Mini-Projet

Pour la reproductibilité des questions numériques, il est conseillé de fixer la « graine » du générateur de nombres pseudo-aléatoires, en haut de votre script, en utilisant la fonction `set.seed` de **R**, par exemple :

```
set.seed(42,kind="Marsaglia-Multicarry")
```

On rappelle le résultat suivant

Théorème 0.1
(loi des grands nombres)
Soit $Z : \Omega \to \mathbb{R}$ *une variable aléatoire sur un espace probabilisé* $(\Omega, \mathcal{F}, \mathbb{P})$ *telle que* $\mathbb{E}(|Z|) < +\infty$*, et soit* $(Z_i)_{i\geq 0}$ *est un échantillon i.i.d. de même loi que* Z*, défini sur le même espace. Il existe* $N \subset \Omega$ *tel que* $\mathbb{P}(N) = 0$ *et*

$$\forall \omega \in \Omega \setminus N, \qquad \frac{1}{n}\sum_{i=1}^{n} Z_i(\omega) \xrightarrow[n\to\infty]{} \mathbb{E}(Z).$$

Autrement dit, la moyenne empirique des Z_i *converge* $\mathbb{P}$*-presque sûrement vers* $\mathbb{E}(Z)$.

Exercice 1 (Dates de défaillance d'un système)**:**
On s'intéresse à la durée de vie X_1 d' une particule radioactive. Il est courant de discrétiser le temps et de considérer que la durée de vie (en secondes, millisecondes, ou autre selon les cas) est une variable aléatoire à valeurs dans $\mathbb{N}^* = \{1, 2, 3, \dots, \}$. Un modèle classique consiste à supposer que, sachant que la particule est encore là, la probabilité qu'elle se désintègre sur l'intervalle de temps suivant vaut $\theta \in]0, 1[$, et qu'il y a 'oubli du passé', c'est à dire que θ reste constant au cours du temps. Ainsi, en notant P_θ la loi de X_1, on a

$$\mathrm{P}_\theta\{k\} = \theta(1-\theta)^{k-1}, \qquad k \in \mathbb{N}^*$$

On observe $X = (X_1, \dots, X_n)$, les durées de vie de n particules ($n \geq 2$), supposées indépendantes et identiquement distribuées selon P_θ, Où $\theta \in]0, 1[$ est le paramètre (inconnu) du modèle. On cherche à estimer la grandeur d'intérêt $g(\theta) = \mathbb{E}_\theta(X_1)$.

1. (2 pt) Calculer $g(\theta)$ pour $0 < \theta < 1$.
2. (1 pt) Par quelle mesure le modèle $\mathcal{P} = \{\mathrm{P}_\theta, \theta \in]0, 1[\}$ est-il dominé ?
3. (2 pt) On admet que le modèle $\{\mathrm{P}_\theta, \theta \in]0, 1[\}$ est régulier, au sens des hypothèses du théorème de Cramér-Rao. On note $T(X) = \frac{1}{n}\sum_{i=1}^{n} X_i$. Monter que la statistique $T(X)$ est un estimateur UVMB (uniformément de variance minimale parmi les estimateurs sans biais) de $g(\theta)$.
4. (2 pt) Soit $h > 0$. On considère le nouvel estimateur

$$S_h(X) = hT(X).$$

Montrer que à θ fixé, pour certaines valeurs de h (que vous préciserez en fonction de θ), et pour le risque quadratique, on a

$$R(\theta, S_h) < R(\theta, T).$$

Donner la valeur optimale $h^*(\theta)$ qui minimise le risque.

5. (2 pt) Existe-t-il h^*, pour $n \in \mathbb{N}$ fixé, qui minimise le risque quadratique $R(\theta, S_h)$ uniformément en θ, c'est-à-dire tel que

$$\forall \theta > 0, \forall h > 0, R(\theta, S_{h^*}) \leq R(\theta, S_h)?$$

6. Illustration numérique (6 pt) : l'objectif est de mettre en évidence numériquement certains aspects des résultats théoriques établis plus haut. On fixe $n = 10$ (taille de l'échantillon). On va simuler indépendemment $M = 100\,000$ échantillons de taille 10 chacun, $(X^j)_{j=1,\dots,M}$, avec $X^j = (X_1^j, \dots, X_{10}^j)$. On va donc pouvoir répéter $M = 10^5$ fois l'expérience consistant à calculer un estimateur de $g(\theta)$ à partir d'un échantillon de taille 10. Le 'vrai' θ pour cette expérience est fixé à $\theta = \theta_0 = 0.2$. Les X_i^j sont donc supposés indépendants. Pour chaque échantillon $X^j (j \leq M)$, on considère les erreurs quadratiques

$$L_1^j = \Big(T(X^j) - g(\theta)\Big)^2 \text{ et } L_h^j = \Big(S_h(X^j) - g(\theta)\Big)^2 ; \qquad (j = 1, \dots, M)$$

