Lecture 13: Design a cluster sample

Prof. Lingling An
University of Arizona

Outline

- Review
- Revisit self-weighting design
- Design a cluster sample
- Systematic sampling

Review: 2-stage equal probability cluster sampling (CSE2)

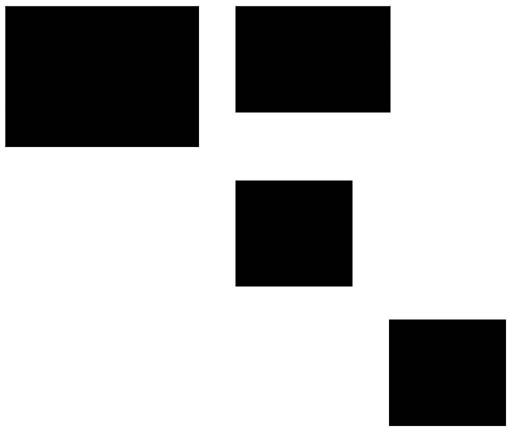
- CSE2 has 2 stages of sampling

Stage 1. Select SRS of n PSUs from population of N PSUs

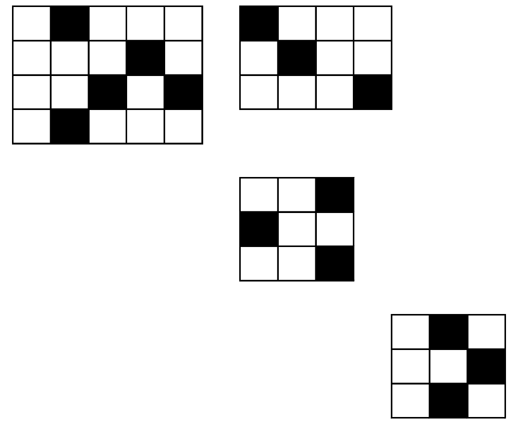
Stage 2. Select SRS of m_i SSUs from M_i elements in PSU i sampled in stage 1

Review: 2-stage cluster sampling

Sample all SSUs in sampled PSUs:



Take an SRS of m_i SSUs in sampled PSU i :



Stage **1** of 2-stage cluster sample (select PSUs)

Stage **2** of 2-stage cluster sample (select SSUs w/in PSUs)

Review: Motivation for 2-stage cluster samples

- Recall motivations for cluster sampling in general
  - Only have access to a frame that lists clusters
  - Reduce data collection costs by going to groups of nearby elements (cluster defined by proximity)

Review: Motivation for 2-stage cluster samples – 2

- Likely that elements in cluster will be correlated
  - May be inefficient to observe all elements in a sample PSU
  - Extra effort required to fully enumerate a PSU does not generate that much extra information
- May be better to spend resources to sample many PSUs and a small number of SSUs per PSU
  - Possible opposing force: study costs associated to going to many clusters

Review: CSE2 unbiased estimation for population total *t*

- Have a <u>sample</u> of elements from a cluster
  - We no longer know the value of cluster parameter, t_i
- Estimate t_i using data observed for m_i SSUs

$$\hat{t}_i = M_i \bar{y}_i = \sum_{j=1}^{m_i} \frac{M_i}{m_i} y_{ij}$$

Review: CSE2 unbiased estimation for population total – 2

- Approach is to plug estimated cluster totals into CSE1 formula
  - CSE1 $$\hat{t}_{unb} = \frac{N}{n}\sum_{j=1}^{n} t_i = \frac{N}{n}\sum_{i=1}^{n} M_i \bar{y}_{iU}$$
  - CSE2 $$\hat{t}_{unb} = \frac{N}{n}\sum_{j=1}^{n} \hat{t}_i = \frac{N}{n}\sum_{i=1}^{n} M_i \bar{y}_i$$

