Lecture 3: Stratified Sampling

Maochao Xu

Department of Mathematics
Illinois State University
mxu2@ilstu.edu

What is Stratified Sampling?

Suppose the population is partitioned into disjoint sets of sampling units called **strata**. If a sample is selected within each stratum, then this sampling procedure is known as **stratified sampling**.

Example: The Federal Deposit Insurance Corporation (FDIC) insures deposits at member banks up to a specified limit. When a bank fails, the FDIC acquires the assets from that bank and uses them to pay the insured depositors. Valuing the assets is time-consuming, so the FDIC selects a sample of the assets in order to estimate the total amount recovered from financial institutions.
Assets from failed institutions fall into several types:

- consumer loans
- commercial loans
- securities
- real estate mortgages
- other owned real estate
- other assets
- net investments in subsidiaries

A simple random sample (SRS) of assets may result in an **imprecise estimate** of the total amount recovered (why?).

why stratified sampling?

- Protected from the possibility of obtaining a really bad sample.
- Want data of known precision for subgroups of the population.

  Ex: Interested in comparing the educational and workforce experiences of male and female graduates. Because there were many more male than female graduates, they sampled a higher fraction of female graduates than male graduates in order to obtain comparable precisions for the two groups.
- A stratified sample may be more convenient to administer and may result in a lower cost for the survey.

  Ex: In a survey of businesses, an Internet survey might be used for large firms while a mail or telephone survey is used for small firms.
- Stratified sampling often gives more precise (having lower variance) estimates for population means and totals.

  Ex: Persons of different ages tend to have different blood pressures, so in a blood pressure study it would be helpful to stratify by age groups.

Example: Number of farm acres per county

Strata: Northeast, North Central, South, and West.

Stratum	Number of Counties in Stratum	Number of Counties in Sample
Northeast	220	21
North Central	1054	103
South	1382	135
West	422	41
Total	3078	300

Region	Sample Size	Average	Variance
Northeast	21	97,629.8	7,647,472,708
North Central	103	300,504.2	29,618,183,543
South	135	211,315.0	53,587,487,856
West	41	662,295.5	396,185,950,266

The following table gives estimates of the total number of farm acres and estimated variance of the total for each of the four strata:

Stratum	Estimated Total of Farm Acres	Estimated Variance of Total
Northeast	21,478,558.2	1.59432×10^{13}
North Central	316,731,379.4	2.88232×10^{14}
South	292,037,390.8	6.84076×10^{14}
West	279,488,706.1	1.55365×10^{15}
Total	909,736,034.4	2.5419×10^{15}

For comparison, the estimate of the population total, using an SRS of size 300, was 916,927,110, with standard error 58,169,381. For this example, stratified sampling ensures that each region of the United States is represented in the sample, and produces an estimate with smaller standard error than an SRS with the same number of observations. The sample variance in Example 2.5 was $s^2 = 1.1872 \times 10^{11}$. Only the West had sample variance larger than s^2; the sample variance in the Northeast was only 7.647×10^9.

Observations within many strata tend to be more homogeneous than observations in the population as a whole, and the reduction in variance in the individual strata often leads to a reduced variance for the population estimate. In this example, the relative gain from stratification can be estimated by the ratio

$$\frac{\text{estimated variance from stratified sample, with } n = 300}{\text{estimated variance from SRS, with } n = 300} = \frac{2.5419 \times 10^{15}}{3.3837 \times 10^{15}} = 0.75.$$

If these figures were the population variances, we would expect that we would need only (300)(0.75) = 225 observations with a stratified sample to obtain the same precision as from an SRS of 300 observations.

