BIOS 662 Fall 2018

Survey Sampling, Part II

David Couper, Ph.D.

david_couper@unc.edu

or

couper@bios.unc.edu

https://sakai.unc.edu/portal

Outline

- Stratified sampling
  - Introduction
  - Notation and estimands
  - Estimators
  - Allocation strategies
  - Example

Stratified Sampling

- *Stratification*: The process of dividing a population of units into distinct sub-populations called strata
- Strata are formed so that each population unit is assigned to only one stratum
- To draw a sample of US counties, we might stratify by region (NE, SE, NW, SW, ...)
- How is stratification used in sample surveys?

Stratified Sampling

- The population is divided into H strata so that each population unit is a member of only one stratum.
- Let N_h denote the number of population units in stratum h for $h = 1, \ldots, H$.
- Thus the total number of units in the population is

$$N = \sum_{h=1}^{H} N_h$$

- Let n_h denote the sample size for stratum h, so that the total sample size is

$$n = \sum_{h=1}^{H} n_h$$

Stratified Sampling

- A sample of size n_h is selected by some probability design (e.g., SRS) from each of the H strata, independent of each other
- Stratum-specific parameters (e.g., means, totals) are estimated separately using data from each of the H strata
- An estimate of the population parameter is produced by appropriately combining the H individual stratum estimates
- If SRS is used within stratum, this is called *stratified random sampling*

Notation and Estimands

- Let y_{hi} denote the variable of interest associated with unit i of stratum h $(i = 1, \ldots, N_h; \; h = 1, \ldots, H)$
- Let $Z_{hi} = 1$ if the corresponding unit is in the sample, 0 otherwise
- Stratum total

$$\tau_h = \sum_{i=1}^{N_h} y_{hi}$$

- Population total

$$\tau = \sum_{h=1}^{H} \tau_h = \sum_{h=1}^{H} \sum_{i=1}^{N_h} y_{hi}$$

Notation and Estimands

- Stratum mean

$$\mu_h = \frac{\tau_h}{N_h} = \frac{\sum_{i=1}^{N_h} y_{hi}}{N_h}$$

- Population mean

$$\mu = \frac{\tau}{N} = \frac{\sum_h \sum_i y_{hi}}{N} = \sum_h \frac{N_h}{N} \cdot \frac{1}{N_h} \sum_i y_{hi} = \sum_h W_h \mu_h$$

where $W_h = N_h/N$ is the proportion of population units in stratum h

Population Mean Estimator

- Estimator of the population mean

$$\bar{y} = \sum_h W_h \bar{y}_h$$

where $\bar{y}_h$ is an estimator of the mean μ_h for stratum h

- $E(\bar{y}_h) = \mu_h$ implies $E(\bar{y}) = \mu$

- Estimator of the variance of $\bar{y}$

$$\widehat{\text{Var}}(\bar{y}) = \sum_h W_h^2 \widehat{\text{Var}}(\bar{y}_h)$$

- $E(\widehat{\text{Var}}(\bar{y}_h)) = \text{Var}(\bar{y}_h)$ implies $E(\widehat{\text{Var}}(\bar{y})) = \text{Var}(\bar{y})$

Population Mean Estimator

- If stratified random sampling, then

$$\bar{y}_h = \frac{\sum_i y_{hi} Z_{hi}}{n_h}$$

and

$$\widehat{\mathrm{Var}}(\bar{y}) = \sum_h W_h^2 \left(\frac{1 - f_h}{n_h} \right) s_h^2$$

where $f_h = n_h / N_h$ is the stratum-specific sampling rate and s_h^2 is the within-stratum sample variance

