

Simulation-Based Systematic PPS Sampling with Unit Substitutions

Changbao Wu

(Joint work with Mary E. Thompson)

Department of Statistics and Actuarial Science
University of Waterloo

November 3, 2006 StatCan Symposium

Outline

1. The ITC-China survey
2. Properties of simulation-based methods
3. Examples: First order inclusion probabilities
4. Second order inclusion probabilities
5. Computational implementation

1. The ITC-China Survey

- The International Tobacco Control (ITC) Policy Evaluation Project for FCTC (Framework Convention for Tobacco Control)
- Participating countries
 - Australia, Canada, United Kingdom, USA (Four countries)
 - Ireland, Thailand, Malaysia
 - Korea, China
 - Still growing

- The ITC-China survey

- A prospective longitudinal study; first wave conducted May-July 2006
- 800 adult smokers and 200 adult non-smokers in each of 7 cities
- Beijing, Shanghai, Guangzhou, Shenyang, Zhenzhou, Changsha and Yinchuan.
- A multistage cluster sampling design
- 45-minute face-to-face interview
- Measures of the demand reduction policies of the FCTC, including labeling, price/taxation, advertising/promotion, smoke-free, cessation, education, etc

- Multistage unequal probability selection of units
 - Ten Street Districts (Jie Dao) within each city, selected using the Randomized Systematic PPS sampling method, with probability proportional to Jie Dao population size
 - Two Residential Blocks (Ju Wei Hui) within each selected Jie Dao, selected using the Randomized Systematic PPS sampling method, with probability proportional to Ju Wei Hui population size
 - 300 Households (HH) within each selected Ju Wei Hui, selected using simple random sampling, and stratified into smoking and non-smoking HHs.
 - Smoking (Non-smoking) HHs are randomly ordered; adult smokers and non-smokers are selected (using the next birthdate method) from the HH list

- Unit Refusals:

Several selected Jie Dao and Ju Wei Hui refuse to participate in the survey

- Unit Substitutions:

Refusing units are substituted by other units, selected from remaining units using the randomized systematic PPS sampling method

To maintain the targeted overall sample size

- The Final Selected Set of Units:

No longer a PPS sample

- Weight Construction:

$$w_{ijkl} = 1/\pi_{ijkl}, \quad \pi_{ijkl} = \pi_{i1}\pi_{j2}\pi_{k3}\pi_{l4}$$

where π_{i1} , π_{j2} , π_{k3} and π_{l4} are inclusion probabilities at each of the four levels of selection

- How to calculate the inclusion probabilities when refusing units are substituted?

⇒ Treating the sampled units as if the inclusion probabilities are STILL proportional to the sizes is NOT correct

2. Simulation-based Methods

- Require that complete design information be available
- Condition on the given set of refusing units
- To estimate $\pi_i = P(i \in s)$ using Monte Carlo simulation:
 - Take K independent simulated samples using the actual sampling procedure
 - Count M_i , the number of samples which include unit i
 - Estimate π_i by $\pi_i^* = M_i/K$
- Estimate $\pi_{ij} = P(i, j \in s)$ by $\pi_{ij}^* = M_{ij}/K$, where M_{ij} is the number of samples which include both i and j

- Result 1: Individual π_i and π_{ij}

- $M_i \sim \text{Binomial}(K, \pi_i)$:

$$E(\pi_i^*) = \pi_i \text{ and } \text{Var}(\pi_i^*) = \pi_i(1 - \pi_i)/K$$

- $K = 25 \times 10^6$ ensures

$$P(|\pi_i^* - \pi_i| < 0.001) \geq 0.99$$

$$P(|\pi_{ij}^* - \pi_{ij}| < 0.001) \geq 0.99$$

for any π_i or π_{ij}

- Result 2: Horvitz-Thompson estimator

- $\hat{T} = \sum_{i \in s} y_i / \pi_i$ and $\tilde{T} = \sum_{i \in s} y_i / \pi_i^*$

- Relative bias: $(\hat{T} - \tilde{T}) / \hat{T}$

- For any response variable $y \geq 0$ and the given sample

$$P\left(\frac{|\hat{T} - \tilde{T}|}{\hat{T}} < \varepsilon\right) \geq 1 - \frac{2(1 + \varepsilon^2)}{K\varepsilon^2} \left(\sum_{i \in s} \frac{1}{\pi_i} - n\right)$$

