Statistique computationnelle

Méthodes d'échantillonnage

Vadim Bertrand

28 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 $x \in [-5, 5]$.

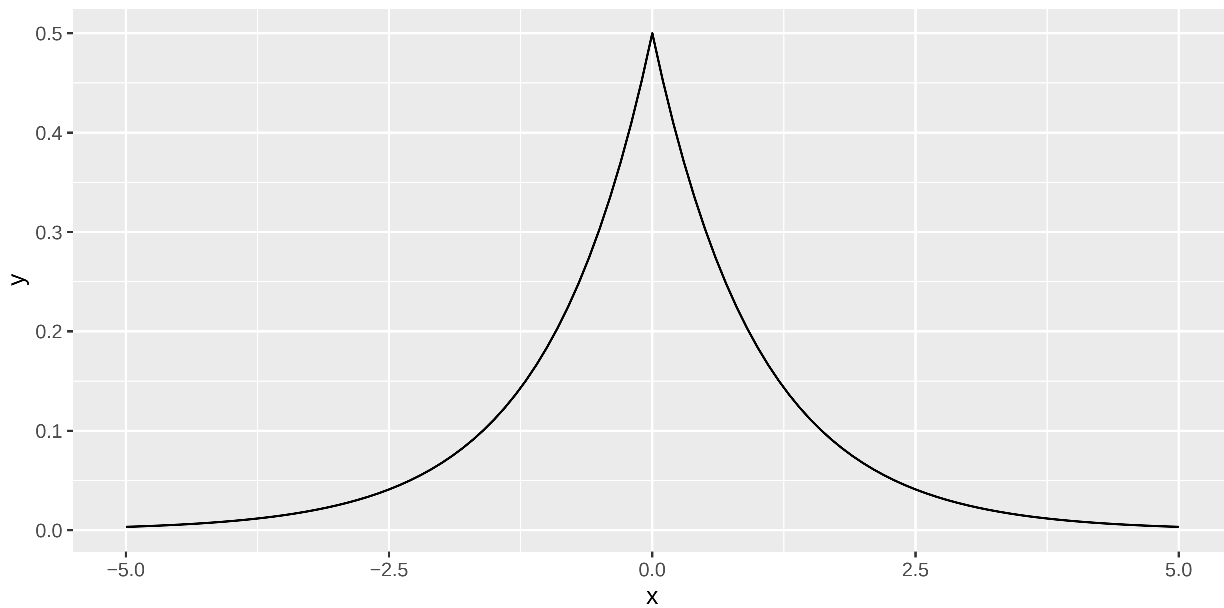


Figure 1: Aperçu de la fonction $g(x) = \frac{1}{2}exp(-|x|)$, pour $x \in [-5, 5]$

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 $x \in \mathbb{R}^-$.

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

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

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

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

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

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

```r
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).

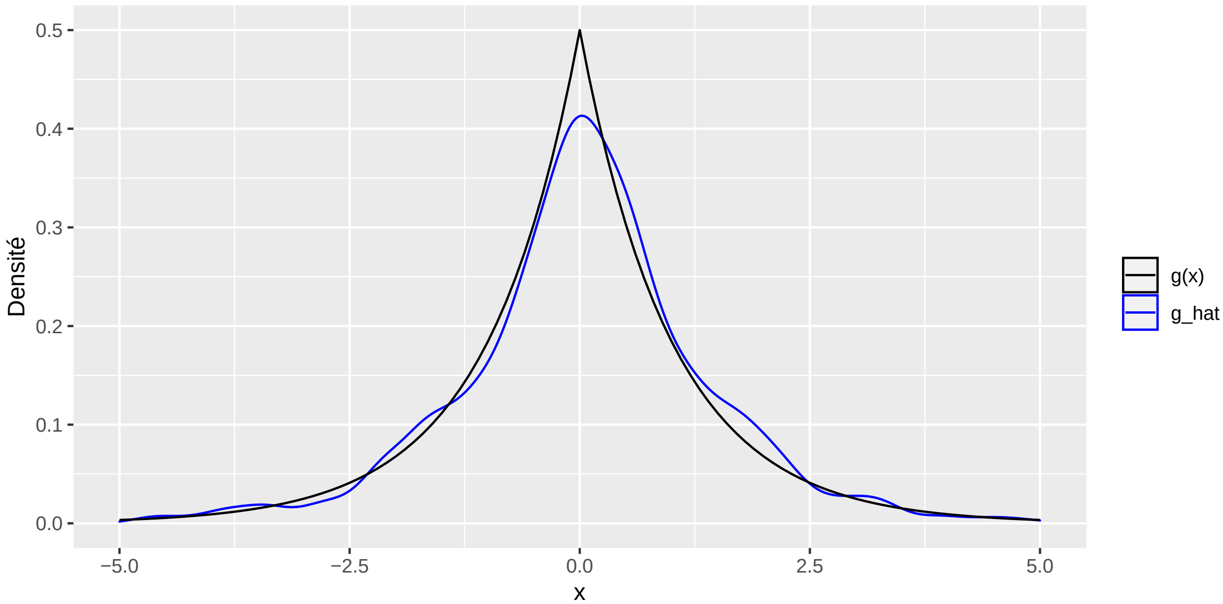


Figure 2: Comparaison de la densité théorique (en noir) et de la densité empirique (en bleu) de g

4.

