

STA 304H1F-1003H Fall 2019

## **Week 11 - Cluster Sampling, Chapter 8**

## Cluster Sampling, Example

Consider that we want to estimate health insurance coverage in Baltimore city. We could take a random sample of 100 households(HH). In that case, we need a sampling list of Baltimore HHs.

If the list is not available, we need to conduct a census of HHs. The complete coverage of Baltimore city is required so that all HHs are listed, which could be expensive. Furthermore, since our sample size is small compared to the numbers of total HHs, we need to sample only few, say one or two, in each block (subdivisions).

Alternatively, we could select 5 blocks (say the city is divided into 200 blocks), and in each block interview 20 HHs. We need to construct HH listing frame only for 5 blocks (less time and costs needed). Furthermore, by limiting the survey to a smaller area, additional costs will be saved during the execution of interviews.

Such sampling strategy is known as "cluster sampling."

The blocks are “Primary Sampling Units” (PSU) – the clusters.

The households are “Secondary Sampling Units” (SSU).

# Cluster Sampling, Definition

## Definition:

A cluster sample is a probability sample in which each sampling unit is a collection or cluster, or element

In cluster sampling, cluster, i.e., a group of population elements, constitutes the sampling unit, instead of a single element of the population.

Need to consider the sampling order:

**Primary sampling units (PSU): clusters**

**Secondary sampling units (SSU): households/individual elements**

We may select the PSU's by using a specific element sampling techniques, such as simple random sampling, or systematic sampling.

We may select all SSU's for convenience or few by using a specific element sampling techniques (such as simple random sampling, or systematic sampling).

# How to Draw a Cluster Sample

## 1. Simple one-stage cluster sample:

List all the clusters in the population, and

from the list, select the clusters – usually with simple random sampling (SRS) strategy.

All units (elements) in the sampled clusters are selected for the survey.

## 2. Simple two-stage cluster sample:

List all the clusters in the population, and

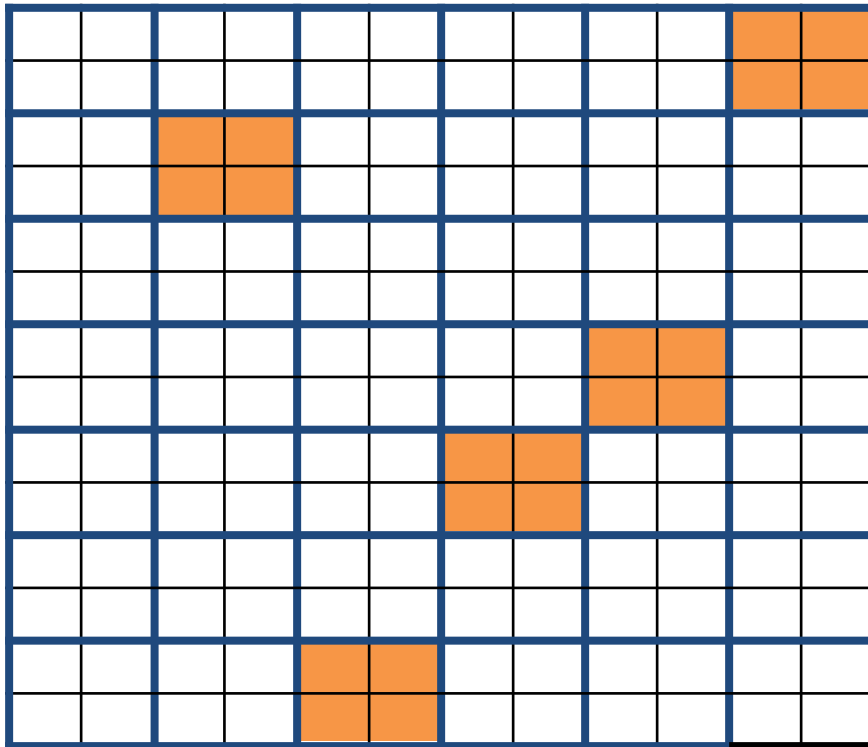
First, select the clusters, usually by simple random sampling (SRS).

The units (elements) in the selected clusters of the first-stage are then sampled in the second-stage, usually by simple random sampling (or often by systematic sampling).

# Cluster Sampling, One Stage (I)

## Remainder on basics

**Basics:** The population is divided into large number of (small) groups (clusters), equal or nonequal. Clusters are selected “at random”. Sample: All elements from selected clusters.

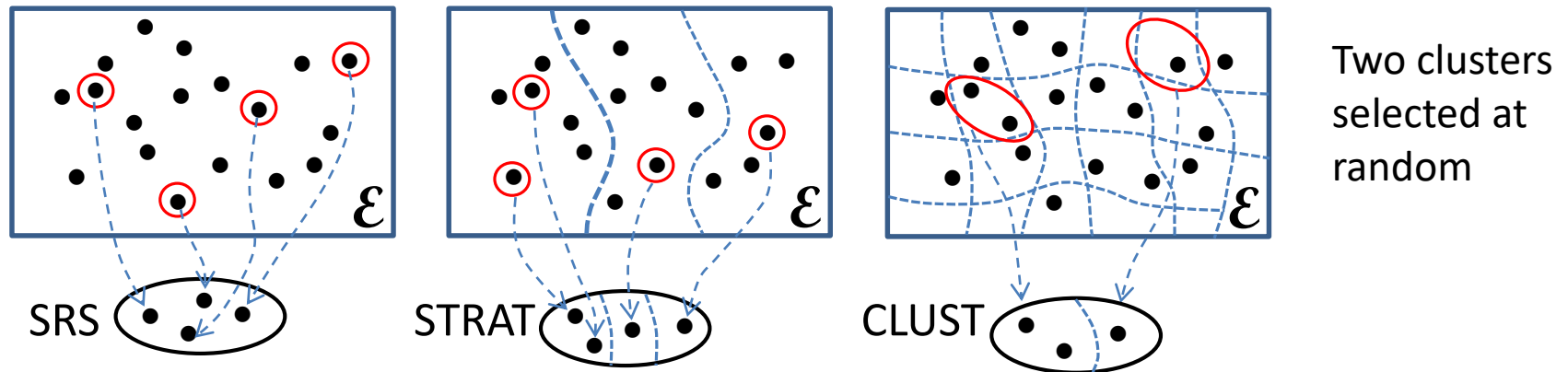


Population size: 168.  
42 clusters of size 4.  
Sample size: 20.  
5 clusters selected  
“at random” from  
42 clusters.