- A practical lower bound for $P(|\hat{T} - \tilde{T}| / \hat{T} < \varepsilon)$

$$\Delta = 1 - \frac{2(N - n)}{K\varepsilon^2}$$

- $\varepsilon = 0.01$, $\Delta = 0.98$ and $N - n = 100$:

The theoretical (conservative) value of K : 10^8

- Result 3: Second order population parameter $Q = \sum_{i=1}^N \sum_{j=1}^N q(y_i, y_j)$

– Two HT type estimators

$$\hat{Q} = \sum_{i \in s} \sum_{i \in s} \frac{q(y_i, y_j)}{\pi_{ij}} \quad \text{and} \quad \tilde{Q} = \sum_{i \in s} \sum_{i \in s} \frac{q(y_i, y_j)}{\pi_{ij}^*}$$

– For any y and the given sample

$$P \left(\frac{|\hat{Q} - \tilde{Q}|}{\hat{Q}} < \varepsilon \right) \geq 1 - \frac{2(1 + \varepsilon^2)}{K\varepsilon^2} \left(\sum_{i \in s} \sum_{i \in s} \frac{1}{\pi_{ij}} - n^2 \right)$$

– A practical lower bound for $P(|\hat{Q} - \tilde{Q}|/\hat{Q} < \varepsilon)$

$$1 - 2 \frac{(N + n)(N - n)}{K\varepsilon^2}$$

3. Examples: First Order Inclusion Probabilities

- Example 1. $N = 22; n = 10$

Randomized systematic PPS sampling WITHOUT substitutions

Information on size variable comes from ITC-China survey

(i) True values of π_i :

0.5413 0.5141 0.6212 0.4941 0.2859 0.2458 0.3642 0.3874 0.
0.5555 0.4999 0.7637 0.6366 0.4142 0.2031 0.3927 0.3874 0.
0.1792 0.5517

(ii) Simulated π_i^* : $K = 10^5$

0.5405 0.5152 0.6216 0.4967 0.2857 0.2450 0.3662 0.3846 0.
0.5566 0.4993 0.7631 0.6352 0.4131 0.2043 0.3925 0.3877 0.
0.1777 0.5522

(iii) Simulated π_i^* : $K = 10^6$

0.5407 0.5141 0.6216 0.4947 0.2857 0.2457 0.3642 0.3871 0.
0.5557 0.5007 0.7636 0.6368 0.4143 0.2031 0.3925 0.3868 0.
0.1794 0.5515

- Example 2. (Cont'd from Example 1) Three refusing units
 - Case 1: Three largest units refuse
 - Case 2: Three smallest units refuse
 - Case 3: One large, one medium and one small unit refuse
 - $K = 10^6$
 - Compare to results as if PPS (Second half in each table)

(i) Three largest units refuse

0.6625 0.6373 0.7343 0.6166 0.3879 0.3392 0.4803 0.5054 0.
0.6754 0.6230 0.0000 0.7474 0.5350 0.2845 0.5117 0.5059 0.
0.2521 0.6722

0.6860 0.6516 0.7873 0.6262 0.3624 0.3116 0.4616 0.4910 0.
0.7040 0.6336 0.0000 0.8068 0.5249 0.2574 0.4978 0.4910 0.
0.2271 0.6992

(ii) Three smallest units refuse

0.5771	0.5491	0.6549	0.5291	0.3152	0.0000	0.3965	0.4201	0.
0.5916	0.5350	0.7922	0.6700	0.4489	0.0000	0.4270	0.4209	0.
0.0000	0.5869							

0.5776	0.5486	0.6629	0.5272	0.3051	0.0000	0.3887	0.4134	0.
0.5928	0.5334	0.8149	0.6793	0.4420	0.0000	0.4191	0.4134	0.
0.0000	0.5887							

(iii) Three units, one large, one small and one medium, refuse

0.6252 0.5990 0.0000 0.5777 0.3548 0.0000 0.4424 0.4676 0.
0.6380 0.5848 0.8223 0.7131 0.4967 0.2580 0.4745 0.4678 0.
0.2284 0.0000

0.6308 0.5992 0.0000 0.5758 0.3332 0.0000 0.4245 0.4515 0.
0.6474 0.5826 0.8900 0.7419 0.4827 0.2367 0.4577 0.4515 0.
0.2089 0.0000

