$Z-$*theorems with Estimated Nuisance Parameters*

Jon A. Wellner

University of Washington

- joint work with:
  - Norman E. Breslow, University of Washington
  - Aad van der Vaart, Vrije Universiteit, Amsterdam
  - Bin Nan, University of Michigan

- joint work with:
  - Norman E. Breslow, University of Washington
  - Aad van der Vaart, Vrije Universiteit, Amsterdam
  - Bin Nan, University of Michigan
- Talk at IMS-China International Conference on Statistics and Probability,
  Weihai, China, July 4, 2009

- joint work with:
  - Norman E. Breslow, University of Washington
  - Aad van der Vaart, Vrije Universiteit, Amsterdam
  - Bin Nan, University of Michigan
- Talk at IMS-China International Conference on Statistics and Probability,
  Weihai, China, July 4, 2009
- 233 Anniversary of Declaration of Independence
  233 = 51st prime

- joint work with:
  - Norman E. Breslow, University of Washington
  - Aad van der Vaart, Vrije Universiteit, Amsterdam
  - Bin Nan, University of Michigan
- Talk at IMS-China International Conference on Statistics and Probability,
  Weihai, China, July 4, 2009
- 233 Anniversary of Declaration of Independence
  233 = 51st prime
- *Email: jaw@stat.washington.edu*
  http: //www.stat.washington.edu/jaw/jaw.research.html

Outline

- Introduction and Motivation:

Outline

- Introduction and Motivation:
- Review: $Z-$ theorems with and without nuisance parameters

Outline

- Introduction and Motivation:
- Review: $Z-$ theorems with and without nuisance parameters
- A new infinite-dimensional $Z-$theorem with nuisance parameters

Outline

- Introduction and Motivation:
- Review: $Z-$ theorems with and without nuisance parameters
- A new infinite-dimensional $Z-$theorem with nuisance parameters
- Semiparametric Models with Missing Data (by design) Horovitz - Thompson (Inverse Probability Weighted) estimators

Outline

- Introduction and Motivation:
- Review: $Z-$ theorems with and without nuisance parameters
- A new infinite-dimensional $Z-$theorem with nuisance parameters
- Semiparametric Models with Missing Data (by design) Horovitz - Thompson (Inverse Probability Weighted) estimators
- Summary; problems and open questions

1. Introduction and Motivation

- **Setting:** semiparametric model, $X \sim P_\theta \in \mathcal{P}$
  $\theta = (\nu, \eta) \in \mathcal{V} \times H$

1. Introduction and Motivation

- **Setting:** semiparametric model, $X \sim P_\theta \in \mathcal{P}$
  $\theta = (\nu, \eta) \in \mathcal{V} \times H$
- parametric part: $\nu \in \mathcal{V} \subset \mathbb{R}^d$

1. Introduction and Motivation

- **Setting:** semiparametric model, $X \sim P_\theta \in \mathcal{P}$
  $\theta = (\nu, \eta) \in \mathcal{V} \times H$
- parametric part: $\nu \in \mathcal{V} \subset \mathbb{R}^d$
- nonparametric part: $\eta \in H \subset \mathcal{B}$, a Banach space.

1. Introduction and Motivation

- **Setting:** semiparametric model, $X \sim P_\theta \in \mathcal{P}$
  $\theta = (\nu, \eta) \in \mathcal{V} \times H$
- parametric part: $\nu \in \mathcal{V} \subset \mathbb{R}^d$
- nonparametric part: $\eta \in H \subset \mathcal{B}$, a Banach space.
- Scores for ν and η:

1. Introduction and Motivation

- **Setting:** semiparametric model, $X \sim P_\theta \in \mathcal{P}$
  $\theta = (\nu, \eta) \in \mathcal{V} \times H$
- parametric part: $\nu \in \mathcal{V} \subset \mathbb{R}^d$
- nonparametric part: $\eta \in H \subset \mathcal{B}$, a Banach space.
- Scores for ν and η:
  - Scores $\dot{l}_\nu$ and $\dot{l}_\eta h = B_{\theta,\eta} h$, $h \in \mathcal{H} \subset \mathcal{B}$
    in a Donsker class $\mathcal{F}$, with (efficient) information matrix for ν positive definite.

1. Introduction and Motivation

- **Setting:** semiparametric model, $X \sim P_\theta \in \mathcal{P}$
  $\theta = (\nu, \eta) \in \mathcal{V} \times H$
- parametric part: $\nu \in \mathcal{V} \subset \mathbb{R}^d$
- nonparametric part: $\eta \in H \subset \mathcal{B}$, a Banach space.
- Scores for ν and η:
  - Scores $\dot{l}_\nu$ and $\dot{l}_\eta h = B_{\theta,\eta} h$, $h \in \mathcal{H} \subset \mathcal{B}$
    in a Donsker class $\mathcal{F}$, with (efficient) information matrix for ν positive definite.
  - Information operator for η: $\dot{l}_\eta^T \dot{l}_\eta = B_0^* B_0$.

