T. D. n° 5
Sondage à probabilités inégales

Exercice 1. Estimation d'une racine. **D'après le livre « Exercices corrigés de méthodes de sondage » de P. Ardilly et de Y. Tillé**
Soit une population de 5 individus. Nous nous intéressons à un caractère d'intérêt y qui prend les valeurs suivantes :

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

Nous définissons 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. Nous nous proposons 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 2. Variance et estimations de variance. **D'après le 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 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 3. Effet de sondage. **Extrait du livre « Exercices corrigés de méthodes de sondage » de P. Ardilly et de Y. Tillé**
Lorsque nous mettons en œuvre des plans de sondage complexes et que nous cherchons à calculer des précisions en utilisant un logiciel, nous obtenons 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 nous obtiendrions si nous effectuions un sondage aléatoire simple de même taille n. Nous notons $\widehat{\overline{Y}}$ la moyenne simple des y_k pour k dans S.

1. En notant $\mathrm{Var}_p\left[\widehat{Y}\right]$ la variance vraie (éventuellement très compliquée) obtenue sous le plan complexe (noté p), donner l'expression du design-effet (noté désormais DEFF).
2. Comment allons-nous naturellement estimer DEFF (on note $\widehat{\mathrm{DEFF}}$ l'estimateur) ?
   Nous nous restreignons 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 $\mathbb{E}\left[s_y^2\right]$). Nous l'exprimerons en fonction de $\mathrm{Var}_p\left[\widehat{\overline{Y}}\right]$, S_y^2, n et N.
5. Considérant le dénominateur de $\widehat{\mathrm{DEFF}}$, montrer que son utilisation introduit un biais que nous exprimons en fonction de n, N et $\mathrm{Var}_p\left[\widehat{\overline{Y}}\right]$. Pour cette question, nous considérons que n est « grand ».
6. En déduire que le dénominateur de $\widehat{\mathrm{DEFF}}$ a une espérance égale à la valeur souhaitée multipliée par le facteur :
$$1 - \frac{1-f}{n}\mathrm{DEFF}.$$
Conclure dans le cas où n est « grand ».

Exercice 4. 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. Nous considérons 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. Nous pouvons 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, nous supposons 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 ?