

5 Stratified Random Sampling

5.1 Introduction

Definition 5.1: A *stratified random sample* is a sample obtained by separating the population elements into nonoverlapping groups, called *strata*, and then selecting a simple random sample from each stratum.

Often, the population sizes of the strata are known, and the sample sizes of the strata are nonrandom.

REASONS FOR USING STRATIFIED SAMPLING:

⊙ Stratified sampling helps protect from the possibility of obtaining a really bad sample. How?

⊙ We may want data of known precision for subgroups.

Example: McIlwee and Robinson (1992, *Women in Engineering: Gender, Power, and Workplace Culture*) sampled graduates from engineering programs at public universities in Southern California. They were interested in comparing the educational and workforce experiences of male and female graduates.

How do you think that the authors conducted their survey?

⊙ A stratified sample may be more convenient to administer and may result in a lower cost for the survey.

Example: In a survey of businesses, one might stratify according to the size of the firm.

Why is stratification convenient here?

- ⊙ Properly performed stratified sampling typically will give more precise (having lower variance) estimators, because the variance *within* each stratum is often *lower* than the variance in the whole population. This lower variance results in a smaller bound, B , on the error of estimation.

Example: How should we sample if we are interested in estimating mean blood pressure?

Example: How should we sample if we are interested in studying the concentrations of plants in an area?

5.2 How to Draw a Stratified Random Sample

- ⊙ Typically, different strata tend to have different characteristics.

Example: How should strata be selected when trying to determine the winner of a Presidential election?

□

- ⊙ First, clearly define the strata.

Example: Suppose that the strata consist of *urban* households and *rural* households.

Is a medium-size city considered *urban* or *rural*?

□

Notation:

- ⊙ L = number of strata
- ⊙ N_i = number of sampling units in stratum i
- ⊙ N = number of sampling units in the population = $N_1 + N_2 + \dots + N_L$

5.3 Estimation of a Population Mean, μ , and Total, τ

Determine a reasonable estimator of the population mean, μ .

We estimate the mean of the i th stratum by the sample mean, \bar{Y}_i , of the stratum, for $i = 1, \dots, L$.

How much *weight* should be given to each sample mean, \bar{Y}_i , of the stratum, when estimating the population mean, μ ?

Suppose stratum #1 represents 20% of the population. How much weight should be given to \bar{Y}_1 when estimating μ ?

The *stratified sample mean*, denoted by \bar{Y}_{st} , is the *weighted* average of the sample means, \bar{Y}_i , of the strata.

$$\bar{Y}_{\text{st}} = \frac{1}{N} \sum_{i=1}^L N_i \bar{Y}_i$$

Determine a reasonable estimator of the VARIANCE of \bar{Y}_{st} .

Recall: If X and U are independent random variables, then $\text{var}(X + U) = \text{var}(X) + \text{var}(U)$.

Recall: If b is a constant, then $\text{var}(bX) = b^2 \text{var}(X)$.

Recall: For *simple random sampling* withOUT replacement,

$$\hat{V}(\hat{\mu}) = \hat{V}(\bar{Y}) = \frac{s^2}{n} \left(\frac{N-n}{N} \right)$$

Thus, for *stratified random sampling*,

$$\hat{V}(\bar{Y}_{\text{st}}) = \frac{1}{N^2} \sum_{i=1}^L N_i^2 \left(\frac{N_i - n_i}{N_i} \right) \frac{s_i^2}{n_i}$$

Error of estimation is $|\bar{Y}_{\text{st}} - \mu|$.

Bound on the error of estimation is $2 \sqrt{\hat{V}(\bar{Y}_{\text{st}})}$.

Sample size allocation is discussed in section 5.5. For the next example, we sample proportionally to stratum size.

Exercise #5.6, p. 160: A school desires to estimate the average score that may be obtained on a reading exam for students in the sixth grade. The students are grouped into three tracks according to ability. Track I is the fast track, and track III is the slow track. The population sizes of these three tracks are 55, 80, and 65. The sample sizes are 14, 20, and 16.