1. Introduction and Motivation

- **Setting:** semiparametric model, $X \sim P_\theta \in \mathcal{P}$
  $\theta = (\nu, \eta) \in \mathcal{V} \times H$
- parametric part: $\nu \in \mathcal{V} \subset \mathbb{R}^d$
- nonparametric part: $\eta \in H \subset \mathcal{B}$, a Banach space.
- Scores for ν and η:
  - Scores $\dot{l}_\nu$ and $\dot{l}_\eta h = B_{\theta,\eta} h$, $h \in \mathcal{H} \subset \mathcal{B}$
    in a Donsker class $\mathcal{F}$, with (efficient) information matrix for ν positive definite.
  - Information operator for η: $\dot{l}_\eta^T \dot{l}_\eta = B_0^* B_0$.
- Possible additional nuisance parameters $\alpha \in \mathcal{A}$
  e.g. for estimating probabilities of selection at second phase of two phase design.

- Nuisance parameter η in model?

- Nuisance parameter η in model?
  - No: Use finite-dimensional score equations for ν classical Huber $Z-$ theorem.

- Nuisance parameter η in model?
  - No: Use finite-dimensional score equations for ν classical Huber $Z-$ theorem.
  - Yes: Information operator $B_0^* B_0$ invertible?

- Nuisance parameter η in model?
  - No: Use finite-dimensional score equations for ν classical Huber $Z-$ theorem.
  - Yes: Information operator $B_0^* B_0$ invertible?
    - Yes: Use infinite-dimensional score equations for ν, η jointly: BKRW - van der Vaart infinite-dimensional $Z-$ theorem.

- Nuisance parameter η in model?
  - No: Use finite-dimensional score equations for ν classical Huber $Z-$ theorem.
  - Yes: Information operator $B_0^* B_0$ invertible?
    - Yes: Use infinite-dimensional score equations for ν, η jointly: BKRW - van der Vaart infinite-dimensional $Z-$ theorem.
    - No: Use finite dimensional score equations for ν: with η estimated by $\hat{\eta}$ (nonparametric or ad-hoc). Newey (1994); Chen-Linton-van Keilegom (2003) estimated nuisance $Z-$theorems

- Nuisance parameter η in model?
    - No: Use finite-dimensional score equations for ν classical Huber $Z-$ theorem.
    - Yes: Information operator $B_0^* B_0$ invertible?
        - Yes: Use infinite-dimensional score equations for ν, η jointly: BKRW - van der Vaart infinite-dimensional $Z-$ theorem.
        - No: Use finite dimensional score equations for ν: with η estimated by $\hat{\eta}$ (nonparametric or ad-hoc). Newey (1994); Chen-Linton-van Keilegom (2003) estimated nuisance $Z-$theorems
- What if additional nuisance parameter α in model?

New $Z-$ theorem: Breslow and Wellner (2008)

2. Review: $Z-$ theorems

- Setting for classical Huber (1967) $Z-$theorem:
  - $\theta \in \Theta \subset \mathbb{R}^d$
  - $\Psi_n : \Theta \to \mathbb{R}^d$, random;
  - $\Psi : \Theta \to \mathbb{R}^d$, deterministic; $\Psi(\theta_0) = 0$.

2. Review: $Z-$ theorems

- Setting for classical Huber (1967) $Z-$theorem:
  - $\theta \in \Theta \subset \mathbb{R}^d$
  - $\Psi_n : \Theta \to \mathbb{R}^d$, random;
  - $\Psi : \Theta \to \mathbb{R}^d$, deterministic; $\Psi(\theta_0) = 0$.
- **Theorem A:** Suppose that $\hat{\theta}_n \to_p \theta_0$, and that:

  A1. $\Psi_n(\hat{\theta}_n) = o_p(n^{-1/2})$

  A2. $\sqrt{n}(\Psi_n(\theta_0) - \Psi(\theta_0)) \to_d \mathbb{Z}$

  A3. Ψ is differentiable at θ_0 with non-singular derivative

  $$\dot{\Psi}_0 = \dot{\Psi}(\theta_0).$$

  A4.

  $$|\sqrt{n}(\Psi_n - \Psi)(\hat{\theta}_n) - \sqrt{n}(\Psi_n - \Psi)(\theta_0)| = o_p(1 + \sqrt{n}|\hat{\theta}_n - \theta_0|).$$

  Then

  $$\sqrt{n}(\hat{\theta}_n - \theta_0) \to_d -\dot{\Psi}_0^{-1}\mathbb{Z}$$

- Setting for BKRW (1993), van der Vaart (1995) infinite-dimensional $Z-$theorem: see van der Vaart and Wellner (1996)
  - $\theta \in \Theta \subset B$, a Banach space
  - $\Psi_n : \Theta \to \ell^\infty(\mathcal{H})$, random;
  - $\Psi : \Theta \to \ell^\infty(\mathcal{H})$, deterministic; $\Psi(\theta_0) = 0$.

- Setting for BKRW (1993), van der Vaart (1995) infinite-dimensional $Z-$theorem: see van der Vaart and Wellner (1996)
  - $\theta \in \Theta \subset B$, a Banach space
  - $\Psi_n : \Theta \rightarrow \ell^\infty(\mathcal{H})$, random;
  - $\Psi : \Theta \rightarrow \ell^\infty(\mathcal{H})$, deterministic; $\Psi(\theta_0) = 0$.