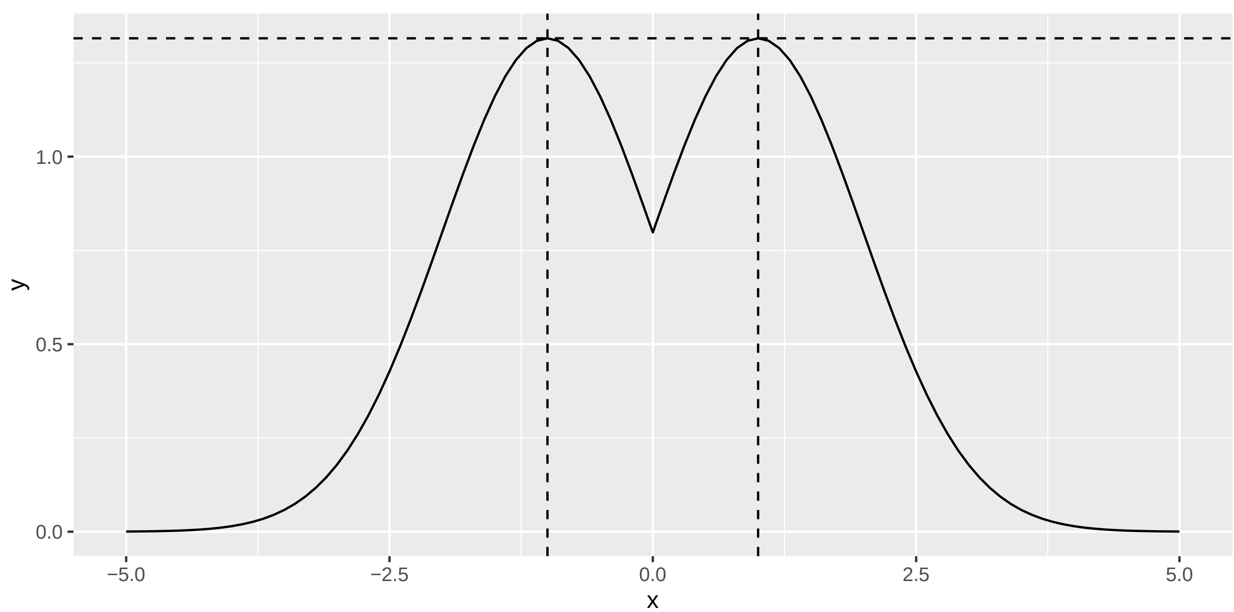


Figure 3: Représentation de $\frac{f(x)}{g(x)}$

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 noire, représentant la rapport entre $f(x)$ et $g(x)$, est inférieure à M (droite horizontale en pointillés) mais tangente en deux points (-1 et 1, droites verticales en pointillés).

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

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

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

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

5.

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

```r
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, rate = rejected/(rejected+n))
}
```

L'idée générale est de répéter M fois :

1. tirer une réalisation x_0 selon g,

2. 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 4),
3. conserver x_0 si $u_0 \leq f(x_0)$ (entre l'axe des abscisses et la courbe noire de la figure 4).

