Statistique computationnelle

Méthodes d'échantillonnage

Vadim Bertrand

15 octobre 2022

Sommaire

1 Exercice 1

1.

$g(x)$ est une densité ssi :

- $g(x)$ est C^0 et positive sur $\mathbb{R}$
- $\int_{-\infty}^{+\infty} g(x)dx = 1$

Or, $exp(u)$ est C^0 et positive sur $\mathbb{R}^-$ (car elle l'est sur $\mathbb{R}$), donc $g(x) = \frac{1}{2}exp(-|x|)$ est C^0 et positive sur $\mathbb{R}$. La figure 1 permet d'illustrer ces propriétés pour $x \in [-5, 5]$.

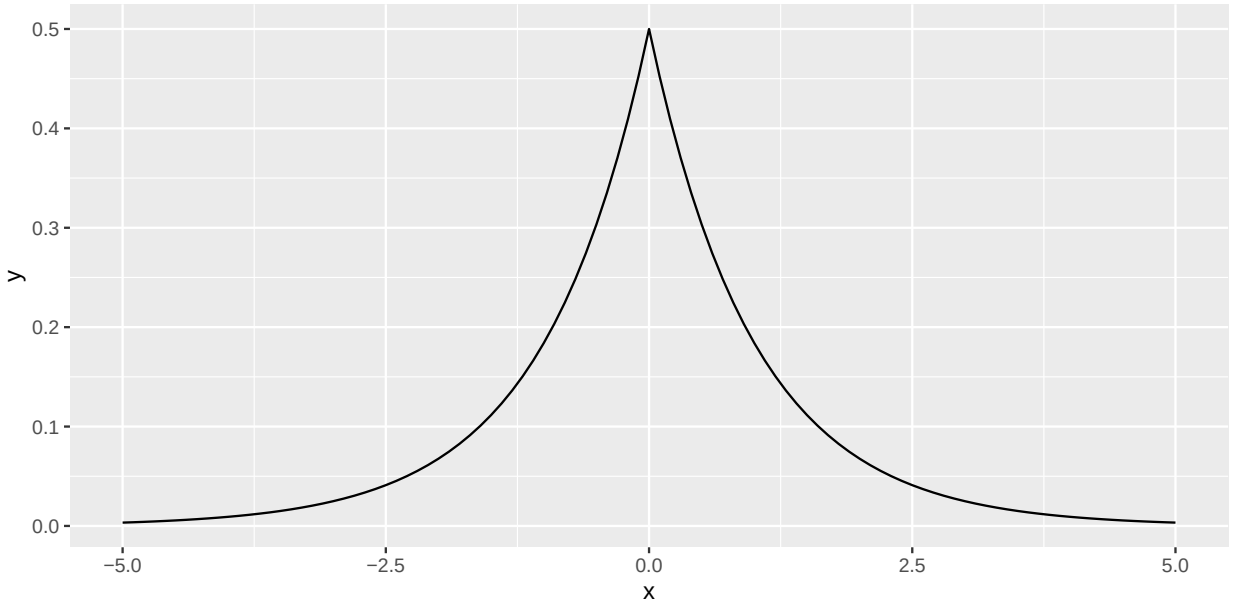


Figure 1: Aperçu de la fonction $g(x) = \frac{1}{2}exp(-|x|)$, pour $x \in [-5, 5]$

Et d'après le développement (1), nous avons bien que $\int_{-\infty}^{+\infty} g(x)dx = 1$:

$$\begin{aligned}\int_{-\infty}^{+\infty} g(x)dx &= \int_{-\infty}^{+\infty} \frac{1}{2}exp(-|x|)dx \\ &= \frac{1}{2}\left(\int_{-\infty}^{0} exp(x)dx + \int_{0}^{+\infty} exp(-x)dx\right) \\ &= \frac{1}{2}\left[[exp(x)]_{-\infty}^{0} + [-exp(-x)]_{0}^{+\infty}\right] \\ &= \frac{1}{2}(1+1) = 1\end{aligned} \tag{1}$$

Ce qui équivaut à vérifier que $g(x)$ est bien une densité.

2.

La méthode d'inversion consiste à simuler un échantillon distribué selon une densité d par le biais du tirage d'un échantillon suivant la loi uniforme $\mathcal{U}[0,1]$, permettant ensuite de revenir à la distribution initialement souhaitée (d) via la réciproque de la fonction de répartition de cette densité (fonction quantile).

Soient $G(x)$ la fonction de répartition de $g(x)$ et $G^{-1}(y)$ sa réciproque.
Par définition, nous avons : $G(x) = \int_{-\infty}^{x} g(u)du$ et $(G^{-1} \circ G)(x) = x$.