En utilisant la loi des grands nombres, nous approcherons les risques quadratiques par les moyennes empiriques

$$\widehat{R}(\theta_0, T) = \frac{1}{M}\sum_{j=1}^{M} L_1^j \quad \text{et} \quad \widehat{R}(\theta_0, S_h) = \frac{1}{M}\sum_{j=1}^{M} L_h^j.$$

Si vous pensez avoir correctement répondu à la question précédente, fixez $h = h^*(\theta_0)$ dans la suite de cet exercice. Sinon, prendre $h = \frac{9}{11}$.

(a) Créer une matrice Z de dimensions 10×10^5, chaque colonne contenant l'échantillon X^j. On pourra utiliser la fonction `matrix` de `R` qui permet de construire une matrice à partir d'un vecteur `v` (en la remplissant colonne par colonne) :

```
Z = matrix(v, nrow = 10)
```

On consultera également l'aide du simulateur de variables géométriques. Attention, dans la convention de `R`, la variable simulée Y est le nombre d'échecs avant le premier succès dans une épreuve de Bernoulli. Autrement dit, le temps de survie dans notre cas est égal en loi à $Y + 1$.

```
?rgeom
```

(b) Construire deux vecteurs (nommés respectivement `Tx` et `Sx`) de taille M, dont les j^{emes} éléments contiennent respectivement les valeurs $T(X^j)$ et $S_h(X^j)$.
Indication : la commande `Z[,j]` renvoie la j^{eme} colonne de Z. On utilisera la fonction `mean`. On pourra éventuellement avoir recours à la fonction `apply` plutôt que d'écrire une boucle `for` pour accélérer l'exécution du code.

(c) Inspectez vos résultats à ce stade : tracez sur le même graphique les histogrammes des valeurs $T(X^j)$ et $S_h(X^j)$, puis affichez la valeur moyenne des estimateurs obtenus, comme ceci (vous devez préalablement définir la variable `theta_0`) :

```
hist(Sx, col="red",probability=TRUE,
     main="Histogrammes de Tx et Sx",
```

```
     breaks=50,
     xlab= "Estimations Sx et Tx de g(theta)",
     ylab ="Probabilite")

hist(Tx,col="blue", probability=TRUE, add=TRUE, density=15, breaks=50)

abline(v = 1/theta_0, lwd=3)

legend("topright", lwd=2, col=c("red", "blue", "black"),
      legend=c( "Sx", "Tx", "g(theta)"))

mean(Sx)
mean(Tx)
```

(d) Construire les vecteur $L_1 = (L_1^j)_{j=1,\dots,M}$ et $L_h = (L_h^j)_{j=1,\dots,M}$.

(e) Calculer numériquement et afficher $R(\theta_0, T)$, $R(\theta, S_h)$ d'une part, et leur approximation $\widehat{R}(\theta_0, T)$ et $\widehat{R}(\theta_0, S_h)$. Commentez ce résultat au vu des résultats de la question 3.

(f) Tracer sur le même graphique :
- Les deux histogrammes des erreurs quadratiques (construits à partir de L_1 et L_h), en utilisant un code couleur permettant de distinguer les deux histogrammes.
- les approximations $\widehat{R}(\theta_0, T)$ et $\widehat{R}(\theta_0, S_h)$ (à représenter par des lignes verticales pleines de deux couleurs différentes)

