PRASC

Project for the Regional Advancement of Statistics in the Caribbean

Projet régional pour l'avancement de la statistique dans les Caraïbes

Funded by the Government of Canada | Canada

Project for the Regional Advancement of Statistics in the Caribbean

PRASC

Component: Household Survey Infrastructure

Funded by the Government of Canada | Canada

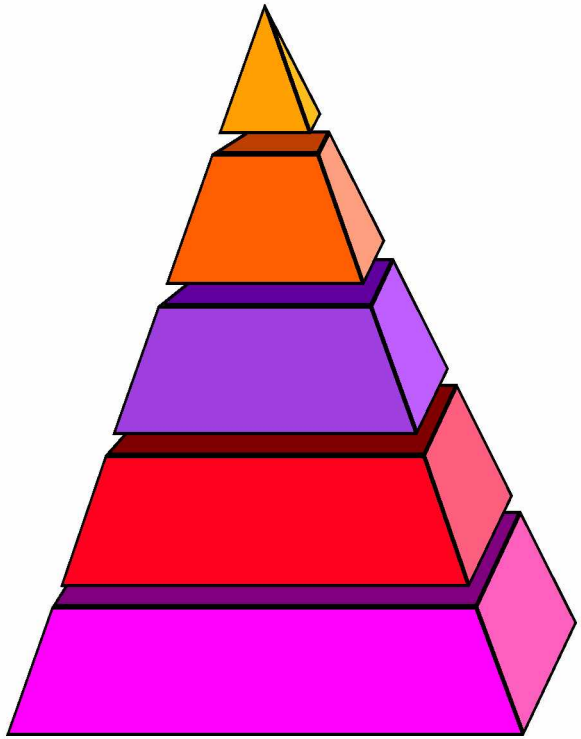
Survey Sampling and Estimation

Denis Malo &
Wisner Jocelyn
Senior Methodologists, SSMD

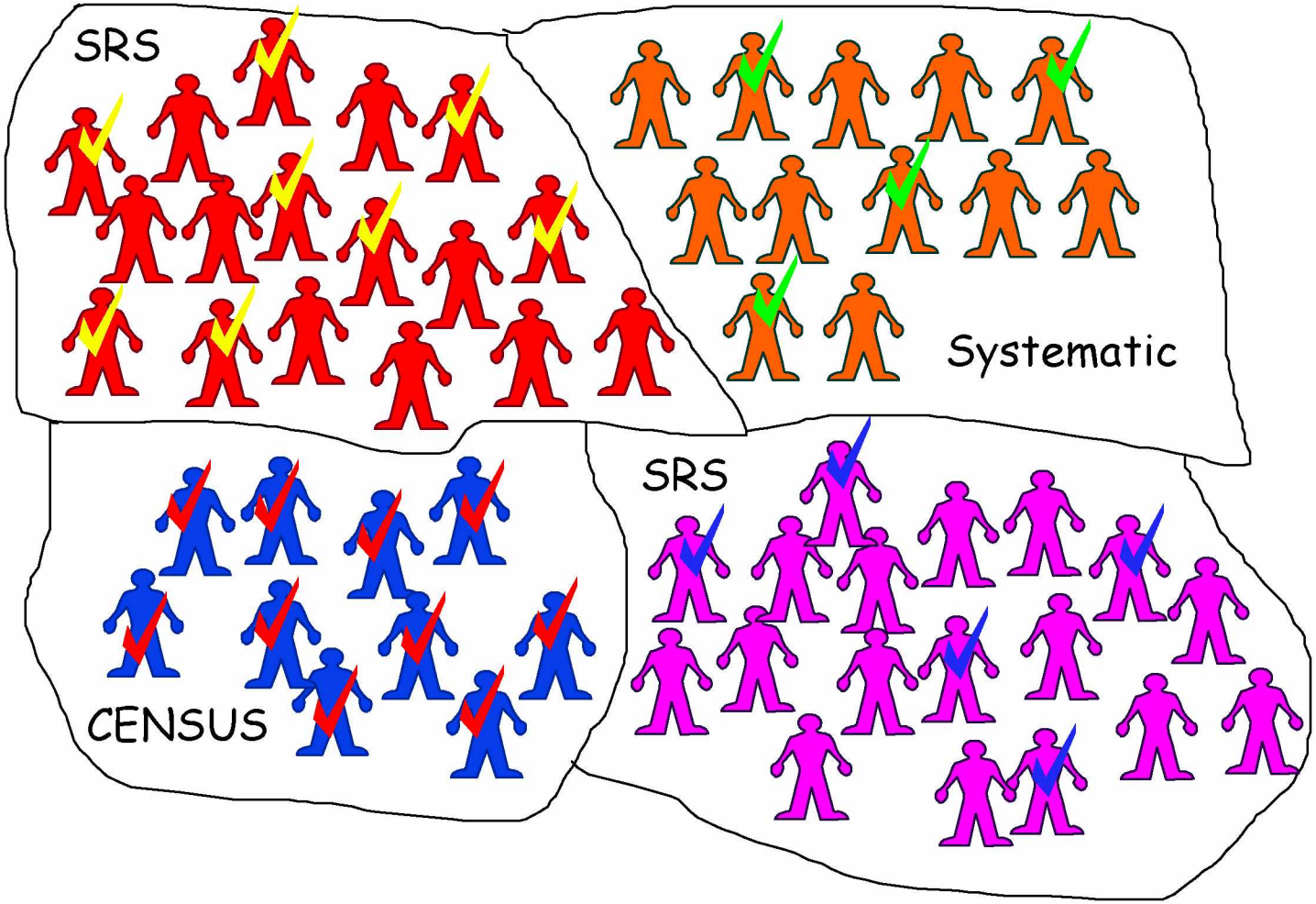
September 30 – October 4
2019

Stratified sampling

- Process of dividing population into *homogeneous* groups, called *strata*, and then selecting independent samples in each *stratum*
- Stratification variables may be geographic and/or socio-economic, etc.
- For Business surveys Stratification by Industrial classification (restaurants, garages), size measures (number of employes, Sales, etc.
- Stratification is limited to the items of information available on the survey frame



Stratified sampling



Stratified sampling

- Four-step process

1. Divide population into homogeneous, non-overlapping groups called strata
2. Allocate the total sample size among strata
3. Select samples independently in each stratum
4. Estimate survey results within each stratum

Stratified sampling

- Sum the stratum estimates to obtain estimated population totals

- Do similar calculations for means and proportions, to obtain numerator and denominator

Stratified sampling

- Stratification is **not** a sample selection method
- It is a method which allows one to impose a structure on the population
- The aim is to establish homogeneity within each stratum, and diversity (heterogeneity) between strata
  - with respect to key survey variables

Stratified sampling

- Strata may correspond to geographic, administrative, social or economic partitions
- Stratification by size of unit is common in economic surveys
- Stratification normally improves the precision of estimates, as compared to SRS; i.e., is more efficient

Stratified sampling

- **Example**: Fuel Consumption Survey
  - **Objective**: To study gasoline consumption of private cars in Canada
  - **Target population**: Private-use cars
