Examen réechantillonage

Lucas Chabeau, Etienne Hamard

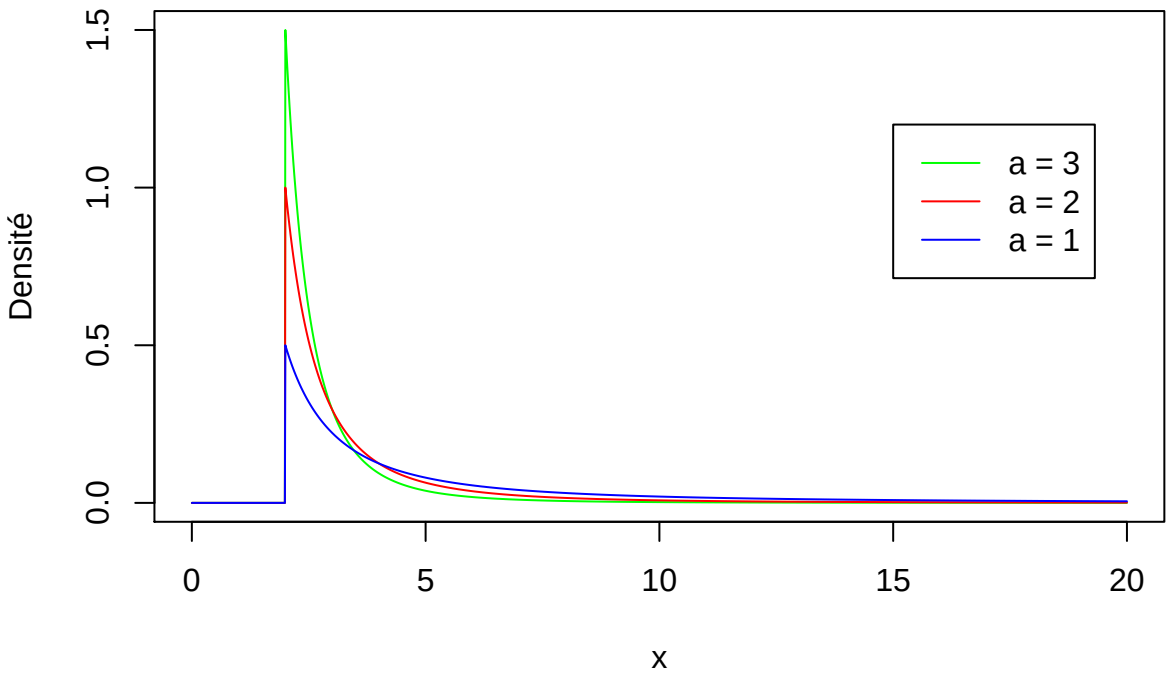
19/11/2019

Exercice 1 - Simulation par inversion

On considère la fonction densité suivante, définie pour $a > 0, b > 0$:

$$f(x) = \begin{cases} \frac{ab^a}{x^{a+1}} & \text{si } x \geqslant b, \\ 0 & \text{sinon.} \end{cases}$$

1. Représenter cette densité pour $b = 2$ et $a \in \{1, 2, 3\}$



2.Implémenterune procédure d'inversion pour simuler une variable aléatoire selon cette densité.

La première étape consiste à donner l'expression de la fonction de répartition $F(x)$ de notre densité $f(x)$:

$$
\begin{aligned}
F(x) &= \int_b^t \frac{ab^a}{t^{a+1}}\,\mathrm{d}t \\
&= ab^a \int_b^t t^{-a-1}\,\mathrm{d}t \\
&= ab^a \left[\frac{t^{-a}}{-a}\right]_b^t \\
&= 1 - (\frac{b}{x})^a
\end{aligned}
$$

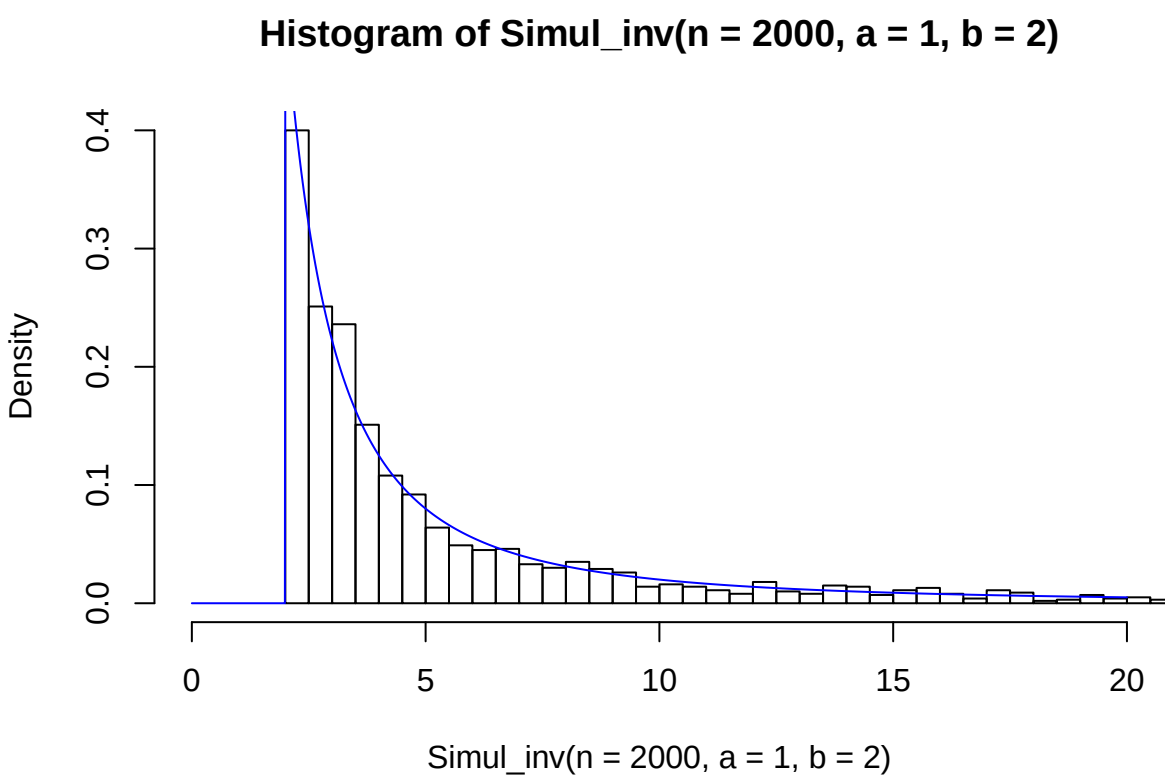
Ensuite il faut inverser la fonction F:

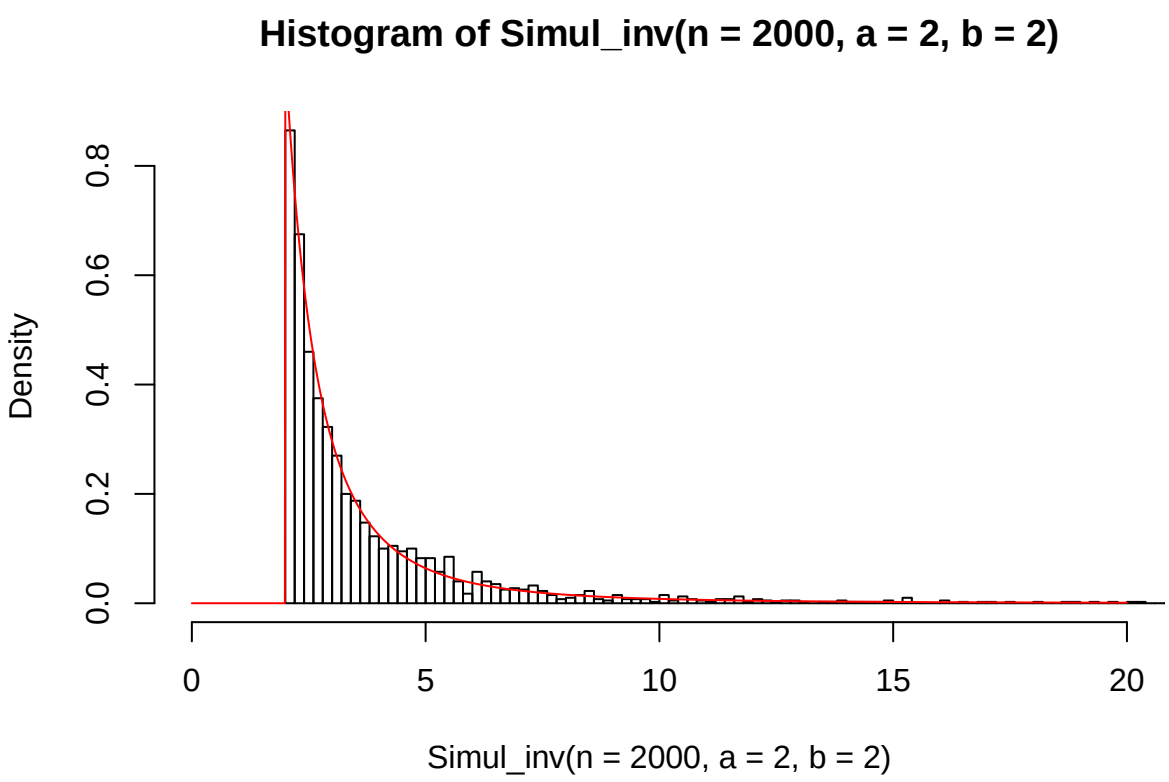
$$
\begin{aligned}
u &= 1 - (\frac{b}{x})^a \\
-u + 1 &= (\frac{b}{x})^a \\
(-u+1)^{\frac{1}{a}} &= \frac{b}{x} \\
\frac{b}{(-u+1)^{\frac{1}{a}}} &= x
\end{aligned}
$$

