

Alternative Designs for the Consumer Expenditure Survey

F. Jay Breidt
Department of Statistics
Iowa State University

November 15, 1999
Federal Committee on Statistical Methodology Research Conference
Arlington, Virginia

*Sponsored by ASA/NSF Senior Research Fellowship
Bureau of Labor Statistics and Census Bureau*

Outline

- Background
 - current design of Consumer Expenditure Survey
 - efficiency of current design
- Redesign problem
 - loss of key stratification variables
 - quantify loss of efficiency?
- Design alternatives
 - alternative stratification schemes
 - two-phase sampling for stratification
 - combination
- Summary

Consumer Expenditure Survey

- Universe: U.S. civilian noninstitutional population
- Two-stage probability sample of U.S. households
- First stage: 101 PSU's
 - counties, groups of counties, or independent cities
- Second stage: interview and diary surveys
 - interview: 8910 households per calendar quarter (large expenses, regular expenses)
- Consumer units: families or other groups
 - make joint expenditure decisions
 - spend pooled income

More on the Second-Stage Design

- From 1990 short form, had home value or contract rent.
- Within PSU's, systematic sample from sorted list frame:

	Vac.	Renter	<\$400
*	Occ.	Renter	<\$400
**	Occ.	Renter	\$400–\$599
	Vac.	Renter	\$400–\$599
	Vac.	Owner	<\$45,000
*	Occ.	Owner	<\$45,000
**	Occ.	Owner	\$45,000–\$79,999
	Vac.	Owner	\$45,000–\$79,999
	Vac.	Renter	\$600–\$749
*	Occ.	Renter	\$600–\$749
**	Occ.	Renter	\$750+
	Vac.	Renter	\$750+
	Vac.	Owner	\$80,000–\$124,999
*	Occ.	Owner	\$80,000–\$124,999
**	Occ.	Owner	\$125,000–\$174,999
	Vac.	Owner	\$125,000–\$174,999
	Vac.	Owner	\$175,000+
*	Occ.	Owner	\$175,000+
	Vac.		

– * sorted by household size: 1, 2, 3, 4+

– ** sorted by household size: 4+, 3, 2, 1

Problem for Redesign

- Changes in the Census questionnaires:
 - property value and rent on Short Form in 1990
 - only on Long Form in 2000
 - alternative stratification?
- Short-form stratification
 - short form will be very sparse
 - no equally useful stratifiers on the short form

Alternative Stratification

- Long-form stratification
 - compute long-form estimates at some small domain level (e.g., block groups)
 - based on LF estimate, all households in domain go into one stratum
 - how much loss of efficiency compared to direct stratification on X_j ?

A Small Domain Superpopulation Model

- Stratification variables: if $j \in U_i$, then

$$X_j = \gamma_i + \eta_j,$$

where

- $\{\gamma_i\}$ iid Normal($0, \sigma_\gamma^2$)
 - $\{\eta_j\}$ iid ($0, \sigma_\eta^2$) (not necessarily normal)
 - $\{(\gamma_i, M_i)\}$ independent of $\{\eta_j\}$
- Dependence structure:
 - within a given domain U_i ,

$$\text{Corr}(X_j, X_k) = \frac{\sigma_\gamma^2}{\sigma_\gamma^2 + \sigma_\eta^2}$$

- independent across domains
- Regression relation:

$$Y_j = \alpha + \beta X_j + \epsilon_j = \alpha + \rho_{xy} \frac{\sigma_y}{\sigma_x} X_j + \epsilon_j$$

Numerical Example

- Take $\rho_{xy} = 0.6$, $\sigma_x^2 = 315.6536$, and μ_M large
- Direct stratification on X_j :

$$\begin{aligned}\text{deff}_y &= 1 - \rho_{xy}^2(1 - \text{deff}_x) \\ &= 1 - (0.6)^2(1 - 0.124) = 0.68464.\end{aligned}$$

- To get same variance stratifying on long form estimate as stratifying on X_j directly, need to increase sample size
- Ratio of new sample size to old sample size is ratio of new deff_y to old deff_y :

