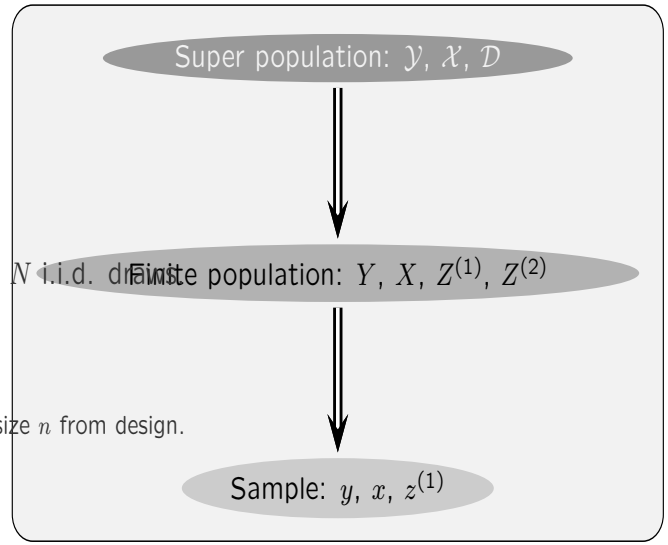
On an Empirical likelihood-based Method for Complex Survey Data with Applications to Non-probability Sampling

Sanjay Chaudhuri

Department of Statistics,
University of Nebraska-Lincoln

SETUP

- We start with some structural assumptions on the sampling design and the scheme.



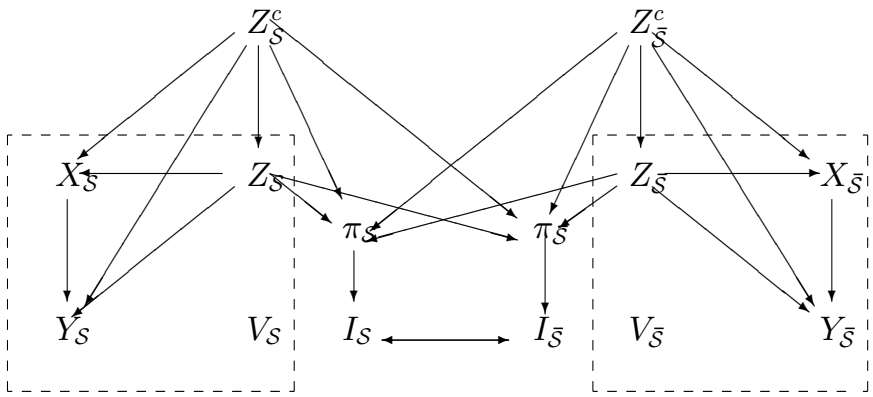
- Consider a "super population" with response Y, auxiliary variables $X = \{X^{(1)}, X^{(2)}, \ldots, X^{(p)}\}$ and design variables D.
- The population $\mathcal{P}$ is an i.i.d. sample of size N from above.
- A random sample $\mathcal{S}$ of n observations is drawn from $\mathcal{P}$ according to 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.

- 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 $V = \{Y\} \cup X \cup Z$.

SAMPLE AND THE DESIGN

- For $S \subseteq \mathcal{P}$ suppose I_S is the random indicator function for $S \subseteq \mathcal{S}$.
- The sample $\mathcal{S}$ 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 $\mathcal{S}$ 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 $D_{\mathcal{P}}$.
- Suppose $\pi_S = Pr_{\mathcal{P}}(I_S = 1 \mid D_{\mathcal{P}})$, where $Pr_{\mathcal{P}}(\cdot)$ is the probability under the population.
- Notice that, π_S is a random variable because of $D_{\mathcal{P}}$.
- 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



Sampling scheme.

- **Basic assumptions:**
  For all $S \subseteq \mathcal{P}$, under the population distribution we assume
  1. $\pi_S \perp\!\!\!\perp (Y_{\mathcal{P}}, X_{\mathcal{P}}) \mid D_{\mathcal{P}}$.
  2. $I_S \perp\!\!\!\perp (X_{\mathcal{P}}, Y_{\mathcal{P}}, D_{\mathcal{P}}) \mid \pi_S$.

