AN EMPIRICAL LIKELIHOOD BASED ESTIMATOR FOR RESPONDENT DRIVEN SAMPLED DATA

Sanjay Chaudhuri
Department of Statistics and Applied Probability,
National University of Singapore

Mark Handcock, Department of Statistics, University of California, Los Angeles
In classical survey sampling the estimators of population means and totals usually use unit-wise inclusion probabilities.

Pairwise inclusion probabilities, when available, are mostly used to estimate the standard errors.

The likelihood based estimators, however, often benefit from correct specification of dependency.

In this talk we inspect a new empirical likelihood based procedure to estimate population parameters from data collected through an complex sampling scheme.

The proposed procedure depends on the pair-wise inclusion probabilities directly and works well for dependent data.
A MOTIVATING EXAMPLE

• There were 4600 counties in United States of America in 2004.

• The county-wise total of cast votes and those in favour of the three presidential candidates namely John Kerry, George W. Bush and Ralph Nader are available.

• Suppose we want to estimate the proportion of counties John Kerry won in the election. The true value is known to be 0.3276. However, we want to estimate this number by drawing samples of size $n = 40$ from the population.

• Usually more populous counties in the urban areas vote democrat, so it natural to consider PPS sampling schemes where sampling probabilities $\pi_i$ are proportional to the total votes cast.

• Three PPS schemes are considered, (i) Tillé Sampling, (ii) Midzuno sampling and (iii) Systematic sampling.

• These three schemes would sample units with various degrees of pairwise dependence $\pi_{ij}$ between units $i$ and $j$. 
Suppose $y_i$ is an indicator function that the $i$th county was won by John Kerry.

We consider three estimators of the population proportion. These are:

- Sample mean $= \frac{1}{n} \sum_{i=1}^{n} y_i$, which ignores the unequal probability sampling.

- Hájek estimator $= \frac{\sum_{i=1}^{n} y_i \pi_i^{-1}}{\sum_{i=1}^{n} \pi_i^{-1}}$, which only involves the first order selection probabilities.

- We propose a new estimator which uses pairwise selection probabilities for estimating the proportion. The estimator is given by:

  $$\text{Proposed estimator} = \frac{1}{2} \frac{\sum_{i=1}^{n} \sum_{j=i+1}^{n} (y_i + y_j) \pi_{ij}^{-1}}{\sum_{i=1}^{n} \sum_{j=i+1}^{n} \pi_{ij}^{-1}}.$$

- We first compare the performances of above three estimators for three PPS sampling schemes by directly sampling from the population. The results presented below are based on 10,000 samples.
RESULTS

### Tillé

<table>
<thead>
<tr>
<th>PPS Sampling Scheme</th>
<th>Sample Mean</th>
<th>Hajek Estimator</th>
<th>Proposed Estimator</th>
</tr>
</thead>
<tbody>
<tr>
<td>dep=0.002</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tillé</td>
<td>0.017</td>
<td>0.017</td>
<td>0.017</td>
</tr>
<tr>
<td>dep=0.010</td>
<td>0.016</td>
<td>0.016</td>
<td>0.016</td>
</tr>
<tr>
<td>Systematic</td>
<td>0.016</td>
<td>0.016</td>
<td>0.016</td>
</tr>
</tbody>
</table>

### Midzuno

<table>
<thead>
<tr>
<th>PPS Sampling Scheme</th>
<th>Sample Mean</th>
<th>Hajek Estimator</th>
<th>Proposed Estimator</th>
</tr>
</thead>
<tbody>
<tr>
<td>dep=0.010</td>
<td>0.017</td>
<td>0.017</td>
<td>0.017</td>
</tr>
</tbody>
</table>

### Systematic

<table>
<thead>
<tr>
<th>PPS Sampling Scheme</th>
<th>Sample Mean</th>
<th>Hajek Estimator</th>
<th>Proposed Estimator</th>
</tr>
</thead>
<tbody>
<tr>
<td>dep=130.79</td>
<td>0.016</td>
<td>0.016</td>
<td>0.016</td>
</tr>
</tbody>
</table>
**Observations**

- In terms of mse, the sample mean, which ignores the PPS, surprisingly performs quite well. It is very biased, but has the smallest variance.

- The Hájek estimator actually performs worst, in terms of the mse. It has the lowest bias for all sampling schemes. But its variance is the highest.

- The proposed estimator does well for the systematic sampling. In fact it has the smallest mse overall, even though it is more biased than the Hájek estimator.

