Sample Survey Methods: Recent Developments and Applications

J.N.K. Rao, Carleton University

Sharon L. Lohr, Arizona State University

Course Outline

- Background
  - Survey design issues
  - Data collection
- Inferential issues
  - Inference on a total: foundations
    - Hansen example
    - calibration estimation
    - confidence intervals
  - conditional inference
  - empirical likelihood

Course Outline

- Analysis of survey data
  - Graphing survey data
  - Accounting for survey design features
  - Variance estimation
  - Software
  - Estimating equation approach
  - Inverse sampling
  - Quantiles

Course Outline

- Dual frame
- Imputation
- Small area estimation
- Wrap-up and evaluations

Course Objectives

- Learn about recent theoretical developments for analyzing complex survey data
- Implement theory using examples from U.S. and Canadian surveys
- Understand principles, estimation methods for dual frame surveys
- Know advantages, drawbacks of methods for computer-intensive variance estimation. Be able to select the best methods for applications.
- Use models to obtain more precise estimates in small areas.

Steps in Taking a Survey

- Survey Design
- Data Collection and Processing
- Estimation and Analysis

Early Milestones

- Neyman (1934): Stratified random sampling; optimal allocation; logic of inference based on CI
- Sukhatme (1935): Pilot samples to implement Neyman allocation
- Hansen-Hurwitz (1943): PPS sampling, stratified multistage sampling
- Mahalanobis (1944): Large-scale sample surveys

Early Textbooks

- Hansen, Hurwitz, and Madow (1953)
- Cochran (1953)
- Sukhatme (1954): agricultural surveys
- Kish (1965): design effects

Recent Textbooks and Monographs

- Lohr (1999). *Sampling: Design and Analysis*. Duxbury Press.
- Korn and Graubard (1999). *Analysis of Health Surveys*. Wiley.
- S.K. Thompson (2002). *Sampling, 2nd ed*. Wiley.
- Särndal, Swensson, Wretman (1992). *Model-Assisted Survey Sampling*. Springer-Verlag.
- M.E. Thompson (1997). *Theory of Sample Surveys*. Chapman & Hall.

Analysis of Survey Data

- Skinner, Holt and Smith (1989). *Analysis of Complex Surveys*. Wiley.
- Chambers and Skinner (2003). *Analysis of Survey Data*. Wiley.
- Lehtonen and Pahkinen (2004). *Practical Methods for Design and Analysis of Surveys, 2nd ed.* Wiley.

Nonsampling Errors

- Biemer & Lyberg (2003). *Introduction to Survey Quality.* Wiley.
- Lessler & Kalsbeek (1992). *Nonsampling Errors in Surveys.* Wiley.
- Groves (1989). *Survey Errors and Survey Costs.* Wiley.
- Groves, Dillman, Eltinge, Little (eds.) (2001). *Survey Nonresponse.* Wiley.

Small Area Estimation

- Schaible (1996). *Indirect Estimators in U.S. Federal Programs.* Springer-Verlag.
- Rao (2003). *Small Area Estimation.* Wiley.

Data Collection

- Sirken et al. (1998). *Cognition and Survey Research.* Wiley.
- Tourangeau, Rips, Rasinski. (2001). *The Psychology of Survey Response.*
- Dillman (1999). *Mail and Internet Surveys: The Tailored Design Method,* 2nd ed.

Survey Design

- Total MSE = $(\text{Bias})^2$ + Response Variance
  + Sampling Variance
- Resource Allocation

  Minimize total error for given cost

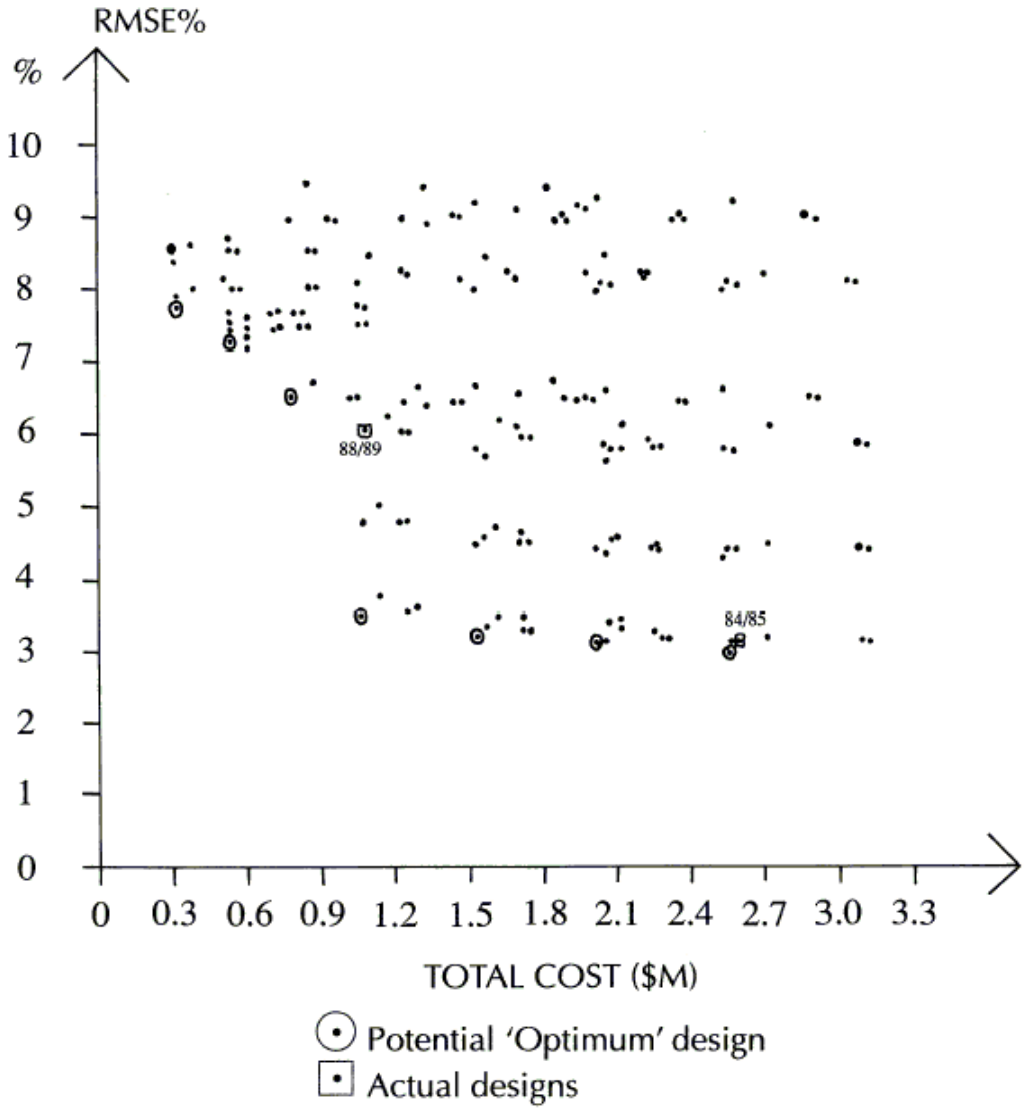


Fig. 1. RMSE% v, total cost for a number of resource allocation options

Linacre and Trewin (1993)

Linacre-Trewin (1993)

- Cost/RMSE graph
- 84/85: cost 2.7 ($M), RMSE = 3%
- 88/89: cost 1.0 ($M), RMSE = 6%
- Option 4: cost 1.0 ($M), RMSE = 3.5%

Fuller (1995)

Zero mean measurement error

$$y_t = \tilde{y}_t + \varepsilon_t, E[\varepsilon_t] = 0, E[\varepsilon_t^2] = \sigma^2$$

$$E(\tilde{y}_t) = \tilde{\mu},\ V(\tilde{y}_t) = \tilde{\sigma}^2$$

$$V(y_t) = \tilde{\sigma}^2 + \sigma^2 = \sigma_0^2$$

Mean $\bar{y}$: $E(\bar{y}) = \tilde{\mu}, E(s_y^2/n) = V(\bar{y})$

Quantiles (under normality)

$$F_y(a) \neq F_{\tilde{y}}(a) \text{ if } a \neq \tilde{\mu} = \text{ median}$$

Naive estimator of quantile $Q(p)$:

$$\hat{Q}(p) = \bar{y} + s_y \Phi^{-1}(p)$$

Replicated observations:

$$y_{it} = \tilde{y}_t + \varepsilon_{it};\ i = 1, \ldots, m\,(m \geq 2); t = 1, \ldots, n_i$$

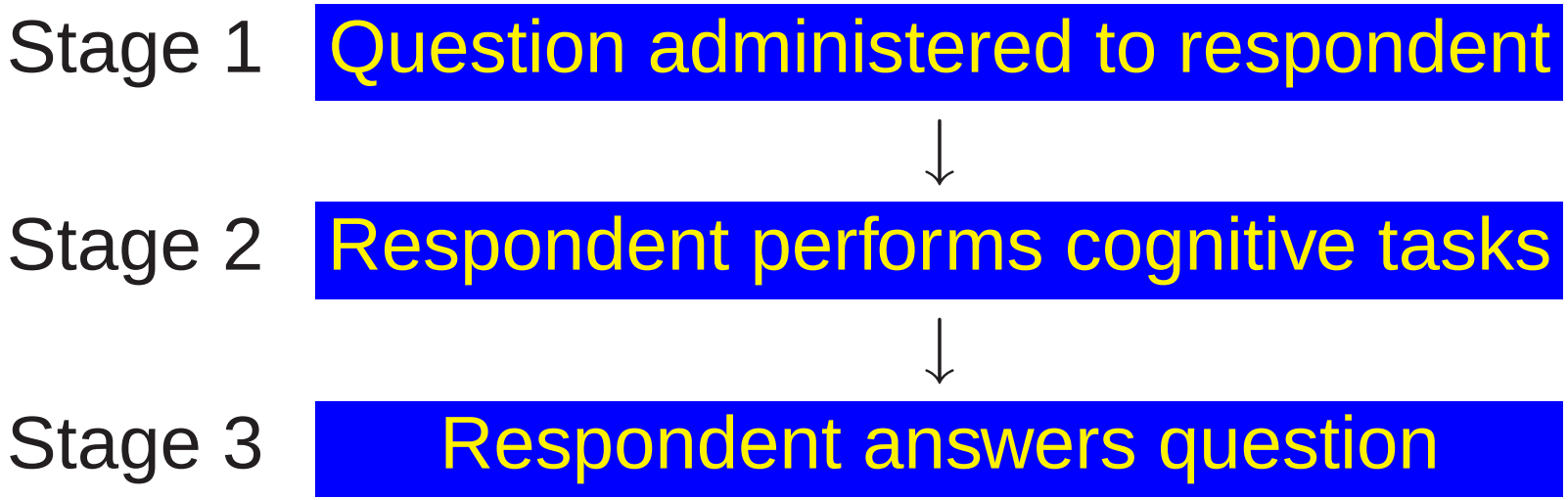
$$\tilde{Q}(p) = \bar{\bar{y}} + (\text{est. } \tilde{\sigma})\Phi^{-1}(p)$$

Moral: Allocate resources to measure measurement error variance σ^2

Data Collection

a) CASM: improve data quality
   Stimulus $\rightarrow$ cognition $\rightarrow$ response
b) Split Questionnaire
   - Decrease nonresponse rate
   - Increase response quality
c) Ordering of questions

3-stage survey response process



Sample Survey Methods: Recent Developments and Applications

Inferential Issues

Inferential Issues

- Total, Y
- Mean, $\bar{Y}$
- Median
- Ratio $R = Y/X$
- Domain Mean
- Regression coefficient
- Income inequality: Lorenz curve, income share, Gini coefficient, low income proportion (Binder & Kovacevic, 1995)

Foundational Aspects of Inference

Population: $U = \{1, 2, \ldots, N\}$
Parameter: $Y = y_1 + \ldots + y_N$
Sampling Design: $\{s, p(s)\}$
Data: $\{(i, y_i), i \in s\}$
Estimator of $Y = \hat{Y}$; std. error of $\hat{Y} = s(\hat{Y})$
Coefficient of variation of $\hat{Y} = s(\hat{Y})/\hat{Y}$

Design-based approach (Neyman, 1934)

"If we are interested in a collective character of a population (say total Y) and use methods of sampling and estimation, allowing us to ascribe to every possible sample(s) a confidence interval $[\widehat{Y}_L(s), \widehat{Y}_U(s)]$ such that the frequency of errors in the statements $\widehat{Y}_L(s) \leq Y \leq \widehat{Y}_U(s)$ does not exceed (1 $-\alpha$) prescribed in advance, *whatever the unknown properties* of the population, I shall call the method of sampling *representative* and the method of estimation *consistent*."

Probability-sampling methods, when carefully applied and with reasonably large samples, provide protection against failures of assumed models ...

Models are appropriately used to guide and evaluate the design of probability samples, but with large samples the inferences should not depend on the model.

Hansen et al. (1983)

Other approaches

- Model-dependent approach:
  Brewer (1963), Royall (1970)
- Model-assisted approach:
  Särndal et al. (1992)
- Conditional design-based approach:
  Holt & Smith (1979), Rao (1985, 1994)
  Robinson (1987), Casady & Valliant (1993)

Probability Sampling (Procedural Inference)

$E_p(\hat{Y}) \approx Y$

$E_p[s^2(\hat{Y})] \approx \text{MSE}_p(\hat{Y})$

Godambe (1955): $\hat{Y} = \sum_{i \in s} d_i(s) y_i$

$d_i(s) =$ design (or sampling) weight

Special case: $d_i(s) = 1/\pi_i$ (H-T estimator)

Non-existence of BLUE, even for SRS

Neyman class: $l_1 y'_1 + \ldots + l_n y'_n \Rightarrow \bar{y}$ BLUE

- **Flat** likelihood (Godambe, 1966):
  Let $\tilde{\mathbf{y}} = (\tilde{y}_1, \ldots, \tilde{y}_N)^T$ denote unknown population vector

  Likelihood $L(\tilde{\mathbf{y}}) =$ Prob(data $\mid \tilde{\mathbf{y}}$)

$$= \begin{cases} p(s) & \text{if } y_i = \tilde{y}_i, i \in s \\ 0 & \text{otherwise} \end{cases}$$

  Data = $\{(i, y_i), i \in s\}$
- Inference: All values $\tilde{y}_i, i \in U - s$ have same likelihood

Resolution

(1) Bayesian route (Ericson, 1969)
Put informative prior on $\tilde{\mathbf{y}}$: exchangeable prior
Posterior distribution of $\tilde{\mathbf{y}}$ given data becomes informative, same for any design

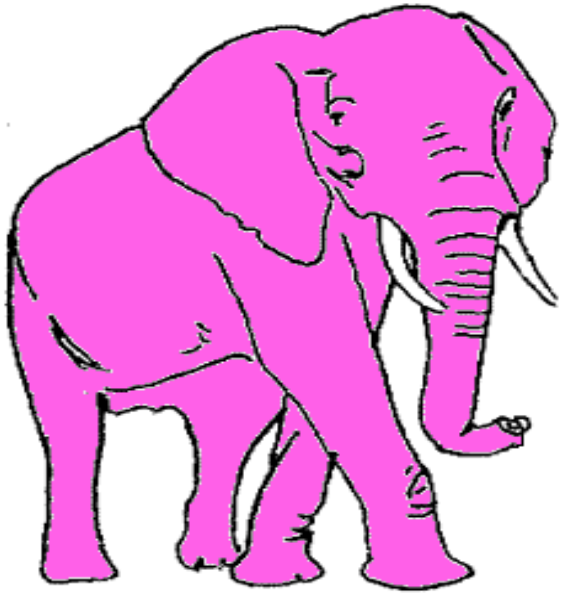
(2) Likelihood route (Hartley & Rao, 1968)
Ignore aspects of data: $\{y_i, i \in s\}$ for SRS
Likelihood informative and (2) similar to (1), but inference depends on design unlike (1).
This likelihood now called **empirical likelihood** (Owen, 1988)

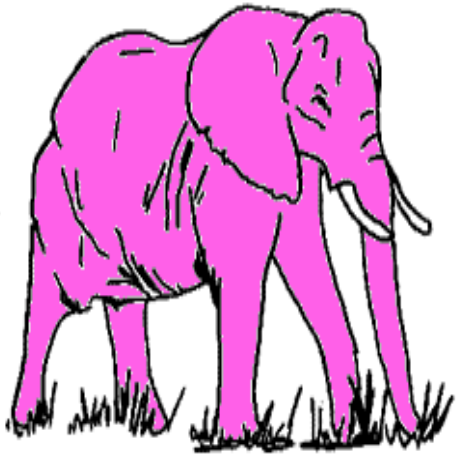
Basu's (1971) Circus Elephants

"The circus owner is planning to ship his 50 adult elephants and so he needs a rough estimate of the total weight of the elephants. As weighing an elephant is a cumbersome process, the owner wants to estimate the total weight by weighing just one elephant. Which elephant should he weigh?"

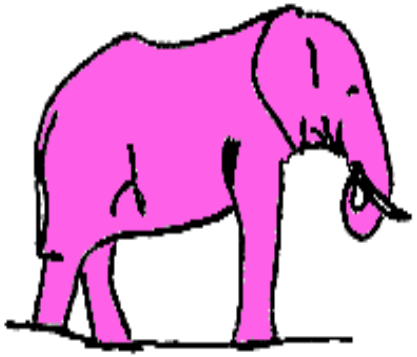
$N = 50, n = 1$

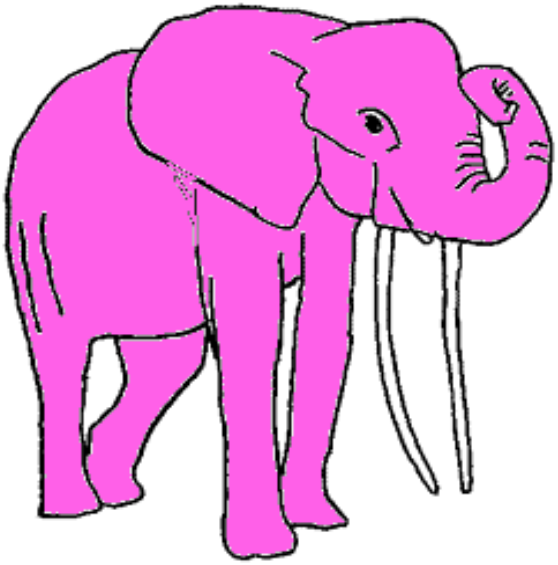
Jumbo

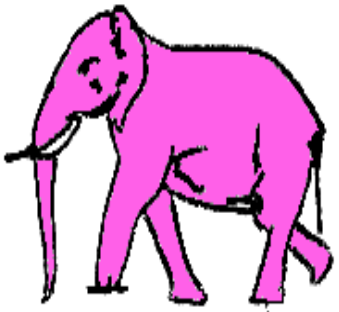




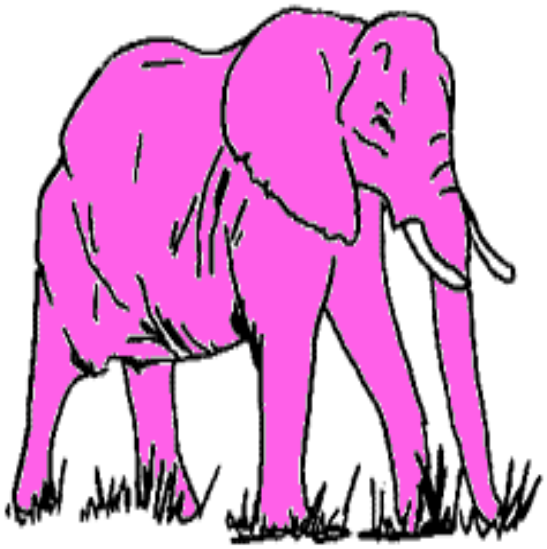
Sambo







Circus Owner's Plan



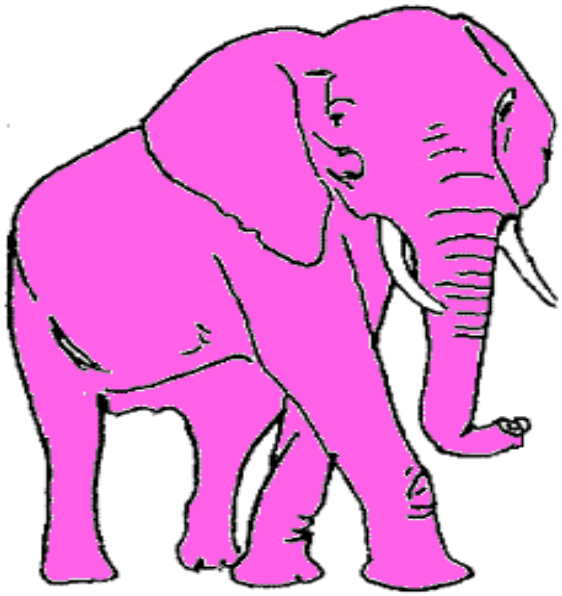
Sambo

- Census 3 years ago
- Sambo average weight in herd ($x_S \approx \bar{X}$)
- Circus owner's plan: weigh Sambo, multiply weight by 50
- $\hat{Y} = 50\, y_S$

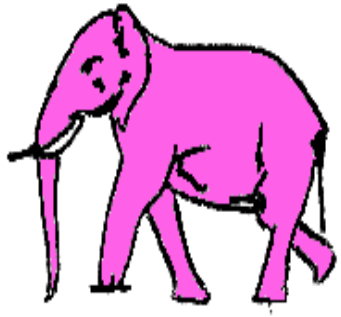
Circus Statistician

- Horrified: not a probability sample
- Unequal probabilities of selection, with
  $\pi_S = 99/100$ (Sambo)
  $\pi_i = 1/4900$ (other 49 elephants)

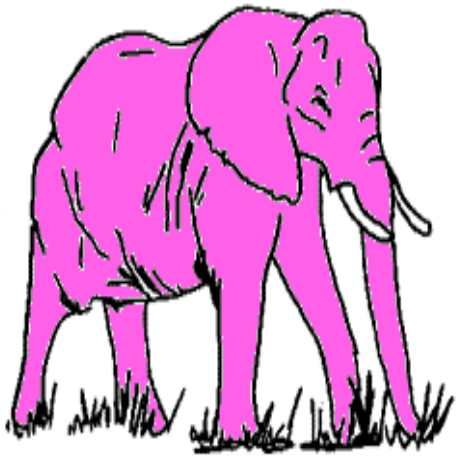
Jumbo, 1/4900



1/4900

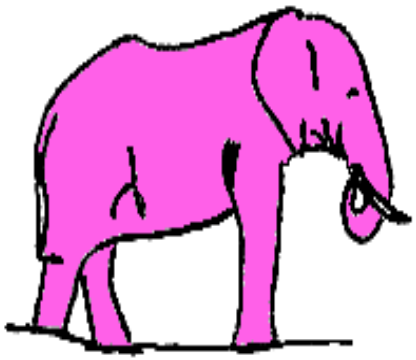


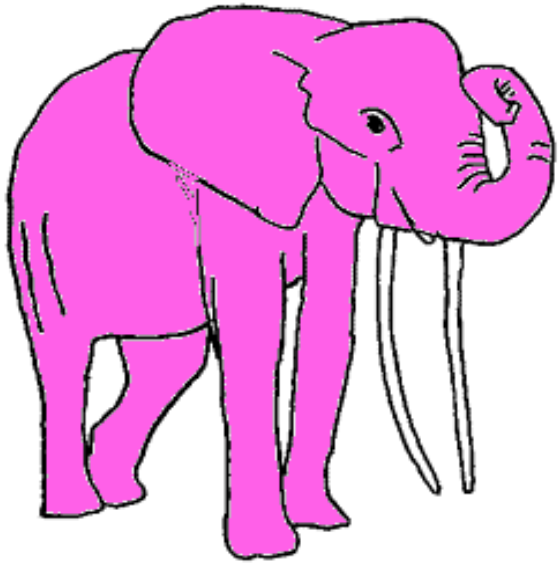
1/4900



Sambo, 99/100

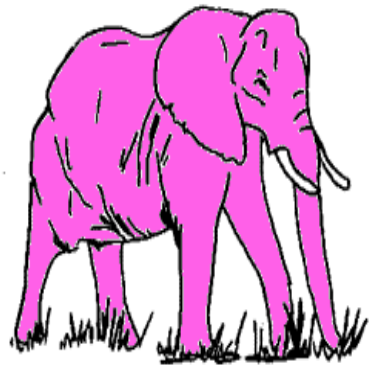
1/4900

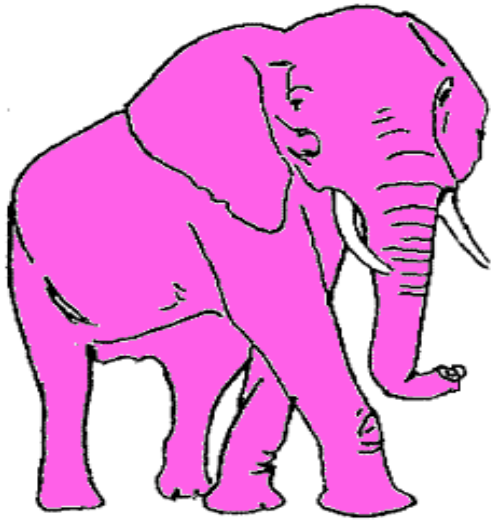




1/4900

Circus Statistician's Plan





- Sambo selected
  $\hat{Y}_{HT} =$ (100/99) $y_S \approx y_S$

- What if Jumbo selected?
  $\hat{Y}_{HT} = 4900\ y_J$

Elephants

- Lindley (*Amer. Statist.*, 1996):
  Basu's "counterexample" essentially
  "destroys frequentist sample survey theory"
- Resolution: Hájek estimator
  $\hat{Y}_H = \frac{y_S/\pi_S}{x_S/\pi_S} X = \frac{y_S}{x_S} X$ if Sambo selected
  Note: y_i, π_i uncorrelated in elephants example (Rao, 1966)

Prediction Approach (Model-dependent)

- Brewer (1963); Royall (1970)
- Superpopulation model on $\mathrm{y} = (y_1, \ldots, y_N)$
  $E_m =$ expectation w.r.t. model
- $E_m(\hat{Y}_s) = E_m(Y) \forall s$: model unbiased
  $E_m[s^2(\hat{Y}_s)] \approx E_m(\hat{Y}_s - Y)^2 = V_m(\hat{Y} - Y) \forall s$
  Inference based on $\hat{Y}_s$ & $s(\hat{Y}_s)$, conditionally on s
- Key assumption: Model holds for the sample s; no selection bias

- Ratio model (auxiliary var. x with known total X)

  $E_m(y_i) = \beta x_i, V_m(y_i) = \sigma_i^2 = \sigma^2 x_i,$

  $\text{Cov}_m(y_i, y_j) = 0$

  SRS: $\hat{Y} = N\bar{y}$ p-unbiased, but

  Model bias of $\hat{Y}_s$: $E_m(\hat{Y}_s - Y) = N\beta(\bar{x}_s - \bar{X})$

  Model bias > 0 if $\bar{x}_s > \bar{X}$; < 0 if $\bar{x}_s < \bar{X}$
- BLUP (Best linear unbiased predictor)

  $\hat{Y}_s = \sum_{i \in s} y_i + \sum_{i \in U-s} \hat{y}_i = \frac{\bar{y}_s}{\bar{x}_s} X$ (ratio est.)

