mathrm{Var}\{y_t^{(\xi)} + y_t^{(\xi')}\} - \mathrm{Var}\{y_t^{(\xi)}\} - \mathrm{Var}\{y_t^{(\xi')}\}]$$

by replacing the respective variance expressions by their corresponding estimators. Thus, an (m x m) estimated covariance matrix V can be determined for the joint set of estimators for the population totals corresponding to m attributes by using a computer program which calculates estimates of variance on individual univariate variables in conjunction with a variable sum operation and the previously indicated identity.

2.7. Estimates for Domain Means and Other Ratio Statistics. The term domain refers to subclasses derived from a particular response variable which is measured during the survey (ie., a posteriori with respect to selection) and which takes on categorical values (either directly as with marital status or indirectly after grouping as with age). The term strata refers to subclasses derived from a factor variable which is presumed known for each individual in the population prior to the undertaking of the survey (eg., region of the country, or urban vs. rural, etc.). Moreover, in most applications, several strata-type variables are directly related to the nature of the selection process with corresponding effects induced on the joint distribution of the U_i in some manner (ie., separate independent random samples are usually obtained from each subpopulation corresponding to appropriate combinations of such strata-type variables).

Estimates of subpopulation totals for both domains as well as strata can be formulated in terms of indicator variables of the type

$$X_{hijt} = \begin{cases} 1 & \text{individual } i \text{ is classified in } j\text{-th domain on } t\text{-th trial measured by } h\text{-th interviewer} \\ 0 & \text{otherwise} \end{cases}$$

where j = 1,2,...,s by forming statistics like

$$y_{jt} = \frac{1}{B}\sum_{i=1}^{N}\sum_{h=1}^{B} W_{hi}\phi_i\lambda_{hi}X_{hijt}Y_{hit}.$$

We shall assume that the domain classification process is affected neither by intrinsic response error nor external response error and is also unbiased in the sense of (2.3); ie., we have

$$X_{hijt} = X_{ij} = \begin{cases} 1 & \text{the individual } i \text{ is always classified correctly in the } j\text{-th domain} \\ 0 & \text{otherwise.} \end{cases}$$

However, it should be recognized that the presence of such response errors is a definite possibility in many survey situations and can have potentially important effects on the statistical properties of estimators like y_{jt} as discussed in Koch [7]; eg., y_{jt} is not necessarily unbiased unless such assumptions apply. Finally, strata-type variables are viewed in this same framework by definition together with the fact that the actual values of the $\{X_{ij}\}$ are known a priori constants for each individual in the population.

As indicated in (2.4) we shall consider the Horvitz-Thompson type estimators

$y_{jt} = \sum_{i=1}^{N}(\frac{1}{\phi_i})U_i G_{ijt}$ which can also be written in

the form $y_{jt} = \sum_{i=1}^{N}(\frac{1}{\phi_i})U_i G_{ijt}$ where $G_{ijt} = X_{ij}Y_{it}$.

However, in this context, the discussion in (2.5)-(2.6) can be applied to obtain an estimated covariance matrix V for the vector of domain total estimators $\underset{\sim}{y}'_t = (y_{1t}, y_{2t}, \ldots, y_{st})$. Similarly, these same considerations can also be applied to the multivariate case of m attribute variables by working with the composite vector

$$\underset{\sim}{y}'_t = [\underset{\sim}{y}_t^{(1)'}, \underset{\sim}{y}_t^{(2)'}, \ldots, \underset{\sim}{y}_t^{(m)'}].$$

In many situations, there is actually greater interest in domain means which are basically ratio estimates of the type

$$\bar{y}_{jt} = \sum_{i=1}^{N}(\frac{1}{\phi_i})U_i G_{ijt} \Big/ \sum_{i=1}^{N}(\frac{1}{\phi_i})U_i X_{ij} = (y_{jt}/x_{jt}).$$

If $\underset{\sim}{y}'_t = (y_{1t}, y_{2t}, \ldots, y_{st})$ and $\underset{\sim}{x}'_t = (x_{1t}, x_{2t}, \ldots, x_{st})$ then the set of domain means $\bar{\underset{\sim}{y}}_t = (\bar{y}_{1t}, \bar{y}_{2t}, \ldots \bar{y}_{st})$ can be written in the compound function framework outlined in Forthofer and Koch [2]

$\bar{\underset{\sim}{y}}_t = \underset{\sim}{R}[\underset{\sim}{\exp}\{\underset{\sim}{K}(\underset{\sim}{\log_e}(\underset{\sim}{A}\begin{matrix}\underset{\sim}{y}_t \\ \underset{\sim}{x}_t\end{matrix})\}]$ where $\underset{\sim}{A} = \underset{\sim}{I}_{2s}$ (which is the (2s x 2s) identity matrix),

$$K = \begin{bmatrix} 1 & 0 & \cdots & 0 & -1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 & 0 & -1 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & 1 & 0 & 0 & \cdots & -1 \end{bmatrix}$$

and $R = I_s$ (which is the (s x s) identity matrix) and the $\log_e$ ( ) vector operation forms the vector of natural logarithms and the exp vector operation forms the vector of anti-logarithms.

