Comprehensive Statistics Assignment 2

bst9911

March 1, 2002

Arnab Bose

Assignment for Rolls 9 and 11

POPULATION TOTAL ESTIMATION FOR NSSO SAMPLING DESIGN

Due to some error in the implementation of Lahiri's scheme for circular systematic sampling with pi-proportional to size `in two half-samples (supposed to be independent), a complex sampling scheme arose and the two half-samples turned out to be dependent. The task is to obtain an unbiased estimate of the population total and a variance estimate for it. The mean squared error of the resulting estimators for the intended (Lahiri's) and the actual sampling schemes should also be compared through simulation.

Preliminaries

For the purpose of this assignment, extend the definition of modulus from integers to real numbers as follows:

mod(a,b)=r s.t. $a=r+b\times n$, where $r\in[0,b)$, n is integer.

Some Sampling Schemes

We are given N units with size measures S_1, S_2, …, S_N. We are to choose a sample of size n from the N units. Let the total of the sizes be $S=\sum S_i$. Define the probability measures p_i as $p_i=S_i/S$. Let $c_0=0$, $c_1=p_1$, $c_2=p_1+p_2$, …, $c_N=1$ be the cumulative probabilities.

Define the function k(u)= min{i: $c_i>u$}.

Sampling Scheme 1

Get an observation u from Uniform(0,1). Select the unit with the index k(u) as the first unit in the sample. Change u to mod(u+1/n, 1), and select the k(u)th unit again into the sample. Repeat this procedure n times, so that a sample of size n is obtained.

Note: This scheme allows a unit to be present more than once in the sample.

This scheme is simple enough to allow straightforward calculation of inclusion probabilities:

$\pi_i= np_i$ if $np_i<1$,

1 otherwise.

This can be also written as

$\pi_i=n$ length({mod(x,1/n): $x\in(c_{i-1}, c_i)$}) where length of a subset of reals has its usual meaning.

Second order inclusion probability can also be obtained as

$\pi_{ij}=n$ length({mod(x,1/n): $x\in(c_{i-1}, c_i)$} $\cap$ {mod(x,1/n): $x\in(c_{j-1}, c_j)$}).

The higher order inclusion probabilities can also be expressed similarly.

Sampling Scheme 2

This scheme is almost same as Scheme 1, except for the fact that repetitions are not allowed here. We essentially follow scheme 1, and to handle repetitions we skip the units that are already included in the sample. In case we have continuously skipped a pre-fixed number of units, we keep the units we already have in the sample and set $u=c_{i-1}$ where the first unit from the top that has not already been included is the i^{th} unit. This will make i^{th} unit to be the unit included next. This procedure is carried out like this as long as a sample of size n is not obtained. The number of continuous skips before we start from the top is usually taken as 15 or 20. Let us call this number MAX_REPEATS.

Note: Although this is the original scheme that was supposed to be used by NSSO, it has an undesirable quality. This scheme gives bias to the inclusion of the first unit if $p_i>1/n$ (i.e. $S_i>S/n$) for some i. Also if $p_i>2/n$ for any i, then we will have $\pi_1=1$ irrespective of size of first unit!

Note that the second scheme reduces to the first scheme and both the schemes are πps schemes when $p_i<1/n$ for all i.

Sampling Scheme 3

Instead of using sampling scheme 2 independently to the first half-sample obtained, a mistake was made and the result was the following scheme for the second half-sample. First we have to have a first half-sample of size n chosen beforehand. Then as before a random number u was observed from the Uniform(0,1) distribution. From then on, scheme 2 was implemented for choosing $2n$ units, pretending that n units (i.e. the first half-sample) are already chosen and remaining n units are needed to be chosen without replacement.

Observations

For the subsequent calculations, let us denote by a_i, a_{ij} the inclusion probabilities for first half-sample, by b_i, b_{ij} the inclusion probabilities for second half-sample, and by π_i, π_{ij} the inclusion probabilities for the combined sample. Also let us denote by A_i the event that i^{th} element is included in 1^{st} half-sample, and by B_i the event that i^{th} element is included in 2^{nd} half-sample.