  $\hat{y}_i$ = model predictor of $y_i = \hat{\beta}_s x_i = (\bar{y}_s / \bar{x}_s) x_i$

Model-assisted Approach
(Särndal et al., 1992)

- Write $Y = \sum_{i \in U} \hat{y}_i + \sum_{i \in U} e_i; \;\; e_i = y_i - \hat{y}_i$

$$\hat{\beta} = \left[\sum_{i \in s} \frac{y_i x_i}{\pi_i (\sigma^2 x_i)} \right] \left[\sum_{i \in s} \frac{x_i^2}{\pi_i (\sigma^2 x_i)} \right]^{-1} = \frac{\hat{Y}_{HT}}{\hat{X}_{HT}},$$

$\hat{Y} = \sum_{i \in U} \hat{y}_i + \sum_{i \in s} e_i / \pi_i$: Difference est.

$= \hat{Y}_{HT} + \hat{\beta}(X - \hat{X}_{HT})$: GREG est.

Model-assisted approach

- $\hat{Y}$ design-consistent and unbiased under "working model"
- $\hat{Y}$ minimum "anticipated" variance asymptotically
- Note: $\hat{Y}$, under ratio model, reduces to ratio estimator $(\hat{Y}_{HT}/\hat{X}_{HT})X$

Hansen et al. (1983)

Model misspecification: $E_m(y_i) = 0.4 + 0.25x_i$

Design: Stratification on x with x–totals approximately same in each stratum, equal strata sample sizes, SRS within strata

Scatter plot: Indicated regression through origin

Confidence interval coverage (n =200; nominal 95%)

	Estimator	L	U	Coverage
Model-assisted:	$(\bar{y}_{st}/\bar{x}_{st})X$	3.1	2.5	94.4
Model-based:	$(\bar{y}/\bar{x})X$	30.0	0	70.0

Conditional Coverage (n =200, nominal 95%)

Group:	1	2	3	4	5	6	7	8	9	10
$(\bar{y}/\bar{x})X$:	74	73	73	75	74	75	76	74	73	74
$(\bar{y}_{st}/\bar{x}_{st})X$:	94	93	94	94	94	95	95	95	94	95
$\hat{Y}_{st} = N\bar{y}_{st}$:	92	94	93	95	95	96	96	94	94	94

Valliant et al. (2000, pp. 88–90): ($n = 200$)
73% of 1000 samples rejected H_0: Intercept $= 0$
if error variance is $\sigma^2 x_i$;
97% for $\sigma^2 x_i^2$; 32% for σ^2.

Confidence Intervals (CI): Model-assisted

Dorfman (1994): "One of the contentions of design-based theory is that design-based estimators that happen to incorporate a model are inferentially satisfactory, despite failure of the model ... The results on coverage of the regression estimator under a quadratic model ... *dramatically* call this contention into question".

Example: Two-phase SRS, linear regression estimator

$\bar{y}_{lr} = \bar{y} + b(\bar{x}' - \bar{x})$

$\bar{x}'$ = first-phase sample mean of x;

$(\bar{y}, \bar{x})$ = second-phase sample means

Normal theory CI on $\bar{Y}$:

$$\left[\bar{y}_{lr} - 2s(\bar{y}_{lr}),\ \bar{y}_{lr} + 2s(\bar{y}_{lr})\right]$$

Model	Skew (y)	Skew (linear residual)
$y_i = x_i + \varepsilon_i$	3.3	0.02
$y_i = 8x_i^2 + \varepsilon_i$	11.0	6.40

CI Coverage Rate (Nominal Level 95%)

$n_1 = 80,\ n_2 = 40;\ \sigma^2 = 0.16$

	Standard	Full
$y_i = x_i + \varepsilon_i$	87.7	92.4
$y_i = 8x_i^2 + \varepsilon_i$	60.2	61.1

CI coverage depends on skewness of residuals, not of y's (or x's). Hence, CI behaves well for $y_i = x_i + \varepsilon_i$.

Model-assisted: Quadratic regression estimator

$$\bar{y}_{qr} = \bar{y} + b_1(\bar{x}' - \bar{x}) + b_2(\bar{z}' - \bar{z});\ z_i = x_i^2$$

C I Coverage Rate (Nominal 95%)

σ^2	Linear fit	Quadratic fit
0.04	63.1	91.5
0.16	62.7	90.9

Rao, Jocelyn and Hidiroglou (2003)

General model-unbiased estimator (ratio model):

$$\hat{Y} = \sum_{i\in s} y_i + \left\{\sum_{j\in s} y_j z_j \Big/ \sum_{j\in s} x_j z_j\right\}(X - \sum_{i\in s} x_i)$$

Predictor $\hat{y}_i = \left\{\sum_{j\in s} y_j z_j \Big/ \sum_{j\in s} x_j z_j\right\} x_i; \quad i \in U - s$

BLUP: $z_j = 1$ (z called instrumental variable)

Asymptotic Design Unbiased: $z_j = \frac{1}{\pi_j} - 1$ (Brewer, 1979)

GREG form: $\hat{Y} = \sum_{i\in s} \frac{y_i}{\pi_i} + \frac{\sum_s y_j(\pi_j^{-1} - 1)}{\sum_s x_j(\pi_j^{-1} - 1)} \left(X - \sum_{i\in s} \frac{x_i}{\pi_i}\right)$

Note: This z_j makes GREG expressible in prediction (or cosmetic) format. $\hat{Y}$ of form $\sum_{i\in s} w_i y_i$ with $w_i > 1$ if $x_i > 0$.

Model-assisted approach

- GREG estimator may be written as

  $\hat{Y} = \sum_{i \in s} w_i y_i$ with $w_i = g_i(s) d_i$

  $g_i(s) = g -$ weight $= X / \hat{X}_{HT}$

  $d_i =$ design weight $= \pi_i^{-1}$
- $\mathbf{x}$ is p-vector with known total $\mathbf{X}$

  Working model: $y_i = \mathbf{x}_i^T \beta + e_i, \ e_i \sim (0, \sigma^2 v(\mathbf{x}_i))$

  GREG: $\hat{Y} = \sum_{i \in s} w_i y_i$

  $g_i(s) = 1 + (\mathbf{X} - \hat{\mathbf{X}})^T (\sum_s d_j \mathbf{x}_j \mathbf{x}_j^T / c_j)^{-1} \mathbf{x}_i / c_i$

  $c_j =$ “tuning” constant; e.g., $c_j = v(\mathbf{x}_j)$

When is BLUP of Y p-consistent? (Fuller, 2002)

Model: $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}, \quad E(\mathbf{e}) = \mathbf{0}, \quad V(\mathbf{e}) = \boldsymbol{\Phi}\sigma^2$

(i) Let one column of $\mathbf{X}$ be $\mathbf{1}$ and $\boldsymbol{\Phi} = \text{diag}\,(\pi) = \mathbf{D}_\pi$

(ii) Let one column of $\mathbf{X}$ be π_i^{-1}–elements and $\boldsymbol{\Phi} = \mathbf{I}$

(iii) Let one column of $\mathbf{X}$ be π_i–elements and $\boldsymbol{\Phi} = \mathbf{D}_\pi^2$

Example:

(a) $\mathbf{X} = \text{col}\,(\pi)$ and $\boldsymbol{\Phi} = \mathbf{D}_\pi^2$, BLUP $= \widehat{Y}_{HT}$

(b) $\mathbf{X} = \mathbf{1}$, $\boldsymbol{\Phi} = \mathbf{D}_\pi$,

BLUP $= N(\sum_s \pi_i^{-1} y_i)/(\sum_s \pi_i^{-1}) =$ Hájek estimator

Case of $(\mathbf{x}_1, \ldots, \mathbf{x}_N)$ known
(nonlinear models):

GLIM Model:

$$E_m(y_i) = \mu(\mathbf{x}_i, \boldsymbol{\beta}) = \mu_i = g(\mathbf{x}_i^T \boldsymbol{\beta})$$

$$\begin{aligned} Y &= \Sigma_{i \in U}\, \mu_i + \Sigma_{i \in U}(y_i - \mu_i) \\ \hat{Y} &= \Sigma_{i \in U}\, \hat{\mu}_i + \Sigma_{i \in s}(y_i - \hat{\mu}_i)/\pi_i \end{aligned}$$

(Model-assisted)

$$= \Sigma_{i \in s}\, y_i/\pi_i + (\Sigma_{i \in U}\, \hat{\mu}_i - \Sigma_{i \in s}\, \hat{\mu}_i/\pi_i)$$

(Difference estimator)

Case of $(\mathbf{x}_1, \ldots, \mathbf{x}_N)$ known (nonlinear models):

$\widehat{Y}$ is BLUP form $\sum_{i \in s} y_i + \sum_{i \in U-s} \widehat{\mu}_i$ if

$\sum_{i \in s} (\pi_i^{-1} - 1)(y_i - \widehat{\mu}_i) = 0$. (Firth & Bennett, 1998)

Wu and Sitter (2001) regress y_i on $\widehat{\mu}_i$.
Local polynomial regression: Breidt and Opsomer (2000)

Calibration estimators (Deville & Särndal, 1992)

Calibrate the design weights d_i to w_i such that

$\sum_{i \in s} w_i \mathbf{x}_i = \mathbf{X}$: user-specified benchmark constraints

Choose w_i to minimise a chi-squared distance

$\sum_{i \in s} c_i (d_i - w_i)^2 / d_i$ subject to above constraints

Solution: $w_i =$ GREG weight (Note: no working model here)

Projection form: If $c_j = \boldsymbol{\nu}' \mathbf{x}_j$ for some $\boldsymbol{\nu}$, GREG simplifies:

$$g_i(s) = \mathbf{X}^T \left(\sum_{j \in s} d_j \mathbf{x}_j \mathbf{x}_j^T / c_j \right)^{-1} \mathbf{x}_i / c_i.$$

Example: $\mathbf{x}_j$ post-strata indicators, $\boldsymbol{\nu} = \mathbf{1}$, $c_j = 1$

Difficulties with calibration:

- (RR) Range restrictions: $L \leq w_k/d_k \leq U;\ k \in s$
  GREG weights satisfy BC but not necessarily RR.
  If number of BC large, some w_k can be < 0 or 1
- Alternative methods:
  (1) Change distance function: Information distance
  $\sum_{i \in s} c_i \{w_i \log(w_i/d_i) - w_i + d_i\}$
  Resulting w_i nonnegative but some w_i can be large.
  (2) Drop constraints (Bankier et al. 1992)

(3) Scaled modified χ^2 dist. (Huang & Fuller, 1978).
Iterative solution satisfying BC at each iteration.
Solution that satisfies BC & RR may not exist; no control on extent of discrepancy in meeting RR.

(4) Shrinkage minimization (Singh & Mohl, 1996).
Same difficulty as (2)

(5) Quadratic programming (Hussain, 1969).
Minimize distance subject to BC & RR
Note: Feasible set of solutions satisfying both BC & RR may be empty.

(6) Ridge weights (Chambers, 1996)

Minimize penalized distance:

$$\sum_{i \in s} \frac{c_i (w_i - d_i)^2}{d_i} + \frac{1}{\lambda} \sum_{j=1}^{p} q_j \left(\sum_i w_i x_{ij} - X_j \right)^2$$

No guarantee BC & RR satisfied.

(7) Ridge shrinkage (Rao & Singh, 1997; Chen, Sitter and Wu, 2002)

Tolerance specification:

$$\left| \sum_s w_i x_{ij} - X_j \right| / X_j \leq \delta_j, \quad j = 1, \ldots, p$$

Iterative method designed to force convergence for specified number of iterations by using a built-in tolerance specification procedure to relax BC while satisfying RR.

Calibration Software & Applications:

(1) GREG: Canadian Labor Force Survey
(2) Alt. method (2): 1991 Canadian Census of Pop. to ensure consistency between short form totals and estimated long form totals (Bankier et al., 1992)
(3) Alt. method (5): estimating census adjustment factors (Isaki et al., 2000)
(4) GREG & method (5): Generalized Estimation System (GES) at Statistics Canada.
(5) Other packages: LIN WEIGHT (Stat. Netherlands)
CALMAR (INSEE, France): different distance fcns
CLAN97 (Statistics Sweden): SAS-program

Stat. Canada Family Expenditure Survey (FAMEX)

BC: age ≤ 15, age > 15

one person household, two or more person household

$n = 797$ (1991 data for Regina); $p = 4$ BCs

$L = 0.5, \quad U = 2.0$

$\delta_{\min}$ for Ridge-shrinkage = 3.9%

Drop some BC method: 13.8% (for dropped BC)

Relative precision (RP) = SE(GREG)/SE(Ridge-shrinkage)

RP = 77% (furniture); RP = 87% (women's clothing)

SE calculated by jackknife

Conditional design-based approach

Hansen et al. (1983): “It is an attractive idea to make inferences that are conditional on the observed sample. This is sometimes legitimate in the framework of probability sampling”.

Post-stratification (SRS): Holt and Smith (1979)

Inference should be conditional on realized poststrata sample sizes $m_1, \ldots, m_p$ $(\sum m_i = n)$. In SRS case, reduces to stratified SRS with strata sample sizes $m_1, \ldots, m_p$.

For general designs, difficult to make inferences conditional on $m_1, \ldots, m_p$.

“Optimal” linear regression estimator

- In SRS case, can regard m_j as sample total of post-stratum indicator variable x_j
- Conditioning on m_j equivalent to conditioning on $\hat{X}_{jHT} = (N/n) \sum_s x_{ij} = (N/n) m_j$
- General designs with known totals $(X_1, \ldots, X_p)$ Condition on $(\hat{X}_{1HT}, \ldots, \hat{X}_{pHT}) = \hat{\mathbf{X}}_{HT}$
- Post-stratification: X_j = post-stratum size N_j (known)
- Robinson (1987) applied to ratio estimator, corrected cond. bias
- GREG est. for general designs cond. biased

Optimal linear regression estimator

$$\hat{Y} = \hat{Y}_{HT} + \hat{\mathbf{B}}_{opt}^{T}(\mathbf{x} - \hat{\mathbf{x}}_{HT})$$

$$\mathbf{B}_{opt} = [V(\hat{\mathbf{x}}_{HT})]^{-1}\text{Cov}\,(\hat{\mathbf{x}}_{HT}, \hat{Y}_{HT})$$

Properties of $\hat{Y}$:

(1) Asymptotically more efficient than GREG

(2) Calibration: $\hat{Y}(\mathbf{x}) = \mathbf{x}$

(3) Cond bias($\hat{Y}$)/Cond SE($\hat{Y}$) $\rightarrow$ 0 as $n \rightarrow \infty$

(4) Single set of weights: $\hat{Y} = \sum_s \tilde{w}_i y_i, \tilde{w}_i = d_i \tilde{g}_i$

(5) Stratified SRS: Include stratum indicators in **x** and define c_i suitably. Resulting GREG = optimal regression estimator (Särndal, 1996). Usual GREG fails to take account of stratification in estimation, unlike optimal regression estimator (Tillé, 1999). As a result, it does not benchmark to known strata sizes unlike optimal regression estimator.

(6) Optimal regression easy to implement for Poisson sampling, stratified SRS, stratified multistage sampling with psu's treated as sampled with replacement (Rao, 1994)

Simulation study

2 strata

($n_1 = n_2 = 100$; $W_1 = 0.2, W_2 = 0.8$)

Ratio model: $\beta_1 = 3, \beta_2 = 1$ (only X known)

Conditional bias ratio: (10 groups)

$\hat{Y}_{HT}$: ranged from -133% to 152%

GREG: 47% (group 1); -39% (group 10)

OPT: $<$ 10%

Conditional error rates (L and U): nominal 5%

$\widehat{Y}_{HT}$:	L ranged from 40% to 0%
	U ranged from 0% to 31%
GREG:	L ranged from 1.7% to 10%
	U ranged from 10.6% to 2.1%
OPT:	L & U close to 5%; no visible trends

Sample Survey Methods: Recent Developments and Applications

Empirical Likelihood

Empirical Likelihood (EL)

Hartley & Rao (1968) “scale-load” approach

y takes finite set of values $y_1^*, \ldots, y_T^*$

$$Y = \sum_{t=1}^{T} N_t y_t^*, \quad N_t = \text{ scale load for } y_t^*,$$

$$\sum_{t=1}^{T} N_t = N$$

$$\bar{Y} = \sum_{t=1}^{T} p_t y_t^*, \quad p_t = N_t/N$$

SRS: observed scale loads $(n_1, n_2, \ldots, n_T)$: data
Likelihood $L(\mathbf{N})$ multiple hypergeometric:

$$\prod_t \binom{N_t}{n_t} \Big/ \binom{N}{n}$$

MLE of N_t : $\hat{N}_t = N\dfrac{n_t}{n} \Rightarrow \hat{Y} = \sum \hat{N}_t y_t^* = N\bar{y}$

MLE of $F(a) = \sum p_t I(y_t^* \leq a)$:

$$\begin{aligned}\hat{F}(a) &= \sum \hat{p}_t I(y_t^* \leq a) \\ &= (1/n) \sum_{i \in s} I(y_i \leq a) = F_n(a)\end{aligned}$$

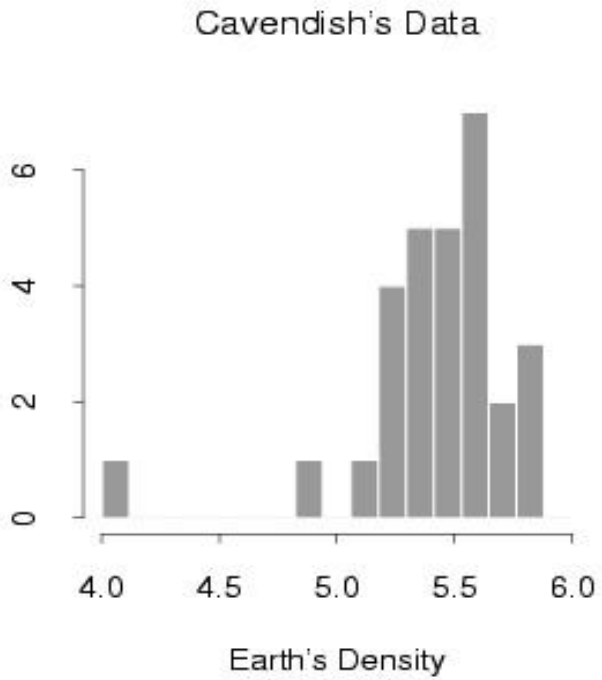
Owen (1988, 2001)

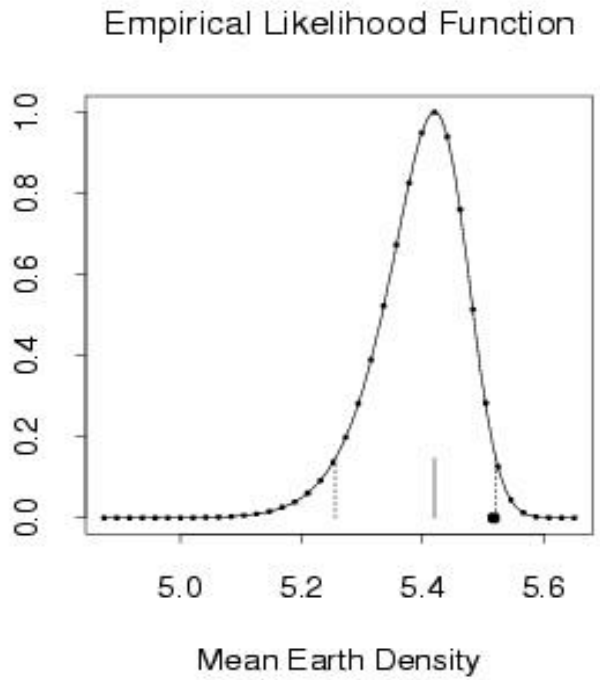
- $y_1, \ldots, y_n$ i.i.d. observations from $F(\cdot)$:
  $L(F) = \prod_1^n p_i$; $l(F) = \log L(F) = \sum_1^n \log p_i$
  $p_i = F(y_i) - F(y_i-) = P[Y = y_i]$
- Maximize $L(F)$ subject to $p_i \geq 0, \sum p_i = 1$:
  $\widehat{p}_i = 1/n$; $i = 1, \ldots, n$
  $\widehat{F}(a) = (1/n) \sum I(y_i \leq y) = F_n(a)$
- Estimation of $\mu = E(Y) = \int y dF(y) = T(F)$
  $\widehat{\mu} = \sum \widehat{p}_i y_i = \bar{y} =$ sample mean

Example (Art Owen)

Cavendish, mean density of Earth relative to water

http://www-stat.stanford.edu/ owen//empirical/





EL confidence intervals on μ:

Advantages:

(1) shape & orientation of CI determined entirely by the data.

(2) CI are range preserving and transformation respecting.

EL confidence intervals on μ:

Profile empirical likelihood ratio function:

$$R(\mu) = \max\left\{\prod_1^n (np_i) | \sum p_i y_i = \mu,\ p_i \geq 0,\ \sum p_i = 1\right\}$$

Note: $\prod(np_i) = L(F)/L(F_n)$

$(1-\alpha)$ level EL confidence interval on μ:

$$\left\{\mu\ | r(\mu)\ = -2\log R(\mu) \leq \chi_1^2(\alpha)\right\}$$

Note: Unlike CI based on normal approximation and standard error of $\widehat{\mu}$, EL intervals do not require standard error evaluation.

EL confidence intervals are particularly useful if balanced tail error rates are needed.

