DATA 606: Statistical Methods in Data Science

—— Stratified sampling

Wenjun Jiang

Department of Mathematics & Statistics
The University of Calgary

Lecture 3

An example

The FDIC[1] was created by the U.S. Congress to supervise banks. When a bank fails, it acquires the assets from that bank and uses them to pay the insured depositors.

- Valuing assets is time-consuming, so FDIC selects a sample of assets.
- The assets contain several categories: *consumer loan, commercial loan, securities, real estate mortgage, other owned real estates*.
- The commercial loan has most monetary value. Suppose using a SRS, and the sample happen to contain no commercial loan.
- The resulting estimate of total asset value would be quite imprecise.

[1] Federal Deposit Insurance Corporation

Intro of stratified sampling

The reason: the variable we are interested in takes different mean values in different subpopulations.

Intro of stratified sampling

The reason: the variable we are interested in takes different mean values in different subpopulations.

- We want to be protected from the probability of selecting a "bad" sample.
- A stratified sample may be more convenient to administer and result in a lower cost.
- It gives more precise estimates for population means and totals.

Another example

- Goal: estimate the average number of farm acres per county.
- Source: 2012 census of agriculture of U.S.
- Use stratified sampling: use the four census regions of the United StatesNortheast, North Central, South, and Westas strata.
- Number of samples in each strata:

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

Another example

- Within-sample average and variance

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

- Estimated strata-level total and variance

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}

Another example

As comparison, we could do simple random sampling with 300 units in the sample.

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

Theory

- Divide population (with N units) into H strata ($h = 1, 2, \ldots, H$), each strata has N_i units such that

$$N_1 + N_2 + \cdots + N_H = N.$$

- Select n_h units from the strata h into the sample. The total sample size is

$$n = n_1 + n_2 + \cdots + n_H.$$

Theory

Notations for stratification:

- y_{hj}: the value of jth unit in stratum h.
- $t_h = \sum_{j=1}^{N_h} y_{hj}$: population total in stratum i.
- $t = \sum_{h=1}^{H} t_h$: population total.
- $\bar{y}_{hU} = \frac{t_h}{N_h}$: population mean of stratum h.
- $\bar{y}_U = \frac{t}{N}$: overall population mean.
- $V_h = \frac{\sum_{j=1}^{N_h}(y_{hj} - \bar{y}_{hU})^2}{N_h - 1}$: population variance in stratum h.

Theory

Estimators within stratum *h* *(S_h is the sample set taken from stratum h)*

- $\bar{y}_h = \frac{1}{n_h} \sum_{j \in S_h} y_{hj}$.
- $\hat{t}_h = N_h \cdot \bar{y}_h$.
- $v_h = \frac{\sum_{j \in S_h} (y_{hj} - \bar{y}_h)^2}{n_h - 1}$.

Overall population total and average estimators

- $\hat{t} = \sum_{h=1}^{H} \hat{t}_h$.
- $\hat{y}_U = \frac{\hat{t}}{N}$.

Theory

- **Unbiasedness**: as an SRS is taken from each stratum, $\mathbf{E}[\bar{y}_h] = \bar{y}_{hU}$, as such

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

- **Variance**: we know each sub-sample set S_h is taken from the stratum h independently, as such

$$\mathrm{Var}(\hat{t}) = \sum_{h=1}^{H} \mathrm{Var}(\hat{t}_h) = \sum_{h=1}^{H} \left(1 - \frac{n_h}{N_h}\right) N_h^2 \frac{V_h}{n_h}.$$

Theory

- **Estimate of variance**: as we do not know V_h, we replace it with v_h:

$$\hat{\mathrm{Var}}(\hat{t}) = \sum_{h=1}^{H} \left(1 - \frac{n_h}{N_h}\right) N_h^2 \frac{v_h}{n_h}.$$

$$\hat{\mathrm{Var}}(\hat{y}_U) = \frac{1}{N^2} \hat{\mathrm{Var}}(\hat{t}) = \sum_{h=1}^{H} \left(1 - \frac{n_h}{N_h}\right) \left(\frac{N_h}{N}\right)^2 \frac{v_h}{n_h}.$$

