Sample Sampling Theory Assignment Question

Instructor - Course

[Total: 14 Marks] - In this question, we will look into the details of *Simple Random Sampling Without Replacement* (SRSWOR), a very simple probablity sampling method to select a sample of fixed size n from a population of size N. Under SRSWOR, every sample $\boldsymbol{S}$ of fixed size n has an equal probability of being selected and the sampling design probability is given by

$$P(\boldsymbol{S}) = \begin{cases} 1/\binom{N}{n} & , \text{ if } \boldsymbol{S} \text{ has size } n \\ 0 & , \text{ otherwise} \end{cases} . \tag{1}$$

(a) **[4 Marks]** For a fixed sample size n, the total number of possible samples is $\binom{N}{n}$ for a population of size N.

<u>Show that</u> the first order inclusion probability is given by

$$\pi_i = \mathbb{P}(i \in \boldsymbol{S}) = \frac{n}{N} \ , \forall\, i \in \boldsymbol{U} \ , \tag{2}$$

and **<u>explain intuitively</u>** why the second order inclusion probability is given by

$$\pi_{ij} = \mathbb{P}(i, j \in \boldsymbol{S}) = \frac{n(n-1)}{N(N-1)} \ , \ i \neq j \ , \forall\, i, j \in \boldsymbol{U} \ . \tag{3}$$

Remark: *Formula (2) suggests that every unit $i \in \boldsymbol{U}$ has an equal probability of being included in the selected sample $\boldsymbol{S}$ under SRSWOR.*

(b) **[2 Marks]** Suppose we collect a sample $\boldsymbol{S} = \{y_1, \ldots, y_n\}$ using SRSWOR. Let $\mu_y = \frac{1}{N}\sum_{i \in \boldsymbol{U}} y_i$ be the population mean and $\bar{y} = \sum_{i \in \boldsymbol{S}} y_i$ be the sample mean.
Let

$$I_i = \begin{cases} 1 \ , \ i \in \boldsymbol{S} \\ 0 \ , \ k \notin \boldsymbol{S} \end{cases} \ , \forall\, i \in \boldsymbol{U} \tag{4}$$

be the inclusion indicator for unit i being selected in the sample $\boldsymbol{S}$. So, I_i follows a Bernoulli distribution with parameter $p = \pi_i$ (you do not need to show this).

Use formula (4) to **<u>show that</u>**

$$\mathbb{E}[\bar{y}] = \mu_y \tag{5}$$

that is, the sample mean $\bar{y}$ is a design-unbiased estimator for the population mean μ_y under SRSWOR.

(c) **[2 Marks]** Based on the definition of I_i given in (4), **show that**

$$\mathbb{E}(I_i I_j) = \begin{cases} \frac{n(n-1)}{N(N-1)} & , \text{ for } i \neq j \\ \frac{n}{N} & , \text{ for } i = j \end{cases} . \tag{6}$$

(d) **[6 Marks]** Let $\sigma_y^2 = (N-1)^{-1} \sum_{i \in \boldsymbol{U}} (y_i - \mu_y)^2$ be the population variance. Use the results from part (a), (b), (c), **<u>derive</u>** the design-based variance of $\bar{y}$ under SRSWOR

$$\text{Var}(\bar{y}) = \left(1 - \frac{n}{N}\right) \frac{\sigma_y^2}{n} \ . \tag{7}$$

Hint: You may use the result $\sum_{i \in \boldsymbol{U}} y_i^2 - N\mu_y^2 = \sum_{i \in \boldsymbol{U}} (y_i - \mu_y)^2$ directly without showing it.

