DATA COLLECTION

PROBABILISTIC SAMPLING

PROBABILISTIC SAMPLING

Probabilistic sample designs are usually more **difficult** and **expensive** to set-up (due to the need for a quality survey frame) and take longer to complete.

They provide **reliable estimates** for the attribute of interest and the **sampling error**, paving the way for small samples being used to draw inferences about larger target populations (in theory, at least; the non-sampling error components can still affect results and generalisation).

SAMPLING DESIGNS

Different **sampling designs** have distinct advantages and disadvantages.

They can be used to compute estimates

- for various population attributes: mean, total, proportion, ratio, difference, etc.
- for the corresponding 95% CI.

We might also want to compute sample sizes for a given **error bound** (an upper limit on the radius of the desired 95% CI), and how to determine the **sample allocation** (how many units to be sampled in various sub-population groups).

SAMPLING UNIVERSE

Target population:

- N units and measurements $\mathcal{U} = \{u_1, \dots, u_N\}$

True population attributes:

- mean μ, variance σ^2, total τ, proportion p

Sample population:

- n units and measurements $\mathcal{Y} = \{y_1, \dots, y_n\} \subseteq \mathcal{U}$

Sample population attributes:

- sample mean $\bar{y}$, sample variance s^2, sample total $\hat{\tau}$, sample proportion $\hat{p}$

PROBABILISTIC SAMPLING DESIGNS

Simple random sampling (SRS)

Stratified random sampling (StS)

Systematic sampling (SyS)

Cluster sampling (ClS)

Replicated sampling (ReS)

Multi-stage sampling (MSS)

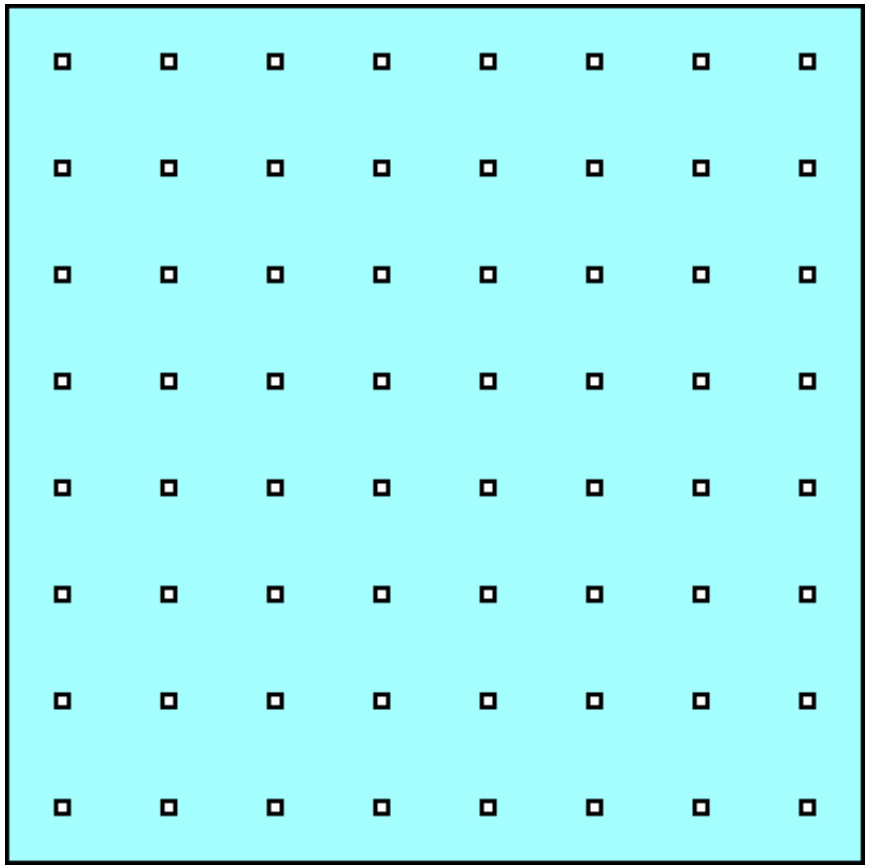
Multi-phase sampling (MPS)

Probability proportional-to-size sampling (PPS)

SAMPLING UNIVERSE

Goal: estimate the true population attributes μ, σ^2, τ, p *via* the sample population attributes $\bar{y}$, s^2, $\hat{\tau}$, $\hat{p}$, n, and the size N of the target population.

We look for **confidence intervals** (typically 95%).



