Empirical Likelihood Inference with Public-Use Survey Data

Changbao Wu
Department of Statistics and Actuarial Science
University of Waterloo

(Joint work with J.N.K. Rao)

June 24, 2016 – NUS EL Workshop

Design-based Inference for Surveys

- Survey population: $\mathbf{U} = \{1, 2, \ldots, N\}$
- $\mathbf{U}$ is treated as fixed
- Measures of variables $(y_i, \boldsymbol{x}_i)$ are non-random; attached to units
- Probability sampling design: $\mathcal{P}(\mathbf{S})$
- The set of sampled units, $\mathbf{S}$, is random
- First and second order inclusion probabilities:

$$\pi_i = P(i \in \mathbf{S}) \,, \quad \pi_{ij} = P(i, j \in \mathbf{S})$$

- Design-based inference: Frequentist interpretation with respect to the probability sampling design for the given finite population

The Horvitz-Thompson Estimator

- The population total of y: $T_y = \sum_{i=1}^{N} y_i$
- The HT estimator of T_y:

$$\hat{T}_{y\mathrm{HT}} = \sum_{i \in \mathbf{S}} \frac{y_i}{\pi_i} = \sum_{i \in \mathbf{S}} d_i y_i$$

- The basic design weights: $d_i = 1/\pi_i$
- The HT estimator is the only design-unbiased estimator in a sub-class of the Godambe class of linear estimators
- Variance estimation: Require π_i and π_{ij} $(\pi_{ii} = \pi_i)$

$$v\left(\hat{T}_{y\mathrm{HT}}\right) = \sum_{i \in \mathbf{S}} \sum_{j \in \mathbf{S}} \frac{\pi_{ij} - \pi_i \pi_j}{\pi_{ij}} \frac{y_i}{\pi_i} \frac{y_j}{\pi_j}$$

Professor V. P. Godambe (June 1, 1926 – June 9, 2016)

Major Practical Issues with Survey Data

- Nonresponse
  - Unit nonresponse: No information is available for any intended measures
  - Item nonresponse: Measures on certain variables are missing
- Calibration
  - The calibration weights w_i minimize a distance measure between $(w_1, \ldots, w_n)$ and $(d_1, \ldots, d_n)$
  - The calibration weights satisfy the benchmark constraints (calibration equations)

$$\sum_{i \in \mathbf{S}} w_i \boldsymbol{x}_i = T_{\boldsymbol{x}}$$

where $T_{\boldsymbol{x}}$ are the known population totals of auxiliary variables $\boldsymbol{x}$
- Why calibration? (1) Efficiency (2) Internal consistency

Production of Public-Use Survey Data Files

- Unit nonresponse adjustment:
  - Ratio adjustment for uniform nonresponse
  - Propensity scores for non-uniform nonresponse
- Calibration weighting
  - Calibration variables are decided at the data file creation stage
  - Control totals for calibration are (typically) NOT available to users
  - It is the final calibration weights w_i, not the basic design weights d_i, that are released in the data file
- Replication weights for variance estimation
  - The second order inclusion probabilities are NOT available for users of survey data files
  - Bootstrap and jackknife, and occasionally BRR, are commonly used replication methods
- Imputation for item nonresponse: A more difficult issue!

A Typical Format for Public-Use Survey Data

i	y_{i1}	y_{i2}	x_{i1}	x_{i2}	x_{i3}	w_i	$w_i^{(1)}$	$\cdots$	$w_i^{(B)}$
1	y_{11}	y_{12}	x_{11}	x_{12}	x_{13}	w_1	$w_1^{(1)}$	$\cdots$	$w_1^{(B)}$
2	y_{21}	y_{22}	x_{21}	x_{22}	x_{23}	w_2	$w_2^{(1)}$	$\cdots$	$w_2^{(B)}$
3	y_{31}	y_{32}	x_{31}	x_{32}	x_{33}	w_3	$w_3^{(1)}$	$\cdots$	$w_3^{(B)}$
4	y_{41}	y_{42}	x_{41}	x_{42}	x_{43}	w_4	$w_4^{(1)}$	$\cdots$	$w_4^{(B)}$
5	y_{51}	y_{52}	x_{51}	x_{52}	x_{53}	w_5	$w_5^{(1)}$	$\cdots$	$w_5^{(B)}$
6	y_{61}	y_{62}	x_{61}	x_{62}	x_{63}	w_6	$w_6^{(1)}$	$\cdots$	$w_6^{(B)}$
7	y_{71}	y_{72}	x_{71}	x_{72}	x_{73}	w_7	$w_7^{(1)}$	$\cdots$	$w_7^{(B)}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
n	y_{n1}	y_{n2}	x_{n1}	x_{n2}	x_{n3}	w_n	$w_n^{(1)}$	$\cdots$	$w_n^{(B)}$