Within / Between	Within-Domain X Correlation	deff Ratio
0.0	1.00	1.00
0.1	0.91	1.08
1.0	0.50	1.35
3.0	0.25	1.43
∞	0.00	1.46

Empirical Research

- Normality of random effects is suspect
 - makes certain computations tractable
- Given frame, can assess stratifications empirically
 - compute deff_x via Monte Carlo
 - repeatedly simulate systematic samples from ordered list frame
- Estimate correlations based on past CE data
- Given deff_x and correlation ρ_{xy} , compute deff_y

Two-Phase Sampling for Stratification

- In 2000 redesign, I_{hij} and N_h are not available on frame
 - two-phase sampling for stratification is standard procedure in this case
- Phase One: long form respondents
 - X_j and hence I_{hij} available for all long form respondents
 - estimate of N_h can be obtained from long form
- Phase Two: subsample long form respondents
 - efficient “old-style” stratification can be used to select subsample
 - Y_j is measured for each subsampled household
- Potential problem: selection bias in long form respondents

Selection Problems

- *Nonresponse*: Some sampled households do not respond to the long form
- *Conversion*: Some non-sampled households do respond to the long form
 - long form may go to wrong address
 - follow-up interviewer may (deliberately or accidentally) go to wrong address
- Problematic if selection probability is related to study variables of interest

Modeling the Selection Mechanism

- Let

$$A_{ij} = \begin{cases} 1, & \text{if household } j \text{ of domain } i \text{ is sampled for the long form,} \\ 0, & \text{otherwise} \end{cases}$$

and

$$R_{ij} = \begin{cases} 1, & \text{if household } j \text{ of domain } i \text{ responds,} \\ 0, & \text{otherwise} \end{cases}$$

- Model:

$$R_{ij} \mid A_{ij} \sim \text{Bernoulli}(\pi_{ij})A_{ij} + \text{Bernoulli}(\kappa_{ij})(1 - A_{ij})$$

- sampled household ($A_{ij} = 1$) responds with probability π_{ij}
- non-sampled household ($1 - A_{ij} = 1$) is converted to a respondent with probability κ_{ij}

- Assume

$$\pi_{ij} = \pi + \frac{\delta_{ij}^{\dagger}}{\sqrt{D}}$$

and

$$\kappa_{ij} = \kappa + \frac{\delta_{ij}^*}{\sqrt{D}}$$

Asymptotic Design Mean Squared Error

- The following approximation holds for large D and $M_i \equiv \mu_M$:

$$\begin{aligned} \text{design mse} &\simeq \left[\frac{1}{\rho} \{ \lambda \text{Cov}(Y_j, \pi_{ij}) + (1 - \lambda) \text{Cov}(Y_j, \kappa_{ij}) \} \right]^2 \\ &+ \frac{1}{\rho^2 N} \text{Var}(Y_j) \lambda (1 - \lambda) (\pi - \kappa)^2 \\ &+ \frac{1}{\rho^2 N} \text{Var}(Y_j) \{ \lambda \pi (1 - \pi) + (1 - \lambda) \kappa (1 - \kappa) \} \\ &+ \sum_{h=1}^H \frac{\omega_h}{\rho N} \text{Var}_{U_h}(Y_j) \frac{1 - \varphi_h}{\varphi_h}, \end{aligned}$$

where

$$\rho = \lambda \pi + (1 - \lambda) \kappa,$$

$$\omega_h = \text{E}[I_{hij}],$$

$$\text{Var}_{U_h}(Y_j) = \frac{\text{E}[Y_j^2 I_{hij}]}{\text{E}[I_{hij}]} - \frac{\text{E}^2[Y_j I_{hij}]}{\text{E}^2[I_{hij}]},$$

and

$$\varphi_h = \frac{1}{\rho \mu_M \text{E}[I_{hij}]} \lim_{D \rightarrow \infty} \frac{m_h}{D}$$

