STA304/1003 H1 F - Summer 2014: Surveys, Sampling, and Observational Data

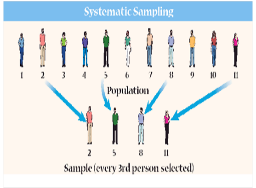
Lecture 9 - Part II: Systematic Sampling

Ramya Thinniyam

June 12, 2014

Systematic Sampling

- Want to take a sample of size n from a population of size N
- $k = N/n$ possible samples each with size n (k is called the sampling interval)
- Randomly select a number from 1 to k - call this number i
- Sample every kth element starting from i: $y_i, y_{i+k}, y_{i+2k}, \ldots, y_{i+(n-2)k}, y_{i+(n-1)k}$ is the sample of measured variables



- Special case of one-stage cluster sampling:
  - Population has k psus each of size n
  - Take an SRS of 1 psu - the set of elements in the selected psu is our sample

Example: Simple Systematic Sample

We want to take a sample of size 3 from elements $1, 2, 3, 4, \ldots, 13, 14, 15$. Setup this problem with Systematic Sampling. Identify the parameters of the systematic sample (ie. N, n, k, psus, etc.)

population = Sampling frame = $\{1, 2, \cdots, 15\}$

Systematic sample of size 3 from this pop

$N=15$, $n=3$, $k=\frac{N}{n}=\frac{15}{3}=5$

PSUS $k=5$ psus/clusters

$S_1=\{1,6,11\}$
$S_2=\{2,7,12\}$
$S_3=\{3,8,13\}$
$S_4=\{4,9,14\}$
$S_5=\{5,10,15\}$

- Every psu has 1/5 possibility of selection
- once psu selected that psu = sample
- $k=5$ possible samples with $P(\text{sample})=1/5$

- randomly choose $i \in \{1:5\}$
- Then S_i is the selected sample one-stage cluster sample w/ 5 psus, select 1 psu

$P(j\text{th unit is in the sample}) = \frac{1}{5} \quad \forall j$

↳ self-weighting.

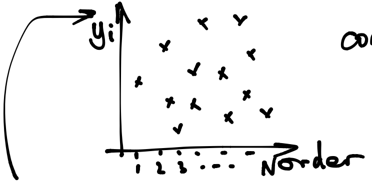
Ex: $\{1,4,8\}$ not possible sample so NOT SRS

Estimation

1. Estimating the Population Mean:
   - Observe mean of psu selected: $\hat{\bar{y}}_{sys} = \bar{y}_i = \bar{y}_{iU}$
   - $E(\hat{\bar{y}}_{sys}) = \bar{y}_U$
   - $V(\hat{\bar{y}}_{sys}) = \left(1 - \frac{1}{k}\right) \frac{S_t^2}{n^2} = \left(1 - \frac{1}{k}\right) \frac{MSB}{n} \approx \frac{S^2}{n}[1 + (n-1)ICC]$
2. Estimating the Population Total:
   - Estimate total using the psu selected: $\hat{\tau}_{sys} = N\hat{\bar{y}}_{sys}$
   - $E(\hat{\tau}_{sys}) = \tau$
   - $V(\hat{\tau}_{sys}) = N^2 \left(1 - \frac{1}{k}\right) \frac{S_t^2}{n^2} = N^2 \left(1 - \frac{1}{k}\right) \frac{MSB}{n} \approx N^2 \frac{S^2}{n}[1 + (n-1)ICC]$

   - If $ICC < 0$: Systematic sampling more precise than SRS of size n. i.e. when variance within possible systematic samples (psus) is larger than overall variance
   - If ICC is large: SRS more precise than Systematic sampling ie. little variation within psus so elements within the sample are similar and give similar information
   - Note: effective sample size is 1 for a simple systematic sample - select 1 psu out of k in a one-stage cluster sample:
   - Cannot estimate variance like before since we have only one cluster total/mean: need population structure

