Small Area Estimation

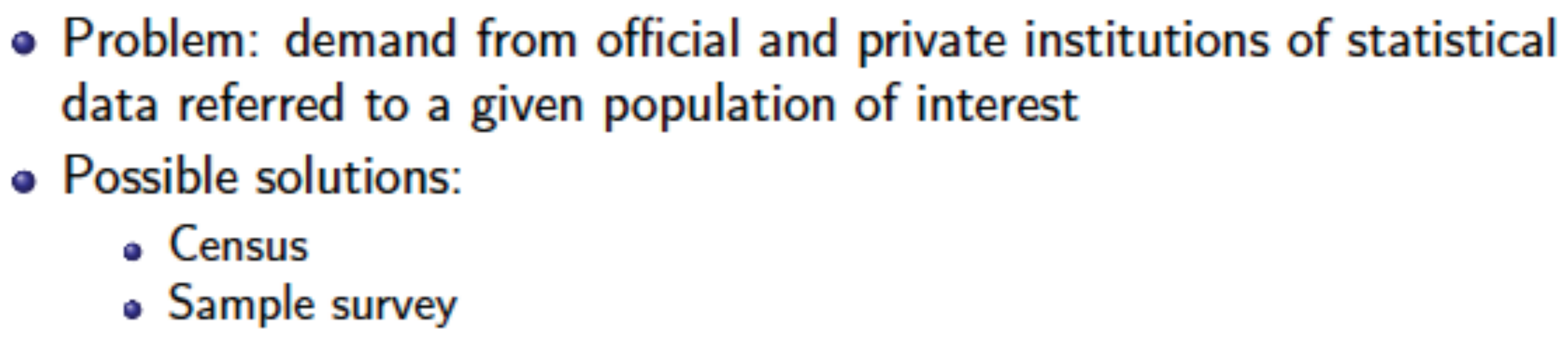
1 - Small area estimation problem

2 - Estimation for domains - Direct estimators – estimation for planned domains

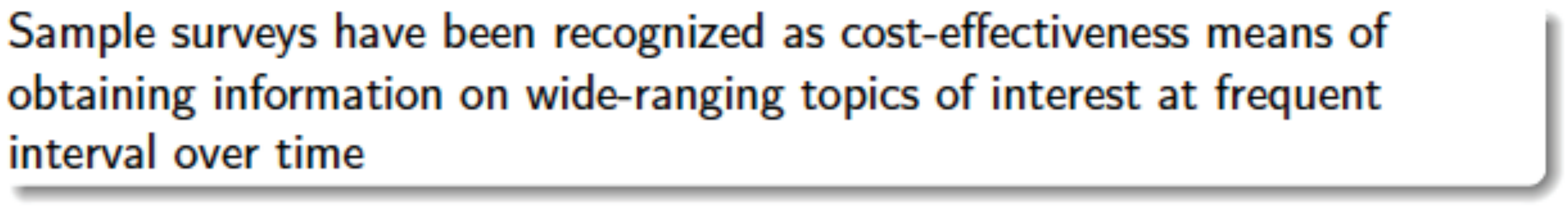
3 – Coefficient of Variation and Minimum level of precision

4- Estimation for unplanned domains and/or where the sample size is not enough for the minimum level of precision

1 - Small Area Estimation problem



- Problem: demand from official and private institutions of statistical data referred to a given population of interest
- Possible solutions:
  - Census
  - Sample survey