- From the histograms, it is seen that for Tillé and Midzuno schemes, the proposed estimator is quite close to the Hájek estimator. However, for systematic sampling with more dependence, the performance of the former improves. Even though its absolute bias increases, there is a marked accumulation near the true value.

- In summary, when the sampling is dependent the proposed estimator may be better. On the other hand, for independent or near independent sampling it probably won’t be worse than the Hájek estimator.

- In what follows we provide an empirical likelihood based justification of our proposed estimator.
Towards a Composite Likelihood Based Method (Setup)

- We discuss a general construction of a composite empirical likelihood below.
- We need some structural assumptions on the sampling design and the scheme.

Variables in $X$ are not in the design. However, $Y$ and $X$ may depend on some variables in $Z$. $A$ has all the explanatory variables in the model. Further, let $V = \{Y\} \cup X \cup Z$.

- Consider a “super population” with response $Y$, auxiliary variables $X = \{X^{(1)}, X^{(2)}, \ldots, X^{(p)}\}$ and design variables $D$.
- The population $\mathcal{P}$ is an i.i.d. sample of size $N$ from above.
- A random sample $S$ of $n$ observations is drawn from $\mathcal{P}$ according to a design depending on $D$.
- The data does not have all of $D$, a subset $Z = \{Z^{(1)}, Z^{(2)}, \ldots, Z^{(m)}\}$ is supplied.

Super population: $Y, X, D$

$N$ i.i.d. draws.

Finite population: $Y, X, D$

Sample of size $n$ from design.

Sample: $Y, X, Z \subseteq D$
For $S \subseteq \mathcal{P}$ suppose $I_S$ is the random indicator function for $S \subseteq S$.

The sample $S$ is the unique largest subset $S$ of $\mathcal{P}$ such that $I_S = 1$.

If the sample units are drawn according to a design, the sampling mechanism may not be ignorable.

That is, the observed distribution of $V$ in the sample $S$ may be different from its distribution in the population and may depend on the particular sample selected.

For any $S \subseteq \mathcal{P}$, the design specifies the conditional probability of $I_S = 1$, given $D_P$.

Suppose $\pi_S = Pr_{\mathcal{P}} (I_S = 1 \mid D_P)$, where $Pr_{\mathcal{P}} (\cdot)$ is the probability under the population.

Notice that, $\pi_S$ is a random variable because of $D_P$.

Though a bit controversial, in practice this assumption is un-avoidable. It also facilitates analysis, which we shall see at later.
Basic assumptions on the design and sample selection

• Basic assumptions:
  For all $S \subseteq \mathcal{P}$, under the population distribution we assume
  1. $\pi_S \perp \perp (Y_P, X_P) | D_P$.
  2. $I_S \perp \perp (X_P, Y_P, D_P) | \pi_S$.

Implications of the basic assumptions:

• Assumption 1. $\Leftrightarrow \pi_S = \pi(S, D_P)$. Note: $\pi$ is user specified.

• $I_S \perp D_P | \pi_S$, $\forall S$. Selection ignorability (Sugden and Smith [1984]).

• $I_S \perp (X_P, Y_P) | D_P$. Basic design assumption of Scott[1977].

• $Pr_P [I_S = 1 | V_S] = E_P [\pi_S | V_S]$.

• $Pr_P [I_S = 1, V_S] = Pr_P [I_S = 1 | V_S] Pr_P [V_S] = E_P [\pi_S | V_S] Pr_P [V_S]$. 
THE LIKELIHOOD

- Suppose \( i_S = \{(i, j) : i < j, i, j \in S\} \).

- Following Pfeffermann et. al., for any \((i, j) \in i_S\) define:

\[
dF^{(S)}_{(i,j)} = \frac{Pr_P[(i, j) \in i_S | V_i, V_j] \; dF^{(P)}_{(ij)}}{Pr_P[(i, j) \in i_S]}
\]

where \( F^{(P)} \) denotes the unknown population distribution.

- Under our assumption we get:

\[
dF^{(S)}_{(i,j)} = \frac{E_P[\pi_{ij} | V_i, V_j] \; dF^{(P)}_{(ij)}}{\int E_P[\pi_{ij} | V_i, V_j] \; dF^{(P)}_{(ij)}}.
\]

- Following Pfeffermann et. al. again, we can construct a composite likelihood of the sample by multiplying \( dF^{(S)}_{(i,j)} \) over the set \( i_S \). The likelihood is given by:

\[
\prod_{(i,j) \in i_S} dF^{(S)}_{(i,j)} = \prod_{(i,j) \in i_S} \frac{E_P[\pi_{ij} | V_i, V_j] \; dF^{(P)}_{(ij)}}{\int E_P[\pi_{ij} | V_i, V_j] \; dF^{(P)}_{(ij)}}.
\]
**Empirical Likelihood**

- Suppose \( \nu_{ij} = E_P[\pi_{ij} | V_i, V_j] \).