Review: CSE2 unbiased estimation for population total – 3

- The variance of $\hat{t}_{unb}$ has 2 components associated with the 2 sampling stages
  1. Variation among PSUs
  2. Variation among SSUs within PSUs

$$\hat{V}\left(\hat{t}_{unb}\right) = \underbrace{N^2\left(1-\frac{n}{N}\right)\frac{s_t^2}{n}}_{\text{among PSU}} + \underbrace{\frac{N}{n}\sum_{i=1}^{n}\left(1-\frac{m_i}{M_i}\right)M_i^2\frac{s_i^2}{m_i}}_{\text{within PSU}}$$

Review: CSE2 unbiased estimation for population total – 4

- In CSE1, we observe all elements in a cluster
  - We know t_i
  - Have variance component 1, but no component 2
- In CSE2, we sample a subset of elements in a cluster
  - We estimate t_i with $\hat{t}_i$
  - Component 2 is a function of estimates variance for $\hat{t}_i$

$$M_i^2\left(1-\frac{m_i}{M_i}\right)\frac{s_i^2}{m_i}$$

Review: CSE2 unbiased estimation for population total – 5

- Estimated variance among cluster totals

$$s_t^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left(\hat{t}_i - \frac{\hat{t}_{unb}}{N} \right)^2$$

- Estimated variance among elements in a cluster

$$s_i^2 = \frac{1}{m_i - 1} \sum_{j=1}^{m_i} \left(y_{ij} - \bar{y}_i \right)^2$$

Review: CSE2 unbiased estimation for population mean $\bar{y}_U$

$$\hat{\bar{y}}_{unb} = \frac{\hat{t}_{unb}}{M_0}$$

$$\hat{V}\left(\hat{\bar{y}}_{unb}\right) = \frac{\hat{V}\left(\hat{t}_{unb}\right)}{M_0^{\,2}}$$

Review: CSE2 ratio estimation for population mean $\bar{y}_U$

$$\hat{\bar{y}}_r = \frac{\sum_{i=1}^{n} \hat{t}_i}{\sum_{i=1}^{n} M_i} = \frac{\sum_{i=1}^{n} M_i \bar{y}_i}{\sum_{i=1}^{n} M_i}$$

Review: CSE2 ratio estimation for population mean – 2

$$\hat{V}\left(\hat{\bar{y}}_r\right) = \frac{1}{\bar{M}_U^2}\left[\left(1-\frac{n}{N}\right)\frac{s_r^2}{n} + \frac{1}{nN}\sum_{i=1}^{n} M_i^2\left(1-\frac{m_i}{M_i}\right)\frac{s_i^2}{m_i}\right]$$

$$s_r^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left[M_i\bar{y}_i - M_i\hat{\bar{y}}_r\right]^2 = \frac{1}{n-1}\sum_{i=1}^{n} M_i^2\left[\bar{y}_i - \hat{\bar{y}}_r\right]^2$$

$\bar{M}_U$ can be estimated by sample mean of M_i or $\bar{M}_S = \frac{1}{n}\sum_{i=1}^{n} M_i$

