BST 542
Sampling Theory and Survey Design in Public Health

Steven E. Rigdon

Spring 2019

Chapter 3
Stratified Random Sampling

What Is Stratified Sampling?

How to Draw a Stratified Random Sample

1. Divide the population into nonoverlapping groups, called *strata*.
2. Take a SRS from each stratum.
3. Combine the data in an appropriate way to estimate the population parameters.

Why Take a Stratified Sample?

- Margin of error may be smaller for a fixed total sample size.
- Cost per observation may be less
- Estimates can be obtained for each stratum that have a desired margin of error.

How can a stratified sample produce estimates with a smaller margin of error?

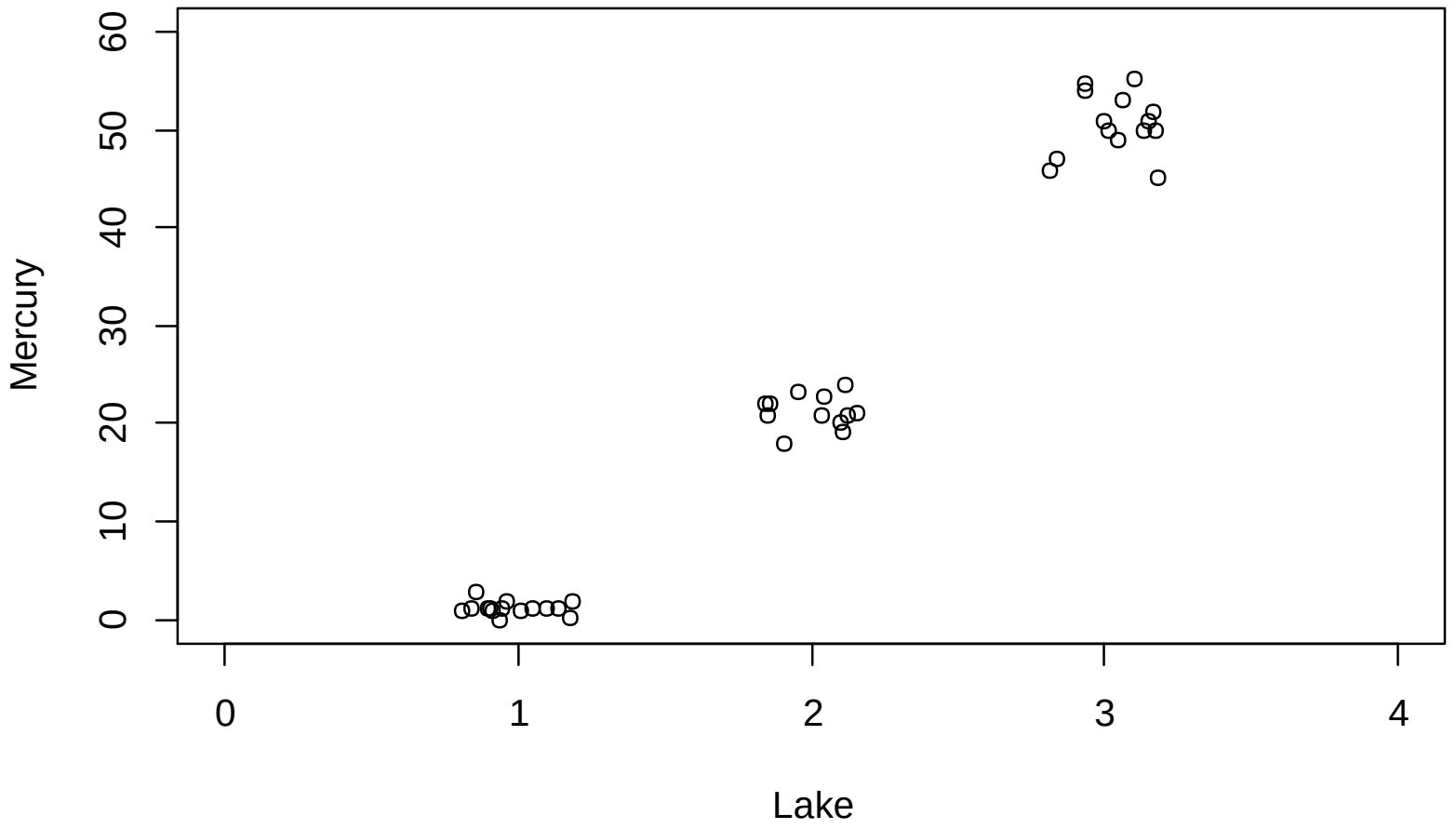
Suppose the population looks like this …

