DOCUMENT RESUME

ED 268 155 TM 860 177

AUTHOR Mislevy, Robert J.
TITLE Inferences about Latent Populations from Complex Samples.
INSTITUTION Educational Testing Service, Princeton, N.J.
REPORT NO ETS-RR-85-41
PUB DATE Dec 85
NOTE 39p.
PUB TYPE Reports - Research/Technical (143)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Algorithms; Data Interpretation; *Estimation (Mathematics); *Latent Trait Theory; Mathematical Models; Prediction; Predictor Variables; Racial Differences; Research Design; *Sampling; Sex Differences; *Statistical Inference; Statistical Studies; Surveys; Vocational Aptitude; Young Adults
IDENTIFIERS Armed Services Vocational Aptitude Battery; *Latent Variables; *Missing Data; Profile of American Youth; Randomization (Statistics)

ABSTRACT

A method for drawing inferences from complex samples is based on Rubin's approach to missing data in survey research. Standard procedures for drawing such inferences do not apply when the variables of interest are not observed directly, but must be inferred from secondary random variables which depend on the variables of interest stochastically. This method allows reasonable inferences to be made. The key is to represent knowledge about latent variables in the form of a predictive distribution, conditional on manifest variables. It is then possible to obtain the expectations of statistics that would have been computed if the values of the latent variables corresponding to sampled units were known, along with variance estimators that account for uncertainty due to both subject sampling and the latency of the variables. (A numerical example is presented, using data from the Profile of American Youth (1980). Possible responses to four arithmetic reasoning items from the Armed Services Vocational Aptitude Battery were studied for Black male and female and White male and female populations). (GDC)

RR-85-41

RESEARCH REPORT

INFERENCES ABOUT LATENT POPULATIONS FROM COMPLEX SAMPLES

Robert J. Mislevy

Educational Testing Service
Princeton, New Jersey
December 1985

Inferences about Latent Populations from Complex Samples

Robert J. Mislevy

Educational Testing Service

Princeton, New Jersey

December 1985

Running head: INFERENCES ABOUT LATENT POPULATIONS

Abstract

Standard procedures for drawing inferences from complex samples do not apply when the variables of interest z are not observed directly, but must be inferred from secondary random variables x that depend on z stochastically. Employing Rubin's (1977) approach to missing data in survey research, we present a procedure by which reasonable inferences can be made in such situations. The key is to represent knowledge about latent variables in the form of a predictive distribution, conditional on manifest variables. It is then possible to obtain the expectations of statistics that would have been computed if the values of the latent variables corresponding to sampled units were known, along with variance estimators that account for uncertainty due to both subject sampling and the latency of z.

Key words: EM algorithm
Incomplete data
Latent structure
Multiple imputation procedures
Sampling designs
Superpopulation models

Inferences about Latent Populations from Complex Samples*

Introduction

While progress has been made in recent years in estimating latent distributions (e.g., Andersen & Madsen, 1977; Dempster, Laird, & Rubin, 1977; Laird, 1978; Mislevy, 1984, 1985; Sanathanan & Blumenthal, 1978), currently available procedures remain limited to simple random samples and are inaccessible to the typical secondary user of survey data.[1] This paper addresses the problem of estimating distributions under conditions that (1) data have been gathered from a finite population under a complex sampling design and (2) one or more variables of interest are not observed directly, but must be inferred from responses which depend upon them stochastically (e.g., "ability" variables under an item response model).

Two basic approaches exist for handling uncertainty due to sampling in a finite population (see Cassel, Särndal, & Wretman, 1977, for an overview). Under the "fixed population" or "randomization" approach, the only source of variation is researcher's random selection of a sample in accordance with probabilities under a given sampling design. Inferences are based on the distribution of an estimator over the samples that can occur under that design. Under the "superpopulation" approach, the finite

*The author would like to thank R. Darrell Bock for calling his attention to the applicability of multiple imputation procedures to the assessment setting, and Henry Braun, Ben King, Paul Rosenbaum, and Don Rubin for comments on earlier drafts of this presentation.

population itself is considered a sample from a hypothetical superpopulation. A structure is assumed for the superpopulation, its parameters are estimated from the sampled units, and inferences are drawn with respect to remaining uncertainty about nonsampled units.

Extension to the latent variable case is possible under both approaches. Attention is restricted here to the randomization approach, although it must be admitted that the unified treatment of uncertainty from all sources in a Bayesian superpopulation solution (e.g., Mislevy, 1985) is more satisfying. Given the overwhelming predominance of the randomization approach in applied work, however, there is clearly a place for a solution within its framework.

The key idea is to represent knowledge about latent variables in the form of a predictive distribution, conditional on manifest variables, in the manner suggested by Rubin (1977) as a way of handling missing responses in survey data. In a manner also suggested by Rubin (1978), this predictive distribution can be approximated numerically by repeated random draws. Standard complete-data procedures may then be employed to obtain the expected value of any statistic that would have been computed, had values of the latent variables been available. An accompanying variance estimator takes into account uncertainty due to both subject sampling and to the latency of the variables of interest.

