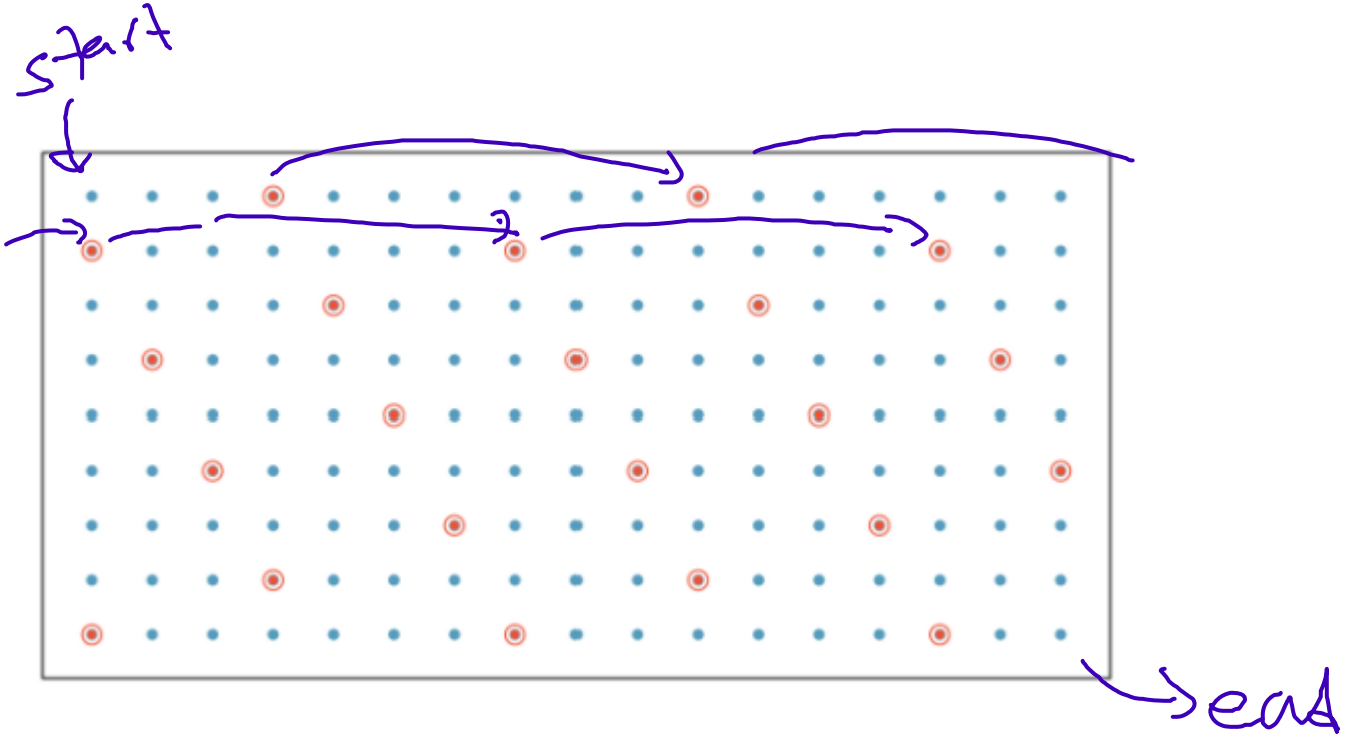
Systematic sampling

Week 6 (5.5)

Stat 260, St. Clair

Design: Systematic Sample

Loosely Defined: Create an ordered list (sample frame) of observation units, randomly select a starting point and add every "k"th unit after.



Design: Systemaic Sample

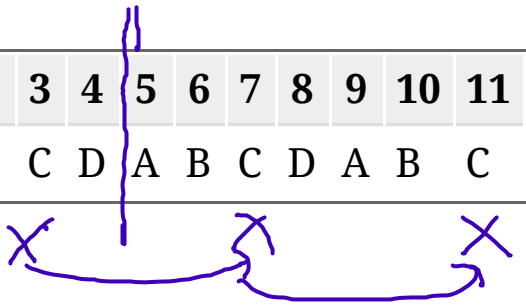
Defined:

- M_0 observation units
- M units is the desired observation unit-level sample size
- Take a SRS of size 1 from the first $N = \frac{M_0}{M}$ units and include every Nth unit after.