Independent Half-Samples

Originally, the first and second half-s1amples were both supposed to be independently chosen using scheme 2. This would ensure the basic necessity for the estimablility of variance of $\hat{Y}_{HTE}$, since independence guarantees that all second order inclusion probabilities will be positive (because the first order inclusion probabilities are positive for each half-sample). We can calculate the combined inclusion probabilities in terms of half-sample inclusion probabilities as follows:

$\pi_i = P(A_i \cup B_i) = P(A_i) + P(B_i) - P(A_i B_i) = a_i + b_i - a_i b_i$ ..... Eqn 1(a)

$\pi_{ij} = P(A_i A_j \cup A_i B_j \cup B_i A_j \cup B_i B_j)$

$= P(A_i A_j) + P(A_i B_j) + P(B_i A_j) + P(B_i B_j)$

$-P(A_i A_j B_i) - P(A_i A_j B_j) - P(A_i A_j B_i B_j) - P(A_i B_j B_i A_j) - P(A_i B_j B_i) - P(B_i A_j B_j)$

$+4P(A_i A_j B_i B_j) - P(A_i A_j B_i B_j)$ (by the inclusion exclusion principle)

$= a_{ij} + a_i b_j + b_i a_j + b_{ij} - (a_{ij} b_j + a_{ij} b_i + a_{ij} b_{ij} + a_{ij} b_{ij} + a_i b_{ij} + a_j b_{ij}) + 3 a_{ij} b_{ij}$

$= a_{ij} + a_i b_j + b_i a_j + b_{ij} - a_{ij}(b_j + b_i) - (a_i + a_j) b_{ij} + a_{ij} b_{ij}$ ..... Eqn 1(b)

Dependent Half-Samples

However, a 'defective' method was used, where first half-sample was chosen by scheme 2 and the second half-sample was chosen (given the first half-sample) by scheme 3, given the first half-sample. Observe that even now all the second order inclusion probabilities are positive since:

$\pi_{ij} > P(\{k(u)=i$ for the first u in scheme 2$\} \cap \{k(u)=j$ for the first u in scheme 3$\})$

$= p_i p_j$ where $p_k = S_k/S$ is the probability measure corresponding to unit k, since the events are independent.

So even in the defective scheme, it is possible to get an unbiased estimator for the variance of $\hat{Y}_{HTE}$.

The Algorithm for Inclusion Probabilities

Observe that for scheme 1, instead of u it suffices to know $\mathrm{mod}(u,1/n)$ in order to find out the sample. This is because after observing u, we choose units $k(\mathrm{mod}(u+i/n,1))$ for $i=1, 2, \ldots, n$. If we change u with $\mathrm{mod}(u,1)$, we will find the same units in a different order.

The same is true for schemes 2 and 3 if it is guaranteed that one cycle will be completed (i.e. n turns will be completed) before getting MAX_REPEATS number of consecutive elements that are already in the sample. This is because if we start with any $\mathrm{mod}(u+i/n,1)$ instead of u, we will complete the same first cycle. If we don't already have n elements, we will not get anything new pursuing the next values of u's. This is because we will get back the 1^{st} element on the $n+1^{st}$ turn, 2^{nd} on the $n+2^{nd}$ turn, and so on, and ultimately exceed MAX_REPEATS turns. After that all the elements we get will clearly be the same as before.

So let us consider this case first, i.e. when the sample is determined by $\mathrm{mod}(u,1/n)$.

If the sample is determined by $\mathrm{mod}(u,1/n)$

Let us therefore wlg take initial u in $(0,1/n)$.

Let us divide the real line from 0 to 1 into proportions given by the S_i's. Then the i^{th} segment formed will have size p_i. Let us also divide the interval into n equal parts.