Theory

Notation for Stratification: The population quantities are:

$$y_{hj} = \text{ value of } j\text{th unit in stratum } h$$

$$t_h = \sum_{j=1}^{N_h} y_{hj} = \text{ population total in stratum } h$$

$$t = \sum_{h=1}^{H} t_h = \text{ population total}$$

$$\bar{y}_{hU} = \frac{\sum_{j=1}^{N_h} y_{hj}}{N_h} = \text{ population mean in stratum } h$$

$$\bar{y}_U = \frac{t}{N} = \frac{\sum_{h=1}^{H}\sum_{j=1}^{N_h} y_{hj}}{N} = \text{ overall population mean}$$

$$S_h^2 = \sum_{i=1}^{N_h} \frac{(y_{hj} - \bar{y}_{hU})^2}{N_h - 1} = \text{ population variance in stratum } h$$

Theory

Using SRS estimators within each stratum, are:

$$\bar{y}_h = \frac{1}{n_h} \sum_{j \in \mathcal{S}_h} y_{hj}$$

$$\hat{t}_h = \frac{N_h}{n_h} \sum_{j \in \mathcal{S}_h} y_{hj} = N_h \bar{y}_h$$

$$s_h^2 = \sum_{j \in \mathcal{S}_h} \frac{(y_{hj} - \bar{y}_h)^2}{n_h - 1}$$

Suppose we only sampled the hth stratum. In effect, we have a population of N_h units and take an SRS of n_h units. Then we would estimate $\bar{y}_{hU}$ by $\bar{y}_h$, and t_h by $\hat{t}_h = N_h \bar{y}_h$. The population total is $t = \sum_{h=1}^{H} t_h$, so we estimate t by

$$\hat{t}_{\text{str}} = \sum_{h=1}^{H} \hat{t}_h = \sum_{h=1}^{H} N_h \bar{y}_h.$$

To estimate $\bar{y}_U$, then, we use

$$\bar{y}_{\text{str}} = \frac{\hat{t}_{\text{str}}}{N} = \sum_{h=1}^{H} \frac{N_h}{N} \bar{y}_h.$$

This is a weighted average of the sample stratum averages; $\bar{y}_h$ is multiplied by N_h/N, the proportion of the population units in stratum h. To use stratified sampling, the sizes or relative sizes of the strata must be known.

Theory

- **Unbiasedness.** $\bar{y}_{\text{str}}$ and $\hat{t}_{\text{str}}$ are unbiased estimators of $\bar{y}_U$ and t.

$$E\left[\sum_{h=1}^{H}\frac{N_h}{N}\bar{y}_h\right]=\sum_{h=1}^{H}\frac{N_h}{N}E[\bar{y}_h]=\sum_{h=1}^{H}\frac{N_h}{N}\bar{y}_{hU}=\bar{y}_U.$$

- **Variance of the estimators.** Since we are sampling independently from the strata, and we know $V(\hat{t}_h)$ from the SRS theory,

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

- **Standard errors for stratified samples.**

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

$$\hat{V}(\bar{y}_{\text{str}})=\frac{1}{N^2}\hat{V}(\hat{t}_{\text{str}})=\sum_{h=1}^{H}\left(1-\frac{n_h}{N_h}\right)\left(\frac{N_h}{N}\right)^2\frac{s_h^2}{n_h}.$$

As always, the standard error of an estimator is the square root of the estimated variance: $\text{SE}(\bar{y}_{\text{str}})=\sqrt{\hat{V}(\bar{y}_{\text{str}})}$.

- **Confidence intervals for stratified samples.** If either (1) the sample sizes within each stratum are large, or (2) the sampling design has a large number of strata, an approximate $100(1-\alpha)\%$ confidence interval (CI) for the population mean $\bar{y}_U$ is

$$\bar{y}_{\text{str}}\pm z_{\alpha/2}\,\text{SE}\,(\bar{y}_{\text{str}}).$$

Proportion

- sample proportion and variance for the stratum

$$\bar{y}_h = \hat{p}_h, \quad s_h^2 = \frac{n_h}{n_h - 1}\hat{p}_h(1 - \hat{p}_h).$$

- sample proportion

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

- variance of estimate

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

- Total

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

Example

Estimate the percentage and number of respondents of the major societies in those seven disciplines that are female.

