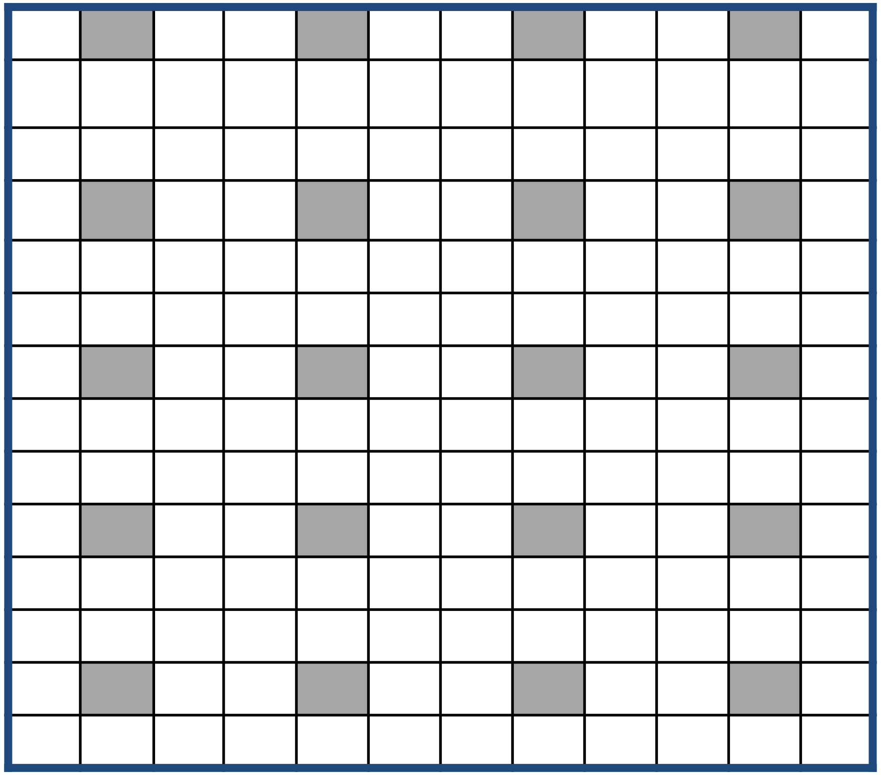
STA 304H1F-1003H Fall 2019

Week 9 - Estimation in Systematic Sampling, Chapter 7

Systematic Sampling (I)

Remainder on basics

Basics: Elements are selected from an ordered sampling frame. First element is selected at random, subsequent elements follow a predetermined pattern, usually an interval.



20 units selected systematically

Example: Supermarket wants to study buying habits of their customers. They can choose every 10th customer entering the supermarket and conduct the study on this sample.

Systematic Sampling (I)

Systematic sample 1 - in - k, with random start (I)

Selection from an ordered list/frame: Select first element/value (y_1) at random out of the first k elements, and then every *k*th element (with step *k*) from the list, y_2, y_3, …, y_n.

The obtained sample is random, but *it is not* a simple random sample: If *k* >1, two consecutive elements cannot appear in the sample; if e_3 is in the sample, e_4 cannot be in the sample.

Notation: *N* – population size,
n – sample size,
k – selection step /interval, $k = N/n$.

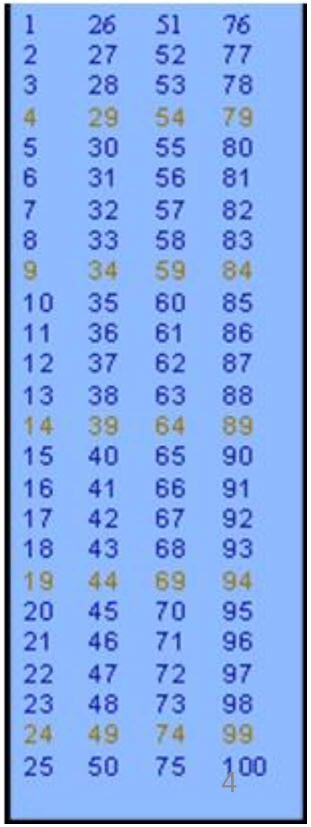
N = 100

want n = 20

N/n = 5

select a random number from 1-5:
chose 4

start with #4 and take every 5th unit



Systematic Sampling (I)

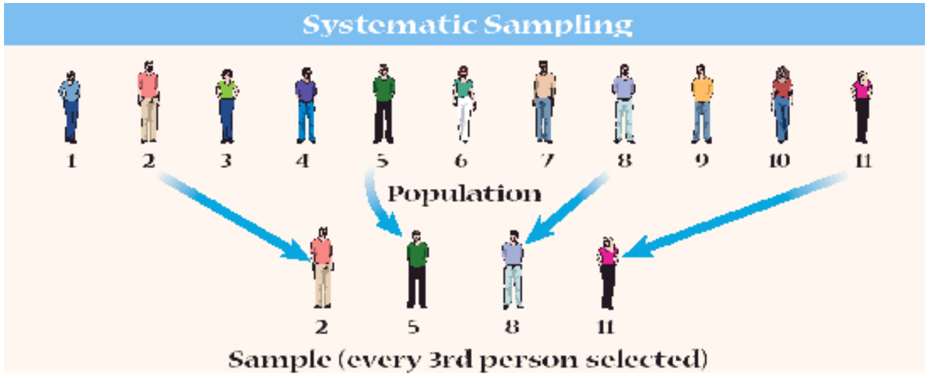
Systematic sample 1 - in - k, with random start (II)

Minor problem: If N = 93, and we want n = 5, k = 93/5 = 18.6.

What to do? If we choose k = 19, we will sometimes finish with 4 elements, not 5 (e_{18} first → e_{37} → e_{56} → e_{75} →e_{94}).

If we choose k = 18, we will sometimes finish with 6 elements, not 5 (e_1 first → e_{19} → e_{37} → e_{55} → e_{73} → e_{91}).

If we round down, we will sometimes finish with n + 1 elements, not n. If n is not very small, this "problem" can be ignored.
We will assume, for simplicity, $\boldsymbol{N = k \times n}$.



Systematic Sampling (II)

Inference: Estimation of mean and total (I) (Ch. 7.3)

Systematic sample: $y_1, y_2, \ldots, y_n$

Sample mean : $\bar{y}_{SYS} = \frac{1}{n}\sum_{i=1}^{n} y_i = \bar{y}$ $\boxed{\hat{\mu}_y = \bar{y}_{SYS} = \bar{y}}$

Main problems: (1) Is $\hat{\mu}_y = \bar{y}_{SYS}$ an unbiased estimator?

