LECTURE 7: SYSTEMATIC SAMPLING 4.2 – 4.4

4.2 Estimation of Population Parameters

Recall that in systematic sampling the sampling interval is $k = N/n$. We will assume that k is an integer.

The sample total is $x = \sum_{i=1}^{n} x_i$, and the point estimator of the population total X is $x' = \frac{N}{n}x = k\,x$. The sample mean is $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, and the sample proportion is $p_x = \frac{1}{n}\sum_{i=1}^{n} x_i$ where $x_i = 0$ or 1.

It is convenient to consider a systematic sample as a cluster of n elements, each obtained with probability $1/k$.

Cluster	Elements
1	$X_1, X_{1+k}, \ldots, X_{1+(n-1)k}$
2	$X_2, X_{2+k}, \ldots, X_{2+(n-1)k}$
	$\ldots$
i	$X_i, X_{i+k}, \ldots, X_{i+(n-1)k}$
	$\ldots$
k	$X_k, X_{2k}, \ldots, X_{nk}$

We will denote by X_{ij} the jth element from the ith cluster,

$$X_{i1} = X_i,\ X_{i2} = X_{i+k}, \ldots, X_{ij} = X_{i+(j-1)k}, \ldots, X_{in} = i + (n-1)k\,.$$

That is, $X_{ij} = X_{i+(j-1)k},\ i = 1, \ldots, k,\ j = 1, \ldots, n\,.$

Proposition (i) The re-scaled sample total x' is an unbiased estimator of the population total X, and

$$\mathbb{V}ar(x') = \Big(\frac{N^2\,\sigma_X^2}{n}\Big)\Big(\frac{N-n}{N-1}\Big)\big(1 + \delta_X(n-1)\big)$$

with the intraclass correlation coefficient

$$\delta_X = \frac{2\sum_{i=1}^{k}\sum_{j=1}^{n}\sum_{l<j}(X_{ij}-\bar{X})(X_{il}-\bar{X})}{nk(n-1)\sigma_X^2}$$

where X_{ij} is the jth element from the ith cluster, and X_{il} is another element from the ith cluster.

(ii) The sample mean $\bar{x}$ is an unbiased estimator of the population mean $\bar{X}$, and

$$\mathbb{V}ar(\bar{x}) = \Big(\frac{\sigma_X^2}{n}\Big)\Big(\frac{N-n}{N-1}\Big)\big(1 + \delta_X(n-1)\big)\,.$$

(iii) The sample proportion p_x is an unbiased estimator of the population proportion P_X, and

$$\mathbb{V}ar(p_x) = \Big(\frac{P_X(1-P_X)}{n}\Big)\Big(\frac{N-n}{N-1}\Big)\big(1 + \delta_X(n-1)\big)\,.$$

The intraclass correlation coefficient δ_X measures homogeneity of observations within a sampled cluster. It is interpreted as proportion of total variation that is attributed to correlation within the cluster.

To estimate δ_X we need measurements from all elements in a population, thus, δ_X is not estimable from observations in a sample. However, if elements in the sampling frame are in random order, then we can assume that δ_X. In this case, we can treat the systematic sampling as a special case of simple random sampling. The point estimates of population total, mean, proportion and their variances are the same as for the simple random sampling. Thus, we have

(i) A $100(1-\alpha)\%$ confidence interval for the population total $X = \sum_{i=1}^{N} X_i$ is

$$x' \pm z_{1-\alpha/2} \frac{N\, s_x}{\sqrt{n}} \sqrt{1 - \frac{1}{k}}$$

(ii) A $100(1-\alpha)\%$ confidence interval for the population mean $\bar{X}$ under the simple random sampling is

$$\bar{x} \pm z_{1-\alpha/2} \frac{s_x}{\sqrt{n}} \sqrt{1 - \frac{1}{k}}\,.$$

(iii) A $100(1-\alpha)\%$ confidence interval for the population proportion has the form

$$p_x \pm z_{1-\alpha/2} \sqrt{1 - \frac{1}{k}} \sqrt{\frac{p_x(1-p_x)}{n-1}}\,.$$