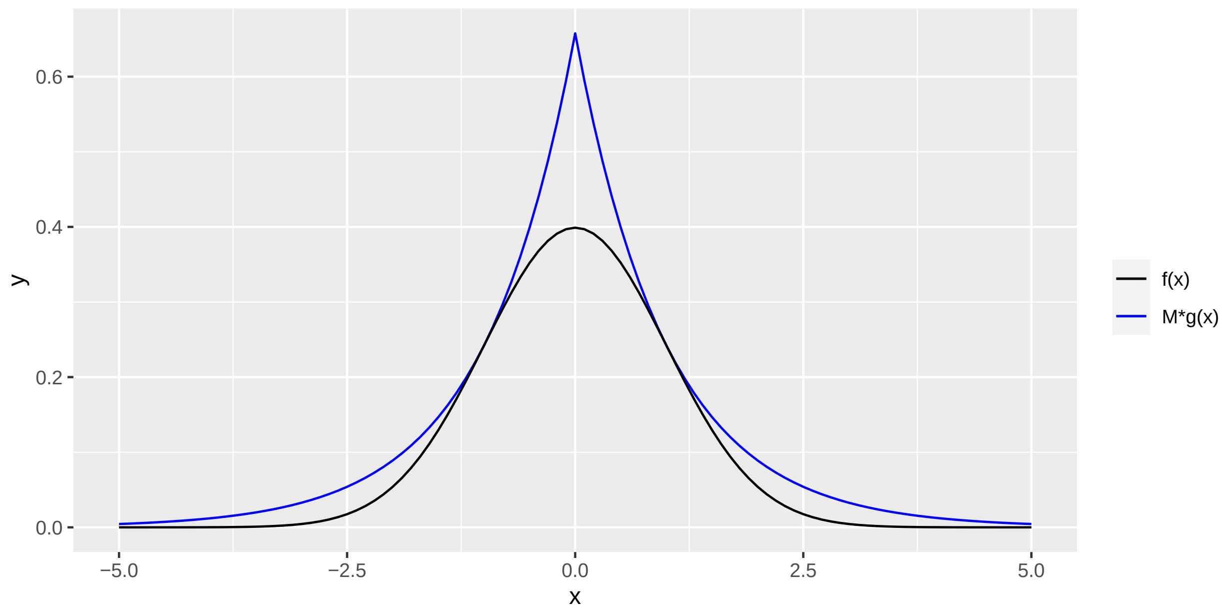


Figure 4: Illustration graphique de la procédure de rejet

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 5 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.

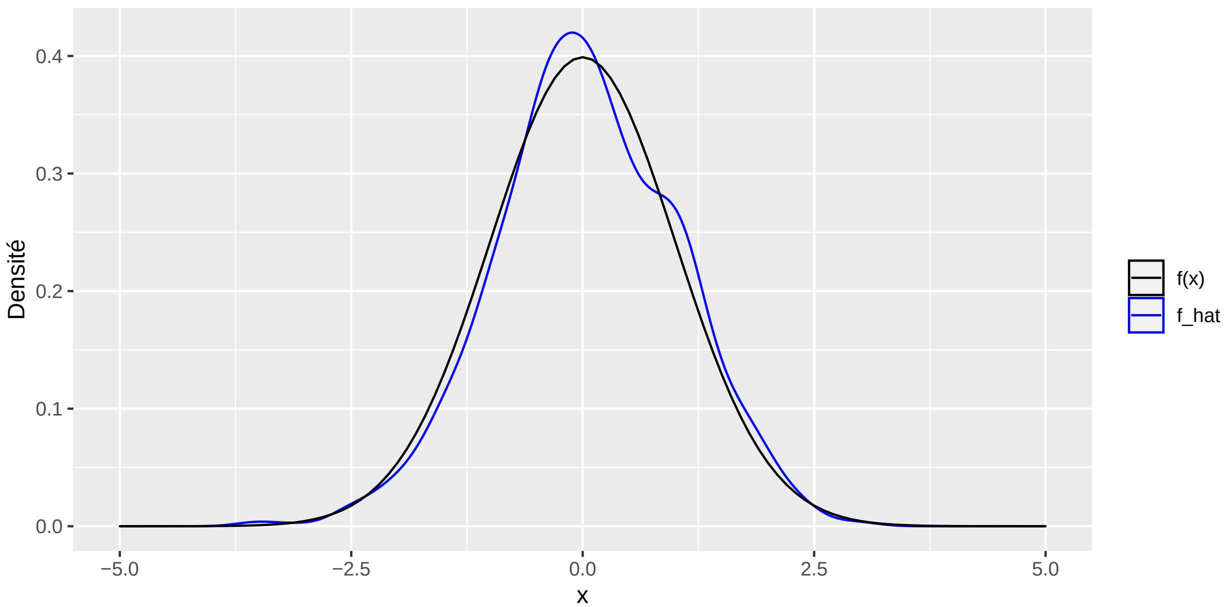


Figure 5: Comparaison de la densité théorique (en noir) et de la densité empirique (en bleu) de f

Le taux de rejet est égale au rapport entre le nombre de tirages rejetés par notre procédure de rejet et le nombre de tirages réalisés total. Au plus ce taux est faible, au plus notre générateur est computationnellement performant.

