

More on Stratified Random Sampling

This handout has some additional information on stratified sampling, compares stratified random sampling to simple random sampling under the setting of proportional allocation, and examines what is known as post-stratification.

Proportions

Extension of the formulas for estimating a population mean with a stratified sample to estimating a population proportion are straightforward. The stratum sample means are the stratum sample proportions \hat{p}_h and the stratum sample variances are $\frac{n_h}{n_h - 1} \hat{p}_h(1 - \hat{p}_h)$.

Comparison to SRS

- Recall that at the beginning of the semester in a small example with a population of size 4, we looked at properties of the mean estimator under both simple random and stratified random sampling. We found that if the strata were heterogeneous between and homogeneous within, then the stratified mean had smaller variance than the SRS-based mean. (We compared the variances since both estimators are unbiased for the population mean.)
- However, we also saw that if the strata are poorly defined, then an SRS can give a smaller variance than a stratified random sample for estimating a mean. In addition, a poor allocation can also make stratified sampling worse (for example, allocating most of the observations to the smallest and least variable strata). Hence, in general, one method will not always be superior to the other; it depends on the degree of heterogeneity between the strata defined and the allocation used.

It is possible to make a more formal comparison of SRS with stratified random sampling in the special case of proportional allocation. First, think about the parallels between the data for an ANOVA, where we are comparing responses across groups, to the data from a stratified random sample where we obtain data from separate groups (strata). Our goal is different; in the latter case, we know (or suspect) there are differences between the strata and our primary interest is not necessarily in that comparison. We're exploiting the stratum differences to estimate a population mean or total. However, the data have the same structure as in an ANOVA. Recall from ANOVA that we can partition the total variability in the whole sample (sum-of-squares total or SST) into two parts: the variability between the group means (sum-of-squares between or SSB) and the pooled variability within groups (sum-of-squares within or SSW). An analogous decomposition can be made for finite populations:

$$\text{SSB} = \sum_{h=1}^L \sum_{i=1}^{N_h} (\mu_h - \mu)^2 = \sum_{h=1}^L N_h (\mu_h - \mu)^2$$

$$\begin{aligned} \text{SSW} &= \sum_{h=1}^L \sum_{i=1}^{N_h} (y_{hi} - \mu_h)^2 = \sum_{h=1}^L (N_h - 1) \sigma_h^2 \\ \text{SST} &= \sum_{h=1}^L \sum_{i=1}^{N_h} (y_{hi} - \mu)^2 = (N - 1) \sigma^2 \end{aligned}$$

where it can be shown that $\text{SST} = \text{SSB} + \text{SSW}$. Recall the forms of the variances for the means in both SRS and stratified random sampling:

$$\begin{aligned} \text{Var}(\bar{y}) &= \left(\frac{N-n}{N} \right) \frac{\sigma^2}{n} = \left(\frac{N-n}{N} \right) \frac{\text{SST}}{n(N-1)} = \frac{1-f}{n(N-1)} (\text{SSW} + \text{SSB}) \\ \text{Var}(\bar{y}_{st}) &= \sum_{h=1}^L \left(\frac{N_h}{N} \right)^2 \left(\frac{N_h - n_h}{N_h} \right) \frac{\sigma_h^2}{n_h} = \frac{1}{N^2} \sum_{h=1}^L N_h (N_h - n_h) \frac{\sigma_h^2}{n_h} \end{aligned}$$

where $f = n/N$. Now assume $n = n_1 + \dots + n_L$ (same total sample size) and also assume proportional allocation. Then $f = \frac{n_h}{N_h} = \frac{n}{N}$ and so using the substitution $n_h = fN_h$,

$$\begin{aligned} \text{Var}(\bar{y}_{st}) &= \frac{1}{N^2} \sum_{h=1}^L N_h (N_h - n_h) \frac{\sigma_h^2}{n_h} = \frac{1}{fN^2} \sum_{h=1}^L (N_h - fN_h) \sigma_h^2 \\ &= \frac{1-f}{nN} \sum_{h=1}^L N_h \sigma_h^2 = \frac{1-f}{nN} \left(\text{SSW} + \sum_{h=1}^L \sigma_h^2 \right) \end{aligned}$$

Comparing this last expression to $\text{Var}(\bar{y})$ above, we can show that $\text{Var}(\bar{y}) < \text{Var}(\bar{y}_{st})$ only if

$$\text{SSB} = \sum_{h=1}^L N_h (\mu_h - \mu)^2 < \sum_{h=1}^L \left(1 - \frac{N_h}{N} \right) \sigma_h^2.$$

Since the left-hand side is large when the strata population sizes are large and/or when the stratum means are very different, it turns out that this condition is rarely satisfied in practice except if the strata sizes are exceptionally small and their means are nearly identical. Hence, stratification with proportional allocation can rarely hurt; obviously, optimal allocation will help even more (if the relative variances can be accurately estimated). The tradeoff is the extra effort needed for stratified sampling and the extra information needed (we need to know the stratum sizes and be able to sample within each stratum).

Stratification when stratum means are not different: Generally, we think of stratification as being beneficial when the stratum means are very different. However, it can be beneficial when the stratum means are not different, but the stratum variances are and we use optimal allocation. Then, we sample more heavily in the strata with large variances which can reduce the variance of the estimator of the population mean as compared to an SRS of the same total size.

Post-Stratification: To illustrate post-stratification, suppose we take a simple random sample (SRS) of adults in Missoula and estimate the proportion of adults who favor a state sales tax. After the fact, we might then decide to stratify by some variable such as age group.

- So we take an SRS of size n from the whole population, and then *note* that there are n_1 from stratum 1, \dots , n_L from stratum L . So, the stratum sizes are not fixed ahead of time and are hence viewed as random.
- With SRS, we would estimate μ with \bar{y} . However, if we know the relative stratum sizes (N_h/N), then we should adjust this estimate for the proportions actually sampled.
- So, the post-stratification estimate is: $\bar{y}_{st} = \sum_{h=1}^L \left(\frac{N_h}{N}\right) \bar{y}_h$. Note that this is just a weighted average of stratum means based on the relative stratum sizes.
- Under proportional allocation (the usual stratification situation), the variance of this post-stratification estimate of the mean μ is just the usual $\text{Var}(\bar{y}_{st})$ plus an additional variance component due to the variability in the stratum sizes. The approximate variance is given below (from equation (5) on page 124 of the text).

$$\begin{aligned} \text{Var}(\bar{y}_{pst}) &\approx \frac{N-n}{nN} \sum_{h=1}^L \left(\frac{N_h}{N}\right) \sigma_h^2 + \frac{1}{n^2} \left(\frac{N-n}{N-1}\right) \sum_{h=1}^L \frac{N-N_h}{N} \sigma_h^2 \\ &= \text{Var}(\bar{y}_{st}) + \frac{1}{n^2} \left(\frac{N-n}{N-1}\right) \sum_{h=1}^L \frac{N-N_h}{N} \sigma_h^2 \quad (\text{using } n_h = nN_h/N). \end{aligned}$$

- To estimate $\text{Var}(\bar{y}_{st})$ for post-stratification, the recommendation is to ignore the second term above and use the same estimate as one would if the sample were pre-stratified. Thus, the recommendation is to carry out an analysis for a post-stratified sample just as if it were a pre-stratified sample with fixed stratum sample sizes.
- Post-stratification is like ratio or regression estimation but with a categorical variable. We are taking advantage of the relationship of the response to an auxiliary variable in the estimation phase. It can be useful when we know the relative stratum sizes at the start, but it is inconvenient or impossible to sample separately within each stratum.