# 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 les résultats suivants

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 $\mathbb{E}(Z)$.
2. **Loi du $\chi^2$ ('Chi 2')**
   Si $Y_1, \ldots, Y_n$ sont des variables aléatoires i.i.d. de loi normale centrée réduite, de moyenne empirique $\bar{Y}$, alors la variable aléatoire $V = (Y_1 - \bar{Y})^2 + \ldots + (Y_n - \bar{Y})^2$ suit une loi du $\chi^2$ à $n-1$ degrés de libertés.
3. **Loi Gamma** Une variable aléatoire $Y$ suit une loi Gamma de paramètres $a$ et $b$ ($a > 0$ et $b > 0$), notée $\mathcal{G}amma(a, b)$, si elle admet une densité par rapport à la mesure de Lebesgue donnée par
   $$f^{\mathcal{G}}_{(a,b)}(y) = \mathbb{1}_{y>0}\frac{b^a}{\Gamma(a)} y^{a-1} e^{-by}.$$
   On rappelle que pour $a > 0$, $\Gamma(a+1) = a\Gamma(a)$. Si $Y \sim \mathcal{G}amma(a, b)$, on a
   $$\mathbb{E}_{a,b}(Y) = \frac{a}{b} \quad ; \quad \mathbb{V}\mathrm{ar}_{a,b}(Y) = \frac{a}{b^2}.$$

On s'intéresse à la distribution de la taille des fichiers stockés dans un répertoire. Le jeu de données se trouve ici : http://perso.telecom-paristech.fr/~bonald/filesize.txt

Ce jeu de données comporte la taille en octets de $n = 400$ fichiers, soit $x = (x_1, \ldots, x_n)$.

*N.B* Les quantiles de la loi log-normale sont disponibles numériquement dans R, tout comme ceux de la loi normale, grâce aux fonctions `qnorm` et `qlnorm`

**Exercice 1** (Analyse exploratoire (2pts))**:**

1. Tracer un histogramme de la loi empirique de la taille des fichiers en échelle logarithmique (soit $\log(x_1), \ldots, \log(x_n)$).
2. superposer l'histogramme (avec l'option `probability = TRUE`) et la densité d'une loi normale de moyenne et variance respectivement égales à la moyenne et la variance

empiriques des $\log(x_i)$.

Au vu de l'exercice 1, On modélise ces données comme des échantillons i.i.d. d'une loi log-normale de paramètres $\mu, \sigma^2$ (la taille de chaque fichier est donc représentée par une variable aléatoire $X$ telle que $\log(X)$ suit une loi normale d'espérance $\mu$ et de variance $\sigma^2$). On note $\theta = (\mu, \sigma^2)$. Certaines questions font appel à la loi du $\chi^2$.

**Exercice 2** (Estimation ponctuelle (7 pts))**:**

1. Calculer la densité par rapport à la mesure de Lebesgue de la loi log-normale de paramètre $\theta = (\mu, \sigma^2)$, en utilisant un changement de variables approprié.
2. Calculer l'estimateur du maximum de vraisemblance $\widehat{\theta} = (\widehat{\mu}, \widehat{\sigma^2})$ de $\theta$. Cet estimateur est-il biaisé ? Si oui, ce biais est-il significatif pour ce jeu de données ?
3. Représenter la loi associée pour le jeu de données considéré sur le même graphique que la loi empirique (toujours en échelle logarithmique).
4. Calculer le risque quadratique associé à l'estimateur $\widehat{\mu}$ de $\mu$. Cet estimateur est-il efficace ?
5. On s'intéresse maintenant à la taille moyenne des fichiers, $g(\theta) = \mathrm{E}_\theta(X)$. L'estimateur $g(\widehat{\theta})$ est-il efficace ? Comparer la valeur obtenue pour ce jeu de données avec la moyenne empirique.
6. Enfin, on s'intéresse au quantile de niveau 0.95, soit la valeur $q(\theta)$ tel que $\mathrm{P}_\theta(X \leq q(\theta)) = 0.95$. On cherche à estimer $\log q(\theta)$. L'estimateur $\log q(\widehat{\theta})$ est-il efficace ? Comparer la valeur obtenue pour ce jeu de données avec le quantile équivalent de la loi empirique.