Nous obtenons un taux de rejet de 0.25, autrement dit : pour 4 tirages, nous en rejetons 1 et nous en acceptons 3. Au premier abord, ce taux peut paraitre étonnamment élevé étant donné que sur la figure 4, 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.
Par ailleurs, si nous nous intéressons aux aires sous les courbes de $f(x)$ et $Mg(x)$ ce taux était même prévisible. Nous savons que $\int_{-\infty}^{+\infty} f(x)dx = \int_{-\infty}^{+\infty} g(x)dx = 1$, donc la "probabilité" (entre guillemets, car ce n'est pas formel) d'accepter un tirage est $\frac{\int_{-\infty}^{+\infty} f(x)dx}{\int_{-\infty}^{+\infty} Mg(x)dx} = \frac{1}{M} \simeq 0.76$ et celle de le rejeter est $1 - \frac{1}{M} \simeq 0.24$: nous retrouvons le taux de rejet.

Si nous souhaitions diminuer le taux de rejet, nous pourrions envisager de :

- considérer l'utilisation d'une densité auxiliaire permettant de mieux approximer f autour de 0, là où sa probabilité est la plus grande,
- utiliser la méthode du rejet adaptatif pour corriger l'approximation de f itérativement.

2 Exercice 2

Dans cet exercice, nous souhaitons nous assurer qu'un taux de bonne classification de 0.95 obtenu par validation-croisée est conservé en confrontant le modèle à de nouvelles données, non utilisées pour son entrainement. Plus précisément, nous souhaitons invalider l'hypothèse que le score de validation est égal à certaines valeurs seuil : $0.8, 0.85, 0.9, 0.92, 0.93$.
Pour cela il est possible de réaliser des tests statistiques comparant les résultats obtenus sur un nouveau jeu de validation aux différents seuils. Sous l'hypothèse de nulle de ces tests, le taux de validation est supposé égal à $p_0 \in 0.8, 0.85, 0.9, 0.92, 0.93$, l'hypothèse alternative étant qu'il vaut $p_1 = 0.95$.
Sous l'hypothèse nulle, nous savons que :

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

Aussi, nous pouvons utiliser la partie gauche de l'expression comme la statistique de test. Et comme nous avons $p_1 > p_0$, le seuil de rejet serait $q_{1-\alpha}^{\mathcal{N}}$, le quantile $1-\alpha$ de la loi normale centrée-réduite. La figure 6 permet de visualiser la région de rejet de ce test.

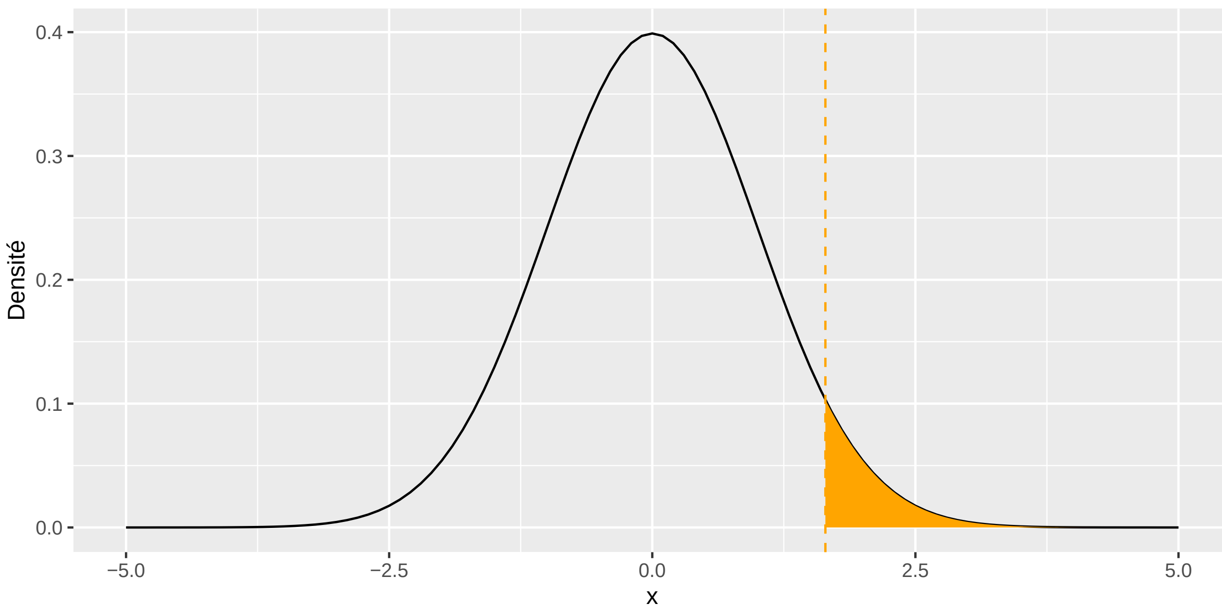


Figure 6: Représentation de la région de rejet du test (en orange)

Afin d'avoir une confiance suffisante dans les résultats des tests nous aimerions connaître leur puissance statistique (c'est à dire la probabilité de rejeter à raison l'hypothèse nulle, et donc d'accepter un taux de bonne classification de 0.95) en fonction de la taille de l'échantillon de validation.