Preliminaries

Consider a population $\mathcal{P}$ of N identifiable units, indexed by i. Each is characterized by a pair of real-valued vectors (Z_i, Y_i); values of z are unknown for all units before observations are taken, although values of some components of Y may be known for all units (e.g., stratification variables). $\underset{\sim}{Z}$ and $\underset{\sim}{Y}$ will refer to the population matrices of these values. Interest lies in a function $S = S(\underset{\sim}{Z}, \underset{\sim}{Y})$ of the population values, but data will be obtained from only a sample of units. A sample design assigns probabilities p(d) of selection to members d of $\mathcal{D}$, the set of the 2^N possible subsets from $\mathcal{P}$, and may effect complexities such as stratification and clustering. Let D be the random variable indicating the units selected in the sample. Correspondingly, $(\underset{\sim}{z}_D, \underset{\sim}{y}_D)$ is a random variable and $(\underset{\sim}{z}_d, \underset{\sim}{y}_d)$ a generic value, representing values of z and y from n_D (or n_d) designated sample units. We shall restrict our attention to noninformative sample designs, or those for which $Pr(D = d)$ does not depend on unknown values of $\underset{\sim}{Z}$ or $\underset{\sim}{Y}$; i.e., letting $\underset{\sim}{y}_d^{(1)}$ represent the prior known components of $\underset{\sim}{y}_d$, we have $Pr(D = d | \underset{\sim}{z}_d, \underset{\sim}{y}_d) = Pr(D = d | \underset{\sim}{y}_d^{(1)})$.

Assumption 1: The estimator $s_D = s(\underset{\sim}{z}_D, \underset{\sim}{y}_D)$ could be used to estimate S if $(\underset{\sim}{z}_D, \underset{\sim}{y}_D)$ were observed. We assume s to be unbiased—i.e., $E_{\mathcal{P}}(s_D - S) = 0$—with variance $V = Var_{\mathcal{P}}(s_D - S)$ estimated by $\hat{V} = \hat{V}(\underset{\sim}{z}_D, \underset{\sim}{y}_D)$. A normal approximation is often employed in practice:

$$(s_D - S) \sim N(0,\hat{V}) \quad .$$

Suppose that observations from sampled unit i consist not of (z_i,y_i), but rather of (x_i,y_i), where x_i is a possibly multidimensional secondary random variable that depends stochastically upon z_i. An example would be the observation of right and wrong answers to test questions, assumed to depend upon a latent ability parameter in an item response model. We shall refer to unobserved variables z in the sequel as the latent variables, the observed variables y as collateral variables, and the observed variables x as item responses.

Assumption 2: Item responses x are governed by a model of known parametric form, characterized by possibly unknown parameters β_1. We assume conditional independence with respect to collateral variables and independence over units:

$$p(\underset{\sim}{x}|\underset{\sim}{z},\underset{\sim}{y};\beta_1) = p(\underset{\sim}{x}|\underset{\sim}{z};\beta_1)$$

$$= \prod_i p(x_i|z_i;\beta_1) \quad .$$

The General Solution

This section provides a general solution for estimating functions of variables in fixed populations, when observations are obtained from only a sample of units and values of one or more variables of interest are not directly observed. The solution proceeds in two stages. The first stage approximates conditional or predictive distributions of the latent variables corresponding to sample units; that is,

$$p(\underset{\sim}{z}_d | \underset{\sim}{x}_d, \underset{\sim}{y}_d) \quad .$$

The second stage obtains marginal distributions of statistics that would have been computed, had values of latent variables been available, conditional on observed values. Of particular interest as an estimator of S is the conditional expectation of s given $\underset{\sim}{x}_d$ and $\underset{\sim}{y}_d$:

$$s_d^* = s^*(\underset{\sim}{x}_d, \underset{\sim}{y}_d)$$

$$= E_{\underset{\sim}{z}_d}(s(\underset{\sim}{z}_d, \underset{\sim}{y}_d) | \underset{\sim}{x}_d, \underset{\sim}{y}_d)$$

$$= \int s(\underset{\sim}{z}_d, \underset{\sim}{y}_d)\ p(\underset{\sim}{z}_d | \underset{\sim}{x}_d, \underset{\sim}{y}_d)\ d\underset{\sim}{z} \quad .$$

First, however, an additional assumption is required to compute the conditional distribution $p(\underset{\sim}{z}|\underset{\sim}{x},\underset{\sim}{y})$:

Assumption 3: The distribution of latent variables given collateral variables, or $p(\underset{\sim}{z}|\underset{\sim}{y};\beta_2)$, follows a known form, with possibly unknown parameters β_2. Furthermore, independence is assumed over units:

$$p(\underset{\sim}{z}|\underset{\sim}{y};\beta_2) = \prod_i p(z_i|y_i;\beta_2) \quad .$$

This assumption resembles those used in superpopulation models for sampling from finite populations (e.g., Ericson, 1969; Royall, 1970).