  - **Sampling frame**: Vehicle registration file for each province
  - **Required outputs**: Quarterly estimates at province level
  - **Sample**: Approx. 750 vehicles per month

Stratified sampling

- **Data collection**: first contact by telephone, followed by mail-out of a “diary” to be filled in and mailed back
- **Sample design**: stratification, with SRS in each stratum
- **Stratification**: by province; year of manufacture; weight/wheelbase/number of cylinders (depending on province); urban/rural

Stratified sampling

- Advantages
  - Increases precision of overall population estimates, for given cost
  - Guarantees that all important groups of the population are represented in the sample
  - Permits efficient estimation at stratum level
  - Operational or administrative convenience
  - Uses auxiliary information
  - May use any selection method in each stratum

Stratified sampling

- Disadvantages
  - Sampling frame is more complex
  - Must know stratification variables for all elements on frame
  - If allocation is far from optimum, can be less efficient than SRS
  - Estimation is more complex
  - Sample design is more costly to prepare
  - Stratification efficient for one variable may not work well for others

Stratified sampling

- Factors that influence efficiency
  - choice of stratification variables
  - number of strata
  - determination of stratum limits
  - allocation of sample to strata

Stratified sampling: example

- Again consider the population of 6 farms
- Suppose we now know the size of each farm in hectares
- Again, we can only afford to select 2 units

Stratified sampling: example

Farm	Hectares	Expenses
1	50	$ 26,000
2	1000	470,000
3	125	63,800
4	300	145,000
5	500	230,000
6	25	12,500
Total	**2000**	**947,300**

Stratified sampling: example

- Step 1 Divide population into 2 strata: large farms and small farms
- Step 2 Allocate total sample **(n=2)** between strata [here, it must be 1 and 1]
- Step 3 Select farm in each stratum
- Step 4 Estimate expenses in each stratum and sum for total

Stratified sampling: example

Farm	Hectares	Expenses
Stratum 1	≤200 ha.	
1	50	$ 26,000
3	125	63,800
6	25	12,500
Stratum 2	>200 ha.	
2	1000	470,000
4	300	145,000
5	500	230,000
Total	**2000**	**947,300**