Example. Population containing many zero values (Chen, Chen & Rao, 2003)

Accounting practice: Amount of money owed to government

$\mu =$ average amount of excessive claims

Labor Force Survey-2000 (Ontario): number of extra hours worked (N=17,415; 3677 non-zeros, 0.21), n =100; 10,000 simulations

LNR= lower non-coverage rate (nominal 2.5%),
LB= lower bound

Normal Approximation		Empirical Likelihood	
LNR	LB	LNR	LB
0.70	9.82	2.11	11.62

Note: Lower bound more accurate for EL method: Noncoverage rate below lower bound closer to target 5%

EL using auxiliary information: $\bar{X}$ known

Hartley & Rao (1968), Chen & Qin (1993): SRS

EL estimator of $\bar{Y} \approx$ regression estimator

$$\bar{y} + b^*(\bar{X} - \bar{x})$$

$$b^* = \sum_{i \in s} (y_i - \bar{y})(x_i - \bar{X}) / \sum_{i \in s} (x_i - \bar{X})^2$$

Stratified SRS

- Zhong & Rao (1996, 2000)
- Data:
  $\{y_{hi}, x_{hi};\ h = 1, \ldots, L;\ i = 1, \ldots, n_h\}$
- Maximise $\sum_h \sum_i \log p_{hi}$ subject to $\sum_i p_{hi} = 1$
  for each h, $p_{hi} \geq 0$, $\Sigma_h \Sigma_i p_{hi} x_{hi} = \bar{X}$
- EL estimator of $\bar{Y} = \Sigma_h \Sigma_i \hat{p}_{hi} y_{hi} \approx$ opt. regression estimator
- EL est. of $F(a) = \sum_h \sum_i \hat{p}_{hi} I(y_{hi} \leq a)$
  monotonic function

General designs: pseudo-EL (Chen & Sitter, 1999)

Use pseudo empirical loglikelihood
$\ell_n(\mathbf{p}) = \sum_{i \in s} d_i \log p_i$; $d_i = \pi_i^{-1}$
Estimate "census" loglikelihood as
$\ell_N(p_i) = \Sigma_{i=1}^{N} \log p_i$
Maximise $\ell_n(\mathbf{p})$ subject to $p_i \geq 0$, $\Sigma_{i \in s} p_i = 1$, $\Sigma_i p_i x_i = \bar{X}$
Pseudo-EL estimator of $\bar{Y} = \Sigma_i \hat{p}_i y_i \approx$ GREG estimator

CI on $\bar{Y}$: unstratified case, no x–information

Let $n^* =$ effective sample size

$= n$/(design effect)

Design effect $= V(\hat{\bar{Y}}_{HT})/V(\bar{y}$: SRS)

$$\hat{\bar{Y}}_{HT} = \sum_i \tilde{d}_i y_i,\ \tilde{d}_i = d_i / \sum_s d_j,\ d_i = \pi_i^{-1}$$

Pseudo empirical loglikelihood

$$= \ell_n(\mathbf{p}) = n^* \sum_{i \in s} \tilde{d}_i \log p_i$$

CI on $\bar{Y}$: unstratified, no x–information

(a) Maximize $\ell_n(p)$ subject to $p_i \geq 0$, $\sum_s p_i = 1$,
$\hat{p}_i = \tilde{d}_i$

(b) Maximize $\ell_n(p)$ subject to $p_i \geq 0$, $\sum_s p_i = 1$,
$\sum_s p_i y_i = \bar{Y} \Longrightarrow \tilde{p}_i(\bar{Y})$

(c) $(1-\alpha)$–level CI on $\bar{Y}$:
$\{\bar{Y} | r(\bar{Y}) = -2\{\ell_n(\tilde{p}) - \ell_n(\hat{p})\} \leq \chi^2_1(\alpha)\}$

95% CI's for the Distribution Function
Rao-Sampford PPS Without Replacement
($\rho = 0.80$)

p	CI	Cov. Prob.	Lower Error Rate	Upper Error Rate	Avg. Length
0.10	NT	88.9	0.2	10.9	0.156
	EL	94.8	1.8	3.4	0.156
0.90	NT	93.1	5.6	1.3	0.130
	EL	94.9	2.7	2.4	0.129

NT=normal theory, EL= empirical likelihood

Integration of surveys (Wu, 2004a):

- Survey 1: $(y_{1i}, x_{1i}, z_i),\ i \in s_1$
- Survey 2: $(y_{2j}, x_{2j}, z_j),\ j \in s_2$
- $\bar{X}_1, \bar{X}_2$ known
- Pseudo empirical loglikelihood:

  Let $d_{ti} = \pi_{ti}^{-1}; t = 1, 2$

  $\ell(\mathbf{p}, \mathbf{q}) = \sum\limits_{i \in s_1} d_{1i} \log p_i + \sum\limits_{j \in s_2} d_{2j} \log q_j$

Integration of surveys (Wu, 2004a):

- Maximize $\ell(\mathbf{p}, \mathbf{q})$ subject to
  $$\sum_{i \in s_1} p_i = 1, \quad \sum_{j \in s_2} q_j = 1, \quad \sum_{i \in s_1} p_i x_{1i} = \bar{X}_1,$$
  $$\sum_{j \in s_2} q_j x_{2j} = \bar{X}_2, \sum_{i \in s_1} p_i z_i = \sum_{j \in s_2} q_j z_j :$$
  ensures consistency between surveys
- $\Rightarrow$ EL est. of $\bar{Y}_1 = \sum \hat{p}_i y_{1i}$;
  EL est. of $\bar{Y}_2 = \sum \hat{q}_j y_{2j}$
- Zieschang (1990), Renssen & Nieuwenbroek (1997): split questionnaire

Simulation: 1996 Statistics Canada Family Expenditure (FAMEX)

- Survey (Ontario) treated as finite population,
  $R = 1000$ simulations, $N = 2396$
- $x_1 =$ number of children age < 15
  $x_2 =$ number of youth (15-24)
  $x_3 =$ number of people in household
  $z =$ total income after taxes
  $y =$ total expenditure
- Survey 1: $x_1,\ x_3$ (known means)
  Survey 2: $x_2,\ x_3$ (known means)
  z from both surveys
  y from survey 1

GREG of $\bar{Y}$: using x_1, x_3 as auxiliary variables

% Relative efficiency of EL est. of $\bar{Y}$
$=$ MSE(GREG)/MSE(EL)

n_1	n_2	RE
240	80	91: Loss in efficiency when $n_2 << n_1$
	160	111
	240	124

Computation in R/SPlus: Wu (2004b)

Sample Survey Methods: Recent Developments and Applications

Variance Estimation

Variance Estimation

Basic Estimator: $\widehat{Y} = \sum_{i \in s} d_i y_i;\ d_i = \pi_i^{-1}$

Operator notation:

$\widehat{Y}(y) = \sum_{i \in s} d_i y_i; \widehat{Y}(z) = \sum_{i \in s} d_i z_i$

$\widehat{Y}(y_1) + \widehat{Y}(y_2) = \widehat{Y}(y_1 + y_2)$: additivity

Variance estimator of $\widehat{Y} = v(y)$

- Fixed sample size:

$$v(y) = \sum_{i<j \in s}\sum \{(\pi_i \pi_j - \pi_{ij})/\pi_{ij}\}(z_i - z_j)^2$$

$$z_i = y_i/\pi_i;\ \pi_{ij} = \Pr(i \in s \text{ and } j \in s)$$

- Inclusion Probabilities Proportional to Size (IPPS or πPS): $\pi_i = n\frac{x_i}{X} = np_i$
- Rao-Sampford method: SAS (sampling software)
- Approximations to π_{ij} in terms of π_i: Brewer & Donadis (2003)

Stratified multistage design (socio-economic surveys):

L strata ($h = 1, \ldots, L$); N_h PSU's in stratum h;
$Y_{hi} = (hi)$–th PSU total;
$d_{hik} =$ basic design weight for $(hik) \in s$

Basic estimator of Y : $\widehat{Y} = \sum_{hik \in s} d_{hik} y_{hik}$

Example: stratified two-stage sampling

πPS first stage; SRS second stage

n_h sample psu's from stratum h

m_{hi} sample elements from M_{hi} pop. elements

$M_{h+} = \sum_i M_{hi} =$ # elements in stratum h

$$d_{hi} = \frac{1}{n_h p_{hi}} \frac{M_{hi}}{m_{hi}}$$

Self-weighting: $p_{hi} = M_{hi}/M_{h+}$, $m_{hi} = \bar{m}$ (equal work loads)

$n_h = \bar{n}(M_{h+}/M_{++}) \Rightarrow d_{hi} = d = M_{++}/(\bar{n}\,\bar{m})$

$\widehat{Y} = M_{++}/(\bar{n}\,\bar{m})$ (sample total)

Variance estimation

- Proper variance estimation involves estimation of variance at each stage & joint inclusion probabilities (Cochran, 1977, p. 300)
- SUDAAN, SAS: Proper variance estimation for some designs
- For variance est. PSU's often regarded as sampled with replacement with probabilities p_{hi} in stratum h. Leads to great simplification regardless of number of sampling stages.

Variance estimation

$$\text{Var. est.}(\hat{Y}) = \sum_h \frac{1}{n_h(n_h-1)} \sum_i (y_{hi} - \bar{y}_h)^2 = v(y_{hi})$$

$y_{hi} = \sum_k (n_h d_{hik}) y_{hik} =$ weighted PSU total

$$\bar{y}_h = \sum_i y_{hi}/n_h$$

Note: Var. est. depends only on PSU totals y_{hi}

$v(y_{hi})$ leads to overestimation, but relative bias small if first-stage sampling fractions, n_h/N_h, are small.

Var. est. for GREG est. of Y:

$\hat{Y}_{GR} = \Sigma_{i \in s} w_i y_i,\ w_i = g_i(s) d_i$

Residual: $e_i = y_i - \hat{\mathbf{B}}^T \mathbf{x}_i$;

$$\hat{\mathbf{B}} = \left(\Sigma_{j \in s} d_j \mathbf{x}_j \mathbf{x}_j^T / c_j \right)^{-1} \left(\Sigma_{j \in s} d_j \mathbf{x}_j \mathbf{y}_j / c_j \right)$$

Taylor linearization:

(1) Var. est. of GREG = $v(e)$: operator notation

(2) Var. est. of GREG = $v(ge)$: operator notation

(2) is a good choice from either design-based or model-based perspective (Särndal, Swensson & Wretman, 1989); also good conditional properties (Valliant, 1993; Yung & Rao, 1996; Royall & Cumberland, 1981).

Stratified multistage design:

$(2) = v(\tilde{e}_{hi})$; $(1) = v(e_{hi})$

$$\tilde{e}_{hi} = n_h \sum_k d_{hik} g_{hik} e_{hik};\ e_{hi} = n_h \sum_k d_{hik} e_{hik}$$

Example: Ratio estimator $\hat{Y}_R = (\hat{Y}/\hat{X})X$

$$g_i(s) = \frac{X}{\hat{X}},\ (2) = \left(\frac{X}{\hat{X}}\right)^2 v(e) = \left(\frac{X}{\hat{X}}\right)^2 (1)$$

If $\hat{X} < X$, then (2) larger than (1) reflecting more uncertainty.
If $\hat{X} > X$, then (2) smaller than (1) reflecting less uncertainty.

Valliant (2002)

Working Model: $E_m(y_i) = \mathbf{x}_i^T\boldsymbol{\beta}$, $V_m(y_i) = v_i$
True Model: $E_m(y_i) = \mathbf{x}_i^T\boldsymbol{\beta}$, $V_m(y_i) = \psi_i$
Poisson sampling: $a_1, \ldots, a_N$ independent with
$P(a_i = 1) = \pi_i$

$$(2) = \sum_{i \in s} \frac{1-\pi_i}{\pi_i^2} g_i^2 e_i^2, \; e_i = y_i - \mathbf{x}_i^T \widehat{\boldsymbol{\beta}};$$

$$\widehat{\boldsymbol{\beta}} = \left[\sum_s \mathbf{x}_i \mathbf{x}_i' / (v_i \pi_i)\right]^{-1} \left[\sum_s \mathbf{x}_i y_i / (v_i \pi_i)\right]$$

$$(1) = \sum_{i \in s} \frac{1-\pi_i}{\pi_i^2} e_i^2$$

Some other variance estimators:

$$(3) = \sum_{i \in s} \left(\frac{g_i}{\pi_i} - 1 \right)^2 e_i^2$$

$$(4) = \sum_{i \in s} \left(\frac{g_i}{\pi_i} \right)^2 e_i^2$$

SUPERCARP (1980): Hidiroglou, Fuller and Hickman

$$(5) = \frac{n}{n-1} \sum_{i \in s} \left[\frac{g_i e_i}{\pi_i} - \frac{1}{n} \sum_{i \in s} \frac{g_i e_i}{\pi_i} \right]^2$$

(3), (4) and (5) similar: approximately model-unbiased under true model and approximately design-unbiased. (2) is also robust if π_i small, but (1) biased under true or even working model.

$$(6) = \sum_{i \in s} \left(\frac{g_i}{\pi_i} - 1 \right)^2 \frac{e_i^2}{1 - h_{ii}},$$

$$h_{ii} = \mathbf{x}_i^T [\textstyle\sum_s \mathbf{x}_i \mathbf{x}_i^T / (v_i \pi_i)]^{-1} \mathbf{x}_i / (v_i \pi_i)$$

$$(7) = \sum_{i \in s} \frac{1 - \pi_i}{\pi_i^2} \frac{g_i^2 e_i^2}{1 - h_{ii}}$$

(6) and (7) also robust.

Unified Taylor linearization variance estimation: Demnati & Rao (2004)

Estimator: $\widehat{\theta} = f(\mathbf{d}(s), \mathbf{A}_y)$

$d_k(s) = 0$ if k is not in s
$\mathbf{A}_y = m \times N$ matrix $(\mathbf{y}_1, \ldots, \mathbf{y}_N)$
$\mathbf{y}_k = (y_{k1}, \ldots, y_{km})^T$

Example: Ratio estimator $\widehat{Y}_R = \frac{\sum d_k(s) y_k}{\sum d_k(s) x_k} X$;
$m = 2$, $y_{1k} = y_k$, $y_{2k} = x_k$

Var. est. of $\widehat{\theta} = v(z)$

$z_k = \frac{\partial f(\mathbf{b})}{\partial b_k}$ eval. at $\mathbf{b} = \mathbf{d}(s)$, $k \in s$

Leads to unique choice, only derivatives involved.

Ratio case:

$$f(\mathbf{b}) = \frac{\sum b_k y_k}{\sum b_k x_k} X; \frac{\partial f(\mathbf{b})}{\partial b_k} = X \left\{ \frac{y_k \sum b_k x_k - x_k \sum b_k y_k}{(\sum b_k x_k)^2} \right\}$$

Hence $z_k = \frac{X}{\widehat{X}}(y_k - \hat{R} x_k) = \frac{X}{\widehat{X}} e_k,$

$\Rightarrow v(z) = v(ge) = (2)$

Resampling Methods for Variance Estimation

(1) Replicated samples

(i) Draw sample $s(1)$ according to a specified design

(ii) Replace $s(1)$ and select $s(2)$ independently by the same design

(iii) Repeat until k samples (random groups) $s(1), \ldots, s(k)$ obtained

$\widehat{\theta}_t =$ estimator based on $s(t)$,

$\widehat{\theta}_. = \frac{1}{k} \sum_{t=1}^{k} \widehat{\theta}_t =$ overall estimator

$\widehat{\theta} =$ overall est. based on combined sample

est. var. $(\widehat{\theta}_.) = \frac{1}{k(k-1)} \sum_{t=1}^{k} (\widehat{\theta}_t - \widehat{\theta}_.)^2 = s_R^2(\widehat{\theta}_.)$

Mahalanobis (1939): Jute acreage in Bengal, India. Mahalanobis called "interpenetrating samples": $s_R^2(\widehat{\theta}_.)$ estimates total variance in presence of measurement errors.

See Lohr (1999, p. 294)

Deming (1956): Replicated samples
Hansen, Hurwitz, Madow (1953): random group method

(a) $s_R^2(\hat{\theta}_\cdot)$ used to estimate variance of $\hat{\theta}$

(b) Use $s_R^2(\hat{\theta}) = \frac{1}{k(k-1)} \Sigma(\hat{\theta}_t - \hat{\theta})^2$ to estimate variance of $\hat{\theta}$

$s_R^2(\hat{\theta}) \geq s_R^2(\hat{\theta}_\cdot)$: $s_R^2(\hat{\theta})$ conservative

(c) SRS: $\widehat{\theta} = \bar{y} = \widehat{\theta}.$

$\mathrm{CV}[s_R^2(\bar{y})] = \left[\frac{\beta_4}{m} + 3\frac{m-1}{m} - \frac{k-3}{k-1}\right]^{1/2} \cdot \frac{1}{\sqrt{k}}$

$\beta_4 =$ Population kurtosis

Number of groups, k, has more impact on CV than group size m : $s_R^2(\widehat{\theta})$ unstable

(d) If $\widehat{\theta}_t$ nonlinear, $\widehat{\theta}.$ can be heavily biased if m small (even for large k or large n)

(2) Jackknife Method (IID case)

- $y_1, \ldots, y_n$ random sample from some $F(\cdot)$
- $\widehat{\theta} =$ estimator of θ
- $\widehat{\theta}_{(j)} =$ estimator of θ when y_j deleted
- Bias reduction: Quenouille (1956)
  Pseudo-values: $\tilde{\theta}_{(j)} = n\widehat{\theta} - (n-1)\widehat{\theta}_{(j)}$
  $\widehat{\theta}_Q = \tilde{\theta}_{(\cdot)} = \frac{1}{n} \sum_{j=1}^{n} \tilde{\theta}_{(j)}$:
  bias of lower order than $\widehat{\theta}$

Example: Ratio (IID case)

$$\widehat{\theta} = \frac{\bar{y}}{\bar{x}}$$

$$\widehat{\theta}_{(j)} = \frac{\bar{y}_{(j)}}{\bar{x}_{(j)}} = \frac{n\bar{y} - y_j}{n\bar{x} - x_j}$$

Bias $(\widehat{\theta}_Q) = O(n^{-2})$, Bias $(\widehat{\theta}) = O(n^{-1})$

Tukey (1958): $\widehat{\theta} = \bar{y},\ \tilde{\theta}_{(j)} = y_j$

Suggests may treat $\tilde{\theta}_{(1)}, \ldots, \tilde{\theta}_{(n)}$ as approx IID

$$
\begin{aligned}
v_J(\widehat{\theta}) &= \text{Jackknife var. est.} \\
&= \frac{1}{n(n-1)} \sum_{j=1}^{n} (\tilde{\theta}_{(j)} - \tilde{\theta}_{(\cdot)})^2 \\
&= \frac{n-1}{n} \sum_{j=1}^{n} (\widehat{\theta}_{(j)} - \widehat{\theta}_{(\cdot)})^2 \\
&= s_J^2(\widehat{\theta})
\end{aligned}
$$

Linear case: $v_J(\bar{y}) = \frac{1}{n} s_y^2 =$ usual var. est.

Example of jackknife (SRS)

- Poll, April 7-9, 2000
- “Do you consider the census an invasion of your privacy, or not?”
- 1006 respondents; 181 “yes”
- Pretend it’s SRS
- Usual 95% CI: $[0.156, 0.204]$

Jackknife weights (SRS)

Set $w_i = 0$; multiply other weights by $n/(n-1)$

y	w	$w_{(1)}$	$w_{(2)}$	$w_{(3)}$	$\ldots$	$w_{(1006)}$
0	1	0	1.000995	1.000995	...	1.000995
0	1	1.000995	0	1.000995	...	1.000995
0	1	1.000995	1.000995	0	...	1.000995
1	1	1.000995	1.000995	1.000995	...	1.000995
1	1	1.000995	1.000995	1.000995	...	1.000995
0	1	1.000995	1.000995	1.000995	...	1.000995
Estimated totals $\sum w_i y_i$:						
	181	181.1801	181.1801	181.1801	...	180.1791

Jackknife for poll data

- $\widehat{\theta}_{(i)} = \begin{cases} 0.1800995 \text{ if } y_i = 0 \\ 0.1791045 \text{ if } y_i = 1 \end{cases}$
- $v_J(\widehat{\theta}) = \frac{1005}{1006} \sum (\widehat{\theta}_{(i)} - \widehat{\theta})^2$
- Same as s^2/n here since mean

Smooth functions: $\widehat{\theta} = g(\bar{y})$

$$\frac{v_J(\widehat{\theta})}{\text{Var}\,(\widehat{\theta})} \longrightarrow_p 1 \text{ as } n \to \infty : v_J(\widehat{\theta}) \text{ consistent}$$

Note: $\widehat{\theta}$ may be used for $\widehat{\theta}_{(\cdot)}$ in $s^2_J(\widehat{\theta})$

Miller (1964); Krewski (1978): U-statistics (SRS)

Shao & Tu (1995 §2.2): Functionals $\widehat{\theta} = T(F_n)$
$T(\cdot)$ "continuously Gâteaux" differentiable: M estimators, smooth L–estimators

95% confidence interval on θ:

$$\{\hat{\theta} - 1.96\, s_J(\hat{\theta}),\ \hat{\theta} + 1.96\, s_J(\hat{\theta})\}$$

Asymptotically ($n \to \infty$) correct coverage

Nonsmooth function: $\hat{\theta}$ = sample median

$v_J(\hat{\theta})$ inconsistent:

$$\frac{v_J(\hat{\theta})}{\text{Var}\,(\hat{\theta})} \longrightarrow_p (\chi^2_2/2)^2 \text{random variable}$$

Remedy, nonsmooth case: Delete–d jackknife

Shao and Tu (1995)

$s_r =$ sample of size $r = n - d$ from s

$\widehat{\theta}(s_r) =$ estimator from s_r

$$v_{J-d}(\widehat{\theta}) = \frac{r}{d} \begin{pmatrix} n \\ d \end{pmatrix}^{-1} \sum_{s_r} [\widehat{\theta}(s_r) - \widehat{\theta}(\cdot)]^2$$

$v_{J-d}(\widehat{\theta})$ consistent if $\frac{d}{n} = \lambda$; $d \to \infty$ as $n \to \infty$

In practice, $\begin{pmatrix} n \\ d \end{pmatrix}$ is very large, so average over

$m \approx n^{3/2}$ independent subsamples s_r.

Example: SRS, $n = 300$ from Gamma(2,2)

- median = $\widehat{\theta} = 3.693$
- Use delete–150 jackknife; $\binom{300}{150} \approx 10^{88}$
- Take 10000 indep subsamples size 150
- $v_{J-150}(\widehat{\theta}) = 0.0279$
- $v_{J-100}(\widehat{\theta}) = 0.0255$
- $v_{J-25}(\widehat{\theta}) = 0.0212$
- $v_{J-1}(\widehat{\theta}) = 0.0101$

Bootstrap (IID case): Efron (1982)

- Draw B samples $(y_1^*, \ldots, y_n^*)$ by SRS with replacement from $\mathbf{y} = (y_1, \ldots, y_n)$
- Calculate bootstrap estimates $\theta^*(1), \ldots, \theta^*(B)$ from the B bootstrap samples.
- Let $\theta^*(\cdot) = \frac{1}{B} \sum_{b=1}^{B} \theta^*(b)$

(i) Bootstrap estimate of bias of $\widehat{\theta} = \theta^*(\cdot) - \widehat{\theta}$
Bias corrected estimator of $\widehat{\theta} = 2\widehat{\theta} - \theta^*(\cdot)$

(ii) Variance estimator:

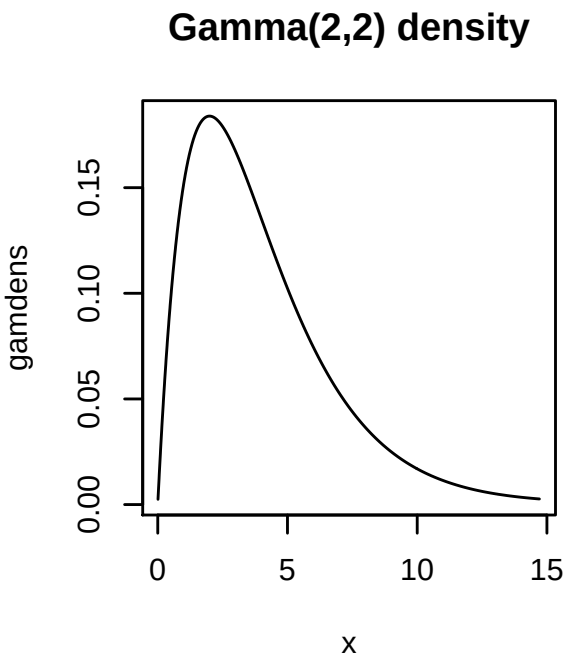
$$v_B(\hat{\theta}) = \frac{1}{B} \sum_{b=1}^{B} [\theta^*(b) - \theta^*(\cdot)]^2$$

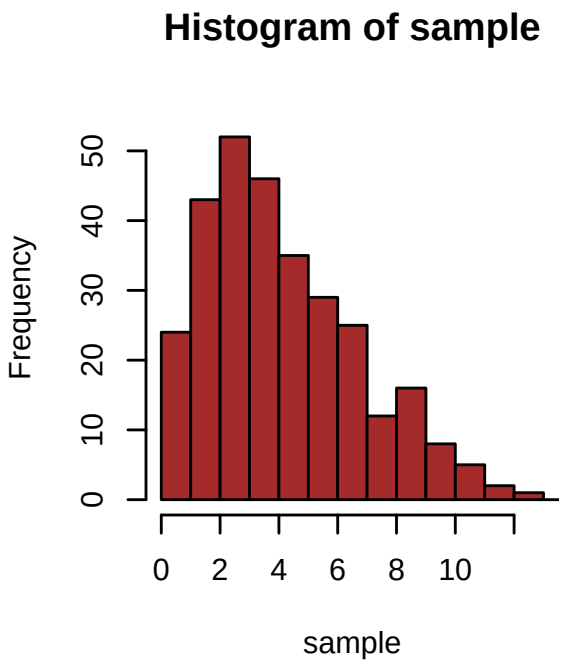
Linear case $\hat{\theta} = \bar{y}$:

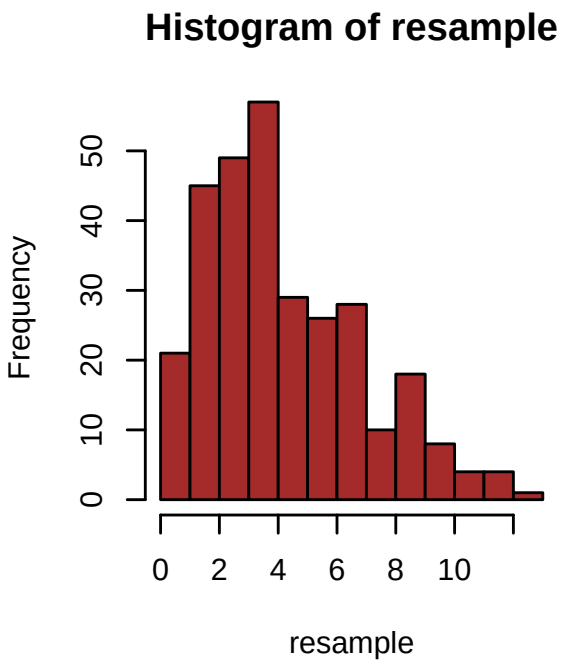
$$v_B(\hat{\theta}) = \frac{n-1}{n} \frac{s_y^2}{n} \approx \frac{s_y^2}{n}$$

For quantiles (median) $\hat{\theta}$, $v_B(\hat{\theta})$ consistent if $E|y|^\alpha < \infty$, $\alpha > 0$.

Example of bootstrap: SRS ($n = 300$)







Bootstrap confidence intervals (CI):

1. Percentile Method:

Distribution of $\widehat{\theta} - \theta$ approximated by bootstrap distribution of $\theta^* - \widehat{\theta}$

Order the bootstrap values as $\theta^*_{(1)} - \widehat{\theta}, \ldots,$ $\theta^*_{(B)} - \widehat{\theta}$, say $B = 1000$.

