

## Stratified Random Sampling for Proportions: An Example

First, recall from the material on estimating proportions (Chap. 5 and pp. 12-13 of notes), the variance of the sample proportion  $\hat{p}$  based on an SRS of size  $n$  is

$$\text{Var}(\hat{p}) = \left(\frac{N-n}{N}\right) \frac{p(1-p)}{n}$$

where  $p$  is the population proportion. The variance is estimated by

$$\widehat{\text{Var}}(\hat{p}) = \left(\frac{N-n}{N}\right) \frac{\hat{p}(1-\hat{p})}{n-1}.$$

Now, suppose we have a stratified random sample from which we are going to estimate a population proportion. If  $p_h$  is the true proportion in stratum  $h$  and, as before,  $N_h$  and  $n_h$  are the stratum size and the sample size, then the stratified estimator of  $p$  with  $L$  strata is

$$\hat{p}_{st} = \sum_{h=1}^L \left(\frac{N_h}{N}\right) \hat{p}_h$$

with variance

$$\text{Var}(\hat{p}_{st}) =$$

The variance would be estimated by

$$\widehat{\text{Var}}(\hat{p}_{st}) =$$

### Example:

Suppose that a large area is divided into 1000 quadrats in three strata. In stratum 1, there are 600 quadrats and it is guessed that about 50% of these contain plants of species X. In stratum 2, there are about 370 quadrats and it is guessed that about 30% contain species X. In stratum 3, there are 30 quadrats and it is guessed that about 10% contain species X. A survey is to be conducted to estimate the proportion of quadrats that contain species X in the whole population of 1000 quadrats. Compare the accuracy of simple random sampling, stratified sampling with proportional allocation, and stratified sampling with optimal allocation.

Based on the guessed percentages, the proportion of all quadrats containing species X is

about  $p = (600 \times 0.5 + 370 \times 0.3 + 30 \times 0.1)/1000 = 0.414$ . The standard deviation of the the sample proportion  $\hat{p}$  based on an SRS of 100 plots would then be estimated to be:

$$SD(\hat{p}) =$$

For stratified random sampling with proportional allocation, the sample sizes will be:

$$n_1 = \qquad \qquad \qquad n_2 = \qquad \qquad \qquad n_3 =$$

The standard deviation of the stratified estimator (based on the guessed proportions) is then

$$SD(\hat{p}_{st}) =$$

The optimum equal-cost allocation is

$$n_h = \frac{nN_h\sigma_h}{\sum_{k=1}^L N_k\sigma_k}, \quad h = 1, 2, \dots, L$$

where  $\sigma_h$  is the standard deviation of the 0-1 values in stratum  $h$ . From Chapter 5, this is equal to

$$\sigma_h = \sqrt{\frac{N_h}{N_h - 1} p_h(1 - p_h)}.$$

Calculating these quantities for the example (using the guessed proportions):

$$\sigma_1 =$$

$$\sigma_2 =$$

$$\sigma_3 =$$

The denominator in the allocation formula is therefore

$$\sum_{k=1}^3 N_k\sigma_k =$$

Hence, the optimal allocation is

$$n_1 =$$

$$n_2 =$$

$$n_3 =$$

Rounding the sample sizes to integers, the standard deviation of the stratified estimator is then

$$\text{Var}(\hat{p}_{st}) =$$

Conclusions? Is there an advantage to stratified sampling over SRS? Is there an advantage to optimal allocation over proportional?

Some remarks:

1. If the cost to sample a unit is the same for all strata, then the gain from stratified random sampling is small unless the proportions vary greatly across strata.
2. Optimum allocation is little better than proportional allocation unless some strata proportions are outside the range 0.1 to 0.9. This is because the standard deviation of 0-1 data is proportional to  $\sqrt{p(1-p)}$  which varies by less than a factor of 2 over the range  $p=0.1$  to 0.9.