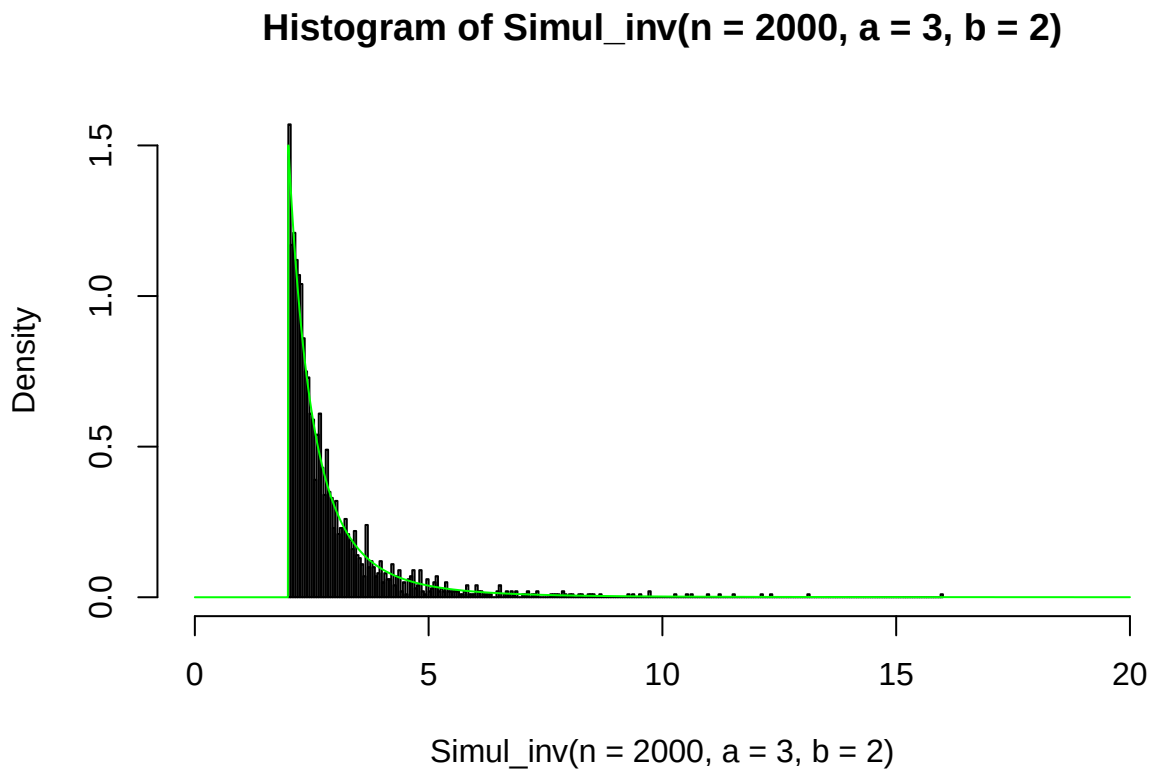
Maintenant que c'est fait nous créons une fonction sur R qui tire un échantillon alétaoire U_n de taille n en suivant une loi uniforme $\mathcal{U}(0,1)$, puis calcule pour chaque tirage i de U $F^{-1}(U_i)$. Ce qui nous donnera un tirage suivant notre densité car $F^{-1}(U) = f(x)$.

3.Simuler un échantillon et comparer graphiquement la densité obtenue empiriquement à la densité théorique

On a simulé un échantillon pour $a = 1$, $a = 2$ et $a = 3$. Nous avons tracé leur histogramme et y avons supperposé la courbe de la fonction de densité correspondante. On voit ça se suit bien. Cette méthode est donc satisfaisante pour simuler un échantillon suivant cette loi.


Histogram of Simul_inv(n = 2000, a = 1, b = 2)
Density
0.0
0.1
0.2
0.3
0.4
0
5
10
15
20
Simul_inv(n = 2000, a = 1, b = 2)


Histogram of Simul_inv(n = 2000, a = 2, b = 2)
Density
0.0
0.2
0.4
0.6
0.8
0
5
10
15
20
Simul_inv(n = 2000, a = 2, b = 2)



Exercice 2 : Approximation de la fonction de répartition de la loi normale centrée réduite par approche MC

On cherche à approximer par une approche Monte-Carlo la fonction de répartition de la loi normale centrée réduite :

$$\phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} exp(-t^2/2)dt$$

pour des valeurs $x > 0$.

1. Proposer une méthode basée sur la loi uniforme, et de préférence la loi $\mathcal{U}(0,1)$.

Nous voulons utiliser une loi $\mathcal{U}(0,1)$ pour résoudre notre problème. Le problème est que notre intégrale est entre $-\infty$ et x. Nous devons faire un changement de variable pour ramener les limites de l'intégrale à $[0,1]$. Nous faisons donc le calcul suivant.

$$I = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} exp(-t^2/2)dt$$

$$I = \int_{-\infty}^{-x} \frac{1}{\sqrt{2\pi}} exp(-t^2/2)dt + \int_{-x}^{x} \frac{1}{\sqrt{2\pi}} exp(-t^2/2)dt$$

$$I = \int_{-\infty}^{-x} \frac{1}{\sqrt{2\pi}} exp(-t^2/2)dt + 2 * \int_{0}^{x} \frac{1}{\sqrt{2\pi}} exp(-t^2/2)dt$$

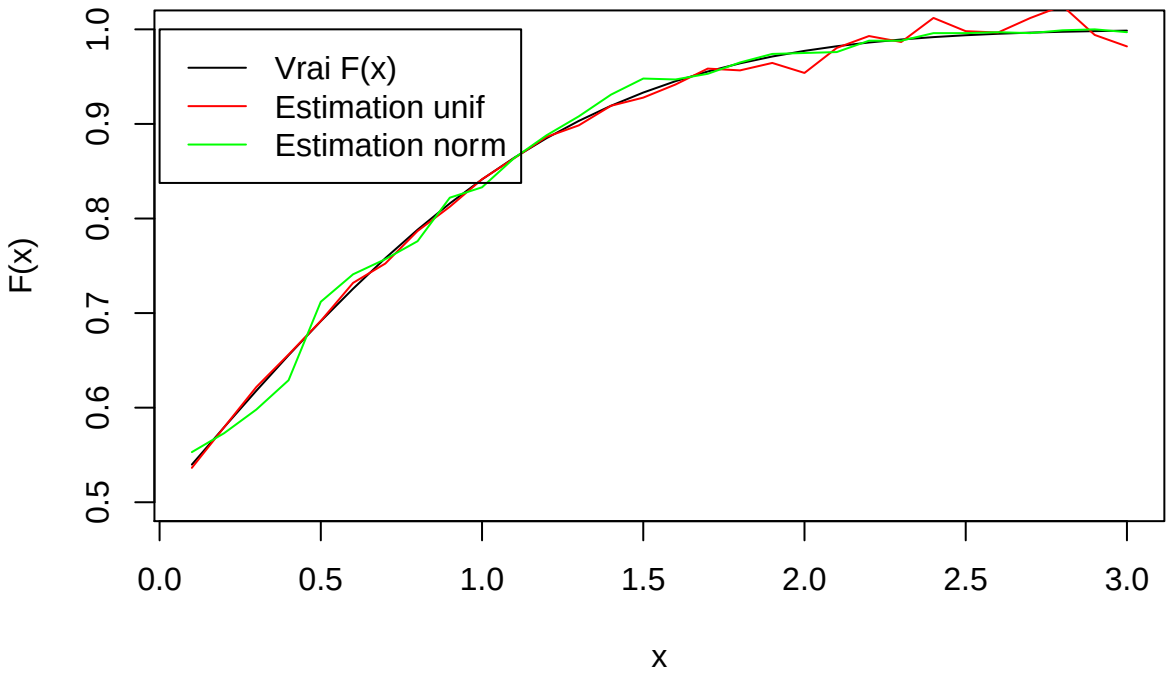
$$I = \int_{0}^{1} \frac{1}{\sqrt{2\pi}} exp(-(-\frac{x}{u})^2/2)\frac{x}{u^2}du + 2 * \int_{0}^{1} \frac{1}{\sqrt{2\pi}} exp(-(ux)^2/2)xdu$$

Une fois cette intégrale posée, nous n'avons plus qu'à tirer m valeurs suivants une loi uniforme $\mathcal{U}(0,1)$ et faire la moyenne des valeurs pour chaque u_i et le quantile x choisi.