Remark: *Based on formula (7), we can replace σ_y^2 by the sample variance $S_y^2 = \frac{1}{n-1}\sum_{i \in \boldsymbol{S}} (y_i - \bar{y})^2$ to obtain a design-unbiased estimator for the variance of the sample mean $\bar{y}$ (you do not need to show this).*

Solution and Grading Scheme

Part (a) - [4 Marks]

The total number of possible samples is $\binom{N}{n}$, and the number of samples containing unit i is $\binom{N-1}{n-1}$. So, the inclusion probability for unit i is

$$\pi_i = \frac{\binom{N-1}{n-1}}{\binom{N}{n}} = \frac{\frac{(N-1)!}{(n-1)!(N-n)!}}{\frac{N!}{n!(N-n)!}} = \frac{(N-1)!}{N!}\frac{n!}{(n-1)!} = \frac{n}{N} \ .$$

Since we use a fixed sample size n, after picking unit i from the population, we still need $n-1$ units from the remaining population of size $N-1$. So, the probability of picking unit j after picking i is $\frac{n-1}{N-1}$. Multiplying the two probabilities together gives

$$\pi_{ij} = \frac{n(n-1)}{N(N-1)} \ .$$

Grader:

1 mark for finding the correct total number of possible samples $\binom{N}{n}$

1 mark for finding the correct number of samples containing unit i $\binom{N-1}{n-1}$

$\star$**1 mark** for showing simplification and arriving to the final expression $\frac{n}{N}$

1 mark for explaining the intuition for finding π_{ij} , anything reasonable and similar to the above is fine

$\star$**Deduct 1 mark** if the student did not get $\binom{N}{n}$ or $\binom{N-1}{n-1}$ but has shown derivation using their own values

$\star$**Deduct no mark** if the student made simplification errors using the correct values $\binom{N}{n}$ and $\binom{N-1}{n-1}$

Part (b) - [2 Marks]

Since $I_i \sim \text{Bernoulli}(\pi_i)$,

$$\mathbb{E}(I_i) = \pi_i = \frac{n}{N} \ .$$

We rewrite the sample mean using the inclusion indicator as

$$\bar{y} = \frac{1}{n}\sum_{i \in \boldsymbol{S}} y_i = \frac{1}{n}\sum_{i \in \boldsymbol{U}} I_i \cdot y_i \ \ ,$$

and taking the expectation gives

$$\mathbb{E}[\bar{y}] = \mathbb{E}\left[\frac{1}{n}\sum_{i \in \boldsymbol{U}} I_i \cdot y_i\right] = \frac{1}{n}\sum_{i \in \boldsymbol{U}} \mathbb{E}[I_i] \cdot y_i = \frac{1}{n}\sum_{i \in \boldsymbol{U}} \frac{n}{N} y_i = \frac{1}{N}\sum_{i \in \boldsymbol{U}} y_i = \mu_y \ .$$

Grader:

1 mark for correctly rewriting the sample mean using indicator I_i

1 mark for correctly evaluating the expectation of I_i

Deduct 1.5 marks if the student wrote something like $\mathbb{E}[y_i] = \mu_y$ and concluded unbiasness from there

Deduct 1.5 marks if the student only wrote sample mean is an unbiased estimator for the population mean

Part (c) - [2 Marks]

Notice that $I_iI_j = 1$ when both $I_i = 1$ and $I_j = 1$, and $I_iI_j = 0$ otherwise. So, when $I_iI_j = 1$, it means both units i and j are selected in $\boldsymbol{S}$.
So,

$$\mathbb{E}(I_iI_j) = \mathbb{P}(I_i = 1, I_j = 1) = \mathbb{P}(i, j \in \boldsymbol{S}) = \pi_{ij} = \frac{n(n-1)}{N(N-1)} \text{ , for } i \neq j \ .$$

When $i = j$, $I_iI_j = I_i^2 = I_i$ since it is an indicator. So, $\mathbb{E}(I_iI_j) = \mathbb{E}(I_i) = \pi_i = \frac{n}{N}$ when $i = j$.

Grader:

1 mark for correctly showing $\mathbb{E}(I_iI_j) = \mathbb{P}(i, j \in \boldsymbol{S}) = \pi_{ij}$

