Systematic Sampling

SurvMeth/Surv 625: Applied Sampling

Yajuan Si

University of Michigan, Ann Arbor

2/19/25

Systematic sampling

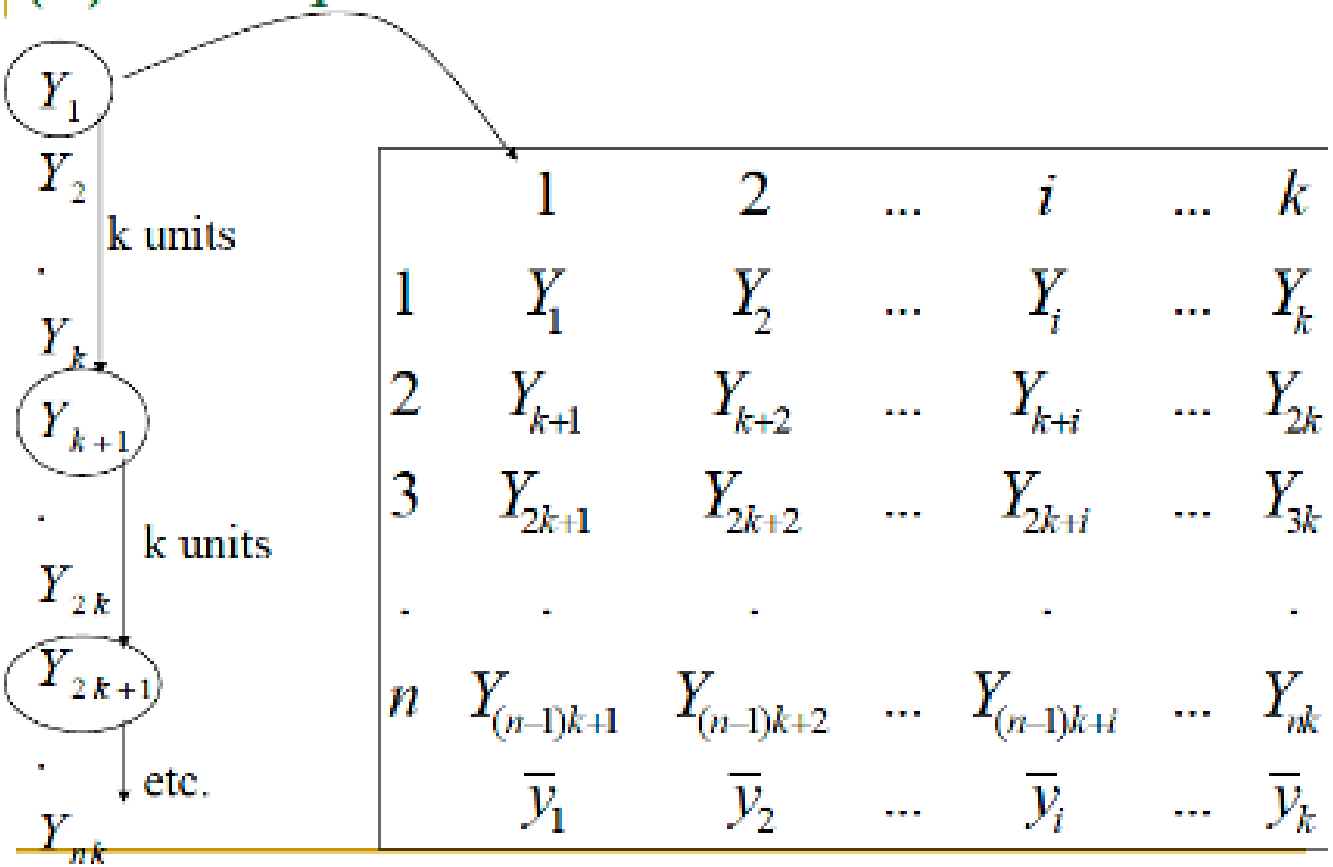
- SRS is difficult to implement and check
- Systematic sampling is clerically easy to do and sometimes used as a proxy for SRS
  - Sorting a list based on auxiliary information on the sampling frame will result in an implicitly stratified list, and group similar elements together in the list
  - There are problems in implementation
  - Systematic sampling does not necessarily give a representative sample, though, if the listing is ordered
- Systematic sampling is really a special case of cluster sampling with implict stratification
- Consider systematic sampling from a numbered list with N population elements, n sample elements, and $f = n/N$