1 Public-Use Survey Data

2 Empirical Likelihood Inference

3 Bayesian Empirical Likelihood

4 Additional Remarks

Public-Use Survey Data File

- Many columns of survey variables (y_i or x_i)
- Single column of the final survey weights w_i:
  Unit nonresponse adjustment and/or calibration weighting
- Additional columns of replication weights $w_i^{(b)}$: $b = 1, \ldots, B$

- Population control totals: Not available to users!
- Detailed design information, π_i and π_{ij}: Not available to users!

- Imputation for item nonresponse: Not considered here!

Finite Population Parameters

- The parameter θ_N is defined as the solution to the "census estimating equation"

$$U_N(\theta) = \sum_{i=1}^{N} g_i(\theta) = 0$$

Estimating function: $g(\theta) = g(y, \boldsymbol{x}; \theta)$

- Population mean $\theta = \mu_y$: $g_i(\theta) = y_i - \theta$
- Distribution function $\theta = F_y(t)$: $g_i(\theta) = I(y_i \le t) - \theta$
- Population quantile $\theta = t_\alpha$: $g_i(\theta) = I(y_i \le \theta) - \alpha$
- Population regression coefficients $\boldsymbol{\theta}_N$:

$$U_N(\boldsymbol{\theta}) = \sum_{i=1}^{N} \boldsymbol{x}_i(y_i - \boldsymbol{x}_i'\boldsymbol{\theta}) = \boldsymbol{0}$$

Assumptions About Public-Use Survey Data

- Assumption 1:

  The final survey weights $(w_1, w_2, \ldots, w_n)$ and the finite population values satisfy that the expansion estimator

  $$\hat{U}_n(\theta_N) = \sum_{i \in \mathbf{S}} w_i \, g_i(\theta_N)$$

  is asymptotically normally distributed with mean zero and variance at the order $O(N^2/n)$.

Assumptions (Cont'd)

- Denote the replicated version of $\hat{U}_n(\theta_N) = \sum_{i \in \mathbf{S}} w_i g_i(\theta_N)$ as

$$\hat{\eta}^{(b)}(\theta_N) = \sum_{i \in \mathbf{S}} w_i^{(b)} g_i(\theta_N)$$

for the bth set of replication weights $(w_1^{(b)}, w_2^{(b)}, \ldots, w_n^{(b)})$.

- Assumption 2:

The replication variance estimator

$$v\{\hat{U}_n(\theta_N)\} = \frac{1}{B} \sum_{b=1}^{B} \left\{ \hat{\eta}^{(b)}(\theta_N) - \hat{U}_n(\theta_N) \right\}^2 \quad (1)$$

is design-consistent for $V\{\hat{U}_n(\theta_N)\}$.

Assumptions (Cont'd)

- Assumption 3:

  The number of replications B is large and the empirical distribution of the B replicated versions

  $$\hat{\eta}^{(1)}(\theta_N), \hat{\eta}^{(2)}(\theta_N), \ldots, \hat{\eta}^{(B)}(\theta_N)$$

  provide an approximation to the sampling distribution of $\hat{U}_n(\theta_N) = \sum_{i \in \mathbf{S}} w_i g_i(\theta_N)$.