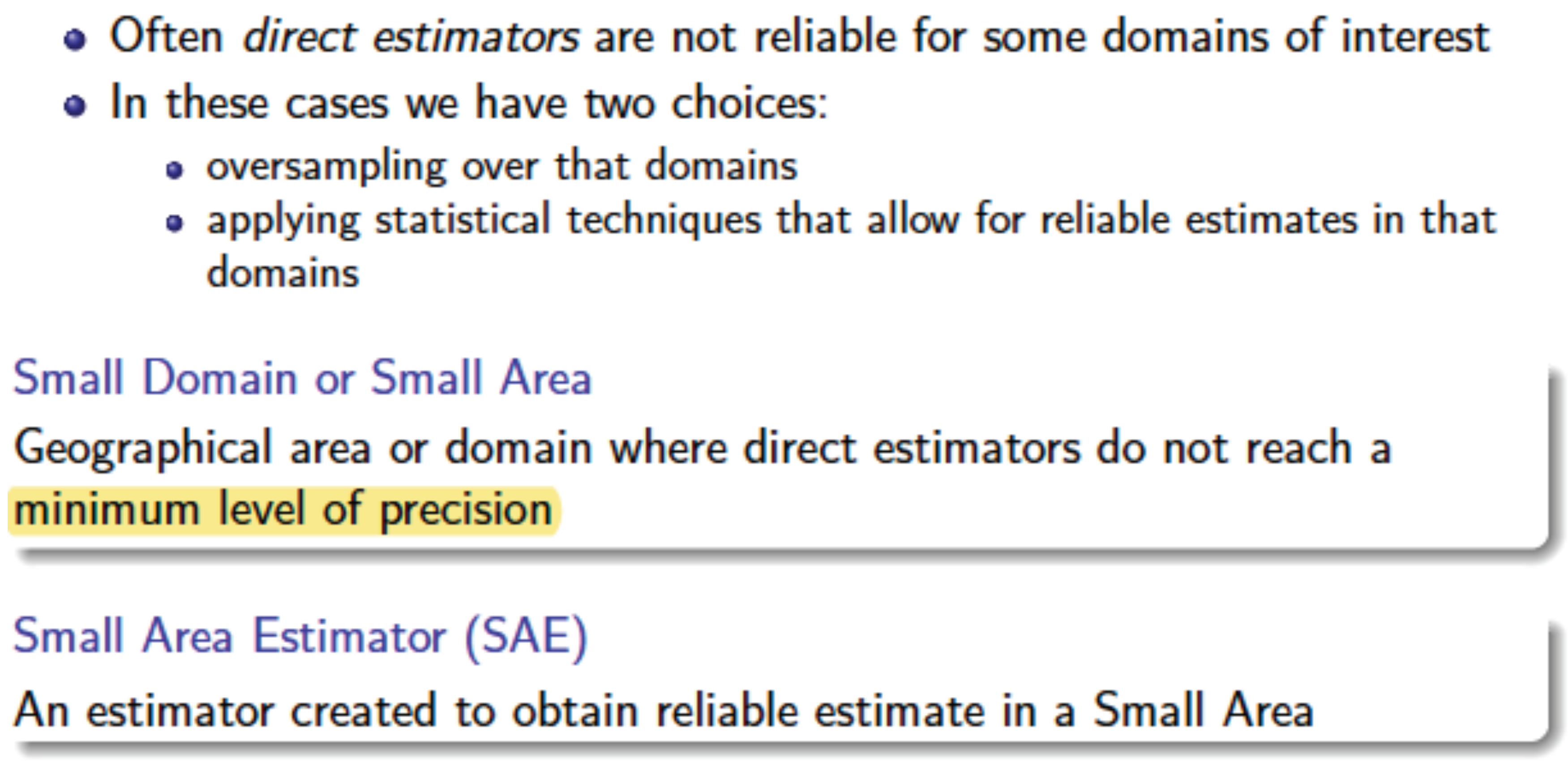


Sample surveys have been recognized as cost-effectiveness means of obtaining information on wide-ranging topics of interest at frequent interval over time

- Population of interest (or target population): population for which the survey is designed

  →*direct estimators* should be reliable for the target population
- Domain: sub-population of the population of interest, they could be planned or not in the survey design
  - Geographic areas (e.g. Regions, Provinces, Municipalities, Health Service Area)
  - Socio-demographic groups (e.g. Sex, Age, Race within a large geographic area)
  - Other sub-populations (e.g. the set of firms belonging to a industry subdivision)

  →we don't know the reliability of *direct estimators* for the domains that have not been planned in the survey design