An estimate of the covariance matrix for y_t which is based on the large sample Taylor series linearized approximation can be obtained with direct matrix multiplication operations as follows

$$\mathrm{Var}\{y_t\} \hat{=} V_y = R D_q K D_a^{-1} A V A' D_a^{-1} K' D_q R'$$

where "$\hat{=}$" means "is estimated by" and where D_a represents the diagonal matrix with elements of the vector $a = A\begin{bmatrix} y_t \\ x_t \end{bmatrix}$ on the main diagonal and D_q represents the diagonal matrix with elements of the vector $q = \exp\{K[\log_e(a)]\}$ on the main diagonal. Estimated covariance matrices for other sets of compounded functions involving estimates of domain totals can be produced in an analogous manner; eg, differences between domain means, post-stratified means, certain types of vital rates based on life-table functions, and rank correlation type measures of association.

In summary, compound function operations can be used to compute the vector $\bar{y}_t$ of estimated means for any given set of domain subpopulations. The corresponding estimated covariance matrix $V_{\bar{y}}$ can be calculated by previously described matrix multiplication operations. Alternatively, an estimated covariance matrix $V_{\bar{y}}$ could also be obtained by using the replication methods described in (2.5b) although there are potentially certain problems with singularities with this approach for reasons which are outlined in [8]. Nevertheless, the estimator $\bar{y}_t$ and its estimated covariance matrix $V_{\bar{y}}$ can be obtained for any specific survey situation within the context of the general framework described in (2.1)-(2.7). The problem now is to consider a general approach for undertaking statistical inference with respect to $\bar{y}_t$.

2.8. Multivariate Analysis for Estimates from Complex Sample Survey Data. Let F denote a (g x 1) vector of statistics like the domain mean estimates $\bar{y}_t$ described in (2.7). Let V_F denote an appropriate valid and consistent estimate of the corresponding (g x g) covariance matrix for F obtained by methods like those described in (2.4)-(2.7).

The relationship between variation among the elements $F_1, F_2, \ldots, F_g$ of the vector F and certain aspects of the nature of various subpopulations (or domains and subdomains) can be investigated by fitting linear regression models to the vector F by the method of weighted least squares. This aspect of statistical analysis can be characterized by writing

$$F = \begin{bmatrix} F_1 \\ F_2 \\ \cdots \\ F_g \end{bmatrix} \hat{=} \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1u} \\ x_{21} & x_{22} & \cdots & x_{2u} \\ \cdots & \cdots & \cdots & \cdots \\ x_{g1} & x_{g2} & \cdots & x_{gu} \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \\ \cdots \\ b_u \end{bmatrix} = X b$$

where X is the pre-specified design (or independent variable) matrix of known coefficients with full rank u, b is the (u x 1) vector of unknown parameters or effects, and "$\hat{=}$" means "is estimated by." This particular model implies the existence of a [(g - u) x g] matrix L which is orthogonal to X such that $\{f = L F \hat{=} L X b = 0\}$ represents a corresponding set of implied constraints. Thus, it follows that the covariance matrix of f can be estimated by $V_f = L V_F L'$. As a result, an appropriate test statistic for the goodness of fit of the model of interest is

$$Q = f' V_f^{-1} f = f'(L V_F L')^{-1} f$$

which is approximately distributed according to the χ^2-distribution with D.F.=(g - u) if the overall sample size n is sufficiently large that the elements of the vector F have an approximate multivariate normal distribution as a consequence of Central Limit Theory. Such test statistics are known as Wald [10] statistics and various aspects of their application to problems involving the multivariate analysis of multivariate categorical data are discussed in [2], [3]. Moreover, since the sample sizes associated with most complex sample surveys are generally very large so that it is reasonable to assume that the resulting estimates of population characteristics tend to have approximately normal distributions (as a consequence of Central Limit Theory), such Wald statistics provide a valid and potentially useful framework for the multivariate analysis of the resulting estimates. However, the actual manner in which this approach is undertaken involves a weighted least squares computational algorithm which is justified on the basis of the fact that

$$Q = f'(L V_F^{-1} L')^{-1} f \equiv (F - X b)' V_F^{-1} (F - X b)$$

where $b = (X' V_F^{-1} X)^{-1} X' V_F^{-1} F$ represent weighted least square estimates for the underlying parameters. In view of this IDENTITY and the large sample validity of the Wald Statistic Q, the weighted least squares estimates b are also regarded as having reasonable statistical properties because of the manner in which they determine Q. With these considerations in mind, it then can be noted that $V_b = (X' V_F^{-1} X)^{-1}$ represents a consistent estimate for the covariance matrix for b