2. Proposer une méthode basée sur la loi normale.

Cette fois-ci nous n'avons pas à modifier la formule de notre intégrale. Il faut juste prendre en compte le fait que la loi normale $\mathcal{N}(0,1)$ donne des valeurs dans $\mathbb{R}$. Or nous voulons une intégrale entre $]-\infty; x]$. Nous ajoutons donc une fonction indicatrice qui fera que seules les valeurs tirées inférieures à x sont prises en compte. Nous n'avons ensuite qu'à calculer la moyenne des valeurs aléatoires tirées inférieures à x.

3. Comparer les valeurs obtenues par les deux procédures à la valeur réelle donnée par R via la fonction pnorm pour des valeurs croissantes de $x \in [0.1, 3]$. On considérera un nombre de tirages $n = 1000$.



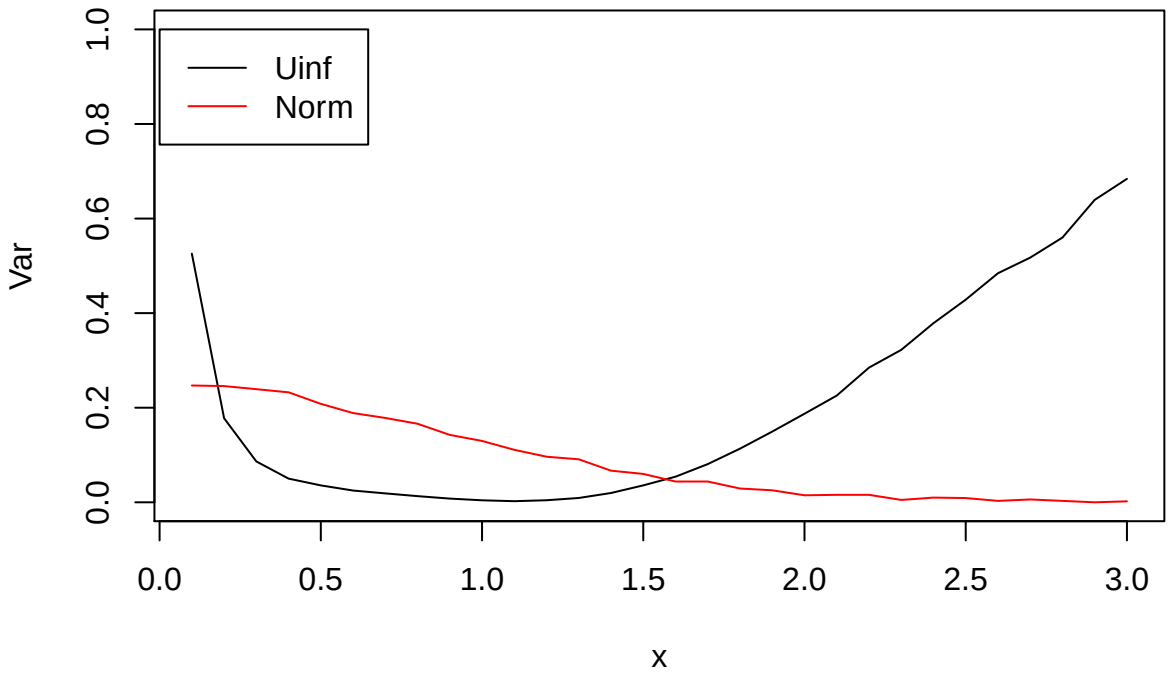
Nos estimations semblent plutôt satisfaisantes tant que x ne séloigne pas trop de 0. Plus on s'éloigne, moins l'estimation par la loi uniforme semble bonne. Pour l'estimation par loi normale en revanche, les estimations semblent meilleures que celles par la méthode basée sur la loi uniforme quand x grandit.

4. Comparer la variance et les intervalles de confiance des deux estimateur pour les valeurs croissantes de $x \in [0.1, 3]$ et interpréter les résultats obtenus.

On rappelle la variance V_n des estimateurs :

$$V_n = \frac{1}{n-1} \sum_{i=1}^{n} (g(Xi) - S_n)\check{s}$$

où S_n est l'estimation ponctuelle.



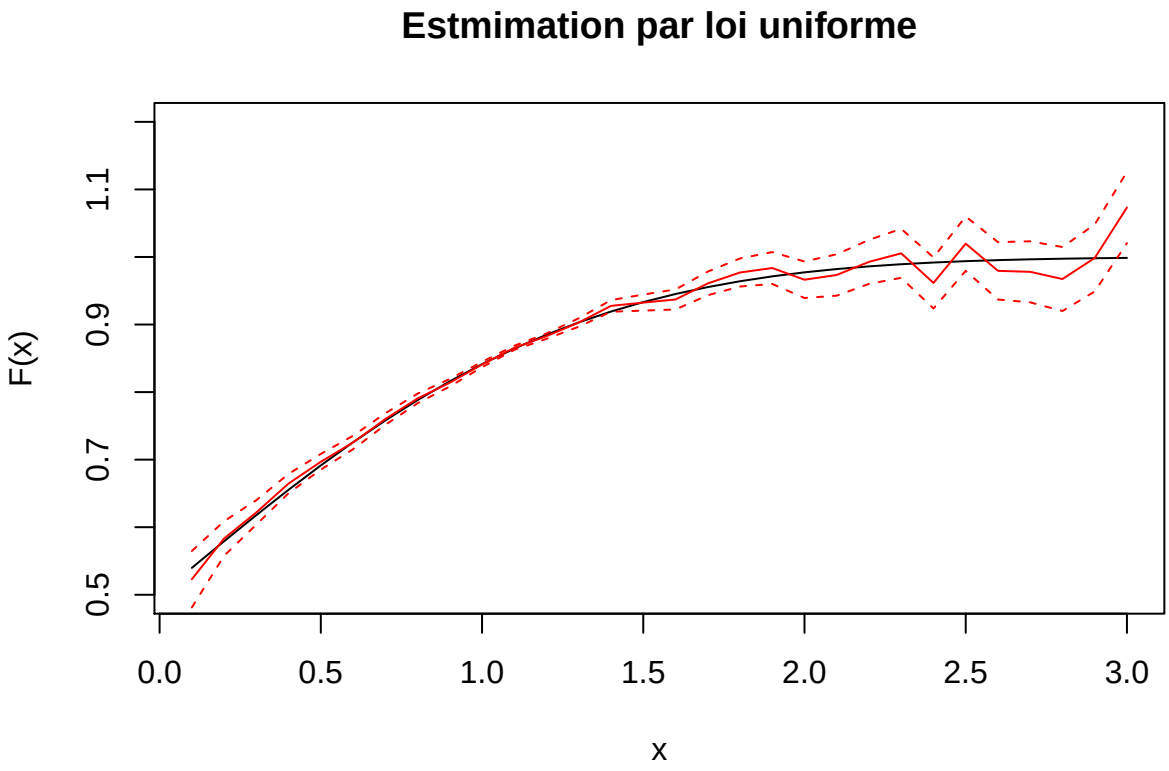
En fonction de la valeur de x la variance sera plus avntageuse pour l'estimation par loi uniforme que par loi normale et inversement.