- Assumption 1 holds for most commonly used designs and populations. It is the foundation for design-based inference
- Assumption 2 does not necessarily require B to be large
- Assumption 3 implies Assumption 2
- Most replication weights are created to satisfy Assumption 2, but not necessarily Assumption 3

Components of Standard Empirical Likelihood

[1] The (nonparametric) empirical (log) likelihood function

$$L(\boldsymbol{p}) = \prod_{i=1}^{n} p_i \qquad \text{or} \qquad \ell(\boldsymbol{p}) = \sum_{i=1}^{n} \log(p_i)\,,$$

where $\boldsymbol{p} = (p_1, \ldots, p_n)$ is a discrete probability measure over the n sampled units

[2] The normalization constraint: $p_i > 0$ and

$$\sum_{i=1}^{n} p_i = 1$$

[3] Constraints induced by parameters and/or known auxiliary information: $E\{g(y, \boldsymbol{x}; \theta)\} = \mathbf{0}$ leads to

$$\sum_{i=1}^{n} p_i\, g(y_i, \boldsymbol{x}_i; \theta) = \mathbf{0}$$

The Wu-Rao PEL (2006) for Public-Use Survey Data

- The Pseudo EL function using the final survey weights w_i

$$l_{\mathrm{WR}}(\boldsymbol{p}) = n \sum_{i \in \mathbf{S}} \tilde{w}_i(\mathbf{S}) \, \log(p_i) \, ,$$

where $\tilde{w}_i(\mathbf{S}) = w_i / \sum_{k \in \mathbf{S}} w_k$

  - $l_{\mathrm{WR}}(\boldsymbol{p})$ reduces to $\sum_{i \in \mathbf{S}} \log(p_i)$ with equal survey weights

- Standard normalization and parameter constraints

$$\sum_{i \in \mathbf{S}} p_i = 1 \quad \text{and} \quad \sum_{i \in \mathbf{S}} p_i \, g_i(\theta) = 0$$

The Wu-Rao PEL (2006) for Public-Use Survey Data

- The PEL ratio statistic for θ

$$r_{\mathrm{WR}}(\theta) = l_{\mathrm{WR}}\{\hat{\boldsymbol{p}}(\theta)\} - l_{\mathrm{WR}}(\hat{\boldsymbol{p}}) = -n \sum_{i \in \mathbf{S}} \tilde{w}_i(\mathbf{S}) \log\{1 + \lambda g_i(\theta)\}$$

- **Result 1**: Under Assumptions 1 and 2, the adjusted pseudo empirical likelihood ratio statistic $-2r_{\mathrm{WR}}(\theta)/\hat{a}_{\mathrm{WR}}$ converges in distribution to a χ^2 random variable with one degree of freedom when $\theta = \theta_N$, where the adjusting factor $\hat{a}_{\mathrm{WR}}$ is computed as

$$\hat{a}_{\mathrm{WR}} = v\{\hat{U}_n(\hat{\theta})\} / \left\{\hat{N} n^{-1} \sum_{i \in \mathbf{S}} w_i \left[g_i(\hat{\theta})\right]^2\right\},$$

with $v\{\hat{U}_n(\hat{\theta})\}$ being the replication variance estimator given in Assumption 2 but replacing θ_N by $\hat{\theta}$, and $\hat{N} = \sum_{i \in \mathbf{S}} w_i$.

The Wu-Rao PEL (2006) for Public-Use Survey Data

- Computating $r_{\rm WR}(\theta)$ (for a given θ) and the adjusting factor $\hat{a}_{\rm WR}$ requires no additional information other than the public-use survey data set
- The $1-\alpha$ level PEL ratio confidence interval for θ_N

$$\mathcal{C}_1 = \left\{ \theta \,\middle|\, -2r_{\rm WR}(\theta)/\hat{a}_{\rm WR} \le \chi^2_1(\alpha) \right\} \qquad (2)$$

- Under Assumption 3, a bootstrap calibrated PEL ratio confidence interval for θ_N can be constructed as

$$\mathcal{C}_2 = \left\{ \theta \,\middle|\, -2r_{\rm WR}(\theta) \le b_{\rm WR}(\alpha) \right\}, \qquad (3)$$

where $b_{\rm WR}(\alpha)$ be the upper α quantile from the empirical distribution of the bootstrap replicated versions $-2r^{(b)}_{\rm WR}(\hat{\theta})$, $b = 1, 2, \ldots, B$, computed in the same way as $-2r_{\rm WR}(\theta)$ at $\theta = \hat{\theta}$ but using the bth replication weights $(w^{(b)}_1, \ldots, w^{(b)}_n)$

The Re-formulated Berger-Torres EL (2016)