Implications of the basic assumptions:

- Assumption 1. $\Leftrightarrow \pi_S = \pi(S, D_{\mathcal{P}})$. **Note:** π is user specified.
- $I_S \perp\!\!\!\perp D_{\mathcal{P}} \mid \pi_S$, $\forall S$. Selection ignorability (Sugden and Smith [1984]).
- $I_S \perp\!\!\!\perp (X_{\mathcal{P}}, Y_{\mathcal{P}}) \mid D_{\mathcal{P}}$. Basic design assumption of Scott[1977].
- $Pr_{\mathcal{P}}[I_S = 1 \mid V_S] = E_{\mathcal{P}}[\pi_S \mid V_S]$.
- $Pr_{\mathcal{P}}[I_S = 1, V_S] = Pr_{\mathcal{P}}[I_S = 1 \mid V_S]\, Pr_{\mathcal{P}}[V_S] = E_{\mathcal{P}}[\pi_S \mid V_S]\, Pr_{\mathcal{P}}[V_S]$.

SOME IMPLICATIONS OF OUR CONSTRUCTION

- From now on we look into the class of sets $S = \{i\}$, $i = 1, 2, \ldots, n$.
- Suppose we denote $\nu_i = E_{\mathcal{P}}[I_i \mid V_i] = E_{\mathcal{P}}[\pi_i \mid V_i]$.
- Note that, because of random sampling, the sum $\mathsf{S}_{i=1}^{n}$ over the sample is a random sum.
- However, for any integrable function $g(V)$ it can be shown that:

$$E_{\mathcal{P}}\left[\mathsf{S}_{i=1}^{n}\frac{g(v_i)}{\nu_i}\right] = E_{\mathcal{P}}\left[\sum_{i=1}^{N} I_i\frac{g(V_i)}{\nu_i}\right] = \sum_{i=1}^{N} E_{\mathcal{P}}\left[E_{\mathcal{P}}\left[I_i\frac{g(V_i)}{\nu_i} \mid V_i\right]\right]$$

$$= \sum_{i=1}^{N} E_{\mathcal{P}}\left[\frac{g(V_i)}{\nu_i}E_{\mathcal{P}}[I_i \mid V_i]\right] = \sum_{i=1}^{N} E_{\mathcal{P}}\left[\frac{g(V_i)}{\nu_i}\nu_i\right] = \sum_{i=1}^{N} E_{\mathcal{P}}[g(V_i)] = N E_{\mathcal{P}}[g(V_1)].$$

SOME MORE RESULTS

- Similarly, for any integrable function $g(V_i, \pi_i)$ we get,

$$E_{\mathcal{P}}\left[\sum_{i=1}^{n} \frac{g(v_i, \pi_i)}{\pi_i}\right] = \sum_{i=1}^{N} E_{\mathcal{P}}\left[E_{\mathcal{P}}\left[I_i \frac{g(V_i, \pi_i)}{\pi_i} \mid V_i, \pi_i\right]\right]$$

$$= \sum_{i=1}^{N} E_{\mathcal{P}}\left[\frac{g(V_i, \pi_i)}{\pi_i} E_{\mathcal{P}}\left[I_i \mid V_i, \pi_i\right]\right] = \sum_{i=1}^{N} E_{\mathcal{P}}\left[\frac{g(V_i, \pi_i)}{\pi_i} E_{\mathcal{P}}\left[I_i \mid \pi_i\right]\right]$$

$$= \sum_{i=1}^{N} E_{\mathcal{P}}\left[g(V_i, \pi_i)\right] = N E_{\mathcal{P}}\left[g(V_1, \pi_1)\right].$$

- This implies if $E_{\mathcal{P}}\left[g(V_1)\right] = 0$, and $E_{\mathcal{P}}\left[g(V_1, \pi_1)\right] = 0$, then both $E_{\mathcal{P}}\left[\sum_{i=1}^{n} g(v_i)/\nu_i\right] = 0$ and $E_{\mathcal{P}}\left[\sum_{i=1}^{n} g(v_i, \pi_i)/\pi_i\right] = 0$ as well.

- From above, it follows that $E_{\mathcal{P}}\left[\sum_{i=1}^{n} 1\right] = N E_{\mathcal{P}}\left[\pi_1\right]$. This implies $E_{\mathcal{P}}\left[\pi_1\right] = n/N$.

- Similarly, $E_{\mathcal{P}}\left[\sum_{i=1}^{n} \pi^{-1}\right] = N$. That is, observed sum of the inverse sampling weights is unbiased for the population size.

AN NEW WAY TO INCLUDE SAMPLING PROBABILITIES

- Note that, under our assumptions,

$$E_{\mathcal{P}}[\pi] = n/N.$$

- Provided The popul;ation size is known, we can use $(\pi_i - n/N)$ as constraint in determining the EL weights.

