Two phase designs with data missing by design: inverse probability weighted estimators; adjustments and improvements

Jon A. Wellner

University of Washington, Seattle – visiting Heidelberg

University of Heidelberg, December 5, 2011

Biometry Seminar, Heidelberg

Based on joint work with:

- Norman Breslow
- Takumi Saegusa

and

- 2011 ISI Review paper: Lumley, Shaw, Dai

Outline

- 1: Survey sampling: the Horvitz-Thompson estimator
- 2: Adjustments of the Horvitz-Thompson estimator:
  - ▷ Regression
  - ▷ Calibration
  - ▷ Estimated weights
- 3: The "paradox"
- 4: Parametric Models & Super-populations
- 5: Limit theory, part 1
- 6: Semiparametric Models & Super-populations
- 7: Limit theory, part 2
- 8: Problems and further questions

Outline

- Part I: Sections 1-4. Based on
  Thomas Lumley, Pamela Shaw, and James Dai (2011).
  Connections between survey and calibration estimators and semiparametric models for incomplete data.
  Int. Statist. Rev. **79**, 200-220.

- Part II: Sections 5-8: Based on
  - ▷ Norman Breslow and Jon A. W. (2007).
    Scand. J. Statist. **34**, 86-102.
  - ▷ Norman Breslow and Jon A. W. (2008).
    Scand. J. Statist. **35**, 83-103.
  - ▷ Takumi Saegusa and Jon A. W. (2011).
    Weighted likelihood estimators with calibration and estimated weights. *Manuscript in progress*.

1. Survey sampling: the Horvitz-Thompson estimator

- First consider a finite population

$$\{(x_i, y_i):\ 1 \le i \le N\}$$

with $y_i \in \mathbb{R}$, $x_i \in \mathbb{R}^p$.

- Suppose: the sampling probability π_i for each individual is known; $R_i = 1$ if item i is sampled, $R_i = 0$ if not, and

$$P(R_i = 1) = \pi_i, \qquad i = 1, \ldots, N.$$

- Bernoulli (independent) sampling: the R_i's are independent, with $P(R_i = 1) = n/N$ for $i = 1, \ldots, N$ with $1 < n < N$. Thus

$$P(R_1 = r_1, \ldots, R_N = r_N) = \left(\frac{n}{N}\right)^{\sum_1^N r_i} \left(1 - \frac{n}{N}\right)^{N - \sum_1^N r_i}$$

for $r_i \in \{0, 1\}$. Note that $\sum_1^N R_i \sim$ Binomial$(N, n/N)$ is random.

1. Survey sampling: the Horvitz-Thompson estimator

- Sampling ($n < N$ items) without replacement: the R_i's are dependent (but exchangeable); $P(R_i = 1) = n/N$ for each $i = 1, \ldots, N$,

$$P(R_i = 1, R_j = 1) = \frac{n}{N}\left(\frac{n-1}{N-1}\right), \quad 1 \le i, j \le N, \quad i \ne j,$$

$$P(\underline{R} = \underline{r}) = \frac{1}{\binom{N}{n}} \quad \text{for} \quad \underline{r} = (r_1, \ldots, r_n), \quad \text{with} \quad \sum_1^N r_i = n.$$

  Note that $\sum_1^N R_i = n$ is fixed (and non-random).

- Goal: Estimate $T \equiv \sum_{i=1}^N y_i$ (or, equivalently, $\mu_N \equiv N^{-1}T = N^{-1}\sum_1^N y_i$).

- An estimator based only on the y_i's: the Horvitz-Thompson estimator of T is

$$\widehat{T} = \sum_{i:R_i=1} \frac{1}{\pi_i} y_i = \sum_{i=1}^N \frac{R_i}{\pi_i} y_i.$$