- Standard EL function

$$l_{\mathrm{BT}}(\boldsymbol{p}) = \sum_{i \in \mathbf{S}} \log(p_i)$$

- Standard normalization constraint

$$\sum_{i \in \mathbf{S}} p_i = 1$$

- Parameter constraint over "transformed" variables

$$\sum_{i \in \mathbf{S}} p_i \{ w_i g_i(\theta) \} = 0 \tag{4}$$

  - With equal survey weights, constraint (4) reduces to

$$\sum_{i \in \mathbf{S}} p_i \, g_i(\theta) = 0$$

The Re-formulated Berger-Torres EL (2016)

- The Berger-Torres EL ratio statistic for θ

$$r_{\mathrm{BT}}(\theta) = l_{\mathrm{BT}}\left\{\hat{\boldsymbol{p}}(\theta)\right\} - l_{\mathrm{BT}}\left(\hat{\boldsymbol{p}}\right) = \sum_{i \in \mathbf{S}} \log\{n\hat{p}_i(\theta)\}$$

- **Result 2**: Under Assumptions 1 and 2, the adjusted pseudo empirical log-likelihood ratio statistic $-2r_{\mathrm{BT}}(\theta)/\hat{a}_{\mathrm{BT}}$ converges in distribution to a χ^2 random variable with one degree of freedom when $\theta = \theta_N$, where the adjusting factor $\hat{a}_{\mathrm{BT}}$ is computed as

$$\hat{a}_{\mathrm{BT}} = v\left\{\hat{U}_n(\hat{\theta})\right\} / \left\{\sum_{i \in \mathbf{S}} \left[w_i g_i(\hat{\theta})\right]^2\right\},$$

with $v\left\{\hat{U}_n(\hat{\theta})\right\}$ being the replication variance estimator given in Assumption 2 but replacing θ_N by $\hat{\theta}$.

The Re-formulated Berger-Torres EL (2016)

- The adjusting factor $\hat{a}_{\rm BT} = 1$ under single-stage unequal probability sampling with replacement (or its asymptotic equivalence) if w_i are the original design weights
- With unit nonresponse adjustment and calibration weighting, the factor $\hat{a}_{\rm BT}$ is typically not 1 no matter what's the original survey design
- The $1-\alpha$ level EL ratio confidence interval for θ_N

$$\mathcal{C}_3 = \left\{ \theta \;\middle|\; -2r_{\rm BT}(\theta)/\hat{a}_{\rm BT} \leq \chi^2_1(\alpha) \right\} \qquad (5)$$

- ${\mathcal{C}_3}^{(1)}$: The naive EL confidence interval treating $\hat{a}_{\rm BT} = 1$
- The bootstrap calibrated EL ratio confidence interval

$$\mathcal{C}_4 = \left\{ \theta \;\middle|\; -2r_{\rm BT}(\theta) \leq b_{\rm BT}(\alpha) \right\} \qquad (6)$$

The Estimating Equation Approach

- V. P. Godambe: One of the main contributors to estimating function (EF) and estimating equation (EE) methodology
- The point estimator $\hat{\theta}$ is the solution to

$$\hat{U}_n(\theta) = \sum_{i \in \mathbf{S}} w_i g_i(\theta) = 0$$

- Confidence intervals for θ can be constructed based on the Wald-type statistic

$$W(\theta) = \left\{\hat{U}_n(\theta)\right\} / \left\{V\left[\hat{U}_n(\theta)\right]\right\}^{1/2}$$

- The variance $V\left[\hat{U}_n(\theta)\right]$ can be handled in two different ways

The Estimating Equation Approach

- **Version 1**: Use the replication variance estimator of $\hat{U}_n(\theta)$ for any given θ. Let $\hat{\eta}^{(b)}(\theta) = \sum_{i \in \mathbf{S}} w_i^{(b)} g_i(\theta)$ and

$$V\left[\hat{U}_n(\theta)\right] = B^{-1} \sum_{b=1}^{B} \left\{ \hat{\eta}^{(b)}(\theta) - \hat{U}_n(\theta) \right\}^2$$

- The profile confidence interval

$$\mathcal{C}_5 = \left\{ \theta \;\middle|\; \{W(\theta)\}^2 \le \chi_1^2(\alpha) \right\} \tag{7}$$