Stage 1: Estimating Conditional Distributions

The task of stage 1 is to approximate the conditional density $p(\underset{\sim}{z}_d|\underset{\sim}{x}_d,\underset{\sim}{y}_d)$. Dropping the subscripts d on $\underset{\sim}{x}$ and $\underset{\sim}{y}$, and denoting (β_1,β_2) by $\underset{\sim}{\beta}$, we note first that

$$p(\underset{\sim}{z}|\underset{\sim}{x},\underset{\sim}{y}) = \int p(\underset{\sim}{z}|\underset{\sim}{x},\underset{\sim}{y};\underset{\sim}{\beta})\, p(\underset{\sim}{\beta}|\underset{\sim}{x},\underset{\sim}{y})\, d\underset{\sim}{\beta}$$

$$= \int p(\underset{\sim}{x}|\underset{\sim}{z},\underset{\sim}{y};\underset{\sim}{\beta})\, p(\underset{\sim}{z}|\underset{\sim}{y};\underset{\sim}{\beta})\, p^{-1}(\underset{\sim}{x}|\underset{\sim}{y};\underset{\sim}{\beta})\, p(\underset{\sim}{\beta}|\underset{\sim}{x},\underset{\sim}{y})\, d\underset{\sim}{\beta}$$

[Bayes theorem]

$$= \int \prod_i p(x_i|z;\beta_1)\, p(z|y_i;\beta_2)\, p^{-1}(\underset{\sim}{x}|\underset{\sim}{y};\underset{\sim}{\beta})\, p(\underset{\sim}{\beta}|\underset{\sim}{x},\underset{\sim}{y})\, d\underset{\sim}{\beta} \quad , \tag{1}$$

[Assumptions 2 and 3]

where

$$p(\underset{\sim}{x}|\underset{\sim}{y};\underset{\sim}{\beta}) = \int \prod_i p(x_i|z;\beta_1)\, p(z|y_i;\beta_2)\, dz \quad .$$

Now by Bayes Theorem,

$$p(\underset{\sim}{\beta}|\underset{\sim}{x},\underset{\sim}{y}) = p(\underset{\sim}{x}|\underset{\sim}{y};\underset{\sim}{\beta})\, p(\underset{\sim}{\beta}|\underset{\sim}{y})\, p^{-1}(\underset{\sim}{x}|\underset{\sim}{y}) \quad ,$$

where

$$p(\underset{\sim}{x},\underset{\sim}{y}) = \int\int p(\underset{\sim}{x}|\underset{\sim}{z},\underset{\sim}{y};\underset{\sim}{\beta})\, p(\underset{\sim}{z}|\underset{\sim}{y};\underset{\sim}{\beta})\, p(\underset{\sim}{\beta}|\underset{\sim}{y})\, d\underset{\sim}{z}\, d\underset{\sim}{\beta} \quad ;$$

this latter quantity does not depend on $\underset{\sim}{\beta}$, so we can write simply

$$p(\underset{\sim}{\beta}|\underset{\sim}{x},\underset{\sim}{y}) = K \int \prod_i p(\underset{\sim}{x}_i|z;\beta_1)\, p(z|y_i;\beta_2)\, d\underset{\sim}{z}\; p(\underset{\sim}{\beta}|\underset{\sim}{y})$$

$$= K\, p(\underset{\sim}{x}|\underset{\sim}{y};\underset{\sim}{\beta})\, p(\underset{\sim}{\beta}|\underset{\sim}{y}) \quad . \qquad (2)$$

Substituting (2) into (1) and noting that $p(\beta|y) \equiv p(\beta)$ by the noninformativity of the sampling design, we obtain

$$p(\underset{\sim}{z}|\underset{\sim}{x},\underset{\sim}{y}) = K \int \prod_i p(x_i|z;\beta_1)\, p(z|y_i;\beta_2)\, p(\underset{\sim}{\beta})\, d\underset{\sim}{\beta} \quad . \qquad (3)$$

Stage 2: Estimating Marginal Distributions

The task of stage 2 is to obtain the expected value of $s(\underset{\sim}{z},\underset{\sim}{y})$ given the observed data $(\underset{\sim}{x},\underset{\sim}{y})$ and $p(\underset{\sim}{z}|\underset{\sim}{x},\underset{\sim}{y})$ from stage 1. A fine

$$s_d^* = s^*(\underset{\sim}{x}_d, \underset{\sim}{y}_d)$$

$$= E[s(\underset{\sim}{z}_d, \underset{\sim}{y}_d) | \underset{\sim}{x}_d, \underset{\sim}{y}_d]$$

$$= \int s(\underset{\sim}{z}_d, \underset{\sim}{y}_d)\ p(\underset{\sim}{z}_d | \underset{\sim}{x}_d, \underset{\sim}{y}_d)\ d\underset{\sim}{z}_d \quad . \qquad (4)$$

In words, s_d^* is the average of $s(\underset{\sim}{z}_d, \underset{\sim}{y}_d)$ over all possible values of $\underset{\sim}{z}_d$ for the sample, with each value weighted by its relative likelihood given the observations. To the extent that s is a reasonable estimator of S, then, so is s* in the latent variable case, since s* is the best quadratic-loss estimator of s_d given $\underset{\sim}{x}_d$ and $\underset{\sim}{y}_d$.

