Assignement 1
MAT 5192
Winter 2021
Hand-in on February 15, 2021
(Please work in groups of 2)

1. (10 marks) Let *U* be a population of size *N*. From *U*, we first select a SRSWOR, S_1, of size n_1. Then, from S_1, we select a SRSWOR, S_2, of size n_2. Show that S_2 is a SRSWOR of size n_2 selected from *U*.

2. (20 marks) Let *U* be a population of size *N*. From *U*, we first select a SRSWOR, S_1, of size n_1. Then, from $U \setminus S_1$, we select a SRSWOR, S_2, of size n_2. The final sample is $S = S_1 \cup S_2$ of size $n = n_1 + n_2$. Let $\bar{y}_1 = \frac{1}{n_1}\sum_{k\in S_1} y_k$ and $\bar{y}_2 = \frac{1}{n_2}\sum_{k\in S_2} y_k$.

(a) (5 marks) Show that *S* is a SRSWOR of size *n* selected from *U*. What is interesting about this result?

(b) (10 marks) Show that $V_p\left(\bar{y}_2\right) = \left(1 - \frac{n_2}{N}\right)\frac{S_y^2}{n_2}$, where $S_y^2 = \frac{1}{N-1}\sum_{k\in U}\left(y_k - \bar{Y}\right)^2$ with $\bar{Y} = \frac{1}{N}\sum_{k\in U} y_k$. **Hint**: You may want to show first that S_2 is a SRSWOR from *U*.

(c) (5 marks) Use (a) and (b) to determine $Cov_p\left(\bar{y}_1, \bar{y}_2\right)$.

3. (15 marks) The files Parishes_pop shows data on a population of parishes in Quebec. The file Parishes_sample shows data from a SRSWOR of size 50 selected from the population.

(a) (5 marks) From the sample data, estimate the total number of births, t_y, in the population of parishes and construct a 95% confidence interval for t_y.

(b) (5 marks) The parish no. 71105 (Saint-Lazare) seems to be particularly large: its population size (in 2001) as well as its number of births are both larger than the corresponding population means. An analyst thinks that this parish should always be included in the sample and that the sample should be completed with 49 parishes selected from the 209 remaining parishes according to SRSWOR. The analyst thinks that this procedure will lead to a more precise estimator. Use the population data to confirm his belief.

(c) (5 marks) Using the sample data, estimate the proportion P_y of parishes in which there had been 10 births or more and determine a 95% confidence interval for P_y.

4. (20 marks) From a population *U*, a sample *S*, of size *n*, is selected according to a given sampling design, $p(S)$. Let θ_N be a finite population parameter and let $\hat{\theta}$ be an estimator of θ_N. As a measure of influence of a unit, we consider the conditional bias of unit *k* defined as

$$B_k = E_p\left(\hat{\theta} - \theta_N \mid I_k = 1\right).$$

(a) (7 marks) If $\theta_N \equiv t_y$ and $\hat{\theta} \equiv \hat{t}_{y,\pi}$, show that

$$B_k = \sum_{l \in U}\left(\frac{\pi_{kl}}{\pi_k \pi_l} - 1\right) y_l = \left(d_k - 1\right) y_k + \sum_{\substack{l \in U \\ l \neq k}}\left(\frac{\pi_{kl}}{\pi_k \pi_l} - 1\right) y_l.$$

What is the value of B_k when $\pi_k = 1$? Interpret this result.

(b) (3 marks) Express the design variance of $\hat{t}_{y,\pi}$ (for a fixed- or random-size sampling design) as a function of B_k, $k = 1, ..., N$.

(c) (5 marks) Give the expression of B_k in the cases of simple random sampling without replacement and Bernoulli sampling.

(d) (5 marks) Suggest a conditionally *p*-unbiased estimator of B_k, $\hat{B}_k$, in the sense that $B_k = E_p\left(\hat{B}_k \mid I_k = 1\right)$.

5. (10 marks) Let *S* be a sample selected by a Bernoulli design with probability π. Let n_s denote the random size of *s*. Show that the conditional probability of obtaining *S* given n_s is the same as the probability of a SRSWOR of the fixed size n_s from *N*.

6. (15 marks) Let *S* be a sample realized by the Bernoulli design with probability π.

(a) (7 marks) To estimate the population mean $\bar{Y} = N^{-1}\sum_{k \in U} y_k$, we consider the following estimator:

$$\bar{y}^* = \begin{cases} \bar{y} & \text{if } n_s \geq 1 \\ 0 & \text{if } n_s = 0 \end{cases}$$

where $\bar{y} = \frac{1}{n_s}\sum_{k \in S} y_k$ denote the sample mean of the *y*-values in the sample. Show that the relative bias of $\bar{y}^*$ is given by

$$RB_p\left(\bar{y}^*\right) = \frac{E_p\left(\bar{y}^*\right) - \bar{Y}}{\bar{Y}} = -P\left(n_s = 0\right) = -\left(1 - \pi\right)^N.$$

Note that the relative bias of $\bar{y}^*$ is negligible if the population size *N* is large.

