Sampling Probability Proportional to Size

SurvMeth/Surv 625: Applied Sampling

Yajuan Si

University of Michigan, Ann Arbor

2/26/25

Controlling sample size

- Unequal-size cluster sampling results in potentially large variation in the sample size

Controlling sample size

- Unequal-size cluster sampling results in potentially large variation in the sample size
    - A ratio mean for design-consistent estimation

Controlling sample size

- Unequal-size cluster sampling results in potentially large variation in the sample size
    - A ratio mean for design-consistent estimation
- Desire more control over sample size. Why?

Controlling sample size

- Unequal-size cluster sampling results in potentially large variation in the sample size
  - A ratio mean for design-consistent estimation
- Desire more control over sample size. Why?
  - Bias and variance of the ratio mean

Controlling sample size

- Unequal-size cluster sampling results in potentially large variation in the sample size
  - A ratio mean for design-consistent estimation
- Desire more control over sample size. Why?
  - Bias and variance of the ratio mean
- Equalize workloads controlling survey costs

Controlling sample size

- Unequal-size cluster sampling results in potentially large variation in the sample size
    - A ratio mean for design-consistent estimation
- Desire more control over sample size. Why?
    - Bias and variance of the ratio mean
- Equalize workloads controlling survey costs
- Avoid large clusters dominating analysis

Controlling sample size

- Unequal-size cluster sampling results in potentially large variation in the sample size
    - A ratio mean for design-consistent estimation
- Desire more control over sample size. Why?
    - Bias and variance of the ratio mean
- Equalize workloads controlling survey costs
- Avoid large clusters dominating analysis
    - Clusters may be a unit of analysis

Controlling sample size

- Unequal-size cluster sampling results in potentially large variation in the sample size
    - A ratio mean for design-consistent estimation
- Desire more control over sample size. Why?
    - Bias and variance of the ratio mean
- Equalize workloads controlling survey costs
- Avoid large clusters dominating analysis
    - Clusters may be a unit of analysis
    - Equal size samples from each cluster provide more efficient analysis

Sampling probability proportional to size (PPS)

- Solution to achieve epsem and equal subsample sizes at the same time
- Select a two stage sample such that we sample the same number of subsamples from each cluster with epsem
  - We want an overall epsem rate $f = \frac{m_0}{M_0}$ such that the sampling rate in the second stage is m/M_i, i.e., always sampling m subsamples

$$f = f_i f_{j|i} = f_i \frac{m}{M_i} = \frac{m_0}{M_0}$$

  - Solving $f_i = \frac{m_0 M_i}{M_0 m} = \frac{n*m*M_i}{\sum_i M_i m} = \frac{n M_i}{\sum_i M_i}$
- Select clusters with probabilities proportionate to their size M_i

R code

```
library(sampling)
# selection of a sample with expected size equal to 200
# the inclusion probabilities are proportional to the average
data(belgianmunicipalities); attach(belgianmunicipalities);
pik=sampling::inclusionprobabilities(averageincome,200)
# draws a sample s using systematic sampling
s=UPsystematic(pik)
```

PPS: Implementation

- Overall selection probability is $f = \frac{nM_i}{\sum_i M_i} * \frac{m}{M_i}$, epsem

PPS: Implementation

- Overall selection probability is $f = \frac{nM_i}{\sum_i M_i} * \frac{m}{M_i}$, epsem
- Selection probability varies at both first & second stages

PPS: Implementation

- Overall selection probability is $f = \frac{nM_i}{\sum_i M_i} * \frac{m}{M_i}$, epsem
- Selection probability varies at both first & second stages
    - Large clusters selected with high probabilities, subsampled at low rate

PPS: Implementation

- Overall selection probability is $f = \frac{nM_i}{\sum_i M_i} * \frac{m}{M_i}$, epsem
- Selection probability varies at both first & second stages
  - Large clusters selected with high probabilities, subsampled at low rate
  - Small clusters selected with low probabilities, subsampled at high rate

How do we actually perform PPS sampling in practice?

- Two ways:

How do we actually perform PPS sampling in practice?

- Two ways:
    - Simple replicated subsampling; random and with replacement (but what about duplicates?)

How do we actually perform PPS sampling in practice?

- Two ways:
  - Simple replicated subsampling; random and with replacement (but what about duplicates?)
  - Systematic sampling without replacement

How do we actually perform PPS sampling in practice?