# Cluster Sampling (I)

## General considerations, definition (I)



Large number of (small) groups

- SRS of  $n$  groups: All elements from selected groups are in the sample
- Groups are called clusters
- Sampling units: Clusters
- Design: One-stage cluster sampling

# Cluster Sampling (I)

## General considerations, notation (II)

$N$  – number of clusters in the population (# of sampling units)

$n$  – number of clusters selected in the sample

$m_i$  –  $i$ th clusters size

$M$  – population size,  $M = m_1 + m_2 + \dots + m_N = \sum_{i=1}^N m_i$

$\bar{M}$  – average cluster size  $\bar{M} = \frac{M}{N} = \frac{1}{N} \sum_{i=1}^N m_i$

$i$ th cluster elements:  $y_{i1}, y_{i2}, \dots, y_{im_i}$

Cluster total  $\tau_i = \sum_{j=1}^{m_i} y_{ij} = y_i$ , cluster mean  $\mu_i = \bar{y}_i = \frac{y_i}{m_i}$

$$\tau = \tau_y = \sum_{i,j} y_{ij} = \sum_{i=1}^N \tau_i = \sum_{i=1}^N y_i, \mu = \mu_y = \frac{\tau_y}{M} = \frac{1}{M} \sum_{i,j} y_{ij}, \tau_y = M\mu_y$$

# Cluster Sampling (I)

## General considerations, notation (III)

Summary, population:

Cluster	1	2	...	$N$	Pop
Size	$m_1$	$m_2$	...	$m_N$	$M$
Total	$\tau_1 = y_1$	$\tau_2 = y_2$	...	$\tau_N = y_N$	$\tau$
Mean	$\mu_1 = \bar{y}_1$	$\mu_2 = \bar{y}_2$	...	$\mu_N = \bar{y}_N$	$\mu$

**Sample:**  $n$  – clusters, sizes  $m_1, m_2, \dots, m_n$ , totals  $y_1, y_2, \dots, y_n$

Means  $\bar{y}_1, \bar{y}_2, \dots, \bar{y}_n$   $\sum_{i=1}^n m_i$  - realized sample size

$\bar{m} = \frac{1}{n} \sum_{i=1}^n m_i$  - average cluster size in the sample

$\bar{y}_t = \frac{1}{n} \sum_{i=1}^n y_i$  - average cluster total in the sample

Sizes  $m_1, m_2, \dots, m_n$ , and totals  $y_1, y_2, \dots, y_n$  are random!



# Cluster Sampling (II)

## Inference: Ratio estimation of mean and total (I)

We consider two methods of estimation:

- 1) Ratio estimation
- 2) Unbiased estimation

### 1) Ratio estimation

Sample mean:

$$\bar{y} = \frac{\sum_{i=1}^n y_i}{\sum_{i=1}^n m_i}$$

$\bar{y}$  – ratio of two  
random variables

$$\hat{\mu} = \bar{y} , \hat{\tau} = M\hat{\mu} = M\bar{y}$$

- Ratio type estimators, biased

- To estimate  $\mu$ , we don't need  $M$ ,
- To estimate  $\tau$ , we need  $M$

# Cluster Sampling (II)

## Inference: Ratio estimation of mean and total (II)

We can do more:

$$\bar{\tau} = \mu_t = \frac{1}{N} \sum_{i=1}^N \tau_i = \frac{1}{N} \sum_{i=1}^N y_i \quad \text{Average cluster total in the population}$$

$$\boxed{\hat{\tau} = \bar{y}_t = \frac{1}{n} \sum_{i=1}^n y_i, \quad \hat{M} = \bar{m}} \quad \text{both unbiased, but}$$

$$\hat{\mu} = \bar{y} = \frac{\sum_{i=1}^n y_i}{\sum_{i=1}^n m_i} = \frac{\frac{1}{n} \sum_{i=1}^n y_i}{\frac{1}{n} \sum_{i=1}^n m_i} = \frac{\bar{y}_t}{\bar{m}} = \frac{\bar{y}_t}{\hat{M}} \quad \text{biased}$$

Still, we know, the bias is usually small, unless for small  $n$

# Cluster Sampling (II)

## Comparison with ratio estimation from Ch. 6

Ratio	Cluster
$y_i$	$y_i = \tau_i$
$x_i$	$m_i$
$\mu_x$	$\bar{M}$
$\tau_x$	$M$

$$r = \frac{\sum y_i}{\sum x_i} = \frac{\bar{y}}{\bar{x}} \text{ - in "ratio" case}$$

$$\bar{y} = \frac{\sum y_i}{\sum m_i} = \frac{\bar{y}_t}{\bar{m}} \text{ - in "cluster" case}$$

Following Ch. 6 results:

$$Var(\hat{\mu}) = Var(\bar{y}) \approx \frac{1}{\bar{M}^2} \frac{N-n}{N-1} \frac{\sigma_r^2}{n}, \sigma_r^2 = \frac{1}{N} \sum_{i=1}^N (y_i - \mu m_i)^2$$

$$\hat{Var}(\hat{\mu}) = \frac{1}{\bar{M}^2} \frac{N-n}{N} \frac{S_r^2}{n} \quad S_r^2 = \frac{1}{n-1} \sum_{i=1}^n (y_i - \hat{\mu} m_i)^2 = \frac{1}{n-1} \sum_{i=1}^n m_i^2 (\bar{y}_i - \hat{\mu})^2$$

$y_i = m_i \bar{y}_i$

$\hat{\bar{M}}^2 = \bar{m}^2$ , if  
 $\bar{M}$  is unknown

$$\hat{Var}(\hat{\tau}) = \hat{Var}(M\hat{\mu}) = \frac{M^2}{\bar{M}^2} \frac{N-n}{N} \frac{S_r^2}{n} = N^2 \frac{N-n}{N} \frac{S_r^2}{n}$$

# Cluster Sampling (II)

## Farms example, cluster sampling, ratio estimation (I)

**Population:** 2072 farms divided into 53 clusters of unequal size (by area),  
 $\bar{M} = 2072/53 = 39.094$ , the average cluster size.

**Goal:** Estimate the average number of cattle per farm.

**Variables:** Number of cattle on farm ( $y$ ).

**Parameters to be estimated:** Average number of cattle per farm ( $\mu_y$ ).

**Sampling design:** One stage cluster sample of 14 clusters.

**Method of estimation:** Ratio estimation.



One cluster



An example of clustering

# Cluster Sampling (II)

## Farms example, cluster sampling, ratio estimation (II)

**Sample:** Sample of 14 clusters out of 53

Sample cluster $i$	Number of farms $m_i$	Number of cattle $y_i$
1	32	351
2	83	906
3	18	316
4	30	287
5	55	914
6	24	284
7	66	598
8	48	359
9	64	784
10	30	393
11	40	489
12	70	516
13	48	793
14	25	401
Total	633	7,391

### Calculation and estimation:

$$\sum m_i = 633, \sum y_i = 7,391, \sum m_i^2 = 33,823,$$

$$\sum y_i^2 = 4,592,931, \sum m_i y_i = 380,249$$

$$N = 53, M = 2072$$

$$n = 14$$

$$\hat{\mu} = \bar{y} = \frac{7391}{633} = 11.676$$

$$S_r^2 = \frac{1}{n-1} \sum (y_i - m_i \bar{y})^2$$

$$= \frac{1}{14-1} [4592931 - 2 \times 11.676 \times 380249 + (11.676)^2 \times 33823] = 24954.716$$

$$\begin{aligned} \hat{\bar{M}} &= \bar{m} \\ &= \frac{633}{14} \\ &= 45.21 \end{aligned}$$

$$\hat{Var}(\hat{\mu}) = \frac{N-n}{N} \frac{1}{\bar{M}^2} \frac{S_r^2}{n} = \frac{53-14}{53}$$

$$\times \frac{1}{39.094^2} \times \frac{24954.716}{14} = 0.8582$$

$$\begin{aligned} \hat{\mu}_t &= \frac{7,391}{14} \\ &= 527.93 \\ &= \frac{7,391}{633} \frac{633}{14} \\ &= \bar{m} \hat{\mu}_y \end{aligned}$$

$$\hat{\sigma}(\hat{\mu}) = \sqrt{0.8582} = 0.9264$$

# Cluster Sampling (III)

## Inference: Unbiased estimation of mean and total (I)

### 2) Unbiased estimation

$N$  - # of clusters,  $M$  – population size,

$$\tau = \sum_{i=1}^N \tau_i = \sum_{i=1}^N y_i, \quad \mu_t = \bar{\tau} = \frac{1}{N} \sum_{i=1}^N \tau_i = \frac{1}{N} \sum_{i=1}^N y_i,$$

$$\mu_t = \frac{\tau}{N} = \frac{M}{N} \frac{\tau}{M} = \bar{M} \mu_y, \quad \mu_y = \frac{\tau}{M} = \frac{\mu_t}{\bar{M}}, \quad \tau = N \mu_t$$