(a) Estimate the average score, μ , for the sixth grade.

```
> # Read in the data.
```

```
> y.mat = scan2( "EXER5_6.DAT", T )
```

```
> score1 = as.numeric( y.mat[1:14, 2] )
```

```
> # We need the sample means, sample sizes, and population sizes.
```

```
> n = c(14, 20, 16)
```

> N = c(55, 80, 65)

> N.total = sum(N)

(a') Place a bound on the error of estimation.

$$B = 2 \sqrt{\frac{1}{N^2} \sum_{i=1}^L N_i^2 \left(\frac{N_i - n_i}{N_i} \right) \frac{s_i^2}{n_i}}$$

> # We need to compute the sample variances.

(b) Construct parallel box plots for the data, and comment on the patterns you see.

(c) Estimate the difference in average scores between track I and track II students. Are track I students significantly better, on the average, than track II students?

> # Compute the bound on the error of estimation (recall section 4.6, pp. 96-99).

$$\hat{V}(\bar{Y}_1 - \bar{Y}_2) = \hat{V}(\bar{Y}_1) + \hat{V}(\bar{Y}_2) = \frac{s_1^2}{n_1} \left(\frac{N_1 - n_1}{N_1} \right) + \frac{s_2^2}{n_2} \left(\frac{N_2 - n_2}{N_2} \right)$$

$$B = 2 \sqrt{\hat{V}(\bar{Y}_1 - \bar{Y}_2)}$$

□

Estimate τ , the population total, for stratified sampling.

$$\hat{\tau}_{\text{st}} = N\bar{Y}_{\text{st}} = \sum_{i=1}^L N_i \bar{Y}_i$$

A reasonable estimate of the variance of $\hat{\tau}$ is

$$\hat{V}(\hat{\tau}_{\text{st}}) = \hat{V}(N \bar{Y}_{\text{st}}) = \sum_{i=1}^L N_i^2 \left(\frac{N_i - n_i}{N_i} \right) \frac{s_i^2}{n_i}$$

Exercise 5.1, p. 171: Data on the population of the United States are given in Appendix D, pp. 446-447, and on the data disk under USPOP. The goal is to estimate the total U.S. population in the 18-24 age group from a sample of states. The states are divided into four geographic regions. Using these regions as strata, select an appropriately sized stratified random sample of states, and use their data on population in the 18- to 24-year-old group to estimate the total U.S. population in that age group. Because the total population is available from the data on all the states, check to see if your estimate is within the margin of error you established for your estimate.

```
> # Scan in the data, withOUT retyping the data.  
> uspop = scan2( "USPOP.DAT", T, T )  
> region = uspop[ , 2 ]  
> pop1 = uspop[ region == 1, 3 ]
```

```
> # Estimate of population total.
```

```
> # Estimate the variance of tau.hat.
```

```
> var.hat = sum( N^2 * (N-n) / N * y.vars / n )
```

```
> # Place a bound on the error of estimation.
```

```
# What is the true population total?
```

□

5.4 Selecting the Sample Size for Estimating Population Means, μ , and Totals, τ

- ⊙ L strata.
- ⊙ Each stratum has its own variance, σ_i^2 , $i = 1, \dots, L$.

- ⊙ *Goal:* Determine the total sample size, n , needed to achieve a particular bound B on the error of estimation with 95% confidence.
- ⊙ Decide in advance what proportion of the sample should be taken from stratum #1, stratum #2, and so on.
- ⊙ Hence, decide in advance $a_1 = (n_1/n)$, $a_2 = (n_2/n)$, \dots , $a_L = (n_L/n)$.

- ⊙ What is $\sum_{i=1}^L a_i$?

These **allocation fractions** may be selected in a variety of ways. One example is **proportional allocation**.

- ⊙ What is the the **proportional allocation** of the required sample size?
- ⊙ The approximate sample size required to estimate μ or τ with a bound B on the error of estimation is:

$$(5.6) \quad n = \frac{\sum_{i=1}^L N_i^2 \sigma_i^2 / a_i}{N^2 D + \sum_{i=1}^L N_i \sigma_i^2},$$

where $D = B^2/4$ when estimating μ , and $D = B^2/(4N^2)$ when estimating τ .

5.5 Allocation of the Sample

Suppose we want to estimate the population mean, μ , or total, τ .

Once the overall sample size, n , has been selected, how should we allocate the individual sample sizes, n_i , for the different strata?

The answer depends on N_i (population size of the stratum), σ_i (the variability within the stratum), and c_i (the cost of sampling from the stratum).