- Clearly \( \int E_P[\pi_{ij} | V_i, V_j] \, dF^{(P)}_{ij} = \int \nu_{ij} \, dF^{(P)}_{ij} = E_P[\pi_{ij}] \).

- Now if we further assume (cf. Godambe, Hartley Rao) \( E_P[\pi_{ij}] = \gamma \), for all \((i, j) \in i_S\), we can write an empirical likelihood as

\[
L(w, \nu) = \frac{\prod_{(i,j) \in i_s} \nu_{ij}w_{ij}}{\left\{ \sum_{(i,j) \in i_s} \nu_{ij}w_{ij} \right\}^{n/2}}.
\]

- Here we estimate the true \( F^{(P)} \) using empirical likelihood. The unknown weights \( w_{ij} \) are the jumps of a distribution \( F \) at the point \((v_i, v_j)\). We estimate \( \gamma \) by \( \sum_{(i,j) \in i_s} \nu_{ij}w_{ij} \).

- This likelihood is same as the Vardi’s likelihood. We are using the pairwise inclusion probabilities.
ESTIMATING THE MEAN

- In order to estimate $\mu_0$, we maximise $L(w, \nu)$ over the set:

$$\mathcal{W} = \bigcup_{\mu \in \mathbb{R}} \left\{ w : w \in \Delta_{(n^2)-1}, \sum_{(i,j) \in i_s} w_{ij} \left\{ (y_i - \mu) + (y_j - \mu) \right\} = 0 \right\},$$

Where $\Delta_{(n^2)-1}$ is the $(N^2-1)$ dimensional simplex.

- Our maximal empirical likelihood estimator of $\mu_0$ is given by:

$$\hat{\mu}^{(n)}_E = \arg \max_{w \in \mathcal{W}} \{ L(w, \nu) \}.$$

- It follows that under our setup,

1. The composite likelihood $L(w, \nu)$ has a unique maximum in $\mathcal{W}$ at

$$\hat{w}_{ij} = \nu^{-1}_{ij} / \sum_{(i,j) \in i_s} \nu^{-1}_{ij}.$$

2. Corresponding estimator $\hat{\mu}^{(n)}_E$ of $\mu_0$ is unique and is given by:

$$\hat{\mu}^{(n)}_E = \frac{1}{2} \cdot \sum_{(i,j) \in i_s} (y_i + y_j) \nu^{-1}_{ij} / \sum_{(i,j) \in i_s} \nu^{-1}_{ij}.$$
ASYMPTOTIC PROPERTIES

• Our estimator is based on non-degenerate U-statistics. So its asymptotic properties can be easily obtained.

• In particular, one can show that under the population distribution, where variables are well behaved and some regularity conditions:

1. \( \hat{\mu}_E^{(N)} \) is strongly consistent for \( \mu_0 \) under the distribution in the population.

2. Suppose \( U_1 = E_P \left[ (Y_1 + Y_2 - 2\mu_0) \nu_{12}^{-1} \middle| V_1 = v_1 \right] \). Then \( \sqrt{N} \left( \hat{\mu}_E^{(N)} - \mu_0 \right) \) converges in distribution to a \( N \left( 0, \sigma^2 \right) \) variable, with \( \sigma^2 = Var_P [U_1] \left\{ E_P [\nu_{12}^{-1}] \right\}^{-2} \).
PREDICTING FINITE POPULATION MEAN

- Our predictor $\hat{\mu}_N$ for the finite population mean is given by $\hat{\mu}_E^{(n)}$.

- The variance of $\hat{\mu}_N$ with a finite population correction is given by

$$V_{ar_P}[\hat{\mu}_N] = \left(1 - \frac{n - 1}{N - 1}\right) \frac{\sigma^2}{n},$$

from which the standard errors can be computed.

- Using $\hat{w}$ and $\hat{\mu}_E^{(n)}$, we can estimate $\sigma^2/n$ from the data.

