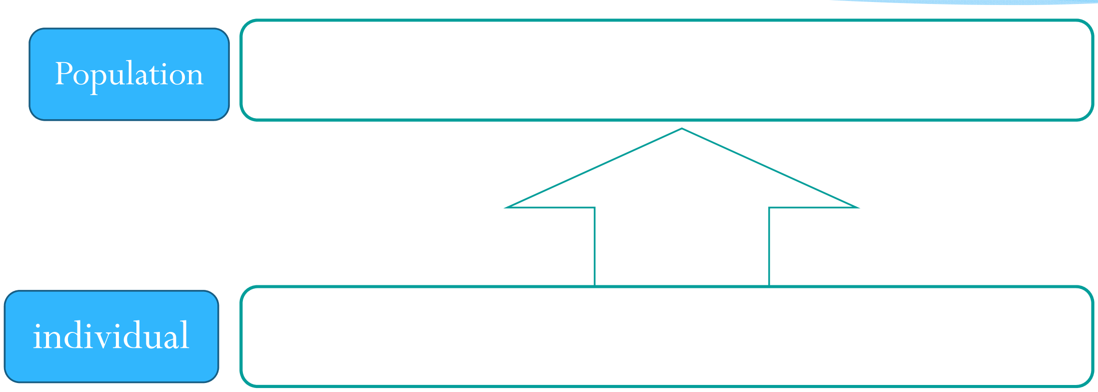
Second Regional Training Course on Sampling Methods for
Producing Core Data Items for Agricultural and Rural Statistics

Module 2: Review of Basics of Sampling Methods: Probability Sampling, Sample Selection and Sample Design and Estimation

Session 2.6: Estimation under different designs

9 – 20 November 2015,
Jakarta, Indonesia

Associate



distribution

statistic

parameter

variable

observation

unknown

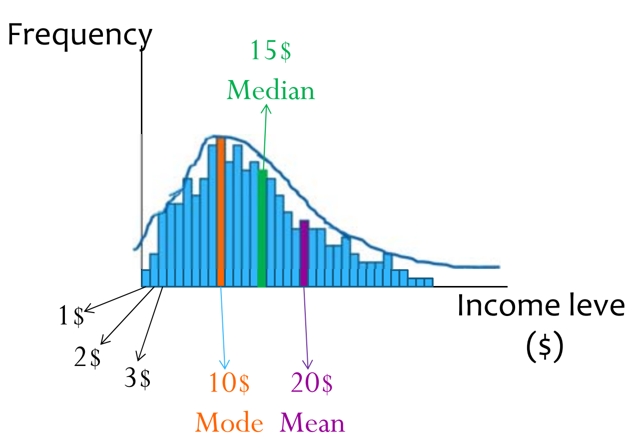
Example

parameter	Mean income in a country
variable	Income
observation	Observed income of each individual
statistic	Observed average Mean income