Example: A university has 6,000 *on-campus* students and 4,000 *off-campus* students.

Suppose we want to sample students to estimate the average weekly alcohol consumption among all students at this university.

How should we allocate our sample?

- (a) Suppose that it is just as easy (and cheap) to sample *on-campus* students as it is to sample *off-campus* students, and the variability (i.e., σ_1) of alcohol consumption among *on-campus* students is the same as the variability (i.e., σ_2) of alcohol consumption among *off-campus* students.

How should we allocate our sample of size, say, $n = 100$?

(b) Now suppose (based on a preliminary sample) that *off*-campus students have greater variability (i.e., σ_2) in alcohol consumption than *on*-campus students do.

Again, how should we allocate our sample of size, say, $n = 100$?

(c) Next suppose that sampling is door-to-door, and that sampling *off*-campus students is more expensive (i.e., time-consuming) than sampling *on*-campus students.

How should we allocate our sample size, based on our limited resources (of time or money)?

□

Approximate allocation that minimizes cost for a fixed value of $V(\bar{Y}_{st})$ or minimizes $V(\bar{Y}_{st})$ for a fixed cost:

$$(5.7) \quad n_i = n \left(\frac{N_i \sigma_i / \sqrt{c_i}}{\sum_{k=1}^L N_k \sigma_k / \sqrt{c_k}} \right)$$

Another perspective: Fix B (the bound on the error of estimation) and determine the *entire* sample size, n , needed to minimize the cost.

$$(5.8) \quad n = \frac{\left(\sum_{k=1}^L N_k \sigma_k / \sqrt{c_k} \right) \left(\sum_{i=1}^L N_i \sigma_i \sqrt{c_i} \right)}{N^2 D + \sum_{i=1}^L N_i \sigma_i^2},$$

where $D = B^2/4$ when estimating μ , and $D = B^2/(4N^2)$ when estimating τ .

Then, the individual sample sizes, n_i , of each stratum are determined by the earlier formula (5.7) on n_i .

These two above formulas (for n and n_i) produce the **optimal allocation** for minimizing cost when B is fixed.

Special case: Neyman allocation occurs when the costs are the same for all strata under optimal allocation.

Then, the above formulas reduce to the following:

$$(5.9) \quad n_i = n \left(\frac{N_i \sigma_i}{\sum_{k=1}^L N_k \sigma_k} \right)$$

$$(5.10) \quad n = \frac{\left(\sum_{i=1}^L N_i \sigma_i \right)^2}{N^2 D + \sum_{i=1}^L N_i \sigma_i^2},$$

where $D = B^2/4$ when estimating μ , and $D = B^2/(4N^2)$ when estimating τ .

5.6 Estimation of a Population Proportion, p

Example: Suppose a university has 65% *female* students and 35% *male* students.

A sample of *female* students is taken, and a sample of *male* students is taken.

The students are asked: “Should students be permitted to carry firearms on campus?”

We want to estimate p , the population proportion of students at this university who would answer “yes.”

Suppose (among the sample of *female* students), $\hat{p}_1 = 22\%$.

Suppose (among the sample of *male* students), $\hat{p}_2 = 38\%$.

How should we estimate p ?

How much *weight* should be given to each sample proportion, \hat{p}_i , of the stratum, when estimating the population proportion, p ?

The *stratified sample proportion*, denoted by \hat{p}_{st} , is the *weighted* average of the sample proportions, \hat{p}_i , of the strata.

$$\hat{p}_{\text{st}} = \frac{1}{N} \sum_{i=1}^L N_i \hat{p}_i$$

Determine a reasonable estimator of the VARIANCE of \hat{p}_{st} .

Is the sample of *female* students independent of the sample of *male* students?

Recall: For *simple random sampling* withOUT replacement,

$$\hat{V}(\hat{p}) = \left(\frac{\hat{p}\hat{q}}{n-1} \right) \left(\frac{N-n}{N} \right)$$

Thus, for *stratified random sampling*,

$$\hat{V}(\hat{p}_{\text{st}}) = \frac{1}{N^2} \sum_{i=1}^L N_i^2 \left(\frac{N_i - n_i}{N_i} \right) \left(\frac{\hat{p}_i \hat{q}_i}{n_i - 1} \right)$$

Error of estimation is $|\hat{p}_{\text{st}} - p|$.