Review: CSE2 ratio estimation for population total *t*

$$\hat{t}_r = M_0 \, \hat{\bar{y}}_r$$

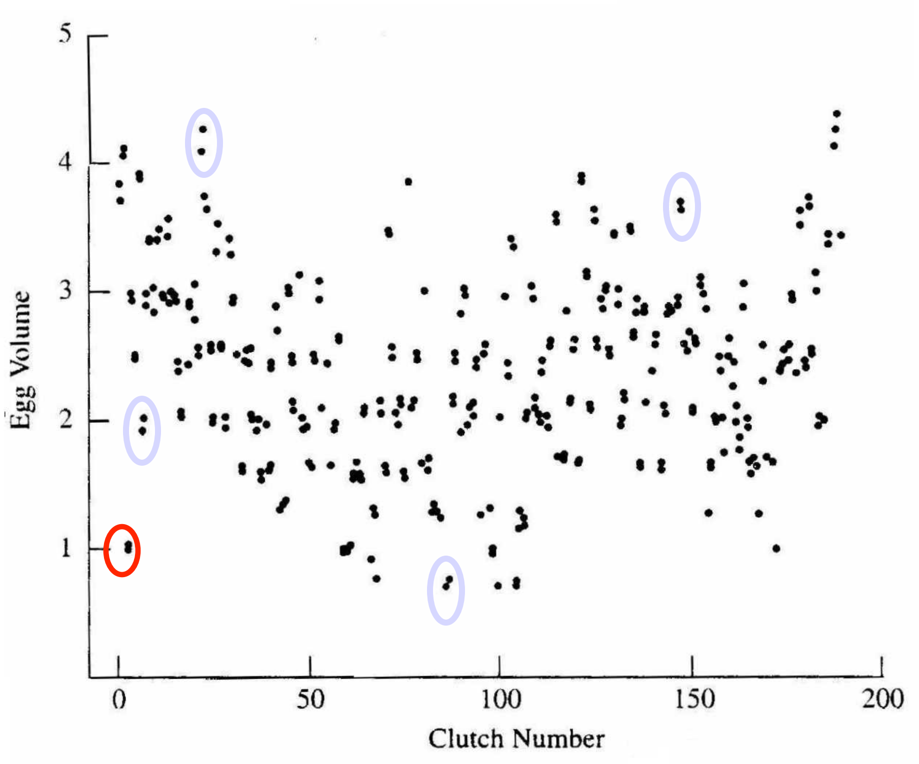
$$\hat{V}\left(\hat{t}_r \right) = M_0^{\,2} \hat{V}\left(\hat{\bar{y}}_r \right)$$

Coots egg example

- Target pop = American coot eggs in Minnedosa, Manitoba
- PSU / cluster = clutch (nest)
- SSU / element = egg w/in clutch
- Stage 1
  - SRS of n = 184 clutches
  - N = ??? Clutches, but probably pretty large
- Stage 2
  - SRS of m_i = 2 from M_i eggs in a clutch
  - Do not know M_0 = ??? eggs in population, also large
  - Can count M_i = # eggs in sampled clutch i
- Measurement
  - y_{ij} = volume of egg j from clutch i

Coots egg example – 2

(y_{ij})