Le développement (2) donne l'expression de $G(x)$ pour $x \in \mathbb{R}^+$ et $x \in \mathbb{R}^-$.

$$\begin{aligned} \forall x \in \mathbb{R}^+, \qquad G_+(x) &= \frac{1}{2}\left(\int_{-\infty}^{0} exp(u)du + \int_{0}^{x} exp(-u)du\right) \\ &= \frac{1}{2}\left[[exp(u)]_{-\infty}^{0} + [-exp(-u)]_{0}^{x}\right] \\ &= 1 - \frac{1}{2}exp(-x) \end{aligned} \tag{2a}$$

$$\begin{aligned} \forall x \in \mathbb{R}^-, \qquad G_-(x) &= \frac{1}{2}\int_{-\infty}^{x} exp(u)du \\ &= \frac{1}{2}[exp(u)]_{-\infty}^{x} \\ &= \frac{1}{2}exp(x) \end{aligned} \tag{2b}$$

L'expression des réciproques de G_+ et G_- est obtenu via le développement (3).

$$\begin{aligned} & G_+(x) = 1 - \frac{1}{2}exp(-x) \\ \Leftrightarrow\ & exp(-x) = 2(1 - G_+(x)) \\ \Leftrightarrow\ & x = -ln(2(1 - G_+(x))) \\ \Rightarrow\ & \forall y \in [0.5, 1], \qquad G_+^{-1}(y) = -ln(2(1-y)) \end{aligned} \tag{3a}$$

$$\begin{aligned} & G_-(x) = \frac{1}{2}exp(x) \\ \Leftrightarrow\ & exp(x) = 2G_-(x) \\ \Leftrightarrow\ & x = ln(2G_-(x)) \\ \Rightarrow\ & \forall y \in [0, 0.5], \qquad G_-^{-1}(y) = ln(2y) \end{aligned} \tag{3b}$$

Munis de ces expressions, nous pouvons implémenter en **R** la fonction de tirage suivante :

```
rg <- function (n) { # retourne les réalisations de g
  Un <- runif(n, min = 0, max = 1) # tirage selon la loi U[0,1]
  Gn <- sapply(Un, function (u) { # tirage selon g via sa fonction quantile
    if (u > 0.5) {
      -log(2*(1-u))
    } else {
      log(2*u)
    }
  })
}
```

3.

Nous avons utilisé notre procédure d'inversion pour générer un échantillon de taille 1000 selon la densité g. La représentation proposée sur la figure 2 nous permet de valider graphiquement cette procédure, étant donné que la densité empirique de l'échantillon (tracée en bleu) est semblable à la densité théorique (en noir).

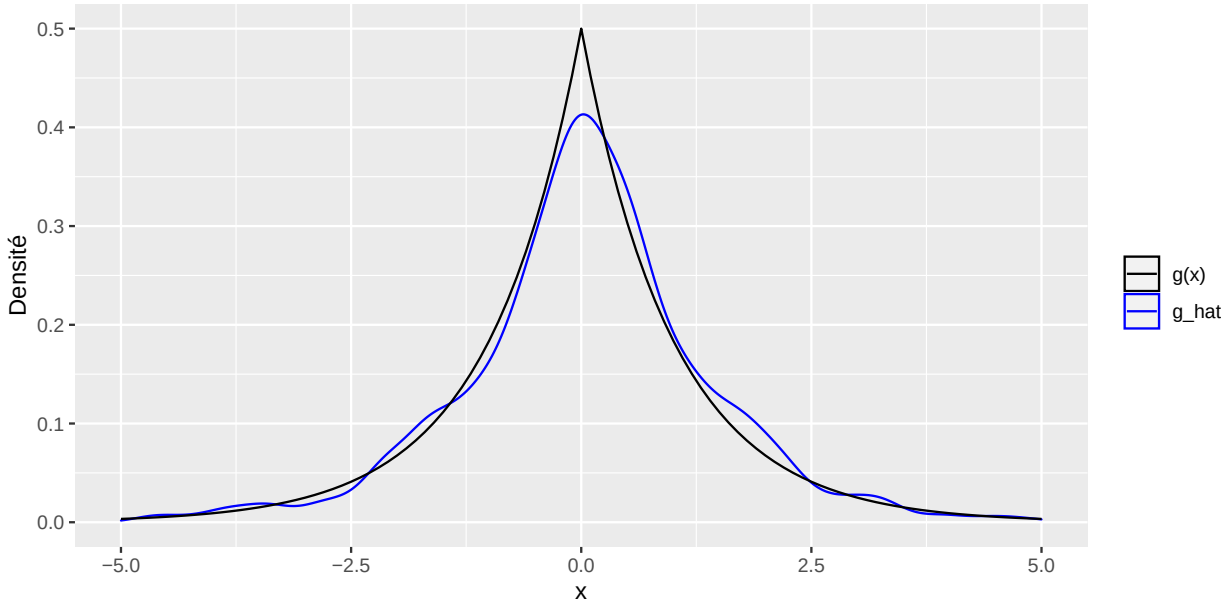


Figure 2: Comparaison de la densité théorique (en noir) et de la densité empirique (en bleu) de g

4.