Bound on the error of estimation is $2 \sqrt{\hat{V}(\hat{p}_{\text{st}})}$.

Note in our above example, we had not yet specified the sample sizes or populations sizes of the strata.

Suppose $N_1 = 6500$ and $N_2 = 3500$.

Suppose the sample sizes were $n_1 = 200$ and $n_2 = 150$.

Determine the estimated variance of \hat{p}_{st} and the bound on the error of estimation.

Example: Continue with the example involving estimating the population proportion of students who support allowing students to carry firearms on campus.

> # Estimate of the population proportion, p .

Determine the estimated variance of \hat{p}_{st} .

Determine B , the bound on the error of estimation when using \hat{p}_{st} to estimate p .

Determine a 95% confidence interval on p .

□

5.7 Selecting the Sample Size, n , and Allocating the Sample, n_i , to Estimate Proportions, p

Recall: A *proportion* is a special case of a *mean*, when the original data have what numerical values?

Recall the formulas in sections 5.4 and 5.5 for *selecting the sample size, n , and allocating the sample, n_i , to estimate a MEAN, μ .*

These formulas contain σ_i , for $i = 1, \dots, L$.

Suppose the data consist of only *zeros* and *ones*, corresponding to *failures* and *successes*.

Then, the data represent Bernoulli(p) random variables.

What is the *mean*, μ , of a Bernoulli(p) random variable?

What is the *standard deviation* of a Bernoulli(p) random variable?

The approximate sample size required to estimate p with a bound B on the error of estimation is:

$$(5.15) \quad n = \frac{\sum_{i=1}^L N_i^2 p_i q_i / a_i}{N^2 D + \sum_{i=1}^L N_i p_i q_i},$$

where a_i is the fraction of observations allocated to stratum i , p_i is the population proportion for stratum i , and $D = B^2/4$.

The approximate allocation that minimizes cost for a fixed value of $V(\hat{p}_{\text{st}})$ or minimizes $V(\hat{p}_{\text{st}})$ for a fixed cost:

$$(5.16) \quad n_i = n \left(\frac{N_i \sqrt{p_i q_i / c_i}}{\sum_{k=1}^L N_k \sqrt{p_k q_k / c_k}} \right),$$

for $i = 1, \dots, L$

What is the drawback to using the above formulas?

Special case: **Neyman allocation**

When is $p_i q_i$ (the standard deviation of the stratum) maximized?

Summary: The following are the steps to obtain optimal allocation with fixed B when estimating p :

- ⊙ First, select B , the bound on the error of estimation.
- ⊙ Use equation (5.16) to obtain the ratios n_i/n .
- ⊙ Set the allocation fractions $a_i = n_i/n$.
- ⊙ Use equation (5.15) to obtain n .
- ⊙ Finally, let $n_i = a_i n$, for $i = 1, \dots, L$.

Example (revisit): Suppose a university has 65% *female* students and 35% *male* students.

A sample of *female* students is taken, and a sample of *male* students is taken.

We want to estimate p , the population proportion of students who support allowing students to carry firearms on campus.

Based on the preliminary sample of 200 *female* students, $\hat{p}_1 = 22\%$.

Based on the preliminary sample of 150 *male* students, $\hat{p}_2 = 38\%$.

Suppose $N_1 = 6500$ and $N_2 = 3500$.

Determine the appropriate sample sizes under Neyman allocation using $B = 3\%$.

(Recall in our preliminary sample that $B = 4.64\%$.)

(a) Use equation (5.16) (with equal costs c_i) to determine n_i/n , which will be set equal to a_i .

$$n_i/n = \left(\frac{N_i \sqrt{p_i q_i}}{\sum_{k=1}^L N_k \sqrt{p_k q_k}} \right)$$

(b) Next, use equation (5.15) (with equal costs c_i) to determine n , based on your values of a_i .

$$n = \frac{\sum_{i=1}^L N_i^2 p_i q_i / a_i}{N^2 D + \sum_{i=1}^L N_i p_i q_i},$$

(c) Next, determine the sample sizes needed for each stratum.

$$n_i = a_i n$$

□

5.8 Additional Comments on Stratified Sampling

Compare *stratified* sampling to *simple random* sampling.

Example (revisit from section 5.7): Suppose a university has 65% *female* students and 35% *male* students.