Mercury levels in fish in lakes:

Lake 1: 2, 1, 0, 1, 1, 1, 2, 1, 1, 3, 0, 1, 1, 1, 1

Lake 2: 23, 21, 20, 19, 22, 21, 21, 23, 18, 22, 21, 24

Lake 3: 45, 51, 55, 51, 52, 49, 50, 47, 53, 54, 55, 50, 50, 46


Mercury
0
10
20
30
40
50
60
0
1
2
3
4
Lake

Notation for Stratified Sampling: Population Parameters

H number of strata

N_h number of sampling units in stratum h (assumed known)

$N = N_1 + N_2 + \cdots + N_H$ (total number of sampling units in population)

y_{hj} = value of j-th unit in stratum h

$\mu_h =$ mean of <u>all</u> units in stratum h (book calls this $\bar{y}_{hU}$)

$\tau_h = N_h \bar{y}_{hU} =$ total of all units in stratum h (book calls this t_h)

$\sigma_h^2 =$ variance from stratum h (book calls this S_h^2)

$\mu =$ population mean

$\tau =$ population total

Notation for Stratified Sampling: Sample Data

$\mathcal{S}_h =$ the sample from stratum h

$n_h =$ sample size from stratum h

$\bar{y}_h =$ average from sample from stratum $h = \frac{1}{n_h}\sum_{j\in\mathcal{S}_h} y_{hj}$

$s_h^2 =$ sample variance from stratum $h = \frac{1}{n_h-1}\sum_{j\in\mathcal{S}_h}\left(y_{hj} - \bar{y}_h\right)^2$

$\hat{\tau}_h =$ estimated total from sample from stratum $h = N_h\bar{y}_h$

Estimating μ and τ

$$\hat{\mu}_h = \bar{y}_h$$

$$\hat{\sigma}_h^2 = s_h^2$$

$$\hat{\mu} = \bar{y}_{\text{str}} = \frac{1}{N}\left[N_1\bar{y}_1 + N_2\bar{y}_2 + \cdots + N_H\,\bar{y}_H\right]$$

$$\hat{\tau} = \sum_{h=1}^{H}\hat{\tau}_h = \sum_{h=1}^{H} N_h\bar{y}_h$$

$$\hat{V}(\hat{\mu}) = \frac{1}{N^2}\left[N_1^2\left(1-\frac{n_1}{N_1}\right)\frac{s_1^2}{n_1} + \cdots + N_L^2\left(1-\frac{n_H}{N_H}\right)\frac{s_H^2}{n_H}\right]$$

$$\hat{V}(\hat{\tau}) = N_1^2\left(1-\frac{n_1}{N_1}\right)\frac{s_1^2}{n_1} + \cdots + N_L^2\left(1-\frac{n_H}{N_H}\right)\frac{s_H^2}{n_H}$$

Estimating Proportions with Stratified Sampling

$\hat{p}_h =$ proportion in stratum h

Use formulas for estimating means, with $\hat{p}_h$ in place of $\bar{y}_h$ and $\frac{n_h}{n_h - 1}\hat{p}_h(1 - \hat{p}_h)$ in place of s_h^2.

$$\hat{p}_{\text{str}} = \sum_{h=1}^{H} \frac{N_h}{N}\hat{p}_h$$

$$\hat{V}(\hat{p}_{\text{str}}) = \sum_{h=1}^{H} \left(1 - \frac{n_h}{N_h}\right)\left(\frac{N_h}{N}\right)^2 \frac{\hat{p}_h(1 - \hat{p}_h)}{n_h - 1}$$

Example, p. 81. Am. Council of Learned Soc.

Estimate proportion female in ACLS

Discipline	Membership	Sample	Percent Female
Literature	9100	636	38
Classics	1950	451	27
Philosophy	5500	481	18
History	10850	611	19
Linguistics	2100	493	36
Political Science	5500	575	13
Sociology	9000	588	26
TOTAL	44000	3835	

Sampling Weights in Stratified Sampling

Estimate of total is

$$\hat{\tau} = \sum_{h=1}^{H} N_h \bar{y}_h = \sum_{h=1}^{H} \sum_{j \in S_h} \frac{N_h}{n_h} y_{hj}$$

$$w_{hj} = \frac{N_h}{n_h} = \frac{1}{\pi_{hj}}$$

Example

- Population: 1600 men, 400 women
- Stratified Sample of 200 men and 200 women.
- Calculate sampling weights.
- $w_{hj} = \frac{1600}{200} = 8$ for men; $w_{hj} = \frac{400}{200} = 2$ for each woman
- Each man in sample “represents” 8 men; each woman in sample “represents” 2 women.

agstrat.csv

- Farm acreage across stratified sample from population of 3078 counties.
- Look at sampling weights

Stratum = Region	# Counties in Stratum	# Counties in Sample
Northeast	220	21
North Central	1054	103
South	1382	135
West	422	41
TOTAL	3078	300

Allocating Observations to Strata

How do you select sample sizes for each stratum?