1. Survey sampling: the Horvitz-Thompson estimator

- Properties of $\widehat{T}$:
  - $E(\widehat{T}) = \sum_{i=1}^{N} \frac{E(R_i)}{\pi_i} y_i = \sum_{i=1}^{N} y_i = T.$
  - Bernoulli sampling (BS):

$$
\begin{aligned}
Var(\widehat{T}_{BS}) &= \sum_{i=1}^{N} \frac{Var(R_i)}{\pi_i^2} y_i^2 = \sum_{i=1}^{N} \frac{\pi_i(1-\pi_i)}{\pi_i^2} y_i^2 \\
&= \sum_{i=1}^{N} \frac{1-\pi_i}{\pi_i} y_i^2 = \frac{1-n/N}{n/N} \sum_{i=1}^{N} y_i^2 \\
&= \frac{N-n}{n} \sum_{i=1}^{N} y_i^2 = n \frac{N}{n} \frac{N-1}{n} \left(1 - \frac{n-1}{N-1}\right) \frac{1}{N} \sum_{i=1}^{N} y_i^2.
\end{aligned}
$$

  - Sampling Without Replacement (SWOR):

$$
Var(\widehat{T}_{SWOR}) = n \frac{N^2}{n^2} \left(1 - \frac{n-1}{N-1}\right) \frac{1}{N} \sum_{i=1}^{N} (y_i - \overline{y}_N)^2.
$$

1. Survey sampling: the Horvitz-Thompson estimator

Something left out: Sampling n from N with replacement? In this case $\underline{R} = (R_1, \ldots, R_N) \sim \text{Mult}_N(n, (1/N, \ldots, 1/N))$, so $R_i \in \{0, 1, \ldots, n\} \supset \{0, 1\}$. Nonetheless, with π_i replaced by $E(R_i) = n/N$:

$$E(\widehat{T}_{SWR}) = T,$$
$$Var(\widehat{T}_{SWR}) = n\frac{N^2}{n^2}\frac{1}{N}\sum_{i=1}^{N}(y_i - \overline{y}_N)^2.$$

2. Adjustments of the Horvitz-Thompson estimator

Regression: Can we use the x_i's to help in estimating T?

- Notation:

$$\begin{array}{lll} X & = ((1, \underline{x}_i),\ i \text{ s.t. } R_i = 1,\ 1 \le i \le N), & n \times (p+1) \text{ matrix;} \\ Y & = (y_i,\ i \text{ such that } R_i = 1,\ 1 \le i \le N), & n \times 1 \quad \text{vector} \\ W & = \operatorname{diag}(1/\pi_i,\ i \text{ such that } R_i = 1,\ 1 \le i \le N), & n \times n \text{ matrix.} \end{array}$$

- Inverse probability weighted estimate of (finite) population Least Squares coefficients β_N are

$$\widehat{\beta} = (X^T W X)^{-1} X^T W Y.$$

- The regression estimator of T is:

$$\widehat{T}_{reg} = \sum_{i=1}^{N} \frac{R_i}{\pi_i}(y_i - x_i\widehat{\beta}) + \sum_{i=1}^{N} x_i\widehat{\beta} = 0 + \sum_{i=1}^{N} x_i\widehat{\beta}.$$

2. Adjustments of the Horvitz-Thompson estimator

- Improvement? Let

$$\rho_N^2 = \text{variance explained by finite population regression,}$$
$$1 - \rho_N^2 = \frac{\sum_1^N (y_i - x_i\beta_N)^2}{\sum_1^N (y_i - \overline{y})^2}.$$

- Decomposition:

$$\begin{aligned}
\widehat{T}_{reg} &= \sum_{i=1}^N \frac{R_i}{\pi_i}(y_i - x_i\beta_N) + \sum_{i=1}^N x_i\beta_N \; + \; \sum_{i=1}^N x_i \left(1 - \frac{R_i}{\pi_i}\right)(\widehat{\beta} - \beta_N) \\
&= \sum_{i=1}^N \frac{R_i}{\pi_i}(y_i - x_i\beta_N) + \sum_{i=1}^N x_i\beta_N \; + \; \sum_{i=1}^N x_i \left(1 - \frac{R_i}{\pi_i}\right) O_p(n^{-1/2}) \\
&\equiv I + II + O_p(N/\sqrt{n}).
\end{aligned}$$

where

$$\begin{aligned}
Var(I) &= (1 - \rho_N^2) Var(\widehat{T}) \\
Var(II) &= 0.
\end{aligned}$$

