Stratified Sampling

SurvMeth/Surv 625: Applied Sampling

Yajuan Si

University of Michigan, Ann Arbor

1/22/25

Stratified sampling

- Implementation
- Inference
- Projection

Implementation

- Dividing our population of elements into subgroups (strata) using auxiliary information that is available *prior* to drawing the sample
- Simple random sampling of elements WITHIN each of the strata (or population subgroups): **Independent** across strata
- Need the auxiliary information on the frame to create mutually exclusive and exhaustive subgroups (strata)

1. Avoid selecting a really bad SRS sample
2. Desire precision for subgroups
3. More convenient to administer and may results in a lower survey cost
4. Often gives more precise estimates for population means and totals

Inference

- We can apply everything that we've learned about for SRS within each of the strata
- Stratum index: $h = 1, \dots, H$
- Denote the variable of interest for i-th element in stratum h as Y_{hi}
- For each population stratum, population mean $\bar{Y}_h = \sum_{i=1}^{N_h} Y_{hi}/N_h$ and element variance $S_h^2 = \frac{1}{N_h - 1} \sum_{i=1}^{N_h} (Y_{hi} - \bar{Y}_h)^2$
- For each stratum in the sample, we can compute $\bar{y}_h, s_h^2, t_h, N_h, n_h,$ etc., in addition to sampling variances, etc.; **all specific to h!**

Inference: Population mean

- We can rewrite the population mean as a weighted sum of the population means for each stratum, where the weight W_h is the relative proportion of the population within each stratum

$$\bar{Y} = \sum_h \frac{N_h}{N} \bar{Y}_h \doteq \sum_h W_h \bar{Y}_h$$

- We can write the sample mean in the same way, assuming that we have good (unbiased) estimates of the means in each stratum

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

Inference: Sampling weight

- Element-level weighting:

$$\bar{y}_w = \frac{\sum_{h=1}^{H} \sum_{i=1}^{n_h} w_{hi} y_{hi}}{\sum_{h=1}^{H} \sum_{i=1}^{n_h} w_{hi}}$$

- Here we have introduced survey weight w_{hi}, the sampling weight for unit i in stratum h
- **The sampling weight in often the reciprocal of the inclusion probability**: $w_{hi} = \frac{1}{\pi_{hi}}$, where π_{hi} is the inclusion probability of unit i in stratum h.
- For stratified sampling, $\pi_{hi} = n_h/N_h$, so we have $w_{hi} = N_h/n_h$ and $\sum_{h=1}^{H} \sum_{i=1}^{n_h} w_{hi} = N$.
- A stratified sample is self-weighting is the sampling fraction n_h/N_h is the same across strata, where the sampling weight for each observation is N/n, exactly the same as in SRS.

Inference: Population mean cont.

Within stratum h: SRS

- Sampling fraction: $f_h = n_h/N_h$
- Mean estimate: $\bar{y}_h$ [or p_h if proportion]
- Element variance estimate: $s_h^2 = \frac{1}{n_h-1}\sum_{i=1}^{n_h}(y_{hi} - \bar{y}_h)^2$ [or $s_h^2 = \frac{n_h}{n_h-1}p_h(1-p_h)$ if proportion]
- Sampling variance estimate: $var(\bar{y}_h) = (1-f_h)\frac{s_h^2}{n_h}$
- Standard error: $se(\bar{y}_h) = \sqrt{var(\bar{y}_h)}$

Inference: Population mean cont.

Combine across strata

The sampling variance of the overall estimated mean is entirely a function of the within-stratum sampling variances only!

$$
\begin{aligned}
var(\bar{y}_w) &= var(\sum_h W_h \bar{Y}_h) = \sum_h var(W_h \bar{Y}_h) \\
&= \sum_h W_h^2 var(\bar{y}_h) = \sum_h W_h^2 (1 - f_h) \frac{s_h^2}{n_h} \qquad (1)
\end{aligned}
$$

Example: Estimating the average number of farm acres per county

Use four U.S. census regions as strata to select counties

Region	# Counties in population	# Counties in sample	Sample mean in region	Sample variance in region
Northest	220	21	?	?
North Central	1054	103	?	?
South	1382	135	?	?
West	422	41	?	?
Total	3178	300		

Analysis of variance (ANOVA)