- Two ways:
  - Simple replicated subsampling; random and with replacement (but what about duplicates?)
  - Systematic sampling without replacement
- For both approaches, we need to accumulate the sizes of the clusters; each cluster in the list (frame) will have an associated cumulative size

How do we actually perform PPS sampling in practice?

- Two ways:
  - Simple replicated subsampling; random and with replacement (but what about duplicates?)
  - Systematic sampling without replacement
- For both approaches, we need to accumulate the sizes of the clusters; each cluster in the list (frame) will have an associated cumulative size
- All of this requires knowing the exact size of each cluster!

How do we actually perform PPS sampling in practice?

- Two ways:
  - Simple replicated subsampling; random and with replacement (but what about duplicates?)
  - Systematic sampling without replacement
- For both approaches, we need to accumulate the sizes of the clusters; each cluster in the list (frame) will have an associated cumulative size
- All of this requires knowing the exact size of each cluster!
- For systematic PPS sampling, we need a variance estimation model

Example

Unit	B_α	Cum. B_α	Unit	B_α	Cum. B_α
1	443	443	6	291	1692
2	162	605	7	64	1756
3	127	732	8	70	1826
4	554	1286	9	232	2058
5	115	1401	10	102	2160

Systematic PPS

- Compute zone size: $M/n = 2160/2 = 1080$, where M is the total size and n is the number of selected clusters

Systematic PPS

- Compute zone size: $M/n = 2160/2 = 1080$, where M is the total size and n is the number of selected clusters
    - Select one subsample from each zone size 1080

Systematic PPS

- Compute zone size: $M/n = 2160/2 = 1080$, where M is the total size and n is the number of selected clusters
    - Select one subsample from each zone size 1080
    - Zone boundary falls 348 "elements" (measure units) through Unit 4, which falls in 2 zones

Systematic PPS

- Compute zone size: $M/n = 2160/2 = 1080$, where M is the total size and n is the number of selected clusters
  - Select one subsample from each zone size 1080
  - Zone boundary falls 348 "elements" (measure units) through Unit 4, which falls in 2 zones
- Select Units by selecting RN from 1 to 1080

Systematic PPS

- Compute zone size: $M/n = 2160/2 = 1080$, where M is the total size and n is the number of selected clusters
  - Select one subsample from each zone size 1080
  - Zone boundary falls 348 “elements” (measure units) through Unit 4, which falls in 2 zones
- Select Units by selecting RN from 1 to 1080
  - Identify selection from cumulative counts

Systematic PPS

- Compute zone size: $M/n = 2160/2 = 1080$, where M is the total size and n is the number of selected clusters
  - Select one subsample from each zone size 1080
  - Zone boundary falls 348 "elements" (measure units) through Unit 4, which falls in 2 zones
- Select Units by selecting RN from 1 to 1080
  - Identify selection from cumulative counts
  - Add the interval $k = 1080$ to the RN

Systematic PPS

- Compute zone size: $M/n = 2160/2 = 1080$, where M is the total size and n is the number of selected clusters
  - Select one subsample from each zone size 1080
  - Zone boundary falls 348 "elements" (measure units) through Unit 4, which falls in 2 zones
- Select Units by selecting RN from 1 to 1080
  - Identify selection from cumulative counts
  - Add the interval $k = 1080$ to the RN
  - Identify next selected cluster

Systematic PPS

- Compute zone size: $M/n = 2160/2 = 1080$, where M is the total size and n is the number of selected clusters
  - Select one subsample from each zone size 1080
  - Zone boundary falls 348 "elements" (measure units) through Unit 4, which falls in 2 zones
- Select Units by selecting RN from 1 to 1080
  - Identify selection from cumulative counts
  - Add the interval $k = 1080$ to the RN
  - Identify next selected cluster
  - Select one subsample from each cluster at the rate m/M_i

Example cont.

- Suppose $RN = 804$, select Cluster 4; Since $RN + k = 804 + 1080 = 1884$, select Cluster 9.

Example cont.

- Suppose $RN = 804$, select Cluster 4; Since $RN + k = 804 + 1080 = 1884$, select Cluster 9.
- For Cluster 4, the selection probability $f_4 = \frac{nM_4}{\sum M_i} = \frac{M_4}{\sum M_i/n} = \frac{M_i}{k} = \frac{554}{1080}$

Example cont.