Discipline	Membership	Number Mailed	Valid Returns	Female Members (%)
Literature	9,100	915	636	38
Classics	1,950	633	451	27
Philosophy	5,500	658	481	18
History	10,850	855	611	19
Linguistics	2,100	667	493	36
Political Science	5,500	833	575	13
Sociology	9,000	824	588	26
Totals	44,000	5,385	3,835	

Example

Here, let N_h be the membership figures, and let n_h be the number of valid surveys. Thus,

$$\hat{p}_{\text{str}} = \sum_{h=1}^{7} \frac{N_h}{N} \hat{p}_h = \frac{9100}{44{,}000} 0.38 + \ldots + \frac{9000}{44{,}000} 0.26 = 0.2465$$

and

$$\text{SE}(\hat{p}_{\text{str}}) = \sqrt{\sum_{h=1}^{7} \left(1 - \frac{n_h}{N_h}\right) \left(\frac{N_h}{N}\right)^2 \frac{\hat{p}_h(1-\hat{p}_h)}{n_h - 1}} = 0.0071.$$

The estimated total number of female members in the societies is $\hat{t}_{\text{str}} = 44{,}000 \times (0.2465) = 10{,}847$, with $\text{SE}(\hat{t}_{\text{str}}) = 44{,}000 \times (0.0071) = 312$. ∎

Sampling weights

The stratified sampling estimator $\hat{t}_{\text{str}}$ can be expressed as a weighted sum of the individual sampling units: ,

$$\hat{t}_{\text{str}} = \sum_{h=1}^{H} N_h \bar{y}_h = \sum_{h=1}^{H} \sum_{j \in \mathcal{S}_h} \frac{N_h}{n_h} y_{hj}.$$

The estimator of the population total in stratified sampling may thus be written as

$$\hat{t}_{\text{str}} = \sum_{h=1}^{H} \sum_{j \in \mathcal{S}_h} w_{hj} y_{hj},$$

where the sampling weight for unit j of stratum h is $w_{hj} = (N_h/n_h)$. The sampling weight can again be thought of as the number of units in the population represented by the sample member y_{hj}.

In a stratified random sample, the population mean is thus estimated by

$$\bar{y}_{\text{str}} = \frac{\sum_{h=1}^{H} \sum_{j \in \mathcal{S}_h} w_{hj} y_{hj}}{\sum_{h=1}^{H} \sum_{j \in \mathcal{S}_h} w_{hj}}.$$

Stratum	N_h	n_h	w_{hj}
A	400	98	4.08
B	30	10	3.00
C	61	37	1.65
D	18	6	3.00
E	70	39	1.79
F	120	21	5.71

Allocating Observations to Strata

- Proportional allocation: the number of sampled units in each stratum is proportional to the size of the stratum, the inclusion probability is the same for all strata

$$\pi_{hj} = \frac{n_h}{N_h} = \frac{n}{N}.$$

  If proportional allocation is used, each unit in the sample represents the same number of units in the population (**self-weighting**).

- ANOVA table

Source	df	Sum of Squares
Between strata	$H-1$	$\text{SSB} = \sum_{h=1}^{H}\sum_{j=1}^{N_h}(\bar{y}_{hU}-\bar{y}_U)^2 = \sum_{h=1}^{H} N_h(\bar{y}_{hU}-\bar{y}_U)^2$
Within strata	$N-H$	$\text{SSW} = \sum_{h=1}^{H}\sum_{j=1}^{N_h}(y_{hj}-\bar{y}_{hU})^2 = \sum_{h=1}^{H}(N_h-1)S_h^2$
Total, about $\bar{y}_U$	$N-1$	$\text{SSTO} = \sum_{h=1}^{H}\sum_{j=1}^{N_h}(y_{hj}-\bar{y}_U)^2 = (N-1)\,S^2$

Allocating Observations to Strata

$$
\begin{aligned}
V_{\text{prop}}(\hat{t}_{\text{str}}) &= \sum_{h=1}^{H} \left(1 - \frac{n_h}{N_h}\right) N_h^2 \frac{S_h^2}{n_h} \\
&= \left(1 - \frac{n}{N}\right) \frac{N}{n} \sum_{h=1}^{H} N_h S_h^2 \\
&= \left(1 - \frac{n}{N}\right) \frac{N}{n} \left(\text{SSW} + \sum_{h=1}^{H} S_h^2\right).
\end{aligned}
$$