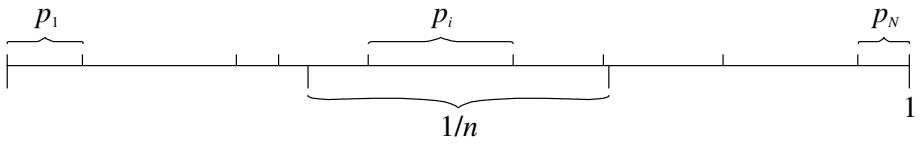

Figure 1: Interval Representation

During the first cycle, units are picked from each interval of size 1/*n*. The process can be represented graphically if we position these intervals one below other. Then if a vertical line is drawn passing through *u*, all units that intersect the line will be chosen.

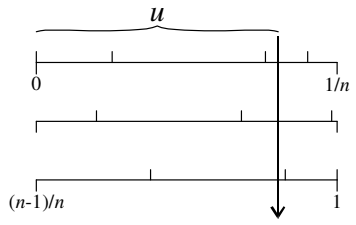

Figure 2: Stacked Interval Representation

If we increase or decrease *u*, there will be a point when this vertical line will cross over one of the size marks. It is clear that when it does so, the points that we pick up in the first *n* turns while forming the sample will no longer be the same as that of the old sample. It is also clear that as long as it does not cross any mark, we will always get the same set of points in the first *n* turns. Since these points comprise of the sample points for with-replacement scheme, the sample we obtain is same as long as *u* does not cross any of the size marks. Also since after getting these points continuous repetitions are bound to occur if we continue a without-replacement scheme - thus fixing the future points of the sample, the same can be said about those schemes.

Hence it will be easy to identify the different possible samples if we superimpose the size marks in these 1/*n*-length intervals. Each of the sub-intervals formed by the superimposed size marks will denote one possible sample. Moreover, the probability of the occurrence of any sample will be the probability of mod(*u*,1/*n*) falling in that sub-interval, which is *n* times the length of that sub-interval.

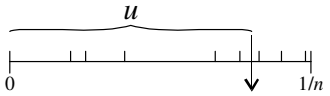

Figure 3: Superimposed Interval Representation

Immediately with the help of this representation we find that the total number of possible samples is *N*, since there are *N* sub-intervals of (0,1/*n*) in the figure. Let us denote the sample corresponding to the i^{th} sub-interval as s_i. Then we have for any function f of the sample,

$$E(f(s_i)) = \sum_{i=1}^{N} f(s_i) P(s_i) \quad \text{.....Eqn 2(a)}$$

In particular, if we take

$f(s)=1_s(i^{th}$ unit), then we get $E(f(s_i))=\pi_i$, and taking

$f(s)=1_s(i^{th}$ unit)$1_s(j^{th}$ unit) gives us $E(f(s_i))=\pi_{ij}$ (for one half-sample only).

Using equations 1(a) and 1(b), we can immediately get the inclusion probabilities for a combined sample formed by two independent half-samples.

For two dependent half-samples

There can be at most N first half-samples, and for each of them there can again be N second half-samples. Let us denote the first half-samples as s_i's second half-samples as t_i's. Then we have $P(s_i \cap t_i)=P(t_i \mid s_i)P(s_i)$. Using the superimposition technique we can calculate theses probabilities.

We have

$$E(f(s_i,t_j)) = \sum_{i=1}^{N}\sum_{j=1}^{N} f(s_i,t_j)P(t_j \mid s_i)P(s_i) \quad \text{..... Eqn 2(b)}$$

giving us the inclusion probabilities for $f(s,t)=1_{s,t}(i^{th}$ unit) and $f(s,t)=1_{s\cup t}(i^{th}$ unit)$1_{s\cup t}(j^{th}$ unit) we get the first and second order inclusion probabilities as the expectations.

If the sample is not determined by mod$(u,1/n)$

This may happen only when there is a unit with size large enough to let us have MAX_REPEATS+1 repetitions of itself, making us start over before completing the first cycle. If this happens for any some initial u, then it will happen for mod$(u,1/n)$ as well. So it is sufficient to test this for u in $(0,1/n)$.