SIMPLE RANDOM SAMPLING (SRS)

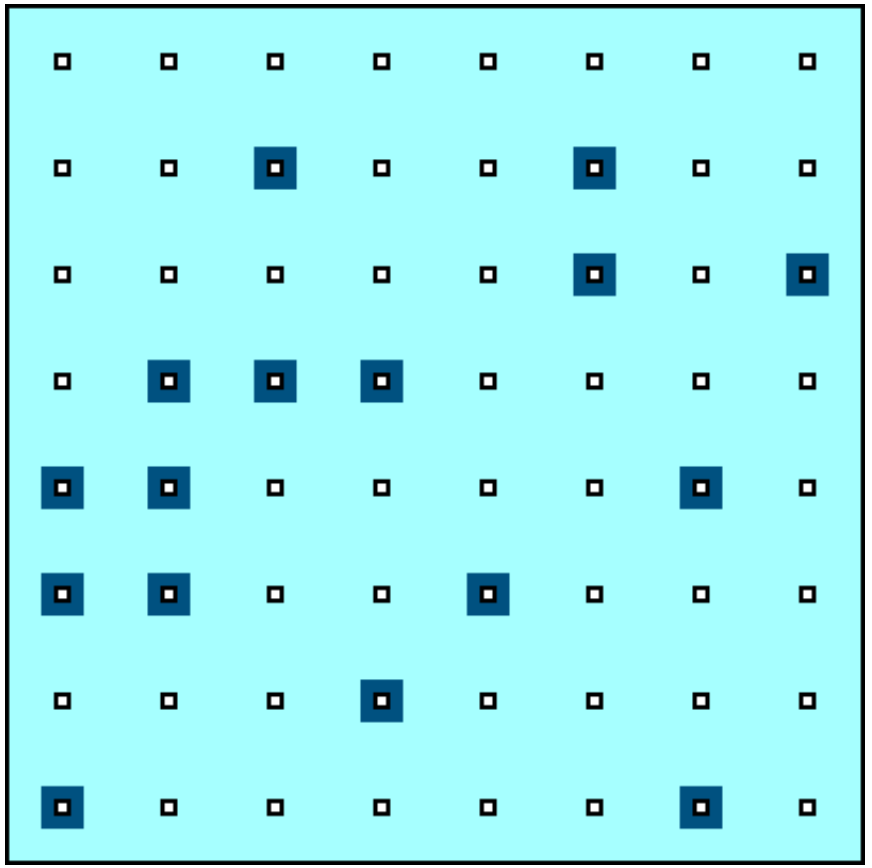
In SRS, we select n units randomly from the frame.

Advantages:

- easiest sampling design to implement
- sampling errors are well-known and easy to estimate
- does not require auxiliary information

Disadvantages:

- makes no use of auxiliary information
- no guarantee that the sample is representative
- costly if sample is widely spread out, geographically



STRATIFIED RANDOM SAMPLING (STS)

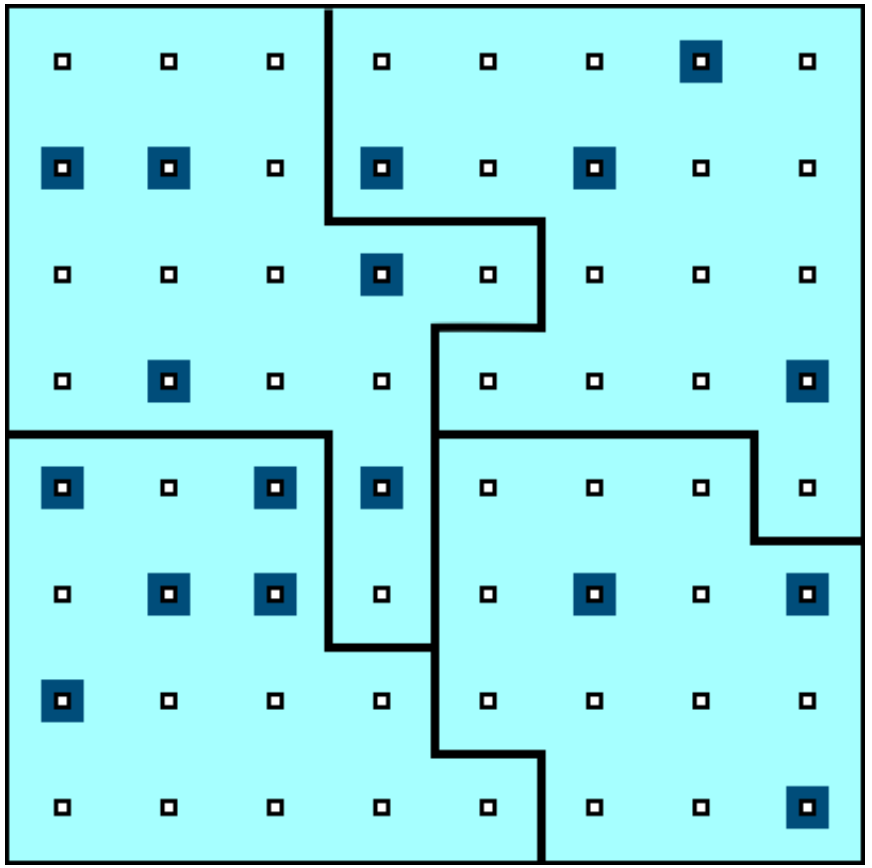
In StS, $n = n_1 + \cdots + n_k$ units are randomly drawn from k **strata**.