- **Version 2**: Use the variance formula at the fixed point $\theta = \hat{\theta}$

$$v_{\mathrm{U}} = V\left[\hat{U}_n(\hat{\theta})\right] = B^{-1} \sum_{b=1}^{B} \left\{ \hat{\eta}^{(b)}(\hat{\theta}) \right\}^2$$

- The resulting confidence interval can be written as

$$\mathcal{C}_6 = \left\{ \theta \;\middle|\; |\hat{U}_n(\theta)| \le v_{\mathrm{U}}^{1/2} Z_{\alpha/2} \right\} \tag{8}$$

Simulation Studies

- Finite population $\{(y_i, x_{i1}, x_{i2}, x_{i3}), i = 1, 2, \ldots, N\}$

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \varepsilon_i$$

- x_1: Gender; x_2: Age
  (x_1 and x_2 are used as calibration variables)
- x_3: Size variable for PPS sampling without replacement
- Three scenarios
  - A. w_i are the original design weights
  - B. w_i are adjusted for unit nonresponse
  - C. w_i are the calibration weights

Simulation Studies

- $N = 20,000$; $n = 400$; $n/N = 2\%$
- Response rate for Scenario B: 67%; Initial sample: $n_0 = 600$
- Replication weights are bootstrap weights constructed for each scenario ($B = 500$)
- Parameters of interest θ: Mean and Proportions

$$\mu_y = N^{-1}\sum_{i=1}^{N} y_i \quad \text{and} \quad F_N(t) = N^{-1}\sum_{i=1}^{N} I(y_i \le t)$$

t at five population quantiles: 5%, 10%, 50%, 90% and 95%

- Results based on 2000 simulation runs

95% Confidence Intervals Under Scenario C

θ		$\mathcal{C}_1$	$\mathcal{C}_2$	$\mathcal{C}_3$	$\mathcal{C}_3{}^{(1)}$	$\mathcal{C}_4$	$\mathcal{C}_5$	$\mathcal{C}_6$
μ_y	AL	0.212	0.221	0.213	0.228	0.225	0.212	0.211
	LE	2.7	2.3	2.4	1.8	2.1	3.4	2.7
	CP	95.6	96.3	95.7	96.8	96.5	95.3	95.4
	UE	1.7	1.4	1.9	1.4	1.4	1.3	1.9
0.05	AL	0.061	0.061	0.062	0.063	0.066	0.061	0.061
	LE	1.9	1.8	2.1	1.6	1.8	0.3	0.7
	CP	93.8	94.1	94.3	94.9	95.5	91.7	91.6
	UE	4.3	4.1	3.6	3.5	2.7	8.0	7.7
0.10	AL	0.080	0.081	0.081	0.084	0.083	0.081	0.080
	LE	2.2	1.9	2.3	2.0	2.1	0.7	1.1
	CP	94.0	94.3	94.4	95.1	95.0	93.6	93.5
	UE	3.8	3.8	3.3	2.9	2.9	5.7	5.4
0.50	AL	0.119	0.123	0.119	0.126	0.123	0.120	0.120
	LE	2.6	2.2	2.6	1.8	2.2	2.1	2.7
	CP	94.4	95.0	94.4	95.9	95.0	94.6	94.3
	UE	3.0	2.8	3.0	2.3	2.8	3.3	3.0

Key Observations

- Confidence intervals $\mathcal{C}_1$ and $\mathcal{C}_3$ based scaled χ^2_1 have excellent performances on almost all cases.
  (**Require Assumptions 1 and 2**)
- Confidence intervals $\mathcal{C}_2$ and $\mathcal{C}_4$ based on the bootstrap approximation to the sampling distribution of $-2r(\theta)$ are very similar to $\mathcal{C}_1$ and $\mathcal{C}_3$ under Scenarios A and B but have slightly inflated length for μ_y under Scenario C.
  (**Require Assumptions 1 and 3**)
- The EL-based confidence intervals $\mathcal{C}_1$ and $\mathcal{C}_3$ have clear advantages over intervals based on the estimating equation theory ($\mathcal{C}_5$) or the normal theory approximation ($\mathcal{C}_6$), especially for small or large population proportions.