- Often *direct estimators* are not reliable for some domains of interest
- In these cases we have two choices:
  - oversampling over that domains
  - applying statistical techniques that allow for reliable estimates in that domains

Small Domain or Small Area

Geographical area or domain where direct estimators do not reach a minimum level of precision

Small Area Estimator (SAE)

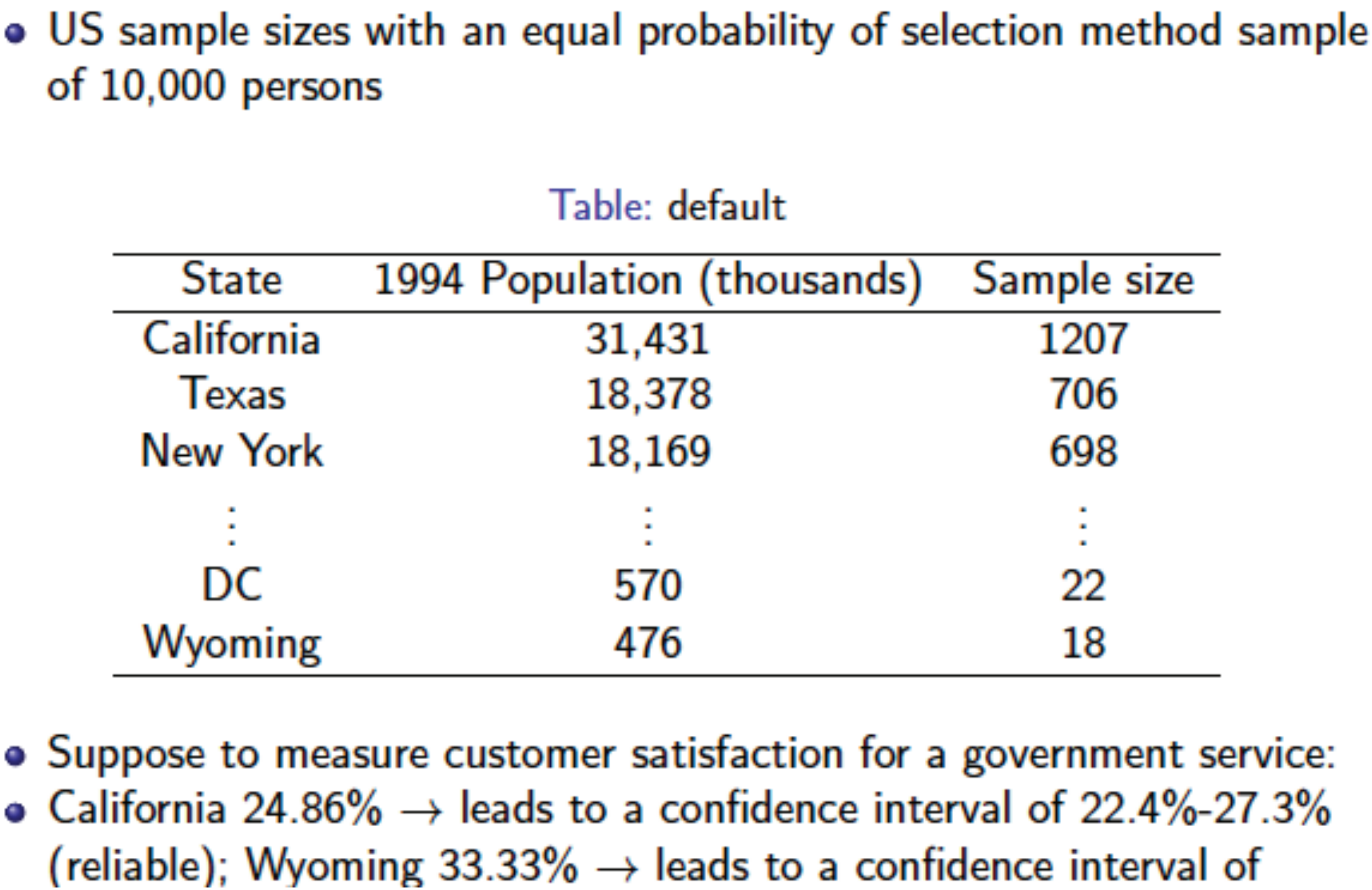
An estimator created to obtain reliable estimate in a Small Area

Oversampling is the practice of selecting respondents so that some groups make up a larger share of the survey sample than they do in the population. Oversampling small groups can be difficult and costly, but it allows polls surveys to shed light on groups that would otherwise be too small to report on.