If we find for some initial $u\in(0,1/n)$ that we will need to start over, then taking mod$(u,1/n)$ will not suffice for all u within that specific sub-interval of $(0,1/n)$. Hence for such u's we will have to consider n possible samples that originate by taking the initial random start as u, $u+1/n$, …, $u+(n-1)/n$. Each of these samples will have a probability equal to the length of the sub-interval (i.e. the normalised size intervals) where u belongs.

We are now in a position to write an algorithm for calculating the inclusion probabilities.

The algorithm

1. Given the population size N, sample size n, and the size measures S_i's, form the probabilities by $p_i=S_i/\sum S_i$. Define $c_0=0$, $c_1=p_1$, $c_2=p_1+p_2$, …, $c_N=1$. Then collect the mod$(c_i,1/n)$'s for $i=0$ to $N-1$ into an array SImp[] (of size $N+1$) and sort it in ascending order. Define SImp[$N+1$]=$1/n$.
2. Enumerate all possible N number of first half-samples by taking $p_k\in$(SImp[k], SImp[$k+1$]). For definiteness, take p_k=(CuS[k]+CuS[$k+1$])/2.
3. Is it known that the schemes are independent? If yes go to step 5.
4. For each of these first half-sample, again generate N number of second half-samples. Use equation 2(b) to find out the inclusion probabilities. Go to step 6.

5. Calculate the inclusion probabilities for the first half-sample using equation 2(a). Enumerate all possible second half-samples now, in the same way as step 2. Calculate the inclusion probabilities for the second half-samples again by using equation 2(a). Now use equations 1(a) and 1(b) to compute the inclusion probabilities for combined sample.
6. End of algorithm.

Once we find the inclusion probabilities for the observed sample, we can estimate the population total and also find exact variance (if we have calculated all the inclusion probabilities) or an unbiased estimate of the variance (if we have calculated inclusion probabilities for the units in observed sample only) of this estimator using the formulae derived by Horvitz and Thompson:

$$\hat{Y}_{HTE} = \sum_{i=1}^{n} \frac{y_i}{\pi_i} \quad \text{.....Eqn 3(a)}$$

$$V(\hat{Y}_{HTE}) = \sum_{i=1}^{n}\sum_{j=1}^{n} y_i y_j \left(\frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j} \right) \text{ (if we take } \pi_{ii}=\pi_i\text{) ..... Eqn 3(b)}$$

$$\hat{V}(\hat{Y}_{HTE}) = \sum_{i=1}^{n}\sum_{j=1}^{n} y_i y_j \left(\frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j \pi_{ij}} \right) \text{ (if we take } \pi_{ii}=\pi_i\text{) .....Eqn 3(c)}$$

Note: Higher order inclusion probabilities can also be calculated if needed by the method mentioned here.

Simulation

A program was written which generates random values for sizes from Uniform distribution 1 to 10000, and the values associated to the units. It then uses the sampling schemes (original and 'defective') to get to pair of half-samples. After getting the samples it generates values associated to those units from some predefined distribution. Then the inclusion probabilities and hence the variance estimates are calculated for only the units included for both the schemes using equation 3(c). The whole process is repeated several times, and number of times the original method gives lesser variance is obtained.

In the beginning, the program was not fast enough with high values of N since the time needed for calculating inclusion probabilities increases by the order of N^2. Same is true for n also. The code very slow - estimated time to complete one for n=20 and N=500 simulation was about 42 hours in my 166MHz PC! That is why some speed increasing algorithms like binary search were needed. After many revisions of the program, the time was drastically reduced to about 1 minute and that too for computing actual variance (using equation 3(b)) instead of variance estimate! Still the speed was not sufficient to work with the magnitude of data in the original problem that NSSO was facing. The table below summarises the results that were obtained.

