TD6 : Estimation dans un problème de sondage

Exercice 1. (Problème de sondages) Soit N le nombre d'habitants d'une commune. Il s'agit de faire un sondage de popularité de deux candidats (candidat A et candidat B) qui se présentent aux élections municipales. On choisit un échantillon de n habitants auxquels on pose la question : "Pour qui voteriez–vous aux élections ?" A l'issue de ce sondage, on obtient les données $X_1, \ldots, X_n$, où

$$X_i = \begin{cases} 1, & \text{si le } i^{\text{ème}} \text{ habitant questionné préfère le candidat } A, \\ 0, & \text{si le } i^{\text{ème}} \text{ habitant questionné préfère le candidat } B, \end{cases} \qquad i = 1, \ldots, n.$$

Pour les raisons évidentes, il est impossible de questionner tous les habitants. Donc $n < N$ (dans la pratique, on a toujours $n \ll N$). Notons θ la part d'habitants de la commune qui préfèrent le candidat A. Le but du sondage est d'estimer θ et de donner un intervalle confiance pour θ.

Définissons les valeurs déterministes $x_1, \ldots, x_N$ par

$$x_j = \begin{cases} 1, & \text{si le } j^{\text{ème}} \text{ habitant préfère le candidat } A, \\ 0, & \text{si le } j^{\text{ème}} \text{ habitant préfère le candidat } B, \end{cases} \qquad j = 1, \ldots, N.$$

On a alors

$$\theta = \frac{1}{N} \sum_{j=1}^{N} x_j.$$

Définissons aussi

$$\sigma^2 = \frac{1}{N} \sum_{j=1}^{N} (x_j - \theta)^2.$$

On appelle θ *moyenne de population* et σ^2 *variance de population*.

1. Proposer un modèle statistique pour ce problème. Pour cela,
   (a) Calculer $\mathbf{P}_\theta\big((X_1, X_2, X_3, X_4) = (0, 1, 1, 0,)\big)$ en utilisant le fait qu'il s'agit d'un tirage au hasard sans remise d'une population de taille N, car chaque habitant peut apparaître au maximum une fois dans l'échantillon.
   (b) Soit $(a_1, \ldots, a_n) \in \{0, 1\}^n$ tel que $a_1 + \ldots + a_n = a$. Deviner la forme de
   $$\mathbf{P}_\theta\big((X_1, \ldots, X_n) = (a_1, \ldots, a_n)\big)$$
   et prouver la formule obtenue par récurrence sur n.
2. Montrer que $\bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$ est un estimateur sans biais de θ.
3. Montrer que

$$\mathbf{Cov}(X_i, X_j) = -\frac{\sigma^2}{N-1} \quad \text{pour} \quad i \neq j \qquad \text{et} \qquad \mathbf{Var}(\bar{X}_n) = \frac{\sigma^2}{n} \left(1 - \frac{n-1}{N-1}\right).$$

4. Calculer $\mathbf{E}[s_n^2]$, où $s_n^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X}_n)^2$, et proposer un estimateur sans biais $\widehat{v}_n^2$ de la variance $\mathbf{Var}(\bar{X}_n)$.

5. On se place maintenant dans le cadre asymptotique où $N \to \infty$, $n = n(N) \to \infty$ et $n/N \to 0$.
   (a) Montrer que $\bar{X}_n$ et $\widehat{v}_n^2$ sont des estimateurs consistants de θ et σ^2.
   (b) On admet que $\sqrt{n}(\bar{X} - \theta)$ converge en loi vers $\mathcal{N}(0, \sigma^2)$. Démontrer la normalité asymptotique
   $$\sqrt{n}\,\frac{\bar{X}_n - \theta}{\widehat{v}_n} \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0,1).$$
   (c) En utilisant le rśultat ci-dessus, trouver un l'intervalle aléatoire $[A, B]$ tel que
   $$\lim_{n\to\infty} \mathbf{P}\big([A,B] \text{ contient } \theta\big) = 95\%.$$

6. *Application numérique :* donner l'intervalle de confiance de niveau asymptotique 95% pour θ lorsque $N = 8000$, $n = 100$, $n_1 = \sum_{i=1}^n \mathbb{1}\{X_i = 1\} = 65$.

Exercice 2. (facultatif) Soient f et g deux densités de probabilité d́finies sur $[0,1]$. On suppose que
$$f(x) \le M\, g(x), \qquad \forall x \in [0,1]. \tag{1}$$
Soient $\{(X_i, U_i) : i \in \mathbb{N}^*\}$ une suite de variables aléatoires iid telles que

- la loi de X_1 a pour densité g,
- U_1 est de loi uniforme sur $[0,1]$,
- X_1 et U_1 sont indépendantes.

On d́finit les variables aléatoires
$$N = \min\big\{n : f(X_n) \ge M\, U_n\, g(X_n)\big\}, \qquad Y = X_N.$$

1. Soit $q = \mathbf{P}(f(X_n) \ge M\, U_n\, g(X_n))$. Montrer que $q = 1/M$.
2. Montrer que N suit la loi géométrique de paramètre q, c'est-à-dire
   $$\mathbf{P}(N = k) = (1-q)^{k-1} q, \qquad k \in \mathbb{N}^*.$$
3. Pour toute fontion mesurable bornée $h : [0,1] \to \mathbb{R}$, montrer que
   $$\mathbf{E}[h(Y)] = \int_0^1 h(y) f(y)\, dy. \tag{2}$$
   En déduire que Y a pour densité f.

Remarque Cet exercice montre que si (1) est vrai et si on est capable de générer des variables aléatoires de densité g, alors on sera également capable de générer des variables aléatoires de desnité f. Cette méthode porte le nom de méthode d'acceptation-rejet. Elle a toutefois un inconvénient : le nombre moyen de tirage de X_i nécessaire pour calculer Y est $\mathbf{E}[N] = M$. Ce nombre peut être excessivement grand lorsque M est trop grand.