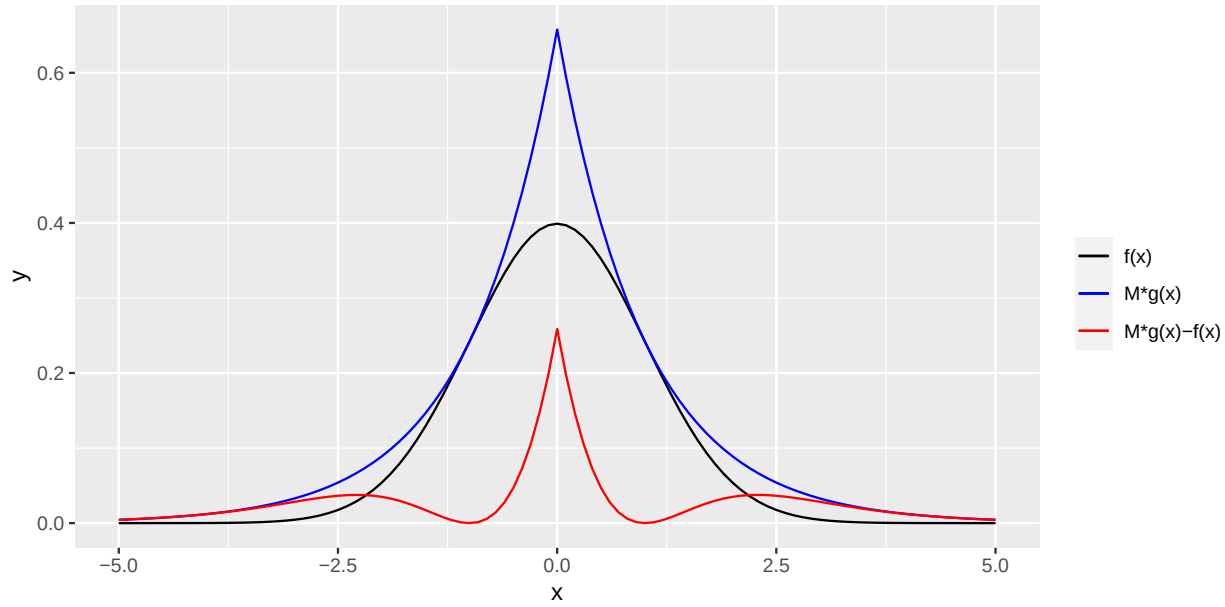


Figure 3: Représentation de $M = \sqrt{\frac{2e}{\pi}}$ plus petit majorant tel que $f(x) \leq Mg(x)$

La figure 3 illustre bien que $M = \sqrt{\frac{2e}{\pi}}$ est le plus petit majorant tel que $f(x) \leq Mg(x)$, $\forall x$ puisque nous

pouvons voir que la courbe rouge, représentant la différence entre $Mg(x)$ et $f(x)$, est tangente à l'axe des abscisses en deux points et positive partout ailleurs.

Cette observation graphique est retrouvée par le calcul selon le développement (4).

$$\begin{aligned} \forall x \in \mathbb{R}, \quad & f(x) \leq Mg(x) \\ & \Leftrightarrow M \geq \frac{f(x)}{g(x)} \qquad (g > 0) \end{aligned} \tag{4a}$$

$$\begin{aligned} \forall x \in \mathbb{R}^+, \quad & \frac{f(x)}{g(x)} = \sqrt{\frac{2}{\pi}} exp(\frac{-x^2}{2} + x) \overset{\text{not}}{=} M_+(x) \\ & \Rightarrow M_+^{'}(x) = \sqrt{\frac{2}{\pi}} exp(\frac{-x^2}{2} + x)(-x+1) \\ & donc, M_+^{'}(x) = 0 \\ & \Leftrightarrow x = 1 \\ & \Rightarrow \min_{x \in \mathbb{R}^+} M_+(x) = M_+(1) = \sqrt{\frac{2e}{\pi}} \end{aligned} \tag{4b}$$

$$\begin{aligned} \forall x \in \mathbb{R}^-, \quad & \frac{f(x)}{g(x)} = \sqrt{\frac{2}{\pi}} exp(\frac{-x^2}{2} - x) \overset{\text{not}}{=} M_-(x) \\ & \Rightarrow M_-^{'}(x) = \sqrt{\frac{2}{\pi}} exp(\frac{-x^2}{2} + x)(-x-1) \\ & donc, M_-^{'}(x) = 0 \\ & \Leftrightarrow x = -1 \\ & \Rightarrow \min_{x \in \mathbb{R}^-} M_-(x) = M_-(-1) = \sqrt{\frac{2e}{\pi}} \end{aligned} \tag{4c}$$

5.