Population Mean Estimator

- CIs

$$\bar{y} \pm t_{1-\alpha/2,df}\sqrt{\widehat{\mathrm{Var}}(\bar{y})}$$

where

$$df = \frac{(\sum_h a_h s_h^2)^2}{\sum_h (a_h s_h^2)^2/(n_h - 1)}$$

and

$$a_h = N_h(N_h - n_h)/n_h$$

- If all the N_h are equal and all the n_h are equal, then

$$df = n - H$$

Population Total Estimator

- Estimator of population total

$$\hat{\tau} = N\bar{y} = \sum_h N_h \bar{y}_h$$

- $E(\bar{y}_h) = \mu_h$ implies $E(\hat{\tau}) = \tau$

- Estimator of variance

$$\widehat{\text{Var}}(\hat{\tau}) = N^2 \widehat{\text{Var}}(\bar{y}) = \sum_h N_h^2 \left(\frac{1 - f_h}{n_h} \right) s_h^2$$

with the second equality assuming stratified random sampling

Population Total Estimator

- $E(\widehat{\mathrm{Var}}(\bar{y}_h)) = \mathrm{Var}(\bar{y}_h)$ implies $E(\widehat{\mathrm{Var}}(\hat{\tau})) = \mathrm{Var}(\hat{\tau})$
- CIs

$$\hat{\tau} \pm t_{1-\alpha/2,df} \sqrt{\widehat{\mathrm{Var}}(\hat{\tau})}$$

where df is as specified on the previous page

Population Total Proportion

- Estimator of population proportion

$$\hat{p} = \sum_h W_h \hat{p}_h$$

where the $\hat{p}_h$ are the stratum-specific estimators;

- $\hat{p}$ is a special case of $\bar{y}$
- Estimator of variance for stratified random sampling

$$\widehat{\mathrm{Var}}(\hat{p}) = \sum_h W_h^2 \left(\frac{1 - f_h}{n_h - 1}\right) \hat{p}_h(1 - \hat{p}_h)$$

Stratification Principle

- Variances depend on within-stratum population variance terms only
- Thus estimators will be more precise the smaller
$$\sigma_h^2 = \sum_i (y_{hi} - \mu_h)^2 / (N_h - 1)$$
- That is, estimation of the population mean or total will be most precise if the population is partitioned into strata in such a way that *within each stratum, the units are as similar as possible*
- For example, in a survey of a plant or animal population, the study area might be stratified into regions of similar habitat or elevation, because we expect abundancies to be more similar within strata than between strata

Stratification Principle: Example

- Suppose $N = 6$; $H = 2$; $N_h = 3$ for $h = 1, 2$
- Stratum 1 values: $0, 1, 2$; stratum 2 values: $4, 5, 9$
- Population variance $\sigma^2 = 10.7$
- Stratum variances $\sigma_1^2 = 1$, $\sigma_2^2 = 7$
- For SRS with $n = 4$,

$$\text{Var}(\bar{y}) = \left(1 - \frac{n}{N}\right)\frac{\sigma^2}{n} = \left(1 - \frac{4}{6}\right)\frac{10.7}{4} = 0.89$$

- For stratified random sampling with $n_1 = n_2 = 2$,

$$\text{Var}(\bar{y}) = \left(\frac{3}{6}\right)^2 \left(\frac{1 - 2/3}{2}\right) 1 + \left(\frac{3}{6}\right)^2 \left(\frac{1 - 2/3}{2}\right) 7 = 0.33$$

Allocation Strategies

- How to choose the sample size n_h for each stratum?
- Four strategies
  - Proportionate: same sampling rates
  - Optimum: most cost efficient
  - Balanced: equal sample sizes
  - Disproportionate: unequal sampling rates (to oversample important domains)

Proportionate Stratified Sampling

- Same sampling rate f_h for all strata:

$$f_h = \frac{n_h}{N_h} = \frac{n}{N} = f$$

- Equivalently

$$W_h = \frac{N_h}{N} = \frac{n_h}{n} = w_h$$

- The proportion of the sample chosen from any given stratum will be the same as the proportion of the population in that stratum