(2) How to estimate $Var(\hat{\mu}_y) = Var(\bar{y}_{SYS})$?

Key question: What is the order of elements on the list? How is it related to the variable of interest? The question is relevant here, because selection of the sample is related to the order of elements on the list. This is not the case in SRS.

One possible answer: If *random order of elements* on the list (in comparison with variable *y*) can be assumed, then the systematic sample can be treated just as an SRS.

Systematic Sampling (II)

Inference: Estimation of mean and total (II)

With random order of elements on the list systematic sampling can be considered just as a convenient, simple way to obtain a *simple random sample* from the population. We then can apply all theory from SRS.

Example: (a) Class list ordered by last name (most common). Variable of interest *y* – test mark. Order not related to *y*.
(b) Class list ordered by mark (e.g., from highest). Variable of interest *y* – test mark. Order directly related to *y*.

Example: Farms ordered by size. (a) Variable of interest - # of cattle on farm. Are they related? (obviously they are)

(b) Variable of interest - income per acre. Are they related? Not clear.

Systematic Sampling: Example (I)

Science and Medicine Library: Collection of statistical books (I)

Sampling design: Systematic sample

Sampling size: 20 shelves

Sampling interval (step): k = *N*/*n* = 161/20 = 8

Sample of shelves: Start from the shelf 5 (randomly selected between 1 and 8), and then every 8th : 5, 13, 21, 29, 37, 45, 53, 61, 69, 77, 85, 93, 101, 109, 117, 125, 133, 141, 149, 157

	1	2	3	4	5
1	1	7	13	19	25
2	2	8	14	20	26
3	3	9	15	21	27
4	4	10	16	22	28
5	5	11	17	23	29
6	6	12	18	24	30

	1	2	3	4	5	6
1	31	38	45	52	59	66
2	32	39	46	53	60	67
3	33	40	47	54	61	68
4	34	41	48	55	62	69
5	35	42	49	56	63	70
6	36	43	50	57	64	71
7	37	44	51	58	65	72

	1	2	3	4	5
1	73	79	85	91	97
2	74	80	86	92	98
3	75	81	87	93	99
4	76	82	88	94	100
5	77	83	89	95	101
6	78	84	90	96	102

	1	2	3	4
1	103	109	115	121
2	104	110	116	122
3	105	111	117	123
4	106	112	118	124
5	107	113	119	125
6	108	114	120	126

	1	2	3	4	5
1	127	134	141	148	155
2	128	135	142	149	156
3	129	136	143	150	157
4	130	137	144	151	158
5	131	138	145	152	159
6	132	139	146	153	160
7	133	140	147	154	161

example

Systematic Sampling: Example (I)

Science and Medicine Library: Collection of statistical books (II)

Systematic Sample: Population size *N* = 161. Sample size *n* = 20. Sample:

sh	5	13	21	29	37	45	53	61	69	77	85	93	101	109	117	125	133	141	149	157
y	18	32	23	22	24	28	19	22	10	19	24	26	28	28	38	9	13	13	18	23
x	12	22	8	6	19	19	16	16	6	10	15	16	16	14	20	6	8	5	6	5

$\Sigma y = 437$, $\Sigma y^2 = 110536$, $\bar{y} = 21.85$, $S_y^2 = 53.397$

Estimation: Consider the sample as an SRS **Discussion**

$$\hat{\mu}_y = \bar{y} = 21.85 \quad \hat{\tau}_y = N\bar{y} = 161 \times 21.85 = 3517.85,$$

$$\hat{Var}(\bar{y}) = \frac{N-n}{N}\frac{S_y^2}{n} = \frac{161-20}{161} \times \frac{53.397}{20} = 2.338$$

$$\hat{Var}(\hat{\tau}_y) = N^2 \hat{Var}(\bar{y}) = 161^2 \times 2.338$$

$$\hat{Sd}(\bar{y}) = 1.529 \qquad \hat{Sd}(\hat{\tau}) = 161 \times 1.529 = 246.169$$

CI for $\mu : \bar{y} \pm B_\mu = 21.85 \pm 3.06 = [18.79, 24.91]$,

CI for $\tau : \hat{\tau} \pm B_\tau = 3517.85 \pm 492.34 = [3025.51,\ 4010.19]$

example

Systematic Sampling: Example (II)

