Simple Random Sample

SurvMeth/Surv 625: Applied Sampling

Yajuan Si

University of Michigan, Ann Arbor

1/15/25

Simple random sample (SRS)

- Implementation: Selection
- Inference: Analysis
- Projection: Future collection
- Sampling frame

Implementation

- SRS is the most basic form of probability sampling and provides the basis for the more complicated forms.
- Select a random sample of size n from a population of N
  1. SRS with replacement (SRSWR): the same unit can be included more than once in the sample, with the equal selection probability $\pi_i = n/N$
  2. SRS without replacement (SRS): all units in the sample are distinct with the equal selection probability $\pi_i = n/N$
- Uniform random number generator
  - R functions: sample(), sampling::srswor(), sampling::srswr(), etc.
- Reproducible: set.seed()

Inference: Population mean

- Support we collect the systolic blood pressure (SBP) measurements of the five individuals selected via SRSWOR from a population of 20 and would like to estimate the population average SBP value: 110, 125, 145, 90, 135
- The finite population correction factor $fpc = 1 - f = 1 - n/N = 1 - 5/20 = 15/20$
- The sample total $t = 110 + 125 + 145 + 90 + 135 = 605$
- The sample mean $\bar{y} = t/n = 605/5 = 121$
- The element variance estimate $s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \bar{y})^2 = 467.5$
- The standard deviation $s = \sqrt{s^2} = 21.622$

Inference: Population mean cont.

- The sampling variance estimate $var(\bar{y}) = (1-f)\frac{s^2}{n} = 70.125$
- The standard error $se(\bar{y}) = \sqrt{var(\bar{y})} = 8.374$
- The 95% confidence interval

$$\bar{y} \pm t_{1-\alpha/2,n-1} * se(\bar{y}) = 121 \pm t_{0.975,t} * 8.374,$$

where the critical t-score $t_{0.975,t} = 2.776$ (NOTE: not 1.96)

Inference: Population total

- Population total estimate: $\hat{t} = N\bar{y} = 20 * 121$
- The sampling variance of the population total estimate:

$$var(\hat{t}) = N^2 var(\bar{y}) = N^2(1-f)\frac{s^2}{n} = 20^2 * 70.125$$

The standard error: $se(\hat{t}) = \sqrt{var(\hat{t})} = N\sqrt{var(\bar{y})} = 20 * 8.374$

- The $1-\alpha$ confidence interval: $\hat{t} \pm t_{1-\alpha/2,n-1} se(\hat{t})$

Inference: Population proportion

- Now we are interested in estimating the proportion of individuals with hypertension, and hypertension indicator values are: 0, 0, 0, 1, 0
- The proportion estimate: $p = \bar{y} = 1/5$
- The theoretical sampling variance (UNKNOWN):

  $Var(p) = (1-f)\frac{S^2}{n} = \frac{1-f}{n}\frac{N}{N-1}P(1-P)$
- The sampling variance estimate:

$$var(p) = \frac{1-f}{n}\frac{n}{n-1}p(1-p) = (1-5/20)/4 * 1/5 * (1-1/5)$$

- The sampling error: $se(p) = \sqrt{var(p)}$
- The $1-\alpha$ confidence interval: $p \pm t_{1-\alpha/2,n-1}se(p)$

R Code

```
library(survey)
library(sampling)
library(SDAResources)
library(tidyverse)
data(agpop)
n <- length(agsrs$acres92)
ybar <- mean(agsrs$acres92,na.rm = T)
ybar
```

```
[1] 297897
```

```
hatvybar<-(1-n/3078)*var(agsrs$acres92,na.rm = T)/n
seybar<-sqrt(hatvybar); seybar
```

```
[1] 18898.43
```

```
# Calculate confidence interval by direct formula using t distribution
Mean_CI <- c(ybar - qt(.975, n-1)*seybar, ybar + qt(.975, n-1)*seybar)
names(Mean_CI) <- c("lower", "upper"); Mean_CI
```

```
   lower    upper
260706.3 335087.8
```

```
# To obtain estimates for the population total,
# multiply each of ybar, seybar, and Mean_CI by N = 3078
seybar*3078; Mean_CI*3078
```

```
[1] 58169381
```

```
     lower      upper
 802453859 1031400361
```

```
# Calculate coefficient of variation of mean
seybar/ybar
```

Projection

- When designing a new survey data collection, investigators often focus on one or two key variables of interest, decide the amount of sampling error, and must balance the estimation precision with survey costs.

