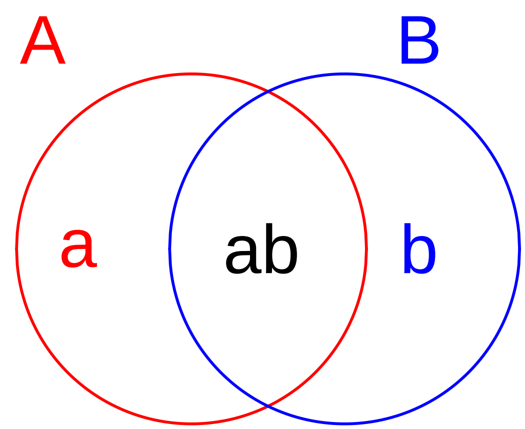
Dual Frame Surveys

- $a \cup ab \cup b = U$

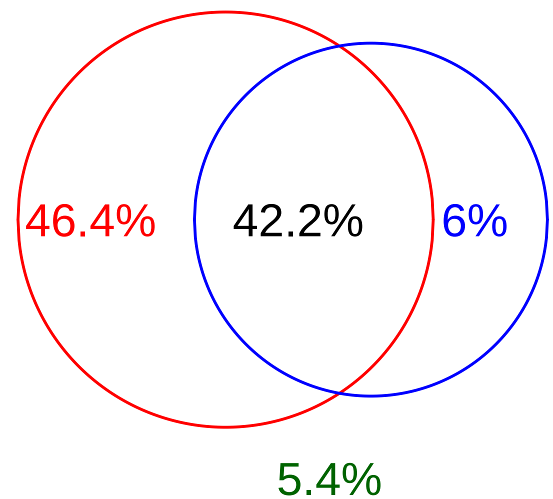


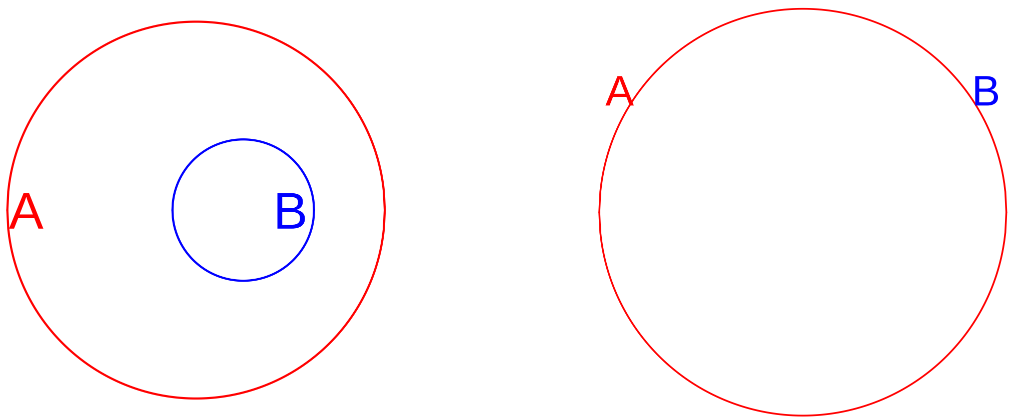
- Independent samples are taken from the two sampling frames
- Used when one frame does not cover whole population of interest, dual frame surveys can provide better coverage and cost less
- May want several survey modes (internet, telephone, personal)

Examples:

- Telephone Surveys

  Tucker et al,.2005
- Frame A: Landlines
- Frame B: Cell Phones
- No Telephones

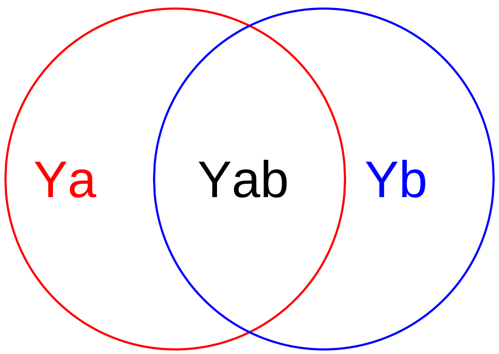




- One Frame is a Proper Subset of Another
  Rare Events (Asthma Patients)
  Frame A: General population health survey
  Frame B: Survey of patients of allergists