- Suppose $RN = 804$, select Cluster 4; Since $RN + k = 804 + 1080 = 1884$, select Cluster 9.
- For Cluster 4, the selection probability $f_4 = \frac{nM_4}{\sum M_i} = \frac{M_4}{\sum M_i/n} = \frac{M_i}{k} = \frac{554}{1080}$
- Within Cluster 4, the subsampling rate $\frac{m}{M_i} = \frac{18}{554}$

Estimated size measures

- Suppose that the exact count of elements, M_i, in each cluster is unknown

Estimated size measures

- Suppose that the exact count of elements, M_i, in each cluster is unknown
- But we know a population measure that approximates the number of units (that is, the size) in each cluster, Measure of Size, MOS_i

Estimated size measures

- Suppose that the exact count of elements, M_i, in each cluster is unknown
- But we know a population measure that approximates the number of units (that is, the size) in each cluster, Measure of Size, MOS_i
- Example:

Estimated size measures

- Suppose that the exact count of elements, M_i, in each cluster is unknown
- But we know a population measure that approximates the number of units (that is, the size) in each cluster, Measure of Size, MOS_i
- Example:
    - Do not know the current exact count of housing units for each unit

Estimated size measures

- Suppose that the exact count of elements, M_i, in each cluster is unknown
- But we know a population measure that approximates the number of units (that is, the size) in each cluster, Measure of Size, MOS_i
- Example:
  - Do not know the current exact count of housing units for each unit
  - Do know the number counted for each Unit at the last payroll one month ago

PPeS sampling: Implementation

- Probabilities Proportionate to estimated Size (PPeS): Overall epsem design in two stages

PPeS sampling: Implementation

- Probabilities Proportionate to estimated Size (PPeS): Overall epsem design in two stages
- Select clusters with the rate $\frac{nMOS_i}{\sum_i MOS_i}$

PPeS sampling: Implementation

- Probabilities Proportionate to estimated Size (PPeS): Overall epsem design in two stages
- Select clusters with the rate $\frac{nMOS_i}{\sum_i MOS_i}$
    - Over-sample large clusters relative to small

PPeS sampling: Implementation

- Probabilities Proportionate to estimated Size (PPeS): Overall epsem design in two stages
- Select clusters with the rate $\frac{nMOS_i}{\sum_i MOS_i}$
  - Over-sample large clusters relative to small
- Then subsample elements at rate $\frac{m^*}{MOS_i}$

PPeS sampling: Implementation

- Probabilities Proportionate to estimated Size (PPeS): Overall epsem design in two stages
- Select clusters with the rate $\frac{nMOS_i}{\sum_i MOS_i}$
  - Over-sample large clusters relative to small
- Then subsample elements at rate $\frac{m^*}{MOS_i}$
- Why m^*?

PPeS sampling: Implementation

- Probabilities Proportionate to estimated Size (PPeS): Overall epsem design in two stages
- Select clusters with the rate $\frac{nMOS_i}{\sum_i MOS_i}$
  - Over-sample large clusters relative to small
- Then subsample elements at rate $\frac{m^*}{MOS_i}$
- Why m^*?
  - The actual number of second stage units selected is unknown; it's a target subsample size

PPeS sampling: Implementation

- Probabilities Proportionate to estimated Size (PPeS): Overall epsem design in two stages
- Select clusters with the rate $\frac{nMOS_i}{\sum_i MOS_i}$
  - Over-sample large clusters relative to small
- Then subsample elements at rate $\frac{m^*}{MOS_i}$
- Why m^*?
  - The actual number of second stage units selected is unknown; it's a target subsample size
  - If $MOS_i = M_i$, then $m^* = m$

Example

Unit	Last payroll	Now
1	443	460
2	162	172
3	127	130
4	554	554
5	115	125
6	291	310
7	64	68
8	70	74
9	232	246
10	102	141
Total	2160	2280

Example cont.