1. Specify the tolerable error: What is expected of the sample, and how much precision do I need? The **desired precision** is often expressed as $P(|\bar{y} - \bar{y}_u| \leq e) = 1 - \alpha$, where e is called the *margin of error*, as one-half of the width of a 95% CI. Sometimes you like to achieve a desired relative precision, such as a desired CV.
2. Find an equation: Relating the sample size n to the desired precision.
3. Estimate any unknown quantities and solve for n.
4. Adjust expectations and re-calculate.

Example of sample size projection

- Desired precision
  1. $var(\bar{y}) = \hat{V} = 2.5$;
  2. Requested 95% CI: $(L, U) = (10, 20)$, we have $e = (U - L)/2 = z_{0.975} * \hat{V}$ and then $\hat{V} = 2.5$
  3. Desired CV: $cv(\bar{y}) = \frac{\hat{V}}{\bar{y}} = 0.5$. If $\bar{y} = 5$, then $\hat{V} = 2.5$
- Find an equation
  1. SRSWR: $n_0 = \frac{\hat{S}^2}{\hat{V}^2} = \frac{z_{0.975}^2 \hat{S}^2}{e^2}$
  2. SRS: $n = \frac{n_0}{1+\frac{n_0}{N}}$
- Solve for n: If $\widehat{S}^2 = 6250$, $\hat{V} = 2.5$, and $N = 100000$, then $n_0 = 1000$ and $n = 991$.

Design effects

- Design effect: Ratio of the variance under a new design to SRS variance with the same sample size
- Depend on the estimated quantity (e.g., mean, regression coefficient)
- Depend on the examined variable
- Often used as a comparison to SRS when evaluating complex sample survey designs
- Inflate the projected SRS sample size by the design effect for complex sample survey sample size projection

Sampling frame

- Frame: Set of materials used to designate a sample of units
- Rule links frame elements to population elements
- Accurate and up-to-date frames located in one location preferred
- Numbered, computerized lists are best
- The population and the list/frame may not match up

Frame problems: Nonconverage

- Some population elements are not on the frame
  - Have zero chance of selection
- Potential solutions
  1. Use supplemental frames that cover noncovered elements
  2. Use noncoverage weighting adjustments (will be discussed later)

Frame problems: Blanks

- Frame elements do not have corresponding population elements
- Know frame element is blank after selection
    - Screening to find eligible list elements
- Potential solutions
    1. Reject blanks by adjusting the sampling rate and size
    2. Substitute with the next element on the listing

Frame problems: Duplicates

- Occur when a single population element is linked to two or more frame elements
- Potential solutions
  1. If only a few readily identified, remove from the list before selection
     - Eliminating duplicates from the sample still leaves unequal probabilities of selection
  2. Choose unique listing
     - First, last, largest, or randomly chosen frame listing
  3. Determine how many duplicates for a given selected element and weight

Frame problems: Clustering

- Occurs when more than one population element can be selected by a sample frame element
- Potential solutions
  1. Take all elements within selected clusters
     - The sample size varies with unequal-size clusters (discuss later)
     - Can adjust the sample size in advance
  2. Use cluster sampling (discuss later)
  3. Weighting adjustment

Objective respondent selection

- In the social sciences, sampling a single element from small clusters of unequal size occurs often
  - Sampling individuals from households
- Techniques for households:
  - Nearest birthday method
  - Objective Respondent Selection proposed by Kish