1. Proportional Allocation

$\pi_{hj} = \frac{n_h}{N_h}$ is the same for all strata

Want a stratified random sample of 200 from a population of 1600 Men, 400 Women.

$N = 2000, n = 200$, so $\frac{n}{N} = \frac{200}{2000} = 0.10$. Take $\frac{n_h}{N_h} = 0.10$ for both men and women. $n_h = 160$ for men and $n_h = 40$ for women.

Allocating Observations to Strata
How do you select sample sizes for each stratum?

2. Optimal Allocation

$S_h^2 =$ variance in stratum h (assumed known !?!)

$c_h =$ cost per observation in stratum h (assumed known)

Optimal allocation: minimize variance for fixed cost or minimize cost for fixed variance

$$n_h = \left(\frac{\frac{N_h S_h}{\sqrt{c_h}}}{\sum_{l=1}^{H} \frac{N_l S_l}{\sqrt{c_l}}} \right) \times n = A_h \, n$$

Allocating Observations to Strata
How do you select sample sizes for each stratum?

2. Optimal Allocation

$$n_h = \left(\frac{\frac{N_h S_h}{\sqrt{c_h}}}{\sum_{l=1}^{H} \frac{N_l S_l}{\sqrt{c_l}}} \right) \times n$$

Take a larger stratum sample when

The stratum size is large.

The variance within the stratum is large.

It is inexpensive to sample in the stratum.

Allocating Observations to Strata
How do you select sample sizes for each stratum?

3. Neyman Allocation

    Optimal allocation with equal costs across strata.

Allocating Observations to Strata
How do you select sample sizes for each stratum?

4. Allocation for Specified Precision within Strata

    Want a specified margin of error for each stratum.

    Use methods for SRS

Required Sample Size for Estimating the Mean or Total with a MOE of e

$$n = \frac{\sum_{h=1}^{H} \frac{N_h^2 S_h^2}{A_h}}{N^2 D + \sum_{h=1}^{H} N_h S_h^2}$$

Not in Lohr's book, but important!

$D = \frac{e^2}{4}$ when estimating μ

$D = \frac{e^2}{4N^2}$ when estimating τ

Example

Estimate TV viewing time in county.

County consists of Town A, Town B, Rural

We want MOE = 2 hours

$S_1 = 5, \;\; S_2 = 15, \;\; S_3 = 10$

$N_1 = 155, \;\; N_2 = 62, \;\; N_3 = 93$

Suppose we allocate equal sample sizes, so $A_1 = A_2 = A_3 = \frac{1}{3}$

We want to estimate the population mean with MOE = 2

$$D = \frac{e^2}{4} = \frac{2^2}{4} = 1$$

Example

$$\sum_{h=1}^{H} \frac{N_i^2 S_i^2}{A_i} = \frac{155^2 \times 5^2}{1/3} + \frac{62^2 \times 15^2}{1/3} + \frac{93^2 \times 10^2}{1/3} = 6{,}991{,}275$$

$$\sum_{h=1}^{H} N_i S_i^2 = 155 \times 5^2 + 62 \times 15^2 + 93 \times 10^2 = 27{,}125$$

$$n = \frac{6{,}991{,}275}{310 \times 1 + 27{,}125} = 56.7 \nearrow 57$$

Round up to a multiple of 3.

Take $n_1 = nA_1 = 19,$ $n_2 = nA_2 = 19,$ $n_3 = nA_3 = 19$

Sample Size Required Using Optimal Allocation

$$n = \frac{\left(\sum_{h=1}^{H} \frac{N_h \sigma_h}{\sqrt{c_h}}\right)\left(\sum_{l=1}^{H} N_l \sigma_l \sqrt{c_l}\right)}{N^2 D + \sum_{h=1}^{H} N_h \sigma_h^2}$$

$D = \frac{e^2}{4}$ when estimating μ

$D = \frac{e^2}{4N^2}$ when estimating τ

Same Example – Use Optimal Allocation

Estimate TV viewing time in county.