1. Public-Use Survey Data
2. Empirical Likelihood Inference
3. Bayesian Empirical Likelihood
4. Additional Remarks

Bayesian Approach

- Bayesian 101:
  Prior Distribution
  Likelihood Function
  Posterior Distribution
- Advantage: Inferences are conditional on the sample data
- Main hurdles with survey data:
  - Specification of likelihood
  - Specification of prior distribution
  - Validity of posterior inference under design-based framework

Non-Parametric Likelihood

- Parameter vector $\tilde{\mathbf{y}} = (\tilde{y}_1, \cdots, \tilde{y}_N)'$; labels i
  Sample data: $\{(i, y_i),\ i \in \mathbf{S}\}$ minimal sufficient
- The flat Godambe likelihood function $L(\tilde{\mathbf{y}})$:
  All possible unobserved $\tilde{y}_i$ have the same

$$L(\tilde{\mathbf{y}}) = P(y_i, i \in \mathbf{S} \mid \tilde{\mathbf{y}}) = \begin{cases} p(\mathbf{S}) & \text{if } \ y_i = \tilde{y}_i \ \text{ for } \ i \in \mathbf{S}\,, \\ 0 & \text{otherwise}\,. \end{cases}$$

The likelihood is **uninformative**: all possible non-observed $y_i, i \notin \mathbf{S}$ lead to the same likelihood.

Bayesian EL: IID Case (Lazar, 2003)

- $y_1, \cdots, y_n$ iid with $\theta = E(y_i)$
- The empirical likelihood (Owen, 1988; 2001): $L(\boldsymbol{p}) = \prod_{i=1}^n p_i$
- The empirical log-likelihood for θ: $(\sum_{i=1}^n p_i = 1, \sum_{i=1}^n p_i y_i = \theta)$

$$l(\theta) = -n\log(n) - \sum_{i=1}^n \log\{1 + \lambda(y_i - \theta)\}$$

where the Lagrange multiplier λ is the solution to

$$h(\lambda) = \sum_{i=1}^n \frac{y_i - \theta}{1 + \lambda(y_i - \theta)} = 0$$

- For a chosen prior $g(\theta)$ on θ, the posterior (Lazar, 2003)

$$\pi(\theta|\boldsymbol{y}) \propto \exp\left[\log\{g(\theta)\} - \sum_{i=1}^n \log\{1 + \lambda(y_i - \theta)\}\right]$$

- It DOES NOT work for survey data even under SRSWOR

Simulation: Effect of Design

- BEL_n: Bayesian equal-tail naïve (IID) credible interval
- BEL_d: Bayesian interval based on pseudo-EL
- 1,000 simulation runs, nominal level 95%, $N = 800$

n	n/N		CP	L	U	AL
40	5%	BEL_n	95.5	2.4	2.1	0.83
		BEL_d	95.0	2.7	2.3	0.81
120	15%	BEL_n	96.0	2.2	1.8	0.48
		BEL_d	94.2	2.8	3.0	0.44
240	30%	BEL_n	98.6	0.9	0.5	0.34
		BEL_d	94.5	2.9	2.6	0.28

Bayesian PEL for Public-Use Survey Data: $\theta = \mu_y$

- Let $n^* = n/\hat{a}_{\mathrm{WR}}$, where $\hat{a}_{\mathrm{WR}}$ is computed based on $g_i(\theta) = y_i - \theta$
- The adjusted PEL function based on public-use survey data:

$$l_{\mathrm{WR}}(\boldsymbol{p}) = n^* \sum_{i \in \mathbf{S}} \tilde{w}_i(\mathbf{S}) \log(p_i)$$

- The (log) PEL function for θ:

$$l_{\mathrm{WR}}(\theta) = n^* \sum_{i \in \mathbf{S}} \tilde{w}_i(\mathbf{S}) \log\{\hat{p}_i(\theta)\}$$

where

$$\hat{p}_i(\theta) = \frac{\tilde{w}_i(\mathbf{S})}{1 + \lambda(y_i - \theta)}$$

and λ solves

$$\sum_{i \in \mathbf{S}} \frac{\tilde{w}_i(\mathbf{S})(y_i - \theta)}{1 + \lambda(y_i - \theta)} = 0$$