- **Theorem B:** Suppose that: $\hat{\theta}_n \rightarrow_p \theta_0$ in B, and that:

  B1. $\Psi_n(\hat{\theta}_n) = o_p(n^{-1/2})$ in $\ell^\infty(\mathcal{H})$

  B2. $\sqrt{n}(\Psi_n(\theta_0) - \Psi(\theta_0)) \Rightarrow \mathbb{Z}$ in $\ell^\infty(\mathcal{H})$

  B3. Ψ is Fréchet differentiable at θ_0 with (continuously) invertible derivative $\dot{\Psi}_0 = \dot{\Psi}(\theta_0)$.

  B4.

$$\|\sqrt{n}((\Psi_n - \Psi)(\hat{\theta}_n) - \sqrt{n}(\Psi_n - \Psi)(\theta_0)\|_{\mathcal{H}} = o_p(1 + \sqrt{n}\|\hat{\theta}_n - \theta_0\|_B).$$

  Then

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \Rightarrow -\dot{\Psi}_0^{-1}\mathbb{Z} \quad \text{in} \quad B.$$

- Setting for Newey '94; Chen-Linton-van Keilegom finite-dimensional $Z-$theorem with estimated parameters:
  - $\theta = (\nu, \eta)$, $\nu \in \mathcal{V} \subset \mathbb{R}^d$, $\eta \in H$ (infinite-dimensional)
  - $\Psi_n : \mathcal{V} \times H \rightarrow \mathbb{R}^p$, $p \geq d$ random;
  - $\Psi : \mathcal{V} \times H \rightarrow \mathbb{R}^p$, deterministic; $\Psi(\nu_0, \eta_0) = 0$.

- Setting for Newey '94; Chen-Linton-van Keilegom finite-dimensional $Z-$theorem with estimated parameters:
  - $\theta = (\nu, \eta)$, $\nu \in \mathcal{V} \subset \mathbb{R}^d$, $\eta \in H$ (infinite-dimensional)
  - $\Psi_n : \mathcal{V} \times H \to \mathbb{R}^p$, $p \geq d$ random;
  - $\Psi : \mathcal{V} \times H \to \mathbb{R}^p$, deterministic; $\Psi(\nu_0, \eta_0) = 0$.
- Conditions:
  Suppose that: $\hat{\nu}_n \to_p \nu_0$ in $\mathbb{R}^d$, and that:
  C0. $\hat{\eta}_n$ is an estimator of η_0 with $P(\hat{\eta}_n \in H) \to 1$,
  $\|\hat{\eta}_n - \eta_0\| = o_p(n^{-1/4})$.
  C1. $\|\Psi_n(\hat{\nu}_n, \hat{\eta}_n)\| = \inf_\nu \|\Psi_n(\nu, \hat{\eta}_n)\| + o_p(n^{-1/2})$.
  C2. $\sqrt{n}(\Psi_n - \Psi)(\nu_0, \eta_0)) + \dot{\Psi}_2(\nu_0, \eta_0)[\hat{\eta}_n - \eta_0] \to_d \mathbb{Z}$ in $\mathbb{R}^p$
  C3. (i) $\nu \mapsto \Psi(\nu, \eta_0)$ is continuously differentiable wrt ν in a neighborhood of ν_0 with non-singular derivative $\dot{\Psi}_1 = \dot{\Psi}_1(\nu_0, \eta_0)$;
  (ii) For some open neighborhood of ν_0, the maps $\eta \mapsto \Psi(\nu, \eta)$ satisfy

$$\|\Psi(\nu, \eta) - \Psi(\nu, \eta_0) - \dot{\Psi}_2(\nu, \eta_0)[\eta - \eta_0]\| \leq c\|\eta - \eta_0\|_H^2$$

- for a derivative map $\dot{\Psi}_2(\nu, \eta_0)$ satisfying

$$\|\dot{\Psi}_2(\nu, \eta_0)[\eta - \eta_0] - \dot{\Psi}_2(\nu_0, \eta_0)[\eta - \eta_0]\| = o(1)\|\nu - \nu_0\|.$$

C4.

$$\begin{aligned} &\|\sqrt{n}(\Psi_n - \Psi)(\hat{\nu}_n, \hat{\eta}_n) - \sqrt{n}(\Psi_n - \Psi)(\nu_0, \eta_0)\| \\ &\quad = o_p(1 + \sqrt{n}\{\|\Psi_n(\hat{\nu}_n, \hat{\eta}_n)\| + \|\Psi(\hat{\nu}_n, \hat{\eta}_n)\|\}). \end{aligned}$$

- for a derivative map $\dot{\Psi}_2(\nu, \eta_0)$ satisfying

$$\|\dot{\Psi}_2(\nu, \eta_0)[\eta - \eta_0] - \dot{\Psi}_2(\nu_0, \eta_0)[\eta - \eta_0]\| = o(1)\|\nu - \nu_0\|.$$

C4.