Possible samples	Observed expenses	Estimated total
(1,2)	496,000	1,488,000
(1,4)	171,000	513,000
(1,5)	256,000	768,000
(2,3)	533,800	1,601,400
(2,6)	482,500	1,447,500
(3,4)	208,800	626,400
(3,5)	293,800	881,400
(4,6)	157,500	472,500
(5,6)	252,500	727,500

9 possible samples **average 947,300 : unbiased**

Possible samples	Estimation SRS	Stratified SRS
(1,2)	1,488,000	1,488,000
(1,3)	269,400	
(1,4)	513,000	513,000
(1,5)	768,000	768,000
(1,6)	115,500	
(2,3)	1,601,400	1,601,400
(2,4)	1,845,000	
(2,5)	2,100,000	
(2,6)	1,447,500	1,447,500
(3,4)	626,400	626,400
(3,5)	881,400	881,400
(3,6)	228,900	
(4,5)	1,125,000	
(4,6)	472,500	472,500
(5,6)	727,500	727,500

15 possible samples **average 947,300 : unbiased**

Sample allocation

- Often, total sample is determined by budget and resources: how to best allocate to strata?
- **Fixed Sample size**
  - Sample size **n** allocated to the strata using a specific allocation (Proportional to size, Optimal allocation, Power allocation)
- Fixed Coefficient of Variation (CV)
  - Sample size **n** unknown
  - Use the square of CV to derive allocation factors (proportion of units within each stratum)

Sample allocation

- **Proportional allocation**
  - Number of units is proportional to the population size of each stratum (N-Proportional) or to a size measure variable **X** (X-proportional)
- **Power allocation (Population size: N^P or size variable X: X^P)**
  - Square root allocation is a particular case of power allocation (P=.5)
  - Provide better estimates at stratum level or for smaller domains

Sample allocation

- **Optimal allocation**
  - Considers cost and variance by stratum
  - Strata with greater variance receive more sample units
  - Strata with greater cost receive fewer sample units

Sample allocation

- If no information is available on cost or variance, proportional allocation gives optimal precision for the whole population
  - but small strata will have very imprecise estimates

Sample allocation

- Allocation of an equal number of sample units per stratum gives equal precision for each stratum
  - but precision is not optimal for whole population

- One compromise, useful in practice, is allocation proportional to the square root of the stratum sizes

Sample allocation

- **Sub-populations**
  - small or localized sub-populations may be of special interest
  - they may be made into strata if the sample frame permits
  - they are then given sufficient sample units to provide reliable estimates

Sample allocation example

Stratum	Population	Equal		N-Proportional		Square root		
		Sample	CV	Sample	CV	√N	Sample	CV
1	100,000	667	3.86%	1,667	2.43%	316	1,242	2.82%
2	15,000	667	3.79%	250	6.28%	122	481	4.49%
3	5,000	667	3.61%	83	10.95%	71	278	5.84%
Total	120,000	2,001	3.26%	2,000	2.22%	509	2,001	2.43%

Note: Suppose n=2,000 to be allocated, SRS in each stratum,
CV is for p=50%

Unequal-probability sampling

- With SRS or other equal-probability methods, the same weight is given to every unit to ensure unbiased estimates
- In many cases, it is desirable to use information on size of units (for example, for the 6 farms)
- Because we suspect a positive correlation between size (in hectares) and certain variables (expenses in $)

Unequal-probability sampling

- We could give each sample unit a weight inversely proportional to its size
- Large farms would have small weights, and *vice versa*
- Estimates should not be too far from the true value
- We would have reduced the variance of the estimates

Unequal-probability sampling

- **Problem**: the average over all possible estimates is not the true value
- Estimation would be **biased**
- Giving the same selection probability to every unit, but with size-based weights, would lead to **over-estimation** of the true value

Farm example

Farm	Hectares	Expenses
1	50	$ 26,000
2	1000	470,000
3	125	63,800
4	300	145,000
5	500	230,000
6	25	12,500
Total	**2000**	**947,300**

Farm example

SRS with **n = 1** and equal weights (= 6)

Farm	Hectares	Est. expenses
1	50	$ 156,000
2	1000	2,820,000
3	125	382,800
4	300	870,000
5	500	1,380,000
6	25	75,000
Average		**947,300**

Farm example

SRS, but with weight proportional to 1/size

Farm	Hectares	Est. expenses
1	50	$1,040,000
2	1000	940,000
3	125	1,020,800
4	300	966,667
5	500	920,000
6	25	1,000,000
Average		**981,245**