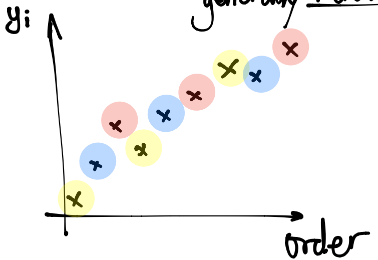
Types of Sampling Frames



completely scattered

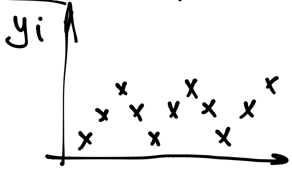
1. Sampling Frame is in random order:
   - Ordering of population not related to characteristics of interest
   - Expect $ICC \approx 0$ (Good for us)
   - Behaves like SRS: Use SRS results and formulas to calculate $\hat{V}(\hat{\bar{y}}_{sys}) = \left(1 - \frac{n}{N}\right)\frac{s^2}{n}$ and $\hat{V}(\hat{\tau}_{sys}) = N^2\left(1 - \frac{n}{N}\right)\frac{s^2}{n}$

"generally monotone"



a little bit problematic

2. Sampling Frame is in increasing/decreasing order:
   - Positive autocorrelation: closer elements are more similar than those that are farther apart
   - $ICC < 0$: variance is lower than in SRS because systematic sample forces the sample values to be spread out whereas an SRS could happen to pick elements with similar values
   - Use SRS variance estimate but it will like overestimate population variance (the true)



If k=period (multiple) estimate of variance using SRS will be 0.
→ sample is equivalent to 1 obs.

3. Sampling Frame has periodic pattern:
   - If interval length = periodicity (or multiple of it): Systematic sampling less precise than SRS
   - Underestimate variance if SRS estimate is used

Using ‘R’ to generate a Systematic Sample

Population with 100 elements, sample size of $n = 20$ with interval length of $k = 5$

1. Read Data into ‘R’:

```
# population in random order
> mydata = round(runif(100,0,20),1)
```

```
> mydata
  [1] 13.2  9.3 12.5 16.8 11.1 10.1 14.9 13.4  2.0 17.4  4.2 15.4
 .
 .
 [77] 13.3  2.1  7.2  2.5  2.2 13.7  8.4 14.6  1.4  9.1 19.5 9.8 4.1
      19.7  0.3  0.8  9.4  0.6 10.0 10.3 12.9 12.8 15.5  2.6
```

2. Generate a random starting position:

```
> start = sample(1:5,1)
> start
[1] 3
```

3. Take the sample:

```
> mysample <- NULL
> for (i in 1:20) {
     mysample[i] = mydata[start + (i-1)*5 ]
     }
```

OR

```
> mysample = mydata[seq(from=start,to=100,by=5)]
```

```
> mysample
 [1] 12.5 13.4  7.9  6.5  1.1  17.9  2.8 11.1 19.8  3.6
       6.1 15.8  8.7 15.0  13.9 2.1   8.4  9.8  9.4 12.8
```

4. Find systematic mean estimate and its standard error:

```
> ybar.sys=mean(mysample)
> SE.ybar.sys = sqrt((1-20/100)*var(mysample)/20)
> ybar.sys
[1] 9.93
> SE.ybar.sys
[1] 1.056311
```

Define as Clusters:

```
> cluster = NULL
> for (i in 1:100) {
      if(i %% 5 ==0) {
         cluster[i]=5
      }
      else {cluster[i] = i %% 5
      }
  }
```

OR

```
> rep(c(1,2,3,4,5), 20)
[1]  1 2 3 4 5 1 2 3 4 5 1 2 3 4 5
      .
      .
[90] 1 2 3 4 5 1 2 3 4 5


> lin.reg <- lm(mydata ~ as.factor(cluster))

> anova(lin.reg)
Analysis of Variance Table

Response: mydata
                   Df Sum Sq Mean Sq F value Pr(>F)
as.factor(cluster)  4  116.9  29.217  0.8548 0.4941
Residuals          95 3247.0  34.179
```

Example: Comparing Systematic to SRS Variances

Ex: 1 Mydata: random order

$N=100$, $n=20$, $k=5$

Use the data/output to calculate *ICC* and compare using Systematic vs. SRS estimates for the variance. Explain your findings.

