Math 408 - Mathematical Statistics

Lecture 19. Stratified Random Sampling

March 6, 2013

Agenda

- Definition of the Stratified Random Sampling (StrRS)
- Basic statistical properties of estimate of μ obtained under StrRS
- Neyman Allocation Scheme
- Summary

Stratified Random Sampling

In stratified random sampling (StrRS), the population is partitioned into subpopulations, or strata, which are then independently sampled.

In many applications, stratification is natural.

Example:

In samples of human populations, geographical areas form natural strata.

Reasons for using StrRS:

- We are often interested in obtaining information about each natural subpopulation in addition to information about the whole population.
- Estimates obtained from StrRS can be considerably more accurate than estimates from simple random sampling if
  - population members within each stratum are relatively homogeneous, and
  - there is considerable variation between strata.

Mathematical Framework of StrRS

Suppose there are L strata. Let N_k be the number of population elements in the k^{th} stratum. The total population size is

$$N = \sum_{i=1}^{L} N_k$$

Denote the mean and variance of the k^{th} stratum by μ_k and σ_k^2, respectively. Let $x_i^{(k)}$ denote the i^{th} value in the k^{th} stratum, then the overall population mean

$$\mu = \frac{1}{N}\sum_{k=1}^{L}\sum_{i=1}^{N_k} x_i^{(k)} = \frac{1}{N}\sum_{k=1}^{L} N_k\mu_k = \sum_{k=1}^{L}\frac{N_k}{N}\mu_k = \sum_{k=1}^{L}\omega_k\mu_k, \qquad \omega_k = \frac{N_k}{N}$$

Thus, the overall population mean is

$$\mu = \sum_{k=1}^{L}\omega_k\mu_k, \qquad \omega_k = \frac{N_k}{N},$$

where ω_k is the fraction of the population in the k^{th} stratum.

Mathematical Framework of StrRS

Within each stratum, a simple random sample $X_1^{(k)}, \dots, X_{n_k}^{(k)}$ of size n_k is taken. The sample mean is

$$\overline{X}_{n_k}^{(k)} = \frac{1}{n_k} \sum_{i=1}^{n_k} X_i^{(k)}, \qquad k = 1, \dots, L$$

Since $\mu = \sum_{k=1}^{L} \omega_k \mu_k$, the natural estimate of μ is

$$\boxed{\overline{X}_n^* = \sum_{k=1}^{L} \omega_k \overline{X}_{n_k}^{(k)}}$$

Remark:

We use star to distinguish $\overline{X}_n^*$ (obtained from stratified random sampling) from $\overline{X}_n$ (obtained from simple random sampling)

Our goal: to study statistical properties of $\overline{X}_n^*$

In particular, we want to find $\mathbb{E}[\overline{X}_n^*]$ and $\mathbb{V}[\overline{X}_n^*]$

Expectation $\mathbb{E}[\overline{X}_n^*]$

Theorem

$\overline{X}_n^*$ *is an unbiased estimate of* μ,

$$\boxed{\mathbb{E}[\overline{X}_n^*] = \mu}$$

Variance $\mathbb{V}[\overline{X}_n^*]$

Theorem

Under stratified random sampling,

$$\boxed{\mathbb{V}[\overline{X}_n^*] = \sum_{k=1}^{L} \omega_k^2 \frac{\sigma_k^2}{n_k}\left(1 - \frac{n_k - 1}{N_k - 1}\right)}$$

Corollary

If the sampling fractions within each stratum are small, i.e. $n_k/N_k \ll 1$*, then*

$$\boxed{\mathbb{V}[\overline{X}_n^*] \approx \sum_{k=1}^{L} \omega_k^2 \frac{\sigma_k^2}{n_k}}$$

Our next goal: to decide how to choose sample sizes $n_1, \ldots, n_L$ efficiently

Neyman Allocation Scheme

So, it was shown that (neglecting the sampling fractions $n_k/N_k \ll 1$)

$$\mathbb{V}[\overline{X}_n^*] = \sum_{k=1}^{L} \omega_k^2 \frac{\sigma_k^2}{n_k}$$

Question:

Suppose that the resources of a survey allow only a total of n units to be sampled. How to choose $n_1, \ldots, n_L$ to minimize $\mathbb{V}[\overline{X}_n^*]$ subject to constraint $\sum n_k = n$?

Optimization problem:

$$\mathbb{V}[\overline{X}_n^*] \to \min \qquad \text{s.t.} \sum_{k=1}^{L} n_k = n \tag{1}$$

Theorem

The sample sizes $n_1, \ldots, n_L$ that solve the optimization problem (1) are given by

$$\boxed{n_k = n \frac{\omega_k \sigma_k}{\sum_{j=1}^{L} \omega_j \sigma_j}} \qquad k = 1, \ldots, L$$

- This optimal allocation scheme is called Neyman allocation

Summary

- Stratified Random Sampling:
  population is partitioned onto strata which are then sampled independently.
- Under stratified random sampling, the estimate of μ is

$$\overline{X}_n^* = \sum_{k=1}^{L} \omega_k \overline{X}_{n_k}^{(k)}$$

- The expectation and variance (assuming $n_k/N_k \ll 1$):

$$\boxed{\mathbb{E}[\overline{X}_n^*] = \mu} \qquad \boxed{\mathbb{V}[\overline{X}_n^*] = \sum_{k=1}^{L} \omega_k^2 \frac{\sigma_k^2}{n_k}}$$

- Neyman Allocation Scheme minimizes $\mathbb{V}[\overline{X}_n^*]$ subject to $\sum_{k=1}^{N} n_k = n$:

$$\boxed{n_k = n \frac{\omega_k \sigma_k}{\sum_{j=1}^{L} \omega_j \sigma_j}} \qquad k = 1, \ldots, L$$