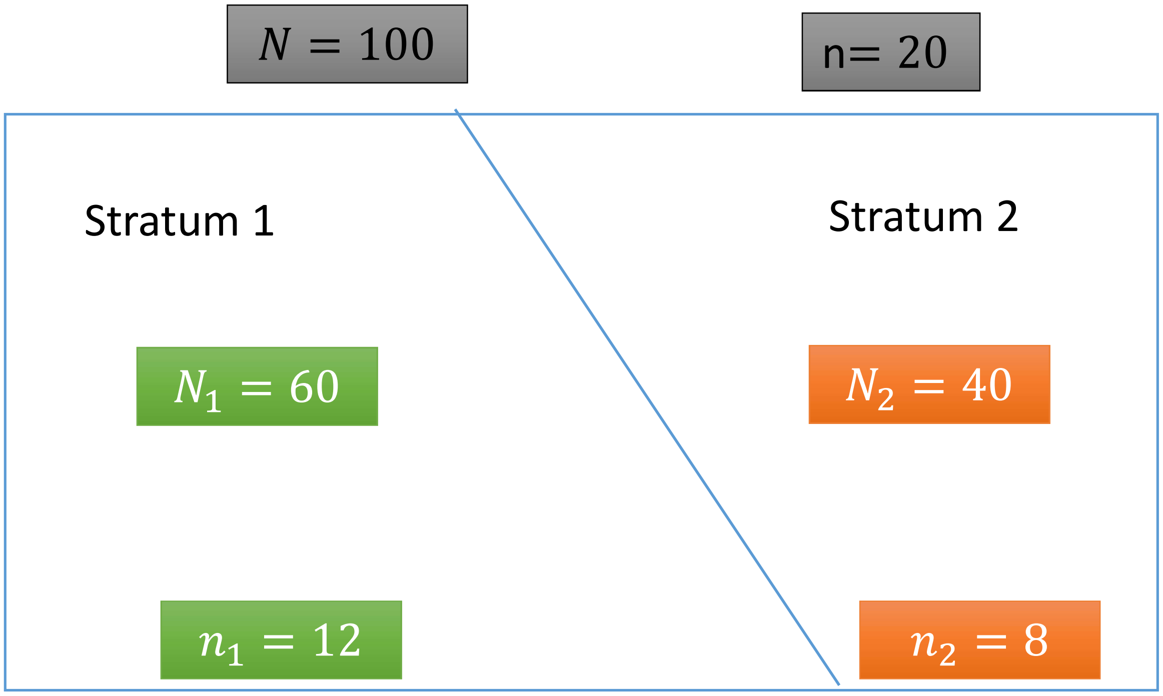
Cluster Sampling

Week	**Date**	**Plan**
08	06/27	Cluster Sampling
09	07/04	Complex designs
10	07/11	Ratio estimation
11	07/18	Categorical data analysis
12	07/25	Project Working Day
13	08/01	Project Presentations Final Review
14	08/08	Final Exam (Inclass portion)
15	08/15	**Final Exam (Take home part)**

Next Live session

- July 04- Independence Day
- July 07 – Friday (6.30p.m.)