2. Adjustments of the Horvitz-Thompson estimator

- Calibration: Note that $\widehat{\beta}$ is a linear function of the sampled y_i's; i.e.

$$\widehat{T}_{reg} = \sum_{i:R_i=1} \frac{g_i}{\pi_i} y_i = \sum_{i:R_i=1} w_i y_i$$

where g_i depends on x and π but not $\underline{y}$: explicitly

$$g_i = 1 + (T_x - \widehat{T}_x)^T (X^T W X)^{-1} x_i \sim 1$$

where $T_x = \sum_1^N x_i$, $\widehat{T}_x = \sum_1^N (R_i/\pi_i) x_i$.
Conclusion: The same $1 - \rho_N^2$ reduction in variance can be achieved by adjustments of the weights:

$$\text{replace } \frac{1}{\pi_i} \text{ by } \frac{g_i}{\pi_i}.$$

2. Adjustments of the Horvitz-Thompson estimator

Since the g_i's do not depend on $\underline{y}$, they are the same if $y_i = x_i$, and for estimating T_x the regression estimator is exact, so the g_i's satisfy the following calibration equation:

$$\sum_{i=1}^{N} x_i = \sum_{i:R_i=1} \frac{g_i}{\pi_i} x_i. \qquad (1)$$

Alternative definition of the g_i's: given a "loss function" $d(a,b)$, choose $\underline{g} = (g_1, \ldots, g_N)$ to minimize

$$\sum_{i:R_i=1} d\left(\frac{g_i}{\pi_i}, \frac{1}{\pi_i}\right)$$

subject to (1).

- g_{reg} corresponding to $\widehat{T}_{reg}$ results from $d(a,b) = (a-b)^2/b$.
- g_{raking} corresponds to $d(a,b) = a(\log a - \log b) + (b-a)$; see Deville and Särndal (1992).

2. Adjustments of the Horvitz-Thompson estimator

- Estimated weights: Robins, Rotnitzky, and Zhao (1994)
  - ▷ Fit a logistic regression model to predict R_i from x_i.
  - ▷ Write $p_i = \widehat{\pi}_i$ for the fitted probability.
  - ▷ The estimating equations for this logistic regression model can be written as

$$\sum_{i=1}^{N} x_i p_i = \sum_{i=1}^{N} x_i R_i \quad \text{or} \quad \sum_{i=1}^{N} x_i \widehat{\pi}_i = \sum_{i=1}^{N} x_i R_i. \tag{2}$$

Since $1/\pi_i$ corresponds to g_i/π_i in calibration, we let $1/\widehat{\pi}_i = h_i/\pi_i$ in this "estimated weights" setting, and then rewrite (2) as

$$\sum_{i=1}^{N} x_i \frac{\pi_i}{h_i} = \sum_{i=1}^{N} R_i x_i. \tag{3}$$

This is similar to the calibration equations (1), but it has the weights on the left side rather than the right side.

2. Adjustments of the Horvitz-Thompson estimator

- Comparison: estimated weights versus calibration
  - ▷ Advantages: The weights h_i always exist and are non-negative.
  - ▷ Disadvantages: All the x_i's are required, as opposed to *sampled* x_i's and population total for calibration.

3. The "paradox"

Even though π_i is known, adjusting the weights from $1/\pi_i$ to g_i/π_i or to $1/\widehat{\pi}_i$ gives an estimator of T with reduced variance. Using estimated weights rather than known weights reduces variances.

One resolution: Compare the regression estimator (special case of calibration) to a decomposition of the Horvitz Thompson estimator:

$$\widehat{T} = \sum_{i=1}^{N} \frac{R_i}{\pi_i}(y_i - x_i\beta_0) + \sum_{i=1}^{N} \frac{R_i}{\pi_i} x_i\beta_0 \tag{4}$$

while

$$\begin{aligned}\widehat{T}_{reg} &= \sum_{i=1}^{N} \frac{R_i}{\pi_i}(y_i - x_i\beta_0) + \sum_{i=1}^{N} x_i\beta_0 \\ &\quad + \sum_{i=1}^{N} x_i \left(1 - \frac{R_i}{\pi_i}\right)(\widehat{\beta} - \beta_0). \end{aligned} \tag{5}$$

3. The "paradox"

- The first terms in the last two displays are the same.
- The second term in (5) involves the known population total $\sum_1^N x_i$ of the x_i's, while the second term in (4) involves the estimated total.
- The third term (smaller order) term in (5) is not present in (4).