When an appropriate model has been determined statistical tests of significance involving b may be performed by standard multiple regression procedures. Linear hypotheses are formulated as H_0: $Cb = 0$, where C is a known (d x u) coefficient matrix and tested using the statistic $Q_C = b'C'[C(X'V_F^{-1}X)^{-1}C']^{-1}C\, b$ which is approximately distributed according to the χ^2-distribution with D.F.=d when the hypothesis H_0 is true.

Successive uses of the goodness of fit tests and the significance tests specified by the C matrices represent ways of partitioning the model components into specific sources of variance. In this context, the Q_C statistics reflect the amount by which the residual sum of squares goodness of fit Wald statistic Q would increase if the basic model were simplified (or reduced) by substitutions based on the additional constraints implied by H_0: $C\, b \hat{=} 0$. This partitioning of total variance into specific sources represents a statistically valid analysis of variance for estimator functions F arising from complex sample survey situations.

Finally, predicted values corresponding to any specific model can be calculated from $\hat{F} = X b = X (X'V_F^{-1} X)^{-1} X'V_F^{-1} F$ and corresponding estimates of variance can be obtained from the diagonal elements of $V_{\hat{F}} = X(X'V_F^{-1}X)^{-1}X'$. Such predicted values not only have the advantage of characterizing essentially all the important features of the variation in the original data, but also represent better estimates than the original function statistics F since they are based on the data from the entire sample (ie., all subdomains combined) as opposed to its component parts. Finally, they are descriptively advantageous in the sense that they make trends more apparent and permit a clearer interpretation of the effects of the respective independent variables comprising X on the vector F.

3. APPLICATIONS AND EXAMPLES

3.1. Strategies for applying the model. In applying the methodology of Section 2 it is necessary to have a data analytic strategy. This strategy depends on taking note of the apparent bimodal nature of the statistical sciences. There are two general types of statistics -- descriptive statistics and inferential statistics. The role of the descriptive statistic is primarily summarization and does not strictly justify any comparative or other type of conclusion being extracted from the data. On the other hand, the inferential statistic is often a dimensionless index which may have only limited descriptive value. Its primary role is an orientation toward decision-making in the sense of being interpreted as either consistent with or in contradiction to a particular hypothesis which is of interest with respect to formulating conclusions from the data. Hence, certain inferential statistics can be used to determine whether any observed differences between two groups of individuals, such as a group of healthy patients and a group of diseased patients, are real or systematic as opposed to being due to chance variation; others can be used similarly to interpret the association among certain variables which are, for example, indicative of clinical status.

The preceding remarks have been directed at some of the objectives of statistical analysis. As formulated here, they appear reasonably clear and concise. However, the various ways in which statisticians operate in accomplishing them often appear heuristic and mystical. This impression results from the fact that "statistics" is in some sense an estranged marriage between "routine data processing" and "abstract mathematical probability theory." The paradox here is that "data processing" exists in the real world and can always be used to produce descriptive measures like arithmetic averages, percentiles, and least squares coefficients from any set of data, no matter how collected in terms of the underlying research design. Such computations constitute what will be called a "numerical analysis" of the data. Of course, the conclusions resulting from this type of approach are entirely limited to the data under consideration and cannot be rigorously generalized to any larger underlying population from which it is a sample. Moreover, a strict "numerical analysis" does not permit us to formally document the precision or reliability of quoted descriptive summary measures.

On the other hand, probability theory is a set of abstract axioms, definitions, and theorems, all of which are very much outside the real world, but which, under suitable assumptions, can provide reasonable mathematical models for characterizing numerically measured quantities. Within this framework, "statistics" is the liaison between a set of data and a suitable mathematical probability model. Hence, the most crucial aspect of any statistical analysis is the validity of the formal assumptions which underlie the corresponding probability model. Indeed, this principle can be underscored in some cases to the extent of identifying those assumptions or conditions which lead to contradictory conclusions and then choosing that conclusion together with its supporting analysis, for which the corresponding set of assumptions seems to be most empirically and/or physically realistic. It is in this context that the paradoxes associated with the sayings dealing with "how to lie with statistics" can be resolved.