L'intervalle de confiance est ainsi défini par

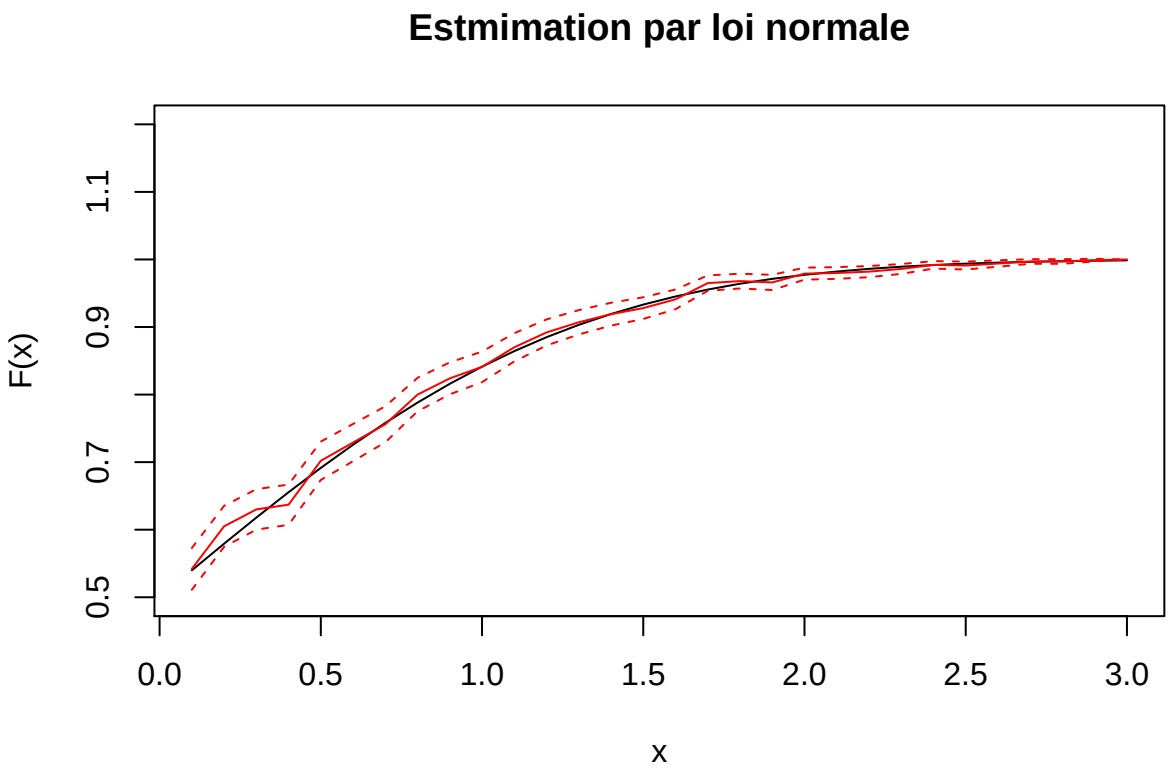
$$[S_n - t_{\alpha/2}\sqrt{V_n/n}; S_n + t_{\alpha/2}\sqrt{V_n/n}]$$

Commençons par regarder les estimations par loi uniforme. Le graphique ci-dessous nous montre en noir la vrai fonction de répartition de la loi normale et en rouge l'estimation ponctuelle entourée par son intervalle de confiance à 95% (en pointillés).

Comme prévu en regardant la variance, nous avons un intervalle de confiance très restreint pour x compris entre 0.5 et 1 puis nous voyons que cet intervalle s'élargit de plus en plus au fur et à mesure que x grandit.



Regardons maintenant le même graphique pour l'estimation par loi normale.



Ici c'est l'inverse, la méthode basée sur la loi normale a un intervalle de confiance qui se restreint au fur et à mesure que x augmente.

Nous concluons que pour des valeurs de x comprises entre 0.2 environ et 1.5, la méthode de Monte Carlo basée sur une loi uniforme sera mailleure que celle basée sur une loi normale. Pour les autres valeurs, ce sera l'inverse.

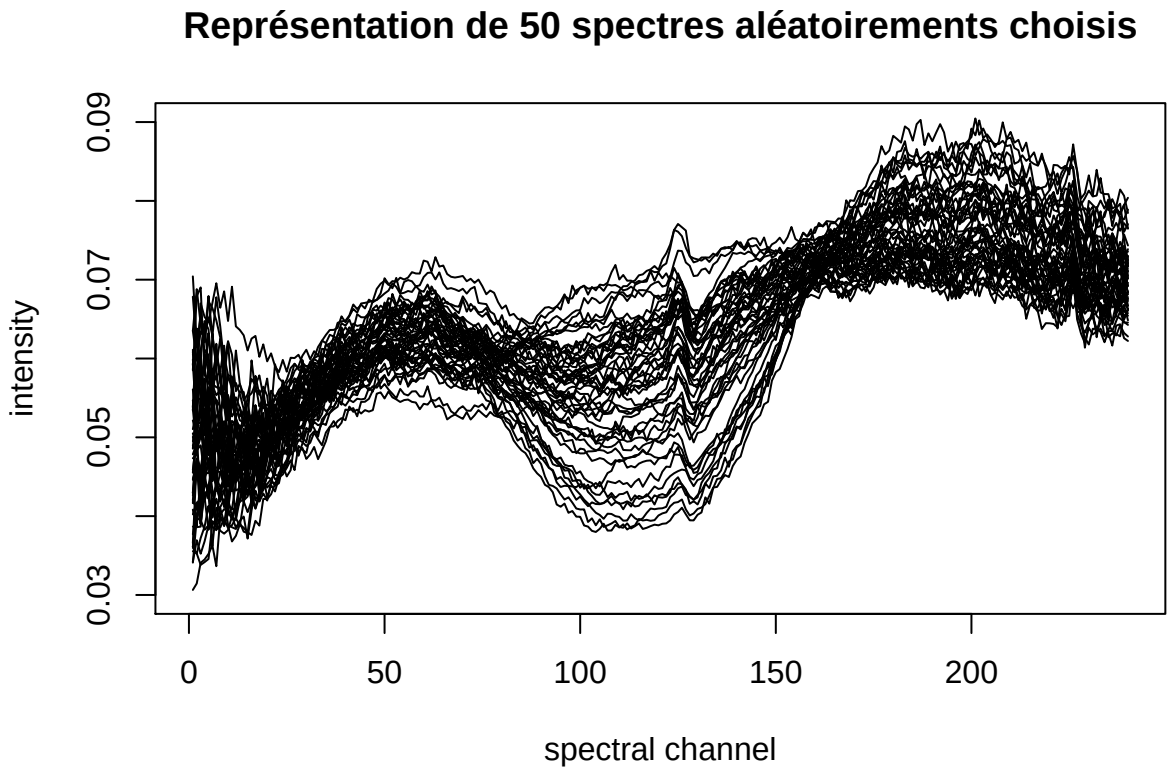
5. Comment modifier la procédure si $x < 0$?

Exercice 3 - bootstrap & ACP

```
## Loading required package: ggplot2
```

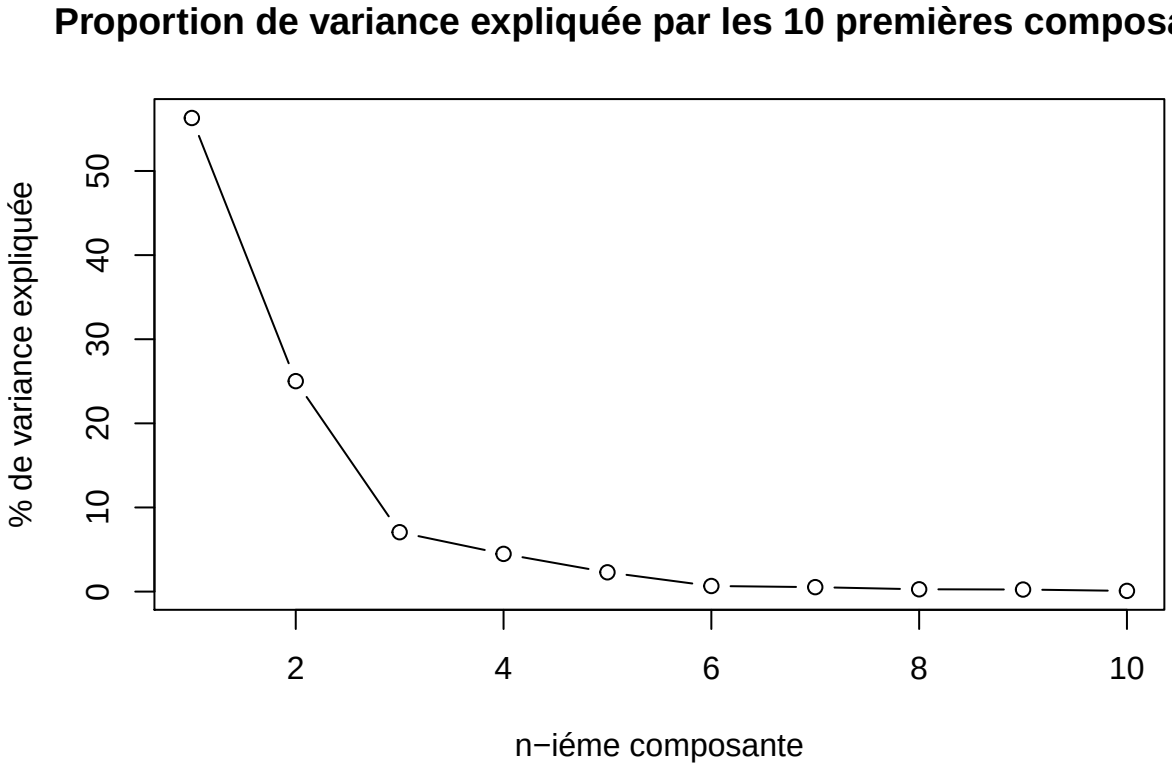
```
## Welcome! Related Books: `Practical Guide To Cluster Analysis in R` at https://goo.gl/13EFCZ
```