$$\begin{aligned}\|\sqrt{n}(\Psi_n - \Psi)(\hat{\nu}_n, \hat{\eta}_n) - \sqrt{n}(\Psi_n - \Psi)(\nu_0, \eta_0)\| \\ = o_p(1 + \sqrt{n}\{\|\Psi_n(\hat{\nu}_n, \hat{\eta}_n)\| + \|\Psi(\hat{\nu}_n, \hat{\eta}_n)\|\}).\end{aligned}$$

- **Theorem C:** (Pakes and Pollard, 1989; Newey, 1994; Chen-Linton-van Keilegom, 2004).
Suppose that C0 - C4 hold. Then

$$\sqrt{n}(\widehat{\nu}_n - \nu_0) \rightarrow_d (\dot{\Psi}_1^T \dot{\Psi}_1)^{-1} \dot{\Psi}_1^T \mathbb{Z}.$$

3. Infinite-dimensional Z theorem with nuisance parameters

- Setting for new theorem:
  - $\theta \in \Theta \subset B$, $\alpha \in \mathcal{A} \subset B'$, B, B' Banach spaces
  - $\Psi_n : \Theta \times \mathcal{A} \to \ell^\infty(\mathcal{H})$, random;
  - $\Psi : \Theta \times \mathcal{A} \to \ell^\infty(\mathcal{H})$, deterministic; $\Psi(\theta_0, \alpha_0) = 0$.

3. Infinite-dimensional Z theorem with nuisance parameters

- Setting for new theorem:
  - $\theta \in \Theta \subset B$, $\alpha \in \mathcal{A} \subset B'$, B, B' Banach spaces
  - $\Psi_n : \Theta \times \mathcal{A} \to \ell^\infty(\mathcal{H})$, random;
  - $\Psi : \Theta \times \mathcal{A} \to \ell^\infty(\mathcal{H})$, deterministic; $\Psi(\theta_0, \alpha_0) = 0$.
- Conditions:

  Suppose that: $\hat{\theta}_n \to_p \theta_0$ in B, and that:

  D0. $\hat{\alpha}_n$ is an estimator of α_0 with $P(\hat{\alpha}_n \in \mathcal{A}) \to 1$,

  D1. $\sup_{h \in \mathcal{H}} |\Psi_n(\hat{\theta}_n, \hat{\alpha}_n)h| = o_p(n^{-1/2})$

  D2. $\mathbb{Z}_n \equiv \sqrt{n}(\Psi_n - \Psi)(\theta_0, \alpha_0) \Rightarrow \mathbb{Z}_0$ in $\ell^\infty(\mathcal{H})$

  D3. (i) $\theta \mapsto \{\Psi(\theta, \alpha)h : \ h \in \mathcal{H}\}$ is uniformly Fréchet differentiable wrt θ in a neighborhood of α_0 with derivative maps $\dot{\Psi}_1(\theta_0, \alpha)$ satisfying

  $\dot{\Psi}_1(\theta_0, \alpha) \to \dot{\Psi}_1(\theta_0, \alpha_0)$ as $\alpha \to \alpha_0$ with $\dot{\Psi}_1(\theta_0, \alpha_0)$ continuously invertible.

- D4.

$$
\begin{aligned}
&\left\| \sqrt{n}(\Psi_n - \Psi)(\hat{\theta}_n, \alpha_0) - \sqrt{n}(\Psi_n - \Psi)(\theta_0, \alpha_0) \right\|_{\mathcal{H}} \\
&= \; o_p^*(1 + \sqrt{n}\|\hat{\theta}_n - \theta_0\|).
\end{aligned}
$$

- D4.

$$\left\|\sqrt{n}(\Psi_n - \Psi)(\hat{\theta}_n, \alpha_0) - \sqrt{n}(\Psi_n - \Psi)(\theta_0, \alpha_0)\right\|_{\mathcal{H}}$$
$$= o_p^*(1 + \sqrt{n}\|\hat{\theta}_n - \theta_0\|).$$

- **Theorem D.** Suppose that D0 - D4 hold.
(i) If $\|\sqrt{n}(\Psi_n(\theta_0, \hat{\alpha}_n) - \Psi(\theta_0, \alpha_0))\|_{\mathcal{H}} = O_p^*(1)$, then

$$\sqrt{n}\|\hat{\theta}_n - \theta_0\| = O_p^*(1), \qquad \text{and}$$

$$\begin{aligned}\sqrt{n}(\hat{\theta}_n - \theta_0) &= -\dot{\Psi}_1^{-1}\sqrt{n}(\Psi_n - \Psi)(\theta_0, \alpha_0) \\ &\quad - \dot{\Psi}_1^{-1}[\sqrt{n}(\Psi_n(\theta_0, \hat{\alpha}_n) - \Psi(\theta_0, \alpha_0))] + o_p^*(1).\end{aligned}$$

Furthermore, if $\hat{\theta}_n^0$ satisfies
$\sup_{h\in\mathcal{H}} |\Psi_n(\hat{\theta}_n^0, \alpha_0)h| = o_p^*(n^{-1/2})$, then

$$\begin{aligned}\sqrt{n}(\hat{\theta}_n - \theta_0) &= \sqrt{n}(\hat{\theta}_n^0 - \theta_0) \\ &\quad - \dot{\Psi}_1^{-1}[\sqrt{n}(\Psi_n(\theta_0, \hat{\alpha}_n) - \Psi(\theta_0, \alpha_0))] + o_p^*(1).\end{aligned}$$