Proportionate Stratified Sampling

- Each unit in the population has the same probability of selection
- This type of design is called a *self-weighting design* because sample estimates of population mean and proportion are simple arithmetic means
- For example, the population mean estimator for proportionate stratified random sampling is

$$\bar{y} = \frac{1}{n} \sum_{h=1}^{H} \sum_{i=1}^{N_H} y_{hi} Z_{hi}$$

Stratification Principle

- Claim: The variance of estimators from proportionate stratified random sampling are always less than or equal to the variance of estimators from SRS

- Sketch of proof (see Cochran 1977, pages 99-100)

$$\begin{aligned}(N-1)\sigma^2 &= \sum_h \sum_i (y_{hi}-\mu)^2 \\ &= \sum_h \sum_i (y_{hi}-\mu_h)^2 + \sum_h N_h(\mu_h-\mu)^2 \\ &= \sum_h (N_h-1)\sigma_h^2 + \sum_h N_h(\mu_h-\mu)^2\end{aligned}$$

implying

$$\sigma^2 \approx \sum_h W_h\sigma_h^2 + \sum_h W_h(\mu_h-\mu)^2$$

Stratification Principle

- Thus

$$\begin{aligned}\mathrm{Var}_{\mathrm{SRS}}(\bar{y}) &= (1-f)\,\sigma^2/n \\ &\approx (1-f)\sum_h W_h \sigma_h^2/n \\ &\quad + (1-f)\sum_h W_h(\mu_h-\mu)^2/n\end{aligned}$$

- Under proportionate stratified random sampling

$$\mathrm{Var}_{\mathrm{pro}}(\bar{y}) = (1-f)\sum_h W_h^2\sigma_h^2/n_h = (1-f)\sum_h W_h\sigma_h^2/n$$

- Therefore

$$\mathrm{Var}_{\mathrm{SRS}}(\bar{y}) \approx \mathrm{Var}_{\mathrm{pro}}(\bar{y}) + (1-f)\sum_h W_h(\mu_h-\mu)^2/n$$

Stratification Principle

- For this reason, proportionate stratified random sampling is often considered the default
- Gains over SRS will be greatest if strata are internally homogeneous

General Guidelines

- The stratification variable should be highly correlated with the principal characteristic being measured in the survey (e.g., age may be a good stratification variable if we are doing a survey on limitation due to chronic illness)
- Strata should be internally homogeneous
- The variance of a population estimate will be smallest (for fixed cost) when each stratum sampling rate is directly related to the variability of units within the stratum and inversely related to the unit cost of data collection in the stratum
- Proportionate stratified sampling is always "safe" in that precision will never be worse than for SRS

General Advantages

- Improved precision of estimates (i.e., smaller variances),which leads to narrower confidence intervals
- Better control of sample sizes for sub-populations which can be defined by strata and for which separate estimates may be sought
- Sampling designs can be made more flexible
- For example, special strata may be established to handle segments of the population that are more difficult to survey (e.g., transient populations in household surveys)

Note of Caution

- Several stratum allocations yield very close to the optimum allocation
- Excessive attempts to determine the actual optimum allocation is almost never cost-effective.

Example of Analysis

- A county with two relatively large communities wants to do a survey on certification of the emergency medical technicians (EMTs) who work in the county and who are required to take special training and pass a competency exam for periodic certification
- Most EMTs work in "City A," which is relatively large and is located in the main urban area of the county; "City B" is smaller and has fewer EMTs; and the rest of the county's EMTs work in smaller towns and in rural areas
- Because of suspected similarities in certification patterns among EMTs in City A and comparable similarities in City B, we decide to divide the county into three strata, "City A", "City B" and "Other"

Example

- We want to estimate:
  - μ, the average number of hours of certification training in the year prior to the last certification
  - μ_1, the average number of hours of certification training in City A in the year prior to the last certification
  - τ, the total number of certification hours for EMTs in the county for the year prior to the last certification.
  - p, the proportion of EMTs who passed their last periodic certification exam on the first attempt