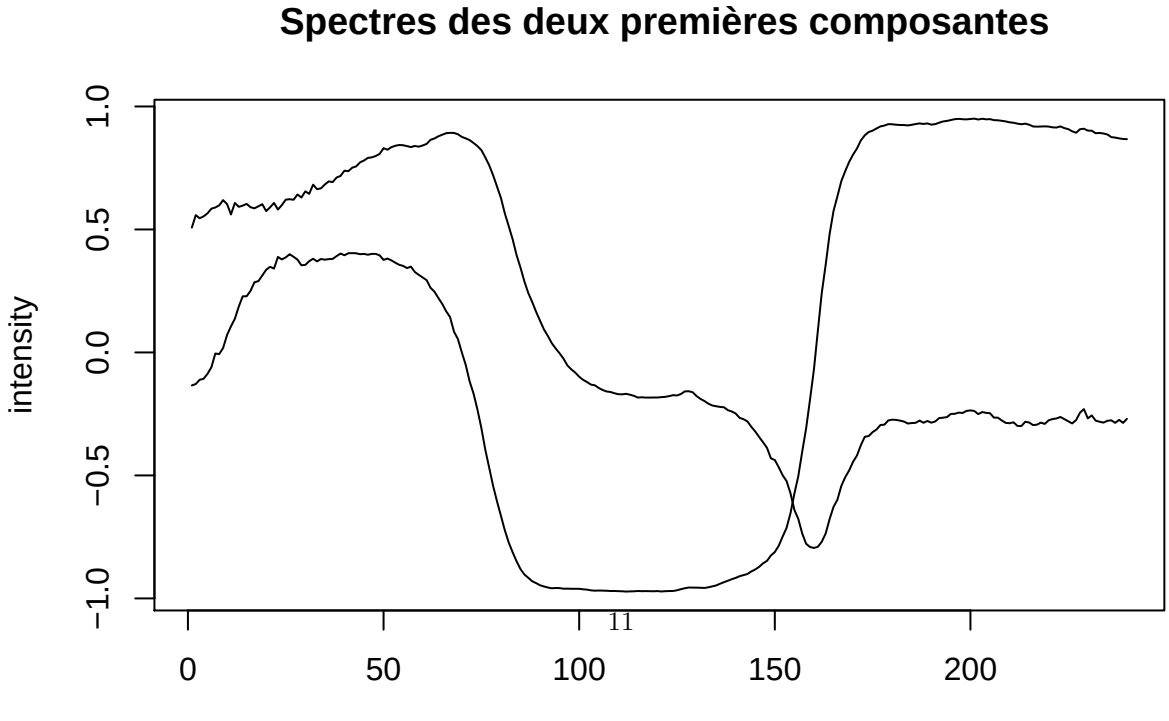
1. Représenter (sur une même figure) un échantillon aléatoire de 50 spectres pour en apprécier leur variabilité, et commenter le résultat obtenu.



On peut voir la variance des spectres augmente entre les indices 100 et 150.

2. Effectuer une ACP et représenter les proportions de variance expliquées par les 10 premières composantes principales : quelle proportion de la variance totale est expliquée par les deux premières composantes principales ? Représenter sous la forme de spectres les deux axes principaux (i.e., les opérateurs linéaires permettant de passer de l'espace des canaux aux deux premières composantes principales) : sont-ils cohérents avec vos observations de la question 1 ?