Cette démarche peut être réalisée par une approche Monte Carlo :

1. répéter M fois :
   (a) générer un n-échantillon 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,
   (b) calculer le taux empirique de bonne classification (sa moyenne empirique $\bar{X_n}$),
   (c) comparer la statistique de test calculée avec $\bar{X_n}$ au seuil de rejet au niveau $\alpha = 0.05$ pour décider du rejet ou non du test.

2. calculer la puissance estimée étant alors le nombre de tests rejetés (à raison), divisé par M.

Nous avons réalisé cette procédure en faisant varier la taille du jeu de validation $n \in [50, 500]$ par pas de 50.

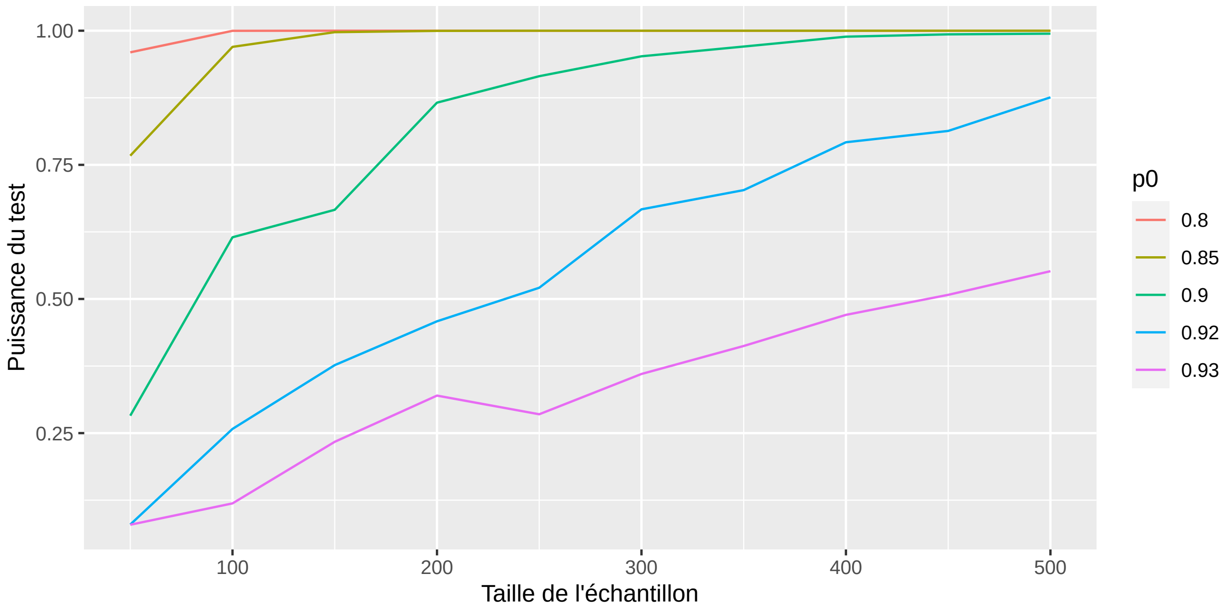


Figure 7: Evolution de la puissance du test selon la taille de l'échantillon, pour différentes valeurs de p_0

La figure 7 présente les résultats obtenus. Nous pouvons observer que la puissance du test augmente avec lorsque la taille de l'échantillon de validation augmente. Cela était attendu étant donné que quand n augmente, la statistique de test augmente également.