Allocating observations to strata

Proportional allocation

- You would like your sample to be a miniature version of the population.
- Using proportional allocation, each unit has probability $\pi_{hj} = \frac{n_h}{N_h} = \alpha$ to be selected into the sample.
- By using proportional allocation, say $n_h = \alpha \cdot N_h$, then

$$\frac{n_i}{N_i} = \alpha = \frac{\alpha(N_1 + \cdots + N_H)}{N_1 + \cdots + N_H} = \frac{n_1 + \cdots + n_H}{N_1 + \cdots + N_H} = \frac{n}{N}.$$

Population 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

Proportional allocation

With proportional allocation

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

Allocating observations to strata

Proportional allocation

With simple random sample (SRS)

$$\begin{aligned}
\mathrm{Var}_{SRS}(\hat{t}) &= \left(1-\frac{n}{N}\right) N^2 \frac{V}{n}, \\
&= \left(1-\frac{n}{N}\right) \frac{N^2}{n} \frac{SSTO}{N-1} \\
&= \left(1-\frac{n}{N}\right) \frac{N^2}{n(N-1)} (\mathrm{SSW}+\mathrm{SSB}) \\
&= \mathrm{Var}_{prop}(\hat{t}) + \left(1-\frac{n}{N}\right) \frac{N}{n(N-1)} \left[N\cdot \mathrm{SSB} - \sum_{h=1}^{H} (N-N_h) V_h \right]
\end{aligned}$$

$$\mathrm{Var}_{prop}(\hat{t}) < \mathrm{Var}_{SRS}(\hat{t}) \iff \mathrm{SSB} > \sum_{h=1}^{H} (1-\frac{N_h}{N}) V_h.$$

Sample size

With proportional allocation

- $\frac{n_h}{N_h} = \lambda$ and $\frac{n_h}{n} = \frac{n_h}{n_1+\cdot+n_H} = \frac{\alpha \cdot N_h}{\alpha(N_1+\cdots+N_H)} = \frac{N_h}{N}$.
- As the variance of $\hat{y}_U$ is

$$\sum_{h=1}^{H} \left(1 - \frac{n_h}{N_h}\right) (\frac{N_h}{N})^2 \frac{V_h}{n_h}$$

$$= \frac{1}{n}(1-\lambda) \sum_{h=1}^{H} (\frac{N_h}{N})^2 \frac{n}{n_h} V_h$$

$$= \frac{1}{n}(1-\lambda) \sum_{h=1}^{H} \frac{N_h}{N} V_h$$

$$= \frac{W}{n}.$$

Sample size (cont.)

- The $(1-\alpha)\%$ confidence interval for $\bar{y}_U$:

$$\left[\hat{y}_U - z_{\alpha/2}\sqrt{\frac{W}{n}}, \hat{y}_U + z_{\alpha/2}\sqrt{\frac{W}{n}}\right].$$

- In other words,

$$\mathbf{P}(|\bar{y}_U - \hat{y}_U| \leq z_{\alpha/2}\sqrt{\frac{W}{n}}) = 1-\alpha.$$

- Set an acceptable error e, then

$$e = z_{\alpha/2}\sqrt{\frac{W}{n}} \iff n = z^2_{\alpha/2}\frac{W}{e^2}.$$

Defining strata

As stratified sampling always gives more precise estimation than SRS, why would anyone not use it?

- Stratification adds complexity to the survey (with extra cost).
- The added complexity may not be worth a small gain in precision.

The principle of defining strata:

we want the subpopulation mean varies significantly from stratum to stratum. Define the strata **such that between-stratum variance is big, while within-stratum variance is small**.

Defining strata

Example 1 (A paradox)

our survey is to estimate total business expenditures on advertising, we would like to put businesses that spent the most on advertising in stratum 1, businesses with the next highest level of advertising expenditures in stratum 2, and so on, until the last stratum contained businesses that spent nothing on advertising. The problem with this scheme is that we do not know the advertising expenditures for all the businesses while designing the survey if we did, we would not need to do a survey at all!

Defining strata

- Most surveys measure more than one variable, so any stratification variable should be related to many characteristics of interest.

Example 2 (continue Example 1)

For estimating total business expenditures on advertising, we might stratify by number of employees or size of the business and by the type of product or service.

For farm income, we might use the size of the farm as a stratifying variable, since we expect that larger farms would have higher incomes

- A general rule to keep in mind is: *the more information, the more strata you should use*[2].

[2] you should use an SRS when little prior information about the target population is available