CSSELECT

This document describes the algorithm used by CSSELECT to draw samples according to complex designs. The data file does not have to be sorted. Population units can appear more than once in the data file and they do not have to be in a consecutive block of cases.

Notation

The following notation is used throughout this chapter unless otherwise stated:

N	Population size
n	Sample size
f	Sampling fraction
h_i	Hit counts of i-th population unit. (i=1,...,N)
M_i	Size measure of i-th population unit. (i=1,...,N)
M	Total size. $M = \sum_{i=1}^{N} M_i$
p_i	$p_i = \frac{M_i}{M}$ is the relative size of i-th population unit (i=1,...,N)

Stratification

Stratification partitions the sampling frame into disjoint sets. Sampling is carried out independently within each stratum. Therefore, without loss of generality, the algorithm described in this document only considers sampling from one population.

In the first stage of selection, the sampling frame is partitioned by the stratification variables specified in stage 1. In the second stage, the sampling frame is stratified by first-stage strata and cluster variables as well as strata variables specified in stage 2. If sampling with replacement is used in the first stage, the first-stage duplication index is also one of the stratification variables. Stratification of the third stage continues in a like manner.

Population Size

Sampling units in a population are identified by all unique level combinations of cluster variables within a stratum. Therefore, the population size N of a stratum is equal to the number of unique level combinations of the cluster variables within a stratum. When a sampling unit is selected, all cases having the same sampling unit identifier are included in the sample. If no cluster variable is defined, each case is a sampling unit.

Sample Size

CSSELECT uses a fixed sample size approach in selecting samples. If the sample size n is supplied by the user, it should satisfy $0 \le n \le N$ for any without replacement design and $n \ge 0$ for any with replacement design.

If a sampling fraction f is specified, it should satisfy $0 < f \le 1$ for any without replacement design and $f > 0$ for any with replacement design. The actual sample size is determined by the formula $n = \text{round}(f * N)$. When the option RATEMINSIZE is specified, a sample size less than RATEMINSIZE is raised to RATEMINSIZE. Likewise, a sample size exceeding RATEMAXSIZE is lowered to RATEMAXSIZE.

Simple Random Sampling

This algorithm selects n distinct units out of N population units with equal probability; see Fan, Muller & Rezucha (1962) for more information.

- Inclusion probability of i-th unit = n/N
- Sampling weight of i-th = N/n

Algorithm

1. If f is supplied, compute n=round(f*N).
2. Set k=0, i=0 and start data scan.
3. Get a population unit and set k=k+1. If no more population units left, terminate.
4. Test if k-th unit should go into the sample
   a) Generate a uniform (0,1) random number U
   b) If $(n-i)/(N-k+1) > U$, k-th population unit is selected and set i=i+1.
   c) If i=n, terminate. Otherwise, go to step 3.

Unrestricted Random Sampling

This algorithm selects n units out of N population units with equal probability and with replacement.

- Inclusion probability of i-th unit = $1-(1-1/N)^n$
- Sampling weight of i-th = N/n. (For use with Hansen-Hurwitz(1943) estimator)
- Expected number of hits of i-th = n/N

Algorithm

1. Set i=0 and initialize all hit counts to zero.
2. Generate an integer k between 1 and N uniformly.
3. Increase hit count of k-th population unit by 1.
4. Set i=i+1.
5. If i=n, then terminate. Otherwise go to step 2.

At the end of the procedure, population units with hit count greater than zero are selected.

Systematic Sampling

This algorithm selects n distinct units out of N population units. If the selection interval (N/n) is not an integer, an exact fractional selection interval is used.

- Inclusion probability of a unit = n/N
- Sampling weight = N/n

Algorithm

1. Draw a uniform (0,1) random number U.
2. Population units with indices {i: i=trunc((U+k)*N/n)+1, k=0,...,n-1} are included in the sample.

Sequential Sampling (Chromy)

See the section on PPS sequential sampling. This algorithm is a special case of PPS Chromy with all size measures M_i equal.