Statistic is an estimate for unknown parameter

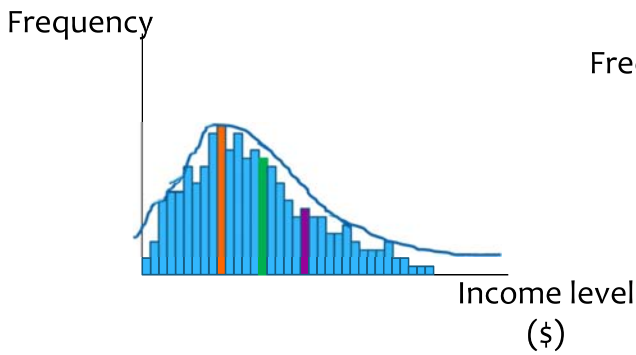
We select from a distribution

Every variable has a distribution

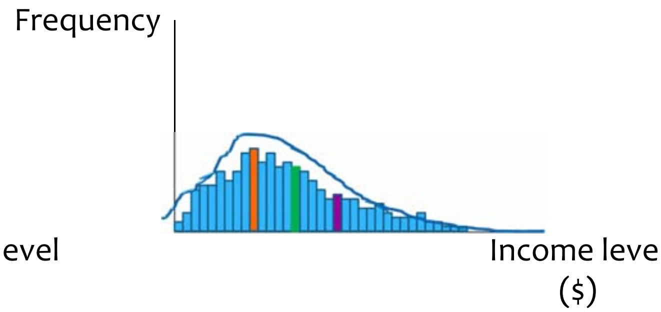


We select from a distribution

Population



Sample



We ideally want to have representatives from all income levels in our sample

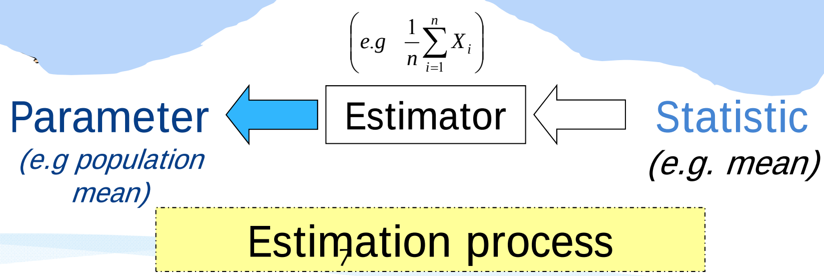
Estimation Procedures

Problem

- Need a scientific 'guess' about the value of an unknown parameter

Solution

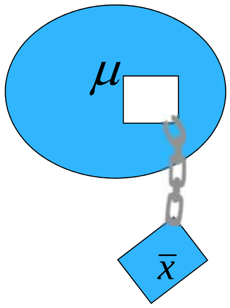
- Based on data from a random sample of elements selected from the population



Two types of estimation

- Point Estimation
  - Compute a single number
- Interval Estimation
  - Compute an *interval of numbers* that has a specified *confidence level* of including the unknown value of the parameter

Theory of point estimation



Real unemployment rate= μ (Unobservable)
Unemployment rate in the sample= $\bar{x}$ (observed from sample)

$\bar{x}$ is ***only one*** point estimation for μ

Because there are many possible samples and so possible estimations

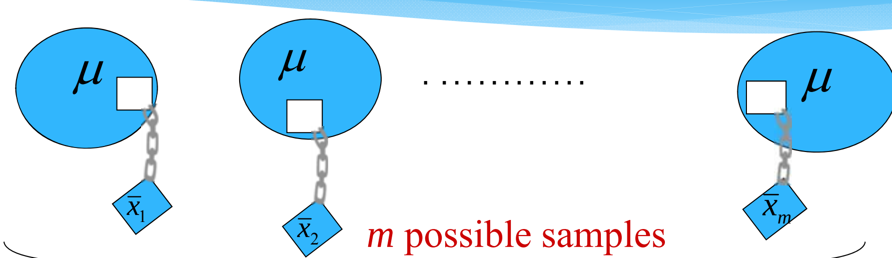
$E(\bar{x}) = \mu$ ⟶ E: expected value

Average over all possible estimations is equal to real value

⇨ $\bar{x}$ is an unbiased estimation for μ

Sampling distribution

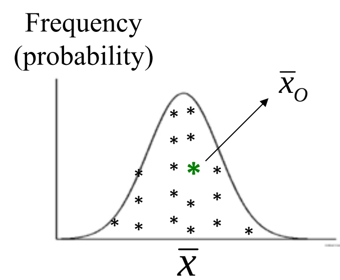
Using same sampling method in each selection



Observed average income from selected sample

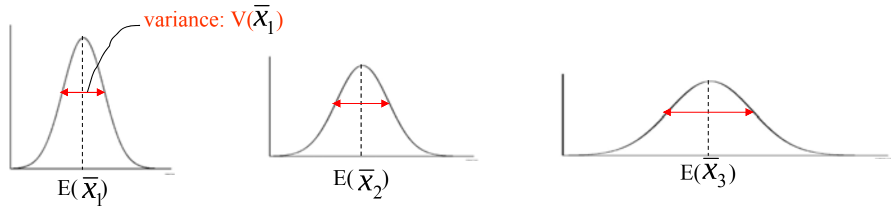
$$\bar{x}_1 \neq \bar{x}_2 \neq \ldots \neq \bar{x}_O \neq \ldots \neq \bar{x}_m$$

$$E(\bar{X}) = \frac{\bar{x}_1 + \bar{x}_2 + \ldots + \bar{x}_m}{m} = \mu$$



Sampling variability

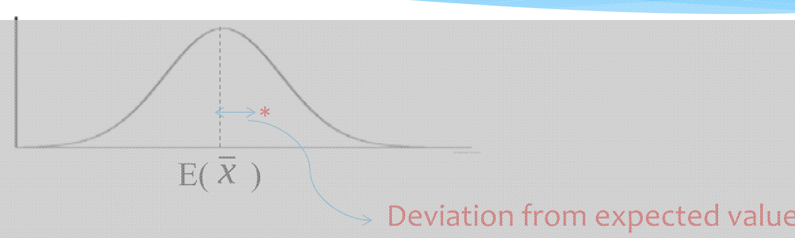
Using three different sampling methods



$E(\bar{x}_1) = E(\bar{x}_2) = E(\bar{x}_3)$: $\bar{x}$ is unbiased estimator of μ