Suppose $1 - \alpha = 0.90$. Get a_P and b_P from the bootstrap histogram of $\theta^*_{(b)} - \hat{\theta}$ such that $100\alpha/2 = 5\%$ of values below a_P and 5% of values above b_P: $a_P = \theta^*_{(50)} - \hat{\theta}$, $b_P = \theta^*_{(950)} - \hat{\theta}$

$1 - \alpha = 0.9$ level interval on θ: $a_P \leq \hat{\theta} - \theta \leq b_P \Longleftrightarrow 2\hat{\theta} - \theta^*_{(950)} \leq \theta \leq 2\hat{\theta} - \theta^*_{(50)}$

Note: Interval not symmetric, but no more accurate than symmetric normal interval.

Bootstrap distribution of $\theta^* - \widehat{\theta}$

$B = 1000$

median $= \widehat{\theta} = 3.693$

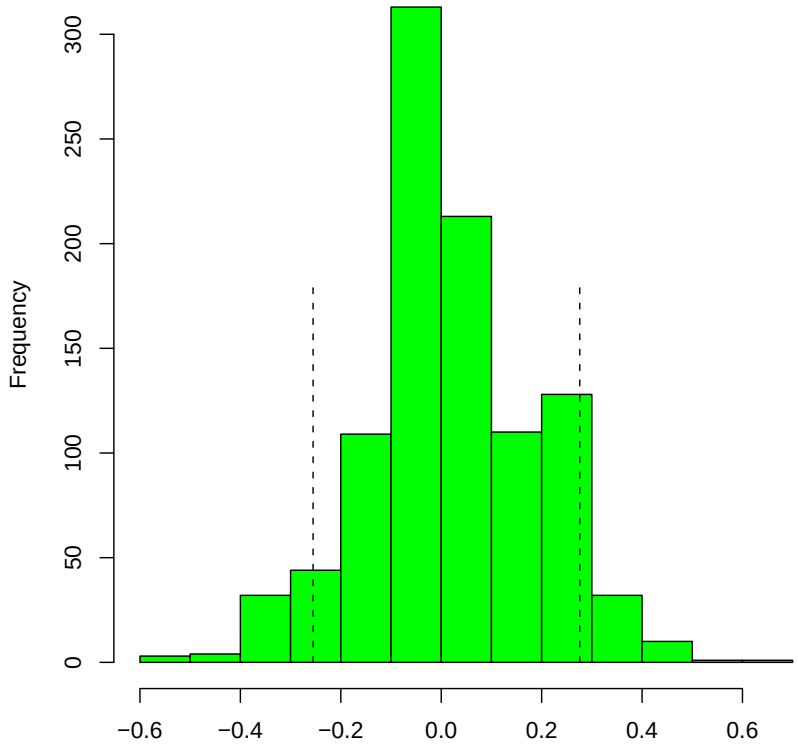
$v_B(\widehat{\theta}) = 0.0277$

$\theta^*_{(50)} - \widehat{\theta} = -0.255$
$\theta^*_{(950)} - \widehat{\theta} = 0.276$

90% CI (percentile method):

$[2\widehat{\theta} - \theta^*_{(950)}, 2\widehat{\theta} - \theta^*_{(50)}]$

$[3.417, 3.948]$



2. Bootstrap-t:

Approximate distribution of $t = \frac{\widehat{\theta}-\theta}{s(\widehat{\theta})}$ by bootstrap distribution of $t^* = \frac{\theta^*-\widehat{\theta}}{s(\theta^*)}$
From the bootstrap histogram of $t_1^*, \ldots, t_B^*$, obtain the points a_t and b_t such that 5% of t^*–values lie below a_t and 5% above b_t.

$1-\alpha = 0.9$ level interval on θ: $a_t \leq t \leq b_t \Longleftrightarrow$
$\widehat{\theta} - b_t s(\widehat{\theta}) \leq \theta \leq \widehat{\theta} - a_t s(\widehat{\theta})$

Choice of $s^2(\hat{\theta})$ and $s^2(\theta^*)$: Use jackknife on both if $\hat{\theta}$ smooth

Use bootstrap on both if $\hat{\theta}$ nonsmooth, or Woodruff-based for quantiles.

Bootstrap-t computationally intensive, but controls error rates in both tails more accurately than normal interval. Length of interval, however, is larger.

Histogram of $t^* = (\theta^* - \hat{\theta})/s(\theta^*)$

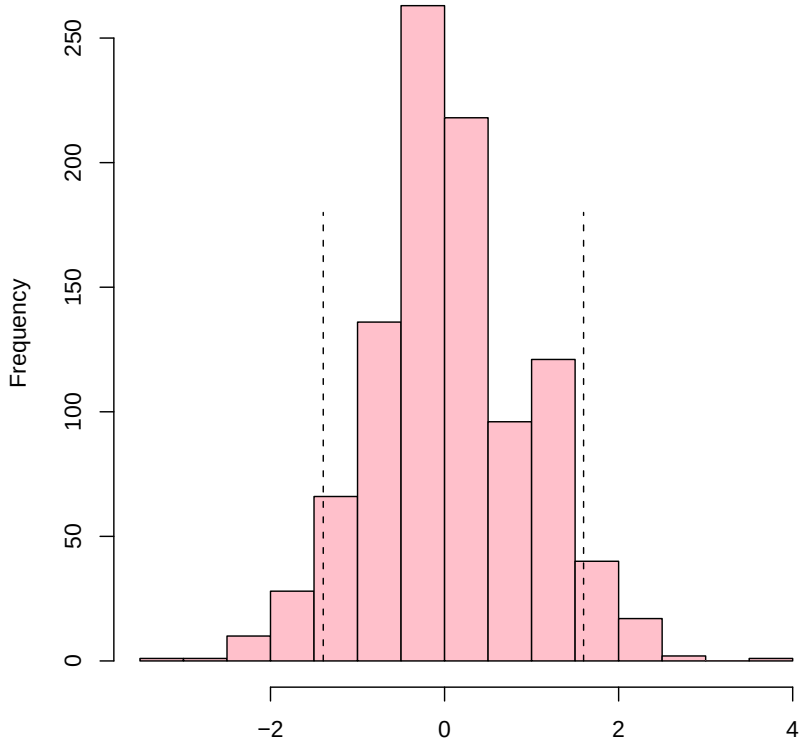
$B = 1000$

median $= \hat{\theta} = 3.693$

$s_B(\hat{\theta}) = 0.1664$

$t^*_{(50)} = -1.394$
$t^*_{(950)} = 1.597$

90% CI (bootstrap t):
$[\hat{\theta} - t^*_{(950)} s_B(\hat{\theta}), \hat{\theta} - t^*_{(50)} s_B(\hat{\theta})]$
$[3.427, 3.925]$



Stratified multi-stage sampling: Jackknife

Basic weights d_{hik}; calibration weight $w_{hik} = d_{hik}\, g_{hik}$

$$g_{hik}(s) = \mathbf{x}^T \left(\sum_{hik \in s} d_{hik}\, \mathbf{x}_{hik}\, \mathbf{x}_{hik}^T \right)^{-1} \mathbf{x}_{hik}$$

$\mathbf{x}_{hik}$ post-strata indicators with known totals $\mathbf{X}$: census projections

Single post-stratifier: $g_{hik} = {}_cM \Big/ \sum_{cs} d_{hik}$

${}_cM = c$-th poststratum size; ${}_cs =$ sample from c-th poststratum

Basic jackknife weights when (gj)-th sample PSU deleted: $d_{hik(gj)} = d_{hik}\, b_{gj}$

$$b_{gj} = \begin{cases} 0 & \text{if } (hi) = (gj) \\ n_g/(n_g - 1) & \text{if } h = g,\ i \neq j \\ 1 & \text{if } h \neq g \end{cases}$$

Replace d_{hik} in w_{hik} by $d_{hik(gj)}$ to get jackknife calibration weights $w_{hik(gj)}$ when (gj)-th PSU deleted: Calculate

$$\hat{Y}_{GR} = \sum_s w_{hik}\, y_{hik};\;\; \hat{Y}_{GR(gj)} = \sum_s w_{hik(gj)}\, y_{hik}$$

Jackknife variance estimator of GREG $\hat{Y}_{GR}$:

$$v_J(\hat{Y}_{GR}) = \sum_{g=1}^{L} \frac{n_g - 1}{n_g} \sum_{j=1}^{n_g} (\hat{Y}_{GR(gj)} - \hat{Y}_{GR})^2$$

Yung & Rao (2000): Asymp. consistency of v_J

Regularity conditions:
(i) No design weight, d_{hik}, is disproportionately large
(ii) $2 + \delta$ moment of PSU totals of residuals bounded for some $\delta > 0$.

Application of $\hat{Y}_{GR}$ and $v_J(\hat{Y}_{GR})$: Canadian Labour Force Survey (LFS)

General parameters θ

$$\begin{aligned}
\widehat{\theta} &= \text{Estimator of } \theta \text{ based on } w_{hik} \\
\widehat{\theta}_{(gj)} &= \text{Estimator of } \theta \text{ based on } w_{hik(gj)} \\
v_J(\widehat{\theta}) &= \sum_{g=1}^{L} \frac{n_g - 1}{n_g} \sum_{j=1}^{n_g} (\widehat{\theta}_{(gj)} - \widehat{\theta})(\widehat{\theta}_{(gj)} - \widehat{\theta})^T
\end{aligned}$$

Data file provides weights w_{hik} and $w_{hik(gj)}$, each (gj).

Note $w_{hik(gj)} = 0$ if $(hi) = (gj)$ and 0–values reveal membership in the same cluster.

Linear case $\widehat{\theta} = \widehat{Y} = \sum_s d_{hik}\, y_{hik}$: $v_J(\widehat{Y}) = v(y_{hi})$.

Most of asymptotic theory for jackknife focussed on estimators based on basic design weights, not on calibration weights: see Shao & Tu (1995, chap. 6); also simulation results. Krewski & Rao (1981): Smooth functions of totals.

Rao & Wu (1985): Linearization v_L and jackknife v_J are asymptotically equivalent to second order if $n_h = 2$ for all h:

$v_J/v_L = 1 + O_p(n^{-2})$ if $n_h = 2$ for all h

Shao (1994): L-statistics: linear combinations of order statistics

Smooth L-statistic: Jackknife consistent

Order $\{y_{hik}\}$ as $\{y_{(\ell)}\}$.

Let $w_{(\ell)} = w_{hik} / \sum_s w_{hik}$ if $y_{(\ell)} = y_{hik}$

$\widehat{\theta} = \sum_{\ell} c_\ell y_{(\ell)}$; $\quad c_\ell = w_\ell J\left(\sum_{t=1}^{\ell} w_t\right)$

Note: $J(\cdot) = 1$ gives $\hat{\theta} =$ weighted sample mean $= \sum_s w_{hik} y_{hik} / \sum_s w_{hik}$

- Trimmed mean: $J(a) = \dfrac{I(\alpha \leq a \leq \beta)}{\beta - \alpha}$; $0 \leq \alpha < \beta \leq 1$ is smooth.
- Other examples: Gini family, Lorenz curve (measures of income inequality).
- Nonsmooth functions $\widehat{\theta}$: Asymptotic consistency of v_J?
- If the sample size within a cluster fairly large, delete-cluster v_J should perform quite well: Rao, Wu and Yue (1992).

Jackknife Linearization (Yung & Rao, 1996)

For GREG estimator $\hat{Y}_{GR}$,

$$\hat{Y}_{GR(gj)} - \hat{Y}_{GR} \approx \frac{1}{n_g - 1}(\tilde{e}_{g.} - \tilde{e}_{gj})$$

$$\tilde{e}_{hi} = \sum_k (n_h w_{hik}) e_{hik}$$

Therefore

$$v_J(\hat{Y}_{GR}) \approx v_{JL}(\hat{Y}_{GR}) = v(\tilde{e}_{hi})$$

v_{JL} is identical to g-weighted linearization variance estimator $v(\tilde{e}_{hi}) = v_L$. Valliant (1993) obtained v_{JL} for the special case of single post-stratifier. Conditional properties of v_J, v_{JL} and v_L studied by Valliant and Yung & Rao. They demonstrated v_J and v_{JL} perform well conditionally unlike v_L.

(1) As noted before, Canadian LFS used $\hat{Y}_{GR}$ and $v_J(\hat{Y}_{GR})$. LFS sample also used for many supplementary household surveys, and users wanted additional calibration constraints. Jackknife is computer intensive because of large number of PSU's. Users wanted a computationally simpler variance estimator which can approximate the LFS v_J very well. This led to v_{JL}.

(2) Canadian Heart Health Surveys in the 10 provinces of Canada used stratified multistage sampling and post-stratification information (age-sex counts). Calibration weights were used to obtain prevalence rates of risk factors and other estimates of totals, means, proportions as well as logistic regression etc. At that time, existing software based on linearization did not incorporate poststratification. As a result, a jackknife software, JACKVAR, was developed to handle complex analyses taking account of post-stratification.

Balanced Repeated Replication: $n_h = 2$

McCarthy (1969)

For $t = 1, \ldots, T$, half-sample t:

$$\delta_h^t = \begin{cases} 1 & \text{if cluster (h1) selected} \\ -1 & \text{if cluster (h2) selected} \end{cases}$$

Orthogonality: $\sum_{t=1}^{T} \delta_h^t \delta_\ell^t = 0, h \neq \ell = 1, \ldots, L$

$T \times T$ Hadamard matrix (δ_h^t);
$L + 1 \leq T \leq L + 4$

Example: $L = 3, T = 4$

	$h = 1$	$h = 2$	$h = 3$
$t = 1$	$+1$	$+1$	$+1$
$t = 2$	$+1$	-1	$+1$	-1
$t = 3$	$+1$	-1	-1	$+1$
$t = 4$	$+1$	$+1$	-1	-1

Choose last three columns

- BRR weights $d_{hik}(t) = d_{hik}\, b_{hi}(t)$

$$b_{hi}(t) = 2 \text{ if cluster } (hi) \text{ in half-sample}$$
$$= 0 \text{ otherwise}$$

- Replace d_{hik} by $d_{hik}(t)$ in calibration weight w_{hik}:
  $w_{hik}(t) = d_{hik}(t) g_{hik}(t)$: BRR calib. wts.
- Calculate $\widehat{\theta}$ based on w_{hik} and $\widehat{\theta}(t)$ based on $w_{hik}(t)$.

- BRR variance estimator:

$$v_{\text{BRR}}(\hat{\theta}) = \frac{1}{T}\sum_{t=1}^{T}(\hat{\theta}(t) - \hat{\theta})^2$$

- Linear case $\hat{Y} = \sum_s d_{hik}\, y_{hik}$:

$$v_{\text{BRR}}(\hat{Y}) = v(y_{hi})$$

- Main advantage of BRR: $v_{\text{BRR}}(\widehat{\theta})$ valid for nonsmooth $\widehat{\theta}$.
- Limitations of BRR: Difficult to construct pseudo-replicates for arbitrary n_h. BRR does not exist if $n_h = 6$.
- Balanced Orthogonal Multi-Arrays (BOMA) Sitter (1993): Allows greater flexibility, fewer replicates, covers a wide range of $n_h = p$ (prime or power of prime)

Fay's Modification of BRR ($n_h = 2$):

$$b_{hi}(t, \varepsilon) = 1 + \varepsilon \text{ if } (hi) \text{ in half-sample}$$
$$= 1 - \varepsilon \text{ otherwise}$$

Modified BRR weights: $d_{hik}(t, \varepsilon) = d_{hik} b_{hi}(t, \varepsilon)$

$$\varepsilon^2 v_{\text{BRR}\,(\varepsilon)}(\hat{\theta}) = \frac{1}{T} \sum_{t=1}^{T} (\hat{\theta}(t, \varepsilon) - \hat{\theta})^2$$

$\varepsilon = 1$: BRR ; $\varepsilon = \frac{1}{2}$ good choice
Utilizes all observations unlike BRR.

Bootstrap: Stratified multistage sampling

- Rao & Wu (1988); Rao, Wu & Yue (1992)
- Draw $m_h = n_h - 1$ clusters (PSUs) from n_h clusters by SRSWR
- Bootstrap weights $d_{hik}(b) = d_{hik} e_{hi}(b)$

$$e_{hi}(b) = \frac{n_h}{m_h} m_{hi}(b)$$

$m_{hi}(b) =$ number of times hi selected in b^{th} bootstrap sample $(b = 1, \ldots, B)$

Bootstrap: Stratified multistage sampling

- Replace d_{hik} by $d_{hik}(b)$ in w_{hik}

$$w_{hik}(b) = d_{hik}(b)g_{hik}(b) :$$

bootstrap calib. wts.

- Calculate $\hat{\theta}$ based on w_{hik} and $\hat{\theta}(b)$ based on $w_{hik}(b)$

- Bootstrap variance estimator:

$$v_B(\hat{\theta}) = \frac{1}{B} \sum_{b=1}^{B} (\hat{\theta}(b) - \hat{\theta})^2$$

- Linear case $\hat{Y} = \sum_s d_{hik}\, y_{hik}$
  $v_B(\hat{Y}) = v(y_{hi})$ if $B = \infty$.
- Balanced bootstrap: exact matching for finite B: Rao and Vijayan (2001)
- Choice of B flexible unlike in Jackknife with $n = \sum n_h$ pseudo-replicates
- Valid for arbitrary n_h unlike BRR

Examples: Replication methods, health surveys

National Health and Nutrition Examination Survey (NHANES) I: Linear Regression

- 4 stages of sampling:
  PSU (county or group of counties)
  census enumeration district
  segment (cluster of households)
  person
- Post-stratified by age, race, sex categories

- Regression of systolic blood pressure on age, race, body mass index for sample of 13573 persons, 2 PSU's per stratum, 35 strata, BRR
- Design effects for reg coefs ranged from 2.69 to 3.61, all variables had highly significant coefs

Canada Heart Health Surveys (1986–92): Modelling and Hypothesis Testing

- Stratified 2-stage sample of about 2000 men and women (18-74) selected in each province using health insurance registries as frames. Health units are PSU's selected by PPS sampling.
- Statistical analyses:
  Compare 2 distributions of proportions
  Test indep., measure association in multiway table
  Logistic regression
  Test for trends in proportions
- Rust and Rao (1996)

Implementing Variance Estimation Methods

- Commercial software implements some methods
- Need to know how to do jackknife, bootstrap for analyses not yet implemented in software
- Keep observations in the same psu together when constructing replicates

Methods in Software

	Lin	BRR	JK	BS	GREG
SUDAAN	X	X	X		some
WesVar		X	X	some	some
SAS/STAT	X			SRS	some
Stata	X		SRS	SRS	some
GES	X		X		X
IVEware	X		X		some
R functions	X	X	X	SRS	some

Variance Estimation: NCVS

- Codebook: generalized variance functions
- Since 1995, public use data files contain pseudo-strata, pseudo-psu's; intended to mirror actual psu structure while maintaining confidentiality
- Public use files: only first-stage information
- Can use lin., BRR, jackknife, bootstrap
- Note: results presented here not "official" victimization statistics

Linearization: 2000 NCVS

- Estimate $\theta =$ proportion of violent victimizations that involve injury
- $y_i = 1$ if injury; $x_i = 1$ if victim of violent crime
- Form $z_i = (y_i - \hat{R}x_i)/\hat{\bar{X}}$

SAS code (linearization)

```
proc surveymeans data=ncvs2000;
   strata v2117;
   cluster v2118;
   weight v3080;
   var sex age violent injury ;
   ratio injury / violent ;
```

SAS Output (linearization)

Data Summary

Variable	N	Mean	Std Error of Mean	Lower 95% CL for Mean	Upper 95% CL for Mean
sex	79280	0.515440	0.001597	0.512283	0.518598
age	79360	41.690112	0.150683	41.392258	41.987966
violent	79360	0.012251	0.000621	0.011025	0.013478
injury	79360	0.003917	0.000331	0.003264	0.004571

Ratio Analysis

Numerator	Denom	N	Ratio	Std Err	95% CI	
injury	violent	79360	0.319736	0.020631	0.278954	0.360518

Jackknife: NCVS

- Form replicate weights deleting one psu at a time
- Can be done in Wesvar, other software
- Easy to implement in R
- $\text{SE}_J(\hat{\theta}) = 0.020634$

Jackknife weights: NCVS

Stratum	psu	w	$w_{(11)}$	$w_{(12)}$	$w_{(21)}$	$w_{(22)}$
1	1	2746	0	5491	2746	2746
1	1	2547	0	5094	2547	2547
1	2	2027	4054	0	2027	2027
1	2	2325	4649	0	2325	2325
2	1	2895	2895	2895	0	5790
2	1	2374	2374	2374	0	4748
2	1	2137	2137	2137	0	4275
2	1	2378	2378	2378	0	4755
2	2	2543	2543	2543	5085	0
2	2	2410	2410	2410	4820	0

Sample Survey Methods: Recent Developments and Applications

Quantiles

Quantiles

Woodruff (1952) Confidence Intervals

Population distribution function: $F(t) = \frac{1}{N} \sum_{i \in U} I[y_i \leq t]$

$$u_i(t) = I[y_i \leq t] = \begin{cases} 1 & \text{if } y_i \leq t \\ 0 & \text{otherwise} \end{cases}$$

qth quantile = $\theta_q = F^{-1}(q)$; median = $\theta_{1/2}$

$\hat{F}(t) = \sum_{i \in s} w_i u_i(t) / \sum_{i \in s} w_i = \hat{U}(t)$

$w_i = d_i$ if only design weight $d_i = 1/\pi_i$ used ($w_i > 0$)

$\hat{\theta}_q = \hat{F}^{-1}(q)$ = smallest y satisfying $\hat{F}(y) \geq q$

Calculation of $\widehat{\theta}_q$

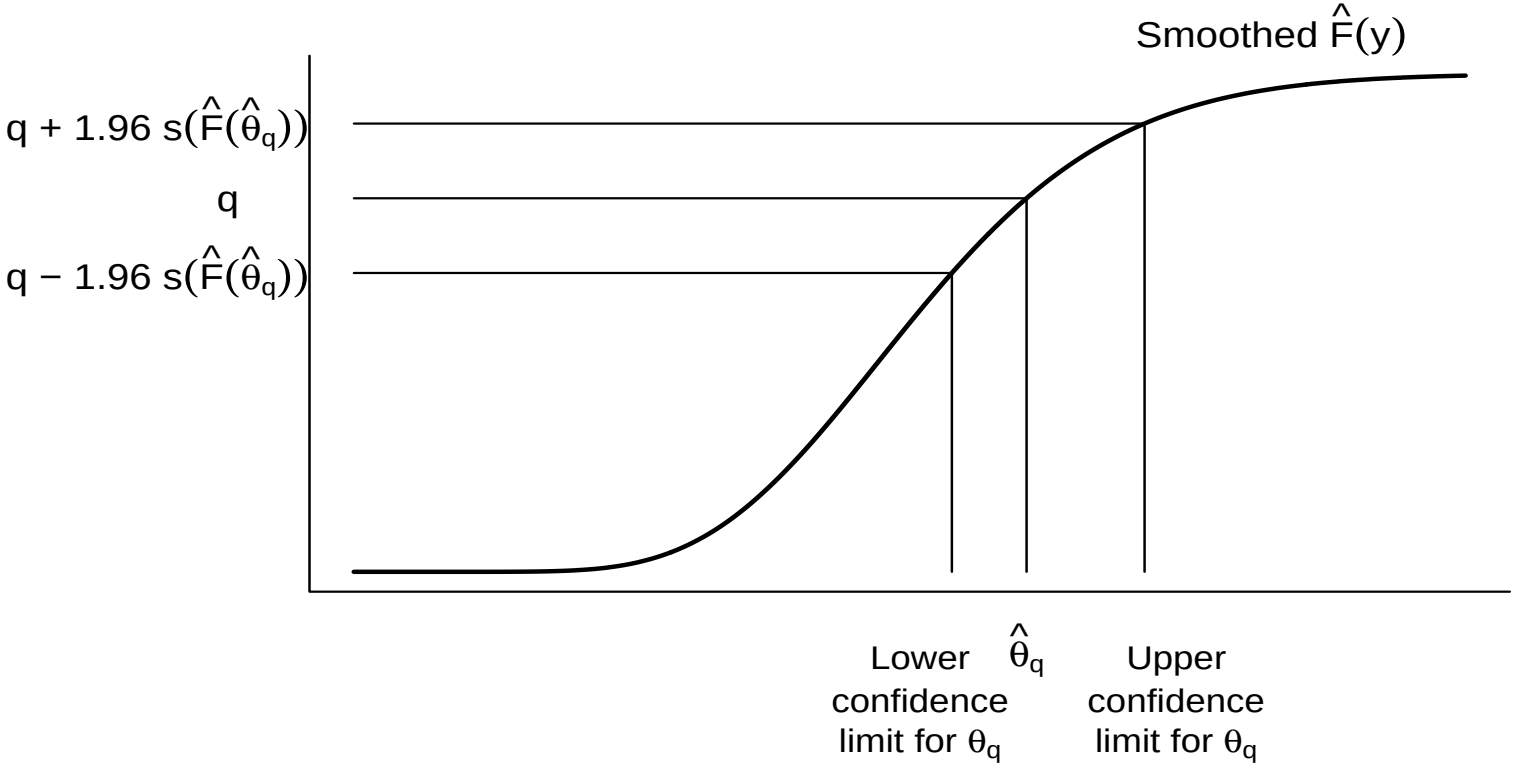
Arrange $y_i, i \in s$ in ascending order, say $y_{(i)}$
Cumulate the associated normalized weights $w_i / \sum_{j \in s} w_j = \tilde{w}_i$ until q is first crossed. The first $y_{(i)}$ encountered after crossing q is taken as $\widehat{\theta}_q$

Taylor linearization variance estimator of $\widehat{F}(t)$:
$s^2[\widehat{F}(t)] = v[\{u(t) - \widehat{F}(t)\}/\hat{N}]$
95% CI on $F(t)$: $\widehat{F}(t) \pm 1.96 s[\widehat{F}(t)]$

95% Woodruff interval on θ_q, $L_q \leq \theta_q \leq U_q$

$L_q = \hat{F}^{-1}\{q - 1.96 s[\hat{F}(\hat{\theta}_q)]\}, U_q = \hat{F}^{-1}\{q + 1.96 s[\hat{F}(\hat{\theta}_q)]\}$

$s[\hat{F}(\hat{\theta}_q)]$ is obtained by replacing θ_q by $\hat{\theta}_q$ in $s[\hat{F}(\theta_q)]$



For small or large q and moderate sample size, normal CI on $F(t)$ performs poorly. As a result, Woodruff interval may be expected to perform poorly, but this is not so (Sitter & Wu, 2001). Woodruff intervals perform surprisingly well for all q: replacing θ_q by $\widehat{\theta}_q$ in $s[\widehat{F}(\theta_q)]$ improved coverage, but inflated length somewhat.