PPS Sampling without Replacement (Hanurav & Vijayan)

This algorithm selects n distinct units out of N population units with probability proportional to size without replacement. This method is first proposed by Hanurav (1967) and extended by Vijayan (1968) to the case of n>2.

- Inclusion probability of i-th unit = np_i
- Sampling weight of i-th unit = $\frac{1}{np_i}$
- Special requirement: $\max M_i \le \frac{M}{n}$.

Algorithm (Case 1)

This algorithm assumes that the population units are sorted by M_i (i.e. $M_1 \le M_2 \le ... \le M_N$) with the additional assumption that $M_{N-n+1} < M_N$.

1. Compute the probabilities $\theta_j = \frac{n(p_{N-n+j-1} - p_{N-n+j})(S + jp_{N-n+1})}{S}$, j=1,...,n, where $S = \sum_{k=1}^{N-n} p_k$.
2. Select one integer from 1,...,n with probability proportional to θ_j.

3. If the integer selected in step 2 is i, then the last (n-i) population units are selected.
4. Define a new set of probabilities for the first (N-n+i) population units.

$$p_j^* = \frac{p_j}{S + ip_{N-n+1}}, \quad 1 \le j \le N-n+1$$

$$= \frac{p_{N-n+1}}{S + ip_{N-n+1}}, \quad N-n+1 < j \le N-n+i$$

5. Define $P_j = \frac{M_j}{(M_{j+1} + ... + M_{N-n+i})}, \quad j = 1,..., N-n+i-1$
6. Set m=1 and select one unit from the first (N-n+1) population units with probability proportional to

$$a_1 = ip_1^*$$

$$a_j = np_j^* \prod_{k=1}^{j-1} [1-(i-1)P_k], \quad j = 2,..., N-n+1$$

7. Denote the index of the selected unit by j_m.
8. Set m=m+1 and select one unit from ($j_{m-1}+1$)-th to (N-n+m)-th population units with the following revised probabilities

$$a_{j_{m-1}+1} = (i-m+1)p^*_{j_{m-1}+1}$$

$$a_j = (i-m+1)p_j^* \prod_{k=j_{m-1}+1}^{j-1} [1-(i-m)P_k], \quad j = j_{m-1}+2,..., N-n+m$$

9. Denote the selected unit in step 8 by j_m.
10. If m=i, terminate. Otherwise, go to step 8.

At the end of the algorithm, the last (n-i) units and units with indices $j_1,..., j_i$ are selected.

Joint Inclusion Probabilities (Case 1)

The joint inclusion probabilities of unit i and unit j in the population ($1 \le i < j \le N$) is given by

$$\pi_{ij} = \sum_{r=1}^{n} \theta_r K_{ij}^{(r)}$$

where

$$K_{ij}^{(r)} = \begin{cases} 1 & \text{if} \quad N-1 \geq i > N-n+r, \\ \dfrac{rp_{N-n+1}}{S+rp_{N-n+1}} & \text{if} \quad N-n+r \geq i > N-n \quad \text{and} \quad j > N-n+r, \\ \dfrac{rp_i}{S+rp_{N-n+1}} & \text{if} \quad N-n \geq i > 0 \quad \text{and} \quad j > N-n+r, \\ \pi_{ij}^{(r)} & \text{if} \quad j \leq N-n+r. \end{cases}$$

$\pi_{ij}^{(r)}$'s are the conditional joint inclusion probabilities given that the last (n-r) units are selected at step 3. They can be computed by the following formula

$$\pi_{ij}^{(r)} = r(r-1)(1-P_1^{(r)})...(1-P_{i-1}^{(r)})P_i^{(r)}p_j^{(r)}$$

where

$$p_k^{(r)} = \begin{cases} \dfrac{p_k}{S+rp_{N-n+1}} & \text{if} \quad k \leq N-n+1 \\ \dfrac{p_{N-n+1}}{S+rp_{N-n+1}} & \text{if} \quad N-n+1 < k \leq N-n+r \end{cases}$$

and

$$P_k^{(r)} = \frac{p_k^{(r)}}{(p_{k+1}^{(r)} + ... + p_{N-n+r}^{(r)})}.$$

Note: There is a typo in (3.5) of Vijayan(1967) and (3.3) of Fox(1989). The factor (1/2) should not be there. See also Golmant (1990) and Watts (1991) for other corrections.