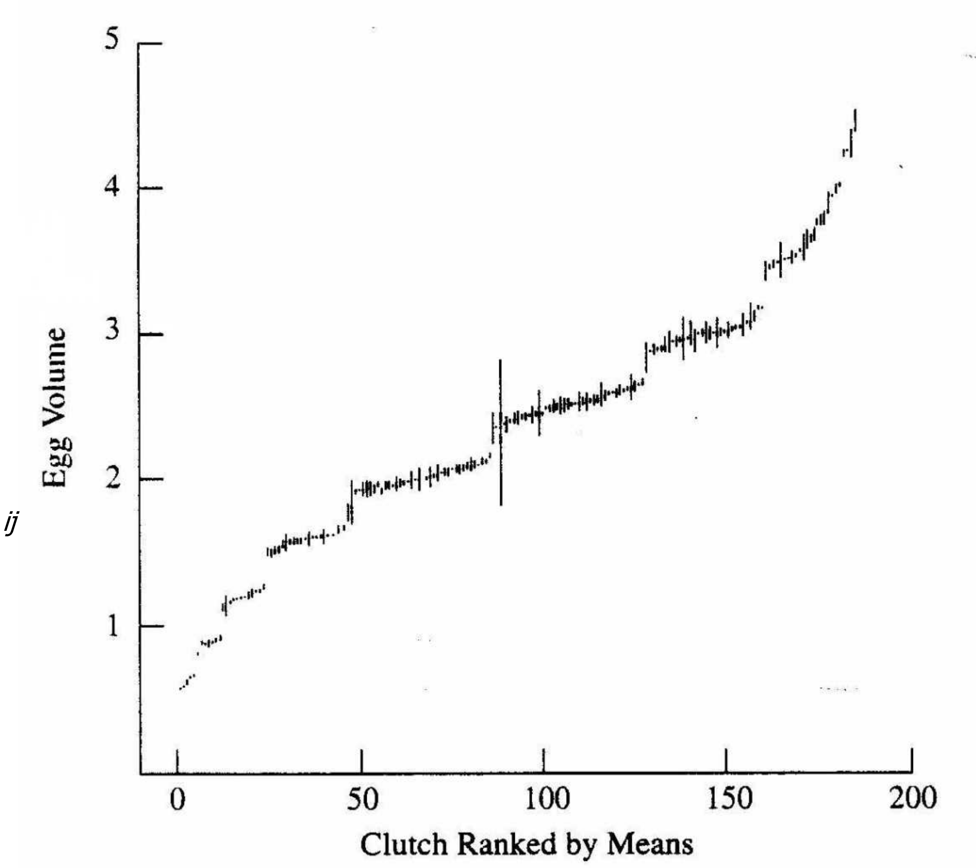


(i)

- Scatter plot of volumes vs. *i* (clutch id)
  - Double dot pattern - high correlation among eggs <u>WITHIN</u> a clutch
  - Quite a bit of clutch to clutch variation
- Implies
  - May not have very high precision unless sample a large number of clutches
  - Certainly lower precision than if obtained a SRS of $\sum_{i=1}^{n} m_i = 368$ eggs

Could use a side-by-side plot for data with larger cluster sizes – PROC UNIVARIATE w/ BY CLUSTER and PLOTS option

Coots egg example – 3

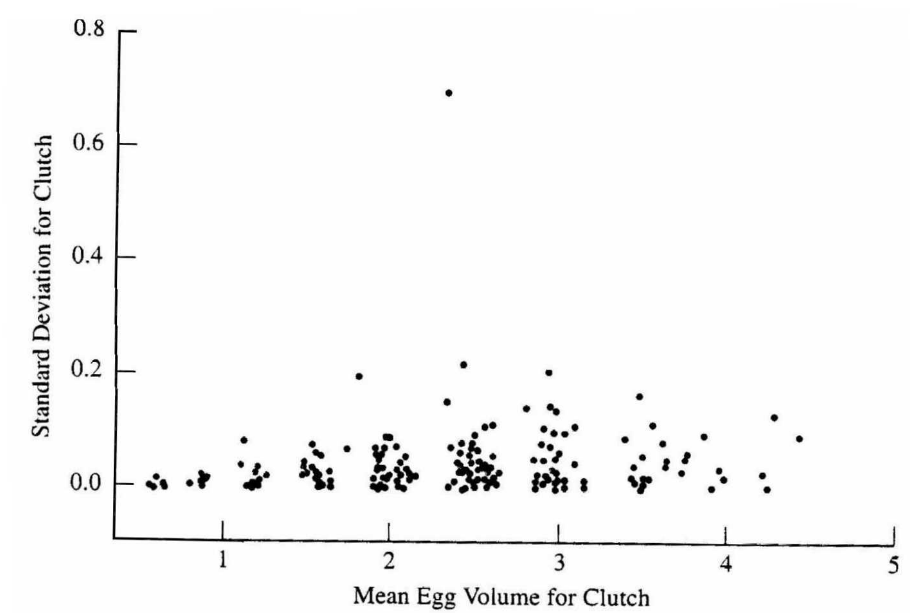


y_{ij}

i sorted by $\bar{y}_i$

- Plot
  - Rank the mean egg volume for clutch *i* , $\bar{y}_i$
  - Plot y_{ij} vs. rank for clutch *i*
  - Draw a line between y_{i1} and y_{i2} to show how close the 2 egg volumes in a clutch are
- Observations
  - Same results as Fig 5.3, but more clear
    - Small within-cluster variation
    - Large between-cluster variation
  - Also see 1 clutch with large WITHIN clutch variation
    - check data (i = 88)

Coots egg example – 4



$\bar{y}_i$

- Plot s_i vs. $\bar{y}_i$ for clutch i
- Since volumes are always positive, might expect s_i to increase as $\bar{y}_i$ gets larger
  - If $\bar{y}_i$ is very small, y_{i1} and y_{i2} are likely to be very small and close -> small s_i
  - See this to moderate degree
- Clutch 88 has large s_i, as noted in previous plot

