Lecture 5 - Key

Stratified Random Sampling

Suppose the population is partioned into disjoint sets of sampling units called **strata**. If a sample is selected within each stratum, then this sampling procedure is known as **stratified sampling.**

If we can assume the strata are sampled independently across strata, then

1. *the estimator of t or $\bar{y}_U$ can be found by combining stratum sample sums or means using appropriate weights*

2. *the variances of the estimators associated with the individual strata can be summed to obtain the variance of an estimator associated with the whole population. (Given independence, the variance of a sum equals the sum of the individual variances.)*

Point (2) implies that only within-stratum variances contribute ot the variance of an estimator. Thus, the basic motivating principle behind using stratification to produce an estimator with small variance is to partition the population so that units within each stratum are as similar as possible. *This is known as the* ***stratification principle.***

Recall the bird dataset used in Lab 3. Stratifying by terrain, or habitat type, greatly reduced the variance of the estimator.

In general ecological studies, it is common to stratify a geographical region into subregions that are similar with respect to a known variable such as elevation, animal habitat type, vegetation type, ect... because it is suspected that the y-values may very greatly across strata while they will tend to be similar within each stratum. Analogously, when sampling people, it is common to stratify on variables such as gender, age groups, income levels, education levels, marital status, ect.

Sometimes strata are formed based on sampling convenience. For example, suppose a large study region appears to be homogenous (that is, there are no spatial patterns) and is stratified based on the geographical proximity of the sampling units. Taking a stratified sample ensures the sample is spread throughout the study region. *It may not, however, lead to any significant reduction in the variance of an estimator.*

If the y-values are spatially correlated (y-values tend to be similar for neighboring units), geographically determined strata can improve estimation of population parameters.

Stratified Sampling Notation

$$
\begin{aligned}
H &= \text{the number of strata} \\
N_h &= \text{number of population units in stratum } h, \text{ where } h = 1, 2, \ldots, H \\
N &= \sum_{h=1}^{H} N_h = \text{ the number of units in the population} \\
n_h &= \text{number of sampled units in stratum } h \\
n &= \sum_{h=1}^{H} n_h = \text{ the total number of units sampled} \\
y_{hj} &= \text{the } y\text{-value associated with unit } j \text{ in stratum } h \\
\bar{y}_h &= \text{the sample mean for stratum } h \\
t_h &= \sum_{j=1}^{N_h} y_{hj} = \text{ the stratum } h \text{ total} \\
t &= \sum_{h=1}^{H} \sum_{j=1}^{N_h} y_{hj} = \sum_{h=1}^{H} t_h = \text{ the population total} \\
\bar{y}_{hU} &= \frac{t_h}{N_h} = \text{ stratum } h \text{ mean} \\
\bar{y}_U &= \frac{1}{N} \sum_{h=1}^{H} \sum_{j=1}^{N_h} y_{hj} = \frac{t}{N} = \text{ the population mean}
\end{aligned}
$$

If a simple random sample is taken within each stratum, then the sampling design is called stratified simple random sampling.

For stratum h, there are $\binom{N_h}{n_h}$ possible SRSs of size n_h. Therefore, there are $\binom{N_1}{n_1}\binom{N_2}{n_2}\ldots\binom{N_H}{n_H}$ possible stratified SRSs for specified stratum sample sizes $n_1, \ldots, n_H$.

If $\mathcal{S}_{strat}$ is a stratified SRS, then the probability of selecting $\mathcal{S}_{strat}$ is

$$P(\mathcal{S}_{strat}) = \prod_{h=1}^{H} \frac{1}{\binom{N_h}{n_h}} = \frac{1}{\binom{N_1}{n_1}\ldots\binom{N_H}{n_H}}$$

Thus, every possible stratified SRS having the sample sizes $n_1, \ldots, n_H$ has the same probability of being selected.

Estimation of $\bar{y}_U$ and t

Because a SRS was taken within each stratum, we can apply the estimator formulas for simple random sampling to each stratum. We can estimate each stratum population mean $\bar{y}_{hU}$ and each stratum population total t_h.

The formulas are:

$$\begin{aligned} \hat{\bar{y}}_{hU} &= \bar{y}_h = *\frac{1}{n_h}\sum_{j=1}^{n_h} y_{hj} * \\ \hat{t}_h &= N_h\bar{y}_h = *\frac{N_h}{n_h}\sum_{j=1}^{n_h} y_{hj}* \end{aligned}$$

Example 1. Considered a smaller version of the bird dataset with only two strata: prairie (stratum 1) and non-prairie (stratum 2). Let $N_1 = 40$, $N_2 = 40$, $n_1 = 4$ and $n_2 = 4$. Then let $\{y_{11}, ..., y_{14}\} = \{7, 4, 9, 10\}$ and $\{y_{21}, ..., y_{24}\} = \{9, 12, 12, 11\}$

Compute:

$\bar{y}_1 = *\frac{30}{4} = 7.5*$

$\hat{t}_1 = 75$

$\bar{y}_2 = \frac{44}{4} = 11$

$\hat{t}_2 = 110$

Because each $\hat{t}_h$ is an unbiased estimator of the stratum total t_h for $i = 1, 2, \ldots, k$ their sum will be an unbiased estimator for the population total t. That is $\hat{t}_{str} = \sum_h \hat{t}_h$ is an unbiased estimator of t.