$\hat{\mu}_t = \bar{y}_t = \frac{1}{n} \sum_{i=1}^n y_i$	$\hat{\tau} = N \hat{\mu}_t = N \bar{y}_t = \frac{N}{n} \sum_{i=1}^n y_i$
--	---

$\hat{\mu}_y = \frac{\hat{\tau}}{M} = \frac{N \hat{\mu}_t}{M} = \frac{N}{M} \bar{y}_t = \frac{\bar{y}_t}{\bar{M}} = \frac{N}{M} \frac{1}{n} \sum_{i=1}^n y_i$
---

All unbiased

- To estimate  $\tau$ , we don't need  $M$
- To estimate  $\mu$ , we need  $M$

# Cluster Sampling (III)

## Inference: Unbiased estimation of mean and total (II)

Following SRS results:

$$Var(\hat{\tau}) = N^2 Var(\hat{\mu}_t) = N^2 \frac{N-n}{N-1} \frac{\sigma_t^2}{n} \quad \sigma_t^2 = \frac{1}{N} \sum_{i=1}^N (y_i - \mu_t)^2$$

variance of  $y_1, y_2, \dots, y_N$

$$S_t^2 = \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y}_t)^2, \hat{\sigma}_t^2 = \frac{N-1}{N} S_t^2$$

$$\hat{Var}(\hat{\tau}) = N^2 \frac{N-n}{N} \frac{S_t^2}{n}$$

$$\hat{Var}(\hat{\mu}) = \hat{Var}\left(\frac{\hat{\tau}}{M}\right) = \left(\frac{N}{M}\right)^2 \frac{N-n}{N} \frac{S_t^2}{n} = \frac{1}{\bar{M}^2} \frac{N-n}{N} \frac{S_t^2}{n}$$

$$\hat{Var}(\hat{\mu}) \approx \frac{1}{\bar{m}^2} \frac{S_t^2}{n} \text{ if ...?}$$


$$\hat{\bar{M}}^2 = \bar{m}^2, \text{ if } \bar{M} \text{ is unknown}$$

# Cluster Sampling (III)

**Inference: Unbiased estimation of mean and total (III)**

**Summary of two types of estimation:**

Ratio	Unbiased
$\hat{\mu} = \hat{\mu}_r = \frac{\bar{y}_t}{\hat{M}} = \frac{\sum_{i=1}^n y_i}{\sum_{i=1}^n m_i}$	$\hat{\mu} = \frac{\bar{y}_t}{\bar{M}} = \frac{N}{M} \frac{1}{n} \sum_{i=1}^n y_i$
$\hat{Var}(\hat{\mu}_r) = \frac{1}{\bar{M}^2} \frac{N-n}{N} \frac{S_r^2}{n}$	$\hat{Var}(\hat{\mu}) = \frac{1}{\bar{M}^2} \frac{N-n}{N} \frac{S_t^2}{n}$

$$\hat{RE}\left(\frac{\hat{\mu}_r}{\hat{\mu}}\right) = \frac{\hat{Var}(\hat{\mu})}{\hat{Var}(\hat{\mu}_r)} = \frac{S_t^2}{S_r^2} = \frac{\sum_{i=1}^n (y_i - \bar{y}_t)^2}{\sum_{i=1}^n (y_i - \hat{\mu} m_i)^2} > 1 ?$$

In many cases  $y_i$  - total in  $i$ th cluster, is correlated with cluster size  $m_i$ , and then the ratio estimator is more efficient.



# Cluster Sampling (III)

## Farms example, cluster sampling, unbiased estimation (I)

**Population:** 2072 farms divided into 53 clusters of unequal size (by area),  
 $\bar{M} = 2072/53 = 39.094$ , the average cluster size

**Goal:** Estimate the average number of cattle per farm.

All same as in the previous case, except

**Method of estimation:** Unbiased estimation.



**Calculation and estimation:**

$$N = 53, M = 2072, n = 14, \sum y_i = 7391, \sum y_i^2 = 4,592,931$$

$$\bar{y}_t = \frac{7391}{14} = 527.929, S_t^2 = \frac{1}{n-1} \sum (y_i - \bar{y}_t)^2 = 53,154.687$$

$$\hat{\mu} = \frac{\bar{y}_t}{\bar{M}} = \frac{N}{M} \frac{1}{n} \sum_{i=1}^n y_i = \frac{527.929}{39.094} = \frac{53}{2072} \frac{1}{14} 7,391 = 13.504$$

$$\hat{Var}(\hat{\mu}) = \frac{N-n}{N} \frac{1}{\bar{M}^2} \frac{S_t^2}{n} = \frac{53-14}{53} \times \frac{1}{39.094^2} \times \frac{53,154.687}{14} = 1.828$$

**Unbiased estimation:**  $\hat{\mu} = 13.504, \hat{\sigma}(\hat{\mu}) = \sqrt{1.828} = 1.352$

**Ratio estimation:**  $\hat{\mu}_r = 11.676, \hat{\sigma}(\hat{\mu}_r) = 0.9264$

What is  
not used  
from the  
available  
data?

Why difference?

# Cluster Sampling (IV)

## Equal cluster sizes, Ch. 8.4 (I)

Assume equal cluster sizes. Often the case in practice.

$N$  - # of clusters,  $m_i = m, i = 1, 2, \dots, N, \bar{m} = m$

$M = N \times m$  – population size known,

Total sample size =  $n \times m$  – known

We have only  
unbiased estimation

$$\bar{\bar{y}}_c = \frac{1}{n \times m} \sum_{i=1}^n y_i = \frac{1}{n} \sum_{i=1}^n \frac{y_i}{m} = \frac{1}{n} \sum_{i=1}^n \bar{y}_i = \frac{1}{m} \times \frac{1}{n} \sum_{i=1}^n y_i = \frac{1}{m} \bar{y}_t$$

$$\hat{\mu} = \bar{\bar{y}}_c$$

$$\hat{\mu}_t = m \bar{\bar{y}}_c$$

All unbiased

$$\hat{Var}(\hat{\mu}) = \hat{Var}(\bar{\bar{y}}_c) = \frac{1}{m^2} \frac{N-n}{N} \frac{S_t^2}{n} = \frac{N-n}{N} \frac{S_{\bar{y}}^2}{n}$$

$$S_t^2 = \frac{1}{n-1} \sum_{i=1}^n (y_i - \hat{\mu}_t)^2 = \frac{1}{n-1} \sum_{i=1}^n (y_i - m \bar{\bar{y}}_c)^2 = \frac{m^2}{n-1} \sum_{i=1}^n (\bar{y}_i - \bar{\bar{y}}_c)^2 = m^2 S_{\bar{y}}^2$$

