T. D. n° 3
Sondage à probabilités inégales

Exercice 1. *Plan et probabilités d'inclusion.* **Extrait du livre « Exercices corrigés de méthodes de sondage » de P. Ardilly et de Y. Tillé**

Soient une population $U = \{1, 2, 3\}$ et le plan suivant :

$$\mathbb{P}(\{1,2\}) = \frac{1}{2}, \quad \mathbb{P}(\{1,3\}) = \frac{1}{4}, \quad \mathbb{P}(\{2,3\}) = \frac{1}{4}.$$

1. Donner les probabilités d'inclusion d'ordre un.
2. Donner la matrice de variance-covariance $\mathbf{\Delta}$ des variables indicatrices d'appartenance à l'échantillon.
3. Donner l'écriture matricielle de la variance de l'estimateur sans biais du total.

Exercice 2. *Variance des indicatrices et plan de sondage.* **Extrait du livre « Exercices corrigés de méthodes de sondage » de P. Ardilly et de Y. Tillé**

Soit la matrice de variance-covariance $\mathbf{\Delta}$ des indicatrices de la présence des unités d'observation dans l'échantillon pour un plan $p(s)$, donnée par :

$$\mathbf{\Delta} = \frac{6}{25} \begin{pmatrix} 1 & 1 & 1 & -1 & -1 \\ 1 & 1 & 1 & -1 & -1 \\ 1 & 1 & 1 & -1 & -1 \\ -1 & -1 & -1 & 1 & 1 \\ -1 & -1 & -1 & 1 & 1 \end{pmatrix}$$

1. Ce plan est-il de taille fixe ?
   Nous rappelons une propriété importante : dans un plan est de taille fixe, la somme de toutes les lignes et la somme de toutes les colonnes de la matrice des Δ_{kl} sont nulles.
2. Ce plan satisfait-il aux conditions de Sen-Yates-Grundy ?
3. Calculer les probabilités d'inclusion de ce plan sachant que :
$$\pi_1 = \pi_2 = \pi_3 > \pi_4 = \pi_5.$$
4. Donner la matrice des probabilités d'inclusion d'ordre deux.
5. Donner les probabilités associées à tous les échantillons possibles.

Exercice 3. *Estimation d'une racine.* **Extrait du livre « Exercices corrigés de méthodes de sondage » de P. Ardilly et de Y. Tillé**

Soit une population de 5 individus. On s'intéresse à un caractère d'intérêt y qui prend les valeurs suivantes :

$$y_1 = y_2 = 1 \quad \text{et} \quad y_3 = y_4 = y_5 = \frac{8}{3}.$$

On définit le plan de sondage suivant :

$$\mathbb{P}\left[\{1,2\}\right] = \frac{1}{2}, \quad \mathbb{P}\left[\{3,4\}\right] = \mathbb{P}\left[\{3,5\}\right] = \mathbb{P}\left[\{4,5\}\right] = \frac{1}{6}.$$

1. Calculer les probabilités d'inclusion aux ordres un et deux.
2. Donner la distribution de probabilités de l'estimateur du total noté $\widehat{T}_{pi}$ dans le cadre de ce plan de sondage.
3. Calculer l'estimateur de la variance de $\widehat{T}_{pi}$ avec une formule du cours. Cet estimateur de la variance est-il biaisé ? Était-ce prévisible ?
4. On se propose d'estimer la racine carrée du total (notée $\sqrt{T}$), par la racine carrée de l'estimateur $\sqrt{\widehat{T}_{pi}}$. Donner la distribution de probabilités de cet estimateur. Montrer qu'il sous-estime $\sqrt{T}$. Était-ce prévisible ?
5. Calculer la variance de $\sqrt{\widehat{T}_{pi}}$.

Exercice 4. *Variance et estimations concurrentes de variance.* **Extrait du livre « Exercices corrigés de méthodes de sondage » de P. Ardilly et de Y. Tillé**

Soient une population $U = \{1,2,3\}$ et le plan suivant :

$$\mathbb{P}\left(\{1,2\}\right) = \frac{1}{2}, \quad \mathbb{P}\left(\{1,3\}\right) = \frac{1}{4}, \quad \mathbb{P}\left(\{2,3\}\right) = \frac{1}{4}.$$

1. Donner la distribution de probabilité du π-estimateur de la moyenne.
2. Donner la distribution de probabilité du ratio de Hájek de la moyenne.
3. Donner les distributions de probabilité des deux estimateurs classiques de variance du π-estimateur au cas où $y_k = \pi_k, k \in U$.

Exercice 5. *Variance de Sen-Yates-Grundy.* **Extrait du livre « Exercices corrigés de méthodes de sondage » de P. Ardilly et de Y. Tillé**

Le but de cet exercice est de montrer que, lorsque la taille de l'échantillon est fixe, la précision d'un tirage sans remise avec probabilités inégales peut s'exprimer sous une forme « sympathique », dite de Sen-Yates-Grundy.

