TP 3 : Estimation bayésienne.

```
suppressMessages(library(tidyverse))
```

1 Échantillons gaussiens de variance connue

1.1 Statistique classique

On s'intéresse ici à un echantillon $X = X_{1:n} \sim \mathcal{N}(\mu, \sigma^2)^{\otimes n}$, où l'on suppose la variance σ^2 connue. Le paramètre d'intérêt est donc $\theta = \mu$. Sous les hypothèses qui précèdent, la moyenne $\bar{X}_n$ de l'échantillon suit la loi gaussienne $\mathcal{N}(\mu, \sigma^2/n)$.

1. En utilisant une boucle `for`, simuler `Nsim` jeux de données suivant le modèle d'échantillon gaussien, et calculer les `Nsim` moyennes de chacun de ces échantillons. On enregistrera ces moyennes dans le vecteur `moyennes`. Notez qu'il est inutile d'enregistrer les échantillons.
2. Tracer un histogramme de la distribution des moyennes des `Nsim` jeux de données simulés. Superposer à cet histogramme la densité de la loi de l'estimateur, rappelée ci-dessus. Expliquez ce que vous observez.
3. Tracer la fonction de répartition empirique des moyennes des `Nsim` jeux de données. Superposer à cette fonction la fonction de répartition de la loi $\mathcal{N}(\mu, \sigma^2/n)$. Expliquez ce que vous observez.
4. Tracer la fonction quantile empirique des moyennes des `Nsim` jeux de données, en fonction des quantiles de la loi $\mathcal{N}(0, 1)$. Expliquez ce que vous observez.

```
# 1.
n0 <- 100
mu0 <- 1
sigma0 <- 1
Nsim <- 10000
moyennes <- rep(0, times = Nsim)
for (i in 1:Nsim) {

}
# 2.

# 3.

# 4.
```

5. On rappelle que l'erreur quadratique moyenne d'un estimateur $\hat{\theta}$ d'un paramètre θ est $EQM(\theta) = \mathbb{E}_\theta(\|\hat{\theta} - \theta\|^2)$. Montrer que cette erreur vaut σ^2/n quelque soit θ. Proposer une méthode de Monte Carlo qui permet d'estimer cette erreur quadratique moyenne en $\mu = 1$ en utilisant les simulations réalisées plus haut.

```
#5.
```

1.2 Statistique bayésienne

On suppose maintenant que la loi a priori, et le modèle statistique sont donnés par

$$\mu \sim \mathcal{N}(m, \tau^2), \quad X_{1:n} \mid \mu \sim \mathcal{N}(\mu, \sigma_0^2)^{\otimes n}.$$

1. Montrer que la loi a posteriori est donnée par $\mu \mid X_{1:n} = x_{1:n} \sim \mathcal{N}(\hat{\mu}(x_{1:n}), v^2)$, où

$$\frac{1}{v^2} = \frac{1}{\tau^2} + \frac{n}{\sigma_0^2}, \quad \hat{\mu}(x_{1:n}) = \frac{\frac{m}{\tau^2} + \frac{n\bar{x}_n}{\sigma_0^2}}{\frac{1}{\tau^2} + \frac{n}{\sigma_0^2}}$$

On s'intéresse à un seul jeu de données dans cette section.

2. Simuler un jeu de données de taille $n_0 = 30$ et de moyenne μ de votre choix, et l'enregistrer dans la variable `jdd_gauss`.

```
#2.
```

3. On suppose que l'on sait a priori que la valeur de μ est autour de 0, et on fixe donc m à 0. On suppose aussi que l'on sait a priori qu'il y a 95% de chance que la vraie valeur soit entre -10 et 10. En déduire une valeur de $\tau = \sqrt{\tau^2}$, la calculer et enregistrer le résultat dans la variable `tau`.
4. Calculer les valeurs numériques de $v = \sqrt{v^2}$ et $\hat{\mu}(x_{1:n})$ et enregistrer les résultats dans les variables `v` et `hat_mu`.

```
# 3.
# 4.
```

5. Sur le même graphique, représenter :

- la densité de la loi a priori de μ en orange,
- la densité de la loi a posteriori en bleu,
- la vraie valeur de μ sous forme d'un trait vertical noir, en pointillé.

On pourra commencer par créer un tableau de données, dont la première colonne est une grille des valeurs de x entre -5 et $\mu + 5$.

```
#5.
```

6. Pourquoi les valeurs de la densité a priori ont-elles l'air d'être si faibles ? Commenter le graphique obtenu.
7. Calculer un intervalle de crédibilité de probabilité 95% sur cet exemple. Calculer aussi un intervalle de confiance de niveau 95%. Comparer les résultats obtenus.

```
#7.
```

2 Échantillon binomial

Dans cette partie, on s'intéresse à un échantillon $X_{1:n} \sim \mathcal{B}(p)^{\otimes n}$, dont le paramètre d'intérêt est $\theta = p$.