The sums of squares add up, with SSTO = SSW + SSB, so the variance of the estimated population total from an SRS of size n is

$$
\begin{aligned}
V_{\text{SRS}}(\hat{t}) &= \left(1 - \frac{n}{N}\right) N^2 \frac{S^2}{n} \\
&= \left(1 - \frac{n}{N}\right) \frac{N^2}{n} \frac{\text{SSTO}}{N-1} \\
&= \left(1 - \frac{n}{N}\right) \frac{N^2}{n(N-1)} (\text{SSW} + \text{SSB}) \\
&= V_{\text{prop}}(\hat{t}_{\text{str}}) + \left(1 - \frac{n}{N}\right) \frac{N}{n(N-1)} \left[N(\text{SSB}) - \sum_{h=1}^{H} (N - N_h) S_h^2\right].
\end{aligned}
$$

Proportional allocation with stratification always gives smaller variance than SRS *unless*

$$
\text{SSB} < \sum_{h=1}^{H} \left(1 - \frac{N_h}{N}\right) S_h^2.
$$

Optimal Allocation

The objective in optimal allocation is to gain the most information for the least cost. We shall then sample heavily within a stratum if

- The stratum accounts for a large part of the population.
- The variance within the stratum is large; we sample more heavily to compensate for the heterogeneity.
- Sampling in the stratum is inexpensive.

For a fixed variance

- Case I: **For a fixed total cost C**, the smallest variance is achieved by choosing n_h such that:

$$n_h = \frac{(C - c_0) N_h S_h / \sqrt{c_h}}{\sum_{h=1}^{H} N_h S_h \sqrt{c_h}}$$

- Case II: **For a fixed variance**, the smallest cost is achieved by first determining the total sample size n such that

$$n = \frac{\left(\sum_{h=1}^{H} N_h S_h \sqrt{c_h}\right)\left(\sum_{h=1}^{H} N_h S_h / \sqrt{c_h}\right)}{N^2 V + \sum_{h=1}^{H} N_h S_h^2}$$

where V is the fixed variance specified by the researcher. Then, the stratum sample size n_h for $h = 1, 2, \ldots, H$ is

$$n_h = \frac{n N_h S_h / \sqrt{c_h}}{\sum_{h=1}^{H} N_h S_h / \sqrt{c_h}}$$

Example: case 1

Suppose there is a fixed total cost $C = \$3000$ and a fixed overhead cost of $c_0 = \$500$. Consider the stratification used in Figure 5b. The unit sampling costs are

$$\begin{aligned} c_1 &= \$20 \text{ per unit from stratum 1} \\ c_2 = c_3 &= \$25 \text{ per unit from stratum 2 or 3} \\ c_4 &= \$30 \text{ per unit from stratum 4} \\ c_5 = c_6 &= \$35 \text{ per unit from stratum 5 or 6} \\ c_7 &= \$40 \text{ per unit from stratum 7} \end{aligned}$$

Then, using s_h^2 as an estimate of S_h^2: $n_h = \dfrac{(C - c_0)N_h s_h/\sqrt{c_h}}{\sum_{h=1}^{H} N_h s_h \sqrt{c_h}} = \dfrac{2500\, N_h s_h/\sqrt{c_h}}{7657.776}$

Stratum	N_h	s_h	c_h	$N_h s_h \sqrt{c_h}$	$N_h s_h/\sqrt{c_h}$	rounded n_h	projected n_h	cost
1	45	3.215	20	647.006	32.350	10.6	11	\$220
2	60	3.847	25	1154.100	46.164	15.1	15	\$375
3	66	4.393	25	1449.690	57.988	18.9	19	\$475
4	58	2.062	30	655.054	21.835	7.1	7	\$210
5	66	2.881	35	1124.919	32.141	10.5	10	\$350
6	60	3.286	35	1166.414	33.326	10.9	11	\$385
7	45	5.132	40	1460.593	36.515	11.9	12	\$480
				7657.776	260.319		85	\$2495

The estimated total cost is $\$2495 + c_0 = \$2495 + \$500 = \$2995$ requiring 85 sampling units.