- (ii) If the map $\alpha \mapsto \{\Psi(\theta_0, \alpha)h : \ h \in \mathcal{H}\}$ is Fréchet differentiable at α_0 with derivative map $\dot{\Psi}_2$ and $\sqrt{n}(\hat{\alpha}_n - \alpha_0) = \mathbb{G}_n\varphi + o_p^*(1)$ satisfies

$$(\mathbb{Z}_n, \sqrt{n}(\hat{\alpha} - \alpha_0)) \Rightarrow (\mathbb{Z}_0, \mathbb{G}\varphi),$$

then,

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \Rightarrow -\dot{\Psi}_1^{-1}(\mathbb{Z}_0 + \dot{\Psi}_2\mathbb{G}\varphi).$$

Furthermore,

$$\begin{aligned}\sqrt{n}(\hat{\theta}_n - \theta_0) \quad &= \quad \sqrt{n}(\hat{\theta}_n^0 - \theta_0) \\ &\quad - \dot{\Psi}_1^{-1}\dot{\Psi}_2\sqrt{n}(\hat{\alpha}_n - \alpha_0) + o_p^*(1).\end{aligned}$$

4. Semiparametric Models with Missing Data

Horovitz - Thompson (IPW) estimators

- Missing data – by design! X not observed for all items / individuals

4. Semiparametric Models with Missing Data

Horovitz - Thompson (IPW) estimators

- **Missing data** – by design! X not observed for all items / individuals
- $\widetilde{X} = \widetilde{X}(X)$ observable part of X in phase 1

4. Semiparametric Models with Missing Data

Horovitz - Thompson (IPW) estimators

- 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

4. Semiparametric Models with Missing Data

Horovitz - Thompson (IPW) estimators

- 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

4. Semiparametric Models with Missing Data

Horovitz - Thompson (IPW) estimators

- 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)

4. Semiparametric Models with Missing Data

Horovitz - Thompson (IPW) estimators

- 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. $P = P_W$

4. Semiparametric Models with Missing Data

Horovitz - Thompson (IPW) estimators

- 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 $\{V_1, \ldots, V_N\}$

4. Semiparametric Models with Missing Data

Horovitz - Thompson (IPW) estimators

- 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 $\{\xi_1, \ldots, \xi_N\}$

4. Semiparametric Models with Missing Data

Horovitz - Thompson (IPW) estimators

- 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 $\{\xi_1, \ldots, \xi_N\}$
  - observe W_i (all of X_i) if $\xi_i = 1$

Many choices for the (phase 2) sampling indicators ξ_i! Here:

- Bernoulli (Manski-Lerman) sampling

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

independent

Many choices for the (phase 2) sampling indicators ξ_i! Here:

- Bernoulli (Manski-Lerman) sampling

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

  independent

- Finite population stratified sampling

Many choices for the (phase 2) sampling indicators ξ_i! Here:

- Bernoulli (Manski-Lerman) sampling

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

  independent

- Finite population stratified sampling
  - Partition $\mathcal{V}$ into J strata $\mathcal{V} = \mathcal{V}_1 \cup \ldots \mathcal{V}_J$.

Many choices for the (phase 2) sampling indicators ξ_i! Here:

- Bernoulli (Manski-Lerman) sampling

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

  independent

- Finite population stratified sampling
  - Partition $\mathcal{V}$ into J strata $\mathcal{V} = \mathcal{V}_1 \cup \ldots \mathcal{V}_J$.
  - Phase 1: Observe $N_j = \sum_{j=1}^{N} 1\{V_i \in \mathcal{V}_j\}$ subjects in stratum j

Many choices for the (phase 2) sampling indicators ξ_i! Here:

- Bernoulli (Manski-Lerman) sampling

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

  independent

- Finite population stratified sampling
  - Partition $\mathcal{V}$ into J strata $\mathcal{V} = \mathcal{V}_1 \cup \ldots \mathcal{V}_J$.
  - 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:

Many choices for the (phase 2) sampling indicators ξ_i! Here:

- Bernoulli (Manski-Lerman) sampling

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

  independent

- Finite population stratified sampling
  - Partition $\mathcal{V}$ into J strata $\mathcal{V} = \mathcal{V}_1 \cup \ldots \mathcal{V}_J$.
  - 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 $\xi_{j,i}$ for subject i in stratum j

Many choices for the (phase 2) sampling indicators ξ_i! Here:

- Bernoulli (Manski-Lerman) sampling

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

  independent

- Finite population stratified sampling
  - Partition $\mathcal{V}$ into J strata $\mathcal{V} = \mathcal{V}_1 \cup \ldots \mathcal{V}_J$.
  - 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 $\xi_{j,i}$ for subject i in stratum j
  - $(\xi_{j,1}, \ldots, \xi_{j,N_j})$ exchangeable with

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

Many choices for the (phase 2) sampling indicators ξ_i! Here:

