64pg ⊳ **Exercice** - *Sondage et confidentialité*

Pour préserver la confidentialité des opinions individuelles, un sondage est effectué avec le protocole suivant. Chaque personne sondée doit, avant de répondre «oui» ou «non» à la question posée, réaliser confidentiellement (elle seule connait le résultat) une variable de Bernoulli de paramètre a, $a \in]0;1[$ donné et connu du sondeur. Si le résultat est 1, la personne doit répondre à la question selon son avis. Si le résultat est 0, la personne doit répondre à la question selon le contraire de son avis.
On suppose que les personnes sondées suivent parfaitement ce protocole et l'on note X_i la variable aléatoire qui vaut 1 si la i-ème personne sondée répond «oui» et 0 si elle répond «non».
On note n le nombre de personnes sondées et p la proportion de personnes dont l'opinion personnelle est «oui» dans la population sondée. Soit q la probabilité qu'une personne sondée réponde «oui». Enfin, on suppose que $a \neq \frac{1}{2}$.

1. Exprimer q en fonction de p et a.

   D'après l'énoncé de la situation et le théorème des probabilités totales, $q = ap + (1-a)(1-p)$.

2. Vérifier que $\overline{X} = \frac{1}{n}\sum_{i=1}^n X_i$ est un estimateur de q sans biais et convergent.

   Chaque variable X_i suit une loi $\mathcal{B}(q)$ donc $\mathbb{E}(\overline{X}) = q$: $\overline{X}$ est un estimateur sais biais de q. Et $V(\overline{X}) = \frac{q(1-q)}{n} \to 0$ donc $\overline{X}$ est un estimateur convergent.

3. On pose
$$T_n = \frac{\overline{X} - 1 + a}{2a-1}$$
   Calculer l'espérance et la variance de T_n.

   Par linéarité et d'après le calcul précédent, on trouve $\mathbb{E}(T_n) = \frac{q-1+a}{2a-1} = p$. D'après les règles de calcul pour la variance,
$$V(T_n) = \frac{1}{(a-1)^2}V(\overline{X} - 1 + a) = \frac{1}{(2a-1)^2}V(\overline{X}) = \frac{q(q-1)}{n(2a-1)^2}$$
   Or en développant le calcul, on trouve $q(q-1) = p(1-p)(2a-1)^2 + a(1-a)$ donc $V(T_n) = \frac{p(1-p)}{n} + \frac{a(1-a)}{n(2a-1)^2}$

4. Démontrer que T_n est un estimateur de p sans biais et convergent.

   Cela permet de voir que $B(T_n) = p - p = 0$ et $V(T_n) \to 0$ quand $n \to +\infty$.

5. Justifier que la variable $\frac{T_n - p}{\sigma(T_n)}$ converge en loi vers la loi normale centrée réduite.

   Pour pouvoir appliquer le théorème central limite, il suffit de vérifier que T_n s'écrit bien sous la forme d'une somme de variables aléatoires indépendantes et identiquement distribuées. En effet, on a
$$T_n = \frac{\frac{1}{n}\left(\sum_{i=1}^n X_i\right) - 1 + a}{2a-1} = \frac{\frac{1}{n}\sum_{i=1}^n (X_i - 1 + a)}{2a-1} = \sum_{i=1}^n \frac{X_i - 1 + a}{n(2a-1)}$$
   Les variables X_i étant i.i.d, les hypothèses sont bien vérifiées et d'après le théorème central limite, la variable $\frac{T_n - \mathbb{E}(T_n)}{\sigma(T_n)}$ converge bien en loi vers une variable suivant une loi normale centrée réduite.

6. Donner une estimation de p par intervalle de confiance au niveau de confiance $1-\alpha$ que l'on notera I_α.

   D'après la question précédente, on peut (en supposant n grand) approcher la variable $\frac{T_n - \mathbb{E}(T_n)}{\sigma(T_n)}$ par une variable Z qui suit une loi $\mathcal{N}(0,1)$. Ensuite, on construit facilement un intervalle de confiance pour $p = \mathbb{E}(T_n)$ au niveau $1-\alpha$ en choisissant dans la table la valeur $u_{\alpha/2}$ permettant d'avoir
$$\mathrm{P}(-u_{\alpha/2} < Z < u_{\alpha/2}) = 1 - \alpha$$
   puis en redéployant l'inégalité autour de p :
$$\mathrm{P}(T_n - u_{\alpha/2}\sigma(T_n) < \mathbb{E}(T_n) < T_n + u_{\alpha/2}\sigma(T_n)) = 1 - \alpha$$
   D'où l'intervalle de confiance pour $p = \mathbb{E}(T_n)$:
$$I_conf = [T_n - u_{\alpha/2}\sigma(T_n) \,;\, T_n + u_{\alpha/2}\sigma(T_n)]$$
   Pour réaliser cet intervalle, il suffit de réaliser l'échantillon et de remplacer T_n et $\sigma(T_n)$ par leurs réalisations. On obtiendra un intervalle par excès en replaçant $\sigma(T_n)$ par un majorant $\frac{1/2}{\sqrt{n(2a-1)^2}}$

7. On fixe $a = \frac{1}{6}$, $n = 1000$ et on considère une réalisation de la variable aléatoire $\overline{X}$ égale à 0,425. Déterminer une réalisation de l'intervalle de confiance utilisé pour l'estimation de p au risque $\alpha = 0{,}05$ (on pourra utiliser la majoration $q(1-q) \leqslant \frac{1}{4}$).

   Application numérique : on trouve $\alpha = 0.05$, on utilise $u_{\alpha/2} = 1.96$ et on trouve
$$I = [0.566 \,;\, 0.658]$$