Based on a preliminary sample of 200 females and 150 males, a reasonable estimate of p , the population proportion of students who support allowing students to carry firearms on campus, is 27.6%.

To estimate p using a **stratified** sample with $B = 3\%$, we found in section 5.7 to sample a total of about 789 students.

How many students should be sampled in a **simple random** sample, also with $B = 3\%$?

Using 27.6% as an initial estimate of p , apply equation (4.18) on p. 94.

$$n = \frac{Npq}{(N-1)D + pq},$$

where $D = B^2/4$

□

Exercise 5.22, p. 164: When does stratification produce large gains in precision over simple random sampling? (Assume costs of observations are constant under both designs.)

□

5.10 Stratification After Selection of the Sample

Example: Suppose a sampling frame lists all households in an area, and you would like to estimate the average monthly household food bill (μ). One desirable stratification variable might be household size. Why?

From U.S. Census data, the distribution of household sizes in the region might be known:

Number of persons in Household	Percentage of Households
1	26
2	31
3	18
4	15
5+	10
total	100

However, the sampling frame does not include information on household size.

The sampling frame lists only the households.

First take a simple random sample of, say, 100 households.

Suppose we obtain the following sample sizes:

$$n_1 = 21, n_2 = 36, n_3 = 19, n_4 = 11, n_5 = 13$$

Suppose the sample mean monthly food bills for the five strata are the following:

$$\bar{Y}_1 = \$270, \bar{Y}_2 = \$420, \bar{Y}_3 = \$560, \bar{Y}_4 = \$690, \bar{Y}_5 = \$810$$

How should we estimate μ ?

□

Formulas for **poststratification** (i.e., stratify AFTER the sample is taken) when estimating μ (p. 151):

Let $A_i = N_i/N$, for $i = 1, \dots, L$.

$$\bar{Y}_{\text{st}} = \sum_{i=1}^L A_i \bar{Y}_i$$

$$(5.18) \quad \hat{V}_p(\bar{Y}_{st}) = \frac{N-n}{Nn} \sum_{i=1}^L A_i s_i^2 + \frac{1}{n^2} \sum_{i=1}^L (1-A_i) s_i^2$$

Example (revisit): Estimate the average monthly household food bill (μ).

Suppose $s_1 = \$150$, $s_2 = \$270$, $s_3 = \$390$, $s_4 = \$470$, and $s_5 = \$560$.

Suppose that this community has 10,000 households.

Determine the poststratified sample mean, the estimated variance of the estimator, the bound on the error of estimation, and a confidence interval on μ .

Interpretation: We are 95% confident that the true population average monthly household food bill (μ) is between \$415 and \$557.

□

5.11 Double Sampling for Stratification

Previously in this chapter, the populations sizes (N_i) of the strata were known. How?

Now, suppose the N_i are unknown, but stratification is still desired.

Suppose you want to administer a lengthy (i.e., expensive!) questionnaire, and you want to stratify by race/ethnicity.

Hence, your sample sizes n_i likely will be small.

Procedure:

- ⊙ Take a large sample size, n' , of the entire population.

This first sample is called the *phase 1 sample*.

- ⊙ From this phase 1 sample, ask the respondent about his or her *race*.
- ⊙ The phase 1 sample sizes (based on race) are n'_1, n'_2, \dots, n'_L .

What is your best estimate of the allocation fractions $N_1/N, N_2/N, \dots, N_L/N$?

- ⊙ For each stratum, subsample the phase 1 sample.
- ⊙ This subsample, which is of size n_i , is called the phase 2 sample.
- ⊙ From this subsample, administer the lengthy questionnaire to collect data.
- ⊙ Based on these data, compute the sample mean, \bar{Y}_i , and sample variance, s_i^2 , for each subsample of the stratum.
- ⊙ This sampling method is called **DOUBLE SAMPLING** or **TWO-PHASE SAMPLING**. □

What is a reasonable estimator of the mean, μ , under stratification (but NOT double sampling)?