Advantages:

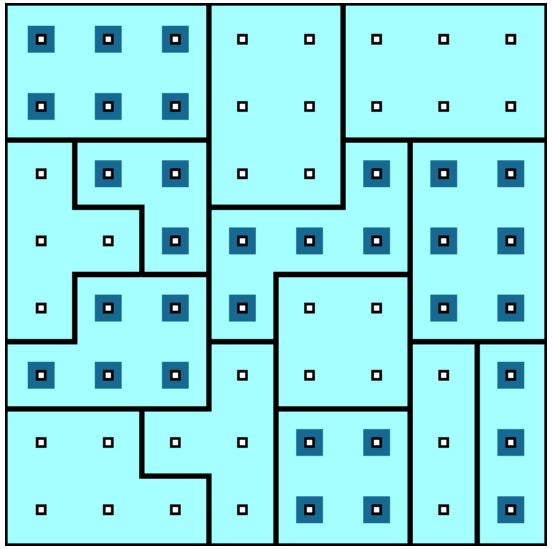
- may produce smaller error bound on estimation than SRS
- may be less expensive if elements are conveniently strat.
- may provide estimates for sub-populations

Disadvantages:

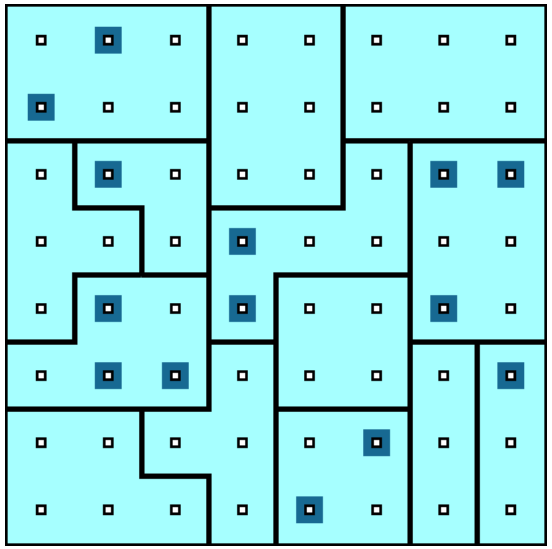
- no major disadvantage
- if there are no natural ways to stratify the frame into homogeneous groupings, StS is roughly equivalent to SRS



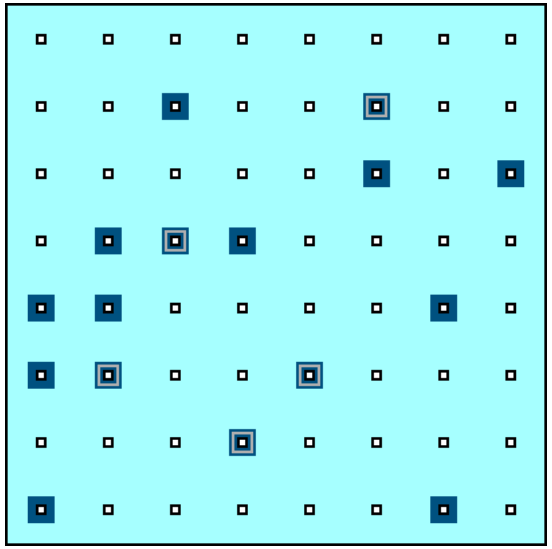
OTHER PROBABILISTIC SAMPLING DESIGNS



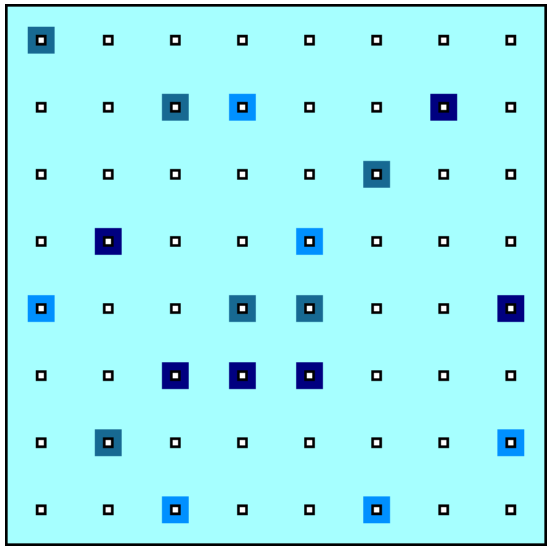
Cluster Sampling (CIS)



Multi-Stage Sampling (MSS)



Multi-Phase Sampling (MPS)



Replicated Sampling (ReS)

DATA COLLECTION

PROBABILISTIC SAMPLING