$V(\bar{x}_1) < V(\bar{x}_2) < V(\bar{x}_3)$: different samples, different variations/precisions

Variance: A Measure of Variation in a sampling



- ***Variance:*** average square deviations from expected value

Number of all possible samples

$$\sigma_{\bar{x}}^2 = \frac{\sum_{i=1}^{m} \left(\bar{x}_i - E(\bar{x})\right)^2}{m}$$

Variance of $\bar{x}$

Population size

$$\sigma_x^2 = \frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}$$

Variance of x

Point Estimators

- Sample Mean $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
- Sample Variance $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$
- Sample Proportion $p = \frac{1}{n}\sum_{i=1}^{n} I_i$

$$I_i = \begin{cases} 1 & ith \text{ case has desired characteristic} \\ 0 & others \end{cases}$$

variance of mean

- Estimate for σ_x^2 is $S_x{}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$

Sample size

- Estimate for $\sigma_{\bar{x}}^2$ is $S_{\bar{x}}{}^2 = (1 - \frac{n}{N}) \times \frac{S_x{}^2}{n}$

$$where\ 1 - \frac{n}{N} = 1 - f$$

$$is\ called\ \ finite\ \ population\ correction\ (fpc)$$

Standard error (SE) and standard deviation (Std)

- Standard deviation for x $= S_x = \sqrt{S_x^2} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$
- Standard error for $\bar{x}$ $= S_{\bar{x}} = \sqrt{S_{\bar{x}}^2} \cong \frac{S_x}{\sqrt{n}}$ *(fpc is ignored for large populations)*
- SE of sample proportion p is $SE(p_s) = \sqrt{\frac{p_s(1-p_s)}{n}}$

$$\text{Precision of estimator} \propto \frac{1}{\text{variance}}$$

Sampling Error

- The error in a estimate that owes to the selection of only a subset (sample) of the total population rather than the entire population.
- Sampling error represents the difference between the estimate and its expected value.
- All sample estimates are subject to sampling error.
- The most commonly used measure of sampling error is *sampling variance*

Estimation Under SRS

* When is Total?

- *Parameter* (Y): *Total household expenditure*
- *Estimator:*

$$\hat{Y} = N \times \bar{y}$$

$$\hat{Y} = N \times \bar{y} = \sum_{i=1}^{n} \frac{N}{n} y_i = \sum_{i=1}^{n} w \, y_i$$

ID	Y	selected
1	10	
2	5	×
3	5	
4	10	×
Total	30	

w = base (sampling) weight for each selected unit or *inflation factor*

Estimation Under SRS

* What happened to weight in estimator for the mean?

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

$$\bar{y} = \sum_{i=1}^{n} \frac{(N/n) y_i}{(N/n) \times n} = \sum_{i=1}^{n} \frac{w y_i}{w \times n} = \frac{w}{w} \times \frac{\sum_{i=1}^{n} y_i}{n}$$

Base (sampling) weight: basic concept

Base weight…

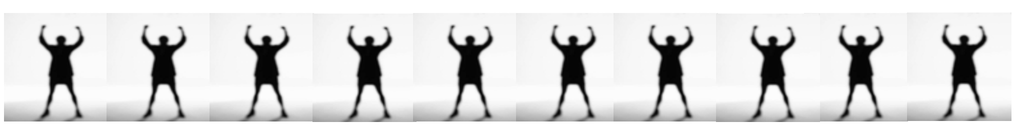
* Is the inverse of the probability of selection
  * Thus, depends on the sampling design and selection method

* Number of units in the population being represented by the sample unit
  * In ideal conditions, the design weights take care of "*representativeness*"
  * But, this is not true in *less than ideal conditions*

Sampling weight: basic concept

In a SRS design: N=10 and n=5

Population:



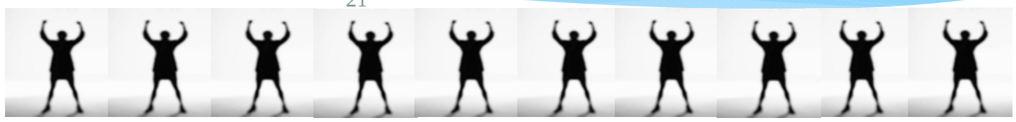
Inclusion probability/probability of selection (chance to be selected in the sample)=

$$\pi = \frac{n}{N} = \frac{5}{10} = \frac{1}{2}$$

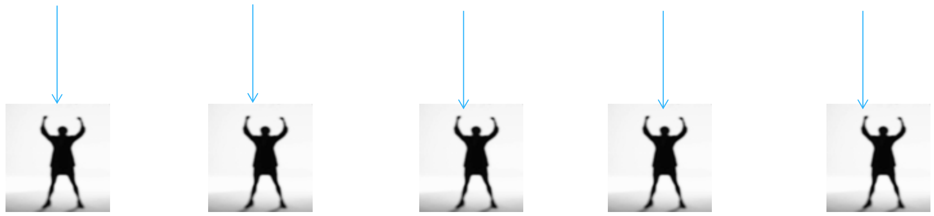
Each individual has 50% chance to be selected in the sample

Sampling weight: basic concept

Population:



SRS sample:



Sampling weight= inverse of inclusion probability:

$$w = \frac{1}{\pi} = \frac{1}{1/2} = 2 \quad OR \quad w = \frac{1}{\left(n/N\right)} = \frac{N}{n} = \frac{10}{5} = 2$$

Base (sampling) weight: more!

* Self-representative sampling units are those with inclusion probability of 1
* In **self-weighting** samples, each sampled unit has the **same design weight**
  * SRS is self weighting design
  * Computation of estimates is further simplified since the weighting factor is a constant number (*w*)

Estimation under systematic sampling

Parameter (Y): total expenditure

Estimator for the total?

Circular SYS and linear when N is divisible by n:

$$\hat{Y} = \sum_{i=1}^{n} kY_i \qquad (w = k = \frac{N}{n})$$

SYS when N is NOT divisible by n:

$$\hat{Y} = \sum_{i=1}^{n} kY_i \qquad (w = k = nearest \quad int\ iger \quad to \frac{N}{n})$$

Estimation under stratified sampling

- Population is divided into *H* strata: *e.g.* 6 regions
- n_h units are selected SRSWOR in stratum *h*(h=1,2..,H)
- Parameter of interest is average of Y $(\overline{Y})$

- Estimator for the mean: $\overline{y}_{st} = \sum_{h=1}^{H} \frac{N_h}{N} \overline{y}_h = \sum_{h=1}^{H} W_h \overline{y}_h$

$$OR \qquad \overline{y}_{st} = \frac{\sum_{h=1}^{H} N_h \overline{y}_h}{\sum_h N_h} = \frac{\sum_{h=1}^{H} \sum_i w_h y_{hi}}{\sum_h \sum_i w_h}$$

$$(W_h = Stratum \quad weight, \qquad w_h = \frac{N_h}{n_h} : sampling \quad weight \quad in \quad stratum \quad h)$$

Estimation under stratified sampling

- Variance of the estimate $\bar{y}_{st}$:

$$V(\bar{y}_{st}) = \sum (1 - f_h) \frac{W_h^2 s_h^2}{n_h} \qquad \left(f_h = \frac{n_h}{N_h},\ W_h = \frac{N_h}{N} \right)$$

- Where $s_h^2 = \frac{1}{(n_h - 1)} \sum_{i=1}^{n_h} (y_{hi} - \bar{y}_h)^2$

- If sampling fractions f_h are negligible:

$$V(\bar{y}_{st}) = \sum \frac{W_h^2 s_h^2}{n_h}$$

Estimation under PPS design

- Inclusion probability is related to an auxiliary variable, Z, that is a measure of "size".
- Selection probability for i^{th} unit is $p_i = Z_i \Big/ \sum_{i=1}^{N} Z_i$
- Sampling weights: $w_i = \frac{1}{np_i} = \frac{\sum_{i=1}^{N} Z_i}{n \times Z_i}$