Design: Systemaic Sample

$M_0 = 12, M = 3, N = 12/3 = 4$

units	1	2	3	4	5	6	7	8	9	10	11	12
possible samples	A	B	C	D	A	B	C	D	A	B	C	D



Possible samples:

A : {1, 5, 9} B : {2, 6, 10}

C : {3, 7, 11} D : {4, 8, 12}

⇒ clusters of A, B, C, D

SRS n = 1 ⇒ obs. all units in selected cluster

Design: Systemaic Sample

Equivalent to a **one-stage** cluster sample of size $n = 1$ from N available clusters.

- Estimated mean:

$$\hat{\bar{y}}_{sys} = \bar{y}$$ = cluster avg.

- SE: for large ~~n~~ M_0

$$SE(\hat{\bar{y}}_{sys}) = \sqrt{\left(1 - \frac{1}{N}\right)\frac{s_t^2}{M^2}} \approx \sqrt{\frac{s^2}{M}(1 + (M-1)ICC)}$$

s_t from 1-stage cluster

s_t^2 = variance of cluster totals t_i in sample

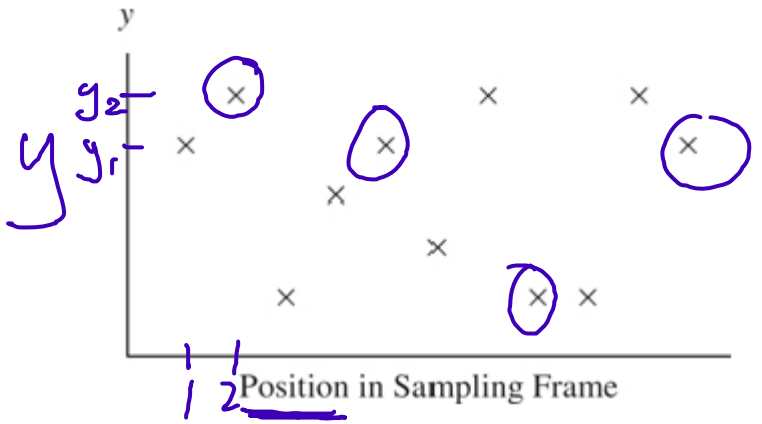
↳ $M \times MSB$

↳ $n = 1$ ⇒ can't compute s_t^2 from data

s^2 = sample variance of data

Design: Systemaic Sample Scenarios

Randomly ordered list with respect to the response y



→ no assoc. btw. cluster & response y

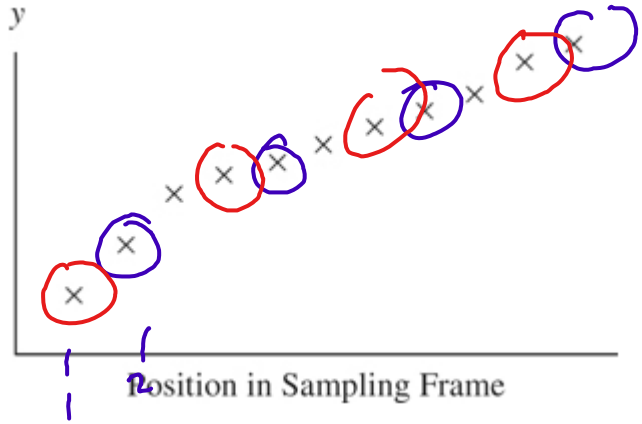
- $SSB \approx 0$ and $ICC \approx 0$ so