4. Second Order Inclusion Probabilities

- The Hartley-Rao (1962) approximation to π_{ij} under randomized systematic PPS sampling WITHOUT substitutions
 - Work well for large N
 - Property UNKNOWN for small N
 - Can assess the goodness-of-approximation by simulation
 - π_{ij} intractable with unit substitutions
- Example: $N = 22$ and $n = 10$, no substitution
 - Simulated π_{ij} (first half of the table)
 - Hartley-Rao Approximation (second half of the table)

(i) Simulated π_{ij} ; $K = 10^6$

0.0000	0.2629	0.3260	0.2521	0.1444	0.1237	0.1839	0.1956	0.
0.2629	0.0000	0.3084	0.2369	0.1366	0.1173	0.1743	0.1861	0.
0.3260	0.3084	0.0000	0.2949	0.1655	0.1422	0.2106	0.2248	0.
0.2521	0.2369	0.2949	0.0000	0.1315	0.1123	0.1676	0.1790	0.
0.1444	0.1366	0.1655	0.1315	0.0000	0.0638	0.0962	0.1026	0.
0.1237	0.1173	0.1422	0.1123	0.0638	0.0000	0.0820	0.0878	0.

(ii) Hartley-Rao approximations to π_{ij}

0.0000	0.2637	0.3224	0.2528	0.1431	0.1225	0.1838	0.1960	0.
0.2637	0.0000	0.3054	0.2395	0.1355	0.1160	0.1741	0.1856	0.
0.3224	0.3054	0.0000	0.2928	0.1657	0.1419	0.2128	0.2269	0.
0.2528	0.2395	0.2928	0.0000	0.1300	0.1113	0.1669	0.1780	0.
0.1431	0.1355	0.1657	0.1300	0.0000	0.0630	0.0945	0.1007	0.
0.1225	0.1160	0.1419	0.1113	0.0630	0.0000	0.0809	0.0863	0.

5. Computational Implementation

- An R/SPLUS function for the randomized systematic PPS sampling method

Input variables:

(1) x : population vector of size variable

(2) n : sample size

Output: A set of sampled unit, s

```
syspps<-function(x,n) {  
N<-length(x)  
U<-sample(N,N)  
xx<-x[U]  
z<-rep(0,N)  
for(i in 1:N) z[i]<-n*sum(xx[1:i])/sum(x)  
r<-runif(1)  
s<-numeric()  
for(i in 1:N) {  
  if(z[i]>=r) {  
    s<-c(s,U[i])  
    r<-r+1  } }  
return(s[order(s)])  
}
```

- An R/SPLUS function for the randomized systematic PPS sampling method with UNIT SUBSTITUTIONS

Input variables:

(1) x : population vector of size variable

(2) n : sample size

(3) `refus`: the set of refusing units

Output: A set of sampled unit, s

```
sysppssub<-function(x,n,refus) {  
s<-syspps(x,n)  
sub<-numeric()  
for (i in 1:n) {  
if (min(abs(s[i]-refus))==0) sub<-c(sub,i)  
}  
m<-length(sub)  
if (m>0) {  
s<-s[-sub]  
U1<-(1:length(x))[-c(refus,s)]  
x1<-x[-c(refus,s)]  
s1<-syspps(x1,m)  
s<-c(s,U1[s1])  
}  
return(s[order(s)])  
}
```

- R/SPLUS codes for simulating π_i with substitutions

```
K<-1000000
pi<-rep(0,N)
for(i in 1:K){
s<-sysppssub(x,n,refus)
for(j in 1:N){
if(min(abs(s-j))==0) pi[j]<-pi[j]+1
}
}
pi<-pi/K
```

- In theory, simulation-based methods work for any sampling design as long as the complete design information is available
- Can handle more complex substitutions (i.e. second and third round refusals) or other types of modifications
- Randomized systematic PPS sampling: Computationally most efficient

- Example: CPU times for computing π_i for all $i \in s$ ($n = 10$ and $K = 10^6$, no substitutions)

Randomized Systematic PPS Versus Rao-Sampford PPS

N	Systematic PPS	Rao-Sampford PPS
200	4.7 hours	7.5 hours
100	2.5 hours	5.0 hours
50	1.6 hours	4.4 hours
20	1.2 hours	8.9 hours