Although the previously described point of view appears somewhat different from that which is concerned with developing standards for justifying statistical statements of a descriptive or inferential nature, there are definite similarities with respect to the underlying philosophy. In particular, the Q statistics in (2.8) represent an analytical procedure for assessing whether there is any variation among a given set of statistics and if so whether it can be partitioned in meaningful manner. The former question is entirely of an inferential nature and can be interpreted as dealing with the multiple comparison problem in a Scheffe simultaneous test procedure sense. The latter is both descriptive and inferential and may also be possibly guided to some extent by substantive considerations pertaining to the subject matter area for which such analysis is being undertaken. Similar remarks can be applied to the predicted values which are generated from various fitted models. These points will all be discussed in more detail for the examples in Section 3.2. In summary, there are two basic goals which are associated with the analysis of complex sample survey statistics as well as any other types of data:

1. Sample statistics which are essentially similar (in the sense of not being statistically different in a significance testing context) should not be reported in tables as different, although raw or unanalyzed data should of course be displayed where appropriate. Alternatively, the same estimate should be reported for each element in such cases. Moreover, one method for obtaining such estimates is weighted least squares as described in (2.8).
2. Sample statistics which are significantly different (at some appropriate level;eg., α=.05) should be reported in terms of correspondingly different estimates. However, attempts should be made to structurally characterize such differences in terms of models which can be fitted by weighted

least squares. In this latter context, it should be recognized that both significance testing inferential considerations as well as percent explained variation descriptive considerations are important.

3.2. Examples of the model. The preceding remarks will guide us in analyzing the following examples. All four are taken from the National Center for Health Statistics, Office of Statistical Methods, (unpublished manuscript). This, rather than the original sources indicated on the examples, was used to facilitate the computation of standard errors. In some cases, the standard errors were provided in the original sources, however in others it could only be computed with difficulty or not at all. In all cases sample correlations between the statistics were assumed to be zero. However, some preliminary investigation suggests that the Q-statistics are conservative in the case of positive equal correlation and anti-conservative in the case of negative equal sample correlation. That is, for positive correlation some differences will go undetected while for negative correlation nonexistent differences will appear.

Our first example analyzes the estimates of the proportion of dentulous persons, ages 18-79, needing to see a dentist prior to next regular visit, in various marital states. Our preliminary model saturates the variation space and the Q-statistic of 18.66 for total variation indicates that significant variation exists among marital states with respect to a χ^2(D.F.=4) distribution. This permits an examination of classes to identify the ones contributing the greatest variation. b_5 which corresponds to the difference between the separated and never married groups generates Q=16.55 which is significant even

Example 1

Comparisons Among Several Subdomains Within a Domain

Subdomain of White Adults	Estimated Proportion Needing to See a Dentist at an Early Date	Estimated Standard Error	Aggressive Model: Smoothed Predicted Values	Aggressive Model: Estimated Standard Errors	Conservative Model: Smoothed Predicted Values	Conservative Model: Estimated Standard Errors
Married	.378	.022	.389	.019	.366	.016
Widowed	.420	.049	.389	.019	.366	.016
Divorced	.426	.058	.389	.019	.366	.016
Separated	.627	.071	.627	.071	.627	.071
Never Married	.318	.027	.318	.027	.366	.016

Preliminary Model

$$\underset{\sim}{X} = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 \end{bmatrix}, \quad \underset{\sim}{b} = \begin{bmatrix} .318 \\ .060 \\ .102 \\ .108 \\ .309 \end{bmatrix}$$

Statistical Tests	D.F.	Q
$b_2 \hat{=} 0$	1	2.97
$b_3 \hat{=} 0$	1	3.32
$b_4 \hat{=} 0$	1	2.85
$b_5 \hat{=} 0$	1	16.55
$b_2 \hat{=} 0, b_3 \hat{=} 0, b_4 \hat{=} 0$	3	5.74
$b_2 - b_3 \hat{=} 0;\ b_2 - b_4 \hat{=} 0$	2	1.06
Total Variation	4	18.66

Final Models

Aggressive

$$\underset{\sim}{X} = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \\ 1 & 0 & 0 \end{bmatrix}, \quad \underset{\sim}{b} = \begin{bmatrix} .318 \\ .071 \\ .238 \end{bmatrix}$$