- Bernoulli (Manski-Lerman) sampling

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

  independent

- Finite population stratified sampling
  - Partition $\mathcal{V}$ into J strata $\mathcal{V} = \mathcal{V}_1 \cup \ldots \mathcal{V}_J$.
  - 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 $\xi_{j,i}$ for subject i in stratum j
  - $(\xi_{j,1}, \ldots, \xi_{j,N_j})$ exchangeable with

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

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

Horovitz-Thompson (or IPW Likelihood) Estimators

- Define Inverse Probability Weighted (IPW) empirical measure:

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

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

Horovitz-Thompson (or IPW Likelihood) Estimators

- Define Inverse Probability Weighted (IPW) empirical measure:

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

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

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

$$\begin{aligned} \mathbb{P}_N^{\pi} \dot{l}_{\nu} &= 0 \qquad \text{in } \mathbb{R}^d \\ \mathbb{P}_N^{\pi} \dot{l}_{\eta} h &= 0 \qquad \forall h \in \mathcal{H} \end{aligned}$$

Horovitz-Thompson (or IPW Likelihood) Estimators

- Define Inverse Probability Weighted (IPW) empirical measure:

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

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

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

$$\begin{aligned} \mathbb{P}_N^\pi \dot{l}_\nu &= 0 \quad \text{in } \mathbb{R}^d \\ \mathbb{P}_N^\pi \dot{l}_\eta h &= 0 \quad \forall h \in \mathcal{H} \end{aligned}$$

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

First Main Result:

- $\widehat{\theta}_N$ solving the IPW estimating equations is asymptotically linear

$$\sqrt{N}(\widehat{\nu}_N - \nu_0) = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \frac{\xi_i}{\pi_i} \widetilde{l}_{\nu_0}(X_i) + o_p(1)$$

$$= \mathbb{G}_N^{\pi}(\widetilde{l}_{\nu_0}) + o_p(1)$$

where $\widetilde{l}_\nu(x)$ is the semparametric efficient influence function for ν (complete data)

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

$$\sqrt{N}(\hat{\nu}_N - \nu_0) = \mathbb{G}_N^{\pi}(\tilde{\ell}_{\nu_0,\eta_0}) + o_p(1) \rightarrow_d N(0, \Sigma)$$

- Asymptotic variances under stratified sampling

$$\Sigma = \begin{cases} \tilde{I}^{-1} + \sum_{j=1}^{J} \nu_j \frac{1-p_j}{p_j} E_j(\tilde{\ell}^{\otimes 2}), & \text{Bernoulli sampling} \\ \tilde{I}^{-1} + \sum_{j=1}^{J} \nu_j \frac{1-p_j}{p_j} Var_j(\tilde{\ell}), & \text{finite popl'n sampling} \end{cases}$$

$$\sqrt{N}(\hat{\nu}_N - \nu_0) = \mathbb{G}_N^\pi(\tilde{\ell}_{\nu_0,\eta_0}) + o_p(1) \rightarrow_d N(0,\Sigma)$$

- Asymptotic variances under stratified sampling

$$\Sigma = \begin{cases} \tilde{I}^{-1} + \sum_{j=1}^J \nu_j \frac{1-p_j}{p_j} E_j(\tilde{\ell}^{\otimes 2}), & \text{Bernoulli sampling} \\ \tilde{I}^{-1} + \sum_{j=1}^J \nu_j \frac{1-p_j}{p_j} Var_j(\tilde{\ell}), & \text{finite popl'n sampling} \end{cases}$$

- Gain from stratified sampling without replacement is centering of efficient scores

$$\sqrt{N}(\hat{\nu}_N - \nu_0) = \mathbb{G}_N^{\pi}(\tilde{\ell}_{\nu_0,\eta_0}) + o_p(1) \rightarrow_d N(0, \Sigma)$$

- Asymptotic variances under stratified sampling

$$\Sigma = \begin{cases} \tilde{I}^{-1} + \sum_{j=1}^{J} \nu_j \frac{1-p_j}{p_j} E_j(\tilde{\ell}^{\otimes 2}), & \text{Bernoulli sampling} \\ \tilde{I}^{-1} + \sum_{j=1}^{J} \nu_j \frac{1-p_j}{p_j} Var_j(\tilde{\ell}), & \text{finite popl'n sampling} \end{cases}$$

- Gain from stratified sampling without replacement is centering of efficient scores
  - Can reduce variance (considerably) via finite popl'n sampling.

$$\sqrt{N}(\hat{\nu}_N - \nu_0) = \mathbb{G}_N^{\pi}(\tilde{\ell}_{\nu_0,\eta_0}) + o_p(1) \rightarrow_d N(0, \Sigma)$$

- Asymptotic variances under stratified sampling

$$\Sigma = \begin{cases} \tilde{I}^{-1} + \sum_{j=1}^{J} \nu_j \frac{1-p_j}{p_j} E_j(\tilde{\ell}^{\otimes 2}), & \text{Bernoulli sampling} \\ \tilde{I}^{-1} + \sum_{j=1}^{J} \nu_j \frac{1-p_j}{p_j} Var_j(\tilde{\ell}), & \text{finite popl'n sampling} \end{cases}$$