An unbiased estimator of $\bar{y}_U$ is a weighted average of the stratum sample means

$$\hat{\bar{y}}_{U,str} = \frac{\hat{t}_{str}}{N} = \frac{1}{N}\sum_{h=1}^{H} N_h\bar{y}_h$$

Now compute the population total and population mean

$\hat{t}_{str} = t_1 + t_2 = 75 + 110 = 185$

$\hat{\bar{y}}_{U,str} = \frac{40}{80}\bar{y}_1 + \frac{40}{80}\bar{y}_2 = 9.25$

Before studying the overall variances $V(\hat{t}_{str})$ and $V(\hat{\bar{y}}_{U,str})$, we need to look at the within-stratum variances.

Because a SRS is taken within each stratum h, we can apply the results for simple random sampling estimators to each stratum. The variances of the stratified SRS estimators of the mean and total are:

$$V(\hat{\bar{y}}_{Uh}) = \left(1 - \frac{n_h}{N_h}\right)\frac{S_h^2}{n_h} \qquad\qquad V(\hat{t}_h) = N_h(N_h - n_h)\frac{S_h^2}{n_h}$$

where S_h^2 is the variance for stratum h.

Because the simple random samples are *independent* across the strata, the variance of $\hat{t}_{str}$ is the sum of the individual stratum variances:

$$*V(\hat{t}_{str}) = \sum_{h=1}^{H} V(\hat{t}_h) = \sum_{h=1}^{H} N_h(N_h - n_h)\frac{S_h^2}{n_h}*$$

Dividing by N^2 gives $V(\hat{\bar{y}}_{U,str})$:

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

Again in most cases, S_h^2 is unknown, so we use s_h^2 to get an unbiased estimator of $V(\hat{t}_h)$, where s_h^2 is the sample variance of the n_h y-values sampled from stratum h. Using s_h^2 in place of S_h^2 above returns $\hat{V}(\hat{t}_{str})$ and $\hat{V}(\hat{\bar{y}}_{U,str})$.

Taking a square root of $\hat{V}(\hat{t}_{str})$ or $\hat{V}(\hat{\bar{y}}_{U,str})$ yield the corresponding standard error, which is used in generating confidence intervals.

Suppose the sample variances of strata 1 and strata 2 are $s_1^2 = 7$ and $s_2^2 = 2$, compute the variance of $\hat{t}_1$, $\hat{t}_2$, and $\hat{t}_{str}$.

$V(\hat{t}_1) = 40 \times (40 - 4)\frac{7}{4} = 2520$

$V(\hat{t}_2) = 40 \times (40 - 4)\frac{2}{4} = 720$

$V(\hat{t}_{str}) = V(\hat{t}_1) + V(\hat{t}_2) = 3240$

Confidence intervals for $\bar{y}_U$ and t

If all of the stratum sample sizes n_h are sufficiently large ($\approx n_h \geq 30$) an approximate 100(1-α)% confidence interval for $\bar{y}_U$ and t are

$$\hat{\bar{y}}_{U,str} \pm z^* \sqrt{\hat{V}(\hat{\bar{y}}_{U,str})} \qquad\qquad \hat{t}_{str} \pm z^* \sqrt{\hat{V}(\hat{t}_{str})}$$

where z^* is the upper $\alpha/2$ critical value from the standard normal distribution.

For smaller sample sizes, the following confidence intervals are recommended:

$$\hat{\bar{y}}_{U,str} \pm t^* \sqrt{\hat{V}(\hat{\bar{y}}_{U,str})} \qquad\qquad \hat{t}_{str} \pm t^* \sqrt{\hat{V}(\hat{t}_{str})}$$

where t^* is the upper $\alpha/2$ critical value form the $t(d)$ distribution. In this case, d is Satterthwaite's approximation of degrees of freedom d where

$$d = \frac{\left(\sum_{h=1}^{H} a_h s_h^2\right)^2}{\sum_{h=1}^{H} (a_h s_h^2)^2/(n_h - 1)} = \frac{(\hat{V}(\hat{t}_{str}))^2}{\sum_{h=1}^{H} (a_h s_h^2)^2/(n_h - 1)}$$

where $a_h = N_h(N_h - n_h)/n_h$.

If the stratum sizes n_h are all equal and the stratum size N_h are all equal, then the degrees of freedom reduces to $d = n - H$ where $n = \sum n_h$ is the total sample size. Some packages (including R) will use $n - H$ degrees of freedom as the default.

One-sided confidence intervals can be generated in the same fashion as before.