This might sound like it would make the survey unrepresentative, but pollsters correct this through weighting. With weighting, groups that were oversampled are brought back in line with their actual share of the population – removing the potential for bias.

- Target population: households who live in an Italian Region
- Variable of interest: Income or other poverty measures
- Survey sample: EUSILC (European Union Statistics on Income and Living Conditions), designed to obtain reliable estimate at Regional level in Italy
  - planned design domains: Regions
  - unplanned design domains: e.g. Provinces, Municipalities

- EUSILC sample size in Tuscany: 1751 households
  - Pisa province 158 households → need SAE (or oversampling)
  - Grosseto province 70 households → need SAE (or oversampling)



- US sample sizes with an equal probability of selection method sample of 10,000 persons

Table: default

State	1994 Population (thousands)	Sample size
California	31,431	1207
Texas	18,378	706
New York	18,169	698
⋮	⋮	⋮
DC	570	22
Wyoming	476	18

- Suppose to measure customer satisfaction for a government service:
- California 24.86% → leads to a confidence interval of 22.4%-27.3% (reliable); Wyoming 33.33% → leads to a confidence interval of

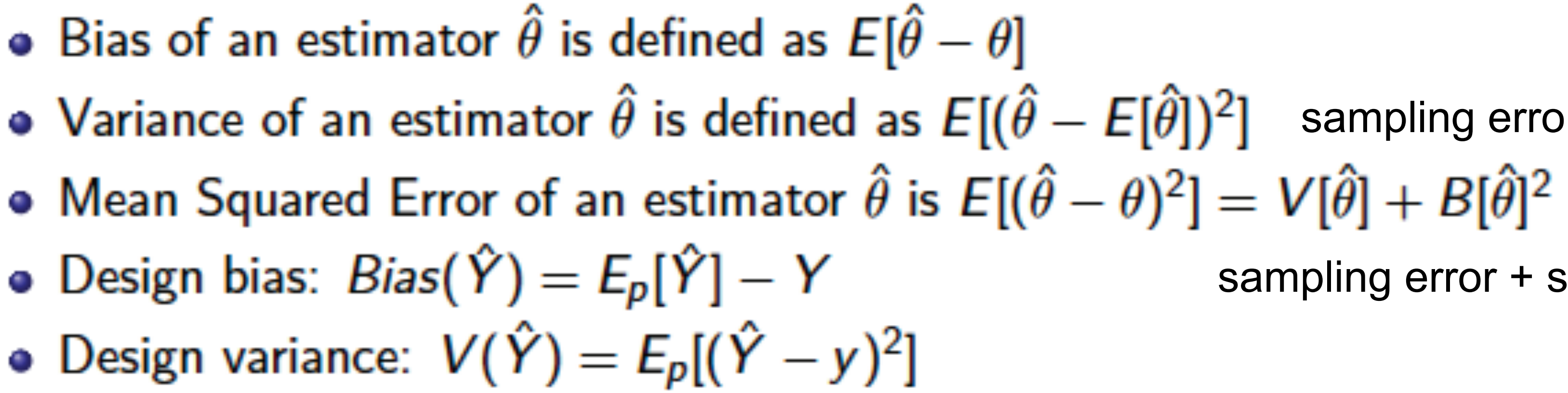
10.9%-55.7% (unreliable)

2 - Estimation for domains – direct estimators

Measures of *precision* *(MSE)* are usually computed to evaluate the quality of a population parameter estimate and to obtain valid inferences.

The estimation method and the sampling design determine the properties of the MSE and of the sampling error .