Conclusion: for large enough n and N, $\widehat{T}_{reg}$ will always be at least as efficient as $\widehat{T}$.

- Other resolutions: via projections of influence functions (Henmi & Eguchi 2004, RRZ 1994)
- Further examples: Lawless, Kalbfleisch, Wild (1999); Zou & Fine (2002) .

4. Parametric Models & Super-populations

- Suppose that $\mathcal{P} = \{P_\theta : \quad \theta \in \Theta \subset \mathbb{R}^d\}$ is a parametric model.
- Suppose that for complete data W_i, $1 \le i \le N$

$$N\mathbb{P}_N \psi(W_i;\theta) = \sum_{i=1}^{N} U_i(\theta)$$

are unbiased estimating equation(s): typically θ_N solving

$$\sum_{i=1}^{N} U_i(\theta_N) = 0$$

satisfies $\sqrt{N}(\theta_N - \theta_0) \rightarrow_d N_d(0,\Sigma)$ where $\Sigma = A^T BA$. Replacing y_i by $U_i(\theta)$ above yields $\widehat{\theta}_N$ satisfying

$$\sum_{i=1}^{N} \frac{R_i}{\pi_i} U_i(\widehat{\theta}) = 0$$

where $\sum_{i=1}^{N} \frac{R_i}{\pi_i} U_i(\theta)$ is the Horvitz-Thompson estimator of $\sum_{i=1}^{N} U_i(\theta)$. (Binder (1983))

4. Parametric Models & Super-populations

- The above $\widehat{\theta}_N =$ Horvitz-Thompson estimator of θ_N.
- Calibration estimators of estimating functions? Yes; see:
  - ▷ Rao, Yung, and Hidiroglou (2002).
  - ▷ Särndal, Swenson, and Wretman (2003).
- Rewrite of regression estimator:

$$\widehat{T}_{reg} = \sum_{i=1}^{N} \frac{R_i}{\pi_i}(y_i - x_i\widehat{\beta}) + \sum_{i=1}^{N} x_i\widehat{\beta} = \sum_{i=1}^{N} \left\{ \frac{R_i}{\pi_i} y_i + \left(1 - \frac{R_i}{\pi_i}\right) x_i\widehat{\beta} \right\}$$

Replacing y_i (real-valued) by $U_i(\theta)$ (vector) yields

$$T_N(\theta) = \sum_{i=1}^{N} \left\{ \frac{R_i}{\pi_i} U_i(\theta) + \left(1 - \frac{R_i}{\pi_i}\right) \phi_i \right\}$$

where ϕ_i is a $d-$vector of arbitrary functions of the data that are available for all N observations.

4. Parametric Models & Super-populations

- Solutions $\widehat{\theta}_N$ of $T_N(\widehat{\theta}_N) = 0$ gives the class of Augmented Inverse Probability Weighted Estimators (AIPW estimators) of Robins, Rotnitzky, and Zhao (1994).

- Superpopulation setting: $\{(x_i, y_i) : 1 \le i \le N\}$ is the realization of a random sample from a population or (hypothetical) super-population. Thus in the simple context of estimating a total or mean, we suppose that $Y_1, \ldots, Y_N$ are i.i.d. P on $\mathbb{R}$ with $\mu = E(Y_1)$ and $\sigma_Y^2 = Var(Y_1) < \infty$.

- We want to study the Horvitz-Thompson estimator $\widehat{\mu} = N^{-1}\widehat{T}$ as an estimator of $\mu = E(Y_1)$.