Coots egg example – 5

- Estimation goal
  - Estimate $\bar{y}_U$, population mean volume per coot egg in Minnedosa, Manitoba
- What estimator?
  - Unbiased estimation
    - Don't know N = total number of clutches or M_0 = total number of eggs in Minnedosa, Manitoba
  - Ratio estimation
    - Only requires knowledge of M_i , number of eggs in selected clutch i , in addition to data collected
    - May want to plot $\hat{t}_i$ versus M_i

Coots egg example – 6

Clutch	M_i	$\bar{y}_i$	s_i^2	$\hat{t}_i$	$\left(1-\frac{2}{M_i}\right)M_i^2\frac{s_i^2}{m_i}$	$\left(\hat{t}_i - M_i\,\hat{\bar{y}}_r\right)^2$
1	13	3.86	0.0094	50.23594	0.671901	318.9232
2	13	4.19	0.0009	54.52438	0.065615	490.4832
3	6	0.92	0.0005	5.49750	0.005777	89.22633
4	11	3.00	0.0008	32.98168	0.039354	31.19576
5	10	2.50	0.0002	24.95708	0.006298	0.002631
6	13	3.98	0.0003	51.79537	0.023622	377.053
7	9	1.93	0.0051	17.34362	0.159441	25.72099
8	11	2.96	0.0051	32.57679	0.253589	26.83682
9	12	3.46	0.0001	41.52695	0.006396	135.4898
10	11	2.96	0.0224	32.57679	1.108664	26.83682
...	...	...	...	...	...	...
180	9	1.95	0.0001	17.51918	0.002391	23.97106
181	12	3.45	0.0017	41.43934	0.102339	133.4579
182	13	4.22	0.00003	54.85854	0.002625	505.3962
183	13	4.41	0.0088	57.39262	0.630563	625.7549
184	12	3.48	0.000006	41.81168	0.000400	142.1994
sum	1757			4375.947	42.17445	11,439.58
var				149.565814		
$\hat{\bar{y}}_r =$		2.490579				

Coots egg example – 7

$$\hat{\bar{y}}_r = \frac{\sum_{i \in S} \hat{t}_i}{\sum_{i \in S} M_i} = \frac{4375.947}{1757} = 2.49$$

Don't know $\bar{M}_U$

$$s_r^2 = \frac{\sum_{i \in S} \left(\hat{t}_i - M_i \, \hat{\bar{y}}_r \right)^2}{n - 1} = \frac{11{,}439.58}{183} = 62.511$$

Don't know N, but assumed large

Use $\bar{M}_S$

$$\bar{M}_S = 1757 \,/\, 184 = 9.549$$

FPC ≈ 1

$$\hat{V}\left(\hat{\bar{y}}_r\right) = \frac{1}{9.549^2} \left[\left(1 - \frac{184}{N} \right) \frac{62.511}{184} + \left(\frac{1}{N} \right) \frac{42.17}{184} \right]$$

2nd term is very small, so approximate SE ignores 2nd

$$SE\left(\hat{\bar{y}}_r \right) \cong \frac{1}{9.549} \sqrt{\frac{62.511}{184}} = 0.061$$

CSE2: Unbiased vs. ratio estimation

- Unbiased estimator can poor precision if
  - Cluster sizes (M_i) are unequal
  - t_i (cluster total) is roughly proportional to M_i (cluster size)
- Biased (ratio estimator) can be precise if
  - t_i roughly proportional to M_i
  - This happens frequently in pops w/cluster sizes (M_i) vary

Summary of CS

- Cluster sampling is commonly used in large survey
  - But with large variance
- If it is much less expensive to sample clusters than individual elements, CS can provide more precision per dollar spent.

