BIOS 7250 Stratification

I. Stratified Random Sampling (STRS)

A sampling plan in which the population elements are separated into non overlapping groups, called **strata**, and then a simple random sample is drawn from each stratum.

A. Advantages of STRS over SRS

1. Homogeneity within each stratum – smaller variation within strata produces stratified sampling estimators with smaller variance than SRS estimators – for the same sample size.
2. Separate estimators for population parameters can be obtained for each stratum.
3. Cost can be less for STRS than SRS if strata represent locations.

B. How to draw a STRS

1. Determine the population we wish to estimate
2. Clearly specify strata so that they are homogeneous within with respect to another variable (eg. location, income, severity of disease, age). Each element belongs to one and only one stratum.
3. Use SRS techniques to select samples from each stratum.

C. Examples:

1. Pages 123-127.

Table 5.1 Population, N = 8.

Table 5.2 for n = 3, $\binom{8}{3} = 56$ possible SRS's, and

$$x' = N\bar{x} = 8\bar{x} \text{ for each sample}$$

Table 5.3 2 strata: Urban, $N_1 = 3, n_1 = 1$; Rural, $N_2 = 5, n_2 = 2$

Table 5.4 $\binom{3}{1}\binom{5}{2} = (3)(10) = 30$ possible STRS's.

In General $\binom{N_1}{n_1}\binom{N_2}{n_2}\cdots\binom{N_L}{n_L} \leq \binom{N}{n}$.

$x_1' = 3\bar{x}_1,\ x_2' = 5\bar{x}_2$, and $x_{STR}' = x_1' + x_2'$

2. Exercises 5.5 and 5.6, pages 140, 141.

D. Population Characteristics for Strata

Notation:

1. L: number of mutually exclusive and exhaustive strata.
2. N: total population size.

3. N_h : size of stratum h (h =1,2,…,L)

$$N = \sum_{h=1}^{L} N_h$$

4. X: variable (characteristic of the population).
5. $X_{h,i}$: value of X for the i^{th} elementary unit within stratum h.
6. X_{h+} : total of X within stratum h,

$$X_{h+} = \sum_{i=1}^{N_h} X_{hi} \qquad (5.1)$$

7. X: total of X for the entire population

$$X = \sum_{h=1}^{L} X_{h+} = \sum_{h=1}^{L} \sum_{i=1}^{N_h} X_{hi} \qquad (5.1)$$

8. $\overline{X}_h$: mean of X for stratum h

$$\overline{X}_h = \frac{\sum_{i=1}^{N_h} X_{hi}}{N_h} = \frac{X_{h+}}{N_h} \qquad (5.2)$$

9. $\overline{X}$: mean of X for the entire population

$$\overline{X} = \frac{X}{N} = \frac{\sum_{h=1}^{L} X_{h+}}{N} = \frac{\sum_{h=1}^{L} N_h \overline{X}_h}{N} = \sum_{h=1}^{L} W_h \overline{X}_h \qquad (5.2)$$

Here $\boxed{W_h = \frac{N_h}{N}}$ is a weight $\Rightarrow$ $\overline{X}$ is a weighted average of stratum means (weights proportional to stratum size).

10. σ_{hx}^2 : Variance of X within stratum h,

$$\sigma_{hx}^2 = \frac{\sum_{i=1}^{N_h} (X_{hi} - \overline{X}_h)^2}{N_h} \qquad (5.4)$$

11. V_{hx} : Coefficient of Variation within stratum h

$$V_{hx} = \frac{\sigma_{hx}}{\overline{X}_h} \qquad (5.5)$$

12. Y_{hi} : the presence or absence of a dichotomous attribute Y for the i^{th} element in stratum h.

$$Y_{hi} = \begin{cases} 1 & \text{if yes for the ith element in stratum } h \\ 0 & \text{if not} \end{cases}$$

13. P_{hy} : the proportion of elements in stratum h with the presence of attribute Y .

$$P_{hy} = \frac{\sum_{i=1}^{N_h} Y_{hi}}{N_h} = \frac{Y_{h+}}{N_h} \qquad (5.3)$$

14. P_y : the proportion of elements in the entire population with the presence of attribute Y .

$$P_Y = \frac{Y}{N} = \frac{\sum_{h=1}^{L} Y_{h+}}{N} = \frac{\sum_{h=1}^{L} N_h P_{hy}}{N} = \sum_{i=1}^{L} W_h P_{hy} \qquad (5.3)$$

15. Example – Pages 130 – 133.

E. Sample Characteristics for Strata

1. n: total sample size.
2. n_h : sample size of stratum h

$$n = \sum_{h=1}^{L} n_h$$

3. $x_{h,i}$: value of X for the i^{th} sampled elementary unit in stratum h.
4. x_{h+} : sample total for stratum h,

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

5. $\bar{x}_h$: sample mean of X for stratum h

$$\bar{x}_h = \frac{\sum_{i=1}^{n_h} x_{hi}}{n_h} = \frac{x_{h+}}{n_h} \qquad (5.7)$$