- Frame A and Frame B are the same
  Frame A: Current Population Survey (CPS)
  Frame B: Survey of Income and Program Participation (SIPP)

Point Estimators in Dual Frame Surveys



- $\hat{Y}_{ab}(\beta) = \beta \hat{Y}_{ab}^{A} + (1-\beta)\hat{Y}_{ab}^{B}$
- $\hat{Y} = \hat{Y}_a + \hat{Y}_{ab} + \hat{Y}_b$
  $= \hat{Y}_a + \beta \hat{Y}_{ab}^{A} + (1-\beta)\hat{Y}_{ab}^{B} + \hat{Y}_b$
- Cross-sectional
- Hartley (1962) (Choose β to minimize variance)
- Fuller & Burmeister (1972) (Optimal)
- Bankier (1986)
- Kalton & Anderson (1986) (Single frame)
- Skinner (1991) (Maximum likelihood)

Problems from Hartley and Fuller & Burmeister Estimators

- Choose β to minimize the variance
- β depend on y's
- Different set of weights for each variable
- Inconsistencies among estimates

  Y_1 = the number of men who are unemployed

  Y_2 = the number of women who are unemployed

  Y_3 = the number of people who are unemployed

  In a complex survey, it will often be the case that $\hat{Y}_1 + \hat{Y}_2 \neq \hat{Y}_3$

Single Frame Estimators

Bankier, 1986, Kalton and Anderson 1986

$$\hat{Y}_S = \sum_{i \in S_A} w_i^* y_i + \sum_{i \in S_B} w_i^* y_i$$

$$w_i^* = \begin{cases} 1/\pi_i^A & i \in a \\ 1/\pi_i^B & i \in b \\ 1/(\pi_i^A + \pi_i^B) & i \in ab \end{cases}$$

- Consider that all observations had been sampled from a single frame with modified weights for the overlap domain observations
- Easy to calculate
- Need to know the inclusion probabilities for both frames. We may not know the frame A inclusion probabilities for sample units selected from frame B that fall in domain ab.
- Single frame estimates depend only on inclusion probabilities and not on variances within the two frames. The resulting estimates can be far from optimal.

Pseudo-Maximum Likelihood (PML) (Skinner & Rao 1996)

- Skinner & Rao (1996)
- Modify MLEs for SRS
- Adjust for complex sampling design
- Single set of weights for all the variables
- Perform similarly to Fuller & Burmeister estimator in many surveys

Common case, N_A and N_B known, but N_{ab} unknown

$$\hat{N}_{ab,H} = \frac{pn_{ab}^{A}N_A}{n_A} + \frac{qn_{ab}^{B}N_B}{n_B} \tag{1}$$

where $p + q = 1$

$$\begin{aligned} Var(\hat{N}_{ab,H}) &= p^2(\frac{N_A}{n_A})^2 Var(n_{ab}^{A}) \\ &+ q^2(\frac{N_B}{n_B})^2 Var(n_{ab}^{B}) \end{aligned}$$

- n_{ab}^A and n_{ab}^B are hypergeometric random variabels

$$n_{ab}^A \sim \frac{\begin{pmatrix} N_{ab} \\ x \end{pmatrix} \begin{pmatrix} N_a \\ n_A - x \end{pmatrix}}{\begin{pmatrix} N_A \\ n_A \end{pmatrix}}$$

$$n_{ab}^B \sim \frac{\begin{pmatrix} N_{ab} \\ x \end{pmatrix} \begin{pmatrix} N_b \\ n_B - x \end{pmatrix}}{\begin{pmatrix} N_B \\ n_B \end{pmatrix}}$$

$$Var(n_{ab}^A) = n_A \frac{N_{ab}}{N_A}(1 - \frac{N_{ab}}{N_A})\frac{N_A - n_A}{N_A - 1}$$

$$Var(n_{ab}^B) = n_B \frac{N_{ab}}{N_B}(1 - \frac{N_{ab}}{N_B})\frac{N_B - n_B}{N_B - 1}$$