Unit	Last payroll	Cumulative
1	443	443
2	162	605
3	127	732
4	554	1286
5	115	1401
6	291	1692
7	64	1756
8	70	1826
9	232	2058
10	102	2160

PPeS: Implementation

- Suppose $m^* = 18, n = 2$, and $\sum MOS_i = 2160$

PPeS: Implementation

- Suppose $m^* = 18, n = 2$, and $\sum MOS_i = 2160$
- $n = 2$ clusters are selected

PPeS: Implementation

- Suppose $m^* = 18, n = 2$, and $\sum MOS_i = 2160$
- $n = 2$ clusters are selected
- One subsample of expected size $m^* = 18$ from each cluster

PPeS: Implementation

- Suppose $m^* = 18, n = 2$, and $\sum MOS_i = 2160$
- $n = 2$ clusters are selected
- One subsample of expected size $m^* = 18$ from each cluster
- Overall $f = \frac{2MOS_i}{\sum MOS_i} \frac{18}{MOS_i} = \frac{36}{2160}$

PPeS: Implementation

- Suppose $m^* = 18, n = 2$, and $\sum MOS_i = 2160$
- $n = 2$ clusters are selected
- One subsample of expected size $m^* = 18$ from each cluster
- Overall $f = \frac{2MOS_i}{\sum MOS_i} \frac{18}{MOS_i} = \frac{36}{2160}$
- With PPeS sampling, subsampling at a specified rate and not selecting a fixed number of elements from each selected cluster

PPeS example cont.

- Suppose $RN = 702$, select Cluster 3; Since $RN + k = 1782$, select Cluster 8.

PPeS example cont.

- Suppose $RN = 702$, select Cluster 3; Since $RN + k = 1782$, select Cluster 8.
- Within Cluster 3, $\frac{m^*}{MOS_3} = \frac{18}{127} = \frac{1}{7.056}$

PPeS example cont.

- Suppose $RN = 702$, select Cluster 3; Since $RN + k = 1782$, select Cluster 8.
- Within Cluster 3, $\frac{m^*}{MOS_3} = \frac{18}{127} = \frac{1}{7.056}$
- If $MOS_3 = M_3$, an exact sample size of 18 will be selected. But $MOS_3 \approx M_3$

PPeS example cont.

- Suppose $RN = 702$, select Cluster 3; Since $RN + k = 1782$, select Cluster 8.
- Within Cluster 3, $\frac{m^*}{MOS_3} = \frac{18}{127} = \frac{1}{7.056}$
- If $MOS_3 = M_3$, an exact sample size of 18 will be selected. But $MOS_3 \approx M_3$
- Since $M_3 = 130$, we have the expected subsample size $x_i = \frac{1}{7.056} * 130 = 18.425$

PPeS example cont.

- Suppose $RN = 702$, select Cluster 3; Since $RN + k = 1782$, select Cluster 8.
- Within Cluster 3, $\frac{m^*}{MOS_3} = \frac{18}{127} = \frac{1}{7.056}$
- If $MOS_3 = M_3$, an exact sample size of 18 will be selected. But $MOS_3 \approx M_3$
- Since $M_3 = 130$, we have the expected subsample size $x_i = \frac{1}{7.056} * 130 = 18.425$
- With a fractional interval 7.056, select a sample of 18 employees with probability 0.575 or a sample of 19 employees with probability 0.425

Stratification

- Independent sampling across strata
- Within strata, specify $\sum_i MOS_i$ n, m^*, etc.,

$$f_h = \frac{n_h MOS_{hi}}{\sum_{i \in h} MOS_{hi}} \frac{m_h^*}{MOS_{hi}} = \frac{n_h m_h^*}{\sum_{i \in h} MOS_{hi}}$$

- Retain epsem for stratified PPS sampling across strata $f = f_h$ for all h

Implicit stratification

- Systematic PPeS sampling implicitly stratifies by selecting within each zone one subsample size m^*
- Stratification notation not necessary with this design
- Zone size is $\frac{\sum_{i \in h} MOS_{hi}}{n_h}$

Example: Stratified PPeS

Stratum 1		Stratum II	
Unit	Mos	Unit	Mos
1	443	5	115
2	162	6	291
3	127	7	64
4	554	8	70
		9	232
		10	102
Total	1286		874

Example cont.

- Select $n = 4$ clusters with a subsample size of $m^* = 18$ from $\sum_i MOS_i = 2160$, then

$$f = \frac{nm^*}{\sum_i MOS_i} = \frac{4 * 18}{2160} = 1/30$$

with a zone size of $2160/4 = 540$

Example cont.