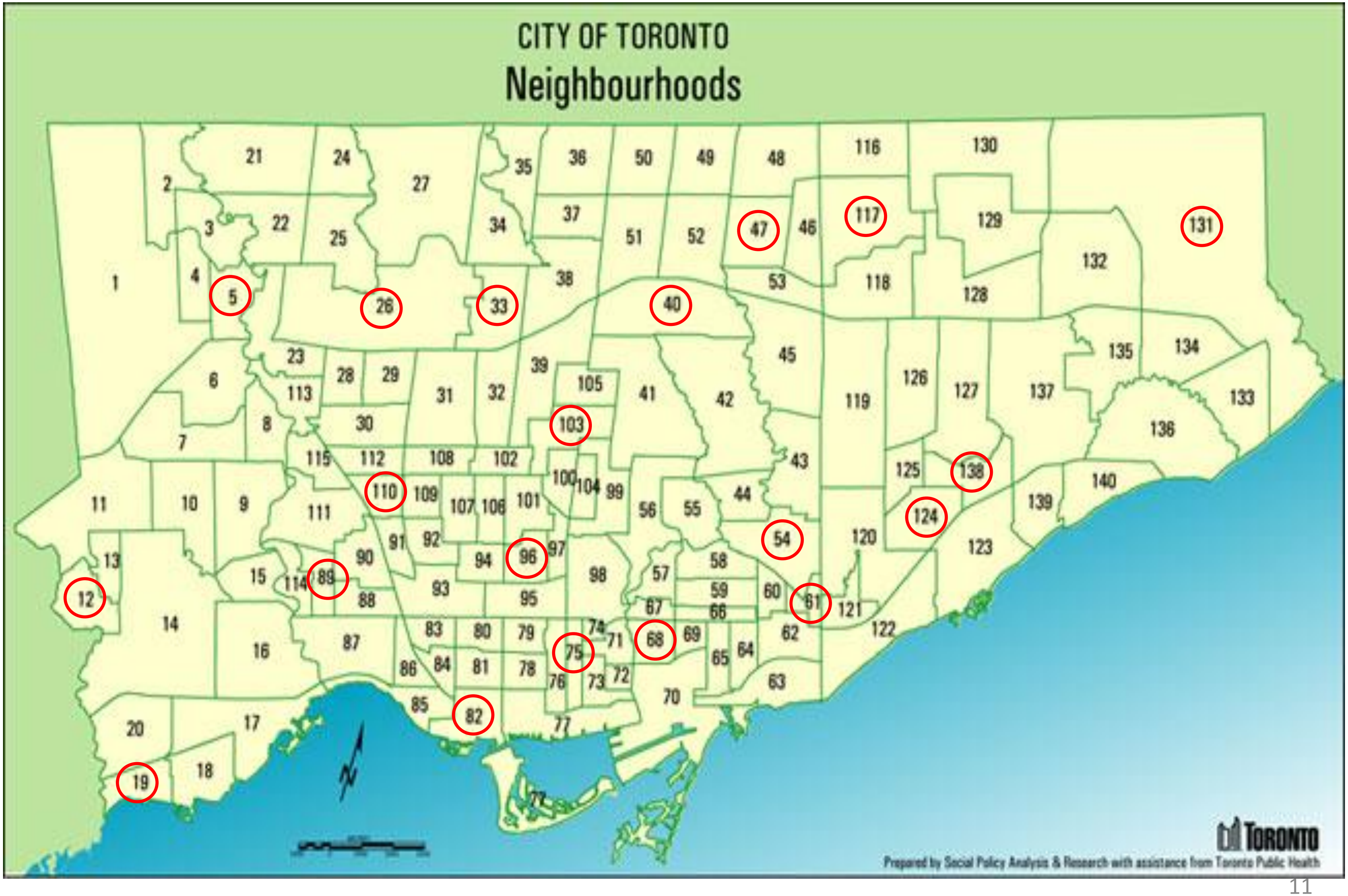
Population of Toronto neighbourhoods (I)

We can select from the list of neighbourhoods we have used for SRS sampling, but this list follows location on the map. Can this be considered as a “random” order?

We can select from the list of neighbourhoods ordered alphabetically. Can this be considered as a “random” order? More “random” than the first list?

Sample I: N = 140. Select n = 20 units from the list by location. Selection interval k = 140/20 = 7. Using TRN, we start at unit 5 (random digit between 1 and 7) and then select every 7th (5, 12, 19, … , 131, 138).

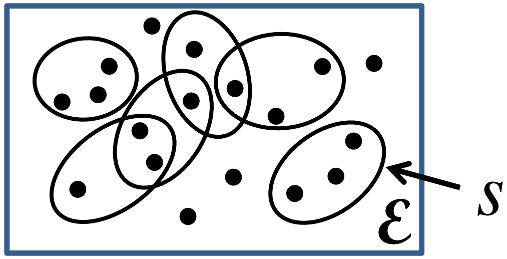
CITY OF TORONTO NEIGHBOURHOODS – Sample I, list by location



Systematic Sampling (II)

General considerations (I)

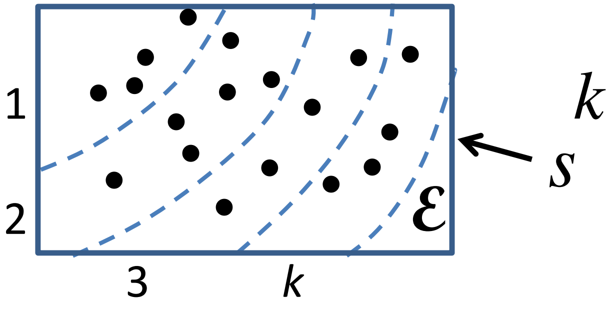
SRS # of samples of size *n*



$$\binom{N}{n}$$

$N = 50, n = 10, \binom{50}{10} \approx 12$ bill.

Systematic # of samples of size *n*



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

$N = 50, n = 10, \frac{50}{10} = 5$

In general, without additional assumptions about the population, $Var(\bar{y})$ cannot be estimated properly from one systematic sample.

Assumptions:

1) **Random order** (relative to variable of interest) – same as SRS

$$\hat{Var}(\bar{y}_{SYS}) = \frac{N-n}{N}\frac{S_y^2}{n}, \; S_y^2 = \frac{1}{n-1}\sum(y_i - \bar{y})^2$$

Systematic Sampling (II)

General considerations (II)

Assumptions:

2) **Ordered population:** Order of elements on the list correlated with variable of interest – mostly *monotonic*.

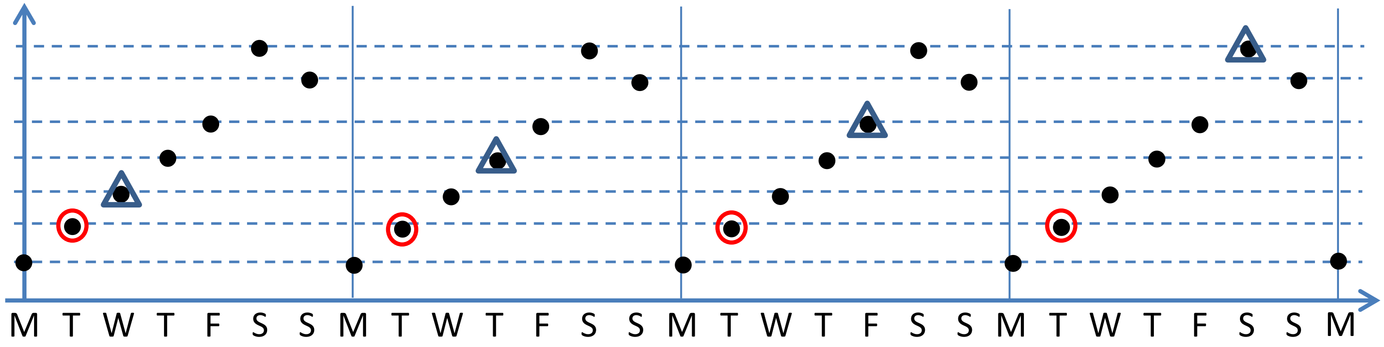
Simple example: Class list ordered by first test results, variable of interest – second test results.

3) **Periodic population:** Such as a list of seasonal variations in daily temperatures, sales, etc., - cases mostly related to time series.