* If we draw a sample of n units with replacement out of N units, with the initial probability of selection of the i^{th} unit as p_i ,the combined unbiased estimator of Y is

$$\hat{Y}_{pps} = \frac{1}{n}\sum_{i=1}^{n}\frac{y_i}{p_i} = \sum_{i=1}^{n}\frac{y_i}{np_i} = \sum_{i} w_i y_i \qquad (p_i = \frac{Z_i}{\sum_i Z_i})$$

$$w_i = \frac{1}{np_i} = \frac{\sum_i Z_i}{nZ_i} \quad (i = 1,2,\dots. n)$$

Sampling Weights under multi-stage design

Inclusion probability/probability of selection

SRS

$$\pi = \frac{n}{N}$$

PPS

$$\pi_i = \frac{nZ_i}{\sum_i Z_i}$$

N=population
n= sample
Z_i=size of unit i

Sampling weight:

$$w = \frac{1}{\pi} = \frac{N}{n}$$

$$w_i = \frac{1}{\pi_i} = \frac{\sum_i Z_i}{nZ_i}$$

Two stage sampling (sub-sampling)

Number of clusters: N

Size of i^{th} cluster: M_i

Sample size in the 1st stage: n ⇒ $\pi_1 = \frac{n}{N}$

Sample size in the 2nd stage in cluster i: m_i ⇒ $\pi_{2i} = \frac{m_i}{M_i}$

⇒ $\pi_i = \pi_1 \times \pi_{2i} = \frac{n}{N} \times \frac{m_i}{M_i}$

Fixed sampling rate

- Fixed sampling rate (k) in the 2nd stage for all clusters
- Probability of selection for all clusters:

$\pi_2 = \pi_1 \times k$ (epsem)

Fixed sample size

- $m_i = m$ is same in all selected clusters
- Probability of selection in cluster i, with size M_i:

$\pi_{2i} = \pi_1 \times \frac{m}{M_i}$ (not epsem)

Two-stage (cluster) sampling

- We select n clusters out of total N clusters (PSUs) in the first stage and a sample of m_i (SSUs) from i^{th} selected cluster with size M_i in the second stage with SRSWOR used in both stages.
- Estimator for total

$$\hat{Y} = \sum_{i=1}^{n} \sum_{j=1}^{m_i} w_i y_{ij} \qquad \left(w_i = \frac{NM_i}{nm_i} \right)$$

Two-stage (cluster) sampling

- Estimator for mean (clusters with unequal size)

$$\hat{\bar{Y}} = \frac{\sum_{i=1}^{n} \sum_{j=1}^{m_i} w_i y_{ij}}{\sum_{i=1}^{n} \sum_{j=1}^{m_i} w_i}$$

Two stage sampling (sub-sampling)

Number of clusters: N

Size of i^{th} cluster: M_i

1^{st} stage: PPS of size n (measure of size M_i) ⇨ $\pi_{1i} = \frac{nM_i}{\sum_i M_i}$

Sample size in the 2^{nd} stage in each cluster: m ⇨ $\pi_{2i} = \frac{m}{M_i}$

$$\pi = \pi_{1i} \times \pi_{2i} = \frac{nM_i}{\sum_i M_i} \times \frac{m}{M_i} = \frac{n \times m}{\sum_i M_i}$$

Self weight

Two-stage (cluster) sampling

- Estimate for Total and Mean

$$\hat{Y} = \sum_i \sum_j w_i y_{ij} \qquad \bar{y} = \sum_i \sum_j w_i y_{ij} \Big/ \sum_i \sum_j w_i$$

$$Where \quad w_i = \frac{\sum_i Z_i}{nZ_i} \times \frac{M_i}{m_i} \qquad (if\ m_i\ fixed\ and\ M = Z)$$

$$then \quad self - weight \quad design \quad : w = \frac{\sum_i M_i}{nm}$$

Exercise 3

- In a two-stage sampling, village (in rural) or city block (in urban) is PSU and household is SSU
- There are 1000 PSUs distributed among 3 regions (strata) as follows

strata	# of PSUs	Total # of HHs
1	200	1500
2	400	3000
3	400	2500

- We allocate a sample size of 40 PSUs proportionally to three strata, and select a PPS sample in the first stage with number of households in each PSU as size variable and SRS sample of 5 HHs from each selected PSU in the 2nd stage.
- Calculate sampling weight in each stratum and PSU?