Within household selection: The Kish method

- Interviewers list eligible household members by gender and age
- Use selection table
- Selection tables “rotated” across households
- Maximum of four eligibles per household can be handled
  - Can be expanded to handle households with five/six eligibles

Respondent selection tables

Table A (1/4)	
If number of eligible subjects is	**Select subject number**
1	1
2	1
3	1
4	1

Table B (1/12)	
If number of eligible subjects is	**Select subject number**
1	1
2	1
3	1
4	2

Table C (1/6)	
If number of eligible subjects is	**Select subject number**
1	1
2	1
3	2
4	2

Table D (1/6)	
If number of eligible subjects is	**Select subject number**
1	1
2	2
3	2
4	3

Table E (1/12)	
If number of eligible subjects is	**Select subject number**
1	1
2	2
3	3
4	3

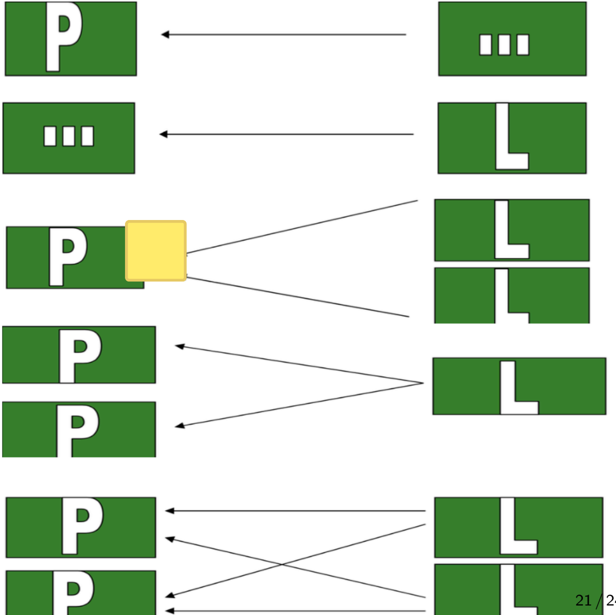
Table F (1/4)	
If number of eligible subjects is	**Select subject number**
1	1
2	2
3	3
4	4

Frame problems: Many-to-many matching

- Occurs when more than one population element can be selected by more than one frame element
- Potential solutions
    - Combinations of weighting and subsampling

Frame problems: Summary

- Non-coverage
- Blanks
- Duplicates
- Clustering
- Many to many



Example: Address-based sample of families

- A survey about families and their use of credit to finance purchases of durable goods (for example, cars, household appliances) is to be given to a sample of families in a large metropolitan area.
- A random sample of $n = 5{,}000$ addresses is selected from the list of $N = 872{,}000$ "Delivery Sequence File" addresses for the metropolitan area purchased from a vendor of the United States Postal Service.
- The addresses are in order by zip code (approximately 15,000 addresses each), carrier route, and delivery sequence (the sequence through a carrier route followed to deliver mail).

Potential frame problems and solutions

- Noncoverage: families at addresses where the USPS does not deliver mail (for example, address uses a PO box only).
  - Remedy: adjust the estimates through weighting to compensate for non-coverage.
- Blanks: addresses that are not residential units, and do not contain families.
  - Remedy: skip the non-residential units, and increase the sample number of addresses slightly to account for blank listings.
- Clustering: addresses might be clusters of families (e.g., an apartment building with one address).
  - Remedy: Select all families at an address.

Summary: When to use SRS?

- SRS is the simplest of all probability sampling methods: Objective, randomized selection
- Every element has an equal chance of selection, called as $epsem$, or self-weighting
- However, the practical use of SRS is rare as it requires
  1. A list of population observation units
  2. Little extra information is available
  3. Analysis assuming independent observations and using SRS formulas
  4. High cost
- Even "bad" samples have an equal chance of being selected