- Key decomposition: Note that:

$$\begin{aligned}
\sqrt{N}(\widehat{\mu} - \mu) &= \sqrt{N}(\mu_N - \mu) + \sqrt{N}(\widehat{\mu}_N - \mu_N) \\
&= \sqrt{N}(\overline{Y}_N - \mu) + \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \left(\frac{R_i}{\pi_i} - 1 \right) Y_i \\
&= \sqrt{N}(\overline{Y}_N - \mu) + \frac{N}{n} \frac{1}{\sqrt{N}} \sum_{i=1}^{N} (R_i - \pi_i)\, Y_i \\
&= I_N + II_N
\end{aligned}$$

where $\mu_N = N^{-1} \sum_{i=1}^{N} Y_i = N^{-1} T$.

5. Limit theory, part 1

- $I_N = \sqrt{N}(\overline{Y}_N - \mu) \to_d \sigma Z_1$ where $Z_1 \sim N(0,1)$.
- In the three cases BS, $SWOR$, and SWR, the second term II_N can be rewritten as

$$II_N \;=\; \sqrt{\frac{N}{n}}\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{N} R_i Y_i - \overline{Y}_N\right).$$

- Assuming $n/N \to \lambda \in (0,1)$,

$$II_N \to_d \begin{cases} \sqrt{\frac{1-\lambda}{\lambda}}\sqrt{E(Y^2)}Z_2, & \text{for } BS, \\ \sqrt{\frac{1-\lambda}{\lambda}}\sigma Z_2, & \text{for } SWOR, \\ \sqrt{\frac{1}{\lambda}}\sigma Z_2, & \text{for } SWR \end{cases}$$

where $Z_2 \sim N(0,1)$ is independent of Z_1.

5. Limit theory, part 1

- Putting the results for I_N and II_N together yields:

$$\sqrt{n}(\widehat{\mu}_N - \mu) \ \to_d \ \begin{cases} N(0, \lambda\sigma^2 + (1-\lambda)E(Y^2), & \text{for } \ BS, \\ N(0, \sigma^2), & \text{for } \ SWOR, \\ N(0, (\lambda+1)\sigma^2), & \text{for } \ SWR. \end{cases}$$

- If $E(Y) \neq 0$, then SWOR wins!

6. Semiparametric Models & Super-populations

- **Setting:**
  - ▷ Semiparametric model, $X \sim P_{\theta,\eta} \in \mathcal{P}$
    - ○ parametric part: $\theta \in \Theta \subset \mathbb{R}^d$
    - ○ nonparametric part: $\eta \in H \subset \mathcal{B}$, a Banach space
- **Assumptions:** To guarantee $\sqrt{N}-$consistency, suppose there exist asymptotically Gaussian ML estimators $(\widehat{\theta}_n, \widehat{\eta}_n)$ of θ and η under i.i.d. random sampling (i.e. complete data).
  - ▷ **1.** Scores $\dot{l}_\theta$ and $\dot{l}_\eta h = B_{\theta,\eta} h$, $h \in \mathcal{H} \subset \mathcal{B}$ in a Donsker class $\mathcal{F}$.
  - ▷ **2.** Scores $L_2(P_0)-$continuous at θ_0, η_0.
  - ▷ **3.** Information operator $\dot{l}_\eta^T \dot{l}_\eta = B_0^* B_0$ continuously invertible on its range.
  - ▷ **4.** $(\widehat{\theta}_N, \widehat{\eta}_N)$ are consistent for (θ_0, η_0).

6. Semiparametric Models & Super-populations

- Missing data – by design! X not observed for all items / individuals
- $\widetilde{X} = \widetilde{X}(X)$ observable part of X in phase 1
- Auxiliary U helps predict inclusion in subsample
  - ▷ $W = (X, U) \in \mathcal{W}$ observable only in validation (phase 2) sample
  - ▷ $V = (\widetilde{X}, U) \in \mathcal{V}$ observable in phase 1 (for all)
- Phase 1: $\{W_1, \ldots, W_N\}$ i.i.d. $P = P_W$
  - ▷ but observe only $\{V_1, \ldots, V_N\}$
- Phase 2: Sampling indicators $\{R_1, \ldots, R_N\}$
  - ▷ observe W_i (all of X_i) if $R_i = 1$

6. Semiparametric Models & Super-populations

Many choices for the (phase 2) sampling indicators R_i; here:

- Bernoulli sampling

$$Pr(R_i = 1|W_i) = Pr(R_i = 1|V_i) = \pi_0(V_i)$$

  conditionally independent given the V_i's.

