AN ALGORITHM FOR CALCULATING JOINT INCLUSION PROBABILITIES UNDER PPS SYSTEMATIC SAMPLING

Susan J. Pinciaro

U.S. Department of Transportation

Given the following definitions:

N = the number of units in the population

n = the number of units in the sample

x_k = the value of the variable x on the k^{th} unit in the population, where k refers to the position held by the unit in a given ordering of the population. The variable x is the "measure of size' variable.

$X = \sum_{k=1}^{N} x_k$ = the total of the x values in the population

$X_{[\ell]} = \sum_{k=1}^{\ell} x_k$ = the total of the x values through unit number ℓ of the population, under the given ordering

$s = X/n$ = the sampling interval

m_{ij} = the number of PPS-systematic samples of size n (out of all possible PPS - systematic samples of size n) in which units numbered i and j occur jointly.

The following algorithm finds the joint inclusion probability, ($\pi_{ij} = m_{ij} \cdot n/X$) for any element i and j ($j=1....N,1....j$) in the population, where i and j refer to the positions held by the specific units in a given ordering of the population. It is only necessary to find m_{ij} for all $i<j$.

The algorithm:

Step 1. Determine whether i and j are too close in the ordering to appear in sample together.

If $d = \left[\left(x_{[j]} - x_{[i-1]}\right) - s\right] \leq 0$, then $m_{ij} = 0$

If $d = \left[\left(x_{[j]} - x_{[i-1]}\right) - s\right] > 0$, proceed

Step 2. For pairs i and j with $d>0$, remove multiple of s from the "distance" between units [i-1] and [j], leaving the remainder, r

$r = \mathrm{MOD}\left[\left(x_{[j-1]} - x_{[i-1]}\right), s\right]$

Step 3. Compare the relative sizes of $(r + x_j)$ and s, defining a measure called a.

If $[(r+x_j)-s] \geq 1$, then $a = (r + x_j)-s$

If $[(r+x_j)-s] < 1$, then $a=0$

Step 4. It is the relative sizes of x_i and r which determines m_{ij}.

If $(x_i-r) = 0$, then $m_{ij} = a$

If $(x_i-r) < 0$, then $m_{ij} = \min(a, x_i)$

If $(x_i-r) > 0$, calculate $t =$

$\min[(x_i-r), x_j]$

If $t = x_j$, then $m_{ij} = x_j$

If $t = x_i-r$, then $m_{ij} = x_i-r + a$

If $x_i-r = x_j$, then $m_{ij} = x_j$

Note 1. The algorithm assumes that certainty units (all units k such that $x_k \geq s$) have been removed. If such cases have not been removed, it is necessary to insert the following step before Step 1.

Step 0.

if $x_i \geq s$ and $x_j < s$, then $m_{ij} = x_j$

if $x_i < s$ and $x_j \geq s$, then $m_{ij} = x_i$

if $x_i \geq s$ and $x_j \geq s$, then $m_{ij} = s$

if $x_i < s$ and $x_j < s$, proceed to Step 1.

Note 2. There should be no units with measure of size equal to zero in the population. If, however, the file has not been edited, and there are units with $x = 0$, then the algorithm calculates $m_{ij} = 0$ for all pairs containing such units.