$S_{\bar{y}}^2$  – sample variance of  $\bar{y}_1, \bar{y}_2, \dots, \bar{y}_n$

# Cluster Sampling (IV)

## Equal cluster sizes, Farms example (I)



**Population:** 167,640 farms divided into 8,382 clusters of equal size, 20.

**Goal:** Estimate the average number of cattle per farm.

**Clusters:** 20 farms grouped by map (location).

**Variables:** Number of cattle on farm ( $y$ ).

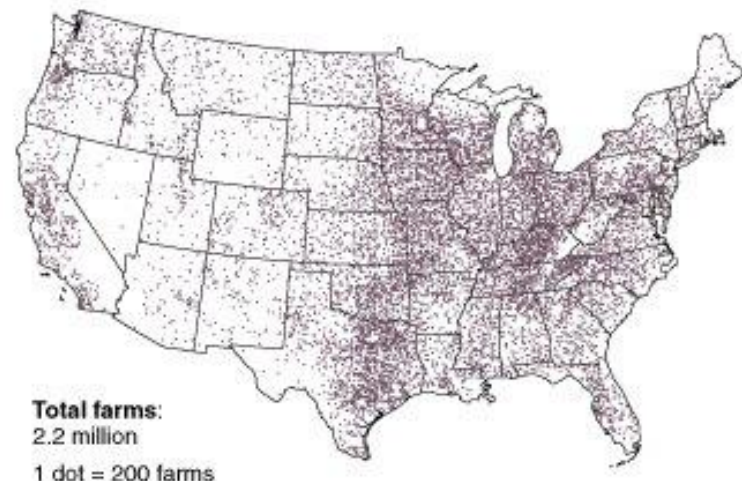
**Parameters to be estimated:** Average number of cattle per farm ( $\mu_y$ ).

**Sampling design:** One stage cluster sample of 87 clusters.

**Method of estimation:** Sample mean.



**Map:** Clusters of 20 farms each



# Cluster Sampling (IV)

## Equal cluster sizes, Farms example (II)



**Sample:** SRS of 87 clusters selected from 8382 clusters

Sample cluster $i$	Total # of cattle $y_i$	Average per farm $\bar{y}_i$
1	183	9.15
2	316	15.80
3	255	.
4	116	.
5	373	.
.	.	.
.	.	.
.	.	.
82	220	.
83	452	.
84	133	.
85	138	6.90
86	117	5.85
87	251	12.55
Total	19,784	989.20

### Calculation and estimation:

$$N = 8,382, M = 167,640, n = 87, \sum y_i = 19,784$$

$$\bar{y}_t = \frac{\sum y_i}{n} = \frac{19,784}{87} = 227.402, \quad m_i = m = 20$$

$$\hat{\mu} = \bar{y}_c = \frac{1}{m} \bar{y}_t = \frac{1}{nm} \sum_{i=1}^n y_i = \frac{19,784}{20 \times 87} = 11.370$$

$$S_t^2 = \frac{1}{n-1} \sum (y_i - \bar{y}_t)^2 = \frac{1}{87-1} \sum (y_i - 227.402)^2$$

$$= 17,348.894 = m^2 S_{\bar{y}}^2 = 20^2 \times 43.372$$

$$\hat{Var}(\hat{\mu}) = \frac{N-n}{N} \frac{1}{m^2} \frac{S_t^2}{n} = \frac{N-n}{N} \frac{S_{\bar{y}}^2}{n}$$

$$= \frac{8382-87}{8382} \frac{43.372}{87} = 0.4934$$

**Unbiased estimation:**  $\hat{\mu} = 11.370, \hat{\sigma}(\hat{\mu}) = \sqrt{0.4934} = 0.702$

# Cluster Sampling (IV)

## Equal cluster sizes and Systematic sampling

**Systematic sampling 1-in-k** – can be considered as a special case of cluster sampling with equal cluster sizes.

- One systematic sample = one cluster
- We have  $N = k$  clusters of size  $m$ , with  $n = 1$  cluster selected

**Repeated systematic sampling** – just as regular cluster sampling with  $n > 1$  clusters selected .

*k* clusters of size *m*

1	11	12	13	...	1 <i>m</i>
2	21	22	23	...	2 <i>m</i>
3	31	32	33	...	3 <i>m</i>
·	...				
<i>k</i>	<i>k</i> 1	<i>k</i> 2	<i>k</i> 3	...	<i>k</i> <i>m</i>

## Intraclass correlation coefficient

– a right moment to look at it again.

# Cluster Sampling (V)

**Equal cluster sizes: ANOVA type calculation, Ch. 8.4, Ch. 7.7 (I)**

$N$  clusters of equal size  $m$ ,  $M = N \times m$

Cluster\el.	1	2	...	$m$	$\mu_i$	$\sigma_i^2$
1	$y_{11}$	$y_{12}$		$y_{1m}$	$\mu_1$	$\sigma_1^2$
2	$y_{21}$	$y_{22}$		$y_{2m}$	$\mu_2$	$\sigma_2^2$
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$	$\vdots$
$N$	$y_{N1}$	$y_{N2}$		$y_{Nm}$	$\mu_N$	$\sigma_N^2$

$$\mu_i = \bar{y}_i = \frac{1}{m} \sum_{j=1}^m y_{ij}, \sigma_i^2 = \frac{1}{m} \sum_{j=1}^m (y_{ij} - \mu_i)^2, \mu = \frac{1}{M} \sum y_{ij} = \frac{1}{N} \sum_{j=1}^m \mu_i$$

**ANOVA identity:** 
$$\sum_{i=1}^N \sum_{j=1}^m (y_{ij} - \mu)^2 = m \sum_{i=1}^N (\mu_i - \mu)^2 + \sum_{i,j} (y_{ij} - \mu_i)^2$$

Sum of squares: 
$$SST_p = SSB_p + SSW_p$$

Mean squares:

$$MSB_p = \frac{SSB_p}{N-1}$$

between clusters

$$MSW_p = \frac{SSW_p}{N(m-1)}$$

within clusters

# Cluster Sampling (V)

## Equal cluster sizes: ANOVA type calculation (II)

$$MSB_p = \frac{N}{N-1} \frac{\sigma_t^2}{m} \Leftarrow \tau_i = m\mu_i, \sigma_t^2 = \frac{1}{N} \sum_{i=1}^N (\tau_i - \bar{\tau})^2 = m^2 \frac{1}{N} \sum_{i=1}^N (\mu_i - \mu)^2$$

$$MSW_p = \frac{1}{N(m-1)} \sum_{i,j} (y_{ij} - \mu_i)^2 = \frac{m}{m-1} \frac{1}{N} \sum_i \frac{1}{m} \sum_j (y_{ij} - \mu_i)^2 = \frac{m}{m-1} \frac{1}{N} \sum_i \sigma_i^2$$

$MSB_p$  - measures homogeneity between clusters

$MSW_p$  - measures homogeneity inside (within) clusters

We can estimate  $MSB_p$  and  $MSW_p$  from cluster sample!

# Cluster Sampling (V)

## Equal cluster sizes: ANOVA type calculation (III)

We can estimate  $MSB_p$  and  $MSW_p$  from cluster sample!

