LECTURE 9: STRATIFIED RANDOM SAMPLING 5.5, 5.6, 6.2, 6.3, 6.5.3, 6.5.4

5.5, 5.6 Sample Statistics for Strata

Suppose a stratified random sample of size n is drawn from a population with L strata. We assume that n_h elements are sampled within each stratum, so that $n = \sum_{h=1}^{L} n_h$. Let $x_{h,i}$ denote the measurement taken on the ith sampled element in stratum h, $i = 1, \ldots, n_h$.
The sample total within stratum h is

$$x_{h+} = \sum_{i=1}^{n_h} x_{h,i} \, .$$

The sample mean within stratum h is

$$\bar{x}_h = \frac{x_{h+}}{n_h} = \frac{\sum_{i=1}^{n_h} x_{h,i}}{n_h} \, .$$

Note that $\bar{x}_h$ is an unbiased estimator of the population mean within stratum h. The re-scaled sample total

$$x'_h = \frac{N_h}{n_h} x_{h+} = N_h \bar{x}_h$$

is an unbiased estimator of the population total within stratum h.

The entire population total X is estimated by

$$x'_{str} = \sum_{h=1}^{L} x'_h \, .$$

The entire population mean $\bar{X}$ is estimated by

$$\bar{x}_{str} = \frac{\sum_{h=1}^{L} N_h \bar{x}_h}{N} \, .$$

The entire population proportion P_X is estimated by

$$p_{str} = \frac{\sum_{h=1}^{L} N_h \, p_h}{N} \, .$$

6.2, 6.3 Confidence Intervals

The estimators $x'_{str}, \bar{x}_{str}$, and p_{str} are unbiased estimators of their respective population parameters. Their variances are

$$\mathbb{V}ar(x'_{str}) = \sum_{h=1}^{L} (N_h^2) \Big(\frac{\sigma_h^2}{n_h} \Big) \Big(\frac{N_h - n_h}{N_h - 1} \Big) \, ,$$

$$\mathbb{V}ar(\bar{x}_{str}) = \frac{1}{N^2}\sum_{h=1}^{L}(N_h^2)\Big(\frac{\sigma_h^2}{n_h}\Big)\Big(\frac{N_h - n_h}{N_h - 1}\Big)\,,$$

and

$$\mathbb{V}ar(p_{str}) = \frac{1}{N^2}\sum_{h=1}^{L}(N_h^2)\Big(\frac{P_h(1-P_h)}{n_h}\Big)\Big(\frac{N_h - n_h}{N_h - 1}\Big)\,.$$

Define the sample variance within stratum h by

$$s_h^2 = \frac{\sum_{i=1}^{n_h}(x_{h,i} - \bar{x}_h)^2}{n_h - 1}\,.$$

We know that it is an unbiased estimator of

$$\sigma_h^2\,\frac{N_h}{N_h - 1}\,.$$

Thus, a $100(1-\alpha)\%$ CI for X is

$$x'_{str} \pm z_{1-\alpha/2}\sqrt{\sum_{h=1}^{L}(N_h^2)\Big(\frac{s_h^2}{n_h}\Big)\Big(\frac{N_h - n_h}{N_h}\Big)}\,.$$

A $100(1-\alpha)\%$ CI for $\bar{X}$ is

$$\bar{x}_{str} \pm z_{1-\alpha/2}\sqrt{\sum_{h=1}^{L}\Big(\frac{N_h}{N}\Big)^2\Big(\frac{s_h^2}{n_h}\Big)\Big(\frac{N_h - n_h}{N_h}\Big)}\,.$$

A $100(1-\alpha)\%$ CI for P_X is

$$p_{str} \pm z_{1-\alpha/2}\sqrt{\sum_{h=1}^{L}\Big(\frac{N_h}{N}\Big)^2\Big(\frac{p_h(1-p_h)}{n_h - 1}\Big)\Big(\frac{N_h - n_h}{N_h}\Big)}\,.$$

6.5.3, 6.5.4 Optimal Allocation

Suppose we want to find $n_1, \ldots, n_L$ such that the variance of x'_{str} is minimized. The sampling method that uses these values of n_h is called optimal allocation. The values are

$$n_h \;=\; (n)\Big(\frac{N_h\sigma_h}{\sum_{i=1}^{L} N_h\,\sigma_h}\Big)\,.$$

Assume that the cost of sampling an element from stratum h is C_h. Then the total cost of sampling n elements is

$$C = \sum_{h=1}^{L} n_h\,C_h\,.$$

For a fixed total cost C, the optimal allocation is given by

$$n_h \;=\; \frac{N_h\,\sigma_h/\sqrt{C_h}}{\sum_{h=1}^{L}(N_h\sigma_h\,\sqrt{C_h})}\,C\,.$$

Example (on pages 160 – 161) It is given that

$N_1 = 150,000,\ N_2 = 110,000,\ C = \$10,000,\ C_1 = \$0.32,\ C_2 = \$0.98,\ \sigma_1 = \sigma_2/2\,.$

We compute

$$n_1 = \frac{150,000(\sigma_2/2)/\sqrt{0.32}}{150,000(\sigma_2/2)\sqrt{0.32} + 110,000(\sigma_2)\sqrt{0.98}}\,10,000 = 8,762\,,$$

and

$$n_1 = \frac{110,000(\sigma_2)/\sqrt{0.98}}{150,000(\sigma_2/2)\sqrt{0.32} + 110,000(\sigma_2)\sqrt{0.98}}\,10,000 = 7,343\,.$$

Note that $(\$0.32)(8,762) + (\$0.98)(7,343) = \$10,000\,.$