MSE Mean Squared Error measures the average of the squares of the errors—
that is, the average squared difference between the estimated values and the actual value.



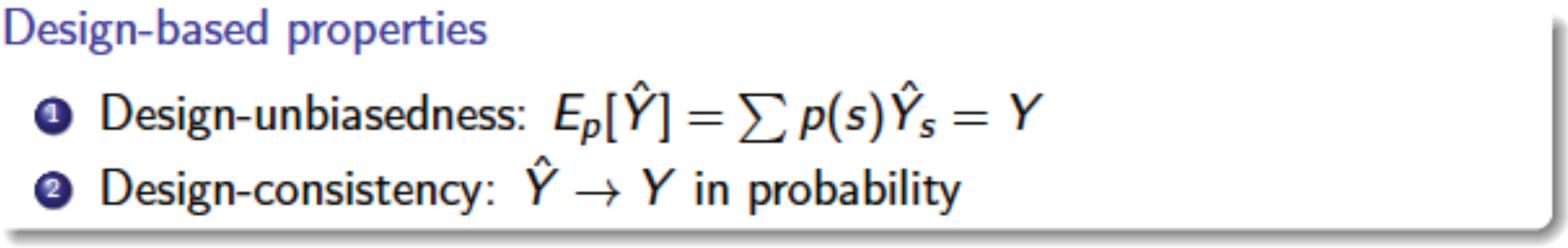
- Bias of an estimator $\hat{\theta}$ is defined as $E[\hat{\theta} - \theta]$
- Variance of an estimator $\hat{\theta}$ is defined as $E[(\hat{\theta} - E[\hat{\theta}])^2]$

sampling error

- Mean Squared Error of an estimator $\hat{\theta}$ is $E[(\hat{\theta} - \theta)^2] = V[\hat{\theta}] + B[\hat{\theta}]^2$
- Design bias: $Bias(\hat{Y}) = E_p[\hat{Y}] - Y$

sampling error + squared bias

- Design variance: $V(\hat{Y}) = E_p[(\hat{Y} - y)^2]$



Design-based properties

1. Design-unbiasedness: $E_p[\hat{Y}] = \sum p(s)\hat{Y}_s = Y$
2. Design-consistency: $\hat{Y} \to Y$ in probability

Expected values "p" are averages of the estimator values weighted
by the probability of occurrence, that is the probability that the sample "s" occurs

- Data $\{y_i\}, i \in s$
- Expansion estimator for the mean:

Horvitz -Thompson estimator

$$\hat{Y} = \frac{\sum_{i \in s} w_i y_i}{\sum_{i \in s} w_i}$$

- $w_i = \pi_i^{-1}$, the basic design weight
- π_i is the probability of selecting the unit i in sample s

Remark: weights w_i are independent from y_i

Estimation for domains: direct estimators

- The basic design-consistent *Horvitz-Thompson estimator* is the most natural estimator to use if there is no auxiliary information available at the estimation stage.

The use of auxiliary data drives to a more complex formulation of the direct estimator:
- ratio estimator
- regression estimator
- calibration estimator

we are not considering these estimatore

Estimation for domains: direct estimators

- Proper estimation conforms to the sampling design.
- Sampling weights are incorporated in the estimation process
- Sampling weights derive from: stratification, clustering, and multi-phase or multi-stage information

Estimation for domains: direct estimators

Use auxiliary data whenever possible to improve the reliability of the estimates (decrease MSE).

Evaluate the use of the auxiliary data.

The use of auxiliary data drives to a more complex formulation of the direct estimator:

- ratio estimator
- regression estimator
- calibration estimator

we are not considering these estimators here

Estimation for domains: direct estimators

Stat Canada suggestion 1:

“Whenever auxiliary data are available for sample units, together with known population totals for such data, consider using calibration estimation (“evolution of HT estimator!”) so that the weighted auxiliary data add up to these known totals. This may result in improved precision and lead to greater consistency between estimates from various sources.”