Systematic sampling: Implementation

1. Calculate the sampling interval $k = 1/f = N/n$
2. Select a random start RN between 1 and k
3. Select population elements: RN, RN+k, RN+2k, ..., RN+(n-1)k

Implementation

(A) Conceptual framework



	1	2	...	i	...	k
1	Y_1	Y_2	...	Y_i	...	Y_k
2	Y_{k+1}	Y_{k+2}	...	Y_{k+i}	...	Y_{2k}
3	Y_{2k+1}	Y_{2k+2}	...	Y_{2k+i}	...	Y_{3k}
.	.	.		.		.
n	$Y_{(n-1)k+1}$	$Y_{(n-1)k+2}$	...	$Y_{(n-1)k+i}$	...	Y_{nk}
	$\overline{Y}_1$	$\overline{Y}_2$	...	$\overline{Y}_i$	...	$\overline{Y}_k$

R code: Example

```
library(sampling)
data(belgianmunicipalities)
Tot=belgianmunicipalities$Tot04
name=belgianmunicipalities$Commune
##defines the inclusion probabilities: sum to n
pik=inclusionprobabilities(Tot,200)
#selects a sample
s=UPsystematic(pik)
#the sample is
#which(s==1)
# extracts the observed data
#getdata(belgianmunicipalities,s)
```

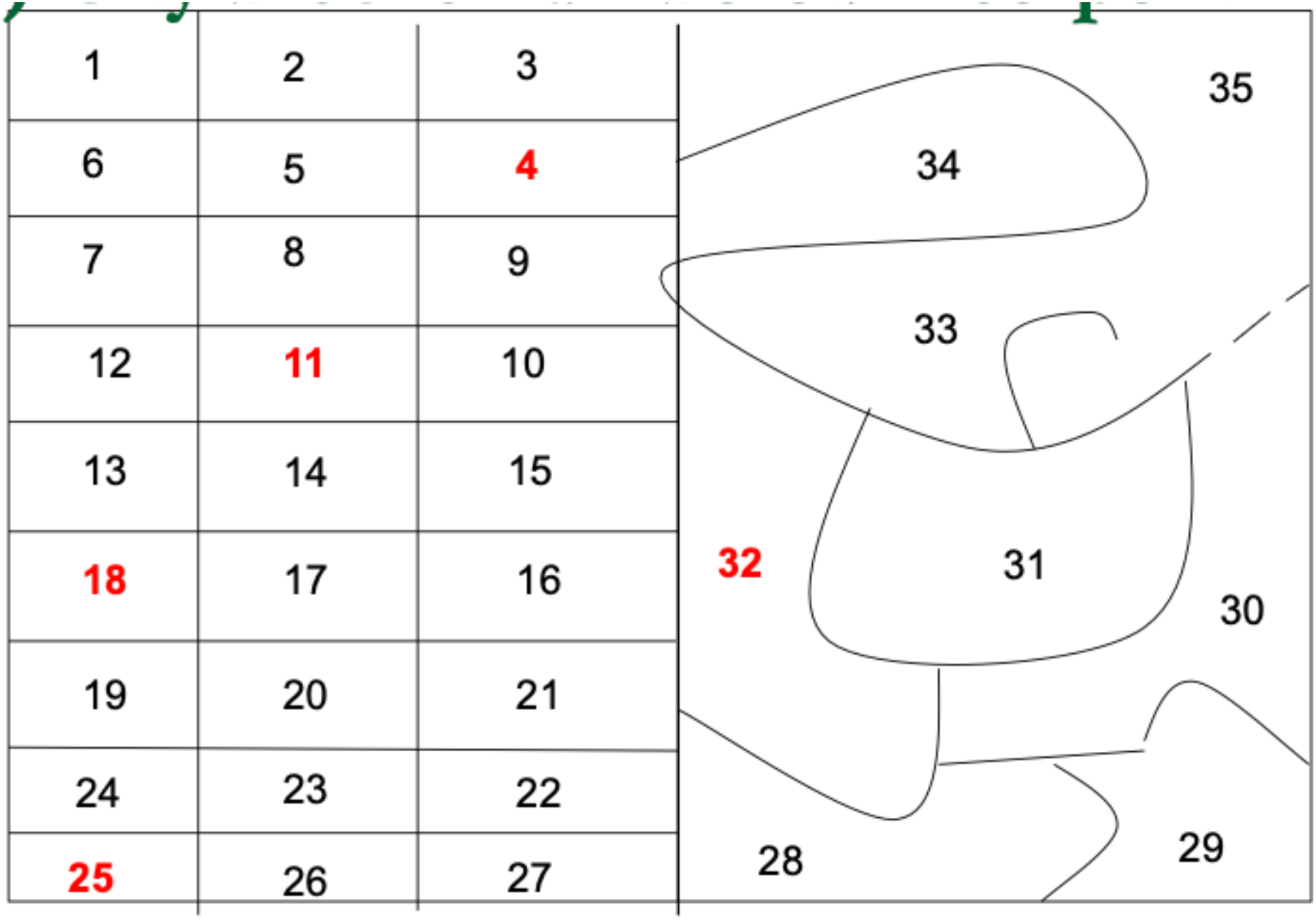
Clusters or strata?