Efficiency of Stratified Simple Random Sampling

Because the variance formulas for $\hat{t}_{str}$ and $\hat{\bar{y}}_{U,str}$ are determined only from within-stratum variances, the precision of the estimators can be improved by forming strata with small S_h^2 values (strata with similar $y - values$ within each stratum). *To contrast stratified random sampling with SRS we will compare* $\hat{V}(\hat{\bar{y}}_U)$ *from a SRS to* $\hat{V}(\hat{\bar{y}}_{str})$ *from a stratified SRS.*

If the variance of the stratified sampling estimator is less than the variance of the SRS estimator, *then we say that* $\hat{\bar{y}}_{U,str}$ ***is more efficient*** *than* $\hat{\bar{y}}_U$*.*

A stratified SRS estimator will be more efficient than the SRS estimator of $\bar{y}_U$ or t *if the variability between stratum means is sufficiently large relative to the within-stratum variability.*

Allocation of Sampling Units

Given that we have enough resources to allocate n units among the H strata, how do we determine the stratum sample sizes n_h?

Situation 1: If all of the strata are the same size and no prior information is available about the population, a reasonable choice would be *to assign equal (or nearly equal) sample sizes to the strata. That is,* $n_h \approx \frac{n}{H}$*.*

Example: *Think about the bike use data from Lab. If we were concerned with bike use by different quadrants in Washington, D.C., equal sample sizes might be a reasonable choice.*

Situation 2: If the strata are not all the same size and no prior information is available about the population, a reasonable choice would be to assign sample sizes *proportional to the sizes of the strata relative to the population size N. That is, $n_h \approx n\frac{N_h}{N}$ This is known as proportional allocation.*

Example: Think about the example in Lab 3. Proportional allocation was used by many groups in this scenario.

Situation 3: The allocation scheme that minimizes $V(\hat{t}_{str})$ is called **optimum allocation** and requires

$$n_h = \frac{N_h S_h}{\sum_{l=1}^{H} N_l S_l} n$$

.

Because the S_h values are unknown, we would need prior point estimates to attempt optimum allocation.

Example: Suppose I gave you the variance terms for the habitat types in the bird survey dataset. Then you would allocate those proportional to the product of the population size and the population standard deviation for each terrain.

Situation 4: In some cases, if the cost of sampling units varies from stratum to stratum, then the total cost of taking a stratified SRS may determine how to allocate units to strata.

We will talk more about this later.

Quota Sampling

When the population is stratified based on more than one factor we are using **multifactor stratification.** *Typically a sample is taken with every combination of strata across the factors.*

A marketing survey was conducted to get information on computer users' use of the internet for shopping. The surveyor suspected there may be differences based on gender (2 levels), household income (6 levels), and age (5 levels). T*here would be a* $2 \times 6 \times 5 = 60$ *multifactor strata to sample from.*

Quota sampling is a form of stratified sampling and typically uses multifactor stratification. Taking a quota sample ensures that data are collected across the population with the belief that doing so will provide a representative sample from the population. *It also allows the researchers to generate estimates related to various subgroups.*

So how does quota sampling differ from stratified simple random sampling? *In quota sampling,* ***the within stratum samples may not be random.*** *Often some element of subjectivity enters into the sampling procedure.*

A typical quota sample is based on:

1. Defining *the multifactor strata.*

2. Determining the stratum sample sizes based on proportional allocation. These are also referred to as the **stratum quotas**. These sample sizes are based on either known or approximate stratum sizes. They can also be based on either known or approximate population proportions associated with each stratum.

3. Data is collected by *predetermined data collection techniques(e.g., phone surveys, mail surveys, personal interviews, ect.) until the stratum quotas are satisfied (that is, until the desired number of responses are collected for each stratum.)*

Although taking a quota sample can save a lot of time and money when compared to simple random sampling, the reseacher must realize that if quota sampling is used, *we cannot be sure that the selection of sampling units would be similar to units collected via simple random sampling.*

Example. Consider online political polling. In principle, use of the estimation formulas based on random sampling techniques on data from a quota sample violates underlying assumptions. Therefore, we cannot be as confident in the results from quota sampling in comparison to results when a probability sampling method is used.

A student organization wants to determine if students favor extending the evening hours that the library remains open. They decided to take a quota sample based on strata related to class standing(first year, sophomore, junior, senior, graduate students). After deciding on the 5 quotas of 25 students for each of the 5 strata, data was collected at the library on consecutive evenings until all 5 quotas were satisfied. What concerns do you have?

You are sampling students who are known to study in the library and not sampling any students who are less likely to study in the library. Therefore, the recorded responses will be biased in support of extending the evening hours that the library remains open.

Nonetheless, quota sampling has proven useful if the quotas are designed properly with careful attention paid to when, how, and where the data are collected.

Poststratification

In some stratified sampling situations, we may not be able to determine from which stratum an observation belongs until it is actually observed.

For example, a fish population may be stratified by age class, body length, body weight, or sex. A random sample of fish is collected. Then each sampled fish is examined and its stratum is recorded along with the response of interest. *Therefore, because the stratum for each fish is determined after sampling, it is not possible to guarantee exact stratum sample sizes n_h for each age group, ect...*

In many studies a simple random sample is taken from the population and then is stratified. *The procedure is known as* ***poststratification*** *or* ***post-hoc stratification****.*

In constrast to traditional stratified simple random sampling, *the stratum sample sizes $n_1, n_2, \ldots, n_h$ are random variables.*