Since Sampling frame is in random order, expect ICC ≈ 0 (low), and expect SRS variance est $\approx$ systematic variance

$ICC = 1 - \frac{n}{n-1}\left(\frac{SSW}{SSTO}\right) = 1 - \frac{20}{19}\cdot\left(\frac{3247}{3247+116.9}\right) \doteq -0.02$

recall the bound of ICC: $\frac{-1}{n-1} \leq ICC \leq 1 \Rightarrow ICC \geq \frac{-1}{19} = -0.05$

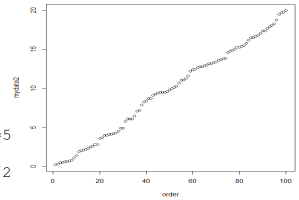
Theoretic Var: $V(\hat{\bar{y}}_{sys}) \doteq \frac{S^2}{n}[1+(n-1)ICC] = \frac{3247+116.9}{99(20)}[1+19(-0.02)] \doteq 1.05$ (using ANOVA)

Using SRS: $\hat{V}_{SRS}(\hat{\bar{y}}_{sys}) = \left(1-\frac{n}{N}\right)\frac{S^2}{n} = 1.12$

↳ approx. same variance, but systematic is better (lower variance)

Example: Ordered Data

```
> mydata2 = sort(mydata)
> order<-seq(1,100,1)
> plot(order,mydata2)

> start2 = sample(1:5,1)
> start2
[1] 2
> mysample2 = mydata2[seq(from=start2,to=100,by=5
> ybar.sys2=mean(mysample2)
> SE.ybar.sys2 = sqrt((1-20/100)*var(mysample2)/2

> ybar.sys2
[1] 9.535

> SE.ybar.sys2
[1] 1.187749


> lin.reg2 <- lm(mydata2 ~ as.factor(cluster))
> anova(lin.reg2)

Analysis of Variance Table

Response: mydata2
                   Df  Sum Sq  Mean Sq  F value  Pr(>F)
as.factor(cluster)  4     7.6     1.89   0.0535  0.9946
Residuals          95  3356.3    35.33
```



Example: Comparing Systematic to SRS Variances

Use the data/output to calculate *ICC* and compare using Systematic vs. SRS estimates for the variance. Explain your findings.

Ex: 2 Ordered data: $N=100, n=20, k=5$

expect $ICC<0$, and systematic variance to be better

$ICC = 1 - \frac{20}{19} \frac{3356.3}{3356.3+7.6} = -0.05$ (min value of ICC)

Theoretic Var: $V(\hat{\bar{y}}_{sys}) = \frac{S^2}{n}(1+(n-1)ICC] = \frac{33.98}{20}[1 + 19(-0.05)] \doteq 0.08$

Using SRS: $\hat{V}_{SRS}(\hat{\bar{y}}_{sys}) = (1-\frac{n}{N})\frac{s^2}{n} = (1.19)^2 \quad SE_{SRS} = 1.19$

Advantages/Disadvantages of Systematic Sampling

Simple Systematic Sample:

- Used when you want to get a representative sample but sampling frame is not constructed in advance
- Commonly used to select elements at the bottom stage of cluster sampling
- In many situations, can be treated as an SRS
- To solve periodicity problem, can use "Interpenetrating Systematic Sample": take several systematic samples from population.

  Each systematic sample is a psu - use formulas from cluster sampling to estimate variance

Advantages:

- Can be cheaper/easier to perform than SRS and STRS
- Cannot use personal bias (unless you are aware of sampling frame pattern)
- Can provide more information per unit cost than SRS can for populations with certain patterns

Disadvantages:

- May be biased/non-representative of population if periodic population
- Variance under/over estimated if population is periodic/ordered respectively