- The first estimator which assumes that $V_{ar_P}[U_1] = E_P[U_1^2]$ is given by

$$V_{ar_P}^{(1)}[\hat{\mu}_N] = \left(1 - \frac{n - 1}{N - 1}\right) \frac{\sigma^2}{n} \frac{\sum_{i=1}^{n} \left\{ \sum_{j=1, j \neq i}^{n} (y_i + y_j - 2\hat{\mu}_E^{(n)}) \nu_{ij} \right\}^2}{\left\{ \sum_{i=1}^{n} \sum_{j=1, j \neq i}^{n} \nu_{ij} \right\}^2}.$$

- Alternatively, one can compute the actual variance of $U_1$ from the data and get an estimate $\hat{V}_{ar_P}^{(2)}[\hat{\mu}_N]$ of $V_{ar_P}[\hat{\mu}_N]$. 
RESULTS FOR THE ELECTION DATA

- As an illustration, we estimate the standard errors and the coverages of 95% confidence intervals for the proposed estimator on the election data we used before. The estimated standard errors are compared with the those obtained by directly sampling from the population as shown in the previous table.

- These results are based on 1000 samples each of size 40. We use the Hartley-Rao estimator of the variance of the Hájek estimator. The Yule-Grundy estimator turns out to be negative for Systematic sampling.

<table>
<thead>
<tr>
<th>Scheme</th>
<th>Proposed Estimator</th>
<th>Hájek Estimator</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\sqrt{Var}$</td>
<td>$\sqrt{Var_P^{(1)}}$</td>
</tr>
<tr>
<td>Tillé</td>
<td>0.155</td>
<td>0.151</td>
</tr>
<tr>
<td>Midzuno</td>
<td>0.161</td>
<td>0.150</td>
</tr>
<tr>
<td>Systematic</td>
<td>0.110</td>
<td>0.139</td>
</tr>
</tbody>
</table>

- From the table it seems that the proposed estimator of the standard error specially $\sqrt{Var_P^{(1)}}$ is quite close to the variance obtained from direct sampling from the population.

- The coverage seems to be good, specially for Systematic sampling.
ESTIMATING STANDARD ERROR WITH DEPENDENCE

- In finite samples $y_i$ and $y_j$ are dependent. The variance formula is not very accurate.

- We use an ad hoc general estimator of the standard error.

- Suppose we define:

$$\hat{w}_{ji} = \hat{w}_{ij}, \text{ for } j < i, \quad \hat{w}_i = \frac{\sum_{j=1, j\neq i}^{n} \hat{w}_{ij}}{\sum_{i=1}^{n} \sum_{j=1, j\neq i}^{n} \hat{w}_{ij}},$$

$$\bar{U}_i = \sum_{j=1, j\neq i}^{n} \frac{\hat{w}_{ij}}{\hat{w}_i} \cdot \frac{y_i + y_j - 2\hat{\mu}_E}{\nu_{ij}}.$$

- With these definition the denominator can be estimated by:

$$\hat{E}_P\left[\nu_{12}^{-1}\right] = \sum_{i=1}^{n} \sum_{j=1, j\neq i}^{n} \frac{\hat{w}_{ij}}{\nu_{ij}}.$$

- If the sampled units were selected independently the estimator takes the form:

$$\frac{\text{Var}_P[U_1]}{n} = \sum_{i=1}^{n} \hat{w}_i^2 \bar{U}_i^2 = \sum_{i=1}^{n} \left\{ \sum_{j=1, j\neq i}^{n} \hat{w}_{ij} \cdot \frac{y_i + y_j - 2\hat{\mu}}{\nu_{ij}} \right\}^2.$$
**ESTIMATING STANDARD ERROR WITH DEPENDENCE**

- If the sampled units are dependent, the above estimator requires a correction which depends on the sampling probabilities. Suppose

\[ s = \text{sign} \left( \sum_{i=1}^{n} \sum_{j=1, j \neq i}^{n} (\pi_{ij} - \pi_i \pi_j) \right). \]

The correction term is given by:

\[
\begin{align*}
&= s \sum_{i=1}^{n} \sum_{j=1, j \neq i}^{n} \left(1 - \frac{\pi_i \pi_j}{\pi_{ij}}\right) \hat{w}_i \hat{w}_j \hat{U}_i \hat{U}_j \\
&= s \sum_{i=1}^{n} \sum_{j=1, j \neq i}^{n} \left(1 - \frac{\pi_i \pi_j}{\pi_{ij}}\right) \left\{ \sum_{k=1, k \neq i}^{n} \hat{w}_{ik} \cdot \frac{y_i + y_k - 2\hat{\mu}_E}{\nu_{ik}} \right\} \left\{ \sum_{l=1, l \neq j}^{n} \hat{w}_{jl} \cdot \frac{y_j + y_l - 2\hat{\mu}_E}{\nu_{jl}} \right\}.
\end{align*}
\]