- Finite population stratified sampling
  - ▷ Partition $\mathcal{V}$ into J strata $\mathcal{V} = \mathcal{V}_1 \cup \ldots \cup \mathcal{V}_J$.
  - ▷ Phase 1: Observe $N_j = \sum_{j=1}^{N} 1\{V_i \in \mathcal{V}_j\}$ subjects in stratum j
  - ▷ Phase 2: Sample n_j of N_j without replacement:
  - ▷ Result: sampling indicators $R_{j,i}$ for subject i in stratum j
  - ▷ $(R_{j,1}, \ldots, R_{j,N_j})$ exchangeable with

$$Pr(R_{ji} = 1|V_1, \ldots, V_N) = n_j/N_j.$$

  - ▷ The vectors $(R_{j,1}, \ldots, \xi_{j,N_j})$, $j = 1, \ldots, J$ are independent.

6. Semiparametric Models & Super-populations

- Define inverse probability weighted (IPW) empirical measure:

$$\mathbb{P}_N^{\pi} = \frac{1}{N} \sum_{i=1}^{N} \frac{R_i}{\pi_i} \delta_{X_i}, \qquad \delta_x = \text{ Dirac measure at } x$$

$$\pi_i = \begin{cases} \pi_0(V_i) & \text{if Bernoulli sampling} \\ \frac{n_j}{N_j} 1\{V_i \in \mathcal{V}_j\} & \text{if finite pop'ln stratified sampling} \end{cases}$$

- Jointly solve the finite - (for θ) and infinite (for η) dimensional equations

$$\begin{aligned} \mathbb{P}_N^{\pi} \dot{l}_\theta &= 0 & \text{in } \mathbb{R}^d \\ \mathbb{P}_N^{\pi} \dot{l}_\eta &= 0 & \text{for all } h \in \mathcal{H}. \end{aligned}$$

- MLE for complete data solves same equations with $\mathbb{P}_N$ instead of $\mathbb{P}_N^{\pi}$.

7. Limit theory, part 2

- Result 1: $\widehat{\theta}_N$ solving the IPW estimating equations is asymptotically linear in that:

$$\sqrt{N}(\widehat{\theta}_N - \theta_0) \;=\; \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \frac{R_i}{\pi_i} \tilde{l}_{\theta_0}(X_i) + o_p(1) = \mathbb{G}_N^{\pi}(\tilde{l}_{\theta_0}) + o_p(1)$$

$\tilde{l}_{\theta}(x)$ is the semparametric efficient influence function for θ (complete data)

$$\mathbb{G}_N^{\pi} = \sqrt{N}(\mathbb{P}_N^{\pi} - P_0).$$

- Notation:
  - ▷ Finite sampling empirical measure for stratum $j \in \{1, \ldots, J\}$:

$$\mathbb{P}_{j,N_j}^{R} = \frac{1}{N_j} \sum_{i=1}^{N_j} R_{ji} \delta_{X_{ji}}$$

  - ▷ Finite sampling empirical process

$$\mathbb{G}_{j,N_j}^{R} = \sqrt{N_j}\left(\mathbb{P}_{j,N_j}^{R} - \frac{n_j}{N_j}\mathbb{P}_{j,N_j}\right),$$

7. Limit theory, part 2

- More notation: $\nu_j \equiv P_0(\mathcal{V}_j)$, for $j \in \{1, \ldots, J\}$.
- Assume: $n_j/N_j \to_p \lambda_j \in (0, 1)$, for $j \in \{1, \ldots, J\}$.
- Key decomposition:

$$\mathbb{G}_N^{\pi} = \mathbb{G}_N + \sum_{j=1}^{J} \frac{N_j}{N} \left(\frac{N_j}{n_j} \right) \mathbb{G}_{j,N_j}^{R};$$

that is

$$\begin{aligned} \sqrt{N}(\mathbb{P}_N^{\pi} - P_0) &= \sqrt{N}(\mathbb{P}_N - P_0) \\ &\quad + \sum_{j=1}^{J} \frac{N_j}{N} \left(\frac{N_j}{n_j} \right) \sqrt{N_j} \left(\mathbb{P}_{j,N_j}^{R} - \frac{n_j}{N_j} \mathbb{P}_{j,N_j} \right). \end{aligned}$$