- Gain from stratified sampling without replacement is centering of efficient scores
  - Can reduce variance (considerably) via finite popl'n sampling.
  - Select strata via covariates so that $\tilde{\ell}$ has small conditional variances on the strata

$$\sqrt{N}(\hat{\nu}_N - \nu_0) = \mathbb{G}_N^\pi(\tilde{\ell}_{\nu_0,\eta_0}) + o_p(1) \rightarrow_d N(0,\Sigma)$$

- Asymptotic variances under stratified sampling

$$\Sigma = \begin{cases} \tilde{I}^{-1} + \sum_{j=1}^J \nu_j \frac{1-p_j}{p_j} E_j(\tilde{\ell}^{\otimes 2}), & \text{Bernoulli sampling} \\ \tilde{I}^{-1} + \sum_{j=1}^J \nu_j \frac{1-p_j}{p_j} Var_j(\tilde{\ell}), & \text{finite popl'n sampling} \end{cases}$$

- Gain from stratified sampling without replacement is centering of efficient scores
  - Can reduce variance (considerably) via finite popl'n sampling.
  - Select strata via covariates so that $\tilde{\ell}$ has small conditional variances on the strata
  - Alternatively: Bernoulli sampling, but model the selection probabilities $\pi_\alpha(V)$ and estimate the α's
    Need to apply the new $Z-$theorem!

Suppose that

$$Pr(\xi_i = 1 | X_i, V_i; \alpha) = Pr(\xi_i = 1 | V_i; \alpha) = \pi_\alpha(V_i)$$

where $\alpha \in \mathbb{R}^q$ is a parameter to be estimated by maximum likelihood from the phase 1 observations $\{V_i, i = 1, \ldots, N\}$. We suppose that the model for π_α is sufficiently regular so that the MLE $\hat{\alpha}$ is consistent and asymptotically normal with influence function

$$\tilde{\ell}_\alpha = \left(\tilde{P}_0 \frac{\dot{\pi}_0^{\otimes 2}}{\pi_0(1-\pi_0)} \right)^{-1} \dot{\pi}_0 \frac{\xi - \pi_0}{\pi_0(1-\pi_0)}.$$

If $\hat{\nu}(\alpha)$ denotes the IPW likelihood estimator under two-phase Bernoulli sampling with 'known' sampling function $\pi_\alpha(V)$, then let $\hat{\nu}_N(\hat{\alpha})$, $\hat{\eta}_N(\hat{\alpha})$ solve

$$\mathbb{P}_N^{\hat{\pi}} \dot{\ell}_\nu = 0, \quad \mathbb{P}_N^{\hat{\pi}} \dot{\ell}_\eta h = 0 \text{ for all } h \in \mathcal{H}.$$

where $\hat{\pi}_i \equiv \pi_{\hat{\alpha}}(V_i)$, $i = 1, \ldots, N$.

Theorem: Breslow and W (2007,2008)
Suppose the semiparametric model $\mathcal{P}$ satisfies A1* and A3* (of BW 2008), and A2, A4, A5 of BW 2007, and that the model $\pi_\alpha(V)$ satisfies the hypotheses of theorem 5.39 of van der Vaart (1998). Suppose, moreover, that π_α satisfy

$$\left| \frac{1}{\pi_\alpha(V)} - \frac{1}{\pi_{\alpha_0}(V)} - \frac{-\dot{\pi}_0^T(v)}{\pi_0^2(v)}(\alpha - \alpha_0) \right| \le \psi(v)|\alpha - \alpha_0|^{1+\xi},$$

for α in a neighborhood of α_0 for some $\xi > 0$ and $E\psi^2(V) < \infty$. Then

$$\sqrt{N}(\hat{\nu}_N(\hat{\alpha}) - \nu_0) \to_d Z \sim N_d(0, \Sigma)$$

where

$$\Sigma = Var\left(\frac{\xi}{\pi_0}\tilde{\ell}_\nu\right) - A^T B^{-1} A$$

and where ...

$$A = \tilde{P}_0 \frac{\tilde{\ell}_\nu \dot{\pi}_0}{\pi_0}, \quad B = \tilde{P}_0 \frac{\dot{\pi}_0^{\otimes 2}}{\pi_0(1-\pi_0)}.$$

- When $\pi_\alpha(V) = \sum_{j=1}^J \alpha_j 1_{\mathcal{V}_j}(V)$, the resulting asymptotic variance of the IPW estimator with $\hat{\alpha}_j = n_j/N_j$ is exactly the same as the finite population sampling variance with $p_j = \alpha_{j,0} \lim(n_j/N_j)$.
- Further efficiency gains possible by modeling $\pi_\alpha(V)$ as a function of V: see e.g. Robins, Rotnitzky, and Zhao (1994)

5. Summary; problems and open questions

- Z-theorem review

5. Summary; problems and open questions

- Z-theorem review
  - classical Huber $Z-$theorem

5. Summary; problems and open questions

- Z-theorem review
  - classical Huber $Z-$theorem
  - Pakes-Pollard (1989) further restrictions $Z-$theorem

5. Summary; problems and open questions

- Z-theorem review
  - classical Huber $Z-$theorem
  - Pakes-Pollard (1989) further restrictions $Z-$theorem
  - infinite dimensional $Z-$theorem: BKRW (1993), van der Vaart (1995), vdV-W (1996)

5. Summary; problems and open questions

- Z-theorem review
  - classical Huber $Z-$theorem
  - Pakes-Pollard (1989) further restrictions $Z-$theorem
  - infinite dimensional $Z-$theorem: BKRW (1993), van der Vaart (1995), vdV-W (1996)
  - Newey (1994), Chen-Linton-van Keilegom (2004) finite-dimensional $Z-$theorem with infinite-dimensional nuisance parameter.