PPS sampling

- **Solution to bias problem**
  - Give each unit (farm) a selection probability proportional to its size
  - This is called “probability proportional to size” sampling (PPS)
  - Then the natural weight is the inverse of the unit’s size, as we tried to “force” above
  - And now the estimator is unbiased

PPS sampling

- This option is attractive if some units are very large and many are quite small
- It permits making selection more likely for units whose values contribute heavily to the total
- These units then carry smaller sampling weights

Farm example

Farm	Estimated expenses	Relative frequency
1	$1,040,000	50/2000
2	940,000	1000/2000
3	1,020,800	125/2000
4	966,667	300/2000
5	920,000	500/2000
6	1,000,000	25/2000

PPS sampling

- **Advantages**
  - Uses auxiliary information
  - Can give dramatic improvements in precision

- **Disadvantages**
  - Requires a stable size measure for all units
  - Estimation is more complex
  - Variance estimation can be complicated

PPS sampling

- **When to use PPS?**
  - When a precise size measure is available
  - When there is a strong correlation between size and the key variables of study

PPS sampling

- **Methods which use PPS**
  - PPS random
  - PPS systematic
  - PPS randomized systematic

PPS sampling

- **PPS random**
  - Calculate cumulative size measures
  - Determine range corresponding to each unit
  - Select random number between **1** and total size
  - Select unit in corresponding range
  - Repeat until **n** units are selected
  - Can be done with or without replacement

PPS sampling

PPS systematic

- Calculate cumulative size measures
- Determine range corresponding to each unit
- Determine sampling interval **K = total size / n**
- Select random number between **1** and **K**
- Select unit in corresponding range
- Select units **R, R+K, R+2K,** etc.

PPS sampling

PPS randomized systematic

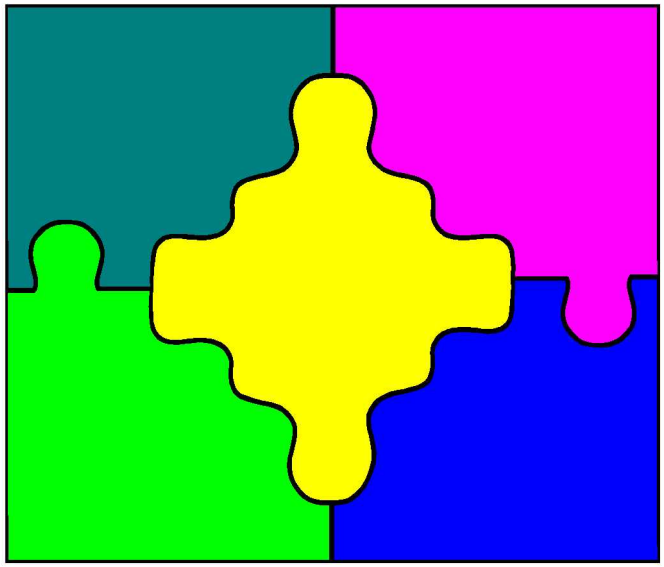
- Randomize order of units on frame
- Then do PPS systematic as described above

List samples and area samples

- **List sample**
  - Selected from a complete list of the units (elements) of the survey population

- **Area sample**
  - Involves selection of geographic areas
  - Need to have an up-to-date list is limited to the selected areas only
  - Methods include cluster sampling and multi-stage sampling

Cluster sampling



- Process of selecting complete groups *(clusters)* of population units
- Used when no satisfactory frame for individual population units exists
- Used to reduce field costs
- May be less precise than SRS because neighbouring units tend to be similar

Cluster sampling

- A two-step process
  - Group the population into clusters which can be identified on maps and in the field
  - Select a sample of clusters and interview all elements within them