7. Limit theory, part 2

- $$\begin{aligned} \sqrt{N}(\mathbb{P}_N^\pi - P_0) &= \sqrt{N}(\mathbb{P}_N - P_0) \\ &\quad + \sum_{j=1}^J \sqrt{\frac{N_j}{N}} \left(\frac{N_j}{n_j}\right) \sqrt{N_j} \left(\mathbb{P}_{j,N_j}^R - \frac{n_j}{N_j}\mathbb{P}_{j,N_j}\right) \\ &\rightsquigarrow \mathbb{G} + \sum_{j=1}^J \sqrt{\nu_j}\sqrt{\frac{1-\lambda_j}{\lambda_j}}\mathbb{G}_j \end{aligned}$$

  where:

- $(\mathbb{G}, \mathbb{G}_1, \ldots, \mathbb{G}_J)$ are all independent, $\mathbb{G}$ is a P_0-Brownian bridge process (indexed by $\mathcal{F}$), $\mathbb{G}_j$ is a $P_j = P_0(\cdot|\mathcal{V}_j)$-Brownian bridge process indexed by $\mathcal{F}$, and

- $$(\mathbb{G}_N, \mathbb{G}_{1,N_1}^R, \ldots, \mathbb{G}_{J,N_J}^R) \rightsquigarrow (\mathbb{G}, \sqrt{\lambda_1(1-\lambda_1)}\mathbb{G}_1, \ldots, \sqrt{\lambda_J(1-\lambda_J)}\mathbb{G}_J).$$

7. Limit theory, part 2

- Upshot, raw weighted likelihood (Horvitz-Thompson):

$$\sqrt{N}(\widehat{\theta}_N - \theta_0) = \mathbb{G}_N^R(\widetilde{\ell}_{\theta_0,\eta_0}) + o_p(1) \to_d N(0,\Sigma)$$

where $\widetilde{l}_0 \equiv \widetilde{\ell}_{\theta_0,\eta_0}$ is the efficient influence function for θ with complete data, and

$$\Sigma = \begin{cases} I_0{}^{-1} + \sum_{j=1}^J \nu_j \frac{1-\lambda_j}{\lambda_j} E_j(\widetilde{\ell}_0^{\otimes 2}), & \text{Bernoulli sampling} \\ I_0{}^{-1} + \sum_{j=1}^J \nu_j \frac{1-\lambda_j}{\lambda_j} Var_j(\tilde{\ell}_0), & \text{SWOR} \\ I_0{}^{-1} + \sum_{j=1}^J \nu_j \frac{1}{\lambda_j} Var_j(\tilde{\ell}_0), & \text{SWR.} \end{cases}$$

- Gain from stratified sampling is centering of efficient scores
  - ▷ Can reduce variance via finite popl'n sampling.
  - ▷ Select strata via covariates so that $\tilde{\ell}_0$ has small conditional variances on the strata
  - ▷ Going further: Improve via calibration or estimated weights!

7. Limit theory, part 2

- Upshot: weighted likelihood with calibration or estimated weights: (and SWOR)

$$
\begin{aligned}
\sqrt{N}(\widehat{\theta}_N - \theta_0) &\to_d Z \sim N(0, \Sigma), \\
\sqrt{N}(\widehat{\theta}_{N,c} - \theta_0) &\to_d Z_c \sim N(0, \Sigma_c), \\
\sqrt{N}(\widehat{\theta}_{N,e} - \theta_0) &\to_d Z_e \sim N(0, \Sigma_e),
\end{aligned}
$$

where

$$
\begin{aligned}
\Sigma &= I_0^{-1} + \sum_{j=1}^{J} \nu_j \frac{1-\lambda_j}{\lambda_j} Var_j(\tilde{\ell}_0), \\
\Sigma_c &= I_0^{-1} + \sum_{j=1}^{J} \nu_j \frac{1-\lambda_j}{\lambda_j} Var_j((I - Q_c)\tilde{\ell}_0), \\
\Sigma_e &= I_0^{-1} + \sum_{j=1}^{J} \nu_j \frac{1-\lambda_j}{\lambda_j} Var_j((I - Q_e)\tilde{\ell}_0),
\end{aligned}
$$