- Cluster sample design is not measurable: $n = 1$
- Order of the list achieves stratification
  - Combination of list order and systematic selection
  - Group similar elements together in list
  - Zones in selection similar to strata: contain elements similar to one another
  - Stratified proportionate allocation

Implicit stratification

- List order combined with systematic selection can improve the efficiency (in terms of variance) of systematic sample designs
- Arrange the list order in advance
- Determines which samples selected
  - Random order: SRS
  - Stratified order
  - Serpentine order

Example: Serpentine ordered city blocks



Example with fractional intervals

- N=23, n=5, and $k = 23/5 = 4.6$

RS	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15
	47	48	49	50	51	52	53	54	55	56	57	58	59	60	61
	93	94	95	96	97	98	99	100	101	102	103	104	105	106	107
	139	140	141	142	143	144	145	146	147	148	149	150	151	152	153
	185	186	187	188	189	190	191	192	193	194	195	196	197	198	199
	231	232	233	234	235	236	237	238	239						

RS	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46
	78	79	80	81	82	83	84	85	86	87	88	89	90	91	92
	124	125	126	127	128	129	130	131	132	133	134	135	136	137	138
	170	171	172	173	174	175	176	177	178	179	180	181	182	183	184
	216	217	218	219	220	221	222	223	224	225	226	227	228	229	230

Inference

- In a population of NM elements, there are N possible choices for the systematic sample, each of size M
- We select one of the N PSUs and observe only one PSU
- The theorectical variance is $Var(\hat{\bar{y}}_{sys}) = (1 - \frac{1}{N})\frac{S_t^2}{M^2}$
- Systematic sampling is not measurable
- One solution: model-assisted variance estimation approach; we assume a model for the population sort order and the sample selection process

Model assisted variance estimation

- Model the population (sample selection process)
- SRS model
  - Are elements in the list are ordered at random?
  - Yes
    - Use SRS variance estimation formula: $var(\bar{y}) = (1 - f)\frac{s^2}{n}$

Stratified random model

- Are elements in the list are ordered at random?
- No
- Can we assume homogeneity across 'rows' (zones), in groups of rows?
- Yes
- Assume random ordering within zones
- Proportionately allocated selection with $n_h = 1$ selected per zone
- Collapse neighboring zones to create "pseudo strata" that have more than one selection
- Use the stratified sampling variance estimation:
  $var(\bar{y}) = (1 - f)\frac{1}{n}\sum W_h s_h^2$

Paired selection model

- Are elements in the list are ordered at random? No
- Can we assume homogeneity across 'rows' (zones), in groups of rows? No
- Is the ordering really almost continuous?
- Yes?
- Stratified random model special case: pair successive rows: $var(\bar{y}) = (1-f)\frac{1}{n^2}\sum(y_{h1} - y_{h2})^2$

