Sondage

Exercice 1. On veut contrôler la qualité d'un lot de N objets. Pour cela, n objets ($1 \leq n \leq N$) sont tirés au hasard successivement et **avec remise** dans le lot, et examinés. On observe donc le vecteur $(X_1, \dots, X_n)$ où

$$X_i = \begin{cases} 1 & \text{si le } i\text{ème objet est défectueux} \\ 0 & \text{sinon} \end{cases}$$

et les variables aléatoires $X_1, \dots, X_n$ sont i.i.d.. On note k le nombre (inconnu) d'objets défectueux dans le lot et $p = k/N$ la proportion d'objets défectueux dans le lot.

(a) Chercher la loi de X_i . Calculer $E(X_i)$ et $Var(X_i)$.

(b) Si $i \neq j$, calculer $Cov(X_i, X_j)$.

(c) On pose $S = \sum_{i=1}^n X_i$. Que représente S ? Calculer $E(S)$ et $Var(S)$.

(d) Construire à l'aide de S un estimateur $\hat{p}$ de p . Déterminer son risque quadratique $E[(\hat{p} - p)^2]$.

(e) Montrer que pour tout $\varepsilon > 0$

$$P(|\hat{p} - p| \geq \varepsilon) \leq \frac{1}{4n\varepsilon^2}.$$

En déduire que

$$P\left(p \in \left[\hat{p} - \frac{5}{\sqrt{n}}, \hat{p} + \frac{5}{\sqrt{n}}\right]\right) \geq 0,99$$

Exercice 2. On veut contrôler la qualité d'un lot de N objets. Pour cela, n objets ($1 \leq n \leq N$) sont tirés au hasard successivement et **sans remise** dans le lot, et examinés. On observe donc le vecteur $(X_1, \dots, X_n)$ où

$$X_i = \begin{cases} 1 & \text{si le } i\text{ème objet est défectueux} \\ 0 & \text{sinon} \end{cases}.$$

On note k le nombre (inconnu) d'objets défectueux dans le lot. On a donc pour $2 \leq i \leq n$:

$$P(X_i = 1 \mid X_1, \dots, X_{i-1}) = \frac{k - \sum_{j=1}^{i-1} X_j}{N - i + 1}.$$

On admet que la suite $(X_1, \dots, X_n)$ est échangeable : pour toute permutation $\sigma : \{1, \dots, N\} \rightarrow \{1, \dots, N\}$ le vecteur $(X_{\sigma(1)}, \dots, X_{\sigma(N)})$ a la même loi que $(X_1, \dots, X_N)$.

(a) Chercher la loi de X_i . Calculer $E(X_i)$ et $Var(X_i)$.

(b) Si $i \neq j$, calculer $Cov(X_i, X_j)$.

(c) On pose $S = \sum_{i=1}^n X_i$. Calculer $E(S)$ et $Var(S)$.

(d) Construire à l'aide de S un estimateur de $p = k/N$. Déterminer son risque quadratique.

Exercice 3. On cherche à évaluer la proportion $p \in [0, 1]$ de personnes téléchargeant illégalement sur internet dans une population donnée. On choisit au hasard (avec remise) n individus dans cette population. La procédure de sondage est la suivante. On demande à chaque individu interrogé de tirer au préalable (sans être vu) une boule dans une urne, et de répondre ensuite par "OUI" ou "NON" à l'une des deux questions : Question 1 : "Est-ce que vous avez déjà téléchargé illégalement ?" si la boule tirée est blanche, Question 2 : "Est-ce que vous n'avez jamais téléchargé illégalement?" si la boule tirée n'est pas blanche. (bien sûr, seul l'individu interrogé a connaissance de la couleur de la boule, afin que personne d'autre ne sache à quelle question il a répondu). La proportion φ de boules blanches dans l'urne est connue et différente de $1/2$. On note $X_1, \dots, X_n$ les variables aléatoires qui correspondent aux réponses des n personnes interrogées (ce sont donc des variables aléatoires à valeurs dans $\{OUI, NON\}$). Soit R le nombre de réponses "OUI" obtenues.

- Quelle est la loi de X_1 ?
- Quelle la loi de R ? Calculer $E(R)$.
- Calculer $Var(R)$ et montrer que $Var(R) \leq n/4$.
- En déduire un estimateur $\hat{p}$ de p tel que $E(\hat{p}) = p$ et $E[(\hat{p} - p)^2] \leq 1/(n4(2\varphi - 1)^2)$.
- En déduire une constante $c_\varphi \in \mathbb{R}_+$ telle que

$$P\left(p \in \left[\hat{p} - \frac{c_\varphi}{\sqrt{n}}, \hat{p} + \frac{c_\varphi}{\sqrt{n}}\right]\right) \geq 0,99$$

Exercice 4. Un gramme de carbone extrait d'un organisme mort contient un nombre θ d'atomes de carbone C14. Ces atomes de C14 ne sont pas stables et on modélise leur durée de vie par des variables aléatoires $X_1, X_2, \dots, X_\theta$ i.i.d. de loi commune la loi exponentielle de paramètre $\lambda = 3,92 \cdot 10^{-12} s^{-1}$. On peut montrer que θ et le temps T écoulé depuis la mort de l'organisme sont liés par la formule

$$T = \frac{\log(N_0) - \log(\theta)}{\lambda} \quad \text{où } N_0 = 6,8 \cdot 10^{10}.$$

Dans la situation où θ est inconnu et que l'on veut l'estimer, on compte le nombre de désintégrations sur l'intervalle de temps $[0, t_0]$ où $t_0 > 0$ est une constante choisie, autrement dit on observe la variable aléatoire

$$S = \sum_{i=1}^{\theta} 1_{\{X_i \leq t_0\}}.$$

- Calculer l'espérance et la variance de S en fonction de θ et de t_0 .
- On pose $\hat{\theta} = S/(1 - e^{-\lambda t_0})$. Démontrer que pour tout $\varepsilon > 0$, pour tout $\theta \in \mathbb{N}$ on a

$$P_\theta\left(|\hat{\theta} - \theta| > \varepsilon\sqrt{\theta}\right) \leq \frac{1}{e^{\lambda t_0} - 1} \frac{1}{\varepsilon^2}.$$

- Pour $m \in [0, 1]$ et $\varepsilon > 0$, résoudre l'inéquation (en l'inconnue x)

$$(x - m)^2 \leq \varepsilon^2 x$$

- En déduire que pour tout $\varepsilon > 0$

$$P\left(\hat{\theta} + \frac{\varepsilon^2}{2} - \frac{\varepsilon}{2}\sqrt{\varepsilon^2 + 4\hat{\theta}} \leq \theta \leq \hat{\theta} + \frac{\varepsilon^2}{2} + \frac{\varepsilon}{2}\sqrt{\varepsilon^2 + 4\hat{\theta}}\right) \geq 1 - \frac{1}{e^{\lambda t_0} - 1} \frac{1}{\varepsilon^2}$$

- Déterminer deux fonctions T_1 et T_2 telles que

$$P\left(T_1(S) \leq T \leq T_2(S)\right) \geq 0,99$$

Application numérique : Pour $t_0 = 3$ jours on a obtenu la valeur $s = 1000$ pour S . Déterminer l'intervalle $[T_1(s), T_2(s)]$.