The magnitude of the uncertainty in s* may be approximated along the line followed by Hertzog and Rubin (1983). There are two sources of variation in s*. First there is variation due to sampling. By Assumption 1,

$$E_{\mathcal{D}}(s_D - S)^2 \approx \hat{V} \quad .$$

Secondly, there is variation due to the latency of z even after the data $\underset{\sim}{x}_d$ have been observed. For any given sample d, we define

$$\hat{W}_d = E_{\underset{\sim}{z}_d} (s_d^* - s_d)^2$$

$$= \int (s_d^* - s(\underset{\sim}{z}_d, \underset{\sim}{y}_d))^2 \, p(\underset{\sim}{z}_d | \underset{\sim}{x}_d, \underset{\sim}{y}_d) \, d\underset{\sim}{z}_d \quad .$$

Herzog and Rubin define the "compromise" estimator $\hat{U}$ of total variance as

$$\hat{U} = \hat{W}_d + \hat{V}_d \quad .$$

In the context of the analysis of nonresponse, Hertzog and Rubin demonstrate good approximation of $s_D^* \sim N(S, \hat{U})$ to nominal probability levels under a linear population model and an ignorable model for the nonresponse process.

Closed-form evaluation of s* and $\hat{U}$ will not be possible except in unusual cases. A numerical approximation with attractive properties for applied work is Monte Carlo integration:

$$s^*(\underset{\sim}{x},\underset{\sim}{y}) = R^{-1} \sum_{r}^{R} s(\underset{\sim}{z}^*_r,\underset{\sim}{y})$$

where

$$\underset{\sim}{z}^*_r = (z^*_{1r},\ldots,z^*_{nr})$$

is a value selected at random from $p(\underset{\sim}{z}|\underset{\sim}{x},\underset{\sim}{y})$. The sampling process is carried out R times to yield R replicate pseudo-data sets of the form $(\underset{\sim}{z}^*_r,\underset{\sim}{y})$. The estimator s is evaluated with each replicate data set in turn, and the results are averaged to provide an estimate of $s(\underset{\sim}{z},\underset{\sim}{y})$ and therefore of $S(\underset{\sim}{Z},\underset{\sim}{Y})$.

Production of the replicate pseudo-data sets can be carried out in two steps. First a value $\underset{\sim}{\beta}^*_r$ is selected at random from $p(\underset{\sim}{\beta})$. Second, because the unit distributions $p(z|x_i,y_i;\underset{\sim}{\beta})$ are independent conditional on $\underset{\sim}{\beta}$, a value z^*_{ir} can be selected at random from $p(z|x_i,y_i,\underset{\sim}{\beta} = \underset{\sim}{\beta}^*_r)$ for each unit in the sample separately. When $\underset{\sim}{\beta}$ is well-determined by $\underset{\sim}{x}_d$ and $\underset{\sim}{y}_d$, the generation of pseudo-data sets with $\underset{\sim}{\beta}^*_r \equiv \hat{\underset{\sim}{\beta}}$, the maximum likelihood estimate of β proves quite adequate.

By similar reasoning,

$$\hat{U} = \hat{W}_d + \hat{V}_d$$

$$\approx R^{-1} \sum_r [s(z^*_r, y) - s^*(x, y)]^2 + R^{-1} \sum_r \hat{V}[s(z^*_r, y)] \quad . \quad (5)$$

Again in words, one approximates the variance of s* by the average of $\hat{V}(s)$ values for s calculated on the R pseudo-data sets, increased by the variance of the pseudo estimates of s. When $\hat{V}(s)$ is given by a resampling scheme such jackknifing or balanced half replication, a less costly approximation for the sampling variance of s is $\hat{V}(s(z^*_r, y))$ as computed from one randomly selected pseudo-data set. These procedures will be recognized as a variation of "multiple imputation" procedures for missing data (Hertzog & Rubin, 1983; Rubin, 1977, 1978), with latent variables considered 100-percent missing--that is, values are not observed from any respondent.

An important practical advantage of the multiple-imputation approach is that the same collection of pseudo-data sets can be used to estimate several different statistics S. A file containing R replicates would thus allow the secondary user to estimate without additional special programming, any statistic he or she would have

liked to calculate had $\underset{\sim}{z}$ been observable, along with an indication of its precision that takes the latency of $\underset{\sim}{z}$ into account.

A Numerical Example

This section applies the procedures outlined above to a small example with data from the Profile of American Youth (U.S. Department of Defense, 1982). For each respondent, the data consist of two demographic variables y (ethnicity and sex) and four responses x to items on an aptitude test, assumed to be governed by a single latent aptitude variable z. The item response model and conditional estimation results are taken from Mislevy (1985); the interested reader is referred to this source for additional detail. A simplified sampling design (though still more complex than simple random sampling) is assumed here for purposes of illustration.