- The optimal weights $\widehat{w}$ are then used to find the maximum empirical likelihood parameter estimate .

- In particular, one seeks the solution of the equation

$$\sum_{i=1}^{n} \widehat{w}_i \psi(v_i, \theta) = 0$$

in θ, where ψ is a user-specified estimation function for the parameter θ.

- This is a two-step estimator.

FINDING THE OPTIMAL EL WEIGHTS

- Let

$$\mathcal{W} = \left\{ w \ : \ \sum_{i=1}^{n} w_i \left(\pi_i - \frac{n}{N} \right) = 0 \right\} \cap \Delta_{n-1}$$

- Two different empirical likelihoods can be constructed here.

- **Likelihood 1.** We forget the unequal sampling, and compute

$$\widehat{w}^{(1)} = \arg \max_{\{w \in W\}} \left\{ \sum_{i=1}^{n} \log w_i \right\}.$$

- It can be easily shown that:

$$\widehat{w}_i^{(1)} = \frac{1}{n} \cdot \frac{1}{1 + \lambda(\pi - \frac{n}{N})},$$

  where λ is the Lagrange multiplier.

LIKELIHOOD 2.

- Alternatively, we can follow Pfeffermann et. al. and maximise an empirical likelihood based version of their sample likelihood.

- Our setup justifies the form of their proposed conditional likelihood of the observations given their selection in the sample.

- The empirical likelihood version is the Vardi's likelihood given by:

$$\widehat{w}^{(2)} = \arg \max_{w \in \mathcal{W}} \left\{ \sum_{i=1}^{n} log(\pi_i w_i) - nlog\left(\sum_{i=1}^{n} \pi_i w_i \right) \right\},$$

- It can be shown that $\widehat{w}^{(2)}$ satisfy:

$$\widehat{w}_i^{(2)} = \frac{\sum_{i=1}^{n} \pi_i \widehat{w}_i^{(2)}}{n} \cdot \frac{1}{\pi_i + \kappa(\pi_i - \frac{n}{N})}.$$

- An interesting property of the proposed constraint is that, one can show $\widehat{w}^{(1)} = \widehat{w}^{(2)}$.

MORE COMMENTS

- As is evident the proposed estimator is quite demanding.
- It is not invariant to the scale of π. One needs to have the sampling probabilities in their original scale.
- Even though might be more readily available, it also requires N.
- The estimate of the population size obtained by maximising the likelihood over N produce biased estimate.
- Any estimate of the population size can be used. The procedure will provide different estimates of the weights for different estimate of the population size.
- When the true population size is known, it is closer in spirit to the Horvitz-Thomson estimator of the mean.
- However, unlike the latter it always remains in range. That is, for example, mean of a 0-1 binary observation will always be in $[0, 1]$.

SOME MORE COMMENTS

- If we use the estimate $\widehat{N} = \sum_{i=1}^{n} \pi_i^{-1}$, the Hajek estimate of the parameters will be obtained.
- It should be easy to compute the standard errors of the parameters.
- Finally, note that, instead of π, ν can also be used.

PERFORMANCE IN MEAN ESTIMATION

- We consider county-wise voting record from 2004 US presidential election.
- The goal is to estimate county-wise average vote cast for John Kerry.

Estimator	mean	sd	Est. sd
pop mean	12213.95	742.98	
samp Mean	168686.73	48754.94	
Hajek	15042.45	6605.63	4994.42
HT-mean	12210.55	916.62	
new.est	12109.76	932.54	513.75
new.est.fitted	11911.25	1020.41	

- In order to follow a super-population model, new finite population of size 4600 was created by bootstrapping the given population. An independent pps sample of size 40 was drawn from the bootstrapped population using the Tille's method.

- The sampling probabilities were proportional to the total number of votes cast.

- The results in the table above are from 1000 replication of the above procedure.

- The ν for new.est.fitted were obtained using Gamma regression.

A TENTATIVE APPLICATION TO NON-PROBABILITY SAMPLING

- This is an ongoing work.
- When the sampling probabilities are not known, the idea is to use the constraint above to estimate the sampling probabilities, empirical likelihood weights and possibly some parameters.
- Let x be a vector of covariates believed to be related to the design variables.
- We set

  $$\nu_i = f(x_i\beta)$$

  for some known function f with unknown parameters β.
- With N known, the basic idea is to solve:

$$\max_{w\in\mathcal{W},\beta} \prod_{i=1}^{n} w_i, \text{ where } \mathcal{W} = \left\{ w \ : \ \sum_{i=1}^{n} w_i \left(f(x_i\beta) - \frac{n}{N} \right) = 0 \right\} \cap \Delta_{n-1}.$$

- Once $\widehat{\nu}$ is known, we can use the predicted weights in a Hajek estimator.