Two cases: Short period, long period

- Don't use interval *k* = multiple of the period (why?)

- Good for sampling if k is small compared with the period.

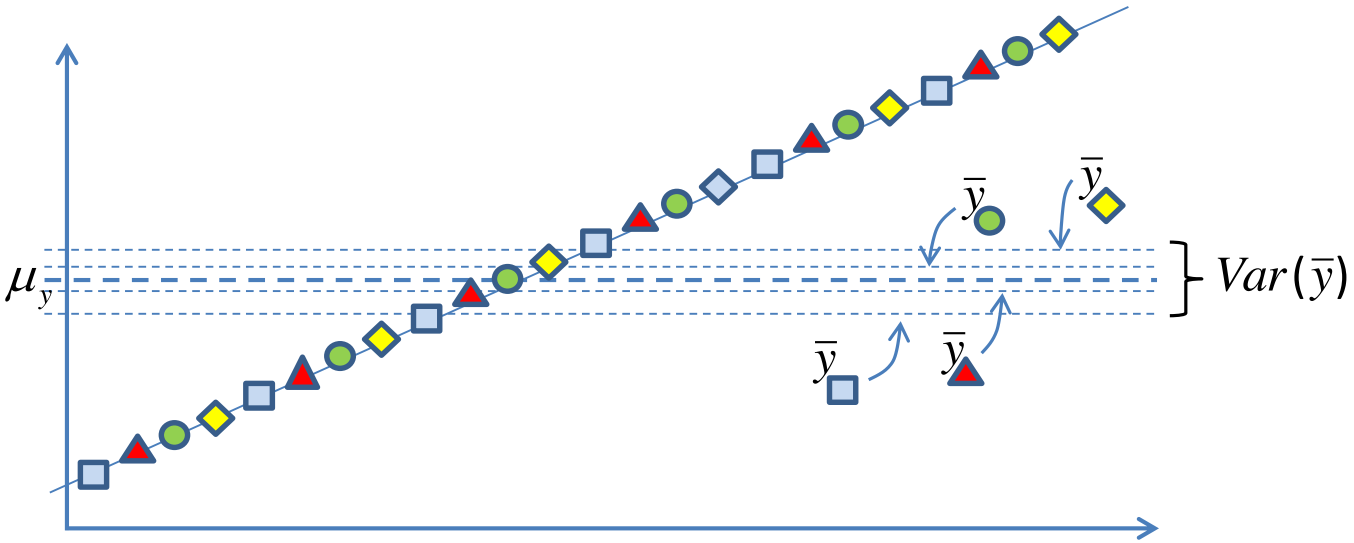


Weekly sales: What *k* should not be, and what can be?

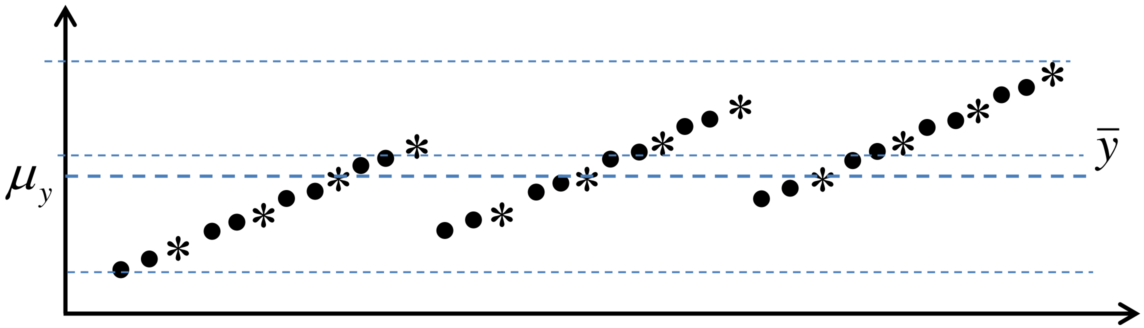
Systematic Sampling (II)

General considerations (III)

Monotonic population: Increasing or decreasing



$Var(\bar{y})$ is about variation of sample means $\bar{y}$, not population values



Ordered population with long period is similar to monotonic population

Systematic Sampling (II)

General considerations (IV)

Assumptions : $N = k \times n$, sample size n, k – selection step, number of possible samples - k.

Example : $N = 12$, sample size 4, $k = \dfrac{12}{4} = 3$ – selection step, number of possible samples $= 3$.

population	1	2	3	4	5	6	7	8	9	10	11	12
variable	2	5	0	6	6	0	1	2	9	8	7	3
samples	•	◊	+	•	◊	+	•	◊	+	•	◊	+

sample	1	2	3	4	$\bar{y}_i$
1	2	6	1	8	4.25
2	5	6	2	7	5.00
3	0	0	9	3	3.00

Start by listing first elements, then second elements, then third elements, …

Systematic Sampling (II)

General considerations (V)

Three possible results of sampling:

$\bar{y}$	$\bar{y}_1$	$\bar{y}_2$	$\bar{y}_3$
value	4.25	5.00	3.00
probability	$1/3$	$1/3$	$1/3$

$$E(\bar{y}) = \frac{1}{3}\times 3 + \frac{1}{3}\times 4.25 + \frac{1}{3}\times 5 = 4.083$$

$$Var(\bar{y}) = \frac{1}{3}\sum_{i=1}^{3}(\bar{y}_i - 4.083)^2 = 0.6833$$

General case:

$$\bar{y}_i = \frac{1}{n}\sum_{j=1}^{n} y_{ij} \qquad \bar{y}_{SYS} : \bar{y}_1, \bar{y}_2, \dots, \bar{y}_k \qquad P(\bar{y}_{SYS} = \bar{y}_i) = \frac{1}{k}, i = 1, 2, \dots, k$$

$$E(\bar{y}_{SYS}) = \sum \bar{y}_i P(\bar{y}_{SYS} = \bar{y}_i) = \frac{1}{k}\sum_{i=1}^{k}\bar{y}_i = \frac{1}{kn}\sum_{i=1}^{k}\sum_{j=1}^{n} y_{ij} = \frac{1}{N}\sum_{i} y_i = \mu$$

$\hat{\mu} = \bar{y}_{SYS}$ - an unbiased estimator of μ

$$\boxed{Var(\bar{y}_{SYS}) = \frac{1}{k}\sum_{i=1}^{k}(\bar{y}_i - \mu)^2}$$

Theoretical variance of $\bar{y}_{SYS}$

$$Var(\bar{y}_{SRS}) = \frac{N-n}{N-1}\frac{\sigma^2}{n}$$

Systematic Sampling (II)

General considerations (V)

Where is the problem?

