Complex Survey Variance and Design Effects in R
using the Rstan and Survey packages

Matt Williams\textsuperscript{1}   Terrance Savitsky\textsuperscript{2}

\textsuperscript{1}National Center for Science and Engineering Statistics
National Science Foundation
mrwillia@nsf.gov

\textsuperscript{2}Office of Survey Methods Research
Bureau of Labor Statistics
Savitsky.Terrance@bls.gov

GASP
Sept 23, 2019
Population Inference from Complex Survey Samples

- **Goal**: perform inference about a finite population generated from an unknown model, \( P_{\theta_0} \).

- **Data**: from under a complex sampling design distribution, \( P_{\nu} \)
  - Probabilities of inclusion \( \pi_i \) are often associated with the variable of interest (purposefully)
  - Sampling designs are “informative”: the balance of information in the sample ≠ balance in the population.

- **Biased Estimation**: estimate \( P_{\theta_0} \) without accounting for \( P_{\nu} \).
  - Use inverse probability weights \( w_i = 1/\pi_i \) to mitigate bias.

- **Incorrect Uncertainty Quantification**:
  - Failure to account for dependence induced by \( P_{\nu} \) leads to standard errors and confidence intervals that are the wrong size.
Variance Estimation

- The de-facto approach:
  - approximate sampling independence of the primary sampling units (Heeringa et al. 2010).
  - within-cluster dependence treated as nuisance

- Two common methods:
  - Taylor linearization and replication based methods.
Taylor Linearization

Let $y_{ij}$, $X_{ij}$, and $w_{ij}$ be the observed data for individual $i$ in cluster $j$ of the sample. Assume the parameter $\theta$ is a vector of dimension $d$ with population model value $\theta_0$.

1. **Approximate** an estimate $\hat{\theta}$, or a ‘residual’ $(\hat{\theta} - \theta_0)$, as a weighted sum: $\hat{\theta} \approx \sum_{i,j} w_{ij} z_{ij}(\theta)$ where $z_{ij}$ is a function evaluated at the current values of $y_{ij}$, $X_{ij}$, and $\hat{\theta}$.

2. **Compute** the weighted components for each cluster (e.g., primary sampling units (PSUs)): $\hat{\theta}_j = \sum_i w_{ij} z_{ij}(\theta)$.

3. **Compute** the variance between clusters:

$$Var(\hat{\theta}) = \frac{1}{J-d} \sum_{j=1}^{J} (\hat{\theta} - \hat{\theta}_j)(\hat{\theta} - \hat{\theta}_j)^T$$

4. For stratified designs, compute $\hat{\theta}_s$ and $\overbrace{Var(\hat{\theta}_s)}$ within strata and sum

$$Var(\hat{\theta}) = \sum_s \overbrace{Var(\hat{\theta}_s)}.$$
Replication

Let $y_{ij}$, $X_{ij}$, and $w_{ij}$ be the observed data for individual $i$ in cluster $j$ of the sample. Assume the parameter $\theta$ is a vector of dimension $d$ with population model value $\theta_0$.

1. Through randomization (bootstrap), leave-one-out (jackknife), or orthogonal contrasts (balanced repeated replicates), create a set of $K$ replicate weights $(w_i)_k$ for all $i \in S$ and for every $k = 1, \ldots, K$.

2. Each set of weights has a modified value (usually 0) for a subset of clusters, and typically has a weight adjustment to the other clusters to compensate: $\sum_{i \in S}(w_i)_k = \sum_{i \in S}w_i$ for every $k$.

3. Estimate $\hat{\theta}_k$ for each replicate $k \in 1, \ldots, K$.

4. Compute the variance between replicates:
\[
\hat{\text{Var}}(\hat{\theta}) = \frac{1}{K-d} \sum_{k=1}^{K}(\hat{\theta} - \hat{\theta}_k)(\hat{\theta} - \hat{\theta}_k)^T.
\]

5. For stratified designs, generate replicates such that each strata is represented in every replicate.
Challenges

There are two notable trade-offs associated with these methods:

- **Taylor linearization**: value $\hat{\theta}$ computed once then used in a plug in for $z_i(\theta)$.
  - Replication methods: estimate $\hat{\theta}_k$ computed $K$ times.
  - Sizable differences in computational effort
- **Replication methods**: no derivatives are needed.
  - Taylor linearization: requires the calculation of a gradient to derive the analytical form of the first order approximation $z_i(\theta)$.
  - This poses significant analytical challenges for all but the simplest models.
Some Improvements

- **Abstraction of Derivatives** (less analytic work for Taylor Linearization)
  - Recent advances in *algorithmic differentiation* (Margossian 2018), allows us to specify the model as a log density but only treat the gradient in the abstract *without* specifying it analytically.
  - Implemented in *Stan* and *Rstan* (Carpenter 2015, Stan Development Team 2016)

- **Hybrid Approach** or Taylor Linearization for replicate designs (less computation for Replication approaches)
  - Survey package (Lumley 2016) to calculate replication variance of gradient
  - Use plug in for $\theta$, only estimate *once*
Example: Design Effect for Survey-Weighted Bayes

- Pseudo posterior $\propto$ Pseudo Likelihood $\times$ Prior

\[
\pi^\pi (\lambda|y, \tilde{w}) \propto \prod_{i=1}^{n} \pi (y_i|\lambda)^{\tilde{w}_i} \pi (\lambda)
\]

- Variances Differ:
  - Weighted MLE: $H^{-1}_{\theta_0} J^\pi_{\theta_0} H^{-1}_{\theta_0}$ (Robust)
  - Weighted Posterior: $H^{-1}_{\theta_0}$ (Model-Based)

- Adjust for Design Effect: $R_2^{-1} R_1$
  - $\hat{\theta}_m \equiv$ sample pseudo posterior for $m = 1, \ldots, M$ draws with mean $\bar{\theta}$
  - $\hat{\theta}_m^a = \left( \hat{\theta}_m - \bar{\theta} \right) R_2^{-1} R_1 + \bar{\theta}$
  - where $R'_1 R_1 = H^{-1}_{\theta_0} J^\pi_{\theta_0} H^{-1}_{\theta_0}$
  - $R'_2 R_2 = H^{-1}_{\theta_0}$
**R Code Schematic**

- **Input** → **R Code** → **Output**

Stan Model

- `sampling (rstan)` → $\hat{\theta}_m$ → $\hat{H}_\theta$

Survey Design

- $\hat{\theta}_m$
- `svrepdesign (survey)` → `withReplicates (survey)` → $\hat{J}_\theta$
- `aaply (plyr)` → $\hat{\theta}_m$


**URL**: http://arxiv.org/abs/1811.05031


**URL**: http://mc-stan.org/