Nous avons donc implémenté la procédure de rejet suivante en utilisant la densité g et le majorant M obtenus précédemment :

```
rf <- function (n) { # retourne les réalisations de f et le taux de rejet
  Xn <- NULL
  rejected <- 0
  while (length(Xn) < n) { # on veut n réalisations
    X0 <- rg(n - length(Xn)) # génération du nombre de réalisations manquantes selon g
    U0 <- sapply(X0, function (x) runif(1, 0, M*g(x))) # tirages dans U[0, g(x0)]
    idx <- U0 <= dnorm(X0) # f=dnorm
    Xn <- c(Xn, X0[idx]) # on conserve les x0 inférieurs ou égaux à f(x0)
    rejected <- rejected + sum(1-idx) # maj du nombre de rejets
  }
  list(Xn = Xn, taux = rejected/(rejected+n))
}
```

L'idée générale est de tirer une réalisation x_0 selon g (sur la courbe noire de la figure 2), puis de tirer une réalisation u_0 selon la loi $\mathcal{U}([0, M * g(x_0)])$ (entre l'axe des abscisses et la courbe bleu de la figure 3) et de conserver x_0 si $u_0 \leq f(x_0)$ (entre l'axe des abscisses et la courbe noire de la figure 3).

De même que pour la procédure d'inversion, nous avons généré un échantillon de taille 1000 selon la densité f via notre méthode de rejet. La figure 4 ci-dessous nous permet de constater que la densité empirique

obtenue (tracée en bleu) est très proche de la densité théorique (représentée en noir), et donc d'attester de la pertinence de notre implémentation.

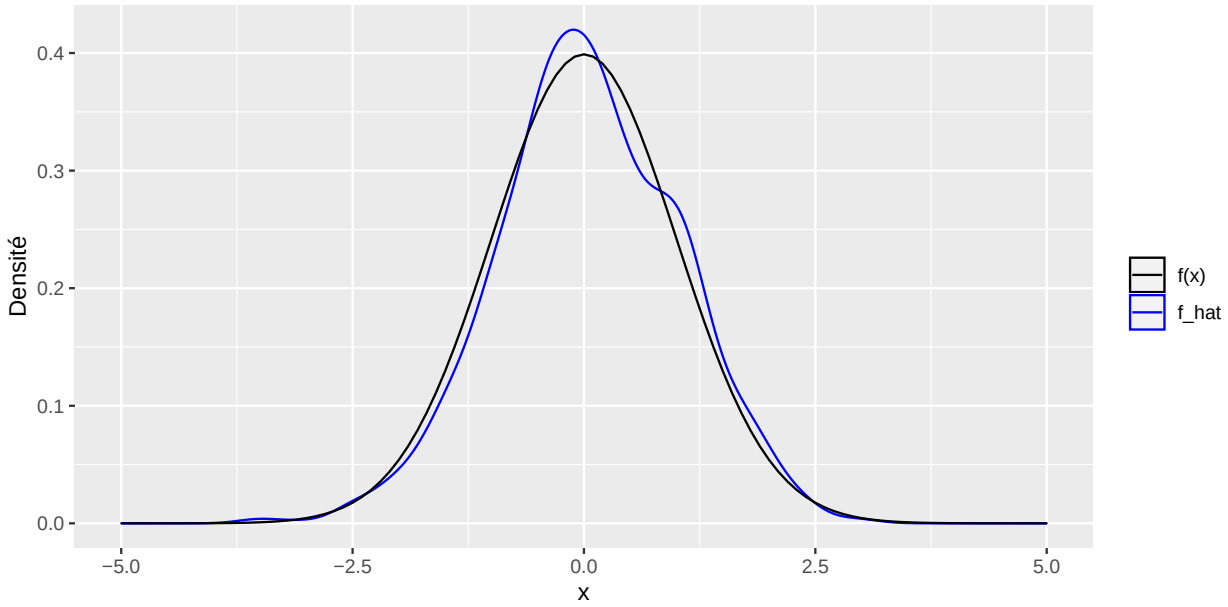


Figure 4: Comparaison de la densité théorique (en noir) et de la densité empirique (en bleu) de f