Sampling Fraction $f_h = \frac{n_h}{N_h}$



$N = 100$

n= 20

Stratum 1

Stratum 2

$N_1 = 60$

$N_2 = 40$

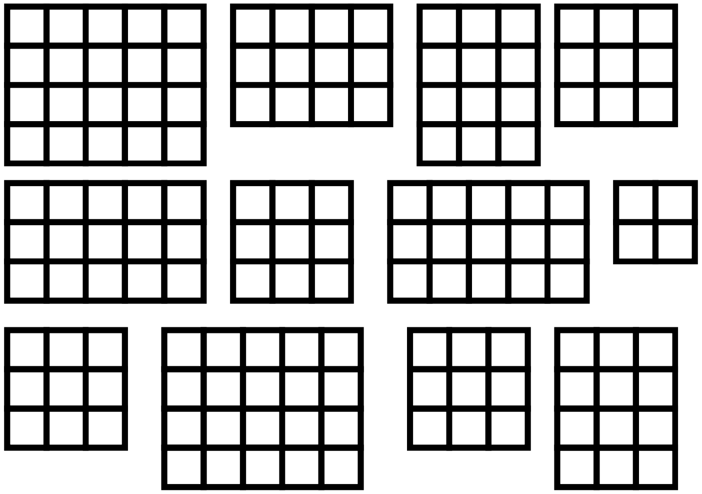
$n_1 = 12$

$n_2 = 8$

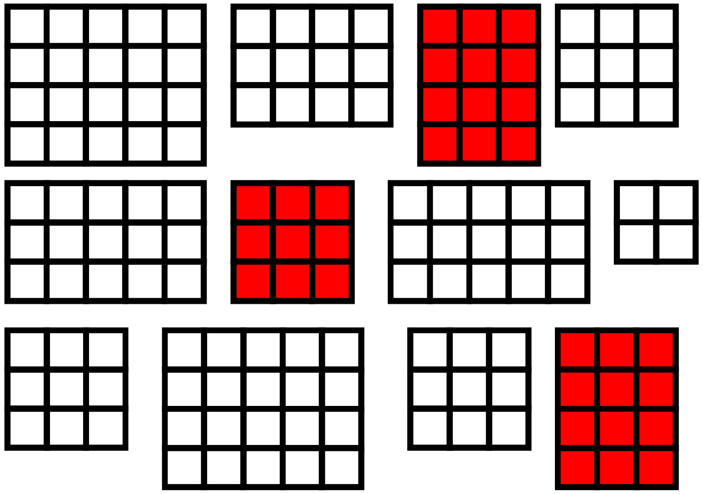
$$n_h = n\left(\frac{N_h}{N}\right)$$

$$\left(\frac{n_h}{N_h}\right) = \left(\frac{n}{N}\right)$$

Cluster Sampling (One-Stage Cluster Sampling)

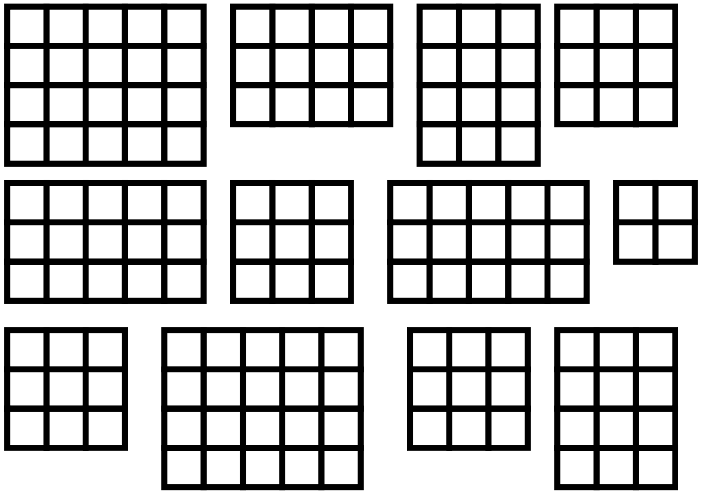


Population of *M* clusters
(M=12)

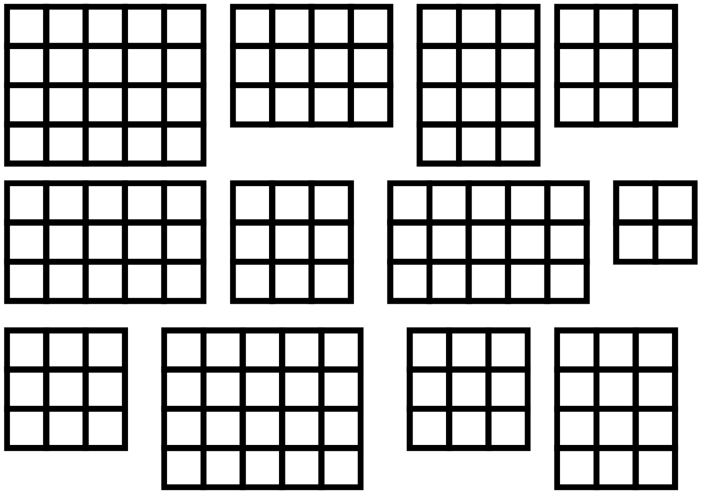


Take srs of *m* clusters, sample every unit in chosen clusters (m=3)

