Loi des grands nombres et tendance vers la loi normale

Semaine 3

Table des matières

Objectifs

Les objectifs de cette séance sont de vérifier certains résultats asymptotiques du calcul des probabilités : loi des grands nombres, tendance vers la loi normale. On mettra ces résultats en oeuvre pour estimer une grandeur caractéristique (π) par la méthode de Monte-Carlo afin de préparer le TP 2.

Loi des grands nombres

Exercice 1

On considère un échantillon i.i.d. de n variables aléatoires $X_1, \ldots, X_n$ de même loi. On suppose que $\mathrm{Var}(X_1) < \infty$. On définit la moyenne empirique de la manière suivante

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i\,.$$

Question 0

Calculer l'espérance et la variance de la variable aléatoire $\overline{X}_n$.

Solution. La solution peut être rédigée sur une feuille libre. On ne reportera que le résultat.

Espérance.

Par linéarité de l'espérance :

$$\mathbb{E}[\overline{X}_n] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[X_i] = \frac{1}{n}\cdot n\mu = \mu = \mathbb{E}(X_1)$$

Variance

Comme les X_i sont i.i.d. et donc indépendantes :

$$\mathrm{Var}(\overline{X}_n) = \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}(X_i) = \frac{1}{n^2}\cdot n\mathrm{Var}(X_1) = \frac{\mathrm{Var}(X_1)}{n}.$$

Question 1

On pose $m = \mathbb{E}[X_1]$. Soit $\epsilon > 0$. En utilisant l'inégalité de Chebyshev, déterminer un majorant pour la probabilité $\mathrm{P}(|\overline{X}_n - m| > \epsilon)$.

Solution. La solution peut être rédigée sur une feuille de papier libre. On ne reportera que le résultat.

Raisonnement

Soit $\varepsilon > 0$. Par l'inégalité de Chebyshev, pour toute variable aléatoire Y de variance finie,

$$\mathbb{P}(|Y - \mathbb{E}[Y]| > \varepsilon) \leq \frac{\mathrm{Var}(Y)}{\varepsilon^2}.$$

En appliquant cela à $Y = \overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ et en utilisant $\mathrm{Var}(\overline{X}_n) = \sigma^2/n$, on obtient

$$\mathbb{P}\big(|\overline{X}_n - m| > \varepsilon\big) \leq \frac{\mathrm{Var}(\overline{X}_n)}{\varepsilon^2} = \frac{\mathrm{Var}(X_1)}{n\,\varepsilon^2}.$$

Question 2

En déduire une preuve de l'énoncé de la loi faible des grands nombres.

Solution. La solution peut être rédigée sur une feuille de papier libre. On ne reportera que les arguments principaux.

Raisonnement

On a établi que

$$0 \leq \mathbb{P}\big(|\overline{X}_n - m| > \varepsilon\big) \leq \frac{\mathrm{Var}(\overline{X}_n)}{\varepsilon^2} = \frac{\mathrm{Var}(X_1)}{n\,\varepsilon^2}.$$

Or $\dfrac{\mathrm{Var}(X_1)}{n\varepsilon^2} \to 0$ lorsque $n \to \infty$.

Par le théorème des gendarmes , on en déduit que

$$\lim_{n\to\infty} \mathbb{P}\big(|\overline{X}_n - m| > \varepsilon\big) = 0.$$

Question 3

On suppose que X_i est de loi uniforme sur $(0,1)$ (loi $\mathcal{U}(0,1)$ pour tout i. Utiliser l'inégalité de Chebyshev afin de proposer une valeur n (la plus petite possible) telle que

$$\mathbb{P}(|\overline{X}_n - \frac{1}{2}| > 0.1) < 0.05\,.$$

Solution. La solution peut être rédigée sur une feuille de papier libre. On ne reportera que le résultat.

Raisonnement

Si $X \sim \mathcal{U}(0,1)$, alors

$$\mathbb{E}[X] = \frac{1}{2}, \qquad \mathrm{Var}(X) = \frac{1}{12}.$$

Pour la moyenne empirique $\overline{X}_n$ on a, par Chebyshev,

$$\mathbb{P}\Big(|\overline{X}_n - \tfrac{1}{2}| > 0.1\Big) \leq \frac{\mathrm{Var}(\overline{X}_n)}{(0.1)^2} = \frac{\frac{1}{12}}{n \cdot \frac{1}{100}} = \frac{100}{12\,n} = \frac{25}{3n}.$$

Donc $\mathbb{P}(|\overline{X}_n - \frac{1}{2}| > 0.1) < 0.05$ si $\frac{25}{3n} < 0.05$

Donc n=167

On trouve donc que le plus petit n pouvant assurer l'inégalité demandé est $n = 167$

Question 4

Simuler 100000 fois la moyenne empirique d'un échantillon de loi uniforme de taille $n = 167$ à l'aide du code suivant.