The Data

The data we consider were obtained as part of the Profile of American Youth, a survey of the aptitudes of a national probability sample of Americans aged 16 through 23 in July, 1980. Table 1 presents counts of the sixteen possible response patterns to four items from the Arithmetic Reasoning subtest of the Armed Services Vocational Aptitude Batter (ASVAB), Form 8A, from samples of white males and females and Black males and females. A 1 denotes a correct response, while a 0 denotes an incorrect response. Though multiple stages of sampling were employed in

the actual design of the study, we shall treat these four groups as a stratified random sample from a target population, with Blacks sampled at a rate of double that of whites.

Insert Table 1 about here

The Item Response Model

Let x_{ij} represent the response of person i to item j. It is assumed that responses are governed by the three-parameter logistic item response model (Birnbaum, 1968), which gives the probability of a correct response as

$$P(x_{ij} = 1 \mid z_i; a_j, b_j, c_j) = P_{ij}$$

$$= c_j + (1 - c_j)/\{1 + \exp[-1.7a_j(z_i - b_j)]\}$$

and the probability of an incorrect response as

$$P(x_{ij} = 0 \mid z_i; a_j, b_j, c_j) = 1 - P_{ij} \quad ,$$

where a_j, b_j, and c_j are parameters that characterize the regression of a correct response to item j on z. These parameters,

over all four items, are denoted by β_1 in the general solution given above. Under the usual assumption of conditional independence, the probability of a vector of item responses x_i from person i is given by

$$P(x_i|z_i;\beta_1) = \prod_j P_{ij}^{x_{ij}}(1 - P_{ij})^{1-x_{ij}} \quad .$$

Estimates of the item parameters, based on responses from an independent sample of 1178 persons and computed with the BILOG computer program (Mislevy & Bock, 1982), appear as Table 2.

Insert Table 2 about here

Conditional Distributions

Conditional multivariate normality under a saturated homoscedastic model is assumed so that

$$p(z|y_i;\underset{\sim}{\gamma},\sigma) = (2\pi\sigma)^{-1/2} \exp[-(z - \underset{\sim}{\gamma}'\underset{\sim}{t}_i)^2/2\sigma^2] \quad ,$$

where $\underset{\sim}{t}_i = (t_{i1},t_{i2},t_{i3},t_{i4})$ is a design vector associated with respondent i, taking values as follows:

$$t_{i1} = 1$$

$$t_{i2} = \begin{cases} .5 \text{ if white} \\ -.5 \text{ if Black} \end{cases}$$

$$t_{i3} = \begin{cases} .5 \text{ if male} \\ -.5 \text{ if female} \end{cases}$$

$$t_{i4} = \begin{cases} .25 \text{ if white male or Black female} \\ -.25 \text{ if Black male or white female;} \end{cases}$$

and where $\underset{\sim}{\gamma} = (\gamma_1, \gamma_2, \gamma_3, \gamma_4)$ represents a constant term, an ethnicity effect, a sex effect, and an ethnicity-by-sex interaction. The common within cell standard deviation is denoted σ. Together, $\underset{\sim}{\gamma}$ and σ play the role of β_2.

Under these assumptions, the conditional likelihood of the data in Table 1 is given by

$$L = \prod_i (x_i | y_i; \underset{\sim}{\beta})$$

$$= \prod_i \int_z p(x_i | z; \beta_1)\ p(z | y_i; \underset{\sim}{\gamma}, \sigma)\ dz \quad .$$

Equating first derivatives of log L to zero yields likelihood equations. For $\underset{\sim}{\gamma}$, after simplification,

$$\hat{\underset{\sim}{\gamma}} = (\underset{\sim}{T}'\underset{\sim}{T})^{-1}\underset{\sim}{T}'\hat{\underset{\sim}{\mu}} \quad , \tag{6}$$

where $\underset{\sim}{T} = (\underset{\sim}{\quad} \ldots \underset{\sim}{\quad}_n)'$ and $\hat{\underset{\sim}{\mu}} = (\hat{\mu}_1,\ldots,\hat{\mu}_n)$ with

$$\hat{\mu}_i = \int_z z\ p(z|x_i,y_i;\beta_1,\underset{\sim}{\gamma},\sigma)\ dz \quad . \tag{7}$$

For σ,

$$\hat{\sigma}^2 = n^{-1} \sum_i \int_z (z - \hat{\mu}_i)^2\ p(z|x_i,y_i;\beta_1,\underset{\sim}{\gamma},\sigma)\ dz \quad . \tag{8}$$

It will be noted that $\underset{\sim}{\gamma}$ and σ appear in the right-hand sides of (7) and (8), necessitating iterative solution. An EM solution proceeds in repeated cycles of the form

E-step: For provisional estimates $\hat{\underset{\sim}{\gamma}}^{(t)}$ and $\hat{\sigma}^{(t)}$, approximate the conditional density by

$$p(z|x_i,y_i;\beta_1,\hat{\underset{\sim}{\gamma}}^{(t)},\hat{\sigma}^{(t)}) \quad .$$

M-step: Taking this approximation as known, evalute (6)-(8) to obtain improved estimates $\hat{\underset{\sim}{\gamma}}^{(t+1)}$ and $\hat{\sigma}^{(t+1)}$.