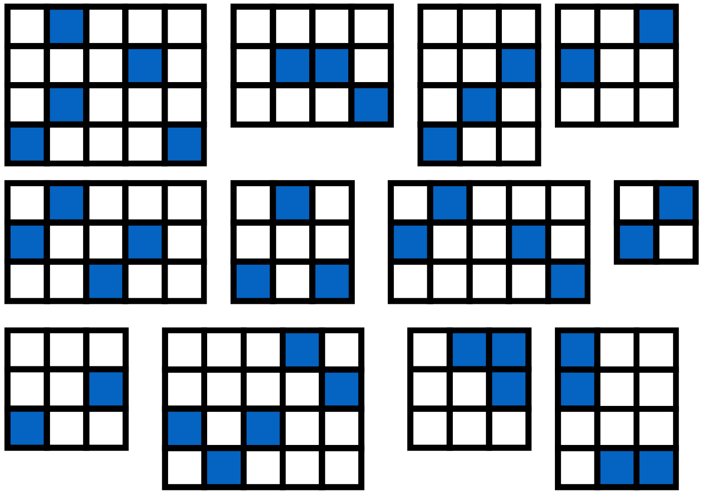
Difference Between Cluster and Stratified Sampling



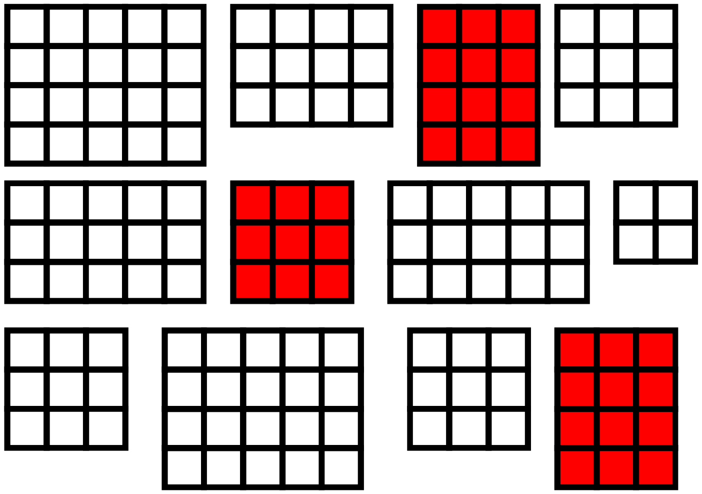
Population of H strata, stratum h contains n_h units