Algorithm (Case 2)

This algorithm assumes that the population units are sorted by M_i with the order $M_1 \leq M_2 \leq ... \leq M_N$ and the additional assumption $M_{N-n+1} = M_N$.

1. Define the probabilities

$$P_j = \frac{M_j}{(M_{j+1} + ... + M_N)}, \quad j = 1,..., N-1$$

2. Select one unit from the first (N-n+1) population units with probability proportional to

$$a_1 = np_1$$

$$a_j = np_j \prod_{k=1}^{j-1} [1-(n-1)P_k], \quad j = 2,..., N-n+1$$

3. Set m=1 and denote the index of the selected unit in step 2 by j_m.
4. Set m=m+1.
5. Select one unit from ($j_{m-1}+1$)-th to (N-n+m)-th population unit with probability proportional to

$$a_{j_{m-1}+1} = (n-m+1)p_{j_{m-1}+1}$$

$$a_j = (n-m+1)p_j \prod_{k=j_{m-1}+1}^{j-1} [1-(n-m)P_k], \quad j = j_{m-1}+2,..., N-n+m$$

6. Denote the index of the unit selected in step 5 by j_m.
7. If m=n, terminate. Otherwise, go to step 4.

At the end of the algorithm, population units with indices $j_1,..., j_n$ are selected.

Joint Inclusion Probabilities (Case 2)

Joint inclusion probabilities π_{ij} of unit i and unit j in the population ($1 \le i < j \le N$) are given by $\pi_{ij} = n(n-1)(1-P_1)...(1-P_{i-1})P_i p_j$.

PPS Sampling with Replacement

This algorithm selects n units out of N population units with probability proportional to size and with replacement. Any units may be sampled more than once.

- Inclusion probability of i-th unit = $1-(1-p_i)^n$
- Sampling weight of i-th unit = $\frac{1}{np_i}$. (For use with Hansen-Hurwitz(1943) estimator)
- Expected number of hits of i-th unit = np_i

Algorithm.

1. Compute total size $M = \sum_{i=1}^{N} M_i$.
2. Generate n uniform (0,M) random numbers $U_1,..., U_n$.

3. Compute hit counts of i-th population unit $h_i = \#\{U_j : M^*_{i-1} < U_j \le M^*_i, j = 1,..., n\}$, where $M_0 = 0$ and $M^*_i = \sum_{k=1}^{i} M_k$.

At the end of the algorithm, population units with hit count $m_i > 0$ are selected.

PPS Systematic Sampling

This algorithm selects n units out of N population units with probability proportional to size. If the size of the i-th unit M_i is greater than the selection interval, the i-th unit is sampled more than once.

- Inclusion probability of i-th unit = np_i
- Sampling weight of i-th unit = $\frac{1}{np_i}$
- Expected number of hits of i-th unit = np_i . In order to have no duplicates in the sample, the condition $\max M_i \le \frac{M}{n}$ is needed.

Algorithm

1. Compute cumulated sizes $M^*_i = \sum_{k=1}^{i} M_k$.
2. Compute the selection interval I=M/n.
3. Generate a random number S from uniform(0,I).
4. Generate the sequence $\{S_j : S_j = S + (j-1)I,\ j = 1,..., n\}$.
5. Compute hit counts of i-th population unit $h_i = \#\{M^*_{i-1} < S_j \le M^*_i, j = 1,..., n\}$, k=1,...,N.

At the end of the algorithm, population with hit counts $h_i > 0$ are selected.

PPS Sequential Sampling (Chromy)

This algorithm selects n units from N population units sequentially proportional to size with minimum replacement. This method is proposed by Chromy (1979).