With β_1 taken as known, the only unknowns are γ and σ, parameters of a distribution in the exponential family; convergence of the EM algorithm is thereby guaranteed (Dempster, Laird, & Rubin, 1977). Resulting estimates are

$$\hat{\underset{\sim}{\gamma}} = (-.13, .92, .13, .43)$$

and

$$\hat{\sigma} = .82 \quad ;$$

implied cell means are

White males	.51
White females	.15
Black males	-.63
Black females	-.55

Generation of Pseudo Data

Let $U_1, \ldots U_{40}$ be a grid of points from -4.875 to 4.875 in equally-spaced steps of .25. The continuous distributions given by $p(z|x_i, y_i; \hat{\underset{\sim}{\beta}})$ for each respondent in the sample may be approximated by discrete distributions over a finite number of points--i.e., histograms--as follows:

$$P(U_q|x_i,y_i;\hat{\beta}) = \frac{p(U_q|x_i,y_i;\hat{\beta})}{\sum_r p(U_r|x_i,y_i;\hat{\beta})} \quad .$$

Five pseudo-data sets were generated by taking five values at random from such a histogram for each respondent in a two-step procedure. In the first step of obtaining z^*_{ir}, a random number t_{ir} from the unit interval was generated to target a block in the histogram, namely that block k_{ir} such that

$$\sum_{q=1}^{k_{ir}-1} p(U_q|x_i,y_i;\hat{\beta}) < t_{ir} \leq \sum_{q=1}^{k_{ir}} p(U_q|x_i,y_i;\hat{\beta}) \quad .$$

In the second step, a second random number s from the unit interval was generated to specify a point in block k_{ir}:

$$z^*_{ir} = U_{k_{ir}} + .25(s - .5) \quad .$$

Table 3 gives likelihoods, a conditional distribution, a predictive distribution, and pseudo values for a typical respondent.

Insert Table 3 about here

Estimation of Marginal Distributions

As noted above, it is desired to estimate the overall mean of the population under the assumption that sampling was random within the strata defined by the cells of the demographic design, with sampling probabilities doubled for Blacks. If values of z had been observed rather than x, the estimate of the mean would have been

$$\bar{z} = \frac{\bar{z}_{11}}{3} + \frac{\bar{z}_{12}}{3} + \frac{\bar{z}_{21}}{6} + \frac{\bar{z}_{22}}{6} \quad ,$$

where subscripts identify cells as follows:

11 = white males,

12 = white females,

21 = Black males, and

22 = Black females.

Ignoring finite population corrections, an estimate of the variance of this estimator is given by

$$\hat{\mathrm{Var}}(\bar{z}) = \frac{s_{11}^2}{9n_{11}} + \frac{s_{12}^2}{9n_{12}} + \frac{s_{21}^2}{36n_{21}} + \frac{s_{22}^2}{36n_{22}} \quad , \qquad (9)$$

where n_{jk} is the sample size in cell jk and s_{jk} is the estimated standard deviation.

Table 4 gives cell means and standard deviations as estimated from the five pseudo-data sets. The expectation of the sample mean $\bar{z}$, given observed data, is the average of the five pseudo-sample means, or .0407. The variance associated with this estimate is given by averaging values of (9) over pseudo-data sets, or .0009, plus the variance among the estimates $\bar{z}^*_r$ or .0008 to yield a final value of .0017.

Insert Table 4 about here

Discussion

A necessary requirement for consistent estimates under the approach outlined above is the correct specification of $p(z|y)$. When the dimensionalities of z and y are low (e.g., five latent variables and five collateral variables), it is possible to obtain a detailed nonparametric approximation of this conditional distribution (Mislevy, 1984). When dimensionalities of z and y run into the hundreds, however, as in a large-scale general-purpose survey such as the National Assessment of Educational Progress (NAEP), simplifications and computing approximations cannot be avoided. This section, therefore, suggests some computing approximations and discusses their effects on the estimation of statistics such as differences in subpopulation means.

Point estimation of $\hat{\beta}$. The integration over β required in (1) to obtain $p(z|x,y)$ can be avoided in large samples when $p(\beta|x,y)$ is well-determined from the data. In such cases the imprecision associated with an individual's value of z that can be attributed to variation in $p(\beta|x,y)$ is negligible, and one may sample values from the more tractible distribution $p(z|x,y;\hat{\beta})$, where $\hat{\beta}$ represents the maximum likelihood or Bayes modal estimate estimate of β.

Solutions can be obtained by means of a generalized EM algorithm (Dempster, Laird, & Rubin, 1977). Bock and Aitkin (1981) give procedures for solving (6) when β_2 is known, and Mislevy (1985) gives procedures for solving (7) when β_1 is known and $p(z|y_i;\beta_2)$ is $MVN(t_i\Gamma,\Sigma)$, with t_i a vector function of y_i expressing the dependence of the conditonal mean upon the effects Γ of collateral variables. These presentations are readily combined to give a joint solution for β_1 and β_2. Such an integrated solution for the special case $p(z|y_i) \sim$ iid $N(\mu,\sigma)$ may be found in Rigdon and Tsutakawa (1983).