- Select $n = 4$ clusters with a subsample size of $m^* = 18$ from $\sum_i MOS_i = 2160$, then

$$f = \frac{nm^*}{\sum_i MOS_i} = \frac{4 * 18}{2160} = 1/30$$

with a zone size of $2160/4 = 540$

- Paired selection from the two strata: $n_h = 2$ with

$$f_1 = \frac{n_1 * m_1^*}{\sum_{i \in h=1} MOS_{1i}} = \frac{2 * m_1^*}{1286} = 1/30$$

$$f_2 = \frac{n_2 * m_2^*}{\sum_{i \in h=2} MOS_{2i}} = \frac{2 * m_2^*}{874} = 1/30$$

Example cont.

- Select $n = 4$ clusters with a subsample size of $m^* = 18$ from $\sum_i MOS_i = 2160$, then

$$f = \frac{nm^*}{\sum_i MOS_i} = \frac{4 * 18}{2160} = 1/30$$

  with a zone size of $2160/4 = 540$

- Paired selection from the two strata: $n_h = 2$ with

$$f_1 = \frac{n_1 * m_1^*}{\sum_{i \in h=1} MOS_{1i}} = \frac{2 * m_1^*}{1286} = 1/30$$
$$f_2 = \frac{n_2 * m_2^*}{\sum_{i \in h=2} MOS_{2i}} = \frac{2 * m_2^*}{874} = 1/30$$

- Adjusting the subsample sizes across strata with $m_1^* = 21.43$ and $m_2^* = 14.57$

Example cont.

- Select $n = 4$ clusters with a subsample size of $m^* = 18$ from $\sum_i MOS_i = 2160$, then

$$f = \frac{nm^*}{\sum_i MOS_i} = \frac{4 * 18}{2160} = 1/30$$

with a zone size of $2160/4 = 540$

- Paired selection from the two strata: $n_h = 2$ with

$$f_1 = \frac{n_1 * m_1^*}{\sum_{i \in h=1} MOS_{1i}} = \frac{2 * m_1^*}{1286} = 1/30$$

$$f_2 = \frac{n_2 * m_2^*}{\sum_{i \in h=2} MOS_{2i}} = \frac{2 * m_2^*}{874} = 1/30$$

- Adjusting the subsample sizes across strata with $m_1^* = 21.43$ and $m_2^* = 14.57$

- Select final subsamples with a fixed rate based on the actual M_i

Recall PPS: Implementation

- Overall selection probability is $f = f_i * f_{j|i} = \frac{nM_i}{\sum_i M_i} * \frac{m}{M_i}$, epsem

Recall PPS: Implementation

- Overall selection probability is $f = f_i * f_{j|i} = \frac{nM_i}{\sum_i M_i} * \frac{m}{M_i}$, epsem
- Selection probability varies at both first & second stages

Recall PPS: Implementation

- Overall selection probability is $f = f_i * f_{j|i} = \frac{nM_i}{\sum_i M_i} * \frac{m}{M_i}$, epsem
- Selection probability varies at both first & second stages
  - Large clusters selected with high probabilities, subsampled at low rate (what if $f_i \geq 1$?)

Recall PPS: Implementation

- Overall selection probability is $f = f_i * f_{j|i} = \frac{nM_i}{\sum_i M_i} * \frac{m}{M_i}$, epsem
- Selection probability varies at both first & second stages
  - Large clusters selected with high probabilities, subsampled at low rate (what if $f_i \geq 1$?)
  - Small clusters selected with low probabilities, subsampled at high rate (what if $f_{j|i} \geq 1$?)

Oversize units

- Oversize units are clusters for which MOS_{hi} is so large that the unit has a certain chance of selection

Oversize units

- Oversize units are clusters for which MOS_{hi} is so large that the unit has a certain chance of selection
  - Often so large that they will be selected multiple times in systematic PPeS

Oversize units

- Oversize units are clusters for which MOS_{hi} is so large that the unit has a certain chance of selection
  - Often so large that they will be selected multiple times in systematic PPeS
  - Eg., Cluster 4 has $MOS_{14} = 554 > 30 * 18$ and can be selected twice with probability of $554/540 = 1.0259$

Oversize units

- Oversize units are clusters for which MOS_{hi} is so large that the unit has a certain chance of selection
  - Often so large that they will be selected multiple times in systematic PPeS
  - Eg., Cluster 4 has $MOS_{14} = 554 > 30 * 18$ and can be selected twice with probability of $554/540 = 1.0259$
- If there are few such clusters and chances of multiple selections small leave them in the list. If one selected k times, select k (different) subsamples

