Stratified Sampling

Survey Sampling
Statistics 4234/5234
Fall 2018

September 20, 2018

What is stratified sampling?

If the variable we are interested in takes on different mean values in different subpopulations, we may be able to obtain more precise estimates of population quantities by taking a **stratified** random sample.

We divide the population into H subpopulations, called **strata**. The strata do not overlap, and they constitute the whole population so that each sampling unit belongs to exactly one **stratum**.

We draw an independent probability sample from each stratum, then pool the information to obtain overall population estimates.

We use stratified sampling for one or more of the following reasons:

1. We want to be protected from the possibility of obtaining a really bad sample.

   Example: Population of size $N = 2000$ consists of 1000 male and 1000 female students. The gender mix in a SRS of size $n = 100$ is likely to be close to 50-50, but there's about a 5% chance the split is 60-40 or worse.

   ```
   Pop <- c(rep(1,1000), rep(0,1000))
   male <- rep(NA, 1e5)
   for(j in 1:1e5){ male[j] <- sum(sample(Pop, 100)) }
   mean(male <= 40 | male >= 60)
   ```

   Eliminate this possibility by taking independent SRSs of 50 males and 50 females.

2. We may want data of known precision for subgroups of the population; these subgroups should be the strata.

   Example: Population of size $N = 2000$ consists of 1800 male and 200 female graduates. If a quantity of interest is the difference in average salary, we should sample a higher fraction of female graduates than male graduates to obtain comparable precision for the two groups.

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

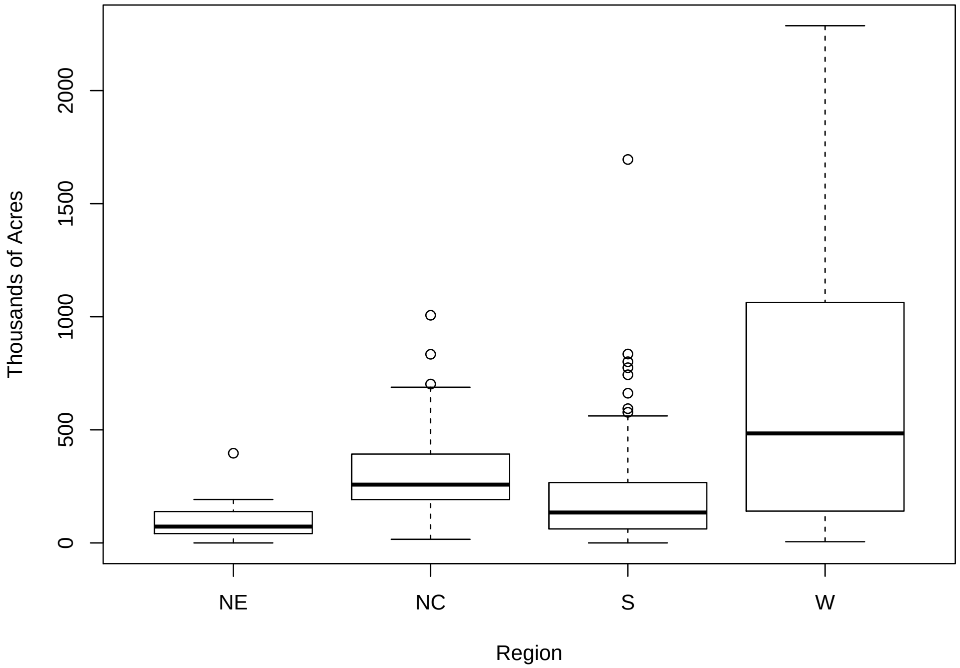
4. Stratified sampling often gives more precise (having lower variance) estimates for population means and totals.

Stratification works for lowering the variance because the variance within each stratum is often lower than the variance in the whole population.

Example: We wish to estimate the total acreage devoted to farms in the United States in 1992. There are $N = 3078$ counties and county-equivalents in the country; we obtain the number of acres devoted to farms in $n = 300$ counties.

Use the four census regions — Northeast, North Central, South, and West — as strata. The data for a stratified random sample using proportional allocation are:

Region	Counties	Sample size	Average	Std Dev
Northeast	220	21	97,630	87,450
North Central	1054	103	300,503	172,099
South	1382	135	211,315	231,490
West	422	41	662,296	629,433



Stratified random sampling

We divide the population of N sampling units into H strata, with N_h sampling units in stratum h; for stratified sampling to work we must known the values of $N_1, N_2, \ldots, N_H$ and must have

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

where N is the total number of units in the entire population.