Multivariate normal conditional distributions. In principle, $p(z|y)$ gives the distribution of the latent variables at all possible values of y. As the dimensionality of z increases, considerations of tractability make it increasingly attractive to model these conditional distributions as multivariate normal (MVN)

with a ·on dispersion matrix. It must be emphasized that this is not the same as assuming MVN marginal distributions among the latent variables z. Indeed, as the number of collateral variables increases, and to the degree they are correlated with z, the estimated marginal distribution of z can become arbitrarily close to a true (smooth) distribution of any form.

Omission of selected interactions. Even under the assumption of conditional multivariate normality, increasing dimensionality of y rapidly overburdens available computing resources if all main effects and interactions of all orders are modeled in $p(z|y)$. A reasonable expedient is to omit all higher level interactions (interactions of order three or higher are rare in behavioral research) and, if necessary, many second order interactions as well. If main effects only are modeled, analyses of pseudo-data sets will capture them correctly but may be in error as to interaction effects. The degree of error is reduced to an extent depending on two factors:

1. It will be recalled that for each respondent, stage 2 combines information from the estimated conditional distribution $p_0(\underset{\sim}{z}|\underset{\sim}{y})$, with information from item responses via $p(\underset{\sim}{x}|\underset{\sim}{z})$ in order to obtain the predictive distribution $p(\underset{\sim}{z}|\underset{\sim}{x},\underset{\sim}{y})$ from which random values are selected. Assuming $p(\underset{\sim}{x}|\underset{\sim}{z})$ is correctly specified, one could use the resulting

pseudo-data set to obtain the empirical distribution $p'(\underset{\sim}{z}|\underset{\sim}{y})$. If $p_0(z|y)$ has been correctly specified, $p_0(\underset{\sim}{z}|\underset{\sim}{y})$ and $p'(\underset{\sim}{z}|\underset{\sim}{y})$ will agree. If $\hat{p}_0(\underset{\sim}{z}|\underset{\sim}{y})$ has not been correctly specified, information from $\underset{\sim}{x}$ will cause $p'(\underset{\sim}{z}|\underset{\sim}{y})$ to differ from $p_0(\underset{\sim}{z}|\underset{\sim}{y})$ value in the direction of the true distribution, by an amount equal to that achieved in one EM cycle of estimation. An approximation of this amount can be obtained by applying the procedures outlined by Dempster et al. (1977, pp. 10-11) to the model that includes the omitted terms.

2. Attenuation of estimates of omitted interactions will also be ameliorated to the extent that such effects are correlated with effects that are not omitted. This follows from results on the consequences of specification errors in linear regression models. If data are generated in accordance with parameter estimates under a model that is misspecified by the omission of certain effects, subsequent analyses of these data with the correct model will yield improved estimates of all effects unless the omitted effects are uncorrelated with those not omitted.

Omission of selected collateral variables. It may be reasonable to omit nonessential variables from the conditional

estimation when the total number of collateral variables is large. Statistics s* based on included variables only will not suffer from this omission; subgroup differences, for example, will be captured 100-percent if these effects were included in the conditioning. For the reasons cited above, the attenuation of statistics based on omitted variables will not be serious when each respondent provides several item responses and as the number of included collateral variables increases; subgroup differences on omitted variables, for example, will suffer negligible attenuation if included variables are chosen carefully.

Use of reduced variables. The careful choice of variables to include in the conditional estimation includes two considerations. First, effects deemed important in their own right should be explicitly modeled if possible so that statistics based on their joint distributions will suffer no attenuation at all. Examples might include key demographic effects, treatment effects, and salient interactions. Second, rather than simply omitting remaining variables it is preferrable to include a few well-chosen linear combinations of remaining variables; e.g., the first four principle components, or factor scores based on the first three principle factors. Such use of reduced variables guarantees efficient use of the limited number of effects that can be modeled in recapturing to a great extent a wide range of potential statistics s*.

References

Andersen, E. B., & Madsen, M. (1977). Estimating the parameters of a latent population distribution. Psychometrika, 42, 357-374.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick, Statistical theories of mental test scores. Reading, MA: Addison-Wesley.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. Psychometrika, 46, 443-459.

Cassel, C. M., Sarndal, C. E., & Wretman, J. H. (1977). Foundations of inference in survey sampling. New York: Wiley.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society, Series B, 39, 1-38.

Ericson, W. A. (1969). Subjective Bayesian models in sampling finite populations. Journal of the Royal Statistical Society, Series B, 31, 195-224.

Hartzog, T., & Rubin, D. B. (1983). Using multiple imputation to handle nonresponse in sample surveys. In W. G. Madow, I. Olkin, & D. B. Rubin (Eds.), Incomplete data in sample surveys. Volume II. Theory and bibliographies. New York: Academic Press.

Laird, N. M. (1978). Nonparametric maximum likelihood estimation of a mixing distribution. Journal of the American Statistical Association, 73, 805-811.

Mislevy, R. J. (1984). Estimating latent distributions. Psychometrika, 49, 359-381.

Mislevy, R. J. (1985a, in press). Estimation of latent group effects. Journal of the American Statistical Association.

Mislevy, R. J. (1985b, in preparation). A Bayesian treatment of latent variables in sample surveys (Research Report). Princeton, NJ: Educational Testing Service.