Nous obtenons un taux de rejet de 0.25. Au premier abord, ce taux peut paraitre étonnamment élevé étant donné que sur la figure 3, la fonction $Mg(x)$ semble assez proche de $f(x)$. Cependant, il faut aussi noter que l'écart entre $Mg(x)$ et $f(x)$ est le plus important pour les valeurs les plus probables de la densité f. Aussi, si nous souhaitions diminuer le taux de rejet, nous pourrions considérer l'utilisation d'une densité auxiliaire permettant de mieux approximer f autour de 0, là où sa probabilité est la plus grande.

2 Exercice 2

Afin de mesurer la puissance du test permettant d'attester du niveau de performance d'un classifieur par une approche Monte Carlo, nous avons généré $M = 10000$ n-échantillons représentant le succès ou non de la classification selon une loi $\mathcal{B}(n, p_1)$, où $p_1 = 0.95$ est le taux de bonne classification observé par validation croisée. Pour chacun de ces n-échantillons, nous calculons le taux empirique de bonne classification (sa moyenne empirique $\bar{X_n}$) et nous déterminons s'il convient de rejeter l'hypothèse nulle du test (que le taux de bonne classification est égale à p_0) en comparant la statistique de test calculée avec $\bar{X_n}$ au seuil de rejet au niveau $\alpha = 0.05$.

Nous avons sous l'hypothèse nulle :

$$\frac{\bar{X_n} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \sim \mathcal{N}(0, 1)$$

Nous utiliserons donc la partie gauche de l'expression comme notre statistique de test. Et comme nous avons $p_1 > p_0$, nous utilisons comme seuil de rejet $q_{1-\alpha}$, le quantile $1 - \alpha$ de la loi normale centrée-réduite. La figure 5 permet de visualiser la région de rejet du test.

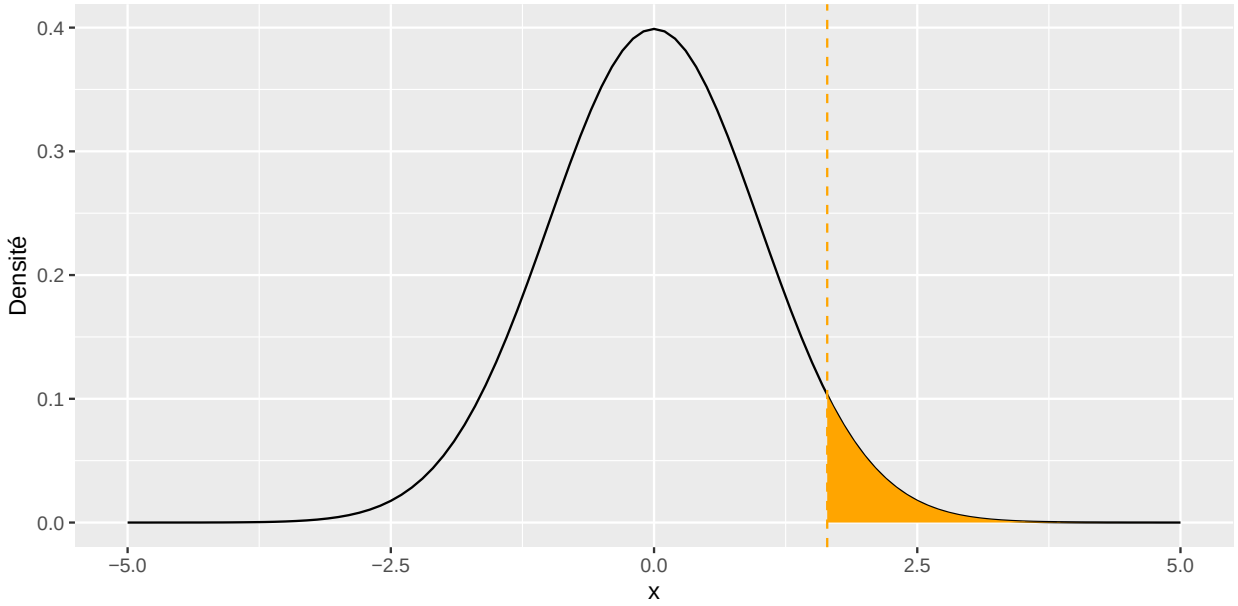


Figure 5: Représentation de la région de rejet du test (en orange)

L'objectif étant de choisir la taille du jeu de validation permettant de démontrer la performance du classifieur, nous avons calculé la puissance de détecter un taux de bonne classification de $p_1 = 0.95$ contre un taux potentiel de bonne classification $p_0 \in 0.8, 0.85, 0.9, 0.92, 0.93$ en faisant varier la taille du jeu de validation $n \in [50, 500]$ par pas de 50.