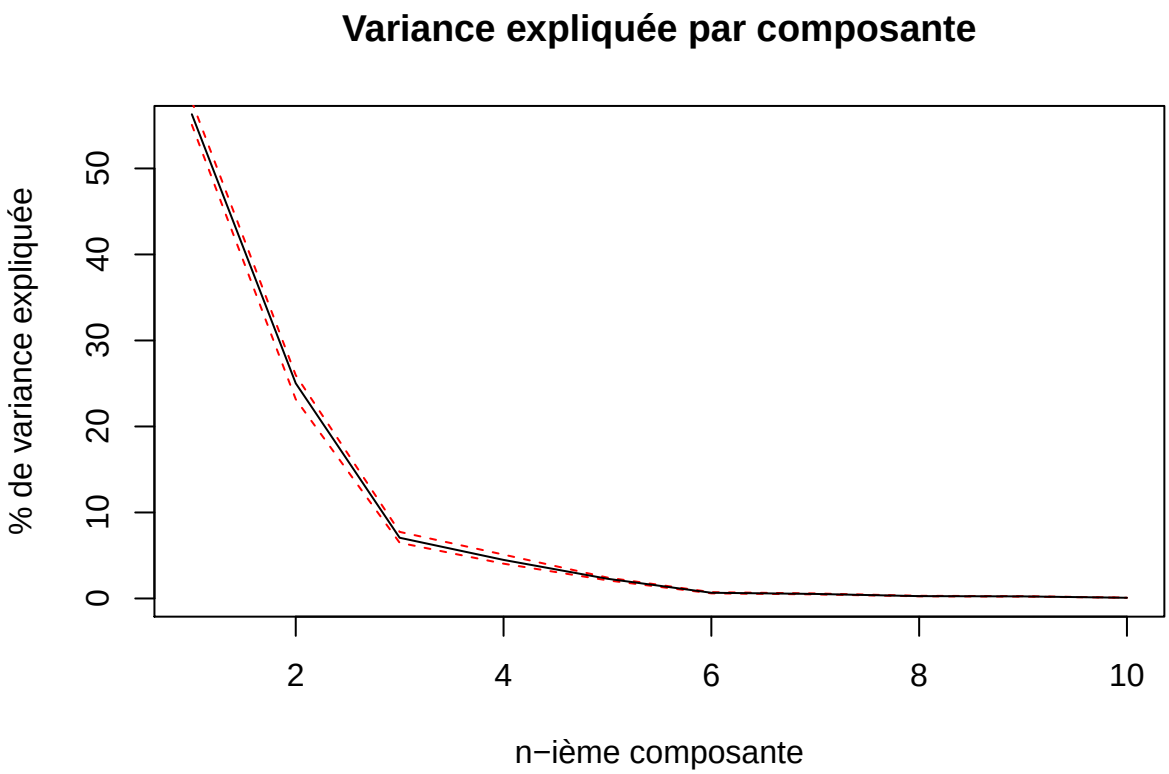




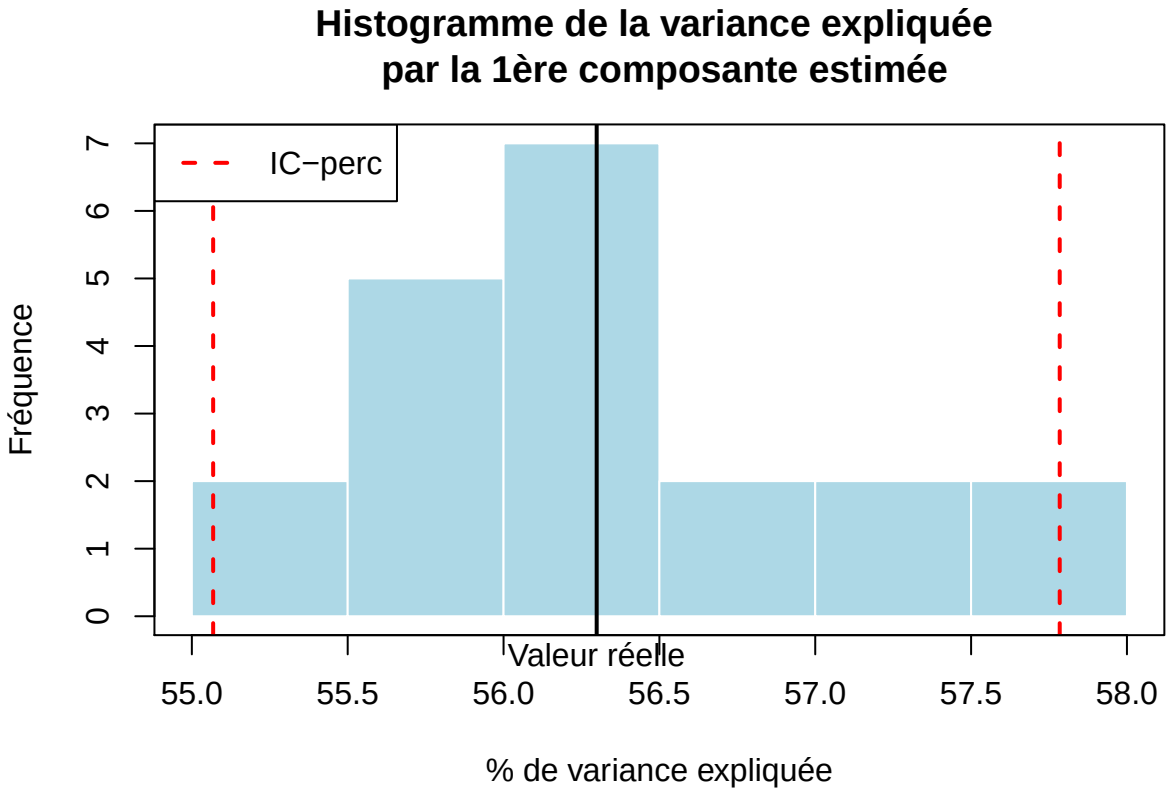
Les deux premières composantes expliquent 81.317% de la variance. La représentation des spectres des deux premières dimensions parait cohérent avec les autres spectres, l'allure est similaire mais les valeurs sont beaucoup plus faibles

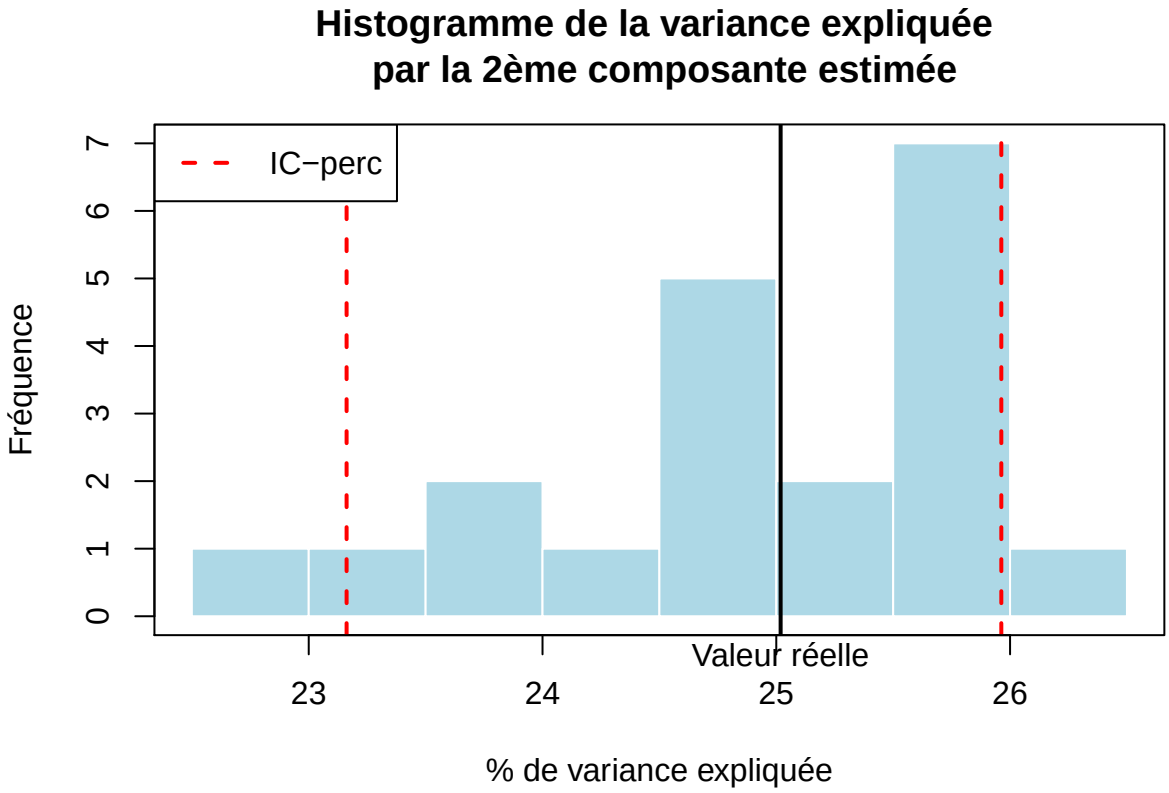
3. Implémenter une procédure bootstrap pour calculer les intervalles de confiance associés aux proportions de variance expliquées par les 10 premières composantes principales, par la méthode de votre choix (e.g., quantile, basique, ...). Proposer une représentation graphique de vos résultats.

```
## Loading required package: bootstrap
```



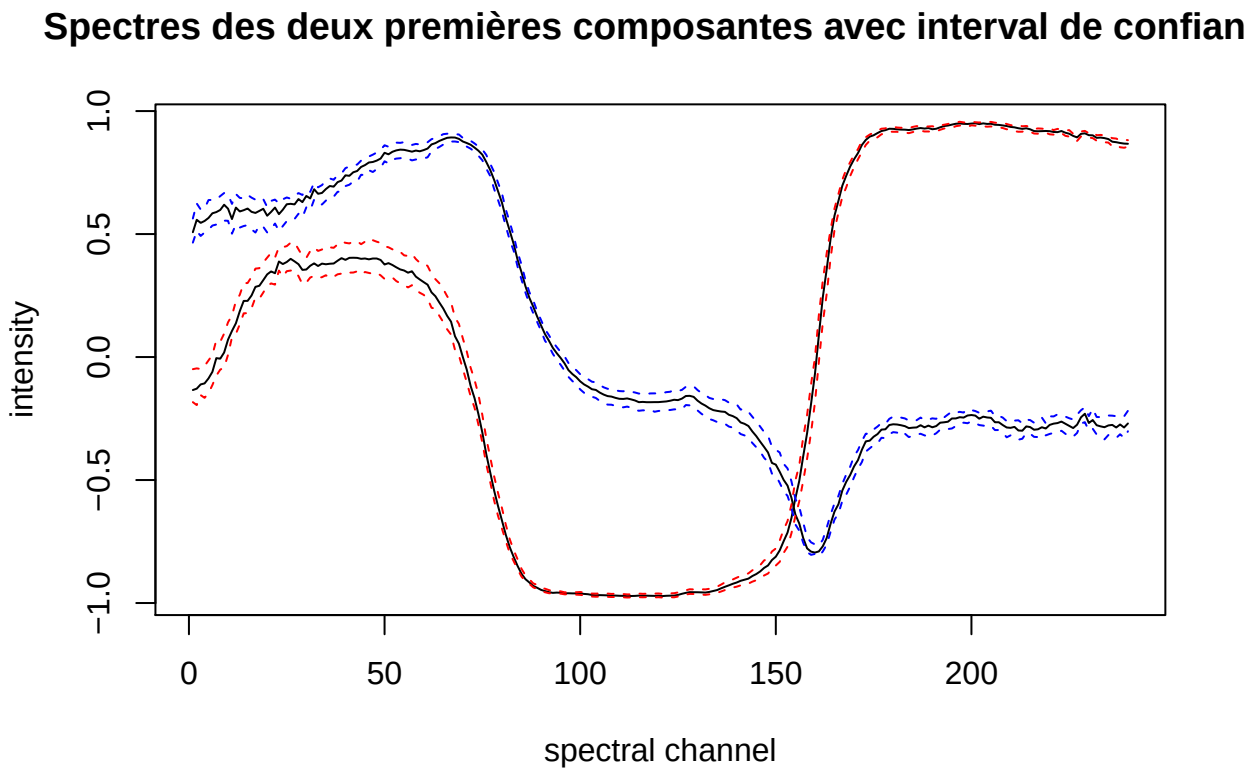
L'écart de l'interval de confiance réduit fortement pour les composantes expliquant le moins de variance.