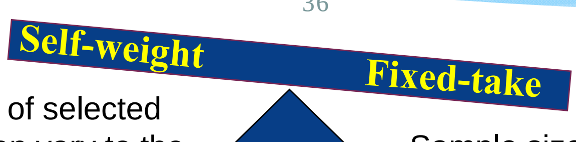
Problem

35

* In a 2-stage design, PSUs are to be selected with PPS (size being the *estimated number* of SSUs in each PSU).
* A random sample of a *fixed size of SSUs* to be selected within each sampled PSU.
* Wish to have a *fixed total sample size* of SSUs.
* *It was found that the actual number of SSUs within each PSU was different from that used in selecting the PPS sample of PSUs.*
* *How to keep the design approximately self-weighting without changing too much the sample size per PSU and the total sample size of SSUs?*

Solution?

36



Number of selected SSUs can vary to the extent measure of size differs from actual size

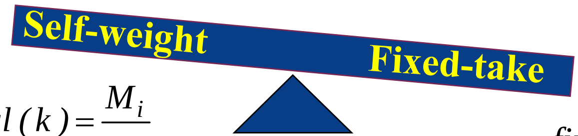
Sample size fixed, but probabilities vary to the extent measure of size differs from actual size

* In general, recommended to maintain self-weight
* Except in heavy surveys that extra workload is unacceptable

Problems with "*fixed-take*"

37

1. Arbitrary variations require further weighting on sample data (*undesirable when changes are large*)



$$fixed\ \ interval\,(k) = \frac{M_i}{m_i}$$

$$w_i = \frac{\sum_i \hat{M}_i}{n\hat{M}_i} \times k$$

$$fixed\ \ size\ m$$

$$w_i = \frac{\sum_i \hat{M}_i}{n\hat{M}_i} \times \frac{M_i}{m}$$

Problems with "*fixed-take*"

38

2. Enumerators prefer fixed interval to fixed size
3. Need to keep record of actual PSU size to calculate base weights
4. Hiding under-coverage and selection problems (preference of enumerators to select small segments, poor listing, etc)
5. Encourages incomplete listing
6. Uncontrolled substitution for non-responding cases
7. In any case, fixed size is not attainable for non-response and other operational problems

Sampling Error

Determining factors

- Sampling error (variance) is affected by a number of factors:
  - variability within the population.
  - sample size - sampling rate
  - sample design

- If sampling principles are applied carefully
  - within the constraints of available resources

  sampling error can be accurately measured and kept to a minimum.

Sampling Error

Sampling Variance and Population Variability

"The sample design and sample size remaining unchanged, the **higher** the **population variance** (*variation in the study variable in the population*) the **higher** is the **sample variance**."

Sampling Error

Sampling Variance and Sample size

"The sample design and population (variance) remaining unchanged, the **higher** the **sample size** the **lower** is the **sample variance**."

Sampling Error

Sampling Variance and Sample Design

"For a given population and sample size,

the **sample variance** depends on the sample design adopted."

The relative precision of a sample design compared to SRSWOR – is measured by ***Design effect*** (*Deff*)

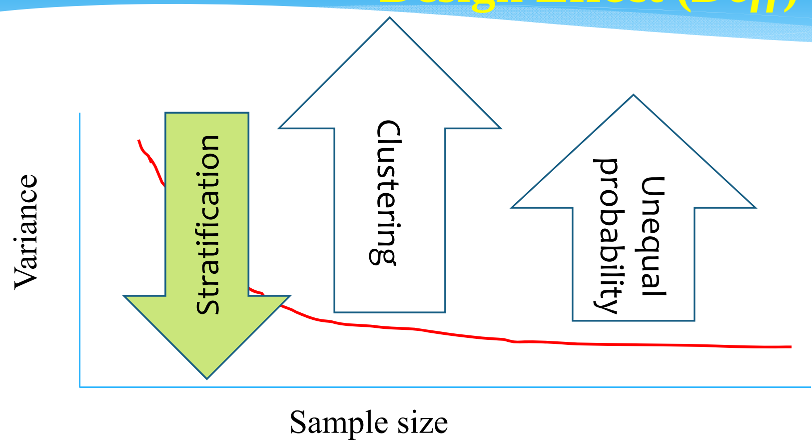
Design Effect (*Deff*)

- Design effect of a sample design, say *D*, is defined as the ratio of the variances of *D* and *SRSWOR*.

$$Deff = \frac{Var\ (design\ \ D)}{Var\ (SRSWOR)}$$

- For the sample designs used in practice (i.e. large scale sample surveys) the *Deff* is usually greater than 1.
- Estimates of *Deff* are often used for determining the required sample size for a given design.

Design Effect (*Deff*)



Deff: overall effect of design on variance

Deff>1 usually in practice