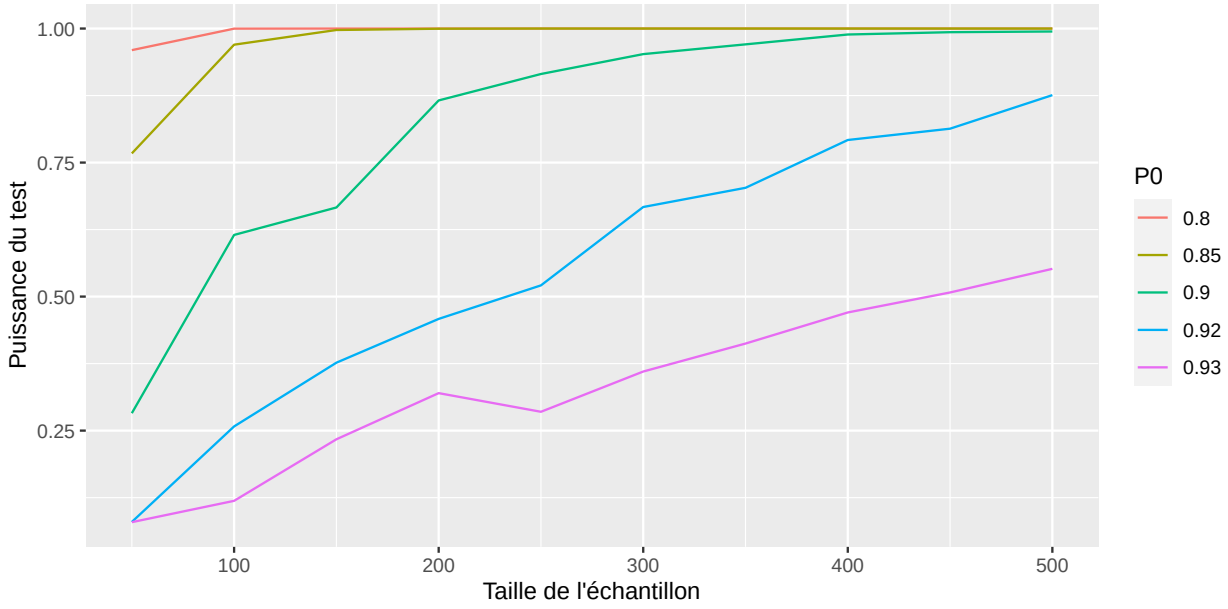
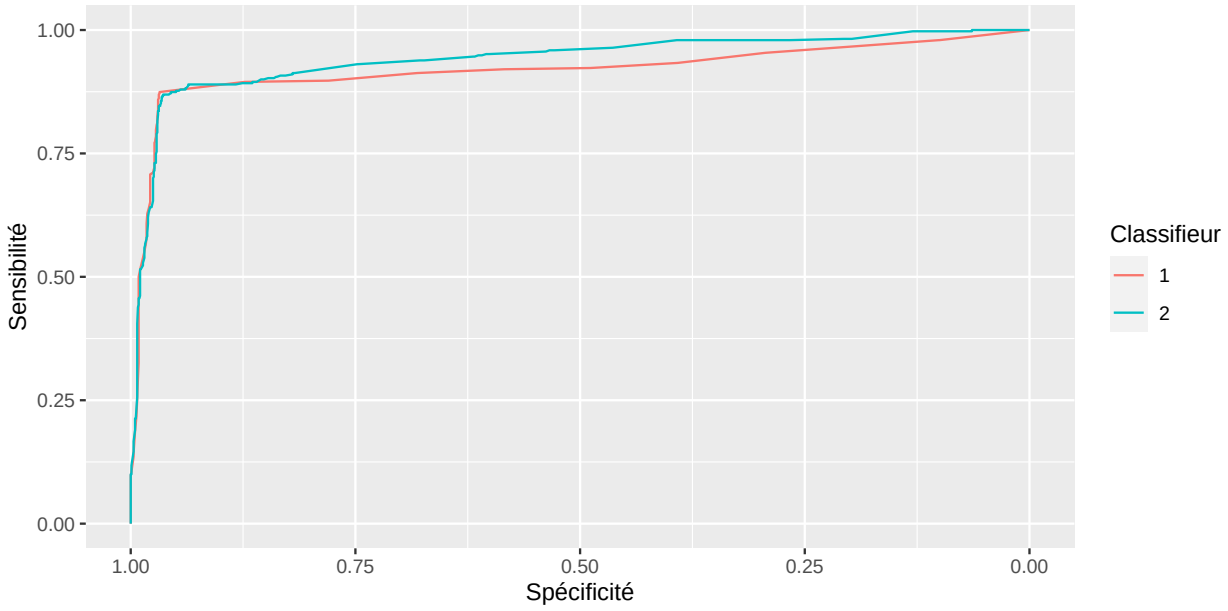