In **stratified random sampling** we independently take an SRS from each stratum, so that n_h observations are randomly selected from the N_h population units in stratum h; the total sample size is $n = n_1 + n_2 + \cdots + n_H$.

Notation for stratification: population quantities

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

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

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

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

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

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

Notation for stratification: sample quantities

Define $\mathcal{S}_h$ to be the set of n_h units in the SRS for stratum h.

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

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

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

Estimation in stratified random sampling

We estimate the population total $t = \sum_{h=1}^{H} t_h$ by

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

and the population mean $\bar{y}_U$ by

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

Properties of estimators

- **Unbiasedness** We have $E(\bar{y}_h) = \bar{y}_{hU}$ in each stratum and thus

$$E\left(\bar{y}_{\text{strat}}\right) = E\left(\frac{N_h}{N}\bar{y}_h\right) = \sum_{h=1}^{H} \frac{N_h}{N} E(\bar{y}_h) = \sum_{h=1}^{H} \frac{N_h}{N}\bar{y}_{hU} = \bar{y}_U$$

- **Variance of the estimators**

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

and

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

- **Standard errors for stratified samples** We obtain unbiased estimators of the variances by substituting sample estimators s_h^2 for the population parameters S_h^2:

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

and

$$\widehat{V}\left(\bar{y}_{\text{strat}}\right) = \sum_{h=1}^{H} \left(\frac{N_h}{N}\right)^2 \frac{s_h^2}{n}\left(1 - \frac{n_h}{N_h}\right)$$

Then

$$\text{SE}\left(\hat{t}_{\text{strat}}\right) = \sqrt{\widehat{V}\left(\hat{t}_{\text{strat}}\right)} \quad \text{and} \quad \text{SE}\left(\bar{y}_{\text{strat}}\right) = \sqrt{\widehat{V}\left(\bar{y}_{\text{strat}}\right)}$$

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

$$\bar{y}_{\text{strat}} \pm z_{\alpha/2}\text{SE}\left(\bar{y}_{\text{strat}}\right)$$

Example: For the farms data (working with thousands of acres) we have

$$\begin{aligned}\hat{t}_{\text{strat}} &= 220(98) + 1054(301) + 1382(211) + 422(662) \\ &= 909{,}736\end{aligned}$$

and the estimated variance is

$$\begin{aligned}\widehat{V}\left(\hat{t}_{\text{strat}}\right) &= 220^2\frac{87.45^2}{21}\left(1-\frac{21}{220}\right)+1054^2\frac{172.1^2}{103}\left(1-\frac{103}{1054}\right)\\ &\quad+1382^2\frac{231.5^2}{135}\left(1-\frac{135}{1382}\right)+422^2\frac{629.4^2}{41}\left(1-\frac{41}{422}\right)\\ &= 50{,}417^2\end{aligned}$$

Converting now to millions of acres we obtain

$$909.736 \pm 1.96\,(50.41725) \quad \Rightarrow \quad [810.9,\ 1008.6]$$

and we are 95% confident that in 1992 there were somewhere between 811 and 1009 million acres of farmland in the United States.

Stratified sampling for proportions

To make inference about a population proportion based on stratified random sampling, proceed as above with

$$\widehat{y}_h = \widehat{p}_h \quad \text{and} \quad s_h^2 = \frac{n_h}{n_h - 1}\widehat{p}_h(1 - \widehat{p}_h)$$

Then

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

and

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

and

$$\text{SE}\left(\widehat{p}_{\text{strat}}\right) = \sqrt{\widehat{V}\left(\widehat{p}_{\text{strat}}\right)}$$

Estimate the total number of population units having a specified characteristic by

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

Then

$$\hat{V}\left(\hat{t}_{\text{strat}}\right) = N^2 \hat{V}\left(\hat{p}_{\text{strat}}\right)$$

Of course

$$\text{SE}\left(\hat{t}_{\text{strat}}\right) = \sqrt{\hat{V}\left(\hat{t}_{\text{strat}}\right)} = N\text{SE}\left(\hat{p}_{\text{strat}}\right)$$