Population of M clusters



Take simple random sample in *every* stratum

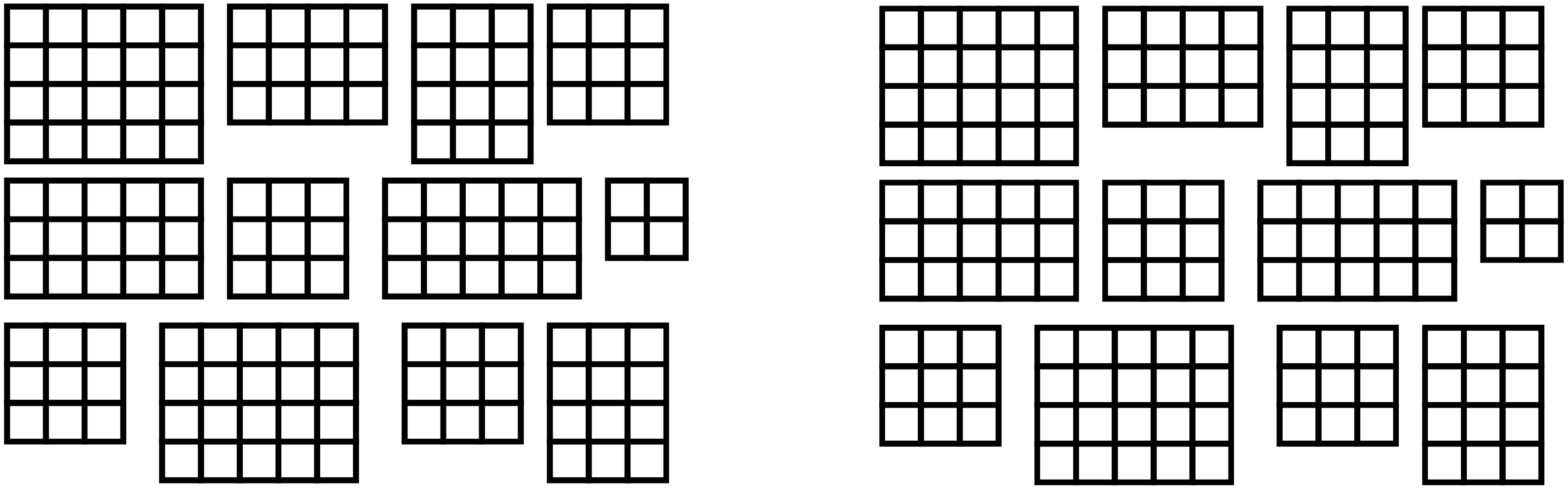


Take srs of m clusters, sample every unit in chosen clusters

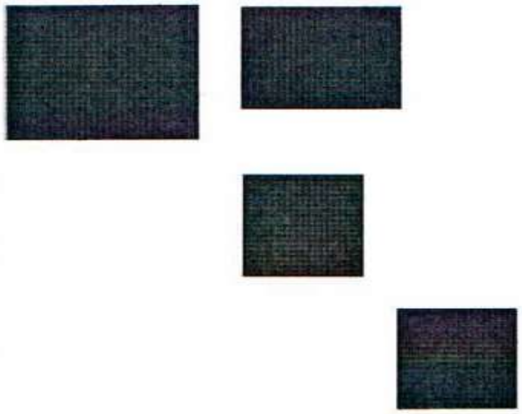
Two-stage designs

- An alternative design is one in which clusters are selected, and then individual units are subsampled from the clusters.
- Is a compromise between a simple random sample and a cluster sample.
- Cluster designs are a special case of two-stage designs.
- Designs may be generalized to multiple stages
- As long as units are selected using randomness at each stage, it is a probability sample
- The clusters are called primary sampling units (PSU's) and the individual units are called secondary sampling units (SSU's)

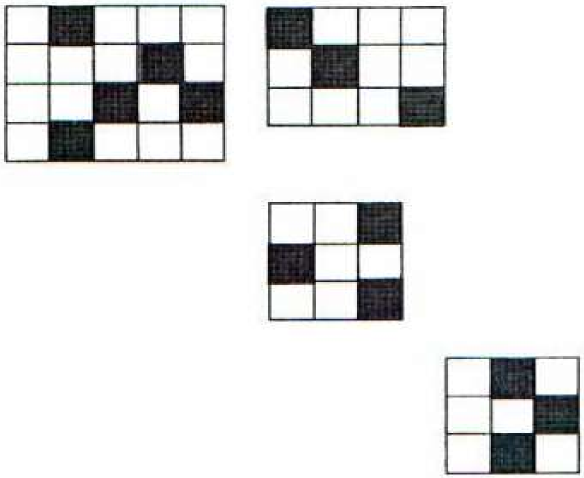
Cluster and Two-Stage Sampling



Population of M clusters



Take srs of m clusters, sample every unit in chosen clusters



Take srs of m clusters, sample n_i units in chosen clusters

Notations

- PSU : primary sampling units
- M= Total number of PSU’s in the population
- m= Number of PSU’s in the selected sample
- SSU : secondary sampling units
- N_i= Total number of SSU’s in the i^{th} PSU
- n_i= Number of SSU’s in the selected sample in the i^{th} PSU.

