Math/Stat 310: Intro to Mathematical Statistics February 12, 2018

Lectures 5: Stratified Sampling

Instructor: Hyunseung Kang *Scribe: Meenmo Kang*

1 Stratified Sampling

When we need to collect a huge amount of data, it would be efficient to divide the data set into small groups. The small groups can be categorized by region, race, gender, or age, for example. Once the data is separated into small groups, then each of small groups is called a stratum, strata for plural. Suppose there is a large data set that was sampled in the US. The data can be categorized by 50 different states. Then the number of total strata L will be 50. After dividing up a dataset, SRS is run on each stratum.

Properties of Stratified SRS
Several properties of stratified SRS are denoted as shown below.

n_l: The number of people in stratum l that were sampled.
N_1: Total number of people in stratum l.

Population: $\{\xi_{lk}, ... \mid l = lth \text{ stratum}, k = kth \text{ individual in stratum } l\} \Rightarrow N = \sum_{l=1}^{L} N_l$

Sample: $X_{lk} = \begin{cases} 1 & \text{If individual } k \text{ in stratum } l \text{ is sampled} \\ 0 & \text{Otherwise} \end{cases}$

Joint and marginal PDF of stratified samplings are formed similarly as those of SRS, since, as I mentioned prior, SRS samplings take place after dividing up each stratum. As a result, joint and marginal PDF of stratified random samplings are denoted as follows.

$$pdf_l(X_{l1}, X_{l2}, X_{l3}, \ldots, X_{lN_l}) = \frac{1}{\binom{N_l}{n_l}} \qquad pdf(X_{lk} = 1) = \frac{n_l}{N_l}$$

For stratum l, there are $\binom{N_l}{n_l}$ total possible ways to sample n_l out of N_l. This means that there are $\binom{N_1}{n_1} \cdot \binom{N_2}{n_2} \ldots \binom{N_l}{n_l}$ possible stratified random samplings. Though each element of the same stratum is dependent on the other, taken by SRS, each stratum is independent. Therefore, the *pdf* can be a product of each stratum. In mathematical notation,

$$pdf(X_{11}, X_{12}, X_{13}, ..., X_{lN_l}) = \prod_{l=1}^{L} pdf_l(X_{l1}, X_{l2}, X_{l3}, \ldots, X_{lN_l}) = \prod_{l=1}^{L} \frac{1}{\binom{N_l}{n_l}}$$

Population Parameters

Population mean is denoted by $\mu = \frac{1}{N} \sum_{l=1}^{L} \sum_{k=1}^{N_l} \xi_{lk}$. Meanwhile, this can be modified using the sample mean formula as below.

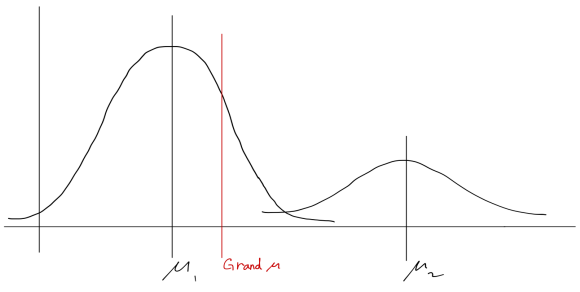
$$\mu_l = \frac{1}{N_l} \sum_{k=1}^{N_l} \xi_{lk} \;\Rightarrow\; \sum_{k=1}^{N_l} \xi_{lk} = \mu_l N_l \qquad \mu = \frac{1}{N} \sum_{l=1}^{L} \mu_l N_l$$

This implies that the sum of each product of mean μ_l and the number of elements N_l of stratum l divided by total number of population N will be population mean.

For variance, there are two different types. One is for unequal variance assumption in T test, denoted as follows.

$$\sigma^2 = \frac{1}{N} \sum_{l=1}^{L} \sum_{k=1}^{N_l} (\xi_{lk} - \mu_l)^2$$

For this case, each stratum has a different size of variance. As depicted below, suppose that the size of one stratum is much larger than that of another one. Then, the grand mean, which is the sum of both means divided by the total number of elements, n_1 and n_2, tends to be biased to the larger one.



Another type of variance is for pooled variance assumption, where the variance of each stratum is identical.

$$\tilde{\sigma}^2 = \frac{1}{N} \sum_{l=1}^{L} \sum_{k=1}^{N_l} (\xi_{lk} - \mu)^2$$

Note that $\tilde{\sigma}^2 = \sigma^2$ if and only if $\mu = \mu_l$ for every l