**Sample** :  $n$  clusters,  $i$ th cluster in the sample :  $y_{i1}, y_{i2}, \dots, y_{im}$ ,

$$\bar{y}_i (= \mu_i) = \frac{1}{m} \sum_{j=1}^m y_{ij}, y_i = \sum_{j=1}^m y_{ij} = m\bar{y}_i, \bar{y}_t = \frac{1}{n} \sum_{i=1}^n y_i$$

$$S_i^2 = \frac{1}{m-1} \sum_{j=1}^m (y_{ij} - \bar{y}_i)^2, \sigma_i^2 = \frac{m-1}{m} S_i^2, S_t^2 = \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y}_t)^2$$

$$MSB = \hat{MSB}_p = \frac{1}{m} S_t^2$$

$$MSW = \hat{MSW}_p = \frac{1}{n} \sum_{i=1}^n S_i^2$$

All unbiased

$$\hat{Var}(\hat{\mu}) = \hat{Var}(\bar{\bar{y}}_c) = \frac{1}{m^2} \frac{N-n}{N} \frac{S_t^2}{n} = \frac{N-n}{N} \frac{MSB}{n \times m}$$



# Cluster Sampling (V)

**Equal cluster sizes: Intraclass correlation coefficient (I)**

**Recall :**  $\rho = \rho_c = \text{Corr}(y', y'')$  - where  $y', y''$  are two different elements selected at random from a randomly selected cluster.

$$\rho = \rho_c = \text{Corr}(y', y'') = \frac{1}{\sigma^2} \frac{1}{N} \sum_{i=1}^N \frac{1}{m(m-1)} \sum_{j \neq l}^m (y_{ij} - \mu)(y_{il} - \mu)$$

Following derivation from Ch. 7, we have

$$\begin{aligned} \sigma_{\bar{y}}^2 &= \frac{1}{N} \sum_{i=1}^N (\bar{y}_i - \mu)^2 = \frac{1}{Nm} SSB_p = \frac{1}{m} \sigma^2 [1 + (m-1)\rho] \\ &= \frac{1}{m} \frac{SST_p}{Nm} [1 + (m-1)\rho] \Rightarrow \rho = \frac{mSSB_p - SST_p}{(m-1)SST_p} \end{aligned}$$

# Cluster Sampling (V)

## Equal cluster sizes: Intraclass correlation coefficient (II)

$$\begin{aligned}\rho &= \frac{(m-1)SSB_p - SSW_p}{(m-1)SST_p} = \frac{(m-1)(N-1)MSB_p - N(m-1)MSW_p}{(m-1)SST_p} \\ &= \frac{(N-1)MSB_p - N \times MSW_p}{SST_p} = \frac{(N-1)MSB_p - N \times MSW_p}{(N-1)MSB_p + N(m-1)MSW_p}\end{aligned}$$

$$N \text{ large} \Rightarrow \rho \approx \frac{MSB_p - MSW_p}{MSB_p + (m-1)MSW_p}$$

$$\hat{\rho} = \frac{\hat{MSB}_p - \hat{MSW}_p}{\hat{MSB}_p + (m-1)\hat{MSW}_p} = \frac{MSB - MSW}{MSB + (m-1)MSW}$$

$$\hat{\rho}: \begin{array}{ll} \leq 0 & \text{if } MSB \leq MSW \\ > 0 & \text{if } MSB > MSW \end{array}$$

Clusters

Outside similar, inside different

Outside different, inside similar

# Cluster Sampling (V)

## Equal cluster sizes: Comparison with SRS (I)

Cluster sample : total sample size  $n' = n \times m$ . Compare with an SRS of the same size  $n'$  from the same population of size  $N' = M = N \times m$ .

Use cluster sample results. (what if we used SRS?)

$$\bar{y}_{SRS} = \frac{1}{n'} \sum_{i=1}^{n'} y'_i, \text{Var}(\bar{y}_{SRS}) = \frac{N' - n'}{N' - 1} \frac{\sigma_y^2}{n'} = \frac{M - nm}{M - 1} \frac{\sigma_y^2}{nm}$$

Find an estimator for  $\sigma_y^2$  from cluster sample :

$$\sigma_y^2 = \frac{1}{M} SST_p = \frac{1}{Nm} (SSB_p + SSW_p) = \frac{1}{Nm} ((N - 1)MSB_p + N(m - 1)MSW_p)$$

$$\Rightarrow \hat{S}^2 = \hat{\sigma}_y^2 = \frac{MSB + (m - 1)MSW}{m}$$

$$\hat{\text{Var}}(\bar{y}_{SRS}) = \frac{M - nm}{M} \frac{\hat{S}^2}{nm}$$

# Cluster Sampling (V)

## Equal cluster sizes: Comparison with SRS (II)

$$\hat{RE}\left(\frac{\bar{\bar{y}}_c}{\bar{y}_{SRS}}\right) = \frac{\hat{Var}(\bar{y}_{SRS})}{\hat{Var}(\bar{\bar{y}}_c)} = \frac{\hat{S}^2}{MSB} = 1 + \frac{m-1}{m} \frac{MSW - MSB}{MSW}$$

$$\hat{RE}\left(\frac{\bar{\bar{y}}_c}{\bar{y}_{SRS}}\right) > 1 \text{ if } MSW > MSB \ (\hat{\rho} < 0)$$

Cluster sampling is more efficient

$$\hat{RE}\left(\frac{\bar{\bar{y}}_c}{\bar{y}_{SRS}}\right) \leq 1 \text{ if } MSW \leq MSB \ (\hat{\rho} \geq 0)$$

SRS is more efficient

# Cluster Sampling (V)



## Equal cluster sizes: Example on apartment houses (I)

**Population:**  $M = 1000$  apartments, divided into  $N = 100$  houses, with  $m = 10$  apartments in each house.

**Variable:** Number of persons living in an apartment ( $y$ ).

**Parameter of interest:** Average number of persons per apartment ( $\mu_y$ ).

**Sampling design:** One stage cluster sample of  $n = 10$  houses.

House	1	2	3	4	5	6	7	8	9	10	Summary
# people	29	41	35	34	28	38	36	30	37	42	$\bar{y}_t = 35.0$
$\bar{y}_i$	2.9	4.1	3.5	3.4	2.8	3.8	3.6	3.0	3.7	4.2	$\bar{\bar{y}}_c = 3.50$
$S_i^2$	0.77	0.75	0.84	0.92	0.81	0.80	0.90	0.90	0.85	0.88	$\sum S_i^2 = 8.42$

$$\bar{y}_t = \text{average per house} = \frac{1}{n} \sum y_i = 35.0, S_t^2 = \frac{1}{n-1} \sum_{i=1}^n (y_i - 35.0)^2 = 23.33$$

$$\bar{\bar{y}}_c = \frac{1}{nm} \sum y_i = \frac{1}{m} \bar{y}_t = 3.50, \quad \boxed{\hat{\mu} = \bar{\bar{y}}_c = 3.50} \quad \boxed{\text{CI: } \hat{\mu} \pm 2\hat{\sigma}(\hat{\mu}) = [3.21, 3.79]}$$

$$\hat{\text{Var}}(\hat{\mu}) = \frac{1}{m^2} \frac{N-n}{N} \frac{S_t^2}{n} = \frac{1}{10^2} \frac{100-10}{100} \frac{23.33}{10} = 0.021, \quad \hat{\sigma}(\hat{\mu}) = 0.145$$

# Cluster Sampling (V)

## Equal cluster sizes: Example on apartment houses (II)



Don't forget, the actual sample is

Appt House	1	2	3	4	5	6	7	8	9	10	$\bar{y}_i$	$s_i^2$
1	3	3	5	3	3	2	3	2	2	3	2.9	0.77
2	4	2	1	.	.				.	3	4.1	0.75
...									.	.	...	.
10	2	3	4	.	.				2	5	4.2	0.88

but we show summaries only. You may get it either way.