Inclusion probability for an element under CSE2 (using SRS at each stage)

- π_i = P{cluster i in sample}
  $= n / N$

- $\pi_{j|i}$ = Pr {element j given cluster i in sample}
  $= m_i / M_i$

- π_{ij} = Pr {element j and cluster i in sample}
  $= \pi_i \, \pi_{j|i}$
  $= (n / N) \times (m_i / M_i)$
  $= nm_i / NM_i$

CSE2 weight for an element (unbiased estimator)

- Estimator for population total

$$\hat{t}_{unb} = \frac{N}{n}\sum_{i=1}^{n}\hat{t}_i = \frac{N}{n}\sum_{i=1}^{n}M_i\bar{y}_i = \frac{N}{n}\sum_{i=1}^{n}\frac{M_i}{m_i}\sum_{j=1}^{M_i}y_{ij}$$

$$= \sum_{i=1}^{n}\sum_{j=1}^{M_i}\boxed{\frac{N}{n}\frac{M_i}{m_i}}y_{ij} = \sum_{i=1}^{n}\sum_{j=1}^{M_i}w_{ij}y_{ij}$$

- Weight for element j in cluster i

$$w_{ij} = \frac{N}{n}\frac{M_i}{m_i} = \frac{1}{\pi_{ij}}$$

CSE2: Self-weighting design

- Stage 1: Select n PSUs from N PSUs in pop using SRS
  - Inclusion probability for PSU i : $\pi_i = \dfrac{n}{N}$

- Stage 2: Choose m_i proportional to M_i so that m_i / M_i is constant, use SRS to select sample
- Sample weight for SSU j in cluster i is constant for all elements

$$w_{ij} = \frac{N}{n}\frac{M_i}{m_i} = \frac{N}{n}c$$

Weight may vary slightly in practice because may not be possible for m_i / M_i to be equal to 1/c for all clusters

Self-weighting designs in general

- Why are self-weighting samples appealing?
- Are dorm student or coot egg samples self-weighting 2-stage cluster samples?
- What self-weighting designs have we discussed?

Self-weighting designs in general – 2

- What is the caveat for variance estimation in self-weighting samples?
  - No break on variance of estimator – must use proper formula for design

- Why are self-weighting samples appealing?
  - Simple mean estimator
  - Homogeneous weights tends to make estimates more precise

Self-weighting designs

- SRS
- SYS

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

- STS with proportional allocation

$$w_{hj} = \frac{N_h}{n_h} = \frac{N}{n}$$

- CSE1

$$w_{ij} = \frac{N}{n}$$

- CSE2 with m_i proportional to M_i or $c = M_i / m_i$

$$w_{ij} = \frac{N}{n}\frac{M_i}{m_i} = \frac{N}{n}c$$

Design a cluster sample

- Need to decide 4 major issues:
  - 1. What overall precision is needed?
  - 2. What size should the PSUs be?
  - 3. How many PSUs should be sampled?
  - 4. How many SSUs should be sampled in each PSU selected for the sample?

Design a cluster sample -2

- Q1 must be faced in any survey design.
- Q2-4: need know the cost of sampling a PSU, the cost of sampling a SSU, measure of homogeneity for the possible sizes of PSU.

Design a cluster sample -2

- Choosing the PSU size
  - The PSU size is often a natural unit.
  - In the case of you need to decide the PSU size, a general principle is:
    - Larger the PSU size, the more variability you expect to see within a PSU.
    - If the PSU size is too large, however, you may lose the cost savings of cluster sampling.

Design a cluster sample -3

- Choosing subsampling sizes:
  - Assume $M_i=M$, and $m_i=m$ for all PSUs,
  - Total cost=$C=c_1n+c_2nm$

$$m_{opt} = \sqrt{\frac{c_1 M(N-1)(1-R_a^2)}{c_2(MN-1)R_a^2}}$$

$$n_{opt} = \frac{C}{c_1 + c_2 m_{opt}}$$

where R_a^2 is defined in eq (5.11) and it measures the homogeneity in general population.