7. Limit theory, part 2

and, with $Z = g(V)$, g known,

$$
\begin{aligned}
Q_c f &\equiv P_0[(\pi_0^{-1}(V) - 1) f Z^T]\{P_0(\pi_0^{-1}(V) - 1) Z^{\otimes 2}\}^{-1} Z, \\
Q_e f &\equiv P_0[\pi_0^{-1}(V) f Z^T] S_0^{-1}(1 - \pi_0(V)) \dot{G}_e(Z^T \alpha_0) Z.
\end{aligned}
$$

8. Problems and further questions

- Give a unified treatment of weighed likelihood estimation with calibration and estimated weights for semiparametric models as above.
- Extend to semiparametric models where there is no $\sqrt{n}-$consistent estimator of η.
- Both of the above are treated in Saegusa and W (2011).
- Extend to more complex sampling designs; e.g. cluster sampling?
  - ▷ Key issue: no general theory for "sampling empirical processes".
- Estimators of variance; e.g. via bootstrap? (Saegusa, 2012).
- Behavior of all these estimators under model miss-specification?
- Incorporate model selection methods for choosing covariates in calibration and estimated weights improvements.

9. Selected references

- Lumley, T., Shaw, P., and Dai, J. (2011).
  Connections between survey and calibration estimators and semiparametric models for incomplete data.
  Int. Statist. Rev. **79**, 200-220.

- Breslow, N. and Wellner, J. A. (2007).
  Weighted likelihood for semiparametric models and two-phase stratified samples with application to Cox regression.
  Scand. J. Statist. **34**, 86-102.

- Breslow, N. and Wellner, J. (2008).
  A Z-theorem with estimated nuisance parameters. *Scand. J. Statist.* **35**, 83-103.

- Saegusa, T. and Wellner, J. A. (2011).
  Weighted likelihood estimators with calibration and estimated weights. *Manuscript in progress*.

9. Selected references

- Robins, J., Rotnitzky, A., and Zhao, L.-P. (1994). Estimation of regression coefficients when some regressors are not always observed. *J. Amer. Statist. Assoc.* **89**, 846-866.
- Deville, J.-C., and Särndal, C.-E. (1992). Calibration estimators in survey sampling. *J. Amer. Statist. Assoc.* **87**, 376-382.
- Särndal, C.-E. (2007). The calibration approach in survey theory and practice. *Survey Methodology* **33**, 99-119.
- Henmi, M. and Eguchi, S. (2004). A paradox concerning nuisance parameters and projected estimating equations. *Biometrika* **91**, 929-941.
- Lumley, T. (2010). *Complex Surveys: A Guide to Analysis Using R*. Wiley, Hoboken.

9. Selected references

- Breslow, N.E., Lumley, T., Ballantyne, C. M., Chambless, L.E., & Kulich, M. (2009a). Improving Horvitz-Thompson estimation ... *Stat. Biosc.* **1**, 32-49.
- Breslow, N.E., Lumley, T., Ballantyne, C. M., Chambless, L.E., & Kulich, M. (2009a). Using the whole cohort in the analysis of case-cohort data. *Amer. J. Epidemiology* **169**, 1398-1405.
- Breslow, N.E. (2009). Lecture notes, *Two Phase Stratified Designs and Analyses for Epidemiology*. (Statistics Alps) http://faculty.washington.edu/norm/software.html .
- van der Vaart, A. W. & Wellner, J. A. (1996). *Weak Convergence and Empirical Processes*. Springer, New York.
- van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge University Press, Cambridge.

Vielen Dank!