1 mark for correctly showing $\mathbb{E}(I_iI_i) = \mathbb{E}(I_i) = \pi_i$

Part (d) - [6 Marks]

$$\begin{aligned}
\mathrm{Var}(\bar{y}) = \mathrm{Var}\left[\frac{1}{n}\sum_{i\in\boldsymbol{S}} y_i\right] &= \frac{1}{n^2}\mathrm{Var}\left[\sum_{i\in\boldsymbol{S}} y_i\right] \\
&= \frac{1}{n^2}\mathrm{Var}\left[\sum_{i\in\boldsymbol{U}} I_i \cdot y_i\right] \\
&= \frac{1}{n^2}\left[\sum_{i\in\boldsymbol{U}} y_i^2 \mathrm{Var}(I_i) \ + \sum_{i,j\in\boldsymbol{U},\ i\neq j} \mathrm{Cov}(I_i\cdot y_i, I_j \cdot y_j)\right] \\
&= \frac{1}{n^2}\left[\frac{n}{N}\left(1-\frac{n}{N}\right)\sum_{i\in\boldsymbol{U}} y_i^2 \ + \sum_{i,j\in\boldsymbol{U},\ i\neq j} y_iy_j \cdot \mathrm{Cov}(I_i, I_j)\right] \\
&= \frac{1}{n^2}\left[\frac{n}{N}\left(1-\frac{n}{N}\right)\sum_{i\in\boldsymbol{U}} y_i^2 \ + \sum_{i,j\in\boldsymbol{U},\ i\neq j} y_iy_j \cdot \Big(\mathbb{E}[I_iI_j] - \mathbb{E}[I_i]\mathbb{E}[I_j]\Big)\right] \\
&= \frac{1}{n^2}\left[\frac{n}{N}\left(1-\frac{n}{N}\right)\sum_{i\in\boldsymbol{U}} y_i^2 \ + \sum_{i,j\in\boldsymbol{U},\ i\neq j} y_iy_j \cdot \left(\frac{n(n-1)}{N(N-1)} - \frac{n}{N}\frac{n}{N}\right)\right] \\
&\quad \vdots \qquad \text{(steps omitted)} \\
&= \frac{1}{n}\frac{N-n}{N^2}\left[\sum_{i\in\boldsymbol{U}} y_i^2 \ - \ \frac{1}{N-1}\left(\sum_{i,j\in\boldsymbol{U}} y_iy_j - \sum_{i,j\in\boldsymbol{U}, i=j} y_iy_j\right)\right] \\
&= \frac{1}{n}\frac{N-n}{N^2}\left[\sum_{i\in\boldsymbol{U}} y_i^2 \ - \ \frac{1}{N-1}\left(\sum_{i\in\boldsymbol{U}}\sum_{j\in\boldsymbol{U}} y_iy_j - \sum_{i\in\boldsymbol{U}} y_i^2\right)\right] \\
&= \frac{1}{n}\frac{N-n}{N^2}\left[\sum_{i\in\boldsymbol{U}} y_i^2 \ - \ \frac{1}{N-1}\left(N^2\mu_y^2 - \sum_{i\in\boldsymbol{U}} y_i^2\right)\right] \\
&= \frac{1}{n}\frac{N-n}{N^2}\left[\frac{N}{N-1}\sum_{i\in\boldsymbol{U}} y_i^2 \ - \ \frac{N^2}{N-1}\mu_y^2\right] \\
&= \frac{1}{n}\frac{N-n}{N}\left[\frac{1}{N-1}\left(\sum_{i\in\boldsymbol{U}} y_i^2 - N\mu_y^2\right)\right] \\
&= \frac{1}{n}\left(1-\frac{n}{N}\right)\left[\frac{1}{N-1}\sum_{i\in\boldsymbol{U}} (y_i - \mu_y)^2\right] \\
&= \left(1-\frac{n}{N}\right)\frac{\sigma_y^2}{n} \ .
\end{aligned}$$

Grader:

$\star$**2 marks** for correctly separating the expression into variances and covariances
$\star$**1 mark** for correctly writing $\text{Cov}(I_i, I_j) = \mathbb{E}[I_i I_j] - \mathbb{E}[I_i]\mathbb{E}[I_j]$
$\star$**1 mark** for correctly rewriting the sum $\displaystyle\sum_{i,j \in \boldsymbol{U},\, i \neq j} y_i y_j = \sum_{i,j \in \boldsymbol{U}} y_i y_j - \sum_{i,j \in \boldsymbol{U}, i=j} y_i y_j$
†**2 marks** for showing simplification (intermediate steps)

†As long as the student shows some derivation, gives them at least **2 marks** even if they are incorrect
$\star$**Deduct no mark** if the student got these three things correct but made simplification errors or did not get to the correct final expression