Standard error of $\widehat{\theta}_q$ using Woodruff interval (Rao & Wu, 1987; Francisco & Fuller, 1991):

$$s(\widehat{\theta}_q) = (U_q - L_q)/(2 \times 1.96)$$
$$= \text{(Length of CI)}/(2 \times 1.96)$$

We used $1 - \alpha = 0.95$ interval above, but any α may be used by changing 1.96 to $z_{(\alpha/2)}$ to get $L_q(\alpha)$ and $U_q(\alpha)$ and

$$s_\alpha(\widehat{\theta}_q) = (U_q(\alpha) - L_q(\alpha))/(2 \times z_{(\alpha/2)})$$

Simulation Study

Stratified random sampling with 32 strata, $n_h = 5$

	% Rel. Bias	Rel. stability
$s_{0.01}(\widehat{\theta}_{1/2})$	3.2	0.40
$s_{0.025}(\widehat{\theta}_{1/2})$	3.7	0.43
$s_{0.05}(\widehat{\theta}_{1/2})$	5.2	0.48
$s_{0.10}(\widehat{\theta}_{1/2})$	4.7	0.57

Customary choice $\alpha = 0.05$ works well

Sample Survey Methods: Recent Developments and Applications

Estimating Equations

General Parameters: Solutions of Census Estimating Equations (EE)

Scalar Parameter θ

Census EE: $S_N(\theta) = \sum_{j \in U} u(y_j, \theta) = 0$

$\Rightarrow \theta_N =$ census parameter

Sample EE: $\widehat{S}(\theta) = \sum_{i \in s} w_i u(y_i, \theta) = 0$

$\Rightarrow \widehat{\theta} =$ GREG est of θ

w_i = calibration weight

Examples

(i) $u(y,\theta) = y - \theta \Rightarrow \theta_N = \frac{1}{N}\Sigma\, y_j = \bar{Y}$

$\widehat{\theta} = \sum_{i\in s} w_i y_i / \sum_{i\in s} w_i = \bar{y}_w$

(ii) $u(y,\theta) = I(y \le t) - \theta \Rightarrow$

$\theta_N = \frac{1}{N}\Sigma\, I(y_j \le t) = F_N(t)$

$\widehat{\theta} = \widehat{F}(t) = \sum_{i\in s} w_i I(y_i \le t) / \sum_{i\in s} w_i$

(iii) Quantile: $u(y,\theta) = I(y \le \theta) - q \Rightarrow$

$\widehat{\theta} = \widehat{F}^{-1}(q)$

Vector $\boldsymbol{\theta}$: $\mathbf{S}_N(\boldsymbol{\theta}) = \sum_{j \in U} \mathbf{u}(y, \boldsymbol{\theta}) = \mathbf{0} \Rightarrow \boldsymbol{\theta}_N$

$\widehat{\mathbf{S}}(\boldsymbol{\theta}) = \sum_{i \in s} w_i \mathbf{u}(y_i, \boldsymbol{\theta}) = 0 \Rightarrow \widehat{\boldsymbol{\theta}}$: GREG est

Examples: (i) Linear regression

$$\mathbf{S}_N(\boldsymbol{\theta}) = \sum \mathbf{z}_j (y_j - \boldsymbol{\theta}^T \mathbf{z}_j) = \mathbf{0}$$
$$\Rightarrow \boldsymbol{\theta}_N = (\sum \mathbf{z}_j \mathbf{z}_j^T)^{-1} (\sum \mathbf{z}_j y_j)$$

$$\widehat{\boldsymbol{\theta}} = (\sum_{i \in s} w_i \mathbf{z}_i \mathbf{z}_i^T)^{-1} (\sum_{i \in s} w_i \mathbf{z}_i y_i)$$

Motivating model:

$$E_m(y_j) = \mu_j(\boldsymbol{\theta}),\ V_m(y_j) = V_{0j} = V_0(\mu_j)$$

Generalized EE:

$$u_\ell(y_j, \boldsymbol{\theta}) = \left[\frac{\partial \mu_j(\boldsymbol{\theta})}{\partial \theta_\ell}\right][y_j - \mu_j(\boldsymbol{\theta})]V_{0j}^{-1}$$

$$\mu_j(\boldsymbol{\theta}) = \mathbf{z}_j^T\boldsymbol{\theta},\ V_{0j} = \sigma^2,$$
$$\mathbf{u}(y_j, \boldsymbol{\theta}) = \mathbf{z}_j(y_j - \mathbf{z}_j^T\boldsymbol{\theta})$$

(ii) Logistic Regression: y binary (0 or 1)

$$\log\left\{\frac{\mu_j(\boldsymbol{\theta})}{1-\mu_j(\boldsymbol{\theta})}\right\} = \mathbf{z}_j^T\boldsymbol{\theta},\ V_{0j} = \mu_j(1-\mu_j)$$

$$\mathbf{u}(y_j, \boldsymbol{\theta}) = \mathbf{z}_j\{y_j - \mu_j(\boldsymbol{\theta})\}$$: $\widehat{\theta}$ by iteration

Newton-Raphson (NR):

$$\widehat{\boldsymbol{\theta}}_r = \widehat{\boldsymbol{\theta}}_{r-1} + [\widehat{\mathbf{J}}(\widehat{\boldsymbol{\theta}}_{r-1})]^{-1}\widehat{\mathbf{S}}(\widehat{\boldsymbol{\theta}}_{r-1})$$

$$\widehat{\mathbf{J}}(\boldsymbol{\theta}) = -\partial\widehat{\mathbf{S}}(\boldsymbol{\theta})/\partial\boldsymbol{\theta}^T$$

Linearization variance estimator:

$\mathbf{u}(y, \theta)$ differentiable function of θ

Scalar θ:

$$s^2(\hat{\theta}) = v[g\hat{J}^{-1}(\hat{\theta})e(\hat{\theta})] = v[g\tilde{e}(\hat{\theta})] = v[e^*(\hat{\theta})],$$

$$\begin{aligned}
e_i(\hat{\theta}) &= u_i(\hat{\theta}) - \hat{\mathbf{B}}(\hat{\theta})^T \mathbf{x}_i, \\
\tilde{e}_i(\hat{\theta}) &= \hat{J}^{-1}(\hat{\theta}) e_i(\hat{\theta}) \\
\hat{\mathbf{B}}(\hat{\theta}) &= \left(\textstyle\sum_{i \in s} w_i \mathbf{x}_i \mathbf{x}_i^T\right)^{-1} \left(\textstyle\sum_{i \in s} w_i \mathbf{x}_i u_i(\hat{\theta})\right)
\end{aligned}$$

$e_i(\hat{\theta})$ are usual resids with y_i changed to $u_i(\hat{\theta})$

Vector $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)^T$

Denote $\mathbf{v}(\mathbf{y})$ as design-based variance estimator of $\hat{\mathbf{Y}}$, vector of estimated totals

$$\text{Var. est.}(\hat{\boldsymbol{\theta}}) = \mathbf{v}[g\hat{\mathbf{J}}^{-1}(\hat{\boldsymbol{\theta}})\mathbf{e}(\hat{\boldsymbol{\theta}})] = \mathbf{v}[g\tilde{\mathbf{e}}(\hat{\boldsymbol{\theta}})]$$

where ℓ–th element of $\mathbf{e}_i(\hat{\boldsymbol{\theta}})$ is

$$e_{i\ell}(\hat{\theta}) = u_{i\ell}(\hat{\boldsymbol{\theta}}) - \hat{\mathbf{B}}_\ell(\hat{\boldsymbol{\theta}})^T \mathbf{x}_i,$$

$$\hat{\mathbf{B}}_\ell(\hat{\boldsymbol{\theta}}) = \left(\sum_{i\in s} w_i \mathbf{x}_i \mathbf{x}_i^T\right)^{-1} \left(\sum_{i\in s} w_i \mathbf{x}_i u_{i\ell}(\hat{\boldsymbol{\theta}})\right).$$

Stratified multistage sampling (scalar θ):

$$\begin{aligned}\text{Var. est.}(\hat{\theta}) &= \sum_h \frac{1}{n_h(n_h-1)} \sum_i [e^*_{hi}(\hat{\theta}) - \bar{e}^*_h(\hat{\theta})]^2 \\ &= v[e^*_{hi}(\hat{\boldsymbol{\theta}})]\end{aligned}$$

$$e^*_{hi}(\hat{\theta}) = n_h \sum_k d_{hik}\, g_{hik} e^*_{hik}(\hat{\boldsymbol{\theta}}),$$

$$\bar{e}^*_h(\hat{\theta}) = \frac{1}{n_h} \sum_i e^*_{hi}(\hat{\theta}).$$

Note: Only cluster totals $e^*_{hi}(\hat{\theta})$ are needed.

Implementation

Estimation package should be able to:

- calculate the g–weights $g_i(s)$ given some auxiliary $\mathbf{x}_i$ with known totals $\mathbf{X}$
- calculate the variance estimator $\mathbf{v}(\mathbf{y})$ of the basic estimator $\hat{\mathbf{Y}} = \sum_s d_i \mathbf{y}_i$ for the given sampling design
- calculate residuals e_i

Step 1. Given basic weights d_i and auxiliary vector $\mathbf{x}_i$ with known $\mathbf{X}$, use estimation package to calculate $g_i(s)$ and the calibration weight $w_i = d_i g_i(s)$.

2. Given the $\mathbf{u}_i(\boldsymbol{\theta})$ and w_i, define sample EE:

$$\hat{\mathbf{S}}(\boldsymbol{\theta}) = \sum_{i \in s} w_i \mathbf{u}_i(\boldsymbol{\theta}) = \mathbf{0}$$

A program is needed to calculate $\hat{\boldsymbol{\theta}}$ from sample EE using NR algorithm. As a by-product to NR, we get $\hat{\mathbf{J}}(\hat{\boldsymbol{\theta}})$. If EE has closed form solution, NR will converge in one iteration.

3. Given $\hat{\boldsymbol{\theta}}$, evaluate $\mathbf{u}_i(\hat{\boldsymbol{\theta}})$ and use estimation package to calculate $\mathbf{e}_i(\hat{\boldsymbol{\theta}})$, $i \in s$.

4. Finally, need a computer program to calculate synthetic residuals $\mathbf{e}_i^*(\widehat{\boldsymbol{\theta}}) = g_i\tilde{\mathbf{e}}_i(\widehat{\boldsymbol{\theta}})$. Use these synthetic residuals in $\mathbf{v}(\mathbf{y})$ by replacing $\mathbf{y}_i$ by $\mathbf{e}_i^*(\widehat{\boldsymbol{\theta}})$ to get variance estimate of $\widehat{\boldsymbol{\theta}}$.

See Binder (1983), Rao, Yung & Hidiroglou (2002).

Note: Existing software does not seem capable of calculating linearization variance estimators for GREG estimator of $\widehat{\boldsymbol{\theta}}$ specified by $\mathbf{u}(\boldsymbol{\theta})$.

Jackknife Variance Estimation

Jackknife may be used with EE: simply replace calibration weights by jackknife calibration weights to get $\widehat{\boldsymbol{\theta}}_{(gj)}$ in linear or logistic regression.

Estimating Function Bootstrap

- Bootstrap PSUs SRS with replacement
- Basic weight: d_{hik}
- Basic bootstrap weight: $d^*_{hik}(b)$
- Calibration weight: w_{hik}
- Bootstrap calibration weight: $w^*_{hik}(b)$

Usual Bootstrap (EE)

- Full sample EE:

  $$\widehat{\mathbf{S}}(\boldsymbol{\theta}) = \sum w_{hik}\mathbf{u}_{hik}(\boldsymbol{\theta}) = \mathbf{0} \quad \Rightarrow \widehat{\boldsymbol{\theta}}$$

- Bootstrap EE:

  $$\mathbf{S}_b^*(\boldsymbol{\theta}) = \sum w_{hik}^*(b)\mathbf{u}_{hik}(\boldsymbol{\theta}) = \mathbf{0}$$

- One-step NR: $\boldsymbol{\theta}^*(b) = \widehat{\boldsymbol{\theta}} + [\mathbf{J}_b^*(\widehat{\boldsymbol{\theta}})]^{-1}\mathbf{S}_b^*(\widehat{\boldsymbol{\theta}})$
- Note: $\boldsymbol{\theta}^*(b)$ may not exist for some b because $\mathbf{J}_b^*(\widehat{\boldsymbol{\theta}})$ may not be invertible

Bootstrap variance estimators

- $\mathbf{v}_{\mathrm{BOOT}}(\hat{\boldsymbol{\theta}}) = \frac{1}{B}\sum_{b=1}^{B}[\boldsymbol{\theta}^*(b) - \hat{\boldsymbol{\theta}}][\boldsymbol{\theta}^*(b) - \hat{\boldsymbol{\theta}}]^T$.
- Estimating function bootstrap:
  Solve $\hat{\mathbf{S}}(\boldsymbol{\theta}) = \mathbf{S}_b^*(\hat{\boldsymbol{\theta}})$ for $\boldsymbol{\theta}$
- One-step NR:
  $\tilde{\boldsymbol{\theta}}(b) = \hat{\boldsymbol{\theta}} - [\hat{\mathbf{J}}(\hat{\boldsymbol{\theta}})]^{-1}\mathbf{S}_b^*(\hat{\boldsymbol{\theta}})$

Note: Only the full-sample inverse of $\hat{\mathbf{J}}(\hat{\boldsymbol{\theta}})$ needed.

- EF Bootstrap variance estimator:

$$\mathrm{v}_{\mathrm{BOOT}}(\hat{\boldsymbol{\theta}}) = \frac{1}{B}\sum_{b=1}^{B}[\tilde{\boldsymbol{\theta}}(b) - \hat{\boldsymbol{\theta}}][\tilde{\boldsymbol{\theta}}(b) - \hat{\boldsymbol{\theta}}]^T.$$

- EF method can also be used with jackknife (Rao and Tausi, 2004)
- EF method can be used to construct EF bootstrap or jackknife calibration (GREG) weights that avoid repeated inversions.

Inverse Sampling

Undo complex survey data structure by repeated subsampling (Hinkins, Oh & Scheuren, 1997; Rao, Scott & Benhin, 2003)

$s =$ full sample

$s^* =$ subsample of size m

Inverse Sampling

Example: Two stage cluster sampling

- s: First stage PPS with replacement of cluster sizes M_i;
  Second stage SRS of size m_i from M_i; $i = 1, \ldots, k$
- s^* select one unit at random from each sample cluster $i = 1, \ldots, k$
  $s^* =$ SRS of size $m = k$.

Estimating Equations

- Draw g subsamples $s_1^*, \ldots, s_g^*$.
- Combined EE (scalar θ)

$$\begin{aligned}
\hat{U}_{gc}(\theta) &= \frac{1}{g}\sum_{j=1}^{g}\left[\frac{N}{m}\sum_{k\in s_j^*} u_k(\theta)\right] \\
&= \frac{1}{g}\sum_{j=1}^{g} U_j^*(\theta) = 0 \\
&\Rightarrow \hat{\theta}_{gc} \\
\hat{J}_{gc}(\theta) &= -\partial\hat{U}_{gc}(\theta)/\partial\theta
\end{aligned}$$

Linearization inverse sampling variance estimator

$$v(\hat{\theta}_{gc}) : \left[\hat{J}_{gc}(\hat{\theta}_{gc})\right]^{-1} v\left[\hat{U}_{gc}(\theta)\right]_{\theta=\hat{\theta}_{gc}} \left[\hat{J}_{gc}(\hat{\theta}_{gc})\right]^{-1}$$

where

$$v\left[\hat{U}_{gc}(\theta)\right] = \frac{1}{g}\sum_{j=1}^{g} v[U_j^*(\theta)] - \frac{1}{g}\sum_{j=1}^{g}(U_j^*(\theta) - \hat{U}_{gc}(\theta))^2$$

$v[U_j^*(\theta)] =$ SRS variance estimator of $U_j^*(\theta)$.

Microdata file consists of multiple subsamples without knowledge of weights, cluster labels, strata labels. Allows reduction of identification risk induced by cluster labels. Estimator $\widehat{\theta}_{gc}$ and variance estimator $v(\widehat{\theta}_{gc})$ can be computed from microdata file. Permute subsamples before reporting.

Example: Two-stage PPS cluster sample

y_{ij} = hectares corn in area seg. j of county i

x_{1ij} = # pixels classified as corn

x_{2ij} = # pixels classified as soybeans

Parameter: census regression coefficients

$$B_1,\ B_2 \text{ in } \mathbf{B} = \left(\sum_U \mathbf{x}_\ell \mathbf{x}_\ell^T\right)^{-1} \left(\sum_U \mathbf{x}_\ell \mathbf{y}_\ell\right)$$

$$\mathbf{x}_\ell = (1\ x_{1\ell}\ x_{2\ell})^T, \quad \ell \in U$$

	Full sample	g 500	1000	10000
$\hat{B}_1$	0.3176	0.3251	0.3171	0.3179
$\hat{B}_2$	-0.1326	-0.1258	-0.1330	-0.1324
$\hat{v}(\hat{B}_1) \times 10^{-3}$:	2.1153	2.2925	1.9127	2.2366
$\hat{v}(\hat{B}_2) \times 10^{-3}$:	2.7369	3.0352	2.7226	2.8038

Inverse sampling can also handle informative cluster sizes
(Benhin, Rao & Scott, 2003)

Super-population view point

	Fixed Pop.	Two-phase
Finite pop. B	yes	yes
Model β	no	yes

Hartley and Sielken (1975)

Informative Sampling

- Pfeffermann & Sverchkov (2003, Chapter 12)
- Terminology:

  $f_p(y_i \mid \mathbf{x}_i) =$ Population distribution
- Sample distribution

$$\stackrel{\text{def}}{=} f(y_i \mid \mathbf{x}_i, i \in s) = f_s(y_i \mid \mathbf{x}_i)$$

$$\stackrel{\text{Bayes}}{=} \frac{P(i \in s \mid y_i, \mathbf{x}_i) f_p(y_i \mid \mathbf{x}_i)}{P(i \in s \mid \mathbf{x}_i)}$$

- Noninformative sampling:
  $f_s(y_i \mid \mathbf{x}_i) = f_p(y_i \mid \mathbf{x}_i)$
  iff $P(i \in s \mid y_i, \mathbf{x}_i) = P(i \in s \mid \mathbf{x}_i)$

- Note: $P(i \in s \mid y_i, \mathbf{x}_i) = E_p(\pi_i \mid y_i, \mathbf{x}_i)$
  $P(i \in s \mid \mathbf{x}_i) = E_p(\pi_i \mid \mathbf{x}_i)$
  $\pi_i = P(i \in s)$

- Sample likelihood: $\prod_{i \in s} f_s(y_i \mid \mathbf{x}_i)$ assuming independence
- $E_p(\pi_i \mid y_i, \mathbf{x}_i) = 1/E_s(w_i \mid y_i, \mathbf{x}_i)$;
  $E_p(\pi_i \mid \mathbf{x}_i) = 1/E_s(w_i \mid \mathbf{x}_i)$
  Conditional expectations of π_i can be evaluated from sample data

Estimating $E_s(w_i \mid \mathbf{x}_i)$

1. Regress w_i on $(y_i, \mathbf{x}_i)$: estimate of $E_s(w_i \mid \mathbf{x}_i, y_i)$
2. Integrate $\int [E_s(w_i \mid \mathbf{x}_i, y_i)]^{-1} f_p(y_i \mid \mathbf{x}_i, \boldsymbol{\beta}) dy_i$ to obtain an estimate of $E_p(\pi_i \mid \mathbf{x}_i)$ as a function of the population parameters $\boldsymbol{\beta}$
3. Compute $E_s(w_i \mid \mathbf{x}_i) = 1/E_p(\pi_i \mid \mathbf{x}_i)$

Linear Regression

$$f_p(y_i \mid x_i) = N(\beta_0 + \beta_1 x_i, \sigma^2)$$

Sampling scheme:

$$E_p(\pi_i \mid y_i, x_i) = \exp[A_0 + A_1 y_i + g_A(x_i)]$$

$$\begin{aligned} f_s(y_i \mid x_i) &= \frac{\exp(A_1 y_i) f_p(y_i|x_i)}{\int \exp(A_1 y_i) f_p(y_i|x_i) dy_i} \\ &= N(\beta_0 + A_1\sigma^2 + \beta_1 x_i, \sigma^2) \end{aligned}$$

Linear Regression

1. Same slope, variance, but $\beta_0 \rightarrow \beta_0 + A_1\sigma^2$
2. Add quadratic term $A_2 y_i^2$ to model on $E_p(\pi_i \mid y_i, x_i)$:

$$f_s(y_i \mid x_i) = N\left[\frac{\beta_0 + A_1\sigma^2 + \beta_1 x_i}{1 - \sigma^2 A_2}, \frac{\sigma^2}{1 - \sigma^2 A_2}\right]$$

Both intercept β_0 and slope β_1 changed

Estimating Equations: Informative Sampling

- Let $q_i = \dfrac{w_i}{E_s(w_i \mid \mathbf{x}_i)}$,
  $u_p(y_i, \boldsymbol{\beta}) =$ estimating equation
- Solve $\mathbf{U}_s(\boldsymbol{\beta}) = \sum_{i \in s} q_i u_p(y_i, \boldsymbol{\beta}) = 0$ for $\boldsymbol{\beta}$: $\widehat{\boldsymbol{\beta}}$
- Note: We can estimate $E_s(w_i \mid \mathbf{x}_i)$ by simply regressing w_i on $\mathbf{x}_i$. SE($\widehat{\boldsymbol{\beta}}$) can be obtained by Liang-Zeger sandwich approach

Simulation Study: Logistic Regression

Population model:

$$P(y_i = 1 \mid x_i) = \frac{1}{C} \exp[\beta_{10} + \beta_{11} x_i]$$

$$P(y_i = 2 \mid x_i) = \frac{1}{C} \exp[\beta_{20} + \beta_{21} x_i]$$

$$P(y_i = 3 \mid x_i) = 1 - P(y_i = 1 \mid x_i) - P(y_i = 2 \mid x_i)$$

$$C = 1 + \exp[\beta_{10} + \beta_{11} x_i] + \exp[\beta_{20} + \beta_{21} x_i]$$

Simulation Study: Logistic Regression

(a) Informative Poisson sampling:

$\pi_i = nz_i/Z,\ Z =$ total of z_i

$z_i =$ Integer$[\frac{5}{9}y_i^2 u_i + 2x_i]$; $u_i \sim U(0,1)$

$N = 3000$, $n = 300$

x-values: $P(x = j) = 0.2, j = 1, \ldots, 5$

(b) Noninformative sampling:

$z_i =$ Integer$[5u_i + 2x_i]$

$R = 1000$ simulation runs

Means of Parameter Estimates: Logistic Regression

		β_{10}	β_{11}	β_{20}	β_{21}
	True values	1.00	0.30	0.50	0.50
Non-	q-weighted	0.99	0.31	0.51	0.51
informative	w-weighted	1.00	0.31	0.52	0.51
Sampling	ordinary MLE	0.99	0.31	0.50	0.51
	q-weighted	0.99	0.31	0.51	0.51
Informative	w-weighted	1.00	0.31	0.53	0.51
Sampling	ordinary MLE	0.29	0.43	0.06	0.59

Estimated SE's: Logistic Regression

		β_{10}	β_{11}	β_{20}	β_{21}
Non-	q-weighted	0.67	0.20	0.68	0.20
informative	w-weighted	0.71	0.22	0.72	0.22
Sampling	ordinary MLE	0.63	0.19	0.63	0.19
	q-weighted	0.63	0.19	0.63	0.19
Informative	w-weighted	0.67	0.21	0.67	0.20
Sampling	ordinary MLE	0.60	0.19	0.60	0.18

Conclusions from Simulation Study

1. q-weighted method performs well in eliminating the sampling effects under informative sampling. But ordinary MLE yields highly biased estimators.
2. w-weighting (or design-weighted) method performs well for eliminating bias, but SE's are consistently larger than those under q-weighting. Ordinary MLE in most cases gives smallest SE's.