Successive differences model

- Are elements in the list are ordered at random? No
- Can we assume homogeneity across 'rows' (zones), in groups of rows? No
- Is the ordering really almost continuous? No
- Is there likely to be a correlation between the i-th and $(i+1)$-th element that is larger than the correlation between the i-th and $(i+2)$-th or the i-th and $(i+3)$-th, etc.? Yes?
- $var(\bar{y}) = \frac{1-f}{2n(n-1)} \sum_{i=1}^{n-1} (y_i - y_{i+1})^2$

Approximate the sampling variance

- How to choose the best model?
- List ordered, deliberately, or accidentally, but achieves implicit stratification

1. Random order: use SRS sampling variance with $df = n - 1$
2. Sorted by a categorical variable: stratified model, possibly collapsing categories, $df = n - H$
3. Sorted by a continuous variable:

- Even sample size: paired selection model, $df = n/2$
- Odd sample size: successive differences model, df is between $n/2$ and $n - 1$

Bias considerations

- There is also a consideration of the bias of the variance estimates if we guess wrong about list order
- Model fails to "capture" stratification represented by systematic selection
- SRS: no stratification effects are accounted for
- Stratified selection: captures stratification
  - Collapses across stratum boundaries
  - Paired selection: two per stratum
  - Successive differences: deepest stratification possible

Example: Selected blocks

RS+k	Block	# Rental	# HUs	g
3	240	23	30	1
13	278	25	33	2
23	288	42	61	3
33	377	0	3	4
43	388	16	27	5
53	398	37	47	6

Mean estimation

- *Epsem* sample

$$\bar{y}_{\#rental} = \frac{\sum y_i}{n}$$

$$= (23 + 25 + 42 + 0 + 16 + 37)/6 = 23.83$$

- What is the estimated average total number of HUs?

Variance estimation (1)

- Is the list order random?
- SRS model

$$\text{var}(\bar{y}) = (1-f)\frac{s^2}{n} = (1-\frac{6}{60})(\frac{1}{6})\frac{(4543-6*23.83^2)}{6-1}$$

$$= (0.90)(0.1667)(226.97) = 34.045$$

Variance estimation (2)

- This list is probably continuously ordered with respect to Y.
- Paired selection model, even # elements

$$\text{var}(\bar{y}) = \frac{(1-f)}{n^2}\sum_{h}^{n/2}(y_{ha}-y_{hb})^2$$

$$= \left(1-\frac{6}{60}\right)\left(\frac{1}{6^2}\right)[(23-25)^2+(42-0)^2+(16-37)^2]$$

$$= (0.9)(0.0278)(4+1764+441) = 55.225$$

Variance estimation (3)

- But maybe we can 'double up' the differences, averaging both before and after differences …
- Successive differences model

$$\text{var}(\bar{y}) = \frac{1-f}{2n(n-1)} \sum_{g}^{n-1} (y_g - y_{g+1})^2$$

$$= \frac{1-6/60}{2\times 6(6-1)} [(23-25)^2 + (25-42)^2 + (42-0)^2 + (0-16)^2 + (16-37)^2]$$

$$= \frac{(0.9)}{60} [4 + 289 + 1764 + 256 + 441] = 41.31$$

Example: Samples of SBP measures

ID	3	8	12	17	21
Age (Frame)	19	23	30	42	58
Systolic Blood Pressure (Y)	115	120	123	132	140

- Sample mean: $\bar{y} = 126; f = 5/23\,; n = 5$
- Use **successive differences** model for variance estimation (why?)
- $var(\bar{y}) = \frac{1-f}{2n(n-1)} \sum_g (y_g - y_{g+1})^2 = \frac{1-5/23}{2\times5\times4}[(115 - 120)^2 + (120 - 123)^2 + \cdots + (132 - 140)^2] = 3.502$
- $se(\bar{y}) = 1.871$
- Use $df = n/2 = 2.5$ to form a confidence interval for the population mean (small df in this example).

Summary

- Systematic sampling is easy to implement
- Systematic sampling is not measurable
- One solution: model-assisted variance estimation approach; we assume a model for the population sort order and the sample selection process
- How to choose the best model?
  - Random order: use SRS sampling variance
  - Sorted by a categorical variable: stratified model, possibly collapsing categories
  - Sorted by a continuous variable:
    - Even sample size: paired selection model
    - Odd sample size: successive differences model