1) From sampling we obtain one single value $\bar{y}_i$, and we cannot estimate the variance of k values $\bar{y}_1, \bar{y}_2, \ldots, \bar{y}_k$ without any additional information about the population.

2) If we have some additional information about the population, such as in random order, or monotonic, we have to find a proper method to estomate $Var(\bar{y}_{SYS})$, such as from SRS.

The question is, to what extend information from one sample gives us information about the population. It could be quite good (monotonic population), or bad (periodic, with sample coinciding with period).

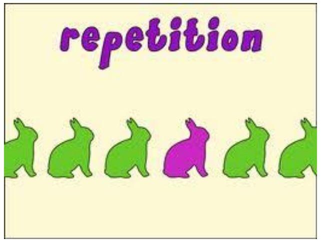
Systematic Sampling (III)

Repeated systematic sampling (I) (Ch. 7.6.)

To avoid problem with additional information entirely, we can use *repeated systematic sampling*, to obtain more information about "population" $\bar{y}_1, \bar{y}_2, \ldots, \bar{y}_k$ needed to estimate $Var(\bar{y}_{SYS})$.

Repeated systematic sampling: Selecting several systematic samples (without repetition), instead of just one.

Toronto neighbourhoods: We can repeat systematic sampling of 1-in-7 units (*n* = 20) three times and get 60 units in the sample. Alternatively, we can repeat systematic sampling of 1-in-14 units (*n* = 10) six times, and also get 60 units in the sample. Which one to use? Or, we should use SRS (if possible), with *n* = 60?



Systematic Sampling (III)

Repeated systematic sampling (II)

Numerical example : $N = 1{,}000$, total sample size allowed $n = 100$

1) One systematic sample, $n = 100$, step $k = 1000 / 100 = 10$.

2) Repeated systematic sampling : $n_s = 5$ (repeat 5 times)

with $n' = 20 (= n/n_s = 100/5)$ elements in each.

Step $k' = 1000/20 = 50$.

3) Repeated systematic sampling : $n_s = 10$ (repeat 10 times)

with $n' = 10 (= n/n_s = 100/10)$ elements in each.

Step $k' = 1000/10 = 100$.

Notation : N - population size, n_s - # of repeated systematic samples, n' - sample size of one systematic sample, $k' = N/n'$ - selection step for one systematic sample, total sample size $n = n_s \times n'$.

Systematic Sampling (III)

Repeated systematic sampling (III)

$$\boxed{\begin{array}{l} k'\times n' = N \\ n_s \times n' = n \end{array}}$$

In practice : 1) $N, n', n_s \Rightarrow k' = N/n', n = n_s \times n'$.

2) $N, n, n_s \Rightarrow n' = n/n_s, k' = N/n'$.

Sample : n_s systematic samples $\Rightarrow \bar{y}_1, \bar{y}_2, \ldots, \bar{y}_{n_s}$ out of possible k'

(an SRS of size n_s from the population $\{\bar{y}_1, \bar{y}_2, \ldots, \bar{y}_{k'}\}$)

Estimation : $\boxed{\hat{\mu} = \bar{y}_{SYS,REP} = \frac{1}{n_s}\sum_{i=1}^{n_s}\bar{y}_i = \frac{1}{n}\sum_{i=1}^{n} y_i = \bar{y}_{SM}}$

unbiased

$$\boxed{\hat{Var}(\bar{y}_{SYS,REP}) = \frac{k'-n_s}{k'}\frac{S^2_{REP}}{n_s} = \frac{N-n}{N}\frac{S^2_{REP}}{n_s}, S^2_{REP} = \frac{1}{n_s-1}\sum_{i=1}^{n_s}(\bar{y}_i - \hat{\mu})^2}$$

$$Var(\bar{y}_{SYS,REP}) = \frac{k'-n_s}{k'-1}\frac{\sigma^2_{REP}}{n_s} = \frac{N-n}{N-n'}\frac{\sigma^2_{REP}}{n_s}, \sigma^2_{REP} = \frac{1}{k'}\sum_{i=1}^{k'}(\bar{y}_i - \mu)^2$$

theoretical variance

Note: $\hat{\sigma}^2_{REP} = \hat{\sigma}^2_{syst,n'} = \frac{k'-1}{k'}S^2_{REP} = \frac{N-n'}{N}S^2_{REP}$

Systematic Sampling (III)

Repeated systematic sampling, example (I)

EXAMPLE: Systematic Sampling and Repeated Systematic Sampling I

Goal: Estimate the average test mark

Population: Class of 40 students

Variable: Test mark (*y*)

Parameters to be estimated: Average test mark (μ_y)

Sampling design: Systematic sample of size 5, and then 4 times repeated systematic sample of size 5.

Method of estimation: Sample mean.

Population ordered by marks:

20	1	1
30	8	1
40	2	1
50	0 0 1 2 3 4 5	7
60	0 0 0 0 5 6 6 7 7 8 8 9 9 9	14
70	0 2 3 3 4 7 9	7
80	0 0 0 1 3 4 5 8	8
90	8	1
	total	40

Population: 21, 38, 42, 50, 50, 51, …, 85, 88, 98

$$\mu_y = 66.425, \sigma_y^2 = 218.1444, \sigma_y = 14.7697$$

For an SRS of size 5, in theory :

$$Var(\bar{y} / SRS) = \frac{40-5}{40-1} \frac{218.14}{5} = 39.15$$

$$Sd(\bar{y} / SRS) = 6.257$$

Systematic Sampling (III)

Repeated systematic sampling, example (II)

List of all possible 8 systematic samples of size 5 from ordered population:

i\j	1 2 3 4 5	$\bar{y}_i$	σ_i^2
1	21 54 66 70 80	58.2	415.35
2	38 55 67 72 80	62.4	214.64
3	42 60 67 73 81	64.6	175.44
4	50 60 68 73 83	66.8	126.16
5	50 60 68 74 84	67.2	135.36
6	51 60 69 77 85	68.4	144.64
7	52 65 69 79 88	70.6	150.64
8	53 66 69 80 98	73.2	227.76