Limitations of the Method

1. Unistage sampling but extensions to random effects (two-stage sampling) are currently under study.
2. Independence of the components $f_s(y_i \mid \mathbf{x}_i)$

Sample Survey Methods: Recent Developments and Applications

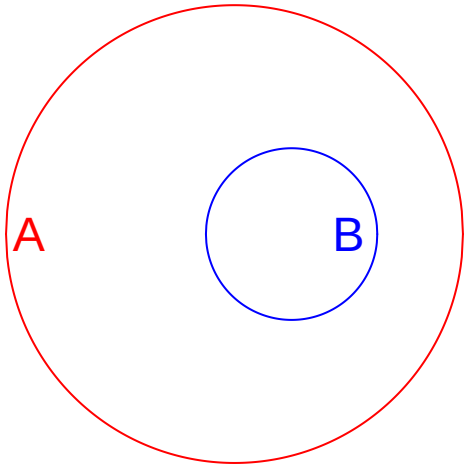
Multiple Frame Surveys

Multiple Frame Surveys

- Why use them?
- Point estimation
- Variance estimation

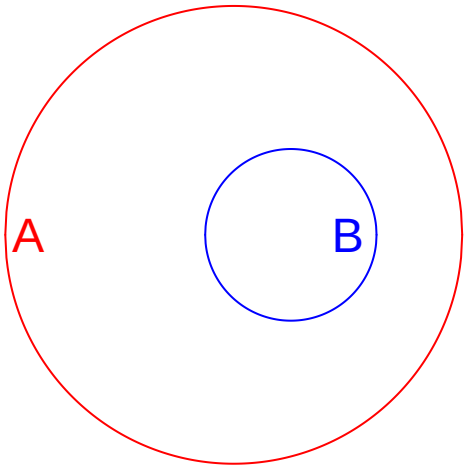
Rare Populations

- Epidemiology
- Frame A: General Population Survey
- Frame B: Survey of Patients
- Kalton & Anderson (1986)



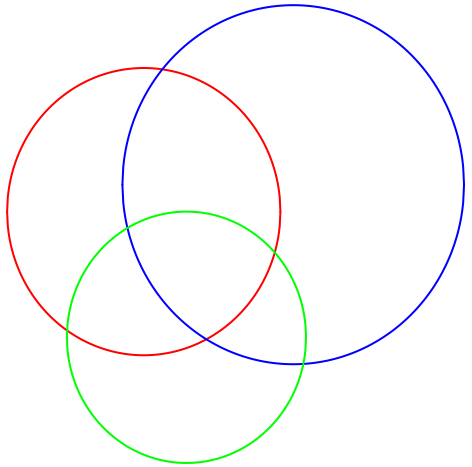
Undercoverage

- Frame A: Random Digit Dialing
- Frame B: List Frame
- Traugott et al. (1987)



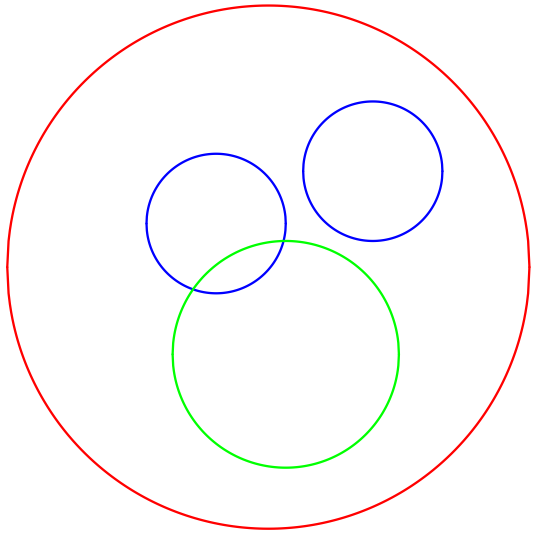
Incomplete Frames

- Sampling the Homeless
- Iachan & Dennis, 1993
- Frame A: Shelters
- Frame B:
  Soup Kitchens
- Frame C: Street



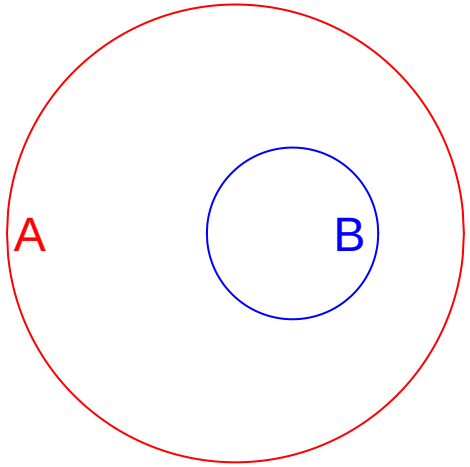
Supplement Large Survey

- Madans et al. (2001)
- Small area estimation
- NHIS (in-person)
- State and Local Area Integrated Telephone Surveys
- Regional Survey



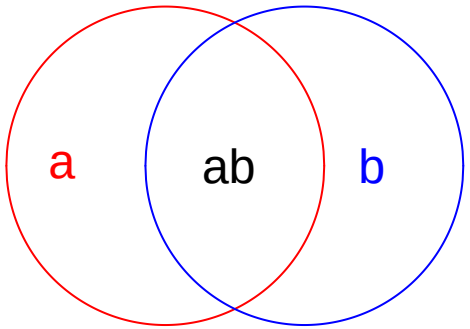
Screening Dual Frame Survey

- Remove units in
  list frame
  from area frame
  before sampling
  or estimation
- Example: SESTAT
- Sum estimates, variances



Dual Frame Assumptions

- a ∪ ab ∪ b = U
- INDEPENDENT probability samples from Frame A and Frame B

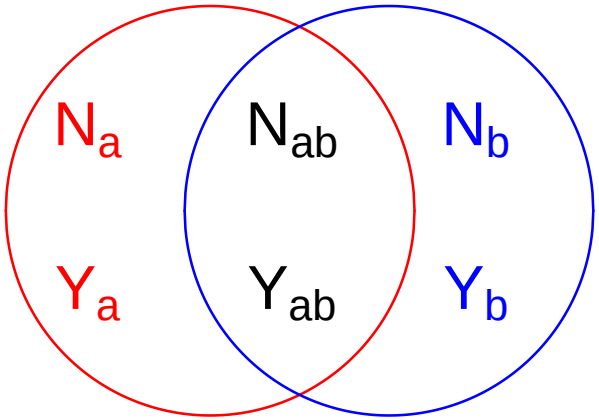


Point Estimation of Totals

$$Y = Y_a + Y_{ab} + Y_b$$

$$\hat{Y} = \hat{Y}_a + \hat{Y}_{ab} + \hat{Y}_b$$

Want to use available info to get best possible estimator of each piece



Estimation of Parts

Use sampling designs to obtain estimates

	Frame A	Frame B
Weights:	w_i^A	w_i^B
Estimators:	$\widehat{N}_{ab}$	$\widehat{N}_{ab}$
	$\widehat{Y}_{ab}$	$\widehat{Y}_{ab}$
	$\widehat{Y}_a$	$\widehat{Y}_b$

“Optimal” Estimators

- Let $\hat{Y}_{ab}(\theta) = \theta \hat{Y}_{ab} + (1-\theta)\hat{Y}_{ab}$
- Hartley (1962, 1974)
  $\hat{Y}_H = \hat{Y}_a + \hat{Y}_{ab}(\theta) + \hat{Y}_b$
- Fuller & Burmeister (1972)
  $\hat{Y}_{FB} = \hat{Y}_a + \hat{Y}_{ab}(\beta_1) + \hat{Y}_b + \beta_2(\hat{N}_{ab} - \hat{N}_{ab})$

Optimal?

- Choose θ_H, (β_1, β_2) to minimize variance
- θ_H, (β_1, β_2) depend on y's
- Different set of weights for each response
- Inconsistencies among estimates

  $\hat{Y}_1 =$ Nursing home costs, age < 85

  $\hat{Y}_2 =$ Nursing home costs, age ≥ 85

  $\hat{Y} =$ Nursing home costs, all patients

  $\hat{Y}_1 + \hat{Y}_2 \neq \hat{Y}$

Single Frame Estimators

- Combine all y_i's into one vector
  $\hat{Y}_S = \Sigma\, w_i^* y_i + \Sigma\, w_i^* y_i$
  $w_i^* = w_i$ if in domain a
  $w_i^* = 1/(1/w_i + 1/w_i)$ if in domain ab
- Must know frame-B probability for units in A-sample
- Bankier (1986), Kalton & Anderson (1986), Rao & Skinner (1996)

Pseudo-maximum Likelihood (PML)

Skinner & Rao (1996)

$$\hat{Y}_{PML} = \frac{\hat{N}_a}{\hat{N}_a}\hat{Y}_a + \frac{\hat{N}_{ab}}{\hat{N}_{ab}(\theta)}\hat{Y}_{ab}(\theta) + \frac{\hat{N}_b}{\hat{N}_b}\hat{Y}_b$$

Adjust weights by better estimate of N_{ab}

PML Implementation

Construct dual frame weights

$$\hat{Y} = \sum w_i' y_i + \sum w_i' y_i$$

$$w_i' = w_i(\hat{N}_a / \hat{N}_a)$$

$$w_i' = w_i(\hat{N}_{ab} / \hat{N}_{ab}(\theta))$$

Which Estimator

- Fuller-Burmeister optimal
- But FB requires new set of weights for each response
- PML equivalent to FB in SRS, performs well in other designs
- Single Frame easy to use, works for any number of frames

Variance Estimation

- Taylor series linearization (Skinner & Rao, 1996)
- Jackknife (Lohr & Rao, 2000)
- Bootstrap

- Assume stratified multistage sample in each frame
- Strata can differ in the frames

Variance Estimation

- $\bar{A}, \bar{B}$ vectors of means
- $\tau = g(\bar{A}, \bar{B}), \hat{\tau} = g(\hat{\bar{A}}, \hat{\bar{B}})$
- S_A, S_B covariance matrices for $(\hat{\bar{A}}, \hat{\bar{B}})$
- Taylor linearization

  $v_L(\hat{\tau}) = g_A^T S_A g_A + g_B^T S_B g_B$

  $g_A =$ vector of partial derivs

Jackknife

- $\hat{\tau} = g(\hat{\bar{A}}, \hat{\bar{B}})$
- Frame A: Eliminate psu i from stratum h

  $\hat{\tau}_{(hi)} = g(\hat{\bar{A}}_{(hi)}, \hat{\bar{B}})$

  $v_A = \sum_h [(n_h - 1)/n_h] \sum_i (\hat{\tau}_{(hi)} - \hat{\tau})^2$
- Frame B: Eliminate psu j from stratum l

  $\hat{\tau}_{(lj)} = g(\hat{\bar{A}}, \hat{\bar{B}}_{(lj)})$

  $v_B = \Sigma_l [(n_l - 1)/n_l] \Sigma_j (\hat{\tau}_{(lj)} - \hat{\tau})^2$
- $v_{JK}(\hat{\tau}) = v_A + v_B$

Bootstrap

- Two forms: separate and combined
- Define bootstrap weights separately for the two frames

Bootstrap

- Frame A: Sample $n_h - 1$ psu's with replacement from stratum h
  $w_i^* = \frac{n_h}{n_h - 1} m_{hi}^* w_i$
- Frame B: Sample $n_l - 1$ psu's with replacement from stratum l
  $w_j^* = \frac{n_l}{n_l - 1} m_{lj}^* w_j$
- $(\hat{\bar{A}}^*, \hat{\bar{B}}^*)$ use bootstrap weights

Bootstrap

- $\hat{\tau}^{*A} = g(\hat{\hat{A}}^*, \hat{\hat{B}})$
- $\hat{\tau}^{*B} = g(\hat{\hat{A}}, \hat{\hat{B}}^*)$
- $\hat{\tau}^* = g(\hat{\hat{A}}^*, \hat{\hat{B}}^*)$
- $$v_s = \frac{1}{R_A - 1} \sum_{r=1}^{R_A} (\hat{\tau}_r^{A*} - \hat{\tau})^2 + \frac{1}{R_B - 1} \sum_{r=1}^{R_B} (\hat{\tau}_r^{B*} - \hat{\tau})^2$$
- $$v_c = \frac{1}{R - 1} \sum_{r=1}^{R} (\hat{\tau}_r^* - \hat{\tau})^2$$

Modification for 2-psu design

- Hartley, FB, PML estimates depend on S_A, S_B
- not smooth function of means
- Can't estimate S_A, S_B on jackknife and bootstrap iterations in 2-psu-per-stratum designs
- Use $\widehat{\theta}$ instead of $\widehat{\theta}_{hi}$, $\widehat{\theta}_{lj}$ in JK iterations
- Use $\widehat{\theta}$ on each bootstrap iteration

Which to use?

- v_{JK}, v_L, v_s, v_c asymptotically equivalent
- Jackknife efficient if stratified multistage; intensive for stratified random sample
- Bootstrap may be more efficient if many psu's
- v_s requires more computation than v_c

Which to use?

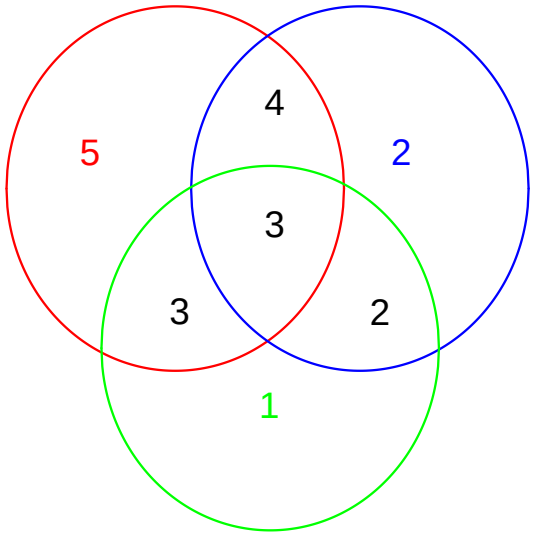
- Full jackknife, bootstrap have less bias, higher coverage probability than modified versions for 2-psu designs
- Confidence intervals: recommend t distribution with min (n_A-1, n_B-1) df
- Combined bootstrap maintains confidentiality for frames, can be used for nonsmooth functions
- Jackknife and bootstrap can be used for any number of frames

More than 2 frames

- Lohr & Rao (2003)
- Generalized Optimal Linear Estimators (Hartley, FB): Solve systems of linear equations
- PML
  MLE of domain sizes in SRS
  Extend to complex surveys using design
  Use effective sample size for $\widehat{N}_{abc}$, etc.
- Single Frame, SF with raking

Simulation study

- Pop size = 10000, each domain
- Means: in picture
- n's: 100 or 200
- SRS or Cluster sample
- 2000 reps



$\sqrt{}$MSE/100 from simulation study

n_A	Clus?	n_B	n_C		H	FB	PML	SF	Srake
100	No	100	100	N	18	18	17	17	17
				Y	83	83	79	90	79
100	Yes	100	100	N	24	24	23	27	27
				Y	143	122	110	165	121
200	No	100	100	N	15	15	15	17	15
				Y	67	68	66	67	66
200	Yes	100	100	N	21	21	21	20	22
				Y	103	95	95	111	97
100	No	200	100	N	15	15	15	17	15
				Y	78	73	70	89	71
100	Yes	200	100	N	22	22	19	29	26
				Y	140	110	101	164	111

Conclusions from simulation study, theory

- PML works well overall
  Close in performance to optimal FB; easier to compute; same weight vector
- Unraked SF performs poorly: increasing a sample size can increase the MSE
- Raked SF better, but can perform poorly in some populations

Operational Problems

(1) Larger overhead than single frame survey: 2 or more operations
(2) Determine domain membership
(3) Misclassification
(4) Some domains may have small sample sizes

Sample Survey Methods: Recent Developments and Applications

Missing Data and Imputation

Outline

1. Types of nonresponse; nonresponse bias
2. Item nonresponse: imputation
3. Multiple imputation, “proper” imputation
4. SRS: Adjusted imputed values, jackknife variance estimator
5. Stratified Multistage sampling: complete response
6. Jackknife, BRR, bootstrap

7. Adjusted imputed values
8. Weighted hot deck: jackknife, BRR
9. Fractionally weighted imputation
10. Nearest neighbour imputation
11. Reimputation
12. “Reverse” approach
13. Alternative approach to variance estimation

Nonresponse Bias in SRS

- $\bar{y}_m$ = mean of y for sample respondents s_r
- Assumption: Response probability
  $p_i = P(R_i = 1) = 1$ or 0
- $B = \text{Bias}(\bar{y}_m) = (1 - W_r)(\bar{Y}_r - \bar{Y}_{nr})$
- $W_r =$ Response rate
- $\bar{Y}_r(\bar{Y}_{nr})$ mean of resp. (nonresp.)
- B decreases as W_r increases.

Remedies

1. Reduce nonresponse rates
2. Subsample nonrespondents
3. Adjust estimators to reduce effect of nonresponse
4. Substitution

1: Cognitive survey methods, Incentives, Repeated call-backs

Types of nonresponse

- **Total (unit) nonresponse**: refusals, not-at-homes
  Remedy: Weighting adjustment within weighting classes

- **Noncoverage**: Sample provides no information about missing elements. External data sources used for weighting adjustment.

- **Item nonresponse**: sensitive item, does not know answer to item, answer to item inconsistent with other answers.

  Remedy: Imputation

	s_{rr}	s_{rm}	s_{mr}	s_{mm}
y_1	$\surd$	$\surd$	$\times$	$\times$
y_2	$\surd$	$\times$	$\surd$	$\times$

Advantages of Imputation

- Complete data file
- Different analyses consistent with each other
- Same survey weight for all items
- Auxiliary information available on all sample units may be used to get good imputed values

Commonly Used Imputation Methods

- Business surveys: mean, ratio, regression, nearest neighbour
- Socio-Economic Surveys: Random donor imputation within imputation classes, common donor
- Construction of weighting or imputation classes: Response probabilities, predicted items

Weighting Adjustment

Assumption: $p_i = p_\nu$ for ν^{th} cell (class)
MCAR within classes

N_ν unknown: Adj. factor $= \sum_{s_\nu} w_i / \sum_{s_{\nu r}} w_i = 1/\hat{p}_\nu$

Adj. wts: $\tilde{w}_{ci} = w_i/\hat{p}_\nu$ if $i \in \nu$-th cell
$\bar{y}_{cm} =$ weighted mean with weights $\tilde{w}_{ci}$
Bias $(\bar{y}_{cm}) \approx 0$ if $p_i = p_\nu$ for each ν

Too many cells: loss of precision

Construction of adjustment cells

1. Natural cells: age, region, household size
2. Response propensities;
   Use $(R_i, \mathbf{x}_i),\ i \in s$ to fit $p_i = g(\mathbf{x}_i'\boldsymbol{\beta})$ using logistic or probit regression $\Rightarrow \widehat{p}_i$, $i \in s$.
   Logit: $\log\{p_i/(1-p_i)\} = \mathbf{x}_i'\boldsymbol{\beta}$: MAR
   Define cells by estimated j/k quantiles of $\widehat{p}_i = g(\mathbf{x}_i'\widehat{\boldsymbol{\beta}}),\ j = 1, \ldots, k-1$: equal-quantile method.
3. Predicted means: Regress y_i on $\mathbf{x}_i$, $i \in s_r$ to get $\widehat{y}_i = \mathbf{x}_i'\widehat{\boldsymbol{\beta}}$, $i \in s$. Equal quantiles on $\widehat{y}_i$.

Choice of number of cells k

U.S. Consumer Expenditure Survey; $y =$ income

k	est$_k$	s.e.	s.e. (est$_k$-est$_1$)	MSE ratio
1	32,967	569	—	—
3	32,736	530	112	1.30
4	32,779	518	122	1.28
$\rightarrow$ 5	32,630	523	138	1.53
10	32,640	514	126	1.58
20	32,634	508	118	1.63

$H_0 : E(\text{est}_1 - \text{est}_5) = 0$ rejected.
$k = 5$: Larger $\widehat{p}_\nu$ has larger cell estimate

$$\$24,333 \ \ (\nu = 1); \$37,057 \ \ (\nu = 5)$$

Eltinge & Yansaneh (1997)

Inference Under Imputation

1. Frequentist approach: Repeated sampling & MCAR within imputation classes
2. Model-based approach: Repeated sampling, model generating the finite population, MAR: response indicator a_i independent of y_i given covariates used for imputation (Särndal 1992)

Imputed Estimator

$$\begin{aligned}\hat{Y}_I &= \sum_{i \in s_r} w_i y_i + \sum_{i \in s_m} w_i y_i^* \\ &= \sum_s w_i a_i y_i + \sum_s w_i (1 - a_i) y_i^* \\ &= \sum_s w_i \tilde{y}_i\end{aligned}$$

where $\tilde{y}_i = a_i y_i + (1 - a_i) y_i^*$

$s_r =$ sample of resp. for item i

$s_m =$ sample of nonresp. for item i

$y_i^* =$ imputed value

Example: Ratio Imputation

- $y_i^* = \hat{R}_r x_i$
  $\hat{R}_r = \sum_s w_i a_i y_i / \sum_s w_i a_i x_i$
- Frequentist: $\hat{Y}_I$ approx p-unbiased for Y
- Model-based with ratio model

$$E_m(y_i) = \beta x_i, \quad V_m(y_i) = \sigma^2 x_i$$

$\hat{Y}_I$ model-unbiased for Y

- Robustness:
  $\widehat{Y}_I$ is robust under ratio imputation in sense of validity under both frequentist and model-based approaches
- Variance estimation: Treating imputed values as if observed can lead to serious underestimation if response rate to item i is low

Multiple Imputation: Rubin (1996)

θ = parameter of interest.

$\widehat{\theta}_I(t)$ = imputed estimator for t-th data set;
$t = 1, \ldots, M$

M = number of imputations (≥ 2)

MI estimator: $\bar{Q}_M = \hat{\theta}_I(\cdot) = \sum_t \hat{\theta}_I(t)/M$

$$
\begin{aligned}
v_{MI}(\bar{Q}_M) &= \frac{1}{M}\sum_{t=1}^{M} v_N(\hat{\theta}_I(t)) \\
&\quad + \frac{M+1}{M}\left\{\frac{1}{M-1}\sum_t \left(\hat{\theta}_I(t) - \hat{\theta}_I(\cdot)\right)^2\right\} \\
&= \bar{U}_M + B_M
\end{aligned}
$$

v_{MI} valid under “proper” imputation in sense of approx unbiased

Note: v_{MI} is inconsistent unless $M \to \infty$

“Proper” Imputation

$(\hat{Q}, \hat{U}) =$ (est., var. est.) from complete data

Conditions: (1) $E_2(\bar{Q}_\infty) = \hat{Q}$

(2) $E_2(\bar{U}_\infty) = \hat{U}$

$\rightarrow$ (3) $E_2(B_\infty) = V_2(\bar{Q}_\infty)$

$(E_2, V_2) =$ expectation, variance over imputation, response mechanism given full sample s

Proper imputation (SRS)

1. Select a bootstrap sample, s_r^*, of size m from s_r by simple random sampling with replacement;

2. Select $n - r = m$ donors from s_r^* by SRS with replacement. This procedure is called approximate Bayesian bootstrap.

Difficulties with MI: Wang and Robins (1998)

i.i.d. case: $Y \sim N(\mu, \sigma^2)$, $\theta = (\mu, \sigma^2)'$
$\pi = P(R = 1|Y) = P(R = 1)$; $\hat{\theta}_{\mathsf{ML}}$

1. Impute from $N(\hat{\mu}_{\mathsf{ML}}, \hat{\sigma}^2_{\mathsf{ML}})$: type B
2. Proper (Rubin): type A

   (a) $\nu_j \sim \chi^2_{n_r-1}$, $\tilde{\sigma}^2_j = \hat{\sigma}^2_{\mathsf{ML}}(n_r - 1)/\nu_j$

   (b) $\tilde{\mu}_j \sim N(\hat{\mu}_{\mathsf{ML}}, n_r^{-1}\tilde{\sigma}^2_j)$

   (c) $Y_j \sim N(\tilde{\mu}_j, \tilde{\sigma}^2_j)$, $j = 1, \ldots, M$

Simulation: $n = 50,\ M = 3,\ \pi = 0.5, 0.25$

	π	cov. prob.	Med. length ratio
Type B	0.5	0.965	1.00
Type A		0.943	1.52
Type B	0.25	0.978	1.00
Type A		0.980	1.87

Single Imputation

Commonly used imputation methods are improper!