Statistical Tests	D.F.	Q
$b_2 \hat{=} 0$	1	4.69
$b_3 \hat{=} 0$	1	10.45
$b_2 + b_3 \hat{=} 0$	1	16.55
Model: $b_2 \hat{=} 0, b_3 \hat{=} 0$	2	17.60
Residual GOF	2	1.
Total Variation	4	18.66

Conservative

$$\underset{\sim}{X} = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 1 & 0 \end{bmatrix}, \quad \underset{\sim}{b} = \begin{bmatrix} .366 \\ .627 \end{bmatrix}$$

Statistical Tests	D.F.	Q
$b_1 - b_2 \hat{=} 0$	1	12.91
Model: $b_2 \hat{=} 0$	1	12.92
Residual GOF	3	5.74
Total Variation	4	18.66

Actual Source: National Center for Health Statistics, Office of Statistical Methods, Manual on Standards for Reviewing Statistical Reports (Unpublished Preliminary Draft), June 1973.
Original Source: National Center for Health Statistics, Vital and Health Statistics, Series 11, Number 36, "Table 5. Percent of dentulous adults who should see a dentist at early date by marital status, race, and sex: United States, 1960-62," p. 14.

in the total variation space, that is with respect to a χ^2 (D.F.=4) distribution. Further, there is no difference among the remaining ever married groups Q=1.06, with respect to a χ^2 (D.F.=2) distribution. The difference between the never married and remaining ever married groups is somewhat more subtle. There are two approaches. The conservative model groups never married, divorced, widowed, and married into one group and distinguishes only the separated group. The residual Q=5.74 is nonsignificant with respect to a χ^2 (D.F.=3) distribution but may conceal an important component of variation as revealed by the aggressive model. The aggressive model has three groups, never married, separated, and other ever married. This model fits very well with a residual Q=1.06 which is nonsignificant even with respect to a χ^2 (D.F.=2) distribution. The smoothed values fulfill our goal of displaying differences only where they 'exist'; the question of which model is to be preferred is one that should be settled on substantive grounds.

Estimates of mean baby birthweight for various family income and mother's employment status categories provide our second example. The total variation, Q=22.83, in the preliminary model is significant with respect to a χ^2(D.F.=5) distribution, so as in example one, an effort to characterize the groups must be made. Parameters b_2 and b_3 compare the first to third and the second to third income groups and b_2 is individually significant. In fact, the Q that examines $b_2-2b_3 \hat{=} 0$ is only .05 so such a characterization of income is plausible. The variation associated with employment Q=14.24 is significant with respect to a χ^2(D.F.=3) distribution so these effects may also be further examined. Moreover, there is no interaction (employment status by income level), Q=1.00, with respect to a χ^2 (D.F.=2) distribution. These considerations lead to our final model and smoothed predicted birthweights which show only significant differences. This model has a small residual, Q=1.18, which is non-significant with respect to a χ^2(D.F.=3) distribution. It is also parsimonious in the sense that it contains only three parameters which permits a reduction in the standard errors of the

Example 2

Comparisons Among the Corresponding Subdomains of Two or More Domains

Domain: Family Income Level	Subdomain: Wife's Employment Status	Estimated Mean Baby Birthweight (grams)	Estimated Standard Error	Smoothed or Predicted Birthweight	Estimated Standard Error
\$3000-4999	Employed	3230	23.46	3232	16.65
	Unemployed	3290	17.96	3293	14.35
\$5000-6999	Employed	3280	22.11	3263	12.82
	Unemployed	3320	18.08	3323	10.47
\$7000 and over	Employed	3280	21.15	3294	15.81
	Unemployed	3360	18.36	3354	14.55

Preliminary Model

$$\underset{\sim}{X} = \begin{bmatrix} 1&1&0&1&0&0\\ 1&1&0&0&0&0\\ 1&0&1&0&1&0\\ 1&0&1&0&0&0\\ 1&0&0&0&0&1\\ 1&0&0&0&0&0 \end{bmatrix}, \quad \underset{\sim}{b} = \begin{bmatrix} 3360\\ -70\\ -40\\ -60\\ -40\\ -80 \end{bmatrix}$$

Statistical Tests	D.F.	Q
$b_2 \hat{=} 0$	1	7.43
$b_3 \hat{=} 0$	1	2.41
$b_4 \hat{=} 0$	1	4.12
$b_5 \hat{=} 0$	1	1.96
$b_6 \hat{=} 0$	1	8.16
$b_4 \hat{=} 0, b_5 \hat{=} 0, b_6 \hat{=} 0$	3	14.24
$b_4 - b_5 \hat{=} 0,\ b_4 - b_6 \hat{=} 0$	2	1.00
$b_2 - 2b_3 \hat{=} 0$	1	.05
Total Variation	5	22.83

