Advanced Methods in Survey Sampling

Homework

1. Verify that $V(I_i(s)) = \pi_i(1 - \pi_i)$, and $V(w_i) = (1 - \pi_i)/\pi_i$.

2. Express the sample size n as sum of random indicator variables, calculate the expected value of this sum and show that $n = \sum_{i=1}^{N} \pi_i$.

3. In a population of four units $U = \{1, 2, 3, 4\}$, the values of a variable y are $y_1 = 10$, $y_2 = 12$, $y_3 = 15$, $y_4 = 22$, and thus the total is $Y = 59$. Considering all six possible samples of size $n = 2$, all with probability of selection equal to $1/6$, calculate the six possible values of the estimator $\hat{Y}$ of the total Y and the differences of these estimates from the true value Y. What do you observe? Calculate the mean of these six estimates. What do you observe?

4. For the population $U = \{1, 2, 3, 4, 5, 6, 7, 8\}$, consider the possible samples of size $n = 4$, $s_1 = \{1, 3, 5, 6\}$, $s_2 = \{2, 3, 7, 8\}$, $s_3 = \{1, 4, 6, 8\}$, $s_4 = \{2, 4, 6, 8\}$, $s_5 = \{4, 5, 7, 8\}$, with respective probabilities $p(s_1) = 1/8$, $p(s_2) = 1/4$, $p(s_3) = 1/8$, $p(s_4) = 3/8$, $p(s_5) = 1/8$. [The rest of the possible samples have $p(s) = 0$.] Find the probability of selection π_i for all units $i \in U$.

5. For simple random sampling (where $\pi_i = n/N$) find the estimator of the mean $\hat{\bar{Y}}$. What do you observe regarding the relationship of sample with population?

6. The file "syc" and its description (found in e-class/έγγραφα) contains selected variables from the *Survey of Youth in Custody*, which collected data from youth who are in custody in correction centers in the United States. The selection probabilities of the sample units are unequal. Use the weights given by the variable *finalwt* to estimate the mean age of first arrest (variable *agefirst*). Compare this estimate with the one obtained without the use of weights (or assuming that the weights are equal).
   Estimate the proportions of the youth that: (a) are of age 14 and lower, (b) are male, (c) are in custody for violent crime (*crimtype*1), (d) have used drugs (*everdrug*), (e) have both characteristics (c) (d). For the calculations omit the missing values in the variables *agefirst* and *everdrug*.

7. For the *ratio estimator*, where we assume $y_i \approx Bx_i$ and $q_i = x_i$, show that $\hat{B} = \hat{Y}/\hat{X}$ and $\hat{Y}^R = \hat{Y}X/\hat{X}$.

8. The data set "cherry" (found together with its description in e-class/έγγραφα) contains measurements of diameter, height and timber volume for a sample of thirty one ($n = 31$) cherry trees.
   (a) Do a scatter plot showing the linear relationship of volume with diameter (y the volume, x the diameter), and compute the correlation coefficient (use cor(y,x) in R).

(b) Suppose that these 31 trees have been selected with simple random sampling from a forest of =2967 cherry trees ($\pi_i = n/N$), and that the sum of the diameters for all trees in the forest is $X = 41835$. Use the ratio estimator, with auxiliary variable x the diameter, to estimate the total timber volume for all cherry trees in the forest.

9. In the special *case* 1 of calibration, verify that $g_i = N_j/\hat{N}_j$ for all $i \in U_j$. (Assume that $q_i = 1$ for all $i \in U$).

10. For simple random sampling (where $\pi_i = n/N$) and for some sub-population $U_d \subset U$, find $\hat{N}_d$, $\hat{Y}_d$, $\hat{\bar{Y}}_d$ and $\hat{P}_d$. What do you observe for $\hat{\bar{Y}}_d$ and $\hat{P}_d$?

11. If calibration of the special *case* 1 has been carried out, what would be the form of $\hat{Y}_d^C$ for some $U_d \subset U$?

12. For the calculated estimator $\hat{Y}$ of exercise 8, calculate the Jackknife variance $\hat{V}_{JK}(\hat{Y})$ with $k = n$, $m = 1$ (find the formula in the slides). Then calculate the Jackknife variance $\hat{V}_{JK}(\hat{Y}^R)$, of the calculated ratio estimator $\hat{Y}^R$, using the formula $[(n-1)/n]\sum_{k=1}^{n}(\hat{\theta}_{(k)} - \hat{\bar{\theta}})^2$, with the suitable $\hat{\theta}_{(k)}$, and with $\hat{\bar{\theta}} = (1/n)\sum_{k=1}^{n}\hat{\theta}_{(k)}$. Compare the jackknife variances $\hat{V}_{JK}(\hat{Y})$ and $\hat{V}_{JK}(\hat{Y}^R)$.

13. In the formula of $\hat{V}_{JK}(\hat{Y})$ in the case of stratified multistage sampling, use the expression of $\hat{Y}$ (splitting s appropriately), analogous to the expression of $\hat{Y}_{(hk)}$ that exist in the slides, and show that

$$\hat{V}_{JK}(\hat{Y}) = \sum_{h=1}^{H}\frac{K_h - 1}{K_h}\sum_{k=1}^{K_h}\left(\hat{Y}_{(hk)} - \hat{Y}\right)^2 = \sum_{h=1}^{H}\frac{K_h - 1}{K_h}\sum_{k=1}^{K_h}\left(\frac{1}{K_h - 1}\hat{Y}_{s_h - s_{hk}} - \hat{Y}_{s_{hk}}\right)^2,$$

where $\hat{Y}_{s_h - s_{hk}} = \sum_{s_h - s_{hk}} w_i y_i$ and $\hat{Y}_{s_{hk}} = \sum_{s_{hk}} w_i y_i$. Interpret the term $\frac{1}{K_h - 1}\hat{Y}_{s_h - s_{hk}} - \hat{Y}_{s_{hk}}$ for stratum h.

14. In the file "syc" (see exercise 8), the variable *everdrug* has 6 missing values. Do imputation with the methods:
(a) Cell random hot-deck imputation, with the 8 cells (or classes) defined by the variables *sex* and *educ* (4 categories, see description of the data file in e-class). To select randomly one unit from a set s of n units, use the R command s[sample(n,1)].
(b) Nearest-neighbor hot-deck imputation, with distance measure defined by the variables *sex* and *educ*. (Use this method for the units with values of the variable *educ* in the interval [01,12]).