Chapter 4: Stratified Random Sampling

The way in which was have selected sample units thus far has required us to know little about the population of interest in advance of selecting the sample. This approach is ideal only if the characteristic of interest is distributed homogeneously across the population. If, however, the characteristic is distributed heterogeneously, then estimates based on these designs will be imprecise relative to several alternative sampling designs. For example, if we have information that we know to be associated with the heterogeneity in the population, we can use that ancillary information to guide alternative strategies for selecting samples that will yield estimates with higher precision that a simple random sample for the same amount of effort. The first of these designs is stratified random sampling.

A stratified random sample is one obtained by dividing the population elements into mutually exclusive, non-overlapping groups of sample units called strata, then selecting a simple random sample from within each stratum (stratum is singular for strata). Every potential sample unit must be assigned to only one stratum and no units can be excluded.

Stratifying involves classifying sampling units of the population into relatively homogeneous groups before (usually) selecting sample units. Strata are based on information other than the characteristic being measured that is known to or thought to vary with the characteristic of interest in such a way that the characteristic is more homogeneous within strata than among strata. Therefore, any feature that explains variation in the characteristic of interest can be used as a basis for defining strata. For example, if our goal is to estimate the number of agaves in an area, and we know from previous work that the agave abundance varies with soil type, we might choose to stratify the population by soil type. Because stratifying the population into relatively homogeneous groups of sampling units reduces sampling error, estimates generated within strata have higher precision than simple random samples drawn from the same population.

Because virtually all ecological systems are heterogeneous, stratifying is used commonly as a way to increase precision in ecological studies. Common strata in ecological studies include elevation, aspect, or other geographic features for studying plant communities and vegetation communities or soils for studying some animal communities. When choosing among several potential strata, seek strata that best minimize variation in the characteristic of interest within strata and that maximize variation among strata.

Stratified random sampling is appropriate whenever there is heterogeneity in a population that can be classified with ancillary information; the more distinct the strata, the higher the gains in precision. The same population can be stratified multiple times simultaneously.

Advantages:

- Higher precision of estimates
- Provides separate estimates for each stratum

Disadvantages:

- Requires ancillary information
- Can be more time consuming to plan and implement

How it is implemented:

- Divide the entire population into non-overlapping strata
- Select a simple random sample from within each strata

L = number of strata
N_i = number of sample units within stratum *i*
N = number of sample units in the population

Estimating the Population Mean

Estimates from stratified random samples are simply the weighted average or the sum of estimates from a series of simple random samples, each generated within a unique stratum. This should be apparent in the estimators below, where the population mean for example is an average of the means from each stratum weighted by the number of sample units measured in each stratum. With only one stratum, stratified random sampling reduces to simple random sampling.

The population mean (μ) is estimated with:

$$\hat{\mu} = \frac{1}{N}\left(N_1\hat{\mu}_1 + N_2\hat{\mu}_2 + \cdots + N_L\hat{\mu}_L\right) = \frac{1}{N}\sum_{i=1}^{L} N_i\hat{\mu}_i$$

where N_i is the total number of sample units in strata *i*, *L* is the number of strata, and *N* is the total number of sample units in the entire population.

Variance of the estimate $\hat{\mu}$ is again just the weighted average of estimated variances of the mean from a series of random samples drawn from strata *i* through *L*, although it looks a bit more cumbersome:

$$\hat{\text{var}}(\hat{\mu}) = \frac{1}{N^2}\left[N_1^2\left(\frac{N_1 - n_1}{N_1}\right)\left(\frac{s_1^2}{n_1}\right) + \cdots + N_L^2\left(\frac{N_L - n_L}{N_L}\right)\left(\frac{s_L^2}{n_L}\right)\right] = \frac{1}{N^2}\sum_{i=1}^{L} N_i^2\left(\frac{N_i - n_i}{N_i}\right)\left(\frac{s_i^2}{n_i}\right)$$

And s_i^2 is an estimate of the overall population variance from each strata *i* through *L*.

Standard error of $\hat{\mu}$ is $\sqrt{\hat{\text{var}}(\hat{\mu})}$: $SE(\hat{\mu}) = \sqrt{\frac{1}{N^2}\sum_{i=1}^{L} N_i^2\left(\frac{N_i - n_i}{N_i}\right)\left(\frac{s_i^2}{n_i}\right)}$.

Estimating the Population Total

Like the population mean, estimating a total for a stratified random sample is a matter of summing individual estimates of the total estimated for each stratum, $N_i\hat{\mu}_i$.