Special Cases

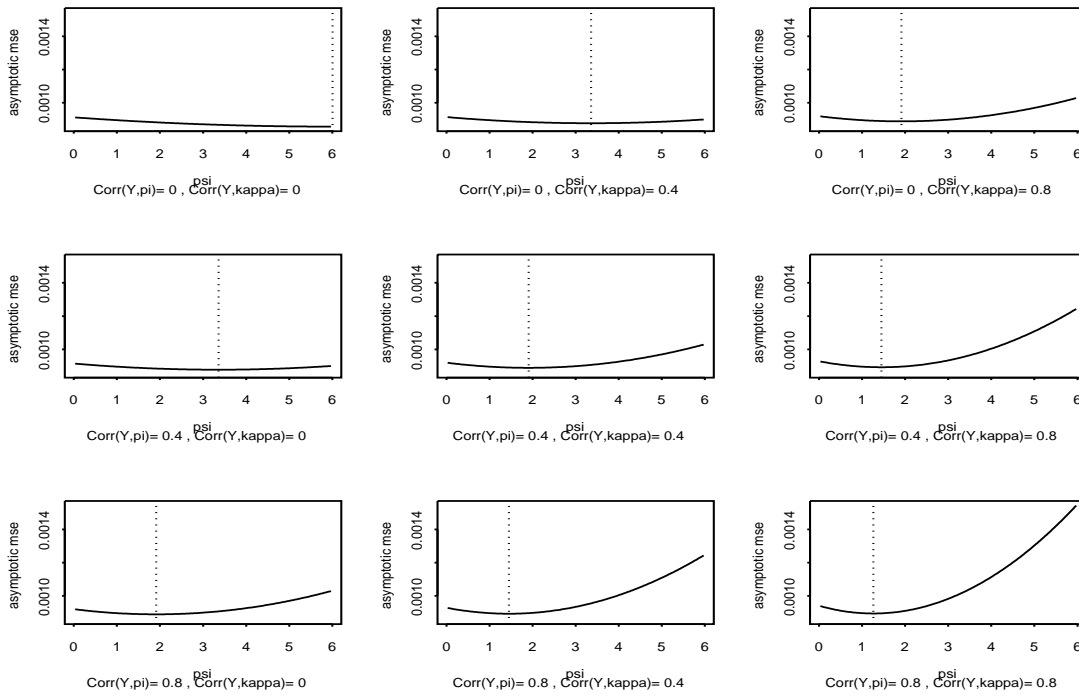
- Simple random sampling
 - if $\pi_{ij} \equiv 1$, $\kappa_{ij} \equiv 0$, and $\varphi_h \equiv 1$, then phase two is a census
- Stratified simple random sampling
 - if $\lambda = 1$ and $\pi_{ij} \equiv 1$, then phase one is a census
- Classical two-phase sampling for stratification (e.g., Cochran 1977)
 - if $\pi_{ij} \equiv 1$ and $\kappa_{ij} \equiv 0$, then full response, no conversions
 - if $\pi_{ij} = 0$ and $\kappa_{ij} = 1$, then no respondents, all conversions

Combining the Two Approaches

- Efficiently stratify long form respondents using X_j, I_{hij}
- Inefficiently stratify all others using $\hat{\gamma}_i, J_{hi}$
- Appropriately combined estimator is unbiased
- Could unequally weight the two parts
 - trade some bias for small variance on the long form subsample
 - can derive optimal weighting, allocation

Numerical Example

- $D = 200, \mu_M = 50, \rho_{xy} = 0.6$



Summary

- Redesign problem for Consumer Expenditure Survey
 - loss of key stratification variables from short form
 - efficiency loss depends on alternative
- Long form stratification
 - loss of efficiency due to domain-level stratification
 - quantified using small domain model
- Two-phase sampling for stratification
 - standard procedure in this context
 - potential biases due to selection problems
 - quantified using Bernoulli mixture
- Combination
 - can be weighted to trade off bias and variance
 - optimal allocation and weighting
 - unbiased weighting never far from optimal