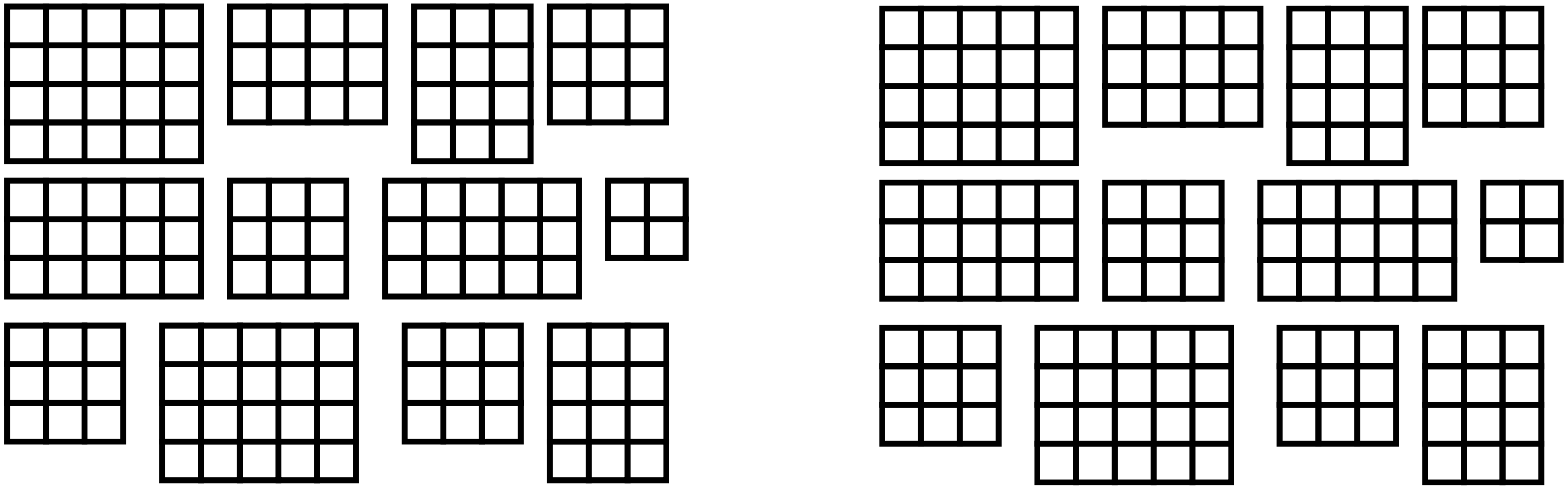
Estimation from two-stage designs

- Suppose there are N_i SSU's in the i^{th} PSU, and M PSU's in the population. If you sample m PSU's and n_i of the N_i SSU's, then the probability of selection is
$$\frac{m}{M}\frac{n_i}{N_i}$$

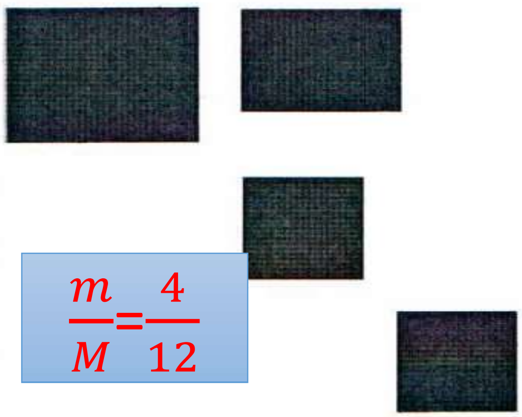
- Then the weight for the jth SSU in the ith PSU is
$$w_{ij} = \frac{M}{m}\frac{N_i}{n_i}$$

Cluster sampling (weight: M/m)