Histogramme de la variance expliquée
par la 1ère composante estimée
IC−perc
Fréquence
0
1
2
3
4
5
6
7
Valeur réelle
55.0
55.5
56.0
56.5
57.0
57.5
58.0
% de variance expliquée



La distribution de l'estimation de la variance expliquée et la valeur réelle sont comprises entre les borne de l'interval de confiance.

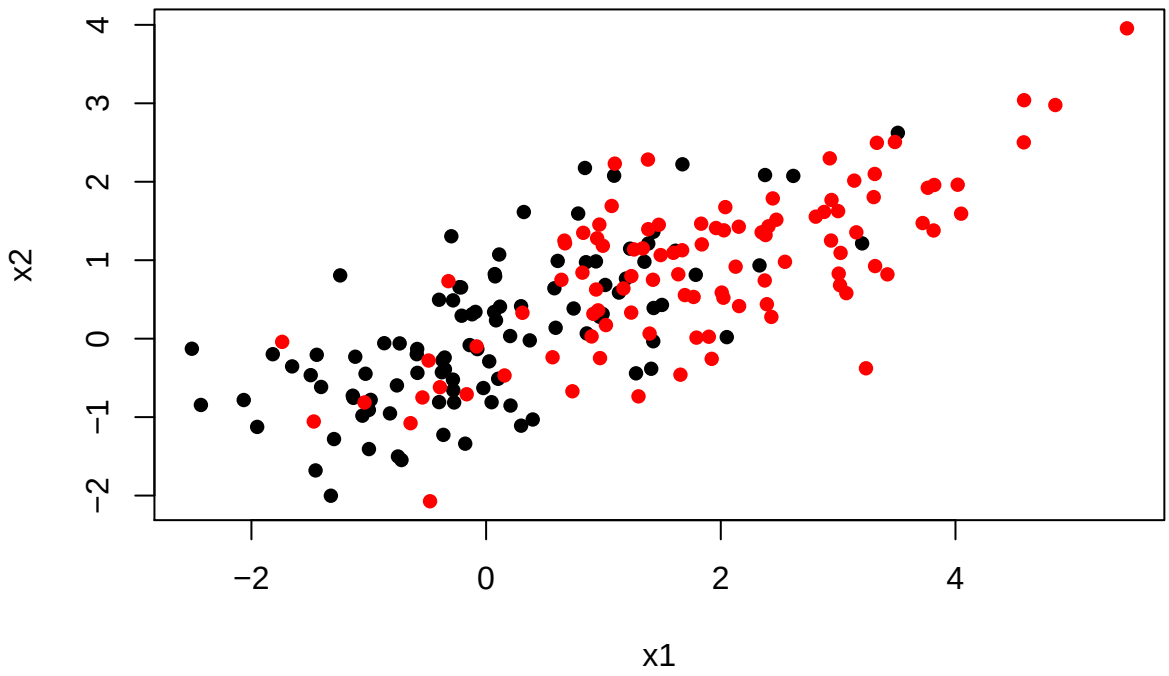
4. Enfin, modifier votre procédure pour pouvoir représenter la variabilité induite par le bootstrap sur les axes principaux de la question 2 et commenter les résultats obtenus.



L'intervalle de confiance encadre bien les spectres originaux.

Exercice 4 : tests par permutation

Charger le jeu de données permutation.Rdata. Il contient une matrice X de taille 200 × 2 représentée ci-dessous, les 100 premières lignes de la matrice correspondant aux points noirs, et les 100 dernières aux points rouges.



On souhaite tester si on est capable de détecter une différence entre ces deux populations, i.e., si on est capable de détecter une différence significative entre les deux nuages de points, stockés dans les 100 premières et les 100 dernières lignes de la matrice. On considère pour cela une statistique basée sur une information de voisinage :

$$T = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}(y_{n(x_i)} = y_i)$$

où $n(x_i) \in \{1, ..., i1, i+1, .., n\}$ est l'indice du plus proche voisin du point x_i, $y_i = 1$ si x_i appartient aux premier ensemble de points (les points noirs) et $y_i = 0$ sinon, et la fonction $\mathbf{1}(.)$ vaut 1 si son argument est vérifié et 0 sinon.

1. Que mesure cette statistique ?

Cette statistique donne un indicatuer de l'hétérogénité des deux groupes. D'un point de vur plus mathématique, elle mesure la proportion de points qui ont leur plus proche vois du même groupe qu'eux. Ainsi, si les deux groupes sont bien hétérogènes, cette statistique sera proche de 1 et si au contraire, les deux groupes sont mélangés, cette statistique se rapprochera de 0.5. Et dans le cas extrême où les groupes sont indisociables (mais cas bizarre pour de l'aléatoire) car chaque voisin le plus proche appartient à l'autre groupe, cette statistique prendrait la valeur 0.

Posons tout de suite les hypothèses de notre problème : - H0 : Les groupes sont indisociables - H1 : Les groupes sont disociables

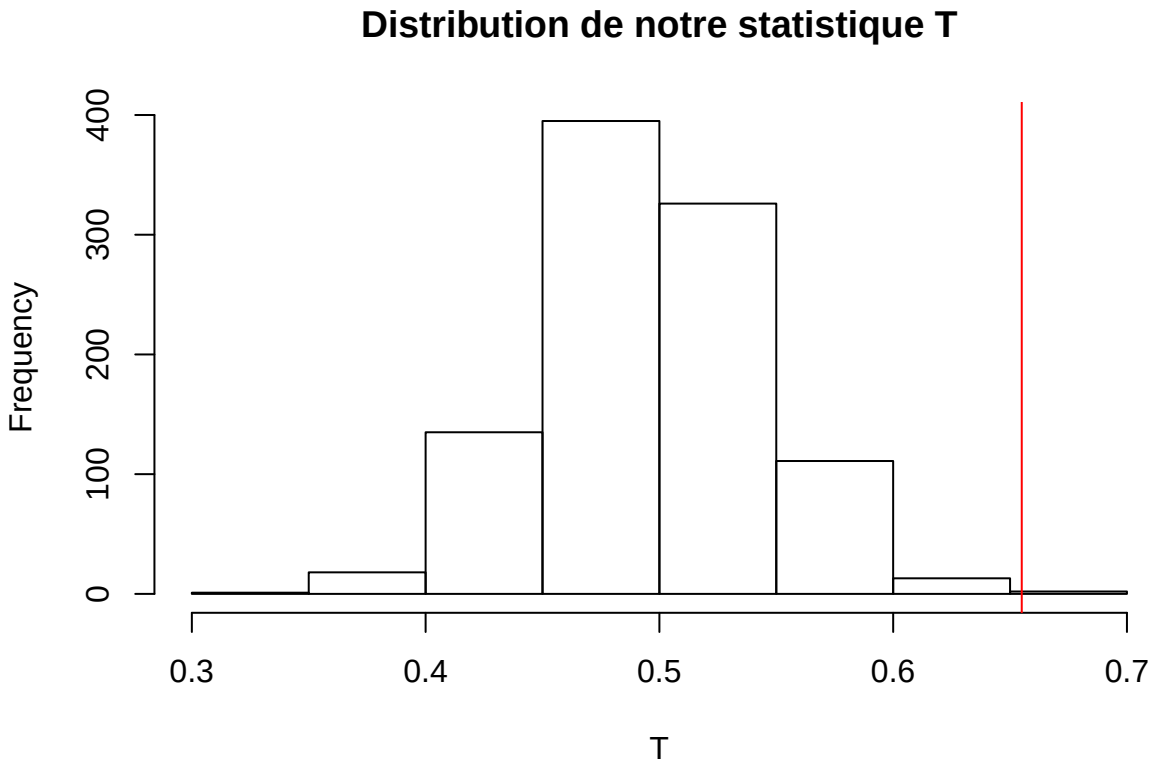
Nous sommes ici dans un cas unilatéral car même si T = 0 serait un cas très étrange pour de l'aléatoire et à creuser, ça ne permet pas en l'état de discriminer les groupes. Nous rejetterons donc l'hypothèse nulle si