Theoretical calculation

$$\mu_y = \tfrac{1}{8}\sum \bar{y}_i = 66.424$$

$$Var(\bar{y}\,/\,syst) = \tfrac{1}{8}\sum(\bar{y}_i - \mu_y)^2 = 19.40$$

$$= \sigma_y^2 - \tfrac{1}{8}\sum \sigma_i^2 = 218.14 - \tfrac{1}{8}(415.35 + 214.64 + \cdots + 227.76)$$

$$Sd(\bar{y}\,/\,syst) = 4.40$$

$$\rho = \frac{1}{n-1}\left(\frac{nVar}{\sigma^2} - 1\right) = \frac{1}{5-1}\left(\frac{5 \times 19.40}{218.14} - 1\right) = -0.1389$$

Negative intracluster correlation

(for ρ, postpone until you see slide 31)

Systematic Sampling (III)

Repeated systematic sampling, example (III)

Repeated systematic samples of size *n*' = 5, with n_s = 4 repetitions from k' = 8 samples (total sample size *n* = 5x4 = 20), ordered population.
Using table of random numbers, 4 out of 8 one digit numbers are selected: 1, 3, 6, 7. Repeated systematic samples:

i\j	1 2 3 4 5	$\bar{y}_i$	σ_i^2
1	21 54 66 70 80	58.2	415.35
3	42 60 67 73 81	64.6	175.44
6	51 60 69 77 85	68.4	144.64
7	52 65 69 79 88	70.6	150.64

Sample calculation

$$\bar{y}_{SYS,R} = 65.45, \hat{\mu}_y = 65.45$$

$$S'^2 = \frac{1}{3}\sum_{i=1}^{4}(\bar{y}_i - 65.45)^2 = 29.5033$$

$$\hat{Var}(\bar{y}_{SYS,R}) = \frac{40-20}{40}\times\frac{29.5033}{4} = 3.688$$

$$\hat{Sd}(\bar{y}_{SYS,R}) = 1.9204$$

$$Var(\bar{y}_{SYS,R}) = \frac{k'-n_s}{k'-1}\times\frac{Var(\bar{y}_{SYS,n'})}{n_s}$$

$$= \frac{8-4}{8-1}\times\frac{19.3944}{4} = 2.77063$$

Theoretical

$$Var(\bar{y}_{SRS}) = \frac{N-n}{N-1}\times\frac{\sigma^2}{n}$$

$$= \frac{40-20}{40-1}\times\frac{218.144}{20} = 5.593$$

$$Sd(\bar{y}_{SRS}) = 2.365 \qquad Sd(\bar{y}_{SYS,R}) = 1.665$$

Systematic Sampling (III)

Repeated systematic sampling, example (IV)

Population (class) ordered alphabetically: Marks (*y*)

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
51	68	77	70	70	60	83	69	81	65	38	72	60	66	74	80	54	52	79	42
21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40
86	68	73	85	53	69	67	60	50	66	69	80	55	80	67	73	21	84	98	50

List of all 8 systematic samples of size 5 from alphabetically ordered population:

i\j	1	2	3	4	5	$\bar{y}_i$	σ_i^2
1	51	81	54	53	55	58.8	124.96
2	68	65	52	69	80	66.8	80.56
3	77	38	79	67	67	65.6	215.04
4	70	72	42	60	73	63.4	135.84
5	70	60	86	50	21	55.4	438.24
6	60	66	68	66	84	68.8	64.96
7	83	74	73	69	98	79.4	107.44
8	69	80	85	80	50	72.8	157.36

Theoretical calculation

$$\mu_y = \tfrac{1}{8}\sum \bar{y}_i = 66.424$$

$$Var(\bar{y}/syst) = \tfrac{1}{8}\sum(\bar{y}_i - \mu_y)^2 = 50.534$$

$$Sd(\bar{y}/syst) = 7.109$$

$$\rho = \frac{1}{n-1}\left(\frac{nVar}{\sigma^2} - 1\right) = \frac{1}{5-1}\left(\frac{5\times 50.534}{218.144} - 1\right)$$

$$= \boxed{0.03957}$$

Week intracluster correlation

$$Var(\bar{y}/SRS) = 39.15$$

$$Sd(\bar{y}/SRS) = 6.257$$

(for ρ, postpone until see slide 31)

Systematic Sampling (III)

Repeated systematic sampling, example (V)

Repeated systematic samples of size *n*' = 5, with n_s = 4 repetitions from *k*' = 8 samples (total sample size *n* = 5x4 = 20), “random” population.

Using same random digits: 1, 3, 6, 7. Repeated systematic sample:

i\j	1	2	3	4	5	$\bar{y}_i$	σ_i^2
1	51	81	54	53	55	58.8	124.96
3	77	38	79	67	67	65.6	215.04
6	60	66	68	66	84	68.8	64.96
7	83	74	73	69	98	79.4	107.44

Sample calculation

$$\bar{y}_{SYS,R} = 68.15, \hat{\mu}_y = 68.15$$

$$S'^2 = \frac{1}{3}\sum_{i=1}^{4}(\bar{y}_i - 68.15)^2 = 73.637$$

$$\hat{Var}(\bar{y}_{SYS,R}) = \frac{40-20}{40}\times\frac{73.637}{4} = 9.205$$

$$\hat{Sd}(\bar{y}_{SYS,R}) = 3.034$$

$$Var(\bar{y}_{SYS,R}) = \frac{8-4}{8-1}\times\frac{50.534}{4} = 7.219$$

Theoretical

$$Var(\bar{y}_{SRS}) = \frac{40-20}{40-1}\times\frac{\sigma^2}{20} = 5.593$$

$$Sd(\bar{y}_{SRS}) = 2.365 \qquad Sd(\bar{y}_{SYS,R}) = 2.687$$

Conclusion: Population ordered by mark values shows better results in estimation than the population ordered alphabetically (close to random order). Compare intracluster correlation coefficients.