Cluster and Two-Stage Sampling

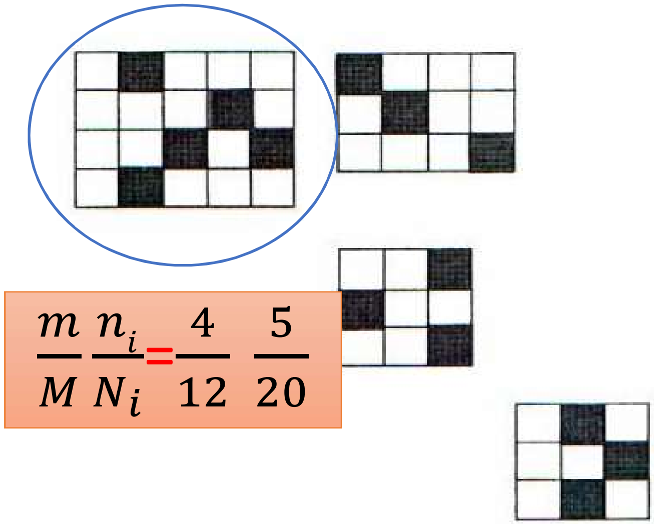


Population of *M* clusters



$$\frac{m}{M}=\frac{4}{12}$$

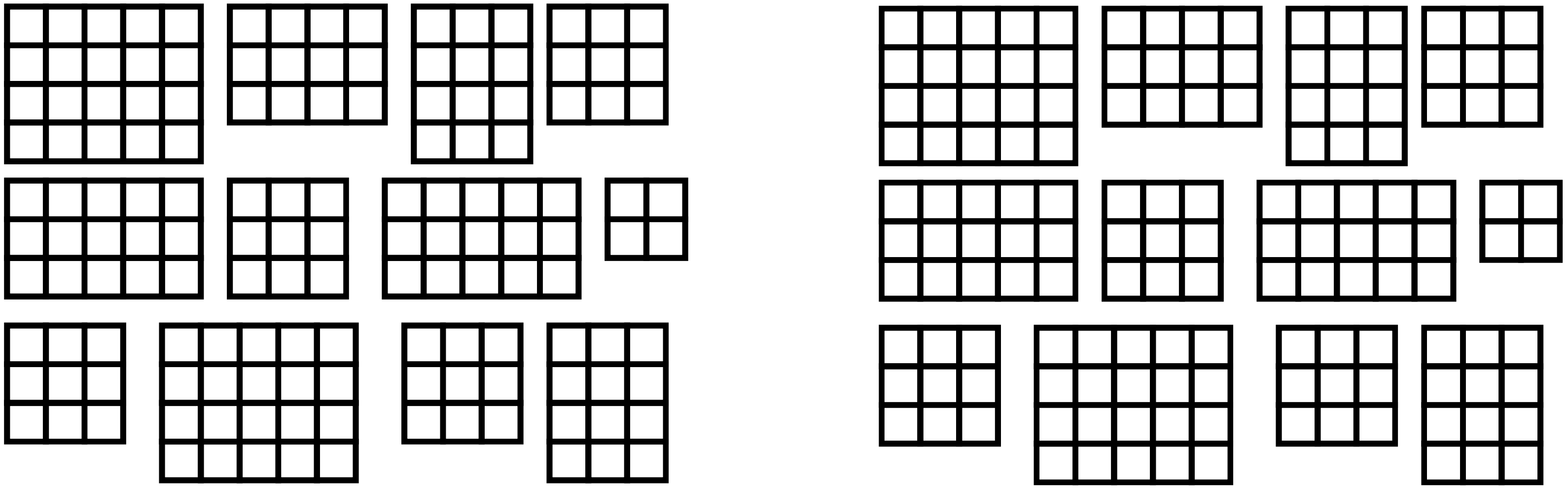
Take srs of *m* clusters, sample every unit in chosen clusters



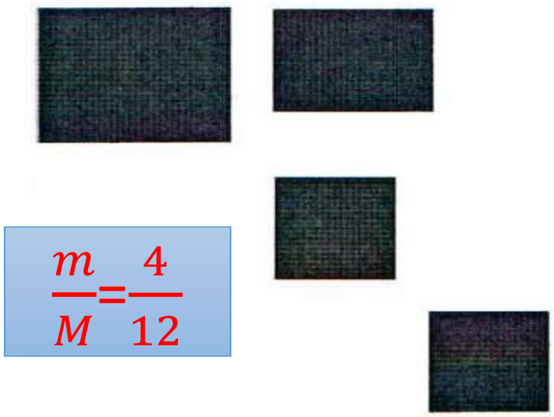
$$\frac{m}{M}\frac{n_i}{N_i}=\frac{4}{12}\frac{5}{20}$$