- So finally the variance of \( \hat{\mu} \) is given by:

\[
\begin{align*}
\text{Var}_P [\hat{\mu}_N] &= \left\{ \sum_{i=1}^{n} \hat{w}_i^2 \hat{U}_i^2 + s \sum_{i=1}^{n} \sum_{j=1, j \neq i}^{n} \left(1 - \frac{\pi_i \pi_j}{\pi_{ij}}\right) \hat{w}_i \hat{w}_j \hat{U}_i \hat{U}_j \right\} \left\{ \sum_{i=1}^{n} \sum_{j=1, j \neq i}^{n} \frac{\hat{w}_{ij}}{\nu_{ij}} \right\}^{-2}.
\end{align*}
\]
EXTENSIONS TO REGRESSION MODELS

- There are several possible extensions to our proposed procedure.
- We discuss the one for regression estimation.
- Suppose $V = (Y, A, Z)$, where $Y$ is the response, $A$ the set of auxiliary variables, which may include some design variables and $Z$ is the set of observed design variables not in $A$.
- A regression model is specified by a vector of functions $\psi_\theta (Y, A)$ depending on $Y$, $A$ and a parameter $\theta \in \mathbb{R}^d$, satisfying
  \[ E_P [\psi_\theta (Y_1, A_1)] = 0. \]
- The composite likelihood $L (w, \nu)$ can be used here. However, the unknown $w$ is determined by maximising the likelihood over the set
  \[ \mathcal{W} = \bigcup_{\theta \in \mathbb{R}^d} \left\{ w : w \in \Delta_s - 1, \sum_{(i,j) \in S} w_{ij} \{ \psi_\theta (y_i, a_i) + \psi_\theta (y_j, a_j) \} = 0 \right\}. \]
  As before, the parameter $\theta$ can be estimated by the maximal argument of $L$ over $\mathcal{W}$.
- The rest like the standard errors etc. can be estimated similarly.
We consider a networked population of high risk HIV positive persons in Colorado Springs in 1990. The network of social ties were collected by the El Paso dept. public health (see Poteratt et. al. 2004).

- The network has 2587 nodes and 10381 edges. Covariate information is available.

- The RDS sample was drawn with 10 randomly selected seeds with probability proportional to their degrees, 2 tickets and was continued till the sample reached size 500.

- In order to find the selection probabilities 50000 such samples were drawn and the proportion of individual pair-wise etc occurrences were computed. The variation in probabilities thus obtained was about 3-4%.

- We are interested in estimating the proportion of non-white members in the network. The actual proportion is known to be .2810205.
• Some results:

<table>
<thead>
<tr>
<th>Estimator</th>
<th>bias</th>
<th>se</th>
<th>rmse</th>
<th>efficiency</th>
<th>se</th>
</tr>
</thead>
<tbody>
<tr>
<td>SS</td>
<td>0.0063</td>
<td>0.09648</td>
<td>0.0967</td>
<td>1.000</td>
<td>0.0965</td>
</tr>
<tr>
<td>Hajek</td>
<td>-0.0007</td>
<td>0.09287</td>
<td>0.0928</td>
<td>0.923</td>
<td>0.0509</td>
</tr>
<tr>
<td>Proposed</td>
<td>-0.0011</td>
<td>0.08884</td>
<td>0.0889</td>
<td>0.844</td>
<td>0.0882</td>
</tr>
</tbody>
</table>

• The SS estimator (Krista Gile) is the current state of the art. It beats many other traditional estimators.

• Clearly, the proposed estimator is better than SS in terms of both bias and the variance.

• Our variance estimator, obtained using the formula underestimates. The variance of SS is usually determined by bootstrap.
CONCLUSION

- Several extensions of the basic methodology described here are possible.

- Population level constraints not depending on the parameter of interest can be easily included. A two step method of estimation would be useful in that case.

- Using U-statistics, empirical likelihood and pairwise sampling probabilities a new estimator of the population variance can also be considered.

- It is possible to handle degenerate U-statistics.

- The pairwise inclusion probabilities are not often available. In such cases we need to use Hartley-Rao, Overton approximations. Performance of the proposed estimator with approximate pairwise weights needs to be looked at.

- There are several unresolved issues about the finite sample and asymptotic behaviours of the estimator.

- Several applications to respondent driven sampling, snowball sampling, adaptive cluster sampling etc. are possible.