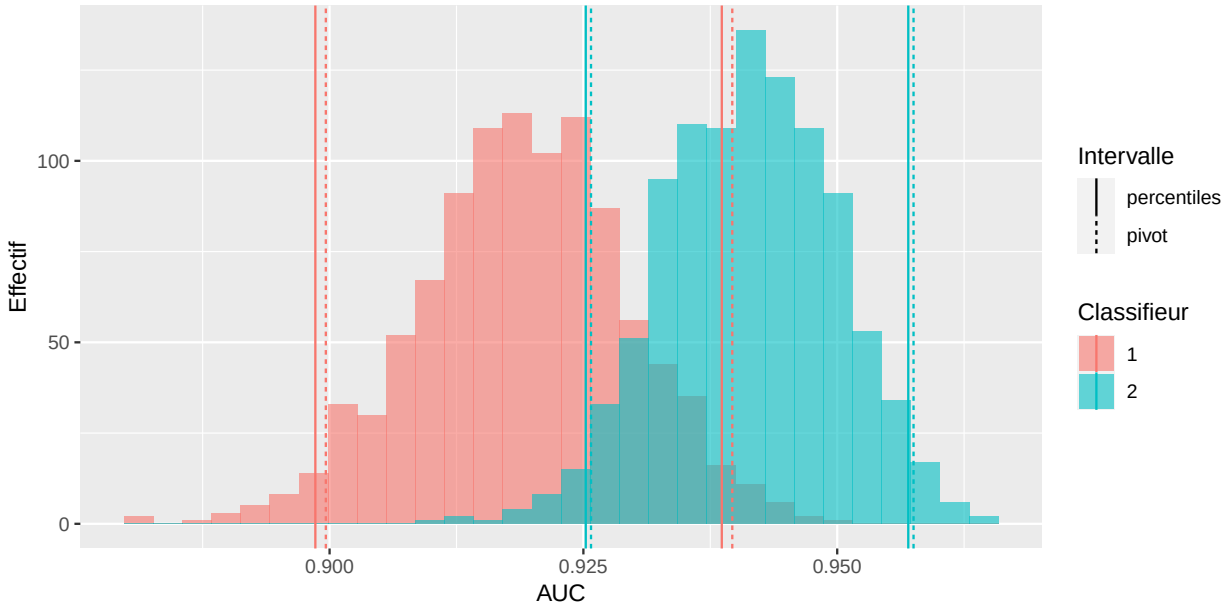