Return to systematic sampling (SYS)

- Have a frame, or list of *N* elements
- Determine sampling interval, *k*
  - *k* is the next integer after *N*/*n*
- Select first element in the list
  - Choose a random number, *R* , between 1 & *k*
  - *R-th* element is the first element to be included in the sample
- Select every *k-th* element after the *R-th* element
  - Sample includes element *R,* element *R* + *k,* element *R* + 2*k, … ,* element *R* + (n-1)k

SYS example

- Telephone survey of members in an organization abut organization's website use
  - N = 500 members
  - Have resources to do n = 75 calls
  - N / n = 500/75 = 6.67
  - k = 7
  - Random number table entry: 52994
    - Rule: if pick 1, 2, …, 7, assign as R; otherwise discard #
  - Select R = 5
  - Take element 5, then element 5+7 =12, then element 12+7 =19, 26, 33, 40, 47, …

SYS – 2

- Arrange population in rows of length
  $k = 7$

R	**1**	**2**	**3**	**4**	**5**	**6**	**7**		***i***
	1	2	3	4	5	6	7		1
	8	9	10	11	12	13	14		2
	15	16	17	18	19	20	21		3
	22	23	24	25	26	27	28		4
	...								...
	491	492	493	494	495	496	497		71
	498	499	500						72

Properties of systematic sampling – 1

- Number of possible SYS samples of size *n* is *k*
- Only 1 random act - selecting *R*
  - After select 1st SU, all other SUs to be included in the sample are predetermined
  - A SYS is a cluster with sample(i.e., cluster) size k
    - Cluster = set of SUs separated by *k* units
- Unlike SRS, some sample sets of size *n* have no chance of being selected given a frame
  - A SU belongs to 1 and only 1 sample

Properties of systematic sampling – 2

- Because only the starting SU of a SYS sample is randomized, a direct estimate of the variance of the sampling distribution can not be estimated
  - Under SRS, variance of the sampling distribution was a function of the population variance, S^2
  - Have no such relationship for SYS

Estimation for SYS

- Use SRS formulas to estimate population parameters and variance of estimator

Estimate pop MEAN $\bar{y}_U$ with $\bar{y}$ and $\hat{V}[\bar{y}] = \frac{s^2}{n}\left(1 - \frac{n}{N}\right)$

Estimate pop TOTAL t with $\hat{t}$ and $\hat{V}[\hat{t}] = N^2\hat{V}[\bar{y}]$

Estimate pop PROPORTION p with $\hat{p}$ and $\hat{V}[\hat{p}] = \frac{\hat{p}(1-\hat{p})}{n-1}\left(1 - \frac{n}{N}\right)$

Properties of systematic sampling – 3

- Properties of SRS estimators depends on frame ordering
  - SRS estimators for population parameters usually have little or no bias under SYS
  - Precision of SRS estimators under SYS depends on ordering of sample frame

Order of sampling frame

- Random order
  – SYS acts very much like SRS
  – SRS variance formula is good approximation
- Ordered in relation to y
  – Improves representativeness of sample
  – SRS formula overestimates sampling variance (estimate is more precise than indicated by SE)
- Periodicity in y = sampling interval k
  – Poor quality estimates
  – SRS formula underestimates sampling variance (overstate precision of estimate)

Example – 3

- Suppose *X* [age of member] is correlated with *Y* [use of org website]
- Sort list by *X* before selecting sample

k	**1**	**2**	**3**	**4**	**5**	**6**	**7**	***X***	***i***
	1	2	3	4	5	6	7	young	1
	8	9	10	11	12	13	14		2
	15	16	17	18	19	20	21		3
	22	23	24	25	26	27	28		4
	…							mid	…
	491	492	493	494	495	496	497		71
	498	499	500					old	72

Practicalities

- Another building block (like SRS) used in combination with other designs
- SYS is more likely to be used than SRS if there is no stratification or clustering
- Useful when a full frame cannot be enumerated at beginning of study
  - Exit polls for elections
  - Entrance polls for parks

Practicalities – 2

- Best if you can sort the sampling frame by an auxiliary variable X that is related to Y
  - Improve representativeness of sample (relative to SRS)
  - Improve precision of estimates
  - Essentially offers implicit form of stratification

Last slide

- Read Sections 5.3-5.5