The sum of squares

$$
\begin{aligned}
\sum_{h=1}^{H}\sum_{i=1}^{N_h}(Y_{hi}-\bar{Y})^2 &= (N-1)S^2 \\
&= \sum_{h=1}^{H}\sum_{i=1}^{N_h}(Y_{hi}-\bar{Y}_h+\bar{Y}_h-\bar{Y})^2 \\
&= \sum_{h=1}^{H}\sum_{i=1}^{N_h}(Y_{hi}-\bar{Y}_h)^2+\sum_{h=1}^{H}\sum_{i=1}^{N_h}(\bar{Y}_h-\bar{Y})^2 \\
&= \sum_{h=1}^{H}(N_h-1)S_h^2+\sum_{h=1}^{H}N_h(\bar{Y}_h-\bar{Y})^2 \\
SSTO &\doteq SSW+SSB
\end{aligned}
$$

ANOVA cont.

We have the simplification as

$$
\begin{aligned}
S^2 &= \sum_{h=1}^{H} \frac{N_h - 1}{N - 1} S_h^2 + \sum_{h=1}^{H} \frac{N_h - 1}{N - 1} (\bar{Y}_h - \bar{Y})^2 \\
&\approx \sum_{h=1}^{H} W_h S_h^2 + \sum_{h=1}^{H} W_h (\bar{Y}_h - \bar{Y})^2 \qquad (2) \\
&= \text{Within-stratum variance} + \text{Between-stratum variance}
\end{aligned}
$$

- The overall S^2 is fixed; if we define strata such that the between-stratum variance component becomes large, the within-stratum variance will necessarily become smaller
- Hence the sampling variance will go down based on Equation (1), which only depends on the within-stratum variance
- Decrease the sampling variance of the mean by making strata heterogeneous between and homogeneous within

Projection

- Always stratify! We give ourselves the potential to reduce the variance of estimates (the same strata will be used in every sample, reducing variance in the estimates across hypothetical samples)
- Expect gains in precision over designs that include the between variance as well
- But gains are not guaranteed. The reductions in the variance of estimates depend on the allocation.
- How do we determine how many elements to sample from each stratum?

Allocation

1. Proportionate allocation: Representative sampling where the sample reflects the population with respect to the stratification variable, $n_h/n = N_h/N = W_h$. We have $\pi_{hi} = n_h/N_h = n/N$, i.e., epsem
2. Equal allocation: $n_h = n/H$, the same sample size across strata; minimize the sampling variance for comparisons $var(\bar{y}_h - \bar{y}_{h'})$
3. Neyman allocation: $n_h \propto W_h S_h$, giving the smallest sampling variance for $\bar{y}_w$
4. Optimum allocation: $n_h \propto W_h S_h / \sqrt{C_h}$, where C_h is the cost related to stratum h

Proportionate allocation

- epsem: $f_h = n_h/N_h = n/N$ and $n_h = nW_h$
- No weighting is needed for the mean $\bar{y}_w = t/n$
- Simplified sampling variance estimate: $var(\bar{y}_w) = \sum_h W_h^2 \frac{1-f_h}{n_h} s_h^2 = \frac{1-f}{n} \sum_h W_h s_h^2 = \frac{1-f}{n} s_w^2$
- Design effect:

$$deff = \frac{var(\bar{y}_w)}{var_{SRS}(\bar{y})} = \frac{\frac{1-f}{n} s_w^2}{\frac{1-f}{n} s^2} = \frac{s_w^2}{s^2} = 1 - \frac{\sum_h W_h (\bar{y}_h - \bar{y}_w)^2}{s^2}$$

- In general deff$<$1

Equal allocation

- Consider $n_h = n/H$ and $f_h = \frac{n/H}{N_h}$, not epsem unless strata are the same size
- Need weighted estimates
- Deff may actually be greater than 1
- Why use it?
  - Suppose $S_h = S_{h'}$ for two different strata $h \neq h'$
  - Equal allocation minimized the sampling variance $var(\bar{y}_h - \bar{y}_{h'})$

Neyman allocation

- Consider $n_h = kW_hS_h$ with $k = n/\sum_h W_hS_h$
- Smallest sampling variance $var(\bar{y}_w)$
- Gains in precison greater than proportionate allocation
- Need estimates of S_h
    - In practice, use reasonable estimates
    - Large gains require variation among S_h
    - Large gains unlikely for proportions
    - Values specific to Y
    - For multipurpose surveys, allocations will vary as S_h's vary across characteristics