# Cluster Sampling (V)

## Equal cluster sizes: Example on apartment houses (III)



### ANOVA type analysis

$$SSW = \sum_1^n \sum_1^m (y_{ij} - \bar{y}_i)^2 = (m-1) \sum_1^n S_i^2 = 9 \times 8.42 = 75.78,$$

$$SSB = m \sum_1^n (\bar{y}_i - \bar{\bar{y}}_c)^2 = m(n-1) S_{\bar{y}}^2 = 10 \times 9 \times 0.2333 = 20.99,$$

$$MSW = \frac{SSW}{n(m-1)} = \frac{9 \times 8.42}{10 \times 9} = 0.842, MSB = \frac{SSB}{n-1} = \frac{20.99}{9} = 2.33,$$

$$\hat{\rho} = \frac{MSB - MSW}{(m-1)MSW + MSB} = \frac{2.33 - 0.842}{9 \times 0.842 + 2.33} = 0.150 > 0,$$

$$\hat{S}^2 = \frac{(m-1)MSW + MSB}{m} = \frac{9 \times 0.842 + 2.33}{10} = 0.991,$$

$$\hat{RE}(\bar{\bar{y}}_c / \bar{y}_{SRS}) = \frac{\hat{S}^2}{MSB} = \frac{0.991}{2.33} = 0.425 < 1.$$

Conclusion: SRS is more efficient than cluster sampling in this problem.  
(Is it more convenient? E.g., could be more expensive.)

# Cluster Sampling (VI)

**Equal cluster sizes: Selecting sample size (see Ch. 8.5)**

**Using ratio estimation:** From

$$\text{Var}(\hat{\mu}) \approx \frac{1}{M^2} \frac{N-n}{N-1} \frac{\sigma_r^2}{n} = \frac{1}{M^2} \frac{N-n}{N} \frac{\tilde{\sigma}_r^2}{n}, \tilde{\sigma}_r^2 = \frac{1}{N-1} \sum_{i=1}^N (y_i - \mu m_i)^2$$

$$n = \frac{N \tilde{\sigma}_r^2}{ND + \tilde{\sigma}_r^2} \approx \frac{\tilde{\sigma}_r^2}{D}, \quad \hat{\sigma}_r^2 = S_r^2 = \frac{1}{n-1} \sum_{i=1}^n (y_i - \hat{\mu} m_i)^2 = \frac{1}{n-1} \sum_{i=1}^n m_i^2 (\bar{y}_i - \hat{\mu})^2$$

**N large**

**Using unbiased estimation:** From

$$\text{Var}(\hat{\mu}) = \frac{1}{M^2} \frac{N-n}{N-1} \frac{\sigma_t^2}{n} = \frac{1}{M^2} \frac{N-n}{N} \frac{\tilde{\sigma}_t^2}{n}, \tilde{\sigma}_t^2 = \frac{1}{N-1} \sum_{i=1}^N (y_i - \mu)^2$$

$$n = \frac{N \tilde{\sigma}_t^2}{ND + \tilde{\sigma}_t^2} \approx \frac{\tilde{\sigma}_t^2}{D}, \quad \hat{\sigma}_t^2 = S_t^2 = \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y}_t)^2$$

**N large**

From presample

$$D = \left( \frac{B_\mu \bar{M}}{2} \right)^2 \text{ for estimating } \mu, \quad D = \left( \frac{B_\tau}{2N} \right)^2 \text{ for estimating } \tau$$



# Cluster Sampling (VII)

**Estimating proportion: See Ch. 8.6, 8.7**

$p$  = proportion of elements with certain property

$a_i$  = # of elements with the property in cluster  $i$ ,  $y_i = a_i$

$\tau_p$  = total # of elements with the property

**Using ratio estimation:**

$$\hat{p} = \bar{y} = \frac{\sum_{i=1}^n a_i}{\sum_{i=1}^n m_i}, \hat{\tau}_p = M\hat{p}$$

$$Var(\hat{p}) \approx \frac{1}{\overline{M}^2} \frac{N-n}{N-1} \frac{\sigma_p^2}{n} = \frac{1}{\overline{M}^2} \frac{N-n}{N} \frac{\tilde{\sigma}_p^2}{n},$$

$$\tilde{\sigma}_p^2 = \frac{1}{N-1} \sum_{i=1}^N (a_i - pm_i)^2$$

$$\hat{Var}(\hat{p}) = \frac{1}{\overline{M}^2} \frac{N-n}{N} \frac{S_p^2}{n}, \hat{\tilde{\sigma}}_p^2 = S_p^2 = \frac{1}{n-1} \sum_{i=1}^n (a_i - \hat{p}m_i)^2 = \frac{1}{n-1} \sum_{i=1}^n m_i^2 (\hat{p}_i - \hat{p})^2$$

$$n = \frac{N\tilde{\sigma}_p^2}{ND + \tilde{\sigma}_p^2} \approx \frac{\tilde{\sigma}_p^2}{D} \quad D = \left( \frac{B_p \overline{M}}{2} \right)^2 \text{ for estimating } p, \quad D = \left( \frac{B_\tau}{2N} \right)^2 \text{ for estimating } \tau$$

$N$  large

# PPS Sampling, Ch. 8.9 (I)

## Sampling with probabilities proportional to size (I)

**Back to basics:** Population  $\mathcal{E} = \{e_1, e_2, e_3, \dots, e_N\}$

Consider sampling with replacement (for simplicity):

In SRS, in one selection everybody is equal:

$$P(e_i) = \frac{1}{N}, \sum_i P(e_i) = N \times \frac{1}{N} = 1$$

$$\text{Total: } \tau = \sum_{i=1}^N y_i, \text{ sample: } \sum_{i=1}^n y_i, \hat{\tau} = \frac{N}{n} \sum_{i=1}^n y_i = \sum_{i=1}^n \frac{N}{n} y_i$$

Every element in the sample "represents"  $\frac{N}{n}$  elements from the population. E.g.,  $N = 100, n = 5$ , every element in the sample represents 20 elements in the population. But, what if elements are quite different in size? Could they be well "represented"?

# PPS Sampling (I)

## Sampling with probabilities proportional to size (II)

**Example:** Population of companies, a small number of large companies, and a large number of small companies (see any annual business report).

An SRS, not very large, would include only small companies, with large probability and underestimate parameters related to company size.

**Exercise:**  $N = 1000$ , 20 – big companies, 980 – small companies. Find probability for  $n = 5$  and  $n = 20$  that the sample will include only small companies.

Sampling with replacement :  $P(\text{all small} \mid n = 5) = \left(\frac{980}{1000}\right)^5 = 0.904$

$$P(\text{all small} \mid n = 20) = \left(\frac{980}{1000}\right)^{20} = 0.668,$$

What to do? Either stratify the population, if possible, or select elements with unequal probabilities.



# PPS Sampling (I)

## Sampling with probabilities proportional to size (III)

Let, in every selection,

$$P(\text{select } e_i) = \pi_i, 0 < \pi_i < 1, \sum_{i=1}^N \pi_i = 1$$

(someone should be selected, and everyone can be selected)

Probability of selection,  $\pi_i$ , should be known (chosen) in advance for every element in the population.

We can select a company with the probability proportional to its profit/number of employees/... last year. We can select a class from a list at U of T with probability proportional to its size (known from register).

Very convenient in cluster sampling!

# PPS Sampling (I)

## Sampling with probabilities proportional to size (IV)

















Our design, population


$e_i$	$e_1$	$e_2$	...	$e_N$
$\pi_i$	$\pi_1$	$\pi_1$	...	$\pi_N$
$y_i$	$y_1$	$y_2$	...	$y_N$

Our design, sample of size  $n$

$i$	1	2	...	$n$
$\pi_i$	$\pi_1$	$\pi_1$	...	$\pi_n$
$y_i$	$y_1$	$y_2$	...	$y_n$

Every sampling is an independent experiment with an outcome  $e_i$  and measurement  $y_i$ , with known probability  $\pi_i$ .