Minimize $Var(\hat{N}_{ab,H})$, we have

$$p_{OH} = \frac{n_A N_b g_B}{n_A N_b g_B + n_B N_a g_A} \tag{2}$$

where

$$g_A = \frac{N_A - n_A}{N_A - 1}$$

and

$$g_B = \frac{N_B - n_B}{N_B - 1}$$

substituting (2) for p into (1). (1) then reduces to a quadratic in $\hat{N}_{ab}$,

$$
\begin{aligned}
& [n_A g_B + n_B g_A] \hat{N}_{ab,s}^2 \\
- & [n_A N_B g_B + n_B N_A g_A \\
+ & n_{ab}^A N_A g_B + n_{ab}^B N_B g_A] \hat{N}_{ab,s} \\
+ & [n_{ab}^A g_B + n_{ab}^B g_A] N_A N_B = 0
\end{aligned}
$$

$$
\begin{aligned}
\hat{Y}_{srs} &= (N_A - \hat{N}_{ab,srs})\hat{u}^A_{a,srs} + \hat{N}_{ab,srs}\hat{u}_{ab,srs} \\
&+ (N_B - \hat{N}_{ab,srs})\hat{u}^B_{b,srs}
\end{aligned}
$$

$$
\hat{u}_{a,srs} = \sum_{s_a} \frac{y_i}{n_a}, \hat{u}^A_{ab,srs} = \sum_{s_A, s_{ab}} \frac{y_i}{n^A_{ab}}
$$

$$
\hat{u}_{b,srs} = \sum_{s_b} \frac{y_i}{n_b}, \hat{u}^B_{ab,srs} = \sum_{s_B, s_{ab}} \frac{y_i}{n^B_{ab}}
$$

$$
\hat{u}_{ab,srs} = (n^A_{ab}\hat{u}^A_{ab,srs} + n^B_{ab}\hat{u}^B_{ab,srs})/(n^A_{ab} + n^B_{ab})
$$

Adjusted to Complex Surveys

$$\hat{u}_{a,srs} \longrightarrow \hat{u}_a = \frac{\hat{Y}_a}{\hat{N}_a}, \hat{u}^A_{ab,srs} \longrightarrow \hat{u}^A_{ab} = \sum_{s_A, s_{ab}} \frac{\hat{Y}^A_{ab}}{\hat{N}^A_{ab}}$$

$$\hat{u}_{b,srs} \longrightarrow \hat{u}_b = \frac{\hat{Y}_b}{\hat{N}_b}, \hat{u}^B_{ab,srs} \longrightarrow \hat{u}^B_{ab} = \sum_{s_B, s_{ab}} \frac{\hat{Y}^B_{ab}}{\hat{N}^B_{ab}}$$

$$\hat{u}_{ab,srs} = (n^A_{ab}\hat{u}^A_{ab,srs} + n^B_{ab}\hat{u}^B_{ab,srs})/(n^A_{ab} + n^B_{ab}) \longrightarrow$$
$$\hat{u}_{ab} = [\frac{n_A}{N_A}\hat{N}^A_{ab}\hat{u}^A_{ab} + \frac{n_B}{N_B}\hat{N}^B_{ab}\hat{u}^B_{ab}]/[\frac{n_A}{N_A}\hat{N}^A_{ab} + \frac{n_B}{N_B}\hat{N}^B_{ab}]$$

Pseudo-Maximum Likelohood (PML)(Skinner and Rao (1996))

$$\begin{aligned} \hat{Y}_{PML} &= (N_A - \hat{N}_{ab,PML})\hat{u}_a^A + \hat{N}_{ab,PML}\hat{u}_{ab} \\ &+ (N_B - \hat{N}_{ab,PML})\hat{u}_b^B \end{aligned}$$