5. Summary; problems and open questions

- Z-theorem review
  - classical Huber $Z-$theorem
  - Pakes-Pollard (1989) further restrictions $Z-$theorem
  - infinite dimensional $Z-$theorem: BKRW (1993), van der Vaart (1995), vdV-W (1996)
  - Newey (1994), Chen-Linton-van Keilegom (2004) finite-dimensional $Z-$theorem with infinite-dimensional nuisance parameter.
  - Breslow - Wellner (2008) infinite dimensional $Z-$theorem with (possibly) infinite-dimensional nuisance parameter

- Application to semiparametric missing data models

- Application to semiparametric missing data models
  - Basic idea: separate calculations for sampling design and for model.

- Application to semiparametric missing data models
  - Basic idea: separate calculations for sampling design and for model.
  - Sampling assumptions give properties of IPW empirical process $\mathbb{G}_N^{\pi}$

- Application to semiparametric missing data models
  - Basic idea: separate calculations for sampling design and for model.
  - Sampling assumptions give properties of IPW empirical process $\mathbb{G}_N^{\pi}$
  - Likelihood calculations for complete data problem give efficient influence function $\tilde{\ell}_{\nu}$ for ν.

- Application to semiparametric missing data models
  - Basic idea: separate calculations for sampling design and for model.
  - Sampling assumptions give properties of IPW empirical process $\mathbb{G}_N^{\pi}$
  - Likelihood calculations for complete data problem give efficient influence function $\tilde{\ell}_{\nu}$ for ν.
  - Estimating the π's can lead to increased efficiency.

- Further problems:

- Further problems:
    - Infinite-dimensional version of Pakes-Pollard $Z-$theorem?

- Further problems:
  - Infinite-dimensional version of Pakes-Pollard $Z-$theorem?
  - Other, more complex sampling designs?

- Further problems:
  - Infinite-dimensional version of Pakes-Pollard $Z-$theorem?
  - Other, more complex sampling designs?
    - Hájek (1964), Rosen (1972a,b), Isaki and Fuller (1982)

- Further problems:
  - Infinite-dimensional version of Pakes-Pollard $Z-$theorem?
  - Other, more complex sampling designs?
    - Hájek (1964), Rosen (1972a,b), Isaki and Fuller (1982)
    - Lin (2000)

- Further problems:
  - Infinite-dimensional version of Pakes-Pollard $Z-$theorem?
  - Other, more complex sampling designs?
    - Hájek (1964), Rosen (1972a,b), Isaki and Fuller (1982)
    - Lin (2000)
  - Can we handle problems with nuisance parameter estimators not converging at rate $\sqrt{N}$ together with finite-sampling or more complex designs?

- Further problems:
  - Infinite-dimensional version of Pakes-Pollard $Z-$theorem?
  - Other, more complex sampling designs?
    - Hájek (1964), Rosen (1972a,b), Isaki and Fuller (1982)
    - Lin (2000)
  - Can we handle problems with nuisance parameter estimators not converging at rate $\sqrt{N}$ together with finite-sampling or more complex designs?
    - Z - theorems of Huang (1995), Wellner and Zhang (2006), Newey (1994).

- Further problems:
  - Infinite-dimensional version of Pakes-Pollard $Z-$theorem?
  - Other, more complex sampling designs?
    - Hájek (1964), Rosen (1972a,b), Isaki and Fuller (1982)
    - Lin (2000)
  - Can we handle problems with nuisance parameter estimators not converging at rate $\sqrt{N}$ together with finite-sampling or more complex designs?
    - Z - theorems of Huang (1995), Wellner and Zhang (2006), Newey (1994).
  - Can we handle both estimating the π's and and finite popl'n sampling?

References:

- Breslow, N. E. 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. E. and Wellner, J. A. (2008). A $Z-$theorem with estimated parameters and correction note for "Weighted likelihood for semiparametric models and two-phase stratified samples, with application to Cox regression". *Scand. J. Statist.* **35**, 186 - 192.
- van der Vaart, A. W. and Wellner, J. A. (2007). Empirical processes indexed by estimated functions. In Asymptotics: particles, processes and inverse problems, *IMS Lecture Notes - Monograph Series* **55**, 234-252.

- Breslow, N. E., Lumley, T., Ballantyne, C. M., Chambless, L. E., and Kulich, M. (2009). Using the whole cohort in the anlysis of case-cohort data. *Amer. J. Epidem.* *169*, 1398 - 1405.
- Breslow, N. E., Lumley, T., Ballantyne, C. M., Chambless, L. E., and Kulich, M. (2009). Improved Horvitz-Thompson estimation of model parameters from two-phase stratified samples: applications in epidemiology. *Statist. Biosc.* *xx*, yy-zz.