Table 1: Summary of Simulations

Y Values	*N*	*n*	Proportion of times 'defective' was better	'Defective' Variances				'Non-defective' Variances			
				Min	Max	Avg.	Std. Dev.	Min	Max	Avg.	Std. Dev.
Normal(50,25) truncated below at 10	1000	40	100%	68.7703	440.9270	186.6980	81.3700	86.6131	450.9410	194.6126	81.5600
Normal(50,25) truncated below at 10	500	20	100%	90.0772	1132.1700	308.4605	227.2100	98.5672	1143.6100	321.9830	226.8092
Exponential with location 25, scale 25	500	20	100%	83.3333	783.9620	313.7694	163.7486	98.3179	795.3010	325.6739	164.6809

For each set-up 100 populations were generated from uniform 1 to 10000 distribution, and the exact variances for the two schemes were calculated using equation 3(b).

Each time, the variance for the 'defective' scheme turned out to be better, although sometimes only marginally. This is not totally unanticipated, since intuitively one should expect less variance for the dependent scheme since there it is ensured that all units chosen will be different, guaranteeing more information than the independent method. This is somewhat like without replacement sampling being better than with replacement sampling. However, it is rather surprising that 'defective' scheme always gave lesser variation, for the entire 300 populations that were generated. This indicates that perhaps, it can be proved analytically that 'defective' scheme is always better.

Appendix

Since the program is taking so long an approximate method is suggested (in fact included in the program, activated by `#define APPROX`). We can arrange the probabilities of N possible samples in ascending order, and then ignore the first few half-samples (for both first and second sampling), which have total probability just less than some predefined α.

The N^2 (combined) samples that we are considering for the algorithm for the defective scheme can be enumerated by the rectangles in the following figure. Note that not all of them will be different, in fact many will actually turn out to be the same sample. The probability of obtaining a sample is proportional to sum of all the areas of the rectangles, which correspond to that sample. The shaded portion of the figure denotes the samples that we are ignoring in our approximation.

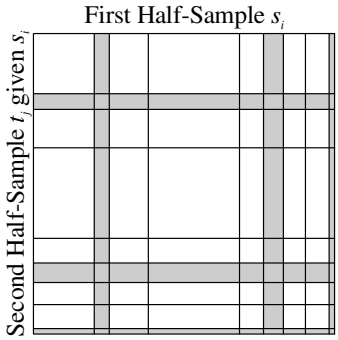

Figure 4: Graphical Representation of Probabilities of Samples

Note: Although the representation gives a feeling of independence, actually that is false. This is because the Second Half Samples will not all be same over a row since they depend on the First Half Sample.

The maximum probability measure of each half-sample that is being ignored is α. The maximum probability measure of combined sample being ignored can be calculated from the figure: $2\alpha-\alpha^2$. The maximum difference in first order or second order inclusion probabilities caused by this approximation is therefore $2\alpha-\alpha^2$. Using this one can come to a value of α that will give desired accuracy.

This approximation technique poses some problems. One is that the accuracies are established using the assumption that sample is determined by mod(u,1/n). Another serious problem is that the approximate inclusion probabilities may sometimes turn out to be zero. Due to these problems and upcoming deadline for submission of the assignment, this line of thought was abandoned.

Supplement

Let us denote by a_i and a_{ij} the inclusion probabilities for the first half-sample, and by d_i and d_{ij} the conditional inclusion probabilities for the second half-sample given the first half-sample. Let s denote the first half-sample and t be the second half-sample.

Then a computationally cheaper estimate of Y is $\hat{Y} = \dfrac{\sum_{i\in s}\frac{y_i}{a_i} + \left(\sum_{j\in t}\frac{y_j}{d_j} + \sum_{i\in s} y_i\right)}{2}$.