$\hat{N}_{ab}^{PML}$ is the smaller of the roots of the quadratic equation

$$px^2 - qx + r = 0 \tag{3}$$

where

$$p = n_A + n_B,$$

$$q = n_A N_B + n_B N_A + n_A \hat{N}_{ab}^A + n_B \hat{N}_{ab}^B,$$

and

$$r = n_A \hat{N}_{ab}^A N_B + n_B \hat{N}_{ab}^B N_A.$$

Optimal Choice of n_A and n_B

$$\hat{\boldsymbol{\eta}} = (\hat{u}_a^A, \hat{u}_{ab}^A, \hat{N}_{ab}^A/N, \hat{u}_b^B, \hat{u}_{ab}^B, \hat{N}_{ab}^B/N)' \quad (4)$$

$$\boldsymbol{\eta} = (u_a, u_{ab}, N_{ab}/N, u_b, u_{ab}, N_{ab}/N)'. \quad (5)$$

Under certain condition, $\hat{\boldsymbol{\eta}}$ is consistent for $\boldsymbol{\eta}$, and

$$\tilde{n}^{1/2}(\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}) \xrightarrow{d} N(0, \boldsymbol{\Sigma}), \quad (6)$$

where $\boldsymbol{\Sigma}$ is a block-diagonal matrix with blocks $\boldsymbol{\Sigma}_A$ and $\boldsymbol{\Sigma}_B$, $\boldsymbol{\Sigma}_A$ is the asymptotic covariance matrix of $\tilde{n}^{1/2}\hat{\boldsymbol{\eta}}_A = \tilde{n}^{1/2}(\hat{u}_a^A, \hat{u}_{ab}^A, \hat{N}_{ab}^A/N)'$ and $\boldsymbol{\Sigma}_B$ is the asymptotic covariance matrix of $\tilde{n}^{1/2}\hat{\boldsymbol{\eta}}_B = \tilde{n}^{1/2}(\hat{u}_b^B, \hat{u}_{ab}^B, \hat{N}_{ab}^B/N)'$.

- $\hat{Y}_{PML}$ depends on n_A and n_B only via the ratio n_A/n_B
- Choose n_A/n_B to minimize $avar(\hat{N}_{ab,PML})$
- By delta method, $\hat{Y}_{PML}$ is asymptotically normal with mean Y and variance $(N^2/n)\sigma^2_{PML}$

- $\sigma^2_{PML} = \nabla' \mathbf{\Sigma} \nabla$
  $\nabla = [N_a/N, \theta N_{ab}/N, \phi(u_{ab} - u_a - u_b),$
  $N_b/N, (1-\theta)N_{ab}/N, (1-\phi)(u_{ab} - u_a - u_b)],$
  $\phi = n_A N_b/(n_A N_b + n_B N_a)$
  and $\theta = n_A N_B/(n_A N_B + n_B N_A)$
- Setting
  $u_a = u_b = 0, u_{ab} = 1$ in ∇
  $\sigma^2_{Ai} = \sigma^2_{Bi} = 0, \sigma_{Aij} = \sigma_{Bij} = 0(i, j = 1, 2)$ in $\mathbf{\Sigma}$
- $avar(\hat{N}_{ab,PML}) = (N^2/n)[\phi^2 \sigma^2_{A3} + (1-\phi)^2 \sigma^2_{B3}]$
- Minimized when $\phi = \sigma^2_{B3}/(\sigma^2_{A3} + \sigma^2_{B3})$
- Equivalently, $n_A/n_B = N_a \sigma^2_{B3}/(N_b \sigma^2_{A3})$

Gross Flow Estimation in Single Frame Survey

- Blumenthal (1968) Multinomial Sampling
- Chen-Fienberg (1974), Stasny (1986) Two stage model, reduced parametrizations of general interest
- Stasny and Fienberg (1984, 1985), Stasny (1983, 1986, 1987, 1988) Continuous time model
- Holt and Skinner (1989) Use survey weights to estimate transition probabilities
- Singh and Rao (1995) Classification error
- Most assume SRS, do not account for survey design

- Lu and Lohr (2010) Gross flow estimation in dual frame surveys

  —-Accounts for survey design

  —-Consider the overlap domain estimation

  —-Model missing data through two-stage procedure