Stratified Sampling, part 3

Survey Sampling

Statistics 4234/5234

Fall 2018

September 27, 2018

Sampling: Design and Analysis, second edition; by Sharon L. Lohr

(Sections 3.4–3.5)

Survey design includes methods for controlling nonsampling as well as sampling error; today we discuss features that affect the sampling error.

Simple random sampling involves one design feature: the sample size.

For stratified random sampling, we need to determine what the strata should be, then decide how many observations to sample in each stratum.

We'll attack these questions in the reverse order.

Allocating observations to strata: proportional allocation

If you are taking a stratified sample in order to ensure that the sample reflects the population with respect to the stratification variable, and you would like your sample to be a miniature version of the population, you should use proportional allocation.

In **proportional allocation**

- the number of sampled units in each stratum is proportional to the size of the stratum, $n_h \propto N_h$;
- the inclusion probability $\pi_{hj} = n_h/N_h$ is the same $(= n/N)$ for all strata.

Example: Population of 2400 men and 1600 women, sample 10% of the population; proportional allocation would mean sampling 240 men and 160 women.

Under proportional allocation, the probability that an individual will be selected is n/N, same as for SRS, but many of the “bad” samples that could occur with SRS are no longer possible.

Example: Under SRS, each unit in the sample represents 10 people in the population. In stratified sampling with proportional allocation, each man in the sample represents 10 men in the population, and each woman represents 10 women in the population.

When the strata are large enough, the variance of $\bar{y}_{\text{strat}}$ under proportional allocation is usually less than the variance of the sample mean from an SRS with the same number of observations.

This is true no matter how silly the stratification scheme may seem.

Proposition: For estimating a population mean or total, stratification with proportional allocation will give smaller variance than SRS *unless*

$$\sum_{h=1}^{H} N_h \left(\bar{y}_{hU} - \bar{y}_U\right)^2 < \sum_{h=1}^{H} \left(1 - \frac{N_H}{N}\right) S_h^2$$

This rarely happens when the N_h are large.

In general, the variance of the estimator of t from a stratified sample with proportional allocation will be smaller than the variance of the estimator of t from an SRS with the same number of observations.

The more unequal the stratum means $\bar{y}_{hU}$, the more precision you will gain by stratifying and using proportional allocation.

If the variances S_h^2 are more or less equal across all the strata, proportional allocation is probably the best allocation for increasing precision.

In cases where the S_h^2 vary greatly, **optimal allocation** can result in smaller costs.

Allocating observations to strata: optimal allocation

In practice, when we are sampling units of different sizes, the larger units are likely to be more variable than the smaller units, and we should sample them with a higher sampling fraction.

Example (accounting): Use recorded book amount to stratify a population of loans; stratum 1 is loans over $1 million, stratum 2 is loans between $500,000 and $999,999, etc; S_h^2 will be much larger in the strata with large loan amounts, so optimal allocation should prescribe a higher sampling fraction for those strata. An error in the recorded amount of a $3 million dollar loan is much more important for the bank to know about that that of a $3000 loan! Maybe even set $n_1 = N_1$ in stratum 1.

The objective in optimal allocation is to gain the most information for the least cost. We want to minimize

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

subject to a constraint on the total cost of the sample, given by

$$c_0 + \sum_{h=1}^{H} c_h n_h \leq C$$

where C denotes the maximum total cost allowed, c_0 represents overhead costs, and c_h is the cost of taking an observation in stratum h.

Using Lagrange multipliers, the solution is to take

$$n_h \propto \frac{N_h S_h}{\sqrt{c_h}}$$

In optimal allocation, we sample heavily within a stratum if

- the stratum accounts for a large part of the population;
- the variance within the stratum is large (sample more heavily to compensate for the heterogeneity);
- sampling in the stratum is expensive.

If all variances and costs are equal, proportional allocation is the same as optimal allocation. If we know the variances within each stratum and they differ, optimal allocation gives a smaller variance for the estimator of $\bar{y}_U$ than proportional allocation.

Neyman allocation is a special case of optimal allocation, used when the costs in the strata (but not the variances) are approximately equal; under Neyman allocation

$$n_h \propto N_h S_h$$

If the variances S_h^2 are specified correctly, Neyman allocation will give an estimator with smaller variance than proportional allocation.

When the stratum variances S_h^2 are approximately known, Neyman allocation gives higher precision than proportional allocation. If the information about the stratum variances is of poor quality, however, disproportional allocation can result in a higher variance than simple random sampling. Proportional allocation, on the other hand, almost always has smaller variance than simple random sampling.

Determining sample sizes

Summing up, we have

- optimal allocation $n_h \propto N_h S_h / \sqrt{c_h}$
- Neyman allocation $n_h \propto N_h S_h$
- proportional allocation $n_h \propto N_h$

The different methods of allocating observations to strata give the relative sample sizes n_h/n. After strata are constructed and observations allocated to strata, the total sample size required to achieve a prespecified margin of error can be determined:

Take

$$n = V \cdot \left(\frac{z_{\alpha/2}}{e}\right)^2$$

where

$$V^* = \frac{1}{N^2} \sum_{h=1}^{H} \frac{n}{n_h} N_h^2 S_h^2$$

Defining strata

Stratification is most efficient when the stratum means differ widely — ideally we would stratify by the values of y.

Although, if we had this information we would not need to do a survey at all!

Instead we try to find some variable closely related to y, and stratify on that.

The number of strata you choose depends on many factors such as the difficulty in constructing a sampling frame with stratifying information, and the cost of stratifying.

A general rule to keep in mind is: The more information, the more strata you should use. Thus, you should use an SRS when little prior information about the target population is available.

You can often collect preliminary data that can be used to stratify your design.

In a survey with more precise information, you will want to use more strata — many surveys are stratified to the point where only two sampling units are observed in each stratum.