To prove that this is an unbiased estimate, let E_1 denote expectation on the first half-sample and E_2 denote conditional expectation on the second half-sample given the first half-sample. Then we have:

$$E(\hat{Y}) = E_1E_2(\hat{Y}) = E_1\left(\left[\sum_{i\in s}\frac{y_i}{a_i} + \left(\sum_{j\in s^c} y_j + \sum_{i\in s} y_i\right)\right]\Big/2\right) = \frac{E_1\left(\sum_{i\in s}\frac{y_i}{a_i}\right) + Y}{2} = Y$$

We can similarly calculate the variance of this estimate. Let V_1 denote variance on the first half-sample only and V_2 denote conditional variance on the second half-sample given the first half-sample. Then,

$$\begin{aligned} V(2\hat{Y}) &= V_1E_2(2\hat{Y}) + E_1V_2(2\hat{Y}) \\ &= V_1\left(\sum_{i\in s}\frac{y_i}{a_i} + Y\right) + E_1\left(\sum_{i\in s^c}\sum_{j\in s^c} y_iy_j\left(\frac{d_{ij}-d_id_j}{d_id_j}\right)\right) \\ &= \sum_{i=1}^{N}\sum_{j=1}^{N} y_iy_j\left(\frac{a_{ij}-a_ia_j}{a_ia_j}\right) + E_1\left(\sum_{i\in s^c}\sum_{j\in s^c} y_iy_j\left(\frac{d_{ij}-d_id_j}{d_id_j}\right)\right) \end{aligned}$$

Here the second expectation is difficult to simplify analytically. However, a computer program can calculate it by using equation 2(a).

Notice that an unbiased estimate of the variance *would* be

$\hat{V}(2\hat{Y}) = \sum_{i\in s}\sum_{j\in s} y_iy_j\left(\frac{a_{ij}-a_ia_j}{a_ia_ja_{ij}}\right) + \sum_{i\in t}\sum_{j\in t} y_iy_j\left(\frac{d_{ij}-d_id_j}{d_id_jd_{ij}}\right)$ *if* all the a_{ij}'s and d_{ij}'s *were* positive. But since most of these probabilities are zero, this is not actually an unbiased estimate for the variance.[1]

The program for simulation was modified to evaluate this variance along with the previous two variances and comparison was made. Out of 100 populations generated with N=1000, n=30, S_i's uniform 1 to 10000, and Y_i's distributed normally with mean 50 and variance 25, in all cases the 'defective' method had the least variance for the

[1] Note that for any actual sample (which has a positive probability of occurrence) this variance estimate is well defined. However, the expectation of this expression will be always greater than the actual variance since it will miss all terms inside the summation with either a_{ij}=0 or b_{ij}=0, from the actual variance.

HTE, and the new estimate for the 'defective' method had the most variance. The summaries of the variances obtained are tabulated below.

Table 2: Summary of simulation with the new estimate

'Defective' HTE Variances			'Non-defective' HTE Variances			New estimate variances		
Avg.	Min	Max	Avg.	Min	Max	Avg.	Min	Max
186.2294	87.1719	538.2910	194.1564	94.6861	541.9180	3304.0110	1730.1700	8890.3200

Thus the new estimate for the 'defective' method has about 18 times more variance than the HTE on average!

We could improve the variance a little by obtaining an unordered estimate using the new estimate. Let $f(s,t)$ be the new ('cheaper') estimate. Then the unordered estimate would be $g(s,t)=\frac{f(s,t)P(s,t)+f(t,s)P(t,s)}{P(s,t)+P(t,s)}$ where $P(s,t)$=probability of obtaining s as the first half-sample and t as the second. This will have better variance by the virtue of being unordered (wrt the first and the second half-samples). Also it will be unbiased since

$$\begin{aligned}
Eg(s,t) &= \sum_{\{s,t\}} \frac{f(s,t)P(s,t)+f(t,s)P(t,s)}{P(s,t)+P(t,s)} \cdot P(\{s,t\}) \\
&= \sum_{\{s,t\}} \frac{f(s,t)P(s,t)+f(t,s)P(t,s)}{P(s,t)+P(t,s)} \cdot \big(P(s,t)+P(t,s)\big) \\
&= \sum_{\{s,t\}} f(s,t)P(s,t)+f(t,s)P(t,s) \\
&= \sum_{s,t} f(s,t)P(s,t) = Ef(s,t) = Y
\end{aligned}$$