- Inclusion probability of i-th unit = np_i
- Sampling weight of i-th unit = $\frac{1}{np_i}$
- Maximum number of hits of i-th unit = $\text{trunc}(np_i) + 1$

- By applying the restriction $\max M_i \le \frac{M}{n}$, we can ensure maximum number of hits is equal to 1.

Algorithm

1. Select one unit from the population proportional to its size M_i. The selected unit receives a label 1. Then assign labels sequentially to the remaining units. If the end of the list is encountered, loop back to the beginning of the list until all N units are labeled. These labels are the index i in the subsequent steps.
2. Compute integer part of expected hit counts $I_i = \text{trunc}(M_i^*)$, where $M_i^* = \sum_{k=1}^{i} M_k$, i=1,...,N.
3. Compute fractional part of expected hit counts $F_i = M_i^* - I_i$, i=1,...,N.
4. Define $I_0 = 0$, $F_0 = 0$ and $T_0 = 0$.
5. Set i=1.
6. If $T_{i-1} = I_{i-1}$, go to step 8.
7. If $T_{i-1} = I_{i-1} + 1$, go to step 9.
8. Determine accumulated hits at i-th step (case 1).
   a) Set $T_i = I_i$.
   b) If $F_i > F_{i-1}$, set $T_i = T_i + 1$ with probability $(F_i - F_{i-1})/(1 - F_{i-1})$.
   c) Set i=i+1.
   d) If i > N, terminate. Otherwise go to step 6.
9. Determine accumulated hits at i-th step (case 2).
   a) Set $T_i = I_i$.
   b) If $F_i > F_{i-1}$, set $T_i = T_i + 1$.
   c) If $F_{i-1} \ge F_i$, set $T_i = T_i + 1$ with probability F_i / F_{i-1}.
   d) Set i=i+1.
   e) If i > N, terminate. Otherwise go to step 6.

At the end of the algorithm, number of hits of each unit can be computed by the formula $h_i = T_i - T_{i-1}$, i=1,...,N. Units with $m_i > 0$ are selected.

PPS Sampford's Method

Sampford's (1967) method selects n units out of N population units without replacement and probabilities proportional to size.

- Inclusion probability of i-th unit = np_i

- Sampling weight of i-th unit = $\frac{1}{np_i}$
- Special requirement: $\max M_i < \frac{M}{n}$

Algorithm

1. If $\max M_i < \frac{M}{n}$, then go to step 2, otherwise go to step 5.
2. Select one unit with probability proportional to p_i, i=1,...,N.
3. Select the remaining (n-1) units with probabilities proportional to $\frac{p_i}{1-np_i}$, i=1,...,N.
4. If there are duplicates, reject the sample and go to step 2. Otherwise accept the selected units and stop.
5. If $N = n$ and M_i's are constant, then select all units in the population and set all sampling weights, 1st and 2nd order inclusion probabilities to 1.

Joint Inclusion Probabilities

First define the following quantities:

$$\lambda_i = \frac{p_i}{(1-np_i)}, \text{ i=1,...,N}$$

$$R_r = \sum_{k=1}^{N} \lambda_k^r, \text{ r=1,...,n}$$

$$L_0 = L_{0,ij} = 1, \text{ i,j=1,...,N}$$

$$L_m = \frac{1}{m}\sum_{k=1}^{m} (-1)^{k-1} R_k L_{m-k}, \text{ m=1,...,n}$$

$$L_{m,ij} = L_m - (\lambda_i + \lambda_j)L_{m-1,ij} - \lambda_i\lambda_j L_{m-2,ij}, \text{ m=1,...,n, i,j=1,...,N}$$

$$K_n = \left(\sum_{k=1}^{n} \frac{kL_{n-k}}{n^k} \right)^{-1}$$

Given the above quantities, the joint inclusion probabilities of i and j-th population unit is

$$\pi_{ij} = K_n \lambda_i \lambda_j \sum_{k=2}^{n} \frac{[k - n(p_i + p_j)]L_{n-k,ij}}{n^{k-2}}$$

PPS Brewer's Method (n=2)

Brewer's (1963) method is a special case of Sampford's method when n=2.