Oversize units

- Oversize units are clusters for which MOS_{hi} is so large that the unit has a certain chance of selection
  - Often so large that they will be selected multiple times in systematic PPeS
  - Eg., Cluster 4 has $MOS_{14} = 554 > 30 * 18$ and can be selected twice with probability of $554/540 = 1.0259$
- If there are few such clusters and chances of multiple selections small leave them in the list. If one selected k times, select k (different) subsamples
- If there are many such clusters comprising a large share of the population, place in separate strata

Oversize units

- Oversize units are clusters for which MOS_{hi} is so large that the unit has a certain chance of selection
  - Often so large that they will be selected multiple times in systematic PPeS
  - Eg., Cluster 4 has $MOS_{14} = 554 > 30 * 18$ and can be selected twice with probability of $554/540 = 1.0259$
- If there are few such clusters and chances of multiple selections small leave them in the list. If one selected k times, select k (different) subsamples
- If there are many such clusters comprising a large share of the population, place in separate strata
  - Each such cluster is now a stratum, a cluster selected with certainty, i.e., self-representing units (SRU)

Oversize units

- Oversize units are clusters for which MOS_{hi} is so large that the unit has a certain chance of selection
  - Often so large that they will be selected multiple times in systematic PPeS
  - Eg., Cluster 4 has $MOS_{14} = 554 > 30 * 18$ and can be selected twice with probability of $554/540 = 1.0259$
- If there are few such clusters and chances of multiple selections small leave them in the list. If one selected k times, select k (different) subsamples
- If there are many such clusters comprising a large share of the population, place in separate strata
  - Each such cluster is now a stratum, a cluster selected with certainty, i.e., self-representing units (SRU)
  - Sampling rate(s) are applied directly within clusters

Undersize units

- Undersize units are clusters for which $MOS_{hi} < m_h^*$, implying sampling within cluster at rate > 1.

Undersize units

- Undersize units are clusters for which $MOS_{hi} < m_h^*$, implying sampling within cluster at rate > 1.
- All elements must be selected in the cluster

Undersize units

- Undersize units are clusters for which $MOS_{hi} < m_h^*$, implying sampling within cluster at rate > 1.
- All elements must be selected in the cluster
- Can include zero measure clusters

Undersize units

- Undersize units are clusters for which $MOS_{hi} < m_h^*$, implying sampling within cluster at rate > 1.
- All elements must be selected in the cluster
- Can include zero measure clusters

1. Link undersize units to form linked units of minimum sufficient size

Undersize units

- Undersize units are clusters for which $MOS_{hi} < m_h^*$, implying sampling within cluster at rate > 1.
- All elements must be selected in the cluster
- Can include zero measure clusters

1. Link undersize units to form linked units of minimum sufficient size
   - Can create clusters with greater heterogeneity

Undersize units

- Undersize units are clusters for which $MOS_{hi} < m_h^*$, implying sampling within cluster at rate > 1.
- All elements must be selected in the cluster
- Can include zero measure clusters

1. Link undersize units to form linked units of minimum sufficient size
    - Can create clusters with greater heterogeneity
    - In area sampling, link geographically contiguous units

Undersize units

- Undersize units are clusters for which $MOS_{hi} < m^*_h$, implying sampling within cluster at rate > 1.
- All elements must be selected in the cluster
- Can include zero measure clusters

1. Link undersize units to form linked units of minimum sufficient size
    - Can create clusters with greater heterogeneity
    - In area sampling, link geographically contiguous units
    - If numerous, place in separate stratum

Undersize units

- Undersize units are clusters for which $MOS_{hi} < m_h^*$, implying sampling within cluster at rate > 1.
- All elements must be selected in the cluster
- Can include zero measure clusters

1. Link undersize units to form linked units of minimum sufficient size
    - Can create clusters with greater heterogeneity
    - In area sampling, link geographically contiguous units
    - If numerous, place in separate stratum
2. Link before selection for the entire frame, especially if the frame is a computerized list

Undersize units

- Undersize units are clusters for which $MOS_{hi} < m^*_h$, implying sampling within cluster at rate > 1.
- All elements must be selected in the cluster
- Can include zero measure clusters