T est suffisament supérieur à 0.5

Ici, nous avons $T = 0.66$

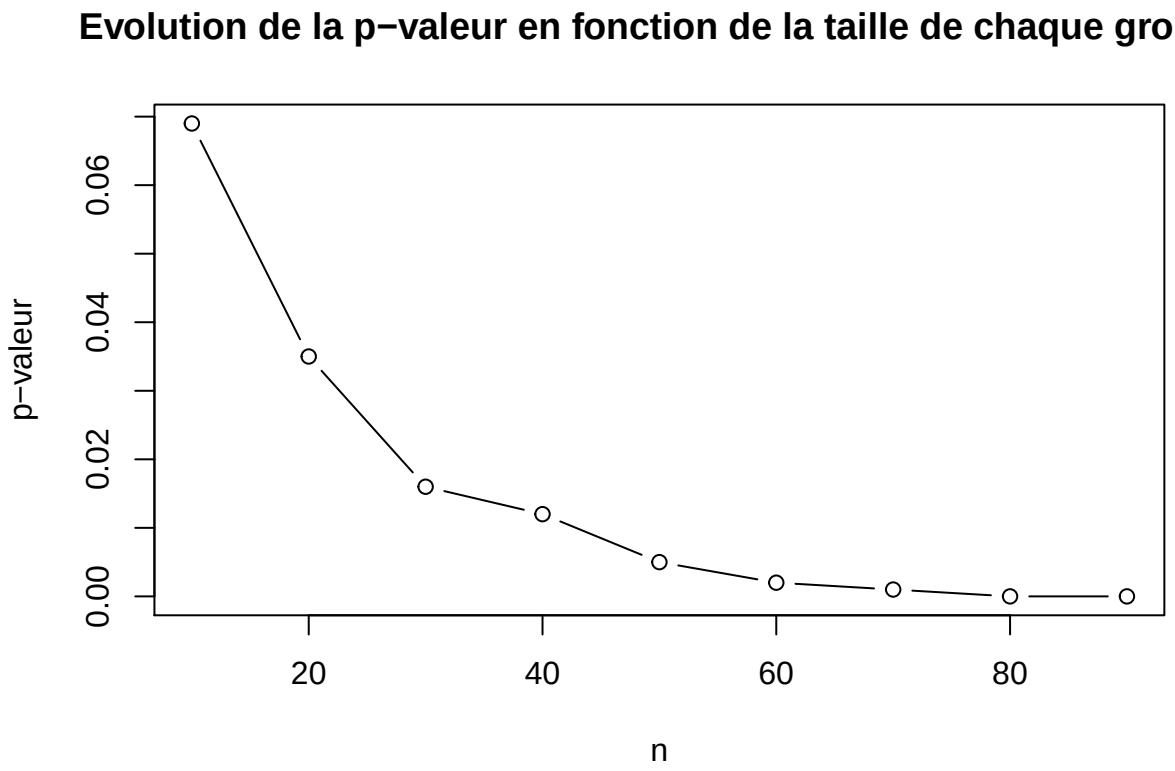
2. Appliquer une procédure par permutation pour B = 1000 tirages et évaluer la p-valeur obtenue. La différence entre les deux nuages de points est-elle significative ?

Nous allons maintenant permuter aléatoirement les positions de nos points (leur position dans le jeu de données, pas leurs coordonnées. C'est à dire que la ligne du 4ème point pourra passer à la 142ème ligne par exemple.). En rappelant que les 100 premiers points correspondent à un groupe et les 100 derniers à l'autre groupe. Cette permutation aura pour effet de changer les groupes de certains points. L'opération étant totalement aléatoire et répété un grand nombre (1000) de fois, nous aurons des valeurs de T qui (sauf si on a vraiment pas de chance mais c'est très improbable) seront distribuées sous la loi de l'hypothèse nulle (Il n'y a pas de différence entre les deux groupes.).



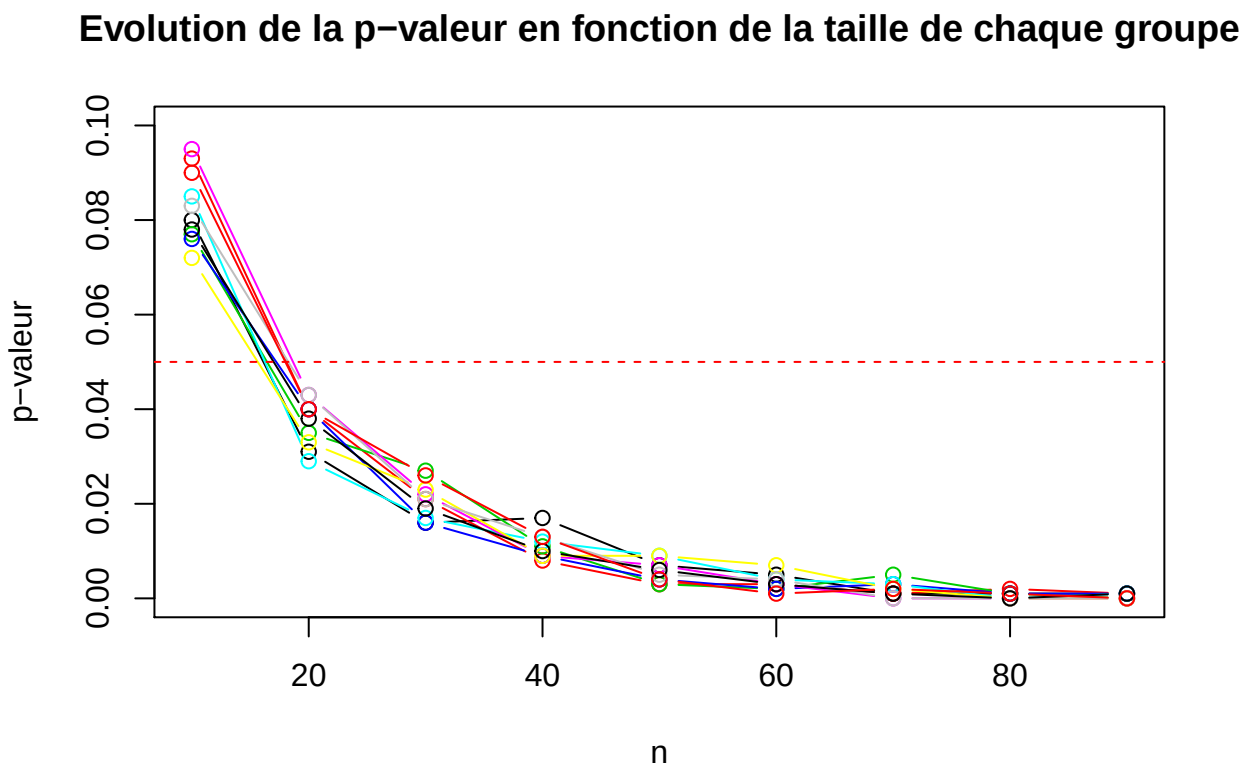
L'histogramme ci-dessus représente la distribution de nos statistiques T obtenues lors de nos 1000 tirages aléatoires. La ligne rouge représente notre statistique T observée sur le vrai jeu de données. Nous voyons que nos T sont centrées autour de 0.5, ce qui est logique puisque dans le cas aléatoire, nous devrions avoir T = 0.5 (Autant de chance que le plus proche voisin soit de la classe 1 que de la classe 2, donc de la même classe que soi.). Rien qu'en regardant l'histogramme, nous voyons que notre T observée est bien supérieure que dans le cas aléatoire. Et que donc, les groupes sont identifiables. La p-valeur de 0 vient confirmer cette impression car inférieure à 0.05. La différence entre les deux nuages est donc bien significative.

3. Reproduire cette analyse en tirant aléatoirement un ensemble de n = 10, 20, ..., 90 points parmi chaque population. Comment la p-valeur évolue t'elle ? Représenter les résultats sous la forme d'un graphique.



On observe que que la p-valeur diminue dans un premier temps lorsque la taille de chaque échantllon augmente. Puis se stabiise lorsque n est assez grand. Dans notre cas, la p-valeur semble se stabiliser à partir de $n = 50$.

4. Enfin, reproduire cette seconde analyse 10 fois et représenter la variabilité dans les p-valeurs obtenues en fonction du nombre de points considérés. A partir de quelle taille d'échantillon est-on capable de détecter une différence significative en considérant que la médiane des p-valeurs obtenues sur les 10 répétitions doit être (et rester) inférieure à 0.05 ?



Plus n est grand, plus la variabilité des p-valeurs pour un n fixé diminue. On constate qu'à partir de $n = 20$ (dans chaque groupe), nous pouvons détécter une différence significative entre les deux groupes. Si on ne se base que sur la médiane des p-valeurs, on peut réduire encore un peu l'échantillon (de peu), mais vu la longueur des calculs, nous n'avons testé les n qu'avec un pas de 10.