PPS Murthy's Method (n=2)

Murthy's (1957) method selects two units out of N population units with probabilities proportional to size without replacement.

- Inclusion probability of i-th unit = $p_i\left(1-\frac{p_i}{1-p_i}+\sum_{k=1}^{N}\frac{p_k}{1-p_k}\right)$
- Sampling weight of i-th unit = inverse of inclusion probability

Algorithm

1. Select first unit from the population with probabilities p_k, k=1,...,N.
2. If the first selected unit has index i, then select second unit with probabilities $p_k/(1-p_i)$, $k \neq i$.

Joint Inclusion Probabilities

The joint inclusion probabilities of population unit i and j is given by

$$\pi_{ij} = p_i p_j (2-p_i-p_j)/(1-p_i)(1-p_j).$$

Saved Variables

STATGEPOPSIZE

STAGEPOPSIZE saves the population sizes of each stratum in a given stage.

STAGESAMPSIZE

STAGESAMPSIZE saves the actual sample sizes of each stratum in a given stage. See the "Sample Size" section for details on sample size calculations.

STAGESAMPRATE

STAGESAMPRATE saves the actual sampling rate of each stratum in a given stage. It is computed by dividing the actual sample size by the population size. Due to the use of rounding and application of RATEMINSIZE and RATEMAXSIZE on sample size, the resulting STAGESAMPRATE may be different from sampling rate specified by the user.

STAGEINCLPROB

Stage inclusion probabilities depend on the selection method. The formulae are given in the individual sections of each selection method.

STAGEWEIGHT

It is equal to the inverse of stage inclusion probabilities.

SAMPLEWEIGHT

It is the product of previous weight (if specified) and all the stage weights.

STAGEHITS

This is the number of times a unit is selected in a given stage. When a WOR method is used the value is always 0 or 1. When a WR method is used it can be any nonnegative integer.

SAMPLEHITS

This is the number of times an ultimate sampling unit is selected. It is equal to STAGEHITS of the last specified stage.

STAGEINDEX

It is an index variable used to differentiate duplicated sampling units resulted from sampling with replacement. STAGEINDEX ranges from one to number of hits of a selected unit.

References

Brewer, K.W.R. (1963). A Model of Systematic Sampling with Unequal Probabilities. *Australian Journal of Statistics*, 5, 93 -105.

Chromy, J.R. (1979). Sequential Sample Selection Methods. *Proceedings of the American Statistical Association, Survey Research Methods Section*, 401 -406.

Fan, C.T., Muller, M.E., and Rezucha, I. (1962). Development of Sampling Plans by Using Sequential (Item by Item) Selection Techniques and Digital Computers. *Journal of the American Statistical Association*, 57, 387 -402.

Fox, D.R. (1989). Computer Selection of Size-Biased Samples. *The American Statistician*, 43(3), 168 -171.

Golmant, J. (1990). Correction: Computer Selection of Size-Biased Samples. *The American Statistician*, 44(2), 194.

Hanurav, T.V. (1967). Optimum Utilization of Auxiliary Information: π_{ps} Sampling of Two Units from a Stratum. *Journal of the Royal Statistical Society, Series B*, 29, 374 -391.

Hansen, M.H., and Hurwitz, W.N. (1943). On the theory of sampling from finite populations. Annals of Mathematical Statistics, 14, 333-362.

Murthy, M.N. (1957). Ordered and Unordered Estimators in Sampling Without Replacement. *Sankhya*, 18, 379 -390.

Sampford, M.R. (1967). On Sampling without Replacement with Unequal Probabilities of Selection. *Biometrika*, 54, 499 -513.

Vijayan, K. (1968). An Exact π_{ps} Sampling Scheme: Generalization of a Method of Hanurav. *Journal of the Royal Statistical Society, Series B*, 30, 556 -566.

Watts, D.L. (1991). Correction: Computer Selection of Size-Biased Samples. *The American Statistician*, 45(2), 172.