Estimation for domains: direct estimators

Stat Canada suggestion 2:

“Incorporate the requirements of small domains of interest at the sampling design and sample allocation stages (Singh, Gambino and Mantel, 1994).

If this is not possible at the design stage, or if the domains are only specified at a later stage, consider special estimation methods (small area estimators) at the estimation stage. These methods “borrow strength” from related areas (or domains) to minimize the mean square error of the resulting estimator (Platek et al., 1987; Ghosh and Rao, 1994; Rao, 1999).”

Definitions and notation - 1

Fixed and finite population $U = \{1,2,...,k,...,N\}$, where k refers to the *label* of population element

The fixed population is said to be generated from a *superpopulation*.

Variable of interest y

For practical purposes, we are interested in one particular realized population U with $(y_1, y_2, ..., y_N)$, not in the more general properties of the process (or model) explaining how the population evolved.

NOTE: In the *design-based* approach, the values of the variable of interest are regarded as *fixed but unknown* quantities. The only source of randomness is the *sampling design*, and our conclusions should apply to hypothetical repeated sampling from the fixed population.

Definitions and notation - 2

Basic parameters for study variable *y* for the whole population:

$$\text{Total } t = \sum_{k \in U} y_k$$

$$\text{Mean } \bar{y} = \sum_{k \in U} y_k / N$$

We discuss here the estimation of totals

In practice, the values y_k of *y* are observed in an *n* element sample $s \subset U$ which is drawn by a sampling design giving probability $p(s)$ to each sample s

NOTE: The sampling design can be *complex* involving stratification and clustering and several sampling stages

Definitions and notation - 4

Variance estimators are derived in two steps:

(1) The theoretical design-based variance $Var(\hat{t})$ (or its approximation if the theoretical design variance is intractable) is derived

(2) The derived quantity is estimated by a design unbiased or design-consistent estimator $\hat{V}(\hat{t})$

NOTE: An estimator is *design consistent* if its design bias and variance tend to zero as the sample size increases

Definitions and notation - 5

Inclusion probability: An observation k is included in the sample with probability $\pi_k = P\{k \in s\}$

The inverse probabilities $a_k = 1/\pi_k$ are called *design weights*

Sample membership indicator:
$I_k = I\{k \in s\}$ with value 1 if k is in the sample and 0 otherwise

Expectation of sample membership indicator $E(I_k) = \pi_k$

Probability of including both elements k and l $(k \neq l)$ is $\pi_{kl} = E(I_k I_l)$ with inverse $a_{kl} = 1/\pi_{kl}$ ($a_{kl} = a_k$ when $k = l$)

The covariance of I_k and I_l is $Cov(I_k, I_l) = \pi_{kl} - \pi_k \pi_l$

Estimation for domains

Domain estimation of totals or averages of variable of interest y over D non-overlapping domains $U_d \subset U$, $d = 1,2,\ldots,d,\ldots,D$, with possibly known domain sizes N_d

Example: Population of a country is divided into D domains by regional classification, with N_d households in domain U_d

The aim is to estimate statistics on household income for the regional areas (domains)

The key parameter is **domain total**: $t_d = \sum_{k \in U_d} y_k$,

where y_k refers to measurement for household k

Why domain totals are important?

Totals are basic and the simplest descriptive statistics for continuous (or binary) study variables

Many other, more complex statistic are functions of totals

Domain ratio: $R_d = \frac{t_{dy}}{t_{dz}} = \frac{\sum_{k \in U_d} y_k}{\sum_{k \in U_d} z_k}$

Estimator: $\hat{R}_d = \frac{\hat{t}_{dy}}{\hat{t}_{dz}} = \frac{\sum_{k \in s_d} a_k y_k}{\sum_{k \in s_d} a_k z_k}$

Domain mean: $\bar{y}_d = t_d / N_d$

Estimator: $\hat{\bar{y}}_d = \hat{t}_d / N_d$ or $\hat{\bar{y}}_d = \hat{t}_d / \hat{N}_d$

Domain totals are important also because many Poverty measures - Laeken indicators P_0 and P_1 are functions of them

.... next lecture topic