Le test 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.

Il existe deux mesures centrales en classification binaire : la sensibilité (le taux de positifs effectivement classés comme positifs) et la spécificité (le taux de négatifs effectivement classés négatifs). Un classifieur parfait aurait une sensibilité et une spécificité égales à 1. En pratique, il faut réaliser un compromis entre sensibilité et spécificité en jouant sur la valeur du seuil de discrimination permettant d'affecter les observations à l'une des deux classes selon leur probabilité d'appartenance à la classe des positifs. La courbe ROC permet de visualiser ce compromis pour différents seuils.

L'AUC correspond à l'aire sous la courbe ROC ; plus l'AUC est proche de 1, meilleur est le compromis sensibilité / spécificité, quel que soit le seuil, et donc meilleur est le classifieur. L'AUC peut également être interprété comme la probabilité que le score d'une observation positive soit supérieur au score d'une observation négative, quelles que soient ces observations. Ainsi, un AUC de 1 signifie que le plus petit score parmi les observations positives est supérieur au plus grand score parmi les négatives, et donc qu'il existe un seuil permettant de discriminer parfaitement les deux classes.

2.

Les deux classifieurs ont un AUC élevé, celui du premier étant égal à 0.92 et celui du deuxième à 0.94. Selon ce critère, le classifieur 2 est donc légèrement plus performant que le 1.

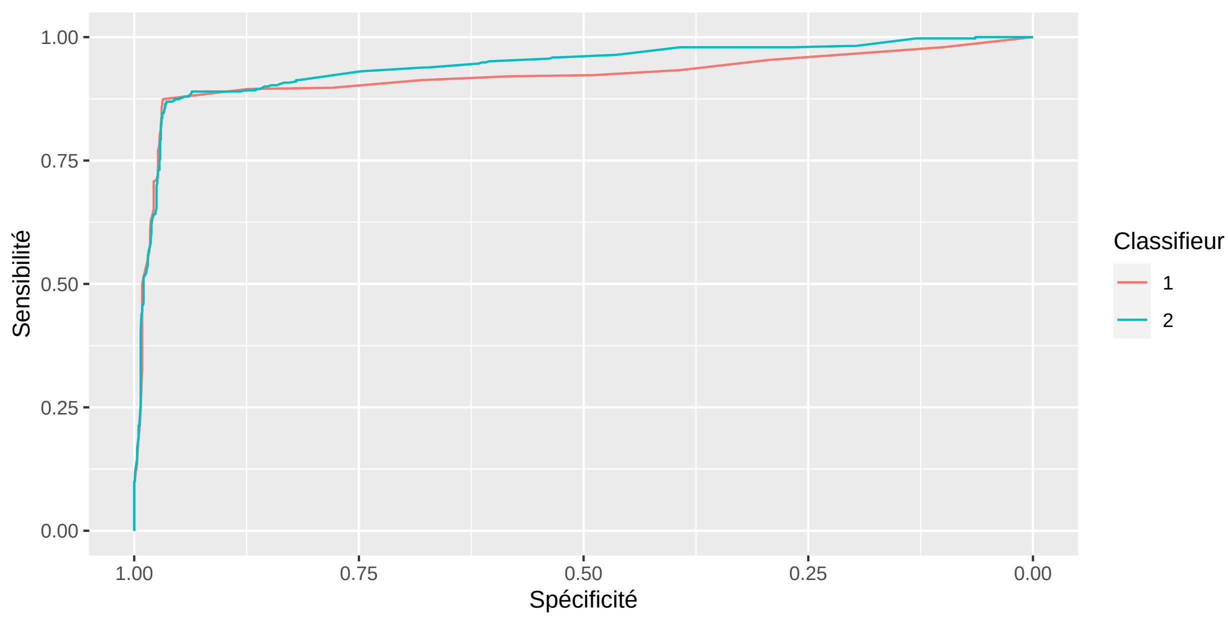


Figure 8: Courbes ROC des deux classifieurs

La figure 8 permet d'arriver à la même conclusion : les deux courbes sont proches des segments ((0,1), (1,1)) et ((1,1), (1,0)), ce qui traduit la bonne qualité des prédicteurs. De plus, la courbe ROC du classifieur 2 (en bleu) est principalement au-dessus de celle du 1 (en rouge), et montre que la sensibilité du classifieur 2 est meilleure que celle du classifieur 1.