Example: case 2

Suppose there is a fixed variance of $V = V(\hat{t}) = .35$. The costs are the same as Case I. Then, using s_h as an estimate of S_h:

$$n = \frac{\left(\sum_{h=1}^{H} N_h S_h \sqrt{c_h}\right)\left(\sum_{h=1}^{H} N_h S_h / \sqrt{c_h}\right)}{N^2 V + \sum_{h=1}^{H} N_h S_h^2} \approx \frac{(7657.776)(260.319)}{(400^2)(.35) + 5254.1} = 37.433$$

Then, substitution yields

$$n_h = \frac{n N_h S_h / \sqrt{c_h}}{\sum_{h=1}^{H} N_h S_h / \sqrt{c_h}} = \frac{(37.433) N_h S_h / \sqrt{c_h}}{260.319} \approx .1438 N_h S_h / \sqrt{c_h}.$$

Stratum	N_h	s_h	c_h	$N_h s_h \sqrt{c_h}$	$N_h s_h / \sqrt{c_h}$	$N_h S_h^2$	rounded n_h	projected n_h	cost
1	45	3.215	20	647.006	32.350	465.0	4.65	5	\$100
2	60	3.847	25	1154.100	46.164	888.0	6.64	7	\$175
3	66	4.393	25	1449.690	57.988	1273.8	8.34	8	\$200
4	58	2.062	30	655.054	21.835	246.5	3.14	3	\$ 90
5	66	2.881	35	1124.919	32.141	547.8	4.62	5	\$175
6	60	3.286	35	1166.414	33.326	648.0	4.79	5	\$175
7	45	5.132	40	1460.593	36.515	1185.0	5.25	5	\$200
				7657.776	260.319	5254.1		38	\$1115

Thus, the minimum cost to achieve V is \$1115 + c_0 = \$1115 + \$500 = \$1615 requiring a total of $n = 38$ sampling units.

Quota Sampling

- **Quota sampling** is a form of stratified sampling and typically uses multifactor stratification. Taking a quota sample ensures that data are collected across the population with the belief that doing so will provide a representative sample from the population. It also allows the researcher to generate estimates related to various subgroups.
- Quota sampling is the primary method of sampling used by many commercial data- collection organizations. It is a common method used in political polls and surveys of consumer attitudes regarding products. It is also common in medical studies in which the researchers will select subjects satisfying the requirements for admittance to a study until a desired number is reached.
- So how does quota sampling differ from stratified simple random sampling? **In quota sampling, the within stratum samples may not be random.** Often some element of subjectivity enters into the sampling procedure.

A typical quota sample

- Defining the multifactor strata.
- Determining the stratum sample sizes based on proportional allocation.
- Data is collected by predetermined data collection techniques (e.g., phone surveys, mail surveys, personal interviews, etc.) until the stratum quotas are satisfied (that is, until the desired number of responses are collected for each stratum).

Although taking a quota sample can save a lot of time and money when compared to simple random sampling, the researcher must realize that if quota sampling is used, we **cannot be sure that the selection of sampling units would be similar to units collected via simple random sampling.**

Example: A student organization wants to determine if students favor extending the evening hours that the library remains open. They decided to take a quota sample based on strata related to class standing (freshmen, sophomores, juniors, seniors, graduate students). After deciding on the 5 quotas of 25 students for each of the 5 strata, data was collected at the library on consecutive evenings until all 5 quotas were satisfied.

Quota Sampling

- Because nonresponse is often ignored in quota sampling, the resulting estimates can be seriously biased.
- Quota sampling has proven useful if the quotas are designed properly with careful attention paid to when, how, and where the data are collected.