Mislevy, R. J., & Bock, R. D. (1982). BILOG: Item analysis and test scoring with binary logistic models [Computer program]. Morresville, IN: Scientific Software.

Royall, R. M. (1970). On finite population sampling theory under under certain linear regression models. Biometrika, 57, 377-387.

Rigdon, S. E., & Tsutakawa, R. K. (1983). Parameter estimation in latent trait models. Psychometrika, 48, 567-574.

Rubin, D. B. (1977). Formalizing subjective notions about the effect of non-respondents in sample surveys. Journal of the American Statistical Association, 72, 538-543.

Rubin, D. B. (1978). Multiple imputations in sample surveys. Proceedings of the Survey Research Methods Section of the American Statistical Association, 20-34.

Rubin, D. B. (1980). Handling nonresponse in sample surveys by multiple imputation. U.S. Bureau of the Census Monograph.

Sanathanan, L., & Blumenthal, N. (1978). The logistic model and latent structure. Journal of the American Statistical Association, 73, 794-798.

Spencer, B. (1984). Simplifying complex samples with the bootstrap. Proceedings of the Survey Research Methods Section of the American Statistical Association, 484-488.

U.S. Department of Defense, Office of the Assistant Secretary of Defense (Manpower, Reserve Affairs, and Logistics). (1982). Profile of American youth. Washington, DC.

Footnote

[1]But see Spencer (1984) on bootstrapping the aforementioned procedures.

Table 1

Counts of Observed Response Patterns

Item Response 1	2	3	4	White Males	White Females	Black Males	Black Females
0	0	0	0	23	20	27	29
0	0	0	1	5	8	5	8
0	0	1	0	12	14	15	7
0	0	1	1	2	2	3	3
0	1	0	0	16	20	16	14
0	1	0	1	3	5	5	5
0	1	1	0	6	11	4	6
0	1	1	1	1	7	3	0
1	0	0	0	22	23	15	14
1	0	0	1	6	8	10	10
1	0	1	0	7	9	8	11
1	0	1	1	19	6	1	2
1	1	0	0	21	18	7	19
1	1	0	1	11	15	9	5
1	1	1	0	23	20	10	8
1	1	1	1	86	42	2	4
TOTAL				263	228	140	145

Table 2

Item Parameters

Item	a	b	c
1	1.27	-.13	.20
2	1.45	.42	.20
3	2.49	.71	.20
4	2.27	.62	.20

Table 3

Likelihood, Conditional Density, and Prediction Density for a Typical Respondent

Collateral variables y: Black, female Item responses x = 1100

U_k	$p(x\|U_k)$	$p(U_k\|y)$	$p(U_k\|x,y)$
-4.875	.026	.000	.000
-4.625	.026	.000	.000
-4.375	.026	.000	.000
-4.125	.026	.000	.000
-3.875	.026	.000	.000
-3.625	.026	.000	.000
-3.375	.026	.000	.000
-3.125	.026	.001	.000
-2.875	.026	.002	.001
-2.625	.026	.005	.002
-2.375	.027	.010	.003
-2.125	.027	.029	.006
-1.875	.028	.032	.011
-1.625	.030	.050	.018
-1.375	.034	.071	.028
-1.125	.039	.092	.043
-0.875	.049	.110	.065
-0.625	.066	.120	.095
-0.375	.092	.121	.134
-0.125	.130	.113	.176
0.125	.168	.097	.194
0.375	.171	.073	.149
0.625	.113	.045	.061
0.875	.042	.022	.012

Table 3 (continued)

1.125	.010	.010	.001
1.375	.002	.005	.000
1.625	.000	.002	.000
1.875	.000	.001	.000
2.125	.000	.000	.000
2.375	.000	.000	.000
2.625	.000	.000	.000
2.875	.000	.000	.000
3.125	.000	.000	.000
3.375	.000	.000	.000
3.625	.000	.000	.000
3.875	.000	.000	.000
4.125	.000	.000	.000
4.375	.000	.000	.000
4.625	.000	.000	.000
4.875	.000	.000	.000

Mean and standard deviation of $P(U_k|x,y)$: -.223, .614

Five randomly selected points: .058, .333, -.352, .009, .176

Table 4

Estimated Population and Subpopulation Means

	Pseudo-Data Set									
	1		2		3		4		5	
Subpopulation	Mean	Var.	Mean	Var.	Mean	Var.	Mean	Var.	Mean	Var.
White males	.4840	.6928	.5276	.8158	.5461	.7547	.5403	.7359	.4964	.6825
White females	.0804	.7570	.2087	.6814	.1964	.6170	.2078	.6973	.1351	.7056
Black males	-.6161	.6054	-.6357	.6527	-.5792	.6156	-.5758	.6935	-.6178	.5573
Black females	-.5509	.5510	-.5866	.5898	-.4833	.6139	-.4911	.5220	-.4878	.6269
Population mean ($\bar{z}$)	-.0064		.0417		.0704		.0716		.0262	
Var ($\bar{z}$)	.0009		.0009		.0008		.0009		.0009	