Exercice 2 (Loi a posteriori du paramètre d'une loi de Géométrique)**:**
On s'intéresse toujours à la durée de vie d'une particule, on adopte la même modélisation qu'à l'exercice précédent. Dans cet exercice, on s'intéresse au problème de l'estimation de θ par une méthode Bayésienne. Pour des raisons pratiques (voir question 1), on choisit comme prior π une loi Bêta, $\pi = \mathcal{B}eta(\alpha, \beta)$ (avec $\alpha, \beta > 0$ fixés par l'utilisateur), c'est-à-dire : π admet comme densité par rapport à la mesure de Lebesgue $\mathrm{d}\theta$ sur $]0,1[$, la fonction (également notée π)

$$\pi(\theta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}$$

On donne l'espérance et la variance d'une telle loi : lorsque $Z \sim \mathcal{B}eta(\alpha, \beta)$, on a

$$\mathbb{E}(Z) = \frac{\alpha}{\alpha+\beta} \qquad ; \qquad \mathbb{V}\mathrm{ar}(Z) = \frac{\alpha\beta}{(\alpha+\beta)^2 - \alpha+\beta+1)}$$

Dans la suite on note θ_0 le "vrai" paramètre, c'est-à-dire le nombre $\theta_0 > 0$ tel que $X_i \sim \mathrm{P}_{\theta_0}$, $i = 1, \dots, n$.

1. (2 pt) Calculer la loi a posteriori $\pi(\theta|x)$, pour $x = (x_1, \dots, x_n) \in \mathbb{N}^n$ et $\theta > 0$.
   Indication : le résultat est à nouveau une loi Beta, dont il faut préciser les paramètres.
2. (1 pt) Calculer l'espérance à posteriori, $\mathbb{E}_\pi(\boldsymbol{\theta}|x)$. Pourquoi peut-on considérer $M(X) = \mathbb{E}_\pi(\boldsymbol{\theta}|X)$ comme un estimateur de θ_0 ?

3. (2pt) On considère une suite infinie d'observations indépendantes $(X_i)_{i\in\mathbb{N}}$, telles que $X_i \sim \mathrm{P}_{\theta_0}$. On note $X^n = (X_1, \dots, X_n)$. En utilisant la loi des grands nombres, montrer que la suite de variables aléatoires $M_n = \mathbb{E}(\boldsymbol{\theta}|X^n)$ converge $\mathbb{P}_{\theta_0}$-presque sûrement vers θ_0.
4. Illustration numérique (6 pt) : Le physicien "croit" que le paramètre θ_0 vaut environ $1/3$, avec une fourchette d'incertitude de l'ordre de $+/-1/4$. Il demande à son équipe de data scientists de mener une étude pour estimer le paramètre θ. L'équipe se donne donc pour prior π sur le paramètre θ une loi $\mathcal{B}eta(1/2, 3/2)$, de sorte que $\mathbb{E}_\pi(\boldsymbol{\theta}) = \alpha/(\alpha+\beta) = 1/3$ et $\mathbb{V}\mathrm{ar}_\pi(\boldsymbol{\theta}) = \frac{\alpha\beta}{(\alpha+\beta)^2 - \alpha+\beta+1)} = 1/16$. En réalité (personne ne le sait encore), le vrai paramètre est $\theta_0 = 0.6$.

   Les données à disposition de l'équipe sont un échantillon de $N = 500$ observations indépendantes, $X = (X_1, \dots, X_N)$, où $X_i \sim \mathrm{P}_{\theta_0}$. On note $X^n = (X_1, \dots, X_n)$ $(n \le N)$ les n premières observations.

   (a) Générérer un échantillon X de taille $N = 500$, distribué selon la "vraie" loi P_{θ_0}.

   Définir une grille de pas $h = 0.01$ entre 0 et 1 (extrémités exclues) : comme ceci

```
grille = seq(0, 1, by = 0.01)
L = length(grille)
grille = grille[-c(1,L)]
```

   On va inspecter l'évolution de la densité à posteriori $\pi(\theta|X^n)$, évaluée sur cette grille, lorsque n augmente.

   (b) Calculer la densité du prior et des lois a posteriori $\pi(\theta|X^n)$, pour $n = 5, 20, 100, 500$, évaluées sur la grille. On utilisera la fonction `dbeta`.
   Tracer sur le même graphique la densité de π et celles de $\pi(\,\cdot\,|X^n)$, avec des codes couleurs légendés permettant de distinguer les différentes courbes. Ajouter une ligne verticale représentant la valeur du vrai paramètre. Commentez vos résultats.