Solution. Code à commenter et à compléter.

```
#on simule n_simulations fois la variable moyenne de n_samples variables aleatoires
#independantes suivant la loi U(0,1)
  rmoy_unif <- function(n_simulations, n_sample){
    replicate(n_simulations, mean(runif(n_sample)))
    }

# moyennes empiriques d'échantillons U(0,1)
  x_bar <-  rmoy_unif(100000, n_sample = 167)
```

Question 5

Vérifier de manière empirique que la probabilité de l'encadrement $(0.4 < \overline{X_n} < 0.6)$ est supérieur à 95% :

$$\mathbb{P}(0.4 < \overline{X_n} < 0.6) > 0.95\,.$$

La valeur n vous paraît-elle être la plus petite possible ? Quelle est cette valeur dans la simulation ?

Solution. Code à commenter et à compléter.

```
# P(0.4<x_bar<0.6)=P(-0.1<x_bar<0.1)=P(|x_bar-1/2|<0.1)
#On fait alors la moyenne de cette probabilité et on test si elle est bien supérieure
#à 0.95
# on calcule la proba que |X_bar(n_samples) - 0.5| < 0.1
mean(abs(x_bar - 0.5) < 0.1 ) > 0.95
```

```
## [1] TRUE
```

```
# Tester d'autres valeurs n = 10, 20, 30, 40, etc
  x_bar1 <-  rmoy_unif(10000, n_sample = 10)
  x_bar2 <-  rmoy_unif(10000, n_sample = 30)
  x_bar3 <-  rmoy_unif(10000, n_sample = 32)
  x_bar4 <-  rmoy_unif(10000, n_sample = 33)
  x_bar5 <-  rmoy_unif(10000, n_sample = 40)
# Evaluer si la condition est vérifiée
mean(abs(x_bar1 - 0.5) < 0.1 ) > 0.95
```

```
## [1] FALSE
```

```
mean(abs(x_bar2 - 0.5) < 0.1 ) > 0.95
```

```
## [1] FALSE
```

```r
mean(abs(x_bar3 - 0.5) < 0.1 ) > 0.95 #parfois affiche false
```

```
## [1] FALSE
```

```r
mean(abs(x_bar4 - 0.5) < 0.1 ) > 0.95
```

```
## [1] TRUE
```

```r
mean(abs(x_bar5 - 0.5) < 0.1 ) > 0.95
```

```
## [1] TRUE
```

Raisonnement

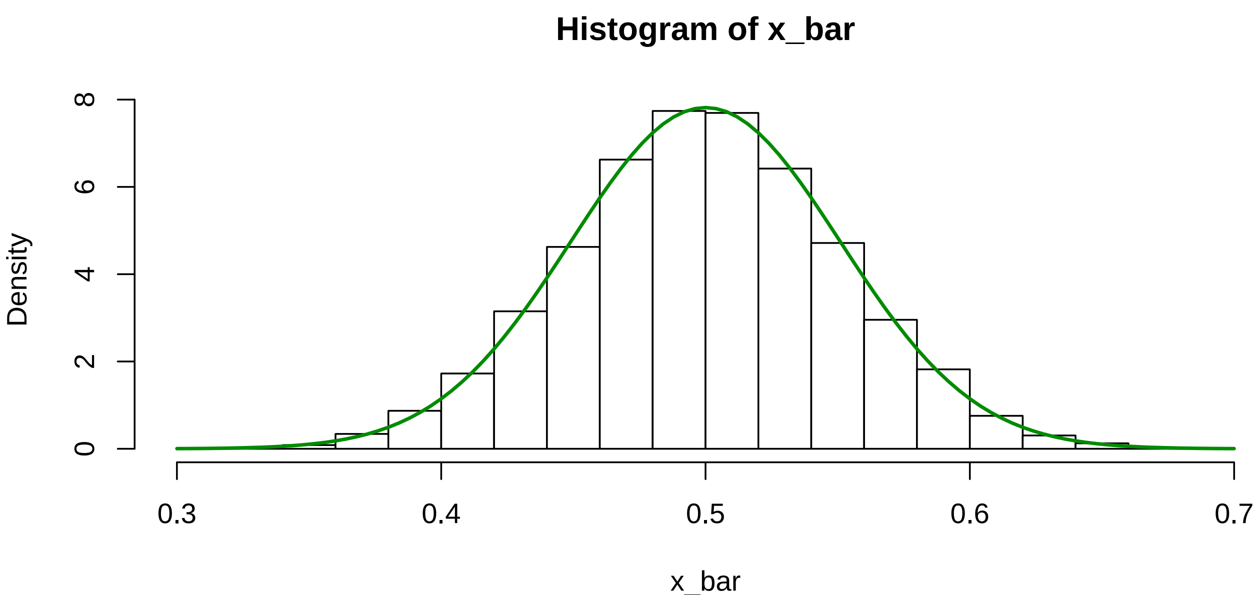
On remarque que pour $n_{sample} < 33$ la condition n'est pas vérifiée. Empiriquement la plus petite valeur de n vérifiant la condition est 33 donc 167 n'est pas la plus petite possible.