2.1 Statistique classique

Reproduire l'étude de la partie 1.1 sur le modèle binomial, avec des échantillons de taille $n = 50$, une vraie valeur $p = 0.1$. On s'intéresse à l'estimateur $\bar{X}_n$ de p.

1. En utilisant une boucle `for`, simuler `Nsim` jeux de données suivant le modèle et calculer les moyennes de ces échantillons.

```
# 1.
n1 <- 50
p1 <- 0.1
Nsim <- 1e4
moyennes <- rep(0, times = Nsim)
for (i in 1:Nsim) {
}
```

2. Tracer un histogramme de la distribution des moyennes simulées. Superposer à cet histogramme l'approximation par une loi normale de sa distribution, c'est-à-dire une loi gaussienne centrée en p et de variance $p(1-p)/n$. Pour tracer cette densité, on utilisera une grille de longueur 1000 sur les nombres entre 0 et 1.
3. Tracer la fonction de répartition empirique des moyennes des `Nsim` jeux de données. Superposer à cette fonction la fonction de répartition de la loi $\mathcal{N}(p, p(1-p)/n)$.
4. Tracer la fonction quantile empirique des moyennes des `Nsim` jeux de données, en fonction des quantiles de la loi $\mathcal{N}(0,1)$. Expliquez ce que vous observez.

```
#2.
#3.
#4.
```

5. Avec une méthode de Monte Carlo, approcher l'erreur quadratique moyenne de l'estimation de $p = 0.1$ par la moyenne de l'échantillon.

2.2 Statistique bayésienne

Reproduire l'étude de la partie 1.2 sur un jeu de données de taille $n = 50$ simulé avec $p = 0.1$. On supposera que la loi a priori est la loi uniforme sur $[0;1]$ (aucune information précise sur la postition de $p\dots$)

On rappelle que, dans ce cas, la loi a posteriori est la loi beta donnée par

$$P \mid X_{1:n} = x_{1:n} \sim \text{Beta}\left(1 + s_n, 1 + n - s_n\right), \quad \text{avec } s_n = \sum_{i=1}^{n} x_i.$$

1. Montrer que la loi a posteriori est donnée par la formule ci-dessus.
2. Simuler un jeu de données de taille $n = 50$ et l'enregistrer dans la variable `jdd_bin`.

```
#2.
```

3. Calculer les valeurs numériques de $\alpha = 1 + s_n$ et $\beta = 1 + n - s_n$.

```
#3.
```

4. Sur le même graphique, représenter :

- la densité de la loi a priori en orange,
- la densité de la loi a posteriori en bleu,
- la vraie valeur de p sous forme d'un trait vertical noir, en pointillé.

On pourra commencer par créer un tableau de données, dont la première colonnne est une grille de valeurs possibles de p entre 0 et 1.

```
#4.
```

5. La densité a posteriori peut-elle être strictement positive pour des valeurs de p strictement négatives ou strictement plus grandes que 1 ? Commenter le graphique obtenu.

6. Calculer un intervalle de crédibilité de probabilité 95 % sur cet exemple. Calculer aussi un intervalle de confiance, et comparer les résultats obtenus.

```
#5.
#6.
```

3 Estimateur de James-Stein

On s'intéresse à l'estimateur du vecteur $\theta = (\theta_1, \ldots, \theta_d)$ de dimension d à partir d'une seule observation vectorielle y de dimension d. Le modèle statistique est : $Y \sim \mathcal{N}_d(\theta, \sigma^2 I_d)$. On supposera σ^2 connu égal à 1, et $d \geq 4$.

L'objectif est de comparer le risque quadratique moyen de deux estimateurs. Le premier estimateur est le maximum de vraisemblance, $\hat{\theta}_{ML} = Y$, et le second est l'estimateur de James-Stein donné par

$$\hat{\theta}_{JS} = \left(1 - \frac{(d-3)\sigma^2}{\|Y\|^2}\right)^+ Y.$$

On peut montrer que l'erreur quadratique moyenne ne dépend de la vraie valeur de θ qu'à travers $\|\theta\|$.

> Proposer, et mettre en œuvre, une méthode de Monte Carlo qui permet d'approcher l'erreur quadratique moyenne en : $\theta = (u, \ldots, u)$ pour $u = 0.1$ ou 0.5, ou 1, ou 2, et $d = 10$.