Take srs of *m* clusters, sample n_i units in chosen clusters

Cluster and Two-Stage Sampling

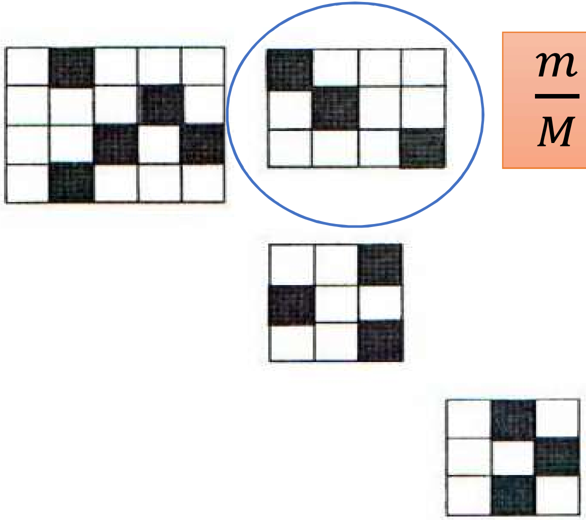


Population of M clusters



$$\frac{m}{M}=\frac{4}{12}$$

Take srs of m clusters, sample every unit in chosen clusters



$$\frac{m}{M}\frac{n_i}{N_i}=\frac{4}{12}\;\frac{3}{12}$$

Take srs of m clusters, sample n_i units in chosen clusters

Design effect: Reminder

- The design effect (deff) is a measure of the quality of a sample design; i.e., how does it compare to a SRS?
- It is defined as

$$deff_{complex} = \frac{V(\bar{y}_{complex})}{V(\bar{y}_{SRS})}$$

for same sample size.

Design effect for cluster designs

The design effect for a cluster design is related both to the size of the clusters (or the number of SSU's sampled from each PSU) and how similar are the units within the clusters.

If clusters were all the same size (*N*), the design effect for estimating the mean, total, or proportion would be

$$deff = 1 + \rho(N - 1),$$

where ρ is the intracluster correlation.

This formula is approximately correct if the clusters sizes vary and $\bar{N}$ is the average cluster size and replaces *N* in the expression above.

Design effect for two-stage designs

The design effect for a two-stage design depends on the # of SSU's sampled from each PSU rather than the number in each PSU, the design effect for estimating the mean, total, or proportion would be

$$deff = 1 + \rho(\bar{n} - 1),$$

where ρ is the intraclass correlation within the cluster.

Example 1

- A sample of adults in Dallas County will be selected to estimate the proportion having medical insurance. The sample design is a cluster design, where the clusters are households, and EVERY adult in the household will included in the sample.

- What is the deff for this design?
- Useful information:
  - Average # of adults/ household = 1.61
  - # of adults in Dallas County is 1.48 million
  - Intra-household correlation for insurance coverage is ~ 0.73

$$deff = 1 + \rho(\bar{n} - 1) = 1 + (0.73)(1.61 - 1) = 1.45$$

Example 2

- The percentage of adults with medical insurance in Dallas County in 2012 was 45%. We would like to estimate current coverage rate with a m.o.e. of 2% for a 95% confidence interval. Which is a more efficient design: a SRS of adults or a cluster sample of households?

- Useful information:
  - Average # of adults/ household = 1.61
  - # of adults in Dallas County is 1.48 million
  - Intra-household correlation for insurance coverage is ~ 0.73
  - Avg. cost for contacting/interviewing 1st subject in each hh = $25
  - Avg. cost for interviewing each additional subject in hh = $2

Steps in calculating sample size for complex design when deff is available:

Step 1. Calculate sample size for srs ***without fpc*** (called $n_{0,srs}$)

Step 2. Calculate $n_{0,complex}$ = $n_{0,srs} * deff$

Step 3. Correct for fpc by calculating

$$n_{complex} = n_{0,complex}/(1 + \frac{n_{0,complex}}{N})$$

Steps in calculating sample size for complex design when deff is available:

Step 1. Calculate sample size for srs ***without fpc***

$$n_{0,srs} = \frac{4*(0.50)(1-0.50)}{0.02^2} = 2500.$$

Step 2. Calculate $n_{0,complex} = n_{0,srs} * deff$

$$= 2500*[1+(0.61)(0.73)] = 3613 \text{ or } \frac{3613}{1.61} = 2244 \text{ hh's.}$$

Step 3. No correction for fpc needed because *M* >> m

- Cost of SRS design is $(\$25)2500 = \$62{,}500$
- Cost of Cluster design is $(\$25)(2244) + (\$2)(3613 - 2244) = \$46{,}200 + \$2738 = \$48{,}938$