Example

- Use proportionately allocated sample sizes

h	Stratum Composition	N_h	n_h
1	City A	155	20
2	City B	62	8
3	Rural Area	93	12
Total		310	40

Example Data

Stratum 1						Stratum 2			Stratum 3		
i	Hours	Passed	i	Hours	Passed	i	Hours	Passed			
1	35	1	11	29	0	1	27	1	1	8	1
2	28	1	12	31	1	2	4	0	2	15	0
3	26	1	13	39	1	3	49	0	3	21	1
4	41	1	14	38	0	4	10	1	4	7	0
5	43	1	15	40	0	5	15	0	5	14	1
6	29	0	16	45	1	6	41	0	6	30	1
7	32	1	17	28	1	7	25	0	7	20	0
8	37	1	18	27	1	8	30	0	8	11	0
9	36	1	19	35	1				9	12	1
10	25	1	20	34	1				10	32	0
									11	34	0
									12	24	1

Example

- Summary statistics

Stratum 1	Stratum 2	Stratum 3
$n_1 = 20$	$n_2 = 8$	$n_3 = 12$
$N_1 = 155$	$N_2 = 62$	$N_3 = 93$
$W_1 = 0.5$	$W_2 = 0.2$	$W_3 = 0.3$
$f_1 = 0.129$	$f_2 = 0.129$	$f_3 = 0.129$
$\bar{y}_1 = 33.900$	$\bar{y}_2 = 25.125$	$\bar{y}_3 = 19.000$
$\hat{p}_1 = 0.8$	$\hat{p}_2 = 0.25$	$\hat{p}_3 = 0.50$
$s_1^2 = 35.358$	$s_2^2 = 232.411$	$s_3^2 = 87.636$

Example

- We want to estimate μ, the average number of hours of certification training in the year prior to the last certification.

- Estimate

$$\bar{y} = \sum_h W_h \bar{y}_h$$

$$= 0.5 \times 33.900 + 0.2 \times 25.125 + 0.3 \times 19.000$$

$$= 27.675 \text{ hours}$$

Example

- Estimated variance

$$
\begin{aligned}
\widehat{\mathrm{Var}}(\bar{y}) &= \sum_h W_h^2 \left(\frac{1-f_h}{n_h}\right) s_h^2 \\
&= 0.5^2 \left(\frac{1-0.129}{20}\right) 35.358 \\
&\qquad + 0.2^2 \left(\frac{1-0.129}{8}\right) 232.411 \\
&\qquad\qquad + 0.3^2 \left(\frac{1-0.129}{12}\right) 87.636 \\
&= 1.97
\end{aligned}
$$

Example

- Estimated standard error $\sqrt{1.97} = 1.40$
- 95% CI using *df* formula

$$27.675 \pm t_{0.975,21.1} 1.4034 = (24.757, 30.593)$$

- SAS uses $df = n - H = 37$

$$27.675 \pm t_{0.975,37} 1.4034 = (24.831, 30.519)$$

Example in SAS

```
data all;
  input stratum id hours pass;
 cards;
1  1 35 1
1  2 28 1
1  3 26 1
.  .  . .
.  .  . .
.  .  . .
3 11 34 0
3 12 24 1
;

data total;
 input stratum _TOTAL_;
 cards;
 1 155
 2 62
 3 93
;
```

Example in SAS

```
proc surveymeans data=all total=total;
  var hours;
  strata stratum;
```

```
                        The SURVEYMEANS Procedure

                              Data Summary

                  Number of Strata                        3
                  Number of Observations                 40

                               Statistics

                                           Std Error
  Variable         N            Mean         of Mean        95% CL for Mean
  ---------------------------------------------------------------------------
  hours           40       27.675000        1.403396    24.8314503 30.5185497
  ---------------------------------------------------------------------------
```