

Name:

Problem 1:

Consider a population U and a sampling design $p(\cdot)$ with inclusion probabilities $\pi_k, k = 1, \dots, N$ and $\pi_{kl} > 0$ for $k, l = 1, \dots, N$. **Assume that the population mean \bar{y}_U is known.**

(a) Propose an unbiased estimator for the population variance, S_U^2 ,

$$S_U^2 = \frac{1}{N-1} \sum_{k \in U} (y_k - \bar{y}_U)^2.$$

(Hint: what is the π -estimator of $\frac{1}{N-1} \sum_{k \in U} y_k$?) (10 pts)

(b) Give an expression for the variance of this estimator as a function of the y_k, π_k, π_{kl} and \bar{y}_U .

(5 pts)

(c) Give an unbiased estimator for the variance in (b). (5 pts)

Problem 2:

In class, we studied the *design efficiency* as a comparison of variances between designs. Another (and equivalent) way to measure the relative precision of different designs is to compare the sample sizes (or expected sample sizes, for random-size designs) required under the designs to achieve a certain variance.

(a) Estimate the variances of the π -estimators of the totals for the following three subpopulations respectively, using the provided sample-based quantities:

$$U_A : \text{SI design, } N = 1000, n = 100, S_{SA}^2 = 5$$

$$U_B : \text{SI design, } N = 8500, n = 150, S_{SB}^2 = 10$$

$$U_C : \text{BE design, } N = 500, \pi = 0.1, \sum_{SC} y_k^2 = 100,$$

where SA , SB and SC are the samples in subpopulation A , B and C respectively. (10 pts)

(b) Assume the population $U = \{U_A, U_B, U_C\}$ and we want to use the stratification design to estimate the whole population total, where the designs in different strata are specified in (a). Using the results from (a), estimate the variance of the π -estimator for the population total under this stratification design. (5 pts)

(c) Assume that the overall population variance is $S_U^2 = 30$, what sample size would be required under an unstratified SI design over the population U to achieve that same variance in (b)? You may ignore the finite population correction in your calculations of the SI variance for this new sample (5 pts).

(d) What design (ST over three subpopulations vs SI over the entire population) is more efficient? Explain carefully what is causing the difference in efficiency between both designs. (Hint: compare the sample sizes under two designs that are used to achieve the same variances) (5 pts).

Problem 3:

A supermarket chain has 450 stores spread over 32 cities. A company official wants to estimate the proportion of stores in the chain that do not meet a specified cleanliness criterion. Because of travel costs, she decides to select a two-stage SISI element sampling design containing (approximately) half the stores within each of 4 cities. The data collected are in the following table:

City	Total # of stores	Sampled stores	Number not meeting criterion
1	25	13	3
2	10	5	1
3	18	9	4
4	16	8	2

(a) Write down a π -estimator for the proportion of stores not meeting the cleanliness criterion, and compute its value. (10 pts)

(b) Write down a variance estimator for the estimator in (a). Just give the formula. [BONUS POINTS: Compute its value. You may only need to calculate the first term which normally accounts for 98% of the total variation. (5 pts)] (10 pts)

(c) Suppose now that the information about the total number of stores in the chain was not available. Propose a new estimate for the proportion of stores not meeting the cleanliness criterion. Is this estimator unbiased? Explain carefully how would you proceed to specify its variance? [BONUS QUESTION: give an expression for its (approximate) variance (5 pts).] (10 pts)

Problem 4:

Suppose that for some finite population $U = \{1, 2, \dots\}$ we draw a sample $S \subset U$ using the sampling design $p(\cdot)$, with first-order inclusion probabilities π_k and second-order inclusion probabilities π_{kl} . You may assume that $\pi_{kl} > 0$ for all $k, l \in U$. We are interested in estimating the following quantity for U ,

$$\hat{\theta} = \hat{t}_{x\pi} \hat{t}_{y\pi} = \sum_{k \in S} \frac{x_k}{\pi_k} \sum_{k \in S} \frac{y_k}{\pi_k}$$

(a) Write down the first-order Taylor series linearization of $\hat{\theta}$. Simplify it to obtain the form

$$\hat{\theta} \approx \text{constant} + \sum_{k \in S} \frac{u_k}{\pi_k}. \quad (10 \text{ pts})$$

(b) Give an approximate expression $AV(\hat{\theta})$ for the variance of $\hat{\theta}$. (5 pts)

(c) Give an estimator of $AV(\hat{\theta})$ above. (5 pts)

(d) Suppose now that the sampling design for this problem was BE with selection parameter π . Rewrite the answers in parts (b) and (c) for this design. (Hint: what are π_k and π_{kl} under BE design?) (5 pts)