(b) (8 marks) To estimate the population variability of the *y*-values, $S_y^2 = \frac{1}{N-1}\sum_{k\in U}\left(y_k - \bar{Y}\right)^2$, we consider the following estimator:

$$s_y^{2*} = \begin{cases} s_y^2 & \text{if } n_s \geq 2 \\ 0 & \text{if } n_s \leq 1 \end{cases}$$

where $s_y^2 = \frac{1}{n_s - 1}\sum_{k\in S}\left(y_k - \bar{y}\right)^2$. Show that the relative bias of s_y^{2*} is given by

$$RB_p\left(s_y^{2*}\right) = \frac{E_p\left(s_y^{2*}\right) - S_y^2}{S_y^2} = -P\left(n_s \leq 1\right) = -\left(1-\pi\right)^{N-1}\left[1 + \left(N-1\right)\pi\right].$$

Once again, note that the relative bias of s_y^{2*} is negligible if the population size *N* is large.

7. (40 marks) Suppose that we want to estimate a population total $t_y = \sum_{k\in U} y_k$. We select a simple random sample *with replacement* of fixed size *m* from *U*. We propose the following estimator of t_y:

$$\hat{t}_{y,HH} = \frac{N}{m}\sum_{k\in U} Q_k y_k,$$

where Q_k is the number of times that unit *k* is selected in the sample, $k = 1,\ldots, N$.

(a) (5 marks) Let $\mathbf{Q} = \left(Q_1,...,Q_N\right)$. What is the distribution of **Q**? Deduce $E_p\left(Q_k\right)$, $V_p\left(Q_k\right)$ and $Cov_p\left(Q_k, Q_l\right)$ for $k \neq l$.

(b) (5 marks) Show that $\hat{t}_{y,HH}$ is design-unbiased for t_y.

(c) (5 marks) Show that $V_p\left(\hat{t}_{y,HH}\right) = \frac{N^2}{m}\tilde{S}_y^2$, where $\tilde{S}_y^2 = \frac{1}{N}\sum_{k\in U}\left(y_k - \bar{Y}\right)^2$.

(d) (5 marks) Find a design-unbiased estimator of $V_p\left(\hat{t}_{y,HH}\right)$.

(e) (20 marks) We consider the following alternative estimator of t_y:

$$\hat{t}_{y,alt} = \frac{N}{n_s}\sum_{k\in S} y_k,$$

where $n_s \le m$ denotes the number of distinct elements and *S* denotes the sample of distinct elements. Note that n_s is a random variable.

(i) (5 marks) We can show that, given n_s, s is a SRSWOR from the population. Une this fact to show that $\hat{t}_{y,alt}$ is design-unbiased for t_y.

(ii) (5 marks) Show that the variance of $\hat{t}_{y,alt}$ is given by

$$V_p\left(\hat{t}_{y,alt}\right) = N^2\left[E_{n_s}\left(\frac{1}{n_s}\right) - \frac{1}{N}\right]S_y^2.$$

(iii) (10 marks) Consider the following variance estimators:

$$\hat{V}_1 = N^2\left\{\frac{1}{n_s} - \frac{1}{N}\right\}s_y^2$$

and

$$\hat{V}_2 = \frac{n_s\left(n_s - 1\right)}{N-1}\left[NE\left(\frac{1}{n_s}\right) - 1\right]\frac{s_y^2}{\bar{\pi}},$$

where $\bar{\pi} = 1 - 2\left(1 - \frac{1}{N}\right)^m + \left(1 - \frac{2}{N}\right)^m$ and $s_y^2 = \frac{1}{n_s - 1}\sum_{k\in S}\left(y_k - \bar{y}\right)^2$ with $\bar{y} = \frac{1}{n_s}\sum_{k\in S} y_k$.

Show that both $\hat{V}_1$ and $\hat{V}_2$ are design-unbiased for $V_p\left(\hat{t}_{y,alt}\right)$.

Hint: You can use the following identities:

$$E_{n_s}\left(n_s\right) = N\left\{1 - \left(\frac{N-1}{N}\right)^m\right\}$$

and

$$V_{n_s}\left(n_s\right) = \frac{\left(N-1\right)^m}{N^{m-1}} + \left(N-1\right)\frac{\left(N-2\right)^m}{N^{m-1}} - \frac{\left(N-1\right)^{2m}}{N^{2m-2}}.$$

8. (40 marks) Construct a simulation program for comparing the performance of the Horvitz-Thompson estimator in terms of bias and efficiency under three sampling designs: simple random sampling without replacement, systematics sampling and Bernoulli sampling. Use the three populations (pop1, pop2 and pop3) provided. As a measure of bias, compute the Monte Carlo relative bias of the Horvitz-Thompson and, as a measure of efficiency, compute its Monte Carlo mean square error. Compute the Monte Carlo design effect for each sampling design. Use *n* = 100. Discuss the Monte Carlo results. Are the results consistent with the theory?