Optimum allocation

- Consider a total fixed cost: $C = \sum_h n_h C_h$
- Minimize the sampling variance under total fixed cost
- Allocate $n_h = kW_h S_h / \sqrt{C_h}$ with $k = \frac{C}{\sum_h W_h S_h \sqrt{C_h}}$
- Neyman allocation is a special case where costs are the same across strata
- Can result in higher precision or lower costs with variation among S_h
- Resulting n_h can be larger than N_h, use N_h
- Values specific to Y

Allocation: Summary

	Proportionate	Equal	Neyman	Optimum
Goal	Representative	minimize the sampling variance for comparisons $var(\bar{y}_h - \bar{y}_{h'})$	Minimize the sampling variance $var(\bar{y}_w)$	Minimize the sampling variance under total fixed cost
n_h	nW_h	n/H	kW_hS_h	$kW_hS_h/\sqrt{C_h}$
epsem?	Yes	No	No	No
deff	< 1	unsure	< 1	< 1
Multi-purpose	All variables	All variables	Per variable	Per variable

Determining the total sample size

- Define a quantity $v = \sum_h \frac{n}{n_h}(\frac{N_h S_h}{N})^2$ as an "average" variability per unit in a stratified random sample with the specified allocation, similar to S^2 as the variability per unit in an SRS
- Ignoring all stratum fpcs, $n_0 = z^2_{\alpha/2} v / e^2$ is the required sample size to give the margin of error e
- It can also be calculated as $n_{SRS} v / S^2$, with n_{SRS} being the required SRS sample size
- If $v < S^2$, as in proportional allocation, stratified sampling allows a desired precision with a smaller sample size than SRS

Number of strata

- Stratification requires discrete categories
  - Stratifying variables may be discrete
  - Continuous stratifying variables divided into categories
- How many categories to capture gains possible?
  - Generally 3-6 strata adequate for a single predictor
  - When more than one stratifier, "coarser" cuts on more variables preferred to "finer" cuts
  - "Deepest" stratification for $n_h = 2$ (or $H = n/2$)
  - "Even deeper" stratification: 1 per stratum, a singleton problem

Paired selection

- Paired selections $n_h = 2$ useful in practice
- "Deepest" stratification possible that allows sampling variances to be estimated without assumptions
- Paired selection is epsem: $N_h = N/H$
- Attraction of paired selection is the simplification in variance estimation
- When the design is epsem, proportionately allocated

Paired selection estimation

- The mean is unweighted, and estimate variance under the proportionate allocation: $\bar{y} = \sum_h \sum_i y_{hi}/n = \frac{\sum_h (y_{h1}+y_{h2})}{n}$

$$
\begin{aligned}
var(\bar{y}) &= \frac{1-f}{n} \sum_h W_h s_h^2 (\text{ where } W_h = 2/n) \\
&= \frac{1-f}{n} \sum_h \frac{2}{n} [(y_{h1} - \frac{y_{h1}+y_{h2}}{2})^2 + (y_{h2} - \frac{y_{h1}+y_{h2}}{2})^2] \\
&= \frac{1-f}{n^2} \sum_h (y_{h1} - y_{h2})^2,
\end{aligned}
$$

as the sum of squares of the differences

- For element sampling, the symmetry of the selection and variance estimation disrupted by 1) Blanks in the list, non-responding elements or 2) Analysis of subclasses, Remedy: collapse strata but with overestimated variance

Poststratification

- Poststratification: variables to be used to create strata are not available at the time of selection
- Stratify after selection using variables collected during the survey
- Gains in precision are possible, with suitable modification to variance estimation
- Population control adjustment
- Poststratification requires
  - Poststrata known for each element
  - Poststratum weights W_h for each poststratum
  - New (approximate) variance estimator

Summary

1. Identify stratifying variables correlated with the measure(s) of interest
2. Choose "cuts" on the stratifying variables and divide the population into strata
3. S_h, $deff$, and S^2 are known, at least approximately, and known that fpc's within strata can be ignored
4. Compute an stratified sample size $n = n_{SRS} * deff$
5. Determine an allocation for the desired n
6. Adjust n based on expected $deff$ & allocate
7. Select sample and compute estimates taking the stratified sample selection into account