**Exercice 3** (Taille de fichiers et modélisation Bayésienne (7 pts))**:**
On considère le même jeu de données qu'à l'exercice précédent et on s'intéresse au paramètre $\sigma^2$ de la loi log-normale de la taille des fichiers. On considérera dans toute la suite que le paramètre $\mu$ est connu, on prendra $\mu = 9.1$ dans les questions numériques.

1. Justifier l'hypothèse '$\mu$ connu, $\mu = 9.1$' : pour cela, estimer l'écart type de l'estimateur du maximum de vraisemblance pour $\mu$ et comparer à une grandeur de référence qui vous parait pertinente.

On se place désormais dans un cadre bayésien pour l'estimation de $\sigma^2$. Pour des raisons pratiques qui apparaîtront ci-dessous, on préfère travailler avec l'inverse de $\sigma^2$, $\lambda = 1/\sigma^2$. On choisit comme prior sur $\lambda$ une loi Gamma $\pi = \mathcal{G}amma(a, b)$ avec $a > 0, b > 0$ des hyper-paramètres fixés par le statisticien.

2. En l'absence d'information pertinente a priori sur la taille des fichiers, on choisit un prior 'large'. Déterminer $a, b$ pour que $\mathbb{E}_\pi[\boldsymbol{\lambda}] = 1$ et $\mathbb{V}\mathrm{ar}_\pi(\boldsymbol{\lambda}) = 10$.
3. Déterminer l'expression de la loi a posteriori de $\boldsymbol{\lambda}$ pour $n$ données $(x_1, \ldots, x_n)$. Calculer numériquement les paramètres de cette loi a posteriori pour le jeu de données fourni.
4. En déduire l'expression de l'estimateur de l'espérance a posteriori pour le paramètre $\boldsymbol{\lambda}$. Comparer avec le résultat obtenu par maximum de vraisemblance.

5. Tracer sur un même graphique la densité de la loi a priori entre 0 et 1, celle de la loi a posteriori. Indiquer par des lignes verticales l'estimateur de l'espérance a posteriori et $1/\widehat{\sigma^2}$.

On veut construire l'espérance a posteriori $\widehat{h} = \mathbb{E}_\pi[h(\boldsymbol{\lambda}) \,|\, x_1, \ldots, x_n]$ de la quantité d'intérêt $h(\lambda) = \log q_\lambda(0.95)$ avec $q_\lambda$ le quantile de la loi log-normale de paramètres $(\mu = 9.1, \sigma^2 = 1/\lambda)$. On ne dispose pas d'expression explicite pour $h(\lambda)$ ni pour $\widehat{h}$. Cependant, comme précisé en introduction du projet, les quantiles de la loi log-normale et de la loi normale sont disponibles numériquement dans `R`.

6. Simuler un échantillon $(\lambda_i)_{i=1,\ldots,M}$ indépendant et identiquement distribué selon la loi a posteriori, avec $M$ suffisamment grand, de manière à approcher $\widehat{h}$ par une moyenne empirique $\tilde{h} = \frac{1}{M}\sum_{i=1}^{M} Z_i$ avec $Z_i$ convenablement construit à partir de $\lambda_i$ et d'une fonction quantile :
   (a) Expliciter $Z_i$ et fournir l'estimation $\tilde{h}$ demandée.
   (b) Donner une estimation de l'écart-type de $\tilde{h}$, conditionnellement à $x_1, \ldots, x_n$.
   (c) Tracer sur le même graphique, en fonction de $M$, l'évolution de $\tilde{h}$ et d'un encadrement de $\tilde{h}$ de largeur 2 écarts-types, pour l'écart-type calculé ci-dessus.

**Exercice 4** (Test d'hypothèses (4 pts))**:**
L'administrateur du réseau cherche à tester l'hypothèse $H_0 : \sigma^2 \leq 8$ contre $H_1 : \sigma^2 > 8$.

1. Construire un test de niveau $\alpha = 0.05$ de $H_0$ contre $H_1$ basé sur la statistique

$$\varphi(X_1, \ldots, X_n) = \sum_{i=1}^{n} (\log(X_i) - \overline{\log(X)})^2$$

   avec $\overline{\log(X)} = \frac{1}{n}\sum_{i=1}^{n} \log(X_i)$. Préciser la région d'acceptation en fonction des quantiles d'une loi que l'on précisera.
2. Quel est le résultat du test sur le jeu de données considéré ?
3. Quel est le seuil minimal $\sigma_0$ tel que l'hypothèse $\tilde{H}_0 : \sigma \leq \sigma_0$ soit rejetée par un test de niveau 0.05 ?