Final Model

$$\underset{\sim}{X} = \begin{bmatrix} 1&-1&1\\ 1&-1&-1\\ 1&0&1\\ 1&0&-1\\ 1&1&1\\ 1&1&-1 \end{bmatrix}, \quad \underset{\sim}{b} = \begin{bmatrix} 3293\\ 31\\ -30 \end{bmatrix}$$

Statistical Tests	D.F.	Q
$b_2 \hat{=} 0$	1	9.47
$b_3 \hat{=} 0$	1	13.24
Model: $b_2 \hat{=} 0, b_3 \hat{=} 0$	2	21.65
Residual GOF	3	1.18
Total Variation	5	22.83

Actual Source: National Center for Health Statistics, Office of Statistical Methods, Manual on Standards for Reviewing Statistical Reports(Unpublished Preliminary Draft), June 1973.
Original Source: National Center for Health Statistics, Vital and Health Statistics, Series 22, Number 8, "Table 7. Average Birth Weight, number of birth, and percent distribution by birth-weight intervals according to family income in 1962 and whether mother was employed during pregnancy; United States, 1963 legitimate live births," p. 19 and Appendix I p. 30.

smoothed cell means. In this model the employment status effect, Q=13.24, is significant at α=.05 in a Scheffe multiple comparison sense with respect to a χ^2 distribution with 5 degrees of freedom in the total variation space, or 3 degrees of freedom in the total employment status subspace, or 2 degrees of freedom in the total reduced model subspace. Similarly the income level effect Q=9.47 is significant at α=.10 with respect to a χ^2 distribution with 5 degrees of freedom, or 4 degrees of freedom which pertains to the total income subspace, or 2 degrees of freedom.

The third example of our sample survey model and the techniques guiding reduction is the estimated mean scores on the Block design subtest for boys and girls ages six to eleven. As before the total variation, Q=1719.36, is significant however we will focus on the sex differential since the age effect is a marked trend. The variation associated with sex, Q=30.49 is significant with respect to a χ^2(D.F.=6) distribution hence further analysis is warranted. The variation associated with the age by sex interaction, Q=10.05, is also significant at the α=.10 level with respect to a χ^2(D.F.=5) distribution. This interaction is not present when only the last 5 age groups are considered jointly, Q=1.88. These considerations produce our final model. Here we have an increasing score with age trend, Q=1377.49, an age x sex interaction term Q=28.53 which combines boys and girls only in the first age group. These effects are significant with respect to a χ^2(D.F.=5) and a χ^2(D.F.=1) distri-

Example 3

Comparisons Among the Corresponding Subdomains of Two or More Domains

Domain Age	Subdomain Sex	Estimated Mean Score on Block Design Subtest	Estimated Standard Error	Smoothed Predicted Value	Estimated Standard Error
6 years	Boys	5.8	0.27	5.7	0.18
	Girls	5.7	0.24	5.7	0.18
7 years	Boys	8.5	0.29	8.6	0.24
	Girls	7.3	0.25	7.2	0.22
8 years	Boys	12.0	0.39	11.8	0.30
	Girls	10.3	0.36	10.5	0.29
9 years	Boys	14.0	0.46	14.0	0.34
	Girls	12.6	0.42	12.6	0.33
10 years	Boys	18.2	0.63	18.6	0.44
	Girls	17.5	0.55	17.2	0.43
11 years	Boys	22.3	0.62	22.0	0.50
	Girls	20.1	0.82	20.6	0.52

Preliminary Model

$$\underset{\sim}{X} = \begin{bmatrix} 1&1&0&0&0&0&1&0&0&0&0&0\\ 1&1&0&0&0&0&0&0&0&0&0&0\\ 1&0&1&0&0&0&0&1&0&0&0&0\\ 1&0&1&0&0&0&0&0&0&0&0&0\\ 1&0&0&1&0&0&0&0&1&0&0&0\\ 1&0&0&1&0&0&0&0&0&0&0&0\\ 1&0&0&0&1&0&0&0&0&1&0&0\\ 1&0&0&0&1&0&0&0&0&0&0&0\\ 1&0&0&0&0&1&0&0&0&0&1&0\\ 1&0&0&0&0&1&0&0&0&0&0&0\\ 1&0&0&0&0&0&0&0&0&0&0&1\\ 1&0&0&0&0&0&0&0&0&0&0&0 \end{bmatrix}, \quad \underset{\sim}{b} = \begin{bmatrix} 20.1\\ -14.4\\ -12.8\\ -9.8\\ -7.5\\ -2.6\\ 0.1\\ 1.2\\ 1.7\\ 1.4\\ 0.7\\ 2.2 \end{bmatrix}$$