3.

Il existe deux procédures bootstrap pour calculer un intervalle de confiance, la méthode des percentiles et celle du pivot. L'approche par les percentiles est la plus intuitive et consiste, pour un intervalle à 95% et dans le cas de l'AUC, à :

1. répéter P fois :
   (a) tirer avec remise un n-échantillon de scores parmis les cas positifs et un n-échantillon parmis les cas négatifs,
   (b) calculer l'AUC à partir de ces deux tirages.
2. calculer les percentiles $\frac{1-95\%}{2}$ et $1 - \frac{1-95\%}{2}$ des P AUC obtenus.

Appliquée aux scores des deux classifieurs étudiés ici, nous obtenons, au niveau 95%, l'intervalle de confiance [0.9, 0.94] pour le classifieur 1 et [0.93, 0.96] pour le 2.

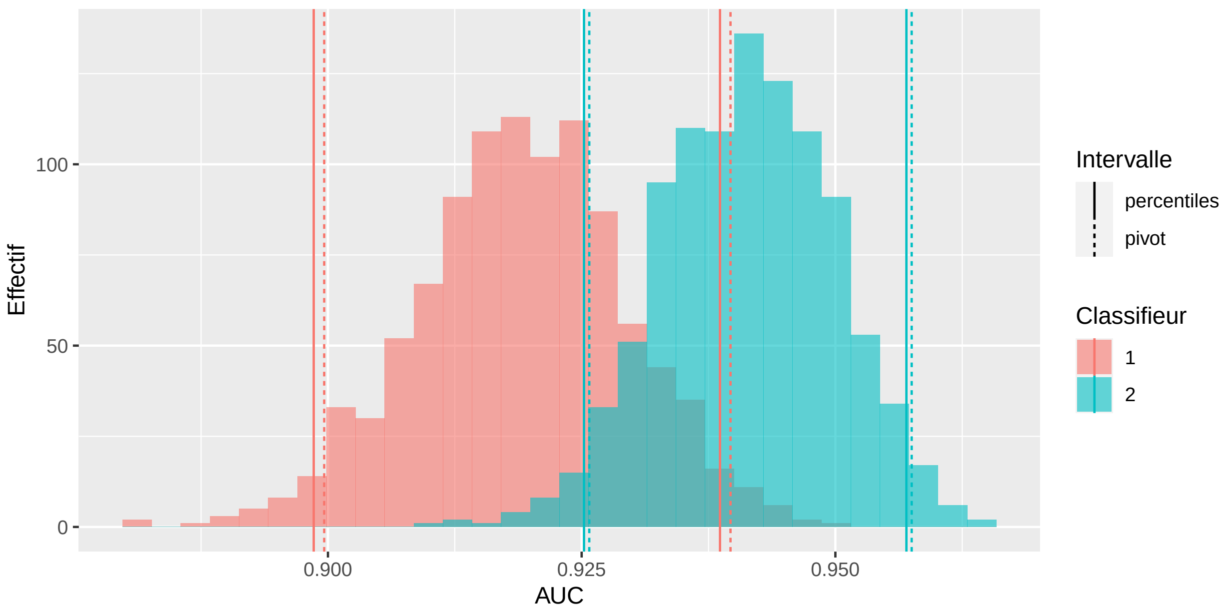