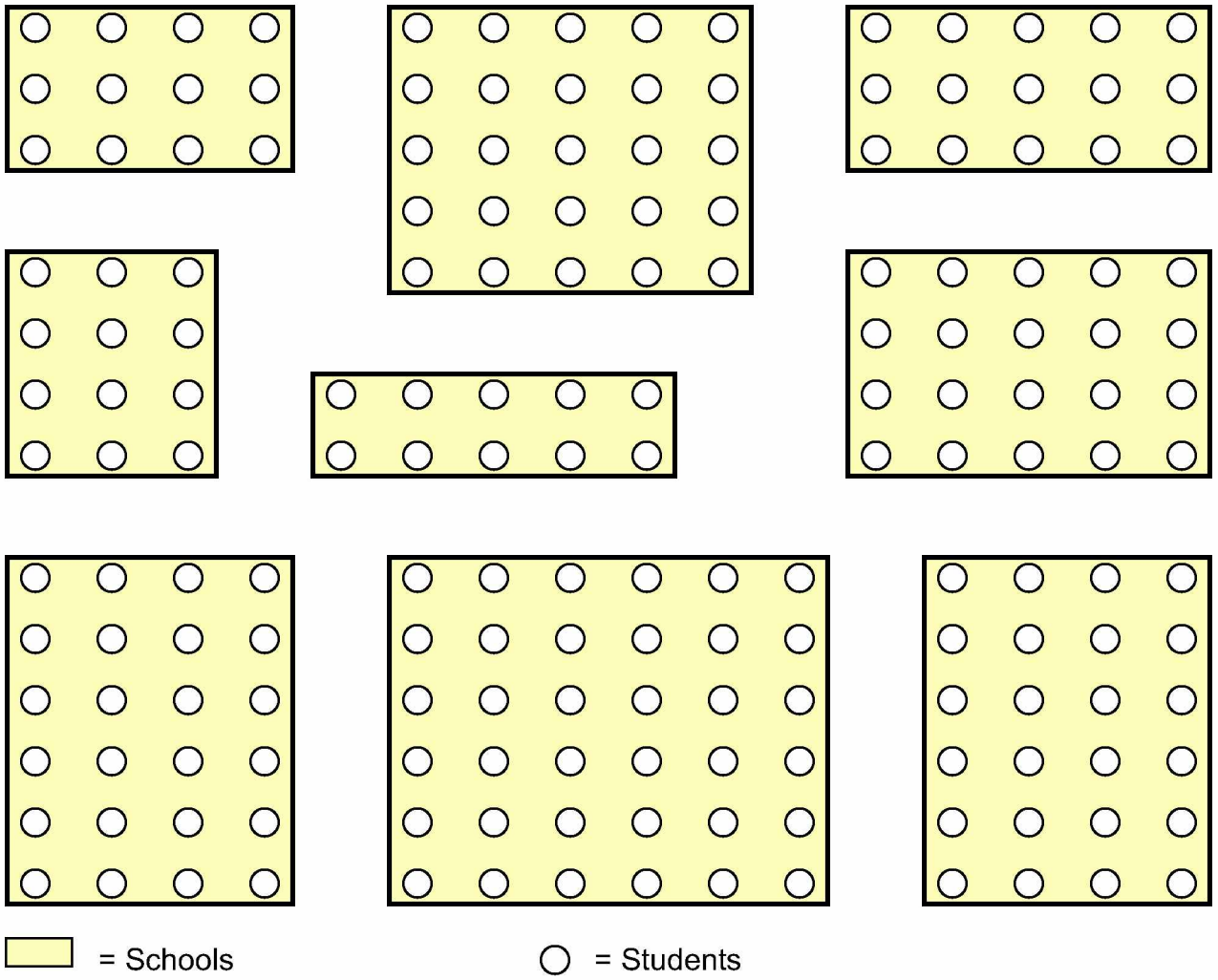
Cluster sampling

- Clusters may be natural or artificial groupings
- Possibly available from sources such as Census (blocks, EAs, etc.)
- Survey designers may have to make them