Statistical Tests	D.F.	Q
$b_7 \hat{=} 0$	1	0.08
$b_8 \hat{=} 0$	1	9.82
$b_9 \hat{=} 0$	1	10.26
$b_{10} \hat{=} 0$	1	5.05
$b_{11} \hat{=} 0$	1	0.70
$b_{12} \hat{=} 0$	1	4.58
$b_7 \hat{=} 0, b_8 \hat{=} 0, \ldots, b_{12} \hat{=} 0$	6	30.49
$b_7 - b_8 \hat{=} 0, \ldots, b_7 - b_{12} \hat{=} 0$	5	10.05
$b_8 - b_9 \hat{=} 0, b_8 - b_{10} \hat{=} 0, b_8 - b_{11} \hat{=} 0, b_8 - b_{12} \hat{=} 0$	4	1.88
Total Variation	11	1719.36

Final Model

$$\underset{\sim}{X} = \begin{bmatrix} 1&1&0&0&0&0&0\\ 1&1&0&0&0&0&0\\ 1&0&1&0&0&0&1\\ 1&0&1&0&0&0&0\\ 1&0&0&1&0&0&1\\ 1&0&0&1&0&0&0\\ 1&0&0&0&1&0&1\\ 1&0&0&0&1&0&0\\ 1&0&0&0&0&1&1\\ 1&0&0&0&0&1&0\\ 1&0&0&0&0&0&1\\ 1&0&0&0&0&0&0 \end{bmatrix}, \quad \underset{\sim}{b} = \begin{bmatrix} 20.6\\ -14.9\\ -13.4\\ -10.2\\ -8.0\\ -3.4\\ 1.4 \end{bmatrix}$$

Statistical Tests	D.F.	Q
Age x Sex: $b_7 \hat{=} 0$	1	28.53
Age: $b_2 \hat{=} 0$, $b_3 \hat{=} 0$, $b_4 \hat{=} 0$, $b_5 \hat{=} 0$, $b_6 \hat{=} 0$	5	1377.49
Nonlinear Age: $b_2-b_3 \hat{=} b_3-b_4$, $b_3-b_4 \hat{=} b_4-b_5$, $b_4-b_5 \hat{=} b_5-b_6$, $b_5-2b_6 \hat{=} 0$	4	45.61
Model	6	1717.41
Residual	5	1.96
Total Variation	11	1719.36

Actual Source: National Center for Health Statistics, Office of Statistical Methods, Manual on Standards for Reviewing Statistical Reports (Unpublished Preliminary Draft), June 1973.
Original Source: National Center for Health Statistics, Vital and Health Statistics, Series 11, No. 107, "Table 3. Mean and standard deviation of raw scores on the Block Design subtest of the WISC by age and sex ... United States, 1963-65" p. 21, and "Table III. Sampling errors for average raw scores on the WISC Vocabulary and Block Design subtests by age, sex, ... United States, 1963-65," p. 38.

Example 4

Evaluation of Trend Effects

Midpoint of Income Class	Estimated Percent Needing Early Dental Visit	Estimated Standard Error	Aggressive Model: Smoothed Predicted Value	Aggressive Model: Estimated Standard Error	Regression Model: Smoothed Predicted Value	Regression Model: Estimated Standard Error
$1000	51.2	3.3	50.2	1.9	50.8	2.0
$3000	50.5	3.1	50.2	1.9	46.6	1.6
$5500	40.3	2.2	41.2	1.2	41.4	1.3
$8500	32.4	2.4	32.3	1.4	35.2	1.2
$15000	23.6	2.6	23.3	2.1	21.7	2.3

Preliminary Model

$$\underset{\sim}{X} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 & 1 \end{bmatrix}, \quad \underset{\sim}{b} = \begin{bmatrix} 51.2 \\ -0.7 \\ -10.2 \\ -7.9 \\ -8.8 \end{bmatrix}$$

Statistical Tests	D.F.	Q
$b_2 \hat{=} 0$	1	.02
$b_3 \hat{=} 0$	1	7.20
$b_4 \hat{=} 0$	1	5.89
$b_5 \hat{=} 0$	1	6.19
$b_2-b_3 \hat{=} 0, b_2-b_4 \hat{=} 0, b_2-b_5 \hat{=} 0,$	3	2.52
$b_3-b_4 \hat{=} 0, b_3-b_5 \hat{=} 0$	2	.164
Total Variation	4	69.45