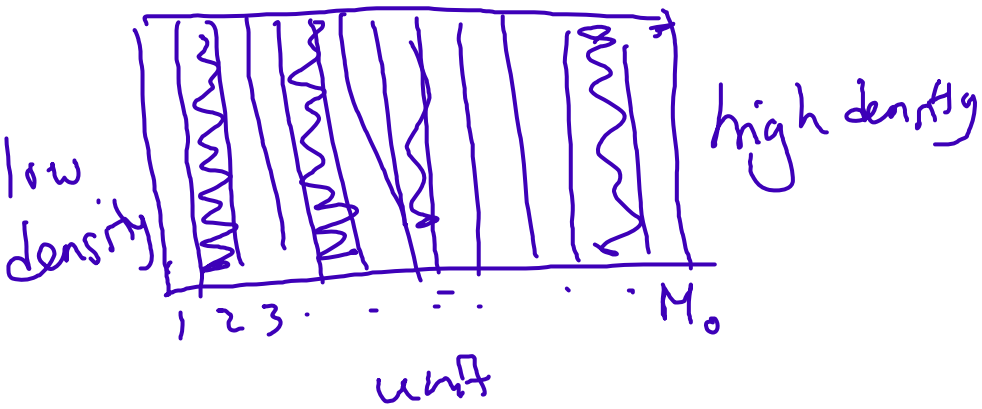
heterogeneous

$$SE(\hat{\bar{y}}_{sys}) \approx \sqrt{\frac{s^2}{M}} \approx SE(\bar{y}_{SRS})$$

Design: Systemaic Sample Scenarios

Trending ordered list with respect to the response y





- $ICC < 0$ so

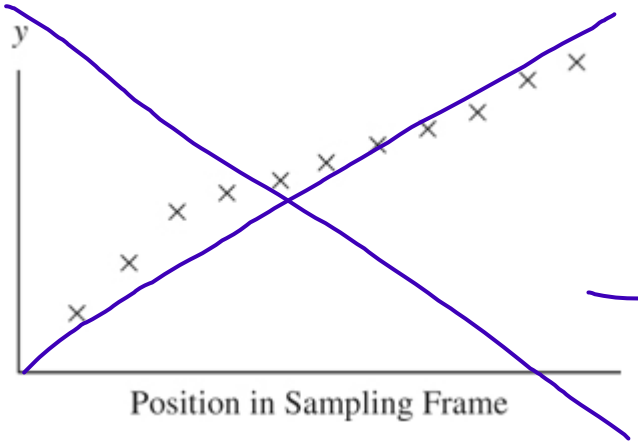
$$SE(\hat{\bar{y}}_{sys}) < \sqrt{\frac{s^2}{M}} = SE(\bar{y}_{SRS})$$

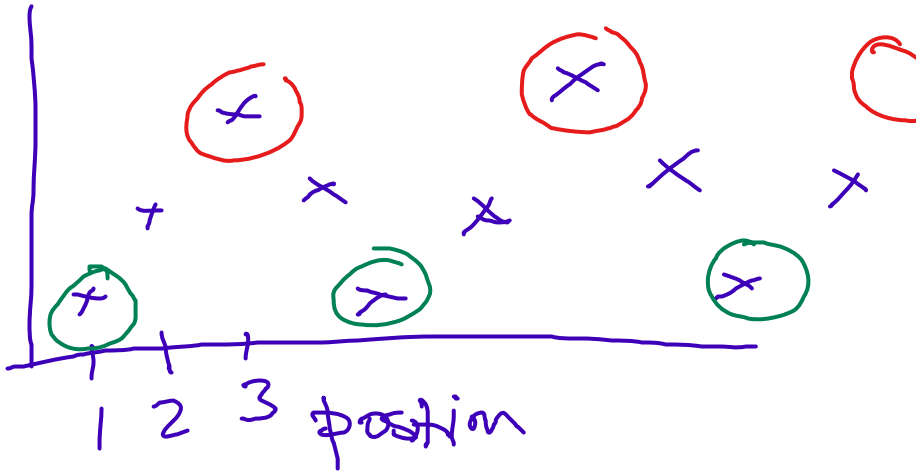
use SRS SE to overestimate SE

all samples see low - high values of y
⇒ artificially heterogeneous

Design: Systemaic Sample Scenarios

Periodic ordered list with respect to the response y





- $SSW \approx 0$ and $ICC \approx 1$ so

$$SE(\hat{\bar{y}}_{sys}) > \sqrt{\frac{s^2}{M}} = SE(\bar{y}_{SRS})$$

possible:
each sample contains ~ same value of y
⇒ homogeneous cluster