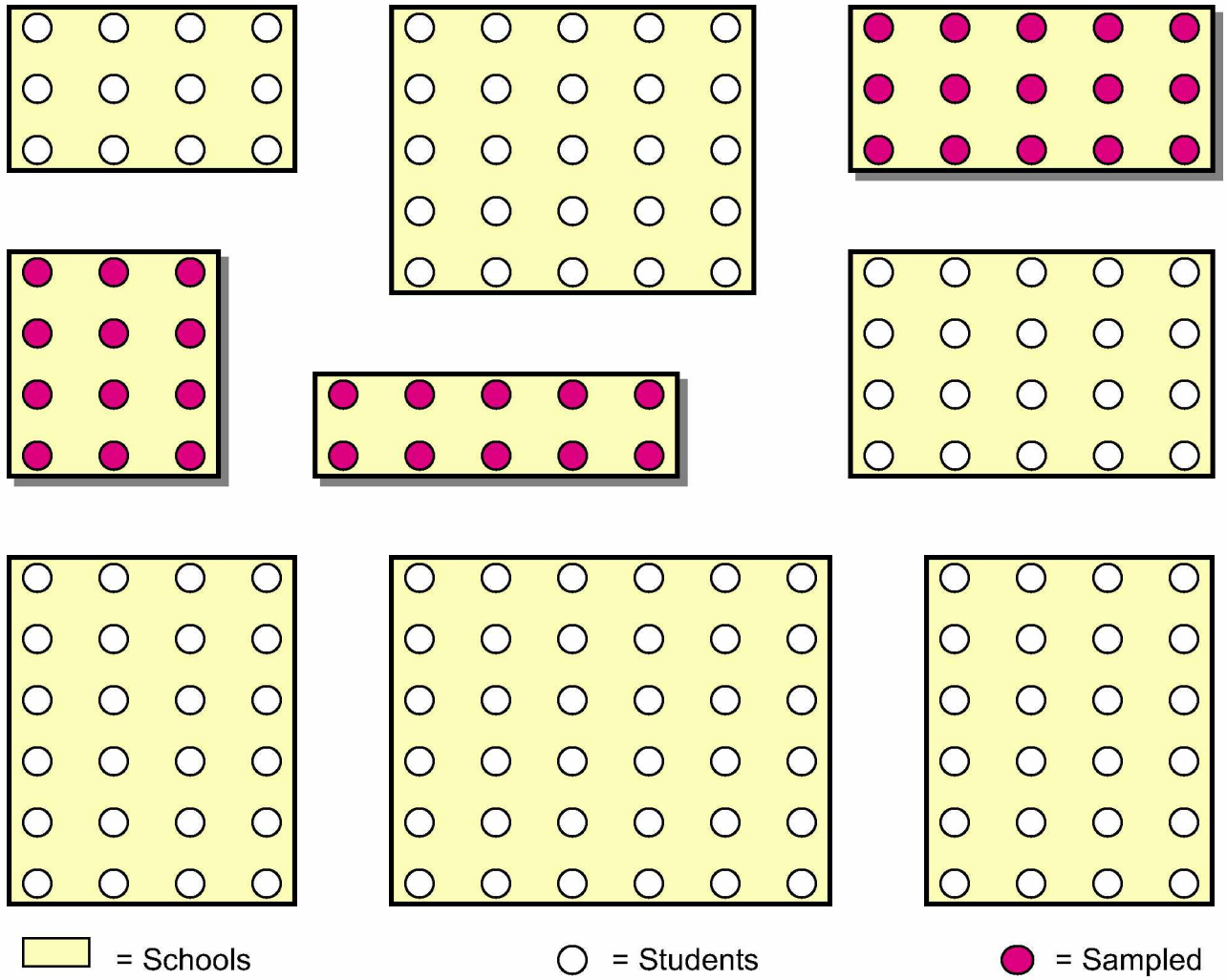
Cluster sampling

- Population is seen as hierarchy of units
  - people live in dwellings
  - dwellings make up blocks
  - many blocks make up a town or city

Student survey



Cluster sample



Cluster sampling

- **Advantages**
  - Can be used even if there is no list of population units
  - For personal interviews, travel time and cost are greatly reduced, particularly for rural populations
  - Only a list of clusters is required
  - Or the possibility of building one

Cluster sampling

- **Disadvantages**
  - Tendency of neighbouring units to be similar reduces efficiency
  - For given **n**, estimates are less precise
  - But if field costs are considered, the possibility of increasing **n** may (partially) offset this loss

Cluster sampling

- If clusters are heterogeneous with respect to study variables, cluster sampling can even be more efficient than SRS

- Example: cluster is household, variables are demographic
  - each household is a “mini-population”

Cluster sampling

- **Considerations**
  - Must have access to a population which can be given a hierarchical structure
  - Clusters must be as heterogeneous as possible
  - Clusters must not be too large (to control interviewing costs and assignment sizes)
  - Preferable to use many small clusters
  - Clusters are selected with SRS, SYS or PPS

Cluster sampling

- **Lack of heterogeneity**
  - Neighbouring farms may resemble each other as to soil type, crops, etc.
  - Households in neighbourhood may have similar income, houses or levels of education

- In such cases, cluster sampling is inefficient

Cluster sampling

- **Sample design effect**
  - ratio of variance under the design to variance under SRS
  - if a cluster design has twice the variance of SRS with the same sample size, the design effect is 2

Multi-stage sampling

- Sample is selected in two or more stages
- Hierarchy of units is required
- Units (e.g. provinces) are selected at the first stage
- Second-stage sub-samples are selected within each selected unit (e.g. cities)

Multi-stage sampling

- First-stage units are called primary sampling units (PSUs)
- Only PSUs selected at the first stage need to be subdivided into second stage units (SSUs)
  - for a two-stage design, SSUs are population elements
- and so on