Final Model

Aggressive Model

$$\underset{\sim}{X} = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 1 \\ 1 & 2 \\ 1 & 3 \end{bmatrix}, \quad \underset{\sim}{b} = \begin{bmatrix} 50.2 \\ -9.0 \end{bmatrix}$$

Statistical Tests	D.F.	Q
Model: $b_2 \hat{=} 0$	1	69.16
Residual	3	0.30
Total Variation	4	69.45

Regression Model

$$\underset{\sim}{X} = \begin{bmatrix} 1 & 1 \\ 1 & 3 \\ 1 & 5.5 \\ 1 & 8.5 \\ 1 & 15. \end{bmatrix}, \quad \underset{\sim}{b} = \begin{bmatrix} 52.8 \\ -2.1 \end{bmatrix}$$

Statistical Tests	D.F.	Q
Model: $b_2 \hat{=} 0$	1	65.71
Residual	3	3.75
Total Variation	4	69.45

Actual Source: National Center for Health Statistics, Office of Statistical Methods, Manual on Standards for Reviewing Statistical Reports, (Unpublished Preliminary Draft), June 1973.
Original Source: National Center for Health Statistics, Vital and Health Statistics Series 11, No. 36, "Table 3. Percent of dentulous adults who should see dentist at early date, by family income, race, and sex: United States, 1960-62," p.13.

bution respectively. The age trend is non linear, Q=45.61, with respect to a χ^2(D.F.=4) distribution. That the final model does not obscure any important variation is seen from the residual Q=1.96, which is non-significant with respect to a χ^2(D.F.=5) distribution. The reported smoothed or predicted scores show all significant differences. Note that boys and girls are equal only at year 6 and boys score higher at later ages. Also there is a substantial reduction in standard errors for the smoothed values which has resulted from the final model's characterization or partioning of the total variation space.

Our last example uses estimates of the percentage of dentulous persons in various income classes needing an early visit to a dentist. The preliminary model displays sufficient total variation(Q=69.45) with respect to a χ^2(D.F.=4) distribution to indicate significant differences among the five income classes. However, the variation between the first two classes(Q=.02) indicates they are the same and the first difference occurs with the third class(Q=7.20). Moreover, contrasting the classes(Q=.164 and Q=2.52) reveals approximately equal declines in the percentages. These considerations lead to two possible models. The regression model uses the midpoint of the income classes and accounts for the significant variation(Q=65.71) with respect to a χ^2(D.F.=4) distribution, in a Scheffé multiple comparison sense. But it could be argued that on substantive grounds the first two income classes are the same and further the income groupings are essentially arbitrary. Given this the aggressive model would be proposed. Again, all significant variation(Q=69.16) is accounted for but groups that are probably alike are not separated.

ACKNOWLEDGMENTS

This research was in part supported by the National Institute of Health (Grants GM-70004 and HD-00371). The authors thank Dr. Monroe G. Sirken and Dr. Daniel G. Horvitz for discussions which were helpful in formulating this methodology.

REFERENCES

1. Cornfield, J., On Samples from Finite Populations. JASA-39.
2. Forthofer, R.N. and Koch, G.G., Analysis for Compounded Funtions of Categorical Data. Biocs-29.
3. Grizzle, J.E., Starmer, C.F., and Koch, G.G., Analysis Categorical Data Linear Models. Biocs-25.
4. Hansen, M.H., Hurwitz, W.N., and Bershad, M.A., Measurement Errors in Censuses and Surveys. I.S.I. Bulletin-38.
5. Horvitz, D.G. and Thompson, D.J., A Generalization of Sampling Without Replacement from a Finite Universe. JASA-47.
6. Kish, L. and Frankel, M.R., Balanced Repeated Replication for Standard Errors. JASA-65.
7. Koch, Gary G., An Alternative Approach to Multivariate Response Error Models for Sample Survey Data with Applications. JASA-68.
8. Koch, Gary G. and Lemeshow, S., An Application of Multivariate Analysis to Complex Sample Survey Data. JASA-67.
9. McCarthy, P.J., Pseudo-Replication: Half Samples. I.S.I. Review-37.
10. Wald, A., Tests of Statistical Hypotheses Concerning Several Parameters When the Number of Observations is Large. Trans. of A.M.S.-54.
11. Wilk, M.B. and Kempthorne, O., Fixed, Mixed, Random Models in Analysis of Variance. JASA-50.