Jackknife variance estimation: Ratio imputation

- Rao & Sitter, 1995

$$y_i^* = (\bar{y}_r/\bar{x}_r)x_i; i \in s_m$$

- When respondent deleted, donor set changed
- Adjusted imputed values:

$$y_i^a(j) = y_i^* + [y_i^*(j) - y_i^*] = y_i^*(j); j \in s_r$$

$$y_i^*(j) = [\bar{y}_r(j)/\bar{x}_r(j)]x_i = \text{ reimputed value}$$

- Imputed est.: $\bar{y}_I, \bar{y}_I(j), \bar{y}_I^a(j); j \in s$

- Naive jackknife var. est.

$$v_{JN}(\bar{y}_I) = \frac{n-1}{n} \sum_{j \in s} [\bar{y}_I(j) - \bar{y}_I]^2$$

- Correct jackknife var. est.

$$v_J(\bar{y}_I) = \frac{n-1}{n} \sum_{j \in s} [\bar{y}_I^a(j) - \bar{y}_I]^2$$

- Under MCAR, $v_J(\bar{y}_I)$ is consistent (under design-based approach)

- Jackknife linearization

$$v_{JL}(\bar{y}_I) = \left(\frac{\bar{x}}{\bar{x}_r}\right)^2 \frac{A}{r} + 2\left(\frac{\bar{x}}{\bar{x}_r}\right)\frac{B}{n} + \frac{C}{n}$$

$A = s_{er}^2,\ e_i = y_i - (\bar{y}_r/\bar{x}_r)x_i$
$B = (\bar{y}_r/\bar{x}_r)s_{exr},\quad C = (\bar{y}_r/\bar{x}_r)^2 s_x^2$

- v_{JL} (or v_J) **valid** under both approaches to inf.
- $\bar{y}_I, v_{JL}(v_J)$ lead to robust inference

Stratified Multistage Sampling

- Rao & Shao (1992)
- $n_h (\geq 2)$ sample psu's from stratum h, $R = \sum n_h$
- $\hat{Y} = \sum_s w_i y_i$; $\hat{Y}(r) = \sum_s w_i(r) y_i$
- Jackknife weights: $w_i(r) = w_i b_i(r); r = 1, \ldots, R$

$$b_i(r) = \begin{cases} 0 & \text{if } i \in \text{ cluster } r \\ n_h/(n_h - 1) & \text{if } i \notin r, r \in \text{ strat } h \\ 1 & \text{if } i, r \text{ in diff strata} \end{cases}$$

$$v_J(\hat{Y}) = \sum_{r=1}^{R} c_r[\hat{Y}(r) - \hat{Y}]^2$$

$$c_r = \frac{n_h - 1}{n_h} \text{if } r \in \text{ stratum } h$$

Imputation

Weighted hot deck: select donors with probability $w_i / \sum_{i \in s_{\nu r}} w_i$ in the ν^{th} class

Adjusted imputed values

$$\begin{aligned} y_i^a(r) &= y_i^* + E_I[y_i^*(r)] - E_I[y_i^*] \\ &= y_i^* + \frac{\sum_{i \in s_{\nu r}} w_i(r) y_i}{\sum_{i \in s_{\nu r}} w_i(r)} - \frac{\sum_{i \in s_{\nu r}} w_i y_i}{\sum_{i \in s_{\nu r}} w_i} \end{aligned}$$

$$v_J(\hat{Y}_I) = \sum_{r=1}^{R} c_r[\hat{Y}_I^a(r) - \hat{Y}]^2$$

$v_J(\hat{Y}_I)$ consistent under MCAR within classes

Reimputation

- Stochastic imputation: independent reimputation leads to overestimation for jackknife
- Deterministic imputation: Reimputation same as adjusted imputation
- Bootstrap (Shao & Sitter, 1996)

  $n_h - 1$ clusters by SRS w/ replacement

  Bootstrap weights:

  $w_i(r) = w_i m_i^* n_h/(n_h - 1)$ for $i \in$ stratum h

- Apply weighted hot deck to bootstrap donors using bootstrap weights $w_i(r)$

$$v_{\mathrm{BOOT}}(\hat{Y}_I) = \frac{1}{R}\sum_{r=1}^{R}(\hat{Y}_I(r) - \hat{Y}_I)^2$$

- Difficulties: Method does not work well if n_h small

- Nearest neighbour imputation (SRS)
  $y_j^* = y_i, i \in s_r$ s.t. $\mid x_i - x_j \mid = \min_{l \in s_r} \mid x_l - x_j \mid$
  $\hat{Y}_I$ not smooth
- Partially adjusted jackknife (Chen & Shao, 2001)

$$\tilde{y}_i^a(j) = y_i^* + t_i[y_i^*(j) - y_i^*], 0 < t_i < 1,$$

t_i depends on frequencies of donors. Leads to a consistent jackknife variance estimator.

- Poststratification (Yung & Rao, 2000)
  Poststrata cut across weighting, imputation classes
  Jackknife & jackknife linearization variance estimators
- Elimination of imputation variance (SRS)
  - Use $\bar{y}_r + (y_i^* - \bar{y}_m^*)$ in place of y_i^*
  - $\bar{y}_I$ reduces to $\bar{y}_r$. Preserves distribution of y-values (Chen, Rao, Sitter, 2000)

Fractionally Weighted Imputation (Fay, 1996)

- Extend Rao & Shao (1992) to $M \geq 2$ multiple imputations
  $\{y^*_{i\ell},\ i \in s_m;\ \ \ell = 1, \ldots, M\}$
- Imputed estimator:

$$\hat{Y}_{IM} = \sum_{i \in s_r} w_i y_i + \sum_{i \in s_r} w_i \left\{ \frac{1}{M} \sum_{\ell=1}^{M} y^*_{i\ell} \right\}.$$

- Adjusted imputed values:
  $y^*_{i\ell} + E^r_*(y^*_{i\ell}) - E_*(y^*_{i\ell})$

- $E_*^r =$ Imputation expectation using r-th pseudo-replicate respondents.
- Jackknife: $E_*^r = E_*^{(gj)}$
- $\hat{Y}_{IM}^a(r) =$ Imputed estimator using adjusted imputed values and replicate weights w_i^r.
- Jackknife:

$$v_J(\hat{Y}_{IM}) = \sum_g \frac{n_g - 1}{n_g} \sum_j \left(\hat{Y}_{IM}^a(gj) - \hat{Y}_{IM}\right)^2$$

Empirical study

- Imputation classes: physician, statistician
- Domains: New York, California
- $y =$ income
- Data generated by Bernoulli (0.5) for each margin of 2×2 table

	V	$\hat{V}$	coverage %	C.I. length
FWI	0.0036	0.0039	95.4	0.120
MI	0.0037	0.0191	100.0	0.294
RS(M=1)	0.0099	0.0100	95.0	0.196

FWI = fractionally weighted imputation

MI = multiple imputation

RS = Rao–Shao method

Reverse Approach

Fay (1991); Shao & Steel (1999)

(a) Traditional: $U \rightarrow s \rightarrow (s_r, s_m)$
(b) Reverse: $U \rightarrow (U_r, U_m) \rightarrow (s_r, s_m)$

(a) $V(\bar{y}_I) = E_p V_r(\bar{y}_I) + V_p E_r(\bar{y}_I)$
(b) $V(\bar{y}_I) = E_r V_p(\bar{y}_I) + V_r E_p(\bar{y}_I)$

Reverse Approach: Ratio Imputation

- $\bar{y}_I = \bar{z}_1(1 + \bar{z}_3/\bar{z}_2) = g(\bar{\mathbf{z}})$
  $z_{1i} = a_i y_i, z_{2i} = a_i x_i, z_{3i} = (1 - a_i)x_i$
- $V_r E_p$ term negligible if n/N small
- est var $(\bar{y}_I) \approx$ est $V_p[g(\bar{\mathbf{z}})]$
- Use jackknife or Taylor linearization
- Method esp. useful for composite imputation

Example: U.S. Transportation Annual Survey

y = current year revenue; $\tilde{y}$ = prior year revenue
x = current year payroll; $\tilde{x}$ = prior year payroll
z = current year adm. annual payroll

1. y_i observed, no imputation
2. $x_i, \tilde{x}_i, \tilde{y}_i$ observed: $y_i^* = x_i \tilde{y}_i / \tilde{x}_i$.
3. x_i observed: ratio imputation: $\hat{R}_{y|x} x_i$.
4. x_i missing: $\hat{R}_{y|x}\ \hat{R}_{x|z}\ z_i$.

1995: naive $\hat{V}$/correct $\hat{V}$ varied from 0.38 to 0.86.

Four indicator variables $a_{1i}, \ldots, a_{4i}$ for cases (1)–(4), a_{5i} for $\hat{R}_{y|x}$ and a_{6i} for $\hat{R}_{x|z}$.

Estimation of Relationships

- $B = \Sigma_U x_i y_i / \Sigma_U x_i^2$
  x_i not missing, y_i missing
- Ex: x_i domain indicator, domain mean $\bar{Y}_d$
- Imputed estimator of B:
  $\hat{B}_I = [\Sigma_s w_i x_i^2]^{-1}[\Sigma_{s_r} w_i x_i y_i + \Sigma_{s_m} w_i x_i y_i^*]$
- Ratio imputation: $y_i^* = (\bar{y}_r / \bar{z}_r) z_i$
- Under frequentist approach (MCAR):
  Bias $\hat{B}_I \approx (1-p)[B_z(\bar{Y}/\bar{Z}) - B]$
  $B_z = [\Sigma_U x_i^2]^{-1}[\Sigma_U x_i z_i]$, $p =$ prob of resp.

Domain Mean

- Bias = 0 if domain ratio $R_d = \bar{Y}_d / \bar{Z}_d$ equals overall ratio $R = \bar{Y} / \bar{Z}$
- Under ratio model:

$$E_m(y_i) = \gamma z_i, \; V_m(y_i) = \tau^2 z_i$$

$\hat{B}_I$ approx unbiased for B

Bias-adjusted Estimator

- Haziza and Rao (2000)
- $\hat{B}_I^a = \hat{p}^{-1}\hat{B}_I + (1-\hat{p}^{-1})\hat{B}_z(\bar{y}_I/\bar{z})$
- $\hat{B}_I^a$ robust: validity under design-based, model-based approaches
- Using reverse approach, consistent variance estimators obtained under both approaches to inference

Extension

$$\mathbf{B} = (\sum_{U} \mathbf{x}_i \mathbf{x}_i^T)^{-1} \sum_{U} \mathbf{x}_i y_i$$

Example:

$\mathbf{x}_i$ indicators of 2 domains $d = 1, 2$

$\mathbf{B}$ = vector of domain means $(\bar{Y}_1, \bar{Y}_2)$

$\theta = \bar{Y}_1 - \bar{Y}_2$ = diff of domain means

- Example: Mean imputation, $z_i = 1$

$$\bar{y}_{1I}^{a} - \bar{y}_{2I}^{a} = \hat{p}^{-1}(\bar{y}_{1I} - \bar{y}_{2I})$$

- Additivity: $\hat{Y}_{1I}^{a} + \hat{Y}_{2I}^{a} = \hat{Y}_{I}$

Domain 1: Female; Domain 2: Male

$y =$ height, $p = 0.8$, CI on $\bar{Y}_1 - \bar{Y}_2 = \theta$

	Coverage Prob (%)	
n	Unadjusted	Adjusted
50	76.3	94.2
80	69.9	94.8
120	59.3	94.9

nominal 95%; unadjusted $\bar{y}_{1I} - \bar{y}_{2I}$

Joint Imputation

- Shao & Wang (2002), Srivastava & Carter (1986)
- Model-based: $y_i = \beta z_i + z_i^{1/2}\varepsilon_i$

  $x_i = \gamma z_i + z_i^{1/2}\eta_i$
- (a_i, b_i): response indicators for (y_i, x_i)
- $\begin{pmatrix} \varepsilon_i \\ \eta_i \end{pmatrix} \underset{\sim}{\text{iid}} \begin{pmatrix} \sigma_\varepsilon^2 & \sigma_{\varepsilon\eta} \\ \sigma_{\varepsilon\eta} & \sigma_\eta^2 \end{pmatrix}$

- Imputation depends on values of (a_i, b_i):
- $(a_i, b_i) = 0$:

$$\begin{pmatrix} \varepsilon_i^* \\ \eta_i^* \end{pmatrix} \overset{\text{iid}}{\sim} \left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \hat{\sigma}_\varepsilon^2 & \hat{\sigma}_{\varepsilon\eta} \\ \hat{\sigma}_{\varepsilon\eta} & \hat{\sigma}_\eta^2 \end{pmatrix} \right]$$

- $y_i^* = \hat{\beta} z_i + z_i^{1/2} \varepsilon_i^*; x_i = \hat{\gamma} z_i + z_i^{1/2} \eta_i^*$

Variance Estimation with Imputed Data

National Immunization Provider Record Check Survey

- Stratified 3-stage sample of households (Nixon et al. 2000)
- Goal: Obtain variance ests. for estimated % of children up-to-date on vaccinations
- HH vaccination data reconciled to provider data to determine “best values.”
- Provider data missing for 378 children (31%) out of 1,230 in the sample

Nonresponse Adjustment: NIPRC

1. Weighting classes (20); weighting adjustment within classes
2. Common donor hot deck within imputation classes (20)

Variance estimation for (1):

Jackknife with replicate weights adjusted for nonresponse within classes (JK-W)

Variance estimation for (2):

(a) Rao-Shao jackknife with adjusted imputed values (JK-RS)

(b) Shao BRR with adjusted imputed values (BRR-S)

(c) Impute within each BRR replicate using replicate donors (BRR-R)

(d) Naive: Treat imputed as observed (Naive)

Comparison: NIPRC

	Method of Variance Estimation				
Vaccination	Naive	JK-RS	BRR-S	BRR-R	JK-W
DTP	1.30	2.22	2.66	2.59	1.93
Polio	1.04	1.93	2.43	1.72	1.66
MMR	0.66	1.69	2.19	2.00	1.64
Hib	0.96	1.74	1.88	1.90	1.51
Hep B	1.61	3.03	3.03	3.08	2.53
431	1.21	2.19	2.62	2.70	1.88
4313	1.21	2.28	2.69	2.75	1.93
43133	1.37	2.66	2.66	3.09	2.40

Conclusions: NIPRC

1. Treating imputed best values as actual values $\Rightarrow$ variance underestimated by factor of 2, or SE's underestimated by 40% or more
2. "Correct" variance ests for imputed ests: 3% to 20% larger than those for weighted ests. This is due to "imputation variance" resulting from random selection of donors within imputation classes with hot deck imputation.
3. JK-RS, BRR-S, BRR-R ests broadly similar

1993-94 Schools And Staffing Survey

Stratified multistage, hot deck within imputation classes (Zhang et al., 2000)

(i) 48 replicate weights using Rao-Wu bootstrap

(ii) Reimpute in each bootstrap sample using donors in the bootstrap sample within imputation classes (Shao & Sitter, 1996)

Naive variance est compared to Shao-Sitter bootstrap est that takes account of imputation variance.

Conclusions: SASS

1. For the total of a continuous variable, imputation does not inflate the variance very much. For variable "number of separate classes taught," SE increased only by 7% even though imputation rate is as high as 27%. For the average estimator, SE increased by 41%.
2. For ratio estimates of continuous variables, increase in SE is small.
3. For the total estimate and percentage estimate of categorical variables, inflation in SE larger than in case of continuous variables.

Implementing Rao-Shao-type Variance Estimation with Replicate Weights, Imputation

Stratified Multistage Sampling (Cohen, 2002)

$(h^0 i^0 j^0)$: unit not responding to item y

$(h'i'j')$: unit responding to y within same class ν

Assumption:

$$E_\nu[y^*_{h^0 i^0 j^0}] = \sum_{(h'i'j') \in A_\nu} a(h'i'j', h^0 i^0 j^0) y_{h'i'j'} \leftarrow$$

$$E^{(r)}_\nu[y^*_{h^0 i^0 j^0}] = \sum_{(h'i'j') \in A_\nu} a^{(r)}(h'i'j', h^0 i^0 j^0) y_{h'i'j'}$$

$a^{(r)}(h'i'j', h^0 i^0 j^0) = 0$ if $(h'i'j')$ not in replicate r

$A_\nu =$ set of respondents to item y in imputation class ν

A record in replicate data file:

$$\text{ID} \quad \text{IC} \quad w_{hij} \quad w_{hij}^{(1)} \quad \ldots \quad w_{hij}^{(R)} \quad \tilde{y}_{hij} \quad \text{IF}_y$$

ID = unit id, IC = imputation class id, IF = imputation flag
$w_{hij} =$ full sample weight, $w_{hij}^{(r)} = r^{th}$ replicate weight
$\tilde{y}_{hij} = y_{hij}$ if observed, y_{hij}^{*} if not observed

Extra records: For each $(h^0 i^0 j^0)$ and $(h'i'j')$ in class ν, create the record

$$\text{ID} \quad \text{IC} \quad 0 \quad w_{h^0 i^0 j^0, h'i'j'}^{(1)} \quad \ldots \quad w_{h^0 i^0 j^0, h'i'j'}^{(R)} \quad \tilde{y}_{h'i'j'} \quad \text{IF}_y$$

Note: full sample weight = 0 for extra records

$$w^{(r)}_{h^0 i^0 j^0, h'i'j'} = [a^{(r)}(h'i'j', h^0 i^0 j^0) - a(h'i'j', h^0 i^0 j^0)] w^{(r)}_{h^0 i^0 j^0}$$

Weighted hot deck:

$$a(h'i'j', h^0 i^0 j^0) = w_{h'i'j'} / \sum_{(hij) \in A} w_{hij}$$

$$a^{(r)}(h'i'j', h^0 i^0 j^0) = w^{(r)}_{h'i'j'} / \sum_{(hij) \in A} w^{(r)}_{hij}$$

Numerical Illustration:

$n = 6$, $r = 3$ respondents ($\text{IF}_y = 0$),

$r = 3$ nonrespondents ($\text{IF}_y = 1$)

ID	IC	w_{hij}	$w_{hij}^{(1)}$	$\ldots$	$w_{hij}^{(R)}$	$\tilde{y}_{hij}$	IF$_y$
001	1	10.1	20.2000		0.0000	5.4	1
002	1	20.3	40.6000		0.0000	5.1	0
003	1	18.4	36.8000		0.0000	5.2	0
004	1	11.1	0.0000		22.2000	5.1	1
005	1	16.3	0.0000		32.6000	5.1	1
006	1	15.4	0.0000		30.8000	5.4	0
001	1	0	3.0162		0.0000	5.1	2
001	1	0	2.7339		0.0000	5.2	2
001	1	0	-5.7501		0.0000	5.4	2
004	1	0	0.0000		-8.3301	5.1	2
004	1	0	0.0000		-7.5505	5.2	2
004	1	0	0.0000		15.8806	5.4	2
005	1	0	0.0000		-12.2325	5.1	2
005	1	0	0.0000		-11.0876	5.2	2
005	1	0	0.0000		23.3201	5.4	2

$\sum w_{hij}\tilde{y}_{hij} = 476.650, \sum w_{hij}^{(1)}\tilde{y}_{hij} = 506.048,$
$\sum w_{hij}^{(R)}\tilde{y}_{hij} = 455.696$
Sums are over **all** records (6 original + 9 extra)

Above values agree with estimates using adjusted imputed values $\tilde{y}_{hij}^{(r)}$, i.e.

$\sum w_{hij}\tilde{y}_{hij} = 476.650, \sum w_{hij}^{(1)}\tilde{y}_{hij}^{(1)} = 506.048,$
$\sum w_{hij}^{(R)}\tilde{y}_{hij}^{(R)} = 455.696$
Sums are over 6 original records

- Disadvantage of the method: Large number of extra records
- Advantage: Standard replicate variance estimation software can be used on the expanded data set without modification

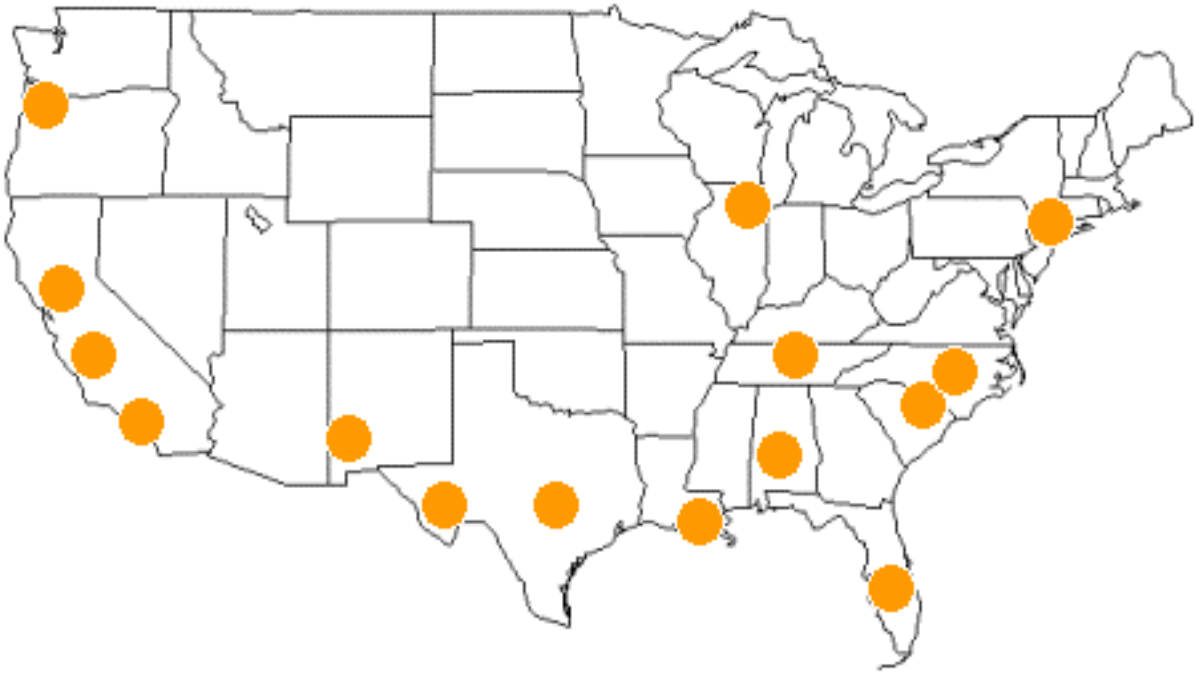
Sample Survey Methods: Recent Developments and Applications

Small Area Estimation

Definition of “small area”

A subpopulation (or domain) is a small area if the number of area specific sample observations is small.

NHANES: possible allocation of psu's (U-shaped)



Terminology:

$Y_i = i^{th}$ area total; $\bar{Y}_i = i^{th}$ area mean

$\widehat{Y}_i =$ direct estimator of Y_i

Indirect estimator: borrows strength from sample observations of related areas to increase effective sample size

Traditional and model-based indirect estimators

Traditional indirect estimators

(i) Synthetic

Implicit "linking" model: $\bar{Y}_i = \bar{Y} = Y/N$

$\hat{Y}_i(\text{syn}) = N_i(\hat{Y}/\hat{N})$, noting $Y_i = N_i\bar{Y}_i$

(ii) Composite

$\hat{Y}_i(\text{comp}) = \phi_i\hat{Y}_i + (1-\phi_i)\hat{Y}_i(\text{syn})$

Sample-size dependent (SSD) estimator:

$$\phi_i = \begin{cases} 1 & \text{if } \hat{N}_i > \delta N_i \\ \hat{N}_i/(\delta N_i) & \text{otherwise} \end{cases}$$

Basic area level model

- Explicit linking model relating $\theta_i = g(Y_i)$ to area level covariates z_i:

$$\theta_i = z_i^T \beta + v_i; \quad v_i \overset{\text{iid}}{\sim} N(0, \sigma_v^2)$$

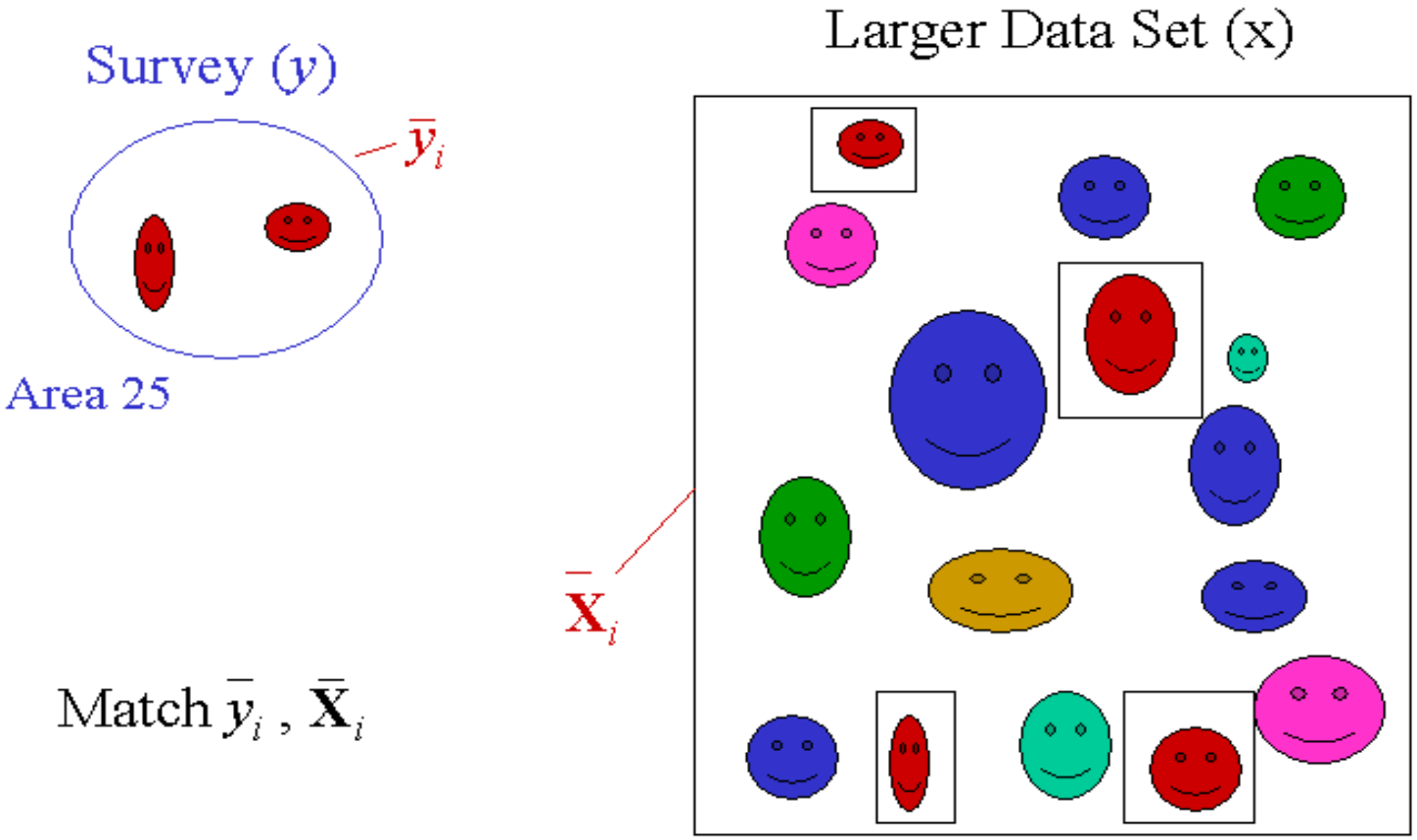
- Matching sampling model:

$$\widehat{\theta}_i = g(\widehat{Y}_i) = \theta_i + e_i; \quad e_i \mid \theta_i \overset{\text{ind}}{\sim} N(0, \psi_i)$$