Bayesian PEL for Public-Use Survey Data: $\theta = \mu_y$

- The likelihood function: $L_{\mathrm{WR}}(\theta) = \exp\{l_{\mathrm{WR}}(\theta)\}$
- Noninformative prior on θ: $g(\theta) \propto 1$
- Posterior distribution of θ is given by

$$\pi(\theta \mid \mathbf{S}) = c(\mathbf{S}) \exp\Big\{ -n^* \sum_{i \in \mathbf{S}} \tilde{w}_i(\mathbf{S}) \log[1 + \lambda(y_i - \theta)] \Big\}$$

- The posterior distribution $\pi(\theta \mid \mathbf{S})$ is asymptotically normal
- The posterior mean matches the design-based estimator of μ_y
- The posterior variance matches the design-based variance of $\hat{\mu}_y$
- Posterior inferences are valid under the design-based framework

Bayesian PEL based on $(p_1, \cdots, p_n)$

- Treating $(p_1, \cdots, p_n)$ as general parameters
- The pseudo empirical likelihood function for $(p_1, \cdots, p_n)$:

$$L_{\mathrm{WR}}(\boldsymbol{p}) = \exp\{l_{\mathrm{WR}}(\boldsymbol{p})\} = \prod_{i \in \mathbf{S}} p_i^{\gamma_i},$$

  where $\gamma_i = n^* \tilde{w}_i(\mathbf{S})$ and n^* depends on the definition of θ

- With the Haldane Dirichlet prior $\pi(\boldsymbol{p}) \propto \prod p_i^{-1}$, the posterior distribution of $(p_1, \cdots, p_n)$ is also Dirichlet:

$$\pi(p_1, \cdots, p_n \mid \mathbf{S}) \propto \prod_{i=1}^{n} p_i^{\gamma_i - 1}$$

The Posterior Distribution of $(p_1, \cdots, p_n)$

- With the Haldane diffuse prior, the posterior distribution is Dirichlet

$$(p_1, \cdots, p_n) \mid \mathbf{S} \;\sim\; D(\gamma_1, \cdots, \gamma_n)$$

- Simulation-based approach:
  - $X_i \;\sim\; f_i(x) = [\Gamma(\gamma_i)]^{-1} x^{\gamma_i - 1} \exp\{-x\}$
  - $X_1, \cdots, X_n$ are independent
  - Let $p_i = X_i / \sum_{i=1}^n X_i$, $i = 1, \cdots, n$. Then

$$(p_1, \cdots, p_n) \;\sim\; D(\gamma_1, \cdots, \gamma_n)$$

- Bayesian bootstrap (simulation-based) approximation to the posterior distribution of θ defined through $\sum_{i=1}^N g_i(\theta) = 0$:

$$\sum_{i \in \mathbf{S}} p_i g_i(\theta) = 0 \quad \longrightarrow \quad \theta = H(p_1, \ldots, p_n \mid \mathbf{S})$$

Problems and Research in Progress ...

- Replication weights satisfy Assumption 2: Current practice
- Methods for creating replication weights to satisfy Assumption 3
- Methods for creating replication weights under imputation for item nonresponses: An important and wide open research area
- Bayesian EL inference with public-use survey data: under further investigation
- EL-based confidence intervals for quantiles and inequality measures with public-use survey data: in progress
- PEL for public-use survey data with vector parameters: under investigation (CANSSI CRT: Zhao, Haziza and Wu)
- Variable selection and regression modelling with public-use survey data: in progress (CANSSI CRT: Zhao, Haziza and Wu)

References

- Wu, C. and Rao, J. N. K. (2006). Pseudo-empirical likelihood ratio confidence intervals for complex surveys. *The Canadian Journal of Statistics* 34, 359–375.
- Berger, Y. G. and De La Riva Torres, O. (2016). Empirical likelihood confidence intervals for complex sampling designs. *Journal of Royal Statistical Society*, Ser. B 78, to appear.
- Rao, J. N. K. and Wu, C. (2016). Empirical likelihood inference with public-use survey data. Under review by *Biometrika*.
- Rao, J. N. K. and Wu, C. (2010). Bayesian pseudo empirical likelihood intervals for complex surveys. *Journal of the Royal Statistical Society*, Ser. B, 72, 533–544.

Acknowledgments: Research Supported By

- Grants from the Natural Sciences and Engineering Research Council (NSERC) of Canada
- Collaborative Research Team (CRT) grant from the Canadian Statistical Sciences Institute (CANSSI)