Figure 6: Evolution de la puissance du test selon la taille de l'échantillon, pour différentes valeurs de p_0

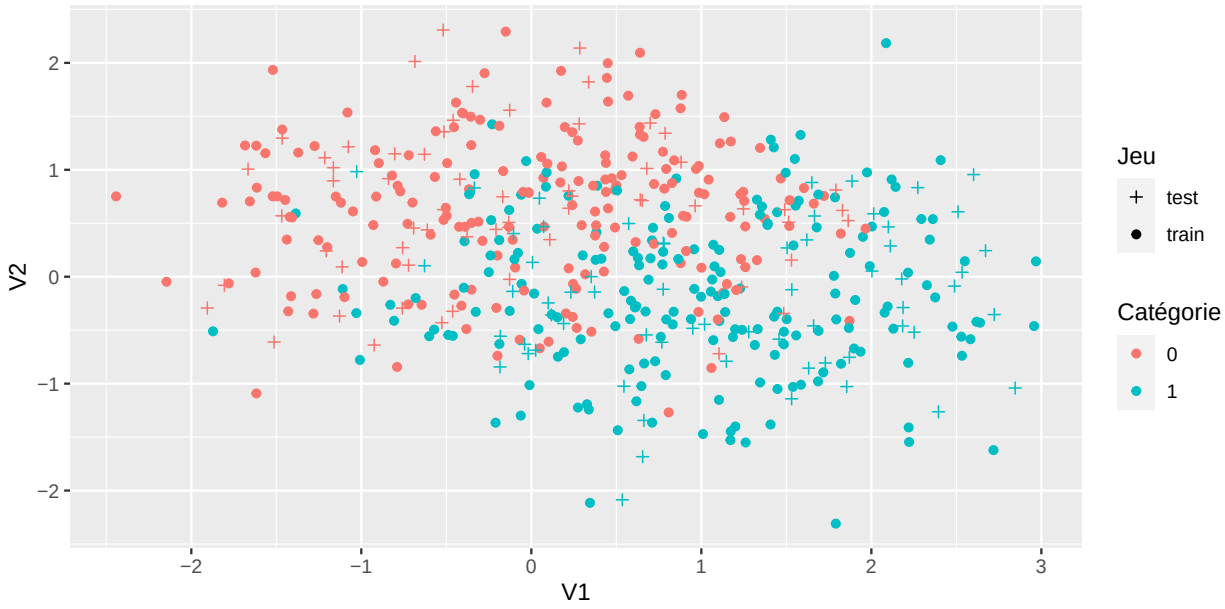
La figure 6 présente les résultats obtenus. Nous pouvons observer que le classifieur permettrait de rejeter l'hypothèse d'un taux de bonne classification égale à 0.8 avec une puissance de 1 (donc une certitude de 100%) à partir d'un jeu de validation de taille 100. En revanche, rejeter l'hypothèse d'un taux à 0.93 avec une grande puissance statistique sera impossible, même avec un jeu de validation de taille 500. S'il est primordial d'avoir une grande confiance dans le fait que le taux du classifieur est supérieur à 0.93 il faudra donc beaucoup plus de données. Si un taux de classification à 0.9 est acceptable, alors choisir un jeu de validation de taille 300 permettra d'avoir une bonne confiance dans les performances attendues du classifieur, étant donné que la puissance statistique du test est alors supérieur à 0.9.

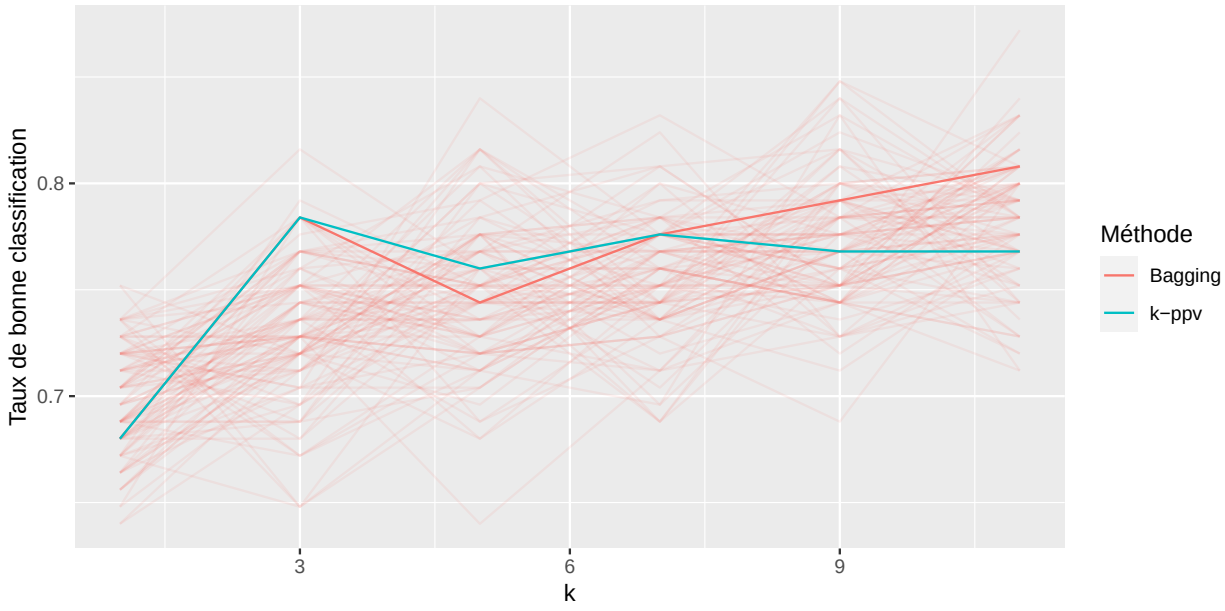
3 Exercice 3


1.00
0.75
0.50
0.25
0.00
Sensibilité
1.00
0.75
0.50
0.25
0.00
Spécificité
Classifieur
1
2


100
50
0
Effectif
0.900
0.925
0.950
AUC
Intervalle
percentiles
pivot
Classifieur
1
2

4 Exercice 4


V2
V1
2
1
0
−1
−2
−2
−1
0
1
2
3
Jeu
test
train
Catégorie
0
1


Taux de bonne classification
k
0.8
0.7
3
6
9
Méthode
Bagging
k−ppv