1. Si on note π_k la probabilité d'inclusion de l'individu k, N la taille de la population, et n la taille fixe de l'échantillon, montrer que :
$$\sum_{k \in U} \pi_k = n.$$
2. Si on note π_{kl} la probabilité d'inclusion double de k et de l, montrer que, pour tout $k \in U$,
$$\sum_{l \in U, l \neq k} \pi_{kl} = (n-1)\pi_k.$$
Indication : utiliser les variables indicatrices.

3. Montrer que, pour tout $k \in U$,

$$\sum_{l \in U, l \neq k} \pi_k \pi_l = \pi_k (n - \pi_k).$$

En déduire que, pour tout $k \in U$,

$$\sum_{l \in U, l \neq k} (\pi_k \pi_l - \pi_{kl}) = \pi_k (1 - \pi_k).$$

4. Mettre la précision de l'estimateur de Horvitz-Thompson $\widehat{T}_Y$ (estimant sans biais le total Y) sous la forme :

$$\text{Var}\left(\widehat{T}_Y\right) = \left[\sum_{k \in U} \left(\frac{y_k}{\pi_k}\right)^2 \sum_{l \in U, l \neq k} (\pi_k \pi_l - \pi_{kl}) - \sum_{k \in U} \sum_{l \in U, l \neq k} \frac{y_k}{\pi_k} \frac{y_l}{\pi_l} (\pi_k \pi_l - \pi_{kl})\right].$$

En déduire :

$$\text{Var}\left(\widehat{T}_Y\right) = \frac{1}{2} \sum_{k \in U} \sum_{l \in U, l \neq k} (\pi_k \pi_l - \pi_{kl}) \left(\frac{y_k}{\pi_k} - \frac{y_l}{\pi_l}\right)^2.$$

Quel est l'intérêt de cette formulation ?

Exercice 6. *Effet de sondage.* **Extrait du livre « Exercices corrigés de méthodes de sondage » de P. Ardilly et de Y. Tillé**

Lorsqu'on met en øeuvre des plans de sondage complexes et que l'on cherche à calculer des précisions en utilisant un logiciel, on obtient en général le calcul d'un rapport appelé « design effect » ou « effet de sondage ». Ce rapport est défini comme le rapport de la variance de l'estimateur du total $\widehat{Y}$ sur la variance de l'estimateur que l'on obtiendrait si on effectuait un sondage aléatoire simple de même taille n. On note $\widehat{\overline{Y}}$ la moyenne simple des y_k pour k dans S.

1. En notant $\text{Var}_p\left(\widehat{Y}\right)$ la variance vraie (éventuellement très compliquée) obtenue sous le plan complexe (noté p), donner l'expression du design-effet (noté désormais DEFF).
2. Comment va-t-on naturellement estimer DEFF (on note $\widehat{\text{DEFF}}$ l'estimateur) ? *On se restreint désormais à des plans complexes p à probabilités égales et de taille fixe.*
3. Dans ces conditions, comment estime-t-on sans biais n'importe quel « vrai » total Y ?
4. Calculer l'espérance de la dispersion s_y^2 dans l'échantillon, sous le plan p (on la note $\mathbb{E}\left(s_y^2\right)$). On l'exprimera en fonction de $\text{Var}_p\left(\widehat{\overline{Y}}\right)$, S_y^2, n et N.
5. Considérant le dénominateur de $\widehat{\text{DEFF}}$, montrer que son utilisation introduit un biais que l'on exprime en fonction de n, N et $\text{Var}_p\left(\widehat{\overline{Y}}\right)$. Pour cette question, on considère que n est « grand ».

6. En déduire que le dénominateur de $\widehat{\text{DEFF}}$ a une espérance égale à la valeur souhaitée multipliée par le facteur :
$$1 - \frac{1-f}{n}\text{DEFF}.$$
Conclure dans le cas où n est « grand ».

Exercice 7. *Ratio de Hájek.* **Extrait du livre « Exercices corrigés de méthodes de sondage » de P. Ardilly et de Y. Tillé**

L'objet de cet exercice est de déterminer certaines conditions dans lesquelles le ratio de Hájek est moins efficace que l'estimateur classique de Horvitz-Thompson. On considère que la taille de l'échantillon est grande et que l'échantillon est de taille fixe.

1. Rappeler, pour l'estimation d'un total Y, les expressions de variance des deux estimateurs en question.
2. On peut toujours écrire, pour tout $k \in U$,
$$y_k = \alpha + \beta x_k + u_k \quad \alpha, \beta \in \mathbb{R},$$
où α et β sont les « vrais » coefficients de régression mais inconnus de y sur x, $\pi_k = nx_k/X$, x_k est une variable de taille, et le tirage est un tirage proportionnel à la taille. Par ailleurs, on suppose que u_k est « petit », c'est-à-dire que x « explique bien » y. Dans ces conditions, que deviennent les expressions de variance des deux estimateurs ?
3. Que vaut approximativement le rapport des deux variances ?
4. En conclusion, dans les conditions d'une forte corrélation linéaire entre x et y (c'est-à-dire u_k petit), quand peut-on considérer « qualitativement » que l'estimateur de Horvitz-Thompson est préférable à celui du ratio de Hájek ?