A TENTATIVE APPLICATION TO NON-PROBABILITY SAMPLING

- We know more about the sampling probabilities which translates to more constraints.
- first of all, since the unit was selected in the sample, each $\widehat{\nu}_i$ should be strictly greater than zero.
- Furthermore, since the units are selected with unequal probability, the variance of $\widehat{\nu}_i$ should not be very small.
- Taking these two information in mind, we solve the problem:

$$\max_{w \in \mathcal{W}', \beta, \sigma^2} \prod_{i=1}^{n} w_i$$

where

$$\mathcal{W}' = \Bigg\{ w \ : \ \sum_{i=1}^{n} w_i \left(f(x_i\beta) - \frac{n}{N} \right) = 0, \sum_{i=1}^{n} w_i \left\{ \left(f(x_i\beta) - \frac{n}{N} \right)^2 - \sigma^2 \right\} = 0, \\ f(x_i, \beta) \geq \epsilon, \forall \ i = 1, 2, \ldots, n \Bigg\} \cap \Delta_{n-1}.$$

MODEL

- Clearly, a basic model would be:

$$logit(\nu) = x\beta. \text{ or equivalently } \nu = (1 + exp(-x\beta))^{-1}.$$

- For computational ease, and more flexibility, we do a bit more:

$$\nu = qbeta\left((1 + exp(-x\beta))^{-1}, a, b\right),$$

  here $qbeta(\cdot, a, b)$ is the quantile function of the beta density with parameters a and b.

- The parameters a and b are chosen such that the expectation is n/N, and the variance is σ^2.

- This formulation follows the formulation of a beta regression.

- Clearly, ν will follow a beta distribution, if for some β and σ^2, $(1 + exp(-x\beta))^{-1}$ is uniformly distributed.

- That is one would expect the optimal w to be close to $(n^{-1}, \ldots, n^{-1})^T$.

SOME RESULTS

Covariates	Sample/Hajek bias	Sample/Hajek rmse	Hajek bias	Hajek rmse	Proposed/Hajek bias	Proposed/Hajek rmse
Intercept+Total Votes	71.56	12.85	1617.1	4667.02	6.22	2.64
Intercept+Votes for Kerry	71.56	12.85	1617.1	4667.02	4.84	2.37
Intercept+Votes for Bush	71.56	12.85	1617.1	4667.02	2.43	1.26
I+Kerry+Total	71.56	12.85	1617.1	4667.02	8.59	19.14
I+Bush+Total	71.56	12.85	1617.1	4667.02	8.07	6.50

- The results in the table are for illustration purposes, and are based on 1000 replications on a sample size of 135.
- The true sampling probabilities are proportional to total votes cast.
- We used $\epsilon = 10^{(-5)}$.
- In terms of rmse the proposed estimator might be better than the Hajek estimator which uses the true sampling probabilities.
- The proposed estimator is less biased than the sample mean.

A PENALISED LIKELIHOOD

- Even though method described above shows promise. A penalised likelihood with direct involvement of the estimated probabilities might be better.

- We define the following penalised negative log-likelihood:

$$l(\beta, \sigma^2) = -2 \max_{w \in \mathcal{W}} \left\{ \sum_{i=1}^{n} \log(n w_i) \right\} - 2 \sum_{i=1}^{n} \log(n p_i) - K \log(n) \log \left(\sigma^2 (- \sum_{i=1}^{n} \nu_i \log(\nu_i)) \right),$$

  where $p_i = \nu_i / \sum_{i=1}^{n} \nu_i$.

- For each value of K we minimise the above negative log-likelihood over β and σ^2.

- The coefficient K is determined by cross-validation.

- In the test set we look for the maximum predicted selection probability.

SOME RESULTS

Covariates	Proposed/Hajek bias	rmse
Intercept+Total Votes	4.92	1.75
Intercept+Votes for Kerry	3.98	1.59
Intercept+Votes for Bush	3.13	1.98

- As before the population mean is estimated by the Hajek estimator with predicted sampling probabilities.

- The above numbers are average of 100 repetitions. Other setups are exactly the same.

- If no reliable estimator of N is available, a Hierarchical Bayesian model can be easily postulated.