Therefore, under double sampling, a reasonable estimator of μ is:

$$(5.22) \quad \bar{Y}'_{st} = \sum_{i=1}^L a'_i \bar{Y}_i$$

An estimator of the variance of \bar{Y}'_{st} is:

$$(5.23) \quad \hat{V}(\bar{Y}'_{st}) = \frac{n'}{n' - 1} \sum_{i=1}^L \left[\left(a_i'^2 - \frac{a'_i}{n'} \right) \frac{s_i^2}{n_i} + \frac{a'_i (\bar{Y}_i - \bar{Y}'_{st})^2}{n'} \right]$$

How should formulas (5.22) and (5.23) change if we are estimating τ rather than μ ?

How should formulas (5.22) and (5.23) change if we are estimating p rather than μ ?

Exercise 5.29, p. 165: A question on a proposed annexation is to be asked of a suburban area, but responses for registered voters could be quite different from those who are not registered. Of 1000 residents of the area who were telephoned, 80% were registered voters. Ten percent of each group (registered and nonregistered) were asked to complete a follow-up questionnaire, on which one question was “Do you favor annexation into the city?” The data are summarized as follows:

	Voters	Nonvoters
n_i	80	20
y_i	60	8

(a) Estimate the proportion of residents who will respond yes to the question of interest.

$$\hat{p}'_{\text{st}} = \sum_{i=1}^L a'_i \hat{p}_i$$

(b) Estimate the variance of \hat{p}'_{st} .

$$\hat{V}(\hat{p}'_{\text{st}}) = \frac{n'}{n' - 1} \sum_{i=1}^L \left[\left(a_i'^2 - \frac{a'_i}{n'} \right) \frac{\hat{p}_i \hat{q}_i}{n_i - 1} + \frac{a'_i (\hat{p}_i - \hat{p}'_{\text{st}})^2}{n'} \right]$$

(c) Estimate B , the bound on the error of estimation of \hat{p}'_{st} .

(d) Construct a 95% confidence interval on p .

□

DOUBLE SAMPLING FOR NONRESPONSE

Example: Traugott (1987, *Public Opinion Quarterly*) analyzed callback data from two 1984 Michigan polls on preference for presidential candidates. The overall response rates for the surveys were about 65%. However, only about 21% of the interviewed sample responded on the first call. Up to 30 attempts were made to reach persons who did not respond on the first call. The later respondents were more likely to be male, older, and Republican than early respondents.

48% of the respondents who answered on the first call supported Reagan, and 45% supported Mondale.

However, 59% of the entire sample supported Reagan, and 39% supported Mondale.

Do the nonrespondents resemble the hard-to-reach (*late*) respondents?

□

However, we might use the *late* respondents to estimate the opinions of the *nonrespondents* in some situations.

Example: Suppose in a telephone interview (regarding political affiliation) in a sample of size $n' = 4000$, we have that 1000 people responded on the first call.

Keeping in the mind the concept of double sampling, what are the two strata?

How do we estimate the allocation fractions?

For our second phase, how many of the 1000 **early** respondents do we sample?

Suppose $\hat{p}_1 = 0.45$, the sample proportion of Republicans among **early** respondents.

Since these **initial nonrespondents** are perhaps more difficult to sample, let us subsample only, say, 800 of these initial nonrespondents.

Among these 800 initial nonrespondents, eventually $n_2 = 500$ of them respond.

These 500 **late** respondents represent all of the original 3000 nonrespondents.

Suppose $\hat{p}_2 = 0.6$, the sample proportion of Republicans among **late** respondents.

(a) How should we estimate p ?

(b) Estimate the variance of \hat{p}'_{st} .

$$\hat{V}(\hat{p}'_{\text{st}}) = \frac{n'}{n' - 1} \sum_{i=1}^L \left[\left(a_i'^2 - \frac{a_i'}{n'} \right) \frac{\hat{p}_i \hat{q}_i}{n_i - 1} + \frac{a_i' (\hat{p}_i - \hat{p}'_{\text{st}})^2}{n'} \right]$$

(c) Estimate B , the bound on the error of estimation of \hat{p}'_{st} .

(d) Construct a 95% confidence interval on p .

□

Example: The **U.S. Census Bureau** tries to enumerate as many people as possible in the decennial census. However, some people are missed, causing population estimates from the census to underestimate the true population count. The undercount is thought to be greater for inner-city areas and minority groups and varies among different regions of the United States. Congressional representatives, billions of dollars of federal funding, and other resources are apportioned based on census results.

For the year 2000 census, a panel of the National Academy of Sciences recommended double sampling. How?

□