- Combined model:

$$\widehat{\theta}_i = z_i^T \beta + v_i + e_i; \quad i = 1, \ldots, m$$

Area-level Model



Match $\bar{y}_i$, $\bar{\mathbf{X}}_i$

- Limitations of basic area level model:
  (i) sampling variances, ψ_i known
  (ii) $E(e_i \mid \theta_i) = 0$
- More realistic sampling model:

$$\hat{Y}_i = Y_i + f_i; \quad E(f_i \mid Y_i) = 0$$

$$V(f_i \mid Y_i) = \sigma_i^2 = Y_i^2 c_i^2; \quad c_i = CV(\hat{Y}_i)$$

- Extensions:
  Correlated sampling errors
  Spatial correlation of model errors v_i
  Time series and cross-sectional data

Disease mapping: Counts

(i) $y_i \overset{\text{ind}}{\sim}$ Poisson$(n_i \lambda_i)$
(ii) $\theta_i = \log(\lambda_i) = z_i^T \beta + v_i; v_i \overset{\text{iid}}{\sim} N(0, \sigma_v^2)$

Age-specific: $\theta_{ij} = \log(\lambda_{ij})$
$\lambda_i =$ true rate; $n_i =$ number exposed

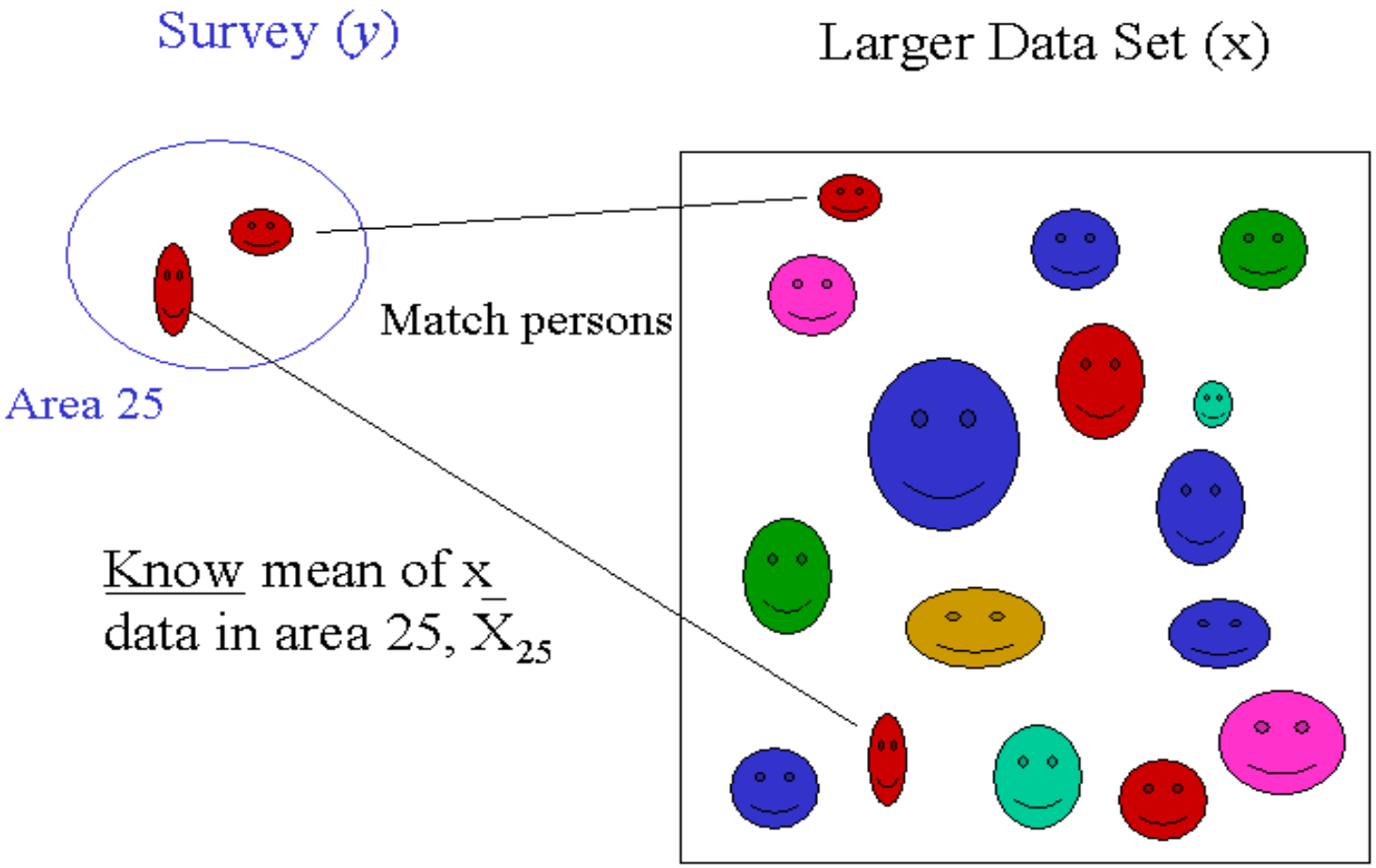
Extension: spatial dependence of θ_i's
CAR spatial model: each θ_i related to a set of neighbourhood area of i.

Basic unit level model

y_{ij} related to x_{ij}:

$$y_{ij} = x_{ij}^T \beta + v_i + e_{ij}$$

$$v_i \overset{\text{iid}}{\sim} N(0, \sigma_v^2), \; e_{ij} \overset{\text{iid}}{\sim} N(0, \sigma_e^2)$$


Survey (y)
Larger Data Set (x)
Match persons
Area 25
Know mean of x
data in area 25, $\bar{X}_{25}$

Extensions: Binary response

(i) $y_{ij} \overset{\text{ind}}{\sim} \text{Bernoulli}(p_{ij})$

(ii) $\log\{p_{ij}/(1-p_{ij})\} = x_{ij}^T\beta + v_i$

$v_i \overset{\text{iid}}{\sim} N(0, \sigma_v^2)$

Malec et al. (1999): random slope β_i

EB estimation: basic area level model

- Best estimator of θ_i: $E(\theta_i \mid \hat{\theta}_i, \beta, \sigma_v^2)$
- Replace β, σ_v^2 by $\hat{\beta}$, $\hat{\sigma}_v^2$ obtained from marginal distribution: $\hat{\theta}_i \overset{\text{ind}}{\sim} N(z_i^T\beta, \sigma_v^2 + \psi_i)$
- EB estimator: $\hat{\theta}_i^{EB} = \hat{\gamma}_i\hat{\theta}_i + (1 - \hat{\gamma}_i)z_i^T\hat{\beta}$

$$\hat{\gamma}_i = \hat{\sigma}_v^2/(\hat{\sigma}_v^2 + \psi_i)$$

- $\hat{\theta}_i^{EB}$ unbiased for θ_i: $E(\hat{\theta}_i^{EB} - \theta_i) = 0$
- $\hat{\theta}_i^{EB}$ design consistent as $\psi_i \to 0$

But $g^{-1}(\hat{\theta}_i^{EB})$ biased for Y_i.

EB estimator of Y_i:

$$\hat{Y}_i^{EB} = E[g^{-1}(\theta_i) \mid \hat{\theta}_i, \beta, \sigma_v^2]_{\beta=\hat{\beta},\sigma_v^2=\hat{\sigma}_v^2}$$

$$E(\hat{Y}_i^{EB} - Y_i) \approx 0$$

Estimation of σ_v^2: Fay-Herriot (1979)

Solve for σ_v^2 iteratively:

$$a(\sigma_v^2) = \sum_{i=1}^{m} [\widehat{\theta}_i - z_i^T \tilde{\beta}(\sigma_v^2)]^2/(\sigma_v^2 + \psi_i) = m - p$$

m: number of small areas
p: dimension of z_i
$\tilde{\beta}(\sigma_v^2) =$ WLS estimator of β

Normality not needed using FH method:

$\widehat{\theta}_i^{EB} =$ EBLUP estimator of θ_i

Other methods: ML, REML, Prasad and Rao (PR)

Mean squared error (MSE)

$$\text{MSE}\,(\hat{\theta}_i^{EB}) = E(\hat{\theta}_i^{EB} - \theta_i)^2$$
$$\approx g_{1i}(\sigma_v^2) + g_{2i}(\sigma_v^2) + g_{3i}(\sigma_v^2)$$

$g_{1i}(\sigma_v^2) = \gamma_i \psi_i$: leading term

$g_{2i}(\sigma_v^2)$: due to estimation of β

$g_{3i}(\sigma_v^2)$: due to estimation of σ_v^2

$$g_{3i}(\sigma_v^2) = \psi_i^2(\sigma_v^2 + \psi_i^2)^{-3} h(\sigma_v^2)$$

$h(\sigma_v^2) =$ asymptotic variance of $\widehat{\sigma}_v^2$

$$h_{\mathsf{ML}}(\sigma_v^2) \leq h_{\mathsf{REML}}(\sigma_v^2) \leq h_{\mathsf{FH}}(\sigma_v^2) \leq h_{\mathsf{PR}}(\sigma_v^2)$$

MSE estimation

$$\text{mse}\,(\hat{\theta}_i^{EB}) = g_{1i}(\hat{\sigma}_v^2) + g_{2i}(\hat{\sigma}_v^2) + 2g_{3i}(\hat{\sigma}_v^2)$$

Valid for REML and PR

Extra term $g_{0i}(\hat{\sigma}_v^2)$ added for ML and FH:
$g_{0i}(\hat{\sigma}_v^2)$ positive for ML, negative for FH

Robustness of mse ($\widehat{\theta}_i^{EB}$) under PR:

Approximate unbiasedness valid under nonnormality of v_i, but e_i normal. (Lahiri and Rao, 1995)

Criticism of mse ($\widehat{\theta}_i^{EB}$):

Not area specific: does not explicitly depend on θ_i

Area specific mse ($\widehat{\theta}_i^{EB}$): Rao (2001)

- Replace $g_{3i}(\widehat{\sigma}_v^2)$ by

  $$\tilde{g}_{3i}(\widehat{\sigma}_v^2, \widehat{\theta}_i) = [\psi_i^2/(\sigma_v^2+\psi_i)^4](\widehat{\theta}_i - z_i^T\widehat{\beta}_i)^2 h(\widehat{\sigma}_v^2)$$

- Simulation study: Datta, Rao & Smith (2004)
  FH-based $\widehat{\theta}_i^{EB}$ maintains good efficiency
  FH outperformed for large ψ_i-variation
  Other methods lead to considerable overes-
  timation of MSE for areas with small ψ_i.
  FH also good in term of CV of MSE estimator

Estimation of conditional MSE

- More appealing to estimate sampling MSE given θ_i's: MSE$_p(\hat{\theta}_i^{EB}) = E_p(\hat{\theta}_i^{EB} - \theta_i)^2$
- Rivest and Belmonte (2000) used PR estimator of σ_v^2 to derive design-unbiased estimator of MSE$_p(\hat{\theta}_i^{EB})$.
- Note: Leading term of this MSE estimator depends on $\hat{\theta}_i$ unlike $g_{1i}(\hat{\sigma}_v^2)$, but it is highly unstable relative to mse $(\hat{\theta}_i^{EB})$ unless $1 - \hat{\gamma}_i$ small.

Unknown sampling variances ψ_i

- Replace ψ_i by an estimator $\widehat{\psi}_i$
- Special case:

$$y_{ij} \overset{\text{iid}}{\sim} N(\theta_i, \sigma_v^2), j = 1, \ldots, n_i (\geq 2)$$

$$\widehat{\theta}_i = \bar{y}_i, \widehat{\psi}_i = s_i^2 / n_i$$

$\bar{y}_i$ independent of $\widehat{\psi}_i$ and

$$\widehat{\psi}_i \approx N(\psi_i, \delta_i = 2\psi_i^2/(n_i - 1))$$

- Wang & Fuller (2003); Rivest & Vandal (2003)
- Additional contribution to MSE estimator: $2\widehat{\delta}_i\widehat{\sigma}_v^4/(\widehat{\psi}_i + \widehat{\sigma}_v^2)^2$
- If n_i small, additional contribution may be large
- If $\widehat{\psi}_i$ is a smoothed estimator of ψ_i based on GVF models, then contribution from the additional term same order, $O(m^{-1})$, as $g_{3i}(\widehat{\sigma}_v^2)$.

Jackknife estimation of MSE

- Binary data
- $y_{ij} \mid p_i \overset{\text{iid}}{\sim}$ Bernoulli(p_i)
- Beta-binomial: $p_i \overset{\text{iid}}{\sim}$ Beta(α, β)
  $\phi = (\alpha, \beta)$
- Logit-normal:
  $$\log\left\{\frac{p_i}{1-p_i}\right\} = z_i^T\beta + v_i,\ v_i \overset{\text{iid}}{\sim} N(0, \sigma_v^2)$$
  $\phi = (\beta, \sigma_v^2)$

MSE: binary data

- Let $y_{i\cdot} = \Sigma_j \, y_{ij}$
- $\widehat{p}_i^B = E(p_i \mid y_{i\cdot}, \phi) \equiv k(y_{i\cdot}, \phi)$
- EB estimator of p_i: $\widehat{p}_i^{EB} = k(y_{i\cdot}, \widehat{\phi})$
- Let $\tilde{g}_{1i}(y_{i\cdot}, \phi) = V(p_i \mid y_{i\cdot}, \phi) =$ MSE $(\widehat{p}_i^B)$
- $\tilde{g}_{1i}(y_{i\cdot}, \widehat{\phi})$ underestimates MSE$(\widehat{p}_i^{EB})$

$$\text{MSE}\,(\widehat{p}_i^{EB}) = E[(\widehat{p}_i^{B} - p_i)^2] + E[(\widehat{p}_i^{EB} - \widehat{p}_i^{B})^2]$$
$$= M_{1i} + M_{2i}$$

$\widehat{p}_i^{EB} = k(y_{i\cdot}, \widehat{\phi})$
$\tilde{g}_{1i}(y_{i\cdot}, \phi) = V(p_i \mid y_{i\cdot}, \phi)$
$\widehat{\phi}(l)$: deleted l^{th} area estimators

$$\widehat{M}_{2i} = \frac{m-1}{m} \sum_{l=1}^{m} [\widehat{p}_i^{EB}(l) - \widehat{p}_i^{EB}]^2$$

$\widehat{p}_i^{EB}(l) = k(y_{i\cdot}, \widehat{\phi}(l))$

Area-specific jackknife MSE estimator

Rao 2003, Lohr & Rao (2004)

$$\tilde{M}_{1i}(y_{i\cdot}) = \tilde{g}_{1i}(y_{i\cdot}, \hat{\phi}) - \frac{m-1}{m} \sum_{i=1}^{m} [\tilde{g}_{1i}(y_{i\cdot}, \hat{\phi}(l)) - \tilde{g}_{1i}(y_{i\cdot}, \hat{\phi})]$$

$$\text{mse}^{*}_{J}(\hat{p}_i^{EB}) = \tilde{M}_{1i}(y_{i\cdot}) + \hat{M}_{2i}$$

Linear mixed model: $\tilde{M}_{1i}(y_{i\cdot}) = g_{1i}(\hat{\sigma}_v^2)$

Non-area-specific jackknife MSE estimator

Jiang, Lahiri, Wan (2002)

$$\hat{M}_{1i} = \tilde{h}_{1i}(\hat{\phi}) - \frac{m-1}{m} \sum_{i=1}^{m} [\tilde{h}_{1i}(\hat{\phi}(l)) - \tilde{h}_{1i}(\hat{\phi})]$$

$$\tilde{h}_{1i}(\phi) = E\{V[p_i \mid y_{i\cdot}, \phi]\}$$

Same as area-specific jackknife in linear model

Simulation Study

- Beta-Binomial model
- # small areas: m =10, 30, 60
- $\alpha, \beta \in \{0.1, 1, 10\}; n_i =$ 1, 2, 3, 4, 5
- 1000 iterations
- Relative Bias = RB = 100 (mse MSE)/MSE
  CV = s(mse) / MSE

$$m = 30, \alpha = \beta = 1$$

	Naive	JLW	AS
Uncond AARB	31.8 (neg)	3.8	2.6
Uncond. ACV	0.30	0.51	0.53
Cond ARB	32.4 (neg)	22.3^*	7.5
Cond ACV	0.24	0.53	0.57

*JLW conditional bias > 0 for big and small $y_{i\cdot}$,
< 0 for others

Relative Performance

- Area-specific jackknife has small bias unconditional and conditional
- JLW jackknife has somewhat smaller CV when large number of areas
- Area-specific jackknife much easier to compute in most models
  Does not need $E\{V[p_i|y_{i\cdot}, \phi]\}$, unlike JLW
  Fewer numerical errors in general cases

Basic unit level model: pseudo-EB

$y_{ij} = x_{ij}^T \beta + v_i + e_{ij}$
$w_{ij} =$ survey weight; $\mu_i = \bar{X}_i^T \beta + v_i$,

Battese, Harter & Fuller (1988): model-based

$$\hat{\mu}_i^{EB} = \hat{\gamma}_i [\bar{y}_i + (\bar{X}_i - \bar{x}_i)^T \hat{\beta}] + (1 - \hat{\gamma}_i) \bar{X}_i^T \hat{\beta}$$

$$\hat{\gamma}_i = \hat{\sigma}_v^2 / (\hat{\sigma}_v^2 + \hat{\sigma}_e^2 / n_i)$$

$\hat{\mu}_i^{EB}$ not design consistent (as $n_i \to \infty$) unless
$w_{ij} = w_i$: self-weighting within areas

Pseudo-EB: You and Rao (2002a)

Let $\bar{y}_{iw} = \sum_{j=1}^{n_i} \tilde{w}_{ij} y_{ij}$; $\tilde{w}_{ij} = w_{ij} / \sum_{j=1}^{n_i} w_{ij}$

$$\bar{x}_{iw} = \sum_{j=1}^{n_i} \tilde{w}_{ij} y_{ij}; \delta_{iw} = \sum_{j=1}^{n_i} \tilde{w}_{ij}^2$$

$$\gamma_{iw} = \sigma_v^2 / (\sigma_v^2 + \sigma_e^2 \delta_{iw}); \hat{\beta}_w = \tilde{\beta}_w(\hat{\sigma}_e^2, \hat{\sigma}_v^2)$$

where $\tilde{\beta}_w(\sigma_e^2, \sigma_v^2)$ solution of

$$\sum_{i=1}^{m} \sum_{j=1}^{n_i} \tilde{w}_{ij} x_{ij} [y_{ij} - x'_{ij}\beta - \gamma_{iw}(\bar{y}_{iw} - \bar{x}'_{iw}\beta)] = 0$$

pseudo-EB (EBLUP) estimator of μ_i:

$$\hat{\mu}_{iw}^{\text{PEB}} = \hat{\gamma}_{iw}[\bar{y}_{iw} + (\bar{X}_i - \bar{x}_{iw})^T \hat{\beta}_w] + (1 - \hat{\gamma}_{iw})\bar{X}_i^T \hat{\beta}_w$$

Advantages of pseudo-EB

(i) Design-consistent as $n_i \to \infty$

(ii) automatic benchmarking:

$$\begin{aligned} \Sigma_{i=1}^{m} N_i \hat{\mu}_{iw}^{\text{PEB}} &= \hat{Y}_w + (X - \hat{X}_w)^T \hat{\beta}_w \\ &= \text{ direct regression estimator} \end{aligned}$$

$$\hat{Y}_w = \sum_i \sum_j w_{ij} y_{ij}; \hat{X}_w = \sum_i \sum_j w_{ij} x_{ij}$$

Limitation:

MSE estimator not robust to deviation from normality of v_i and e_{ij}

Semi-nonparametric density of v_i: Zhang and Dravidian (2001); Finite mixture of normals for v_i: Maiti (2001)

HB approach: basic area level model

- Let $\widehat{\boldsymbol{\theta}} =$ vector of direct estimators $\widehat{\theta}_i$
- $\delta = (\beta^T, \sigma_v^2)$: prior density $f(\delta)$
- HB estimator: $\widehat{\theta}_i^{HB} = E(\theta_i \mid \widehat{\boldsymbol{\theta}})$
- Posterior variance: $V(\theta_i \mid \widehat{\boldsymbol{\theta}})$
- Calculation of $\widehat{\theta}_i^{HB}$, $V(\theta_i \mid \widehat{\boldsymbol{\theta}})$ involves integration w.r.t. posterior density $f(\beta, \sigma_v^2 \mid \widehat{\boldsymbol{\theta}})$
- MCMC generates J samples from $f(\boldsymbol{\theta} \mid \widehat{\boldsymbol{\theta}})$:

$$\{\theta_1^{(j)}, \ldots, \theta_m^{(j)}; j = 1, \ldots, J\}; Y_i^{(j)} = g^{-1}(\theta_i^{(j)})$$

HB estimator of Y_i

$\hat{Y}_i^{HB} = \frac{1}{J} \Sigma_j Y_i^{(j)}$

$V(Y_i \mid \hat{\boldsymbol{\theta}}) = \frac{1}{J} \Sigma_j (Y_i^{(j)} - Y_i^{(\cdot)})^2$

Choice of prior $f(\delta) = f(\beta) f(\sigma_v^2)$

"Matching" prior: $E[V(\theta_i \mid \hat{\boldsymbol{\theta}})] \approx \text{MSE}\,(\hat{\theta}_i^{HB})$

$$f_i(\sigma_v^2) \propto (\sigma_v^2 + \psi_i)^2 \sum_{l=1}^{m} (\sigma_l^2 + \psi_l)^{-2}$$

$$f(\beta) \propto 1$$

See Datta, Rao, and Smith (2004)

If $\psi_i = \psi$ we get $f_i(\sigma_v^2) = f(\sigma_v^2) \propto 1$

US Poverty Counts State model

EB: $\widehat{\sigma}_v^2 = 0 \Rightarrow \widehat{\gamma}_i = 0$

HB: $f(\beta) \propto 1, f(\sigma_v^2) \propto 1$ (Bell, 1999)

$\Rightarrow$ Positive weights $\widehat{\gamma}_i^{HB}$, but

$\max(\psi_i)/\min(\psi_i) \approx 20$

HB inference may not be well calibrated

Canadian census: You and Rao (2002b)

$C_i =$ census count; $Y_i =$ number missing

$\widehat{Y}_i =$ post-census survey estimator

Unmatched sampling and linking models:

$\widehat{Y}_i \mid Y_i \overset{\text{ind}}{\sim} N(Y_i, \sigma_i^2)$; σ_i^2 known

$$\theta = \log\{Y_i/(Y_i+C_i)\} = \beta_0+\beta_1 \log(C_i)+v_i$$

$$v_i \overset{\text{iid}}{\sim} N(0,\sigma_v^2)$$

HB estimates of Y_i and undercoverage rates $U_i = Y_i/(Y_i+C_i)$ and associated CV's (based on posterior variance) calculated using MCMC

Limitations of HB

1. Choice of prior
2. Non-existent posterior: MCMC may not detect
3. Failure of convergence diagnostics for MCMC

Model diagnostics

- Bayes factors, posterior predictive density, cross-validation predictive density (Rao, 2003, section 10.2.6)
- Posterior predictive probability used to check overall model fit.
- Sinharay and Stern (2003): difficult to detect deviations from normality of v_i using this criterion, unless the extent of violation is huge.

Recent applications

1. Basic area level model
   U.S. county counts of poor school-age children: Y_i

   Model:
   $\widehat{\theta}_i = \theta_i + e_i; \theta_i = \log(Y_i) = z_i^T \beta + v_i$

   $\widehat{\theta}_i =$ CPS direct estimator

   $\widehat{Y}_i^{EB}$ taken as $\exp\{\widehat{\theta}_i^{EB}\}$

2. Basic unit level model
   Battese, Harter and Fuller (1988): county crop areas

   x_{ij} = LANDSAT satellite values
3. Time series and cross-sectional model

   - EBLUP: median income of four-person families for U.S. states

     CPS direct estimates, census covariates (Datta, Lahiri and Maiti, 2002)

- HB: monthly unemployment rates, U.S. states

  CPS direct estimates, UI claims covariates, linking model accounted for seasonal variation (Datta, Lahiri, Maiti and Lu, 1999)
- HB: Canadian monthly unemployment rates for Census Metropolitan Areas

  LFS direct estimates, short time series (You, Rao and Gambino, 2003)

4. Disease mapping models

- $y_i \mid \lambda_i \overset{\text{ind}}{\sim} \text{Poisson}(n_i \lambda_i)$
  $\theta_i = \log(\lambda_i) \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$
  Spatial dependence of θ_i's: CAR
  HB: Lip cancer incidence for Scottish counties (Maiti, 1998)
- Age-specific mortality rates for US Health Service areas
  HB: Nandram et al. (1999)

5. Logistic linear mixed model

- HB: Malec et al (1997)
  U.S. National Health Interview Survey for direct estimates of health-related proportions.
- HB: Malec, Davis and Cao (1999)
  Estimate overweight prevalence: U.S. National Health & Nutrition Examination Survey
- Folsom, Shah and Vaish (1999): survey weights and pseudo-HB
  Drug use prevalence rates: U.S. National Household Survey on Drug Abuse

Practical Issues

1. Design issues: Rao (2003, Chapter 2)
   Use of list frame, many small strata, compromise sample allocations, integration of surveys, multiple frames, ”rolling” samples

   “The client will always require more than is specified at the design stage.” (Fuller, 1999)

Practical Issues

2. Model selection and validation:
   Good auxiliary information, model diagnostics: (Rao, 2003, Chapter 6)
   - residual analysis
   - selection of auxiliary variables
   - case-deletion to detect influential observations

   HB: Avoid “plug and play” implementation via MCMC

Practical Issues

3. Area level vs. unit level model:

   Sample selection bias, known sampling variances

4. “Triple-goal” estimation:

   Good ranks, good histogram, good area-specific estimators

Practical Issues

5. Nonsampling errors:

   HB nonresponse models for binary data: Nandram and Choi (2002)

   Measurement errors in response: Fuller (1995)

6. How to handle when explicit data pooling prohibited: Reiter (2000)