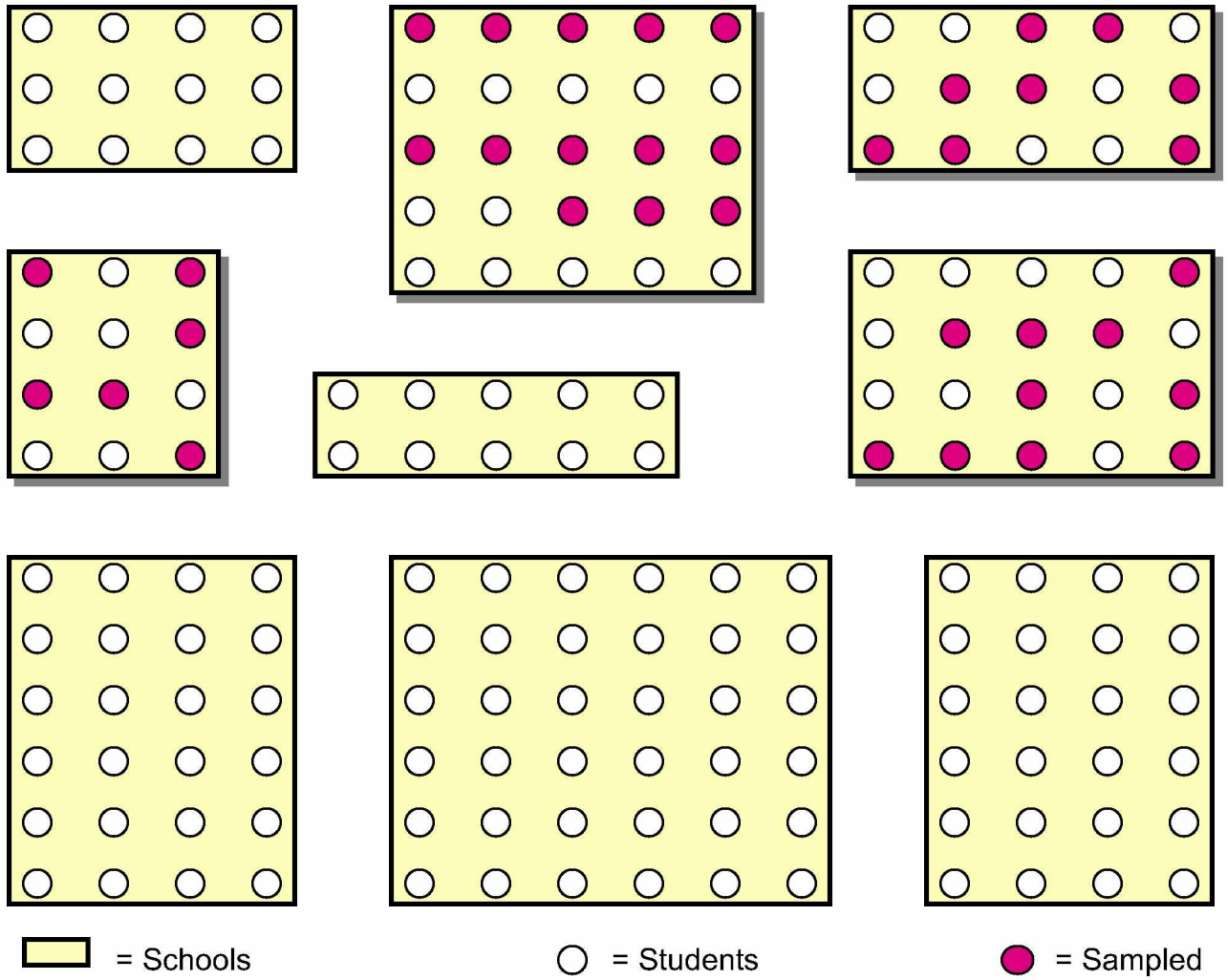
Multi-stage sampling

- A good sampling frame is required at each stage for the units to be selected at that stage
- Generally, frames for higher-level units (PSUs) are more stable than those for lower-level units (farms, dwellings, ...)