5					
4					
3					
2					
1					

**Example:** A class with 5 rows and 16 students. We want to select rows. In SRS,  $P(\text{row } i) = 1/5$ . If we select with probability proportional to # of students ( $y_i = \#$   )

$i$	1	2	3	4	5
$\pi_i$	5/16	3/16	2/16	2/16	4/16
$y_i$	3	1	1	1	2

# PPS Sampling (I)

## Example: How to select a sample using PPS

How to simulate an experiment with these outcomes (to select a sample) using TRN?

$i$	1	2	3	4	5
$\pi_i$	5/16	3/16	2/16	2/16	4/16

In this problem, we first need to look for outcomes with prob. 1/16, and then to adjust to our distribution. We can use 6 groups of two digits from TRN:  $6 \times 16 = 96 < 100$  for 1, 2, ..., 16, and then adjust to  $\pi_i$ .

$i$	1	2	3	4	5
TRN	01	31	49	61	73
	02	32	50	62	74
	.	.	.	.	.
	30	48	60	72	96
	30/96	18/96	12/96	12/96	24/96
$\pi_i$	5/16	3/16	2/16	2/16	4/16

(obviously, we ignore 97, 98, 99, 00)

E.g., for  $n = 3$ , and first two digits from groups of 5 in row 11 (p. 383), we get 28, 69, 88. Sample:

Row	1	4	5
$y_i$	3	1	2
$\pi_i$	5/16	2/16	4/16

Don't forget, repetitions are possible.

We report the values and selection probabilities

# PPS Sampling (II)

## Estimation of mean and total using PPS (I)

For convenience, define a variable

$$z = \frac{y}{N\pi}, \text{ with values } z_i = \frac{y_i}{N\pi_i}, i = 1, 2, \dots, N$$

PPS sample will produce sample  $z_1, z_2, \dots, z_n$  of i.i.d. random variables

We can use :  $\hat{\mu}_{PPS} = \bar{z} = \frac{1}{n} \sum_{i=1}^n z_i = \frac{1}{n} \sum_{i=1}^n \frac{y_i}{N\pi_i} = \frac{1}{N} \frac{1}{n} \sum_{i=1}^n \frac{y_i}{\pi_i}$

$$\hat{\tau}_{PPS} = N\hat{\mu}_{PPS} = N\bar{z} = \frac{1}{n} \sum_{i=1}^n \frac{y_i}{\pi_i}$$

**Example:** In SRS (with replacement)

$$\pi_i = \frac{1}{N}, N\pi_i = 1, z_i = \frac{y_i}{N\pi_i} = y_i, \bar{z} = \frac{1}{n} \sum_{i=1}^n y_i = \bar{y}, \hat{\mu} = \bar{y}, \hat{\tau} = N\bar{y}$$

# PPS Sampling (II)

## Estimation of mean and total using PPS (II)

$\hat{\mu}_{PPS}$  and  $\hat{\tau}_{PPS} = N\hat{\mu}_{PPS}$  are unbiased estimators

**Proof** is simple:  $z_1, z_2, \dots, z_n$  are i.i.d. random variables, and then

$$E(\hat{\mu}_{PPS}) = E(\bar{z}) = E\left(\frac{1}{n} \sum z_i\right) = E(z) = \sum z_i p(z_i) = \sum_{i=1}^N \frac{y_i}{N\pi_i} \pi_i = \frac{1}{N} \sum_{i=1}^N y_i = \mu_y$$

$$E(\hat{\tau}_{PPS}) = NE(\hat{\mu}_{PPS}) = N\mu_y = \tau_y$$

Variance follows immediately:

$$Var(\hat{\mu}_{PPS}) = Var(\bar{z}) = \frac{1}{n} \sigma_z^2,$$

$$\sigma_z^2 = Var(z) = \sum (z_i - \mu)^2 p(z_i) = \sum_{i=1}^N \left(\frac{y_i}{N\pi_i} - \mu\right)^2 \pi_i = \frac{1}{N^2} \sum_{i=1}^N \left(\frac{y_i}{\pi_i} - \tau\right)^2 \pi_i,$$

$$Var(\hat{\tau}_{PPS}) = N^2 Var(\hat{\mu}_{PPS}) = \frac{1}{n} \sum_{i=1}^N \left(\frac{y_i}{\pi_i} - \tau\right)^2 \pi_i.$$



# PPS Sampling (II)

## Estimation of mean and total using PPS (II)

$$\hat{Var}(\hat{\mu}_{PPS}) = \frac{1}{n} \hat{\sigma}_z^2 = \frac{1}{n} S_z^2, S_z^2 = \frac{1}{n-1} \sum_{i=1}^n (z_i - \bar{z})^2 = \frac{1}{n-1} \sum_{i=1}^n \left( \frac{y_i}{N\pi_i} - \hat{\mu}_{PPS} \right)^2$$

$$\hat{Var}(\hat{\tau}_{PPS}) = N^2 \frac{1}{n} \hat{\sigma}_z^2 = \frac{1}{n} \hat{\sigma}_{Nz}^2 = \frac{1}{n} S_{Nz}^2, S_{Nz}^2 = \frac{1}{n-1} \sum_{i=1}^n \left( \frac{y_i}{\pi_i} - \hat{\tau}_{PPS} \right)^2$$

From  example:

$$\begin{aligned} \hat{\tau}_{PPS} &= 5 \frac{1}{3} \sum_{i=1}^3 \frac{y_i}{5\pi_i} = \frac{1}{3} \left( \frac{y_1}{\pi_1} + \frac{y_2}{\pi_2} + \frac{y_3}{\pi_3} \right) = \frac{1}{3} \left( \frac{3}{(5/16)} + \frac{1}{(2/16)} + \frac{2}{(4/16)} \right) \\ &= \frac{128}{15} = 8.53 \quad (\tau = 8) \quad \hat{Var}(\hat{\tau}_{PPS}) = \frac{1}{3} S_{Nz}^2 = \frac{1}{3} \times \frac{1}{3-1} \sum_{i=1}^3 \left( \frac{y_i}{\pi_i} - \hat{\tau}_{PPS} \right)^2 \end{aligned}$$

Row	1	4	5
$y_i$	3	1	2
$\pi_i$	5/16	2/16	4/16

$$= \frac{1}{3} \times 0.85333 = 0.2844$$

$$\hat{Sd}(\hat{\tau}_{PPS}) = \sqrt{0.2844} = 0.5333$$

# PPS Sampling (III)

## Construction of selection probabilities $\pi$ (I)

How to construct  $\pi_i, i = 1, 2, \dots, N$  :

Let  $x$  be a variable with values  $x_i > 0, i = 1, 2, \dots, N$ , and  $\tau_x = \sum x_i$ .

We define  $\pi_i$  to be proportional to "size" of  $x_i$  :  $\pi_i = \frac{x_i}{\tau_x}$

$$0 < \pi_i < 1, \sum \pi_i = \sum \frac{x_i}{\tau_x} = \frac{\sum x_i}{\tau_x} = 1$$

We try to find  $x_i$  proportional to  $y_i$  so that  $\pi_i = \frac{x_i}{\tau_x} \approx \frac{y_i}{\tau_y}$ . Then

$$\text{Var}(\hat{\tau}_{PPS}) = \sum_{i=1}^N \left( \frac{y_i}{\pi_i} - \tau_y \right)^2 \pi_i = \sum_{i=1}^N \left( \frac{y_i}{x_i} \tau_x - \tau_y \right)^2 \pi_i \approx \sum_{i=1}^N \left( \frac{\tau_y}{\tau_x} \tau_x - \tau_y \right)^2 \pi_i = 0$$

Obviously, an "ideal" choice is  $\pi_i^* = \frac{y_i}{\tau_y}$ , when  $\text{Var}(\hat{\tau}_{PPS}) = 0$ ,

but this is not possible in practice.