Systematic Sampling (IV)

Another formula for variance in systematic sampling (I)

Consider k systematic samples 1-in-k of size n (N=n x k) as k strata of size n:

1	1 2 3 … n
2	1 2 3 … n
3	1 2 3 … n
.	. . .
k	1 2 3 … n

$$L = k, N_i = n, W_i = \frac{1}{k}, \mu_i = \bar{y}_i, \mu = \sum_{i=1}^{k} W_i \mu_i = \frac{1}{k}\sum_{i=1}^{k} \bar{y}_i$$

$$\sigma^2 = \sum_{i=1}^{L} W_i \sigma_i^2 + \sum_{i=1}^{L} W_i(\mu_i - \mu)^2 = \frac{1}{k}\sum_{i=1}^{k} \sigma_i^2 + \frac{1}{k}\sum_{i=1}^{k} (\bar{y}_i - \mu)^2$$

$$\boxed{Var(\bar{y}_{SYS}) = \sigma^2 - \frac{1}{k}\sum_{i=1}^{k} \sigma_i^2 = \sigma^2 - \overline{\sigma^2}, \quad \overline{\sigma^2} = \frac{1}{k}\sum_{i=1}^{k} \sigma_i^2}$$

Discussion : 1) $\max_{\text{syst}} Var(\bar{y}_{SYS}) = \sigma^2$, if $\overline{\sigma^2} = 0$, or all $\sigma_i^2 = 0$.

Bad

All systematic samples are homogeneous, all $y_{ij} = \bar{y}_i$!

2) $\min_{\text{syst}} Var(\bar{y}_{SYS}) = 0$, if $\overline{\sigma^2} = \sigma^2$, or all $\sigma_i^2 = \sigma^2$.

Good

Examples here!

All systematic samples are as heterogeneous as the population !

Systematic Sampling (IV)

Difference method for variance estimation (I)

All this theory is nice, but how to estimate $Var(\bar{y}_{SYS})$ from one systematic sample?

Difference method is useful for ordered population :

Sample : $y_1,\ y_2,\ y_3,\ \ldots, y_{n-1},\ y_n$

Differences : $y_2 - y_1, y_3 - y_2, \ldots, y_n - y_{n-1}$

$$d_i = y_{i+1} - y_i, i = 1,2,\ldots,n-1$$

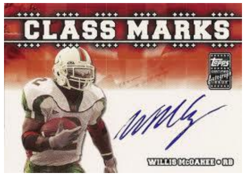
$$\boxed{D = S_d^2 = \frac{1}{2(n-1)} \sum_{i=1}^{n-1} d_i^2}$$

$$\boxed{\hat{Var}(\bar{y}_{SYS}) = \hat{V}_d(\bar{y}_{SYS}) = \frac{N-n}{N} \frac{S_d^2}{n}}$$

Not very precise

Can be used in SRS sampling, but less efficient than the regular estimator.

Prove in SRS : $E(S_d^2) = \sigma^2$



Systematic Sampling (IV)

Difference method for variance estimation (II)

List of all possible 8 systematic samples of size 5 from ordered population (class) with the variance estimated using difference method and SRS method, for comparison.

i\j	1 2 3 4 5	$\hat{V}_d(\bar{y})$	$\hat{V}_{SRS}(\bar{y})$
1	21 54 66 70 80	29.51	90.86
2	38 55 67 72 80	11.42	46.95
3	42 60 67 73 81	10.35	38.38
4	50 60 68 73 83	6.32	27.60
5	50 60 68 74 84	6.56	29.61
6	51 60 69 77 85	6.34	31.64
7	52 65 69 79 88	8.01	32.95
8	53 66 69 80 98	13.63	49.82
	Average $\hat{V}$	11.52	43.48

Difference method, sample 1:

$$S_d^2 = \frac{1}{2(5-1)}\sum_{i=1}^{5-1} d_i^2 = \frac{1}{8}((54-21)^2 + (66-54)^2$$

$$+(70-66)^2 + (80-70)^2) = 168.625$$

$$\hat{V}_d(\bar{y}) = \frac{40-5}{40}\frac{S_d^2}{5} = \frac{7}{40}\times 168.625 = 29.51$$

SRS method, sample 1:

$$S_{SRS}^2 = \frac{1}{5-1}\sum_{i=1}^{5}(y_i - \bar{y})^2 = 519.2$$

$$\hat{V}_{SRS}(\bar{y}) = \frac{40-5}{40}\frac{S_{SRS}^2}{5} = \frac{7}{40}\times 519.2 = 90.86$$

$$Var(\bar{y}\,/\,syst) = \tfrac{1}{8}\sum(\bar{y}_i - \mu_y)^2 = 19.40$$

$\hat{V}_d(\bar{y})$ underestimates and $\hat{V}_{SRS}(\bar{y})$ overestimates the true variance $Var(\bar{y}_{SYS})$

Systematic Sampling (V)

Intracluster correlation coefficient (I)

We know that homogeneity of systematic samples greatly affects the variance : $Var(\bar{y}_{SYS}) = \sigma^2 - \frac{1}{k}\sum_{i=1}^{k}\sigma_i^2$. A useful measure of homogeneity is the "*intracluster correlation coefficient*":

$\rho = \rho_c = Corr(y', y'')$ - where y', y'' are two different elements selected at random from a randomly selected systematis sample.

Usefulness of this definition is in the convenient formula :

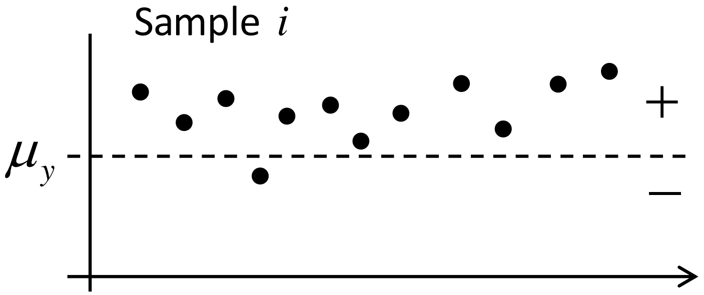
$$\boxed{Var(\bar{y}_{SYS}) = \frac{\sigma^2}{n}[1 + (n-1)\rho]}$$

This formula will help us to discuss accuracy of systematic sampling, depending on the population type.