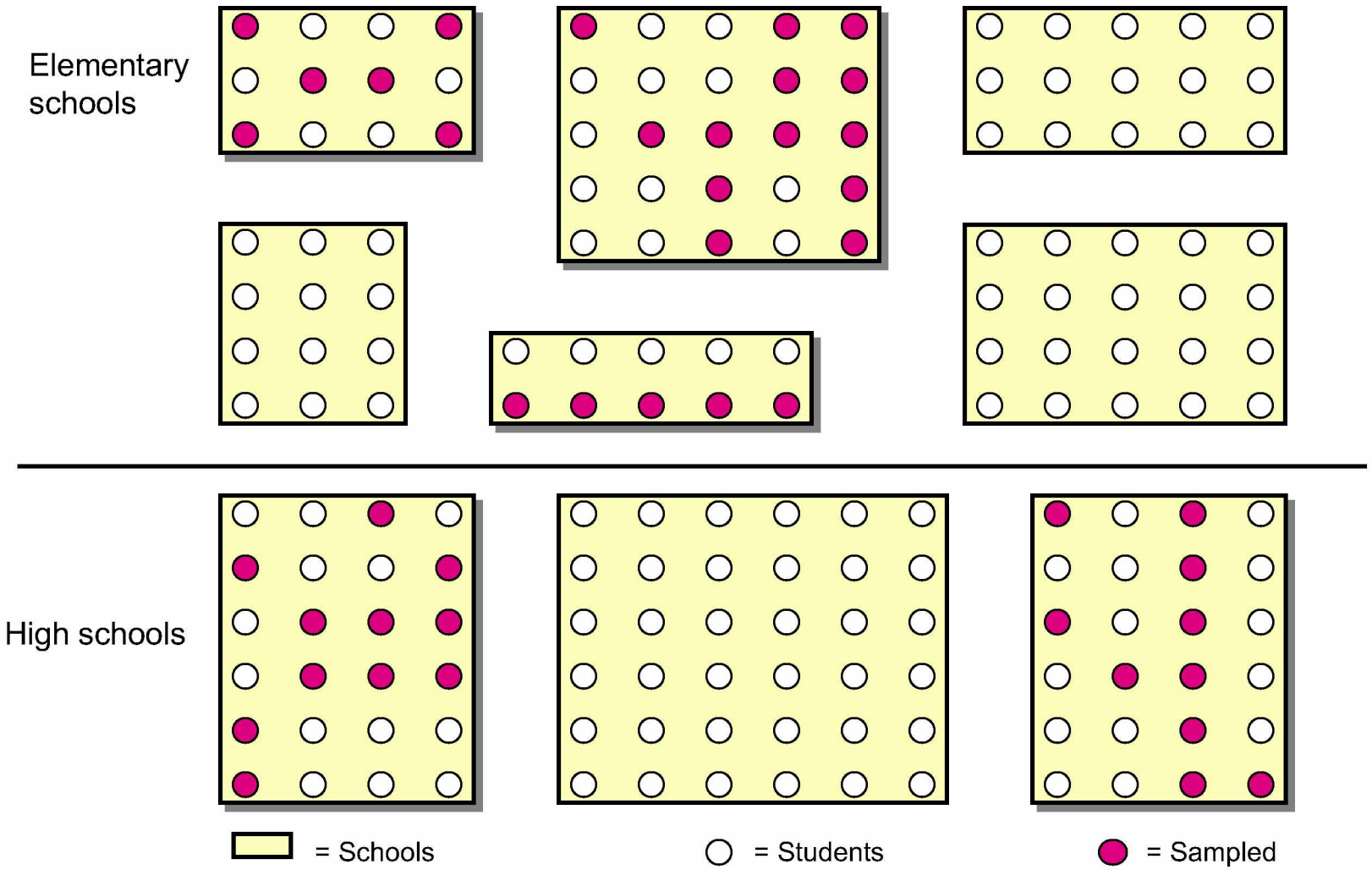
Multi-stage sampling

- Lower-level units are usually established on the basis of field counts
- Method is useful when there is no population list
- It may be the only way to access individuals
- Cost-effective method of distributing the sample throughout the population

Multi-stage sample



Stratified multi-stage sample



Multi-stage sampling

- **Example**: two-stage household sample
  - First stage: selection of urban blocks with $\mathbf{p_1 = 1/100}$
  - Second stage: selection of dwellings with $\mathbf{p_2 = 1/5}$
  - Probability of selection for all dwellings in the city is $\mathbf{p = p_1 * p_2 = 1/500}$

Multi-stage sampling

- Hints for efficiency
  - limit Stages: 2 stages should be the norm
  - Make first stage cluster small: pick more clusters (with fewer units within each selected cluster
  - for a fixed sample size:
    - schools; classes within schools; students within classes
    - schools; students within schools
    - classes; students within classes

Multi-stage sampling

- **Advantages**
  - Collection is concentrated and costs reduced
  - Can be used without complete frame
  - Selection methods can vary from one part of the frame to another

Multi-stage sampling

- **Disadvantages**
  - Design effect and loss of efficiency
  - Construction of sampling frame is complex
  - Estimation becomes more complex

Multi-stage sampling

- **General procedure**
  - preparation of list of PSUs
  - stratification, sample allocation and selection of PSUs
  - preparation of lists of SSUs within PSUs selected for the sample
  - stratification, sample allocation and selection of SSUs
  - etc.

You can contact the PRASC team at:

statcan.prasc-prasc.statcan@canada.ca

Canada