# PPS Sampling (III)

## Construction of selection probabilities $\pi$ (II)

$$\hat{\mu}_{PPS} = \frac{1}{n} \sum_{i=1}^n \frac{y_i}{Nx_i} \tau_x = \mu_x \frac{1}{n} \sum_{i=1}^n \frac{y_i}{x_i}, \hat{\tau}_{PPS} = N\hat{\mu}_{PPS} = \tau_x \frac{1}{n} \sum_{i=1}^n \frac{y_i}{x_i}$$

$$\hat{Var}(\hat{\mu}_{PPS}) = \frac{1}{n} \hat{\sigma}_z^2 = \frac{1}{n} S_z^2, S_z^2 = \frac{1}{n-1} \sum_{i=1}^n \left( \frac{y_i}{x_i} \mu_x - \hat{\mu}_{PPS} \right)^2$$

$$\hat{Var}(\hat{\tau}_{PPS}) = N^2 \hat{Var}(\hat{\mu}_{PPS}) = \frac{1}{n} S_{Nz}^2, S_{Nz}^2 = \frac{1}{n-1} \sum_{i=1}^n \left( \frac{y_i}{x_i} \tau_x - \hat{\tau}_{PPS} \right)^2$$

**Recall:** We can select a company with the probability proportional to its profit/number of employees/... last year ( $x$ ), to estimate profit this year ( $y$ ). These two variables are likely proportional.

# PPS Sampling (IV)

## Application to one-stage cluster sampling (I)

Use cluster size as the variable  $x$ ,  $x_i = m_i, i = 1, 2, \dots, N$ ,

$\tau_x = \sum_{i=1}^N m_i = M$ , that is, make the probability of selecting a cluster *proportional to its size* :

$$\pi_i = \frac{m_i}{M}$$

$$\hat{\tau}_{PPS} = M \frac{1}{n} \sum_{i=1}^n \frac{y_i}{m_i} = \frac{M}{n} \sum_{i=1}^n \bar{y}_i = M \hat{\mu}_{y,PPS}, \hat{\mu}_{y,PPS} = \frac{\hat{\tau}_{PPS}}{M} = \frac{1}{n} \sum_{i=1}^n \bar{y}_i$$

**Warning** : If we use the formula  $\hat{\mu}_{PPS} = \frac{1}{n} \sum_{i=1}^n \frac{y_i}{Nx_i} \tau_x = \frac{1}{n} \sum_{i=1}^n \frac{y_i}{Nm_i} M$   
 $= \frac{M}{N} \frac{1}{n} \sum_{i=1}^n \frac{y_i}{m_i} = \bar{M} \frac{1}{n} \sum_{i=1}^n \bar{y}_i = \bar{M} \hat{\mu}_{y,PPS}$ , we would estimate the average  
per sampling unit - cluster ( $N = \#$  sampling unit - here clusters), not  
the population average!

# PPS Sampling (IV)

## Application to one-stage cluster sampling (II)

$$\hat{Var}(\hat{\tau}_{PPS}) = \frac{1}{n} S_{Nz}^2, \quad S_{Nz}^2 = \frac{1}{n-1} \sum_{i=1}^n \left( \frac{y_i}{m_i} M - \hat{\tau}_{PPS} \right)^2 = M^2 \frac{1}{n-1} \sum_{i=1}^n (\bar{y}_i - \hat{\mu}_{PPS})^2$$

$$\hat{Var}(\hat{\mu}_{PPS}) = \frac{1}{M^2} \hat{Var}(\hat{\tau}_{PPS}) = \frac{1}{n} S_{\bar{y}}^2, \quad S_{\bar{y}}^2 = \frac{1}{n-1} \sum_{i=1}^n (\bar{y}_i - \hat{\mu}_{PPS})^2$$

Summary of estimation in cluster sampling (for  $\tau$ )

1) Unbiased:  $\hat{\tau} = N \frac{1}{n} \sum_{i=1}^n y_i$

2) Ratio:  $\hat{\tau} = M \frac{\sum_{i=1}^n y_i}{\sum_{i=1}^n m_i}$

3) PPS:  $\hat{\tau} = \frac{M}{n} \sum_{i=1}^n \frac{y_i}{m_i}$

SRS of clusters

$y_i$  - total over cluster

PPS of clusters

Some discussion on comparison

# PPS Sampling (V)

## Farms example, PPS sampling of clusters (I)



**Population:** 2072 farms divided into 53 clusters of unequal size (by area),  $\bar{M} = 2072/53 = 39.094$ , the average cluster size.

**Goal:** Estimate the average number of cattle per farm.

**Variables:** Number of cattle on farm ( $y$ ).

**Parameters to be estimated:** Average number of cattle per farm ( $\mu_y$ ) and total number of cattle on farms.

**Sampling design:** One stage cluster sample (with replacements) of 14 clusters with probabilities of selection proportional to cluster size (number of farms).

**Method of estimation:** Unbiased PPS estimation.

**Sample:** PPS sample of 14 clusters selected from 53 clusters with probabilities proportional to number of farms (clusters in the sample ordered by number of farms). (see next slide)



# PPS Sampling (V)



## Farms example, PPS sampling of clusters (II)

Sample cluster, $i$	Number of farms, $m_i$	Number of cattle, $y_i$	Aver. cattle per cluster, $\bar{y}_i = y_i/m_i$	Selection Probability $\pi_i = m_i/M$
1	19	66	3.47	19/2072
2	28	326	11.64	28/2072
3	28	392	14.00	28/2072
4	29	350	12.07	.
5	31	331	10.68	.
6	31	331	10.68	.
7	46	697	15.15	.
8	51	586	11.49	.
9	53	739	13.94	.
10	55	914	16.62	.
11	61	619	10.15	.
12	64	784	12.25	.
13	83	906	10.92	83/2072
14	83	906	10.92	83/2072
Total	662	7,947	163.98	

### Estimation:

$$\hat{\mu}_{pps} = \frac{1}{n} \sum_1^n \bar{y}_i = \frac{163.98}{14} = 11.713$$

$$\hat{\tau}_{pps} = M\hat{\mu} = 2072 \times 11.713 = 24269.3$$

$$\hat{Var}(\hat{\mu}_{pps}) = \frac{1}{n} \left[ \frac{1}{(n-1)} \sum_1^n (\bar{y}_i - \hat{\mu}_{pps})^2 \right]$$

$$= \frac{1}{n} \hat{Var}(\bar{y}_i) = \frac{9.2565}{14} = 0.6612$$

$$\hat{\sigma}(\hat{\mu}) = \sqrt{0.6612} = 0.8131$$

$$\hat{\sigma}(\hat{\tau}_{pps}) = 2072 \times \hat{\sigma}(\hat{\mu}) = 1,684.8.$$

### Compare with ratio estimation:

$$\hat{\mu} = 11.676, \hat{\sigma}(\hat{\mu}) = 0.9264$$

**Another possibility:** PPS sampling with selection probabilities proportional to the cluster area (not number of farms). This may give slightly better results (why?).