The population total τ is estimated with: $\hat{\tau} = N_1\hat{\mu}_1 + N_2\hat{\mu}_2 + \cdots + N_L\hat{\mu}_L = \sum_{i=1}^{L} N_i\hat{\mu}_i$

Variance of the estimated total $\hat{\tau}$ is: $\hat{\text{var}}(\hat{\tau}) = N^2\ \hat{\text{var}}(\hat{\mu}) = \sum_{i=1}^{L} N_i^2\left(\frac{N_i - n_i}{N_i}\right)\left(\frac{s_i^2}{n_i}\right)$

Standard error of $\hat{\tau}$ is the square root of $\hat{\text{var}}(\hat{\tau})$.

Example.

Estimating the Population Proportion

Similarly, estimating the proportion of the population with a particular trait (*p*) using stratified random sampling involves combining estimates from multiple simple random samples, each generated within a stratum. The population proportion is estimated with the sample proportion:

$$\hat{p} = N_1 \hat{p}_1 + N_2 \hat{p}_2 + \cdots + N_L \hat{p}_L = \sum_{i=1}^{L} N_i \hat{p}_i$$

Variance of the estimate $\hat{p}$ is:

$$\hat{\text{var}}(\hat{p}) = \frac{1}{N^2} \sum_{i=1}^{L} N_i^2 \, \hat{\text{var}}(\hat{p}_i) = \frac{1}{N^2} \sum_{i=1}^{L} N_i^2 \left(\frac{N_i - n_i}{N_i} \right) \left(\frac{\hat{p}_i (1 - \hat{p}_i)}{n_i - 1} \right)$$

Standard error of $\hat{p}$ is the square root of $\hat{\text{var}}(\hat{p})$.

Allocating Sampling Effort among Strata

Using stratified random sampling requires that we decide how to divide a fixed amount of sampling effort among the different strata; that process is called ***allocation***. When deciding where to allocate sampling effort, the question becomes how best to allocate effort among strata so that the sampling process will provide the most efficient balance of effort, cost, and estimate precision. Should we allocate the same sampling effort to each stratum? If strata are of different sizes, as is usually the case, should we allocate more effort to larger stratum?

There are several strategies for allocating sampling effort, and the more information available about the population of interest, the more efficient the allocation strategy can be. Information on variability within each stratum, relative cost of obtaining and measuring a sample unit from each stratum, and the number of sample units in each stratum can all help to increase overall sampling efficiency; these elements provide the foundation for common allocation strategies. All strategies function by generating a simple proportional multiplier by which a fixed number of samples can be allocated among strata.

Uniform Allocation

The simplest allocation strategy is to select the same number of samples from each stratum, which is an ideal approach if there is no information available about variability of units within strata, the cost of sampling is similar for all strata, and strata are of similar size.

Example.

Allocation Proportional to Size or Variation

The number of sample units to select from each stratum can be made proportional to the number of sample units (or size) within each stratum. Variation in a stratum often increases with a the size of a stratum, so in some cases this can be considered a rough approach for allocating more effort to strata that are likely to be more variable strata. To allocate proportional to stratum size:

$$n_i = n\left(\frac{N_i}{\sum_{i=1}^{L} N_i}\right) = n\left(\frac{N_i}{N}\right)$$

where *n* is the total number of sample units available for allocation, and *n*$_i$ is the number of sample units to allocate to stratum *i*.

Example.

To allocate proportional to the amount of variation among elements within each stratum, as measured by the estimated standard deviation within each stratum:

$$n_i = n\left(\frac{s_i}{\sum_{i=1}^{L} s_i}\right)$$

This approach relies on estimates generated from a previous study or alternatively by the ability to gauge relative differences in variation among strata, such as expecting one stratum to have 1.5 times the variation as another stratum.

Example.

<u>Optimal Allocation</u>

Both allocation approaches above are special cases of the optimal allocation strategy which estimates the population mean or total with the lowest variance for a given sample size in stratified random sampling. The number of samples selected from each stratum is proportional to the size, variation, as well as the cost (c_i) of sampling in each stratum. More sampling effort is allocated to larger and more variable strata, and less to strata that are more costly to sample.

$$n_i = n\left(\frac{\dfrac{N_i s_i}{\sqrt{c_i}}}{\dfrac{\sum_{k=1}^{L} N_k s_k}{\sqrt{c_k}}}\right)$$

where *k* indexes the *L* strata.

Example.