   (c) Calculer, pour $n = 1, \dots, 500$, l'espérance a posteriori $\mathbb{E}_\pi(\boldsymbol{\theta}|X^n)$. Tracer l'évolution de cette quantité en fonction de n (on pourra utiliser la fonction `cumsum`) Rajouter une ligne horizontale représentant la valeur du vrai θ_0. Commentez votre graphique au regard du résultat de la question 4.

Exercice 3 (Evaluation d'une politique commerciale)**:**
Le montant X_1 des achat d'un consommateur naviguant sur un site e-commerce est modélisé par une loi log-normale de paramètre $(\mu_0, \sigma^2 = 1)$, c'est -à-dire la variable $Y_1 = \log(X_1)$ suit une loi normale $\mathcal{N}(\mu_0, 1)$. L'historique du site permet de déterminer la valeur de μ_0, et on prendra $\mu_0 = 0$ dans cet exercice. Une nouvelle politique d'affichage des annonces est proposée. On suppose que l'effet de la nouvelle politique est une modification du paramètre de localisation, c'est-à-dire qu'après l'application de la politique, on a $\log(X_1) \sim \mathcal{N}(\mu, 1)$, et on espère que $\mu > 0$. Puisque après tout, il est possible que la nouvelle politique soit nuisible, on considère le test de l'hypothèse nulle $H_0 : \mu = 0$ contre l'alternative $H_1 : \mu \neq 0$. On effectue une étude sur n clients, dont les achats $X = (X_1, \dots, X_n)$ sont supposés indépendants et uniformément distribuées selon la loi de X_1.

1. (1 pt)Préciser l'espace des paramètres Θ du modèle (en choisissant une pramétrisation qui ne fait intervenir que les paramètres inconnus de l'expérience), ainsi que les ensembles Θ_0 et Θ_1 correspondant aux deux hypothèses.

2. (2 pt) On veut construire un test statistique $\delta(X)$ de H_0 contre H_1, tel que la probabilité de rejeter à tort H_0 soit inférieure ou égale à $\alpha = 5/100$. Pour cela, on utilisera la statistique $S_n(X) = \frac{1}{n}\sum_{i=1}^{n} \log(X_i)$ et on prendra une région d'acceptation de type

$$\mathcal{X}_0 = \{x \in \{0,1\}^n : S_n(x) \in]-A, A[\}$$

avec $A > 0$.
   - Quelle est la loi de $S_n(X)$ sous l'hypothèse nulle ? Quelle est la loi de $\sqrt{n}S_n(X)$, toujours sous l'hypothèse nulle ?
   - Exprimer la borne A de la région d'acceptation en fonction de n et d'un quantile d'une loi que l'on précisera.
     précision : le quantile d'ordre p d'une fonction de répartition F est

$$F^{\leftarrow}(p) := \inf\{x \in \mathbb{R} : F(x) \geq p\}$$

Application numérique : Dans la suite, on note $X^n = (X_1, \ldots, X_n)$ un échantillon de taille n et $A(n)$ la borne supérieure de l'intervalle d'acceptation calculé précédemment.

3. (2 pt) Calculer numériquement et afficher $A(10), A(100), A(1000)$. On pourra utilisera la fonction `qnorm`.

On s'intéresse maintenant au risque de deuxième espèce pour ce test, pour $\mu = 0.1$.

4. (2 pt) Calculez (théoriquement) l'espérance de X dans le modèle log-normal décrit précédemment, en fonction de μ.
   indication : on utilisera le fait que $\mathbb{E}(X) = \mathbb{E}(e^{\log X})$.
   Que vaut (numériquement) cette espérance pour $\mu = 0.1$?
5. (2 pt) On appelle δ_n le test construit à partir de X^n en utilisant la borne $A(n)$ calculée plus haut. Tracer la courbe du risque $R(\mu = 0.1, \delta_n)$, pour n variant de 50 à 1000, sur une grille de pas $h = 50$. On pourra utiliser la fonction `pnorm`. En déduire une première approximation de la plus petite valeur de n (notée n_0) telle que la puissance β du test vérifie $\beta(\mu = 0.1, \delta_n) \geq 0.95$.
6. (2 pt) Au vu de la question précédente, affiner la grille pour obtenir la valeur exacte de n_0. (*i.e.* la taille de l'échantillon test nécessaire pour certifier que le risque de première espèce et le risque de deuxième espèce pour $\mu = 0.1$, soient simultanément plus petits que 0.05.)