County consists of Town A, Town B, Rural

We want MOE = 2 hours

Stratum	Stratum Size N_h	Standard Dev. S_h	Cost per Obs. c_h
Town A	155	5	9
Town B	62	15	9
Rural	93	10	16
TOTAL	310		

$$n = \frac{\left(\sum_{h=1}^{H} \frac{N_h S_h}{\sqrt{c_h}}\right)\left(\sum_{h=1}^{H} N_h S_h \sqrt{c_h}\right)}{N^2 D + \sum_{h=1}^{L} N_h S_h^2} = \frac{800.83 \times 8835}{310^2 \times 1 + 27{,}125} = 57.42$$

$$\sum_{h=1}^{H} \frac{N_h \sigma_h}{\sqrt{c_h}} = \frac{N_1 S_1}{\sqrt{c_1}} + \frac{N_2 S_2}{\sqrt{c_2}} + \frac{N_3 S_3}{\sqrt{c_3}} = \frac{155 \times 5}{\sqrt{9}} + \frac{62 \times 15}{\sqrt{9}} + \frac{93 \times 10}{\sqrt{16}} = 800.83$$

$$\sum_{h=1}^{H} N_h S_h \sqrt{c_h} = 155 \times 5\sqrt{9} + 62 \times 15\sqrt{9} + 93 \times 10\sqrt{16} = 8835$$

$$D = \frac{e^2}{4} = 1$$

$$\sum_{h=1}^{H} N_h S_h^2 = N_1 S_1^2 + N_2 S_2^2 + N_3 S_3^2 = 155 \times 25 + 62 \times 225 + 93 \times 100 = 27{,}125$$

Example Allocation

$$A_1 = \frac{\frac{N_1 S_1}{\sqrt{c_1}}}{\frac{N_1 S_1}{\sqrt{c_1}} + \frac{N_2 S_2}{\sqrt{c_2}} + \frac{N_3 S_3}{\sqrt{c_3}}} = \frac{\frac{155 \times 5}{\sqrt{9}}}{800.3} = 0.323$$

$$A_2 = \frac{\frac{N_2 S_2}{\sqrt{c_2}}}{\frac{N_1 S_1}{\sqrt{c_1}} + \frac{N_2 S_2}{\sqrt{c_2}} + \frac{N_3 S_3}{\sqrt{c_3}}} = \frac{\frac{62 \times 15}{\sqrt{9}}}{800.3} = 0.387$$

$$A_3 = \frac{\frac{N_3 S_3}{\sqrt{c_3}}}{\frac{N_1 S_1}{\sqrt{c_1}} + \frac{N_2 S_2}{\sqrt{c_2}} + \frac{N_3 S_3}{\sqrt{c_3}}} = \frac{\frac{93 \times 10}{\sqrt{16}}}{800.3} = 0.291$$

Example Sample Sizes

$$n = 58$$

$$n_1 = 0.323 \times 58 = 18.7 \nearrow 19$$

$$n_2 = 0.387 \times 58 = 22.45 \nearrow 23$$

$$n_3 = 0.290 \times 58 = 16.8 \nearrow 17$$

When costs are the same for each stratum we have Neyman allocation.

$$n_h = n \frac{\frac{N_h S_h}{\sqrt{c_h}}}{\frac{N_1 S_1}{\sqrt{c_1}} + \frac{N_2 S_2}{\sqrt{c_2}} + \cdots + \frac{N_H S_H}{\sqrt{c_H}}} = n \frac{N_h S_h}{N_1 S_1 + N_2 S_2 + \cdots + N_H S_H}$$

$$n = \frac{\left(\sum_{h=1}^{H} \frac{N_h S_h}{\sqrt{c_h}}\right)\left(\sum_{l=1}^{H} N_l S_l \sqrt{c_l}\right)}{N^2 D + \sum_{h=1}^{H} N_h S_h^2} = \frac{\left(\sum_{h=1}^{H} N_h S_h\right)^2}{N^2 D + \sum_{h=1}^{H} N_h S_h^2}$$

What if costs *and* variances are the same?

Use white board to derive result.

Problem

$$c_1 = \$1, \quad c_2 = \$1, \quad c_3 = \$4$$

$$S_1 = 10, \quad S_2 = 15, \quad S_3 = 10$$

$$N_1 = 1000, \quad N_2 = 1000, \quad N_3 = 1000$$

Budget is $400. Find optimal allocation and sample sizes.