1. Link undersize units to form linked units of minimum sufficient size
   - Can create clusters with greater heterogeneity
   - In area sampling, link geographically contiguous units
   - If numerous, place in separate stratum
2. Link before selection for the entire frame, especially if the frame is a computerized list
3. Linking after selection

Linking after (during) selection

1. Identify selected unit. If the selected unit and next on the list are minimum sufficient size, STOP

Linking after (during) selection

1. Identify selected unit. If the selected unit and next on the list are minimum sufficient size, STOP
2. If selected unit or next are not of minimum sufficient size,

Linking after (during) selection

1. Identify selected unit. If the selected unit and next on the list are minimum sufficient size, STOP
2. If selected unit or next are not of minimum sufficient size,
    a. Move forward in list until first unit of minimum sufficient size is encountered.

Linking after (during) selection

1. Identify selected unit. If the selected unit and next on the list are minimum sufficient size, STOP
2. If selected unit or next are not of minimum sufficient size,
    a. Move forward in list until first unit of minimum sufficient size is encountered.
    b. Cumulate units backwards until a linked unit of minimum sufficient size is created.

Linking after (during) selection

1. Identify selected unit. If the selected unit and next on the list are minimum sufficient size, STOP
2. If selected unit or next are not of minimum sufficient size,
   a. Move forward in list until first unit of minimum sufficient size is encountered.
   b. Cumulate units backwards until a linked unit of minimum sufficient size is created.
   c. Continue process until the selected unit is linked.

Example: Linking after selection

- When the sampled cluster on the sorted list and the immediate next cluster are BOTH of sufficient size (50)

Example 1

ID	STATE	COUNTY	TRACT	BLKGRP	BLOCK	Housing units: Occupied
346	**26**	**077**	**000500**	**1**	**1002**	**286**
347	26	077	000500	2	2000	73

- ID #346 is the selected block
- No linking necessary: the selected block and the immediate subsequent block are both of sufficient size

Example cont.

- When the sampled cluster on the sorted list is of sufficient size, but the immediate next cluster is NOT of sufficient size

Example 2

ID	STATE	COUNTY	TRACT	BLKGRP	BLOCK	Housing units: Occupied
320	**26**	**077**	**000300**	**4**	**4002**	**136**
321	26	077	000300	4	4003	14
322	26	077	000300	4	4004	32
323	26	077	000300	4	4005	19
324	26	077	000300	4	4006	28
325	26	077	000300	4	4007	20
326	26	077	000300	4	4008	16
327	26	077	000300	4	4009	0
328	26	077	000300	4	4010	17
329	26	077	000300	5	5000	56

- ID #320 is selected block (has sufficient size, 136 housing units)
- Immediate subsequent block is NOT of sufficient size (14)
- Go down the list until ID #329 (next block of sufficient size), and link <u>backwards</u> to form units of sufficient size: 325-328 (53 units), 322-324 (79 units), 320-321 (150 units)
- We would then subsample from the two linked blocks that include our sampled block (320-321) at the second stage

Example cont.

- When the sampled cluster on the sorted list is NOT of sufficient size

Example 3

ID	STATE	COUNTY	TRACT	BLKGRP	BLOCK	Housing units: Occupied
50	26	077	000100	3	3000	62
51	**26**	**077**	**000100**	**3**	**3001**	**4**
52	26	077	000100	3	3002	3
53	26	077	000100	3	3003	2
54	26	077	000100	3	3004	9
55	26	077	000100	3	3005	1
56	26	077	000100	3	3006	0
57	26	077	000100	3	3007	4
58	26	077	000100	3	3008	58

- ID #51 is selected block (not of sufficient size)
- Go down list until next unit of sufficient size (ID #58)
- Link backwards, forming units of sufficient size (e.g., if the number of housing units in ID #55 was 41 instead of 1, you would form one unit including ID #54-57, which would have 54 housing units total, then proceed with ID #53, 52, etc.)
- In this case, we combine ID #57 all the way through ID #50, so that the selected block is part of a linked unit with minimum size; then we would subsample from that linked unit

Summary

- PPS sampling goals:

  1). Maintain epsem to avoid weights;

  2). Control over sample sizes across clusters that will minimize the bias and variance of the ratio mean estimator
- PPeS can maintain epsem across two stages of selection, using sampling rates defined by the same fractions, with the target m^*
- Need to handle oversize or undersize clusters