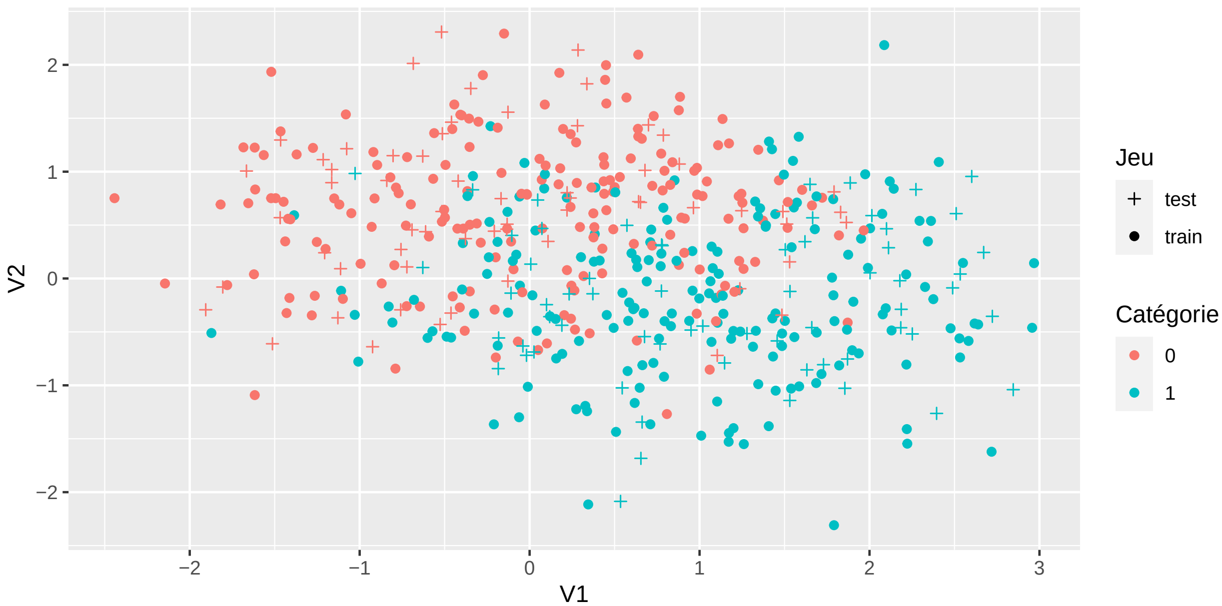


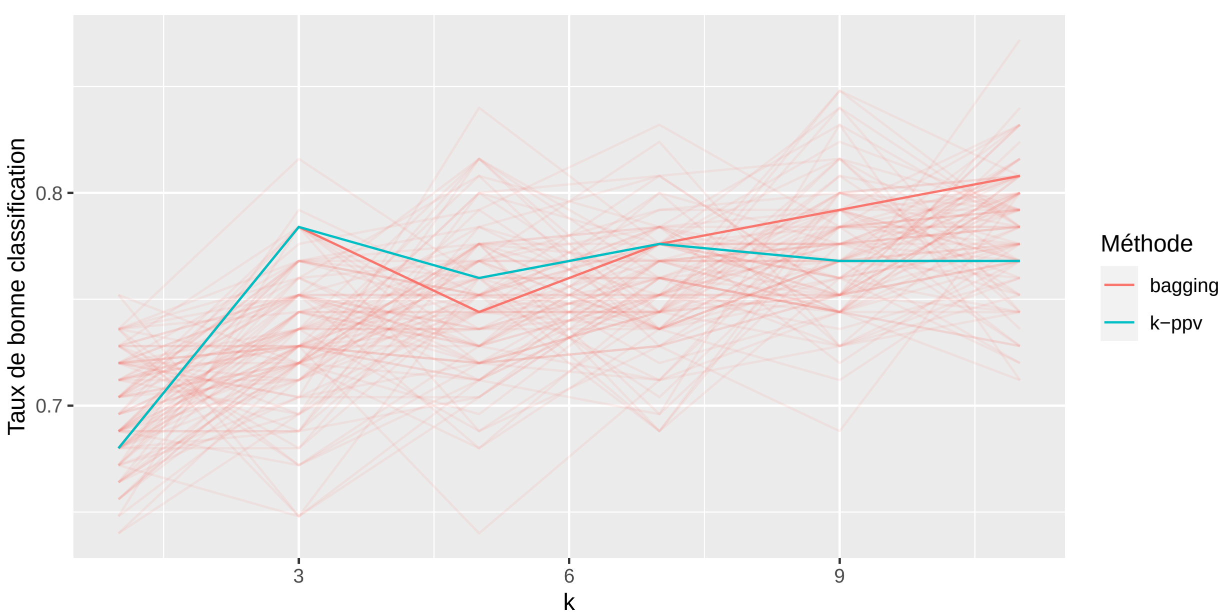
Figure 9: Distribution des AUC obtenus par bootstrap et les intervalles de confiance correspondants

La figure 9 permet de visualiser les distributions des AUC calculées par bootstrap et les intervalles de confiance associés (par la méthode des percentiles décrite précédemment et par celle du pivot). Nous pouvons remarquer que les AUC obtenus pour le classifieur 2 sont légèrement moins étendus que ceux obtenus pour le 1. Par ailleurs, les intervalles de confiance ne sont pas disjoints, et donc que les performances des classifieurs ne sont pas si éloignées.

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
0.8
0.7
3
6
9
k
Méthode
bagging
k−ppv