Systematic Sampling (V)

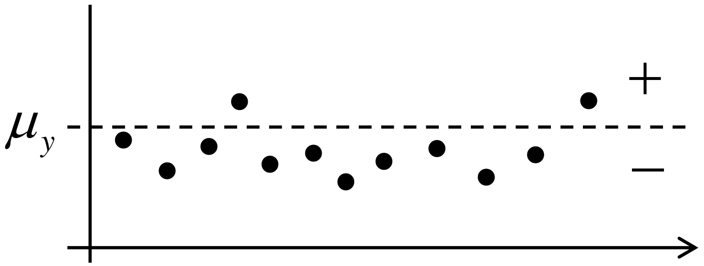
Intracluster correlation coefficient, justification (IV)

Consider $\rho_i = \frac{1}{\sigma^2} Cov(y_i', y_i'') = \frac{1}{\sigma^2 n(n-1)} \sum_{j \neq l}^{n} (y_{ij} - \mu)(y_{il} - \mu)$



$$\underset{+}{(y_{ij} - \mu)}\underset{+}{(y_{il} - \mu)} > 0$$

$$\rho_i > 0$$

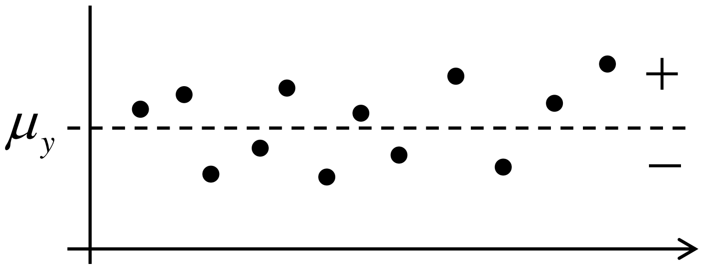


$$\underset{-}{(y_{ij} - \mu)}\underset{-}{(y_{il} - \mu)} > 0$$

$$\rho_i > 0$$

If most of systematic samples tend to be above μ, or below μ, then

$$\rho_c = \frac{1}{k} \sum_{i=1}^{k} \rho_i > 0$$



$$\underset{+(-)}{(y_{ij} - \mu)}\underset{-(+)}{(y_{il} - \mu)} \gtrless 0$$

$$\rho_i \leq 0$$

If most of systematic samples tend to be around μ, then

$$\rho_c = \frac{1}{k} \sum_{i=1}^{k} \rho_i \leq 0$$

Systematic Sampling (V)

Intracluster correlation coefficient (V)

$$\boxed{Var(\bar{y}_{SYS}) = \frac{\sigma^2}{n}[1+(n-1)\rho]} \quad \Rightarrow 1+(n-1)\rho \geq 0 \Rightarrow -\frac{1}{n-1} \leq \rho \leq 1$$

$$Var(\bar{y}_{SYS}) - Var(\bar{y}_{SRS}) = \frac{n-1}{n}\sigma^2\left[\rho + \frac{1}{N-1}\right]$$

$$\boxed{Var(\bar{y}_{SYS}) - Var(\bar{y}_{SRS}) \begin{cases} \leq 0, \quad \text{if } -\dfrac{1}{n-1} \leq \rho \leq -\dfrac{1}{N-1} \text{(negative correlation)} \\ > 0, \text{if } -\dfrac{1}{N-1} < \rho \leq 1 \text{(small or positive correlation)} \end{cases}}$$

If $$-\frac{1}{N-1} \approx 0 \Rightarrow \begin{matrix} Var(\bar{y}_{SYS}) < Var(\bar{y}_{SRS}) \text{ if } \rho < 0 \\ Var(\bar{y}_{SYS}) \geq Var(\bar{y}_{SRS}) \text{ if } \rho \geq 0 \end{matrix}$$

Compare with sl. 28

Systematic Sampling (V)

Intracluster correlation coefficient (VI)

See slides 22, 24 for examples and comparison

$n = 5, \sigma^2 = 218.14$; for ordered population

$$Var(\bar{y}_{SYS}) = 19.40,\ -0.25 \le \rho \le 1, \rho = -0.1389$$

For "random" order

$$Var(\bar{y}_{SYS}) = 50.534,\ -0.25 \le \rho \le 1, \rho = 0.03957$$

Estimation of ρ is not simple in systematic sampling. If you can estimate $Var(\bar{y}_{SYS})$, such as using difference method, or repeated systematic sampling, and σ^2, then

$$\hat{\rho} = \frac{1}{n-1}\left(\frac{n\hat{V}ar(\bar{y}_{SYS})}{\hat{\sigma}^2} - 1\right)$$

We will say more about ρ in cluster sampling

optional

Systematic Sampling (V)

Intracluster correlation coefficient (ICC), an exercise

Let the population consists of elements 1, 2, 3, …, *N*, and the variable $Y(i) = \mathrm{a}i + b$.

(a) Find μ_y and σ_y^2.

(b) Let $N = k \times n$.Find $\bar{y}_i$ and σ_i^2 in $1-$in$-k$ systematic sampling.

(c) Prove that $Var(\bar{y}_{SYS}) = a^2 \frac{k^2-1}{12}$.

(d) Using σ_y^2 from (a) and $Var(\bar{y}_{SYS})$ from (c), find ICC ρ.

(e) Find ρ when $N \to \infty$ and n is fixed.

You may try

Systematic Sampling (VI)

Example: bigger class, 144 test marks, ordered by last name

1	2	3	4	5	6	7	8	9	10
65	47	75	76	63	72	57	62	44	66
60	61	78	62	74	76	64	72	36	59
57	77	54	75	51	64	82	71	46	49
62	79	83	54	87	54	60	45	83	72
85	66	63	58	63	72	54	60	60	57
40	71	61	79	63	52	97	62	75	72
55	68	83	60	67	54	76	72	72	67
69	54	56	56	68	70	74	75	64	66
59	65	45	72	49	86	67	59	77	50
93	70	51	83	82	79	78	53	70	
77	54	64	76	56	49	67	61	70	
69	60	38	70	67	75	65	66	75	
61	71	54	69	52	69	78	37	90	
67	83	80	78	70	67	51	58	76	
44	61	35	64	71	80	66	57	63	

Ordered by columns, 1, 2, … ,10.

You may try

Systematic Sampling (VI)

Example: bigger class, 144 test marks, ordered by mark

1	2	3	4	5	6	7	8	9	10
35	51	56	60	63	67	70	72	77	83
36	51	56	60	63	67	70	74	77	83
37	51	56	60	64	67	70	74	78	83
38	52	57	61	64	67	70	75	78	85
40	52	57	61	64	67	71	75	78	86
44	53	57	61	64	67	71	75	78	87
44	54	57	61	64	67	71	75	79	90
45	54	58	61	65	68	71	75	79	93
45	54	58	62	65	68	72	75	79	97
46	54	59	62	65	69	72	76	80	
47	54	59	62	66	69	72	76	80	
49	54	59	62	66	69	72	76	82	
49	54	60	63	66	69	72	76	82	
49	54	60	63	66	70	72	76	83	
50	55	60	63	66	70	72	77	83	

Ordered by columns