Question 6

On réplique 10000 fois l'expérience précédente pour des échantillons de taille $n = 32$. Commenter le code suivant et le résultat produit (expliquer d'où provient la valeur 0.05103103).

Solution. Code à commenter et à compléter.

```r
# 10 000 moyennes empiriques de 32 variables aléatoires suivant la loi U(0,1)
 x_bar <-  rmoy_unif(10000, n_sample = 32)
 #On affiche la répartition de la variable moyenne (X bar)
 hist(x_bar, col="white", prob = TRUE)

#Affichage de la courbe de la densité d'une loi normale: d'après le théorème centrale,
#x_bar suit approximativement une loi normale d'espérance 1/2 et de variance 1/(12n)
#lorsque n est grand
#sd est l'écart type de xbar: sqrt(1/(32*12))
 curve(dnorm(x, m = 0.5, sd = 0.05103103),
       from = 0.3, to = 0.7, lwd = 2,
       add = TRUE, col = "green4")
```




```r
 #on a superposé les moyennes empiriques de x_bar avec la courbe de densité d'une
 #loi de normale. L'histogramme peut s'interpréter comme une estimation de la densité.
```

Méthode de Monte-Carlo

Exercice 2 - Estimation de π par tirage aléatoire.

Soient (U_n) et (V_n) deux suites de variables indépendantes de loi uniforme $\mathcal{U}(0,1)$. On suppose que U_n et V_n sont indépendantes pour tout n.

On pose

$$X_n = \begin{cases} 1 & \text{si } U_n^2 + V_n^2 < 1 \\ 0 & \text{sinon.} \end{cases}$$

Pour tout $n \geq 1$, on pose $Y_n = 4\overline{X_n}$.

Question 1

Déterminer la loi de variable X_n.

Indication (admis) : Calculer $\mathbb{P}(X_n = 1)$ revient à calculer l'aire correspondant à la condition $(U_n^2 + V_n^2 < 1)$, car les tirages sont de loi uniforme dans le carré unité.

Solution. On ne reportera que le résultat.

L'ensemble des points correspondant à la condition $(U_n^2 + V_n^2 < 1)$ sont dans un quart de cercle unité, donc l'aire correspondante vaut $\pi/4$

On a : $\mathbb{P}(X_n = 1) = \frac{\pi}{4}$ et $\mathbb{P}(X_n = 0) = \frac{3\pi}{4}$

De plus, X_n prend comme valeurs 0 et 1 avec la probabilité d'un succès égale à $\pi/4$. Donc X_n suit une loi bernouilli de paramètre $\pi/4$

Question 2

Calculer l'espérance et la variance de Y_n. En déduire que la suite (Y_n) converge vers π.

Solution. La solution peut être rédigée sur une feuille libre. On ne reportera que le résultat.

$\mathbb{E}(Y_n) = \pi$ et $\mathbb{V}(Y_n) = \frac{\pi(4-\pi)}{n}$

En utilisant l'inégalité de Bienyamé- Tchebychev et le fait que la variance de Y_n tend vers 0, on en déduit que:

$$Y_n \underset{n \to +\infty}{\longrightarrow} \pi$$

Question 3

Vérifier à l'aide de la simulation d'un échantillon de taille $n = 10000$ que la valeur empirique Y_n est proche de π. À combien de décimale près peut-on prévoir que l'estimation est précise ?

Indication : on peut augmenter n pour répondre à cette question.

Solution. Commenter les codes ci-dessous.

```r
# On implémante la taille de l'échantillon
  n = 10000

# on définit les 2 VAR de loi uniforme sur [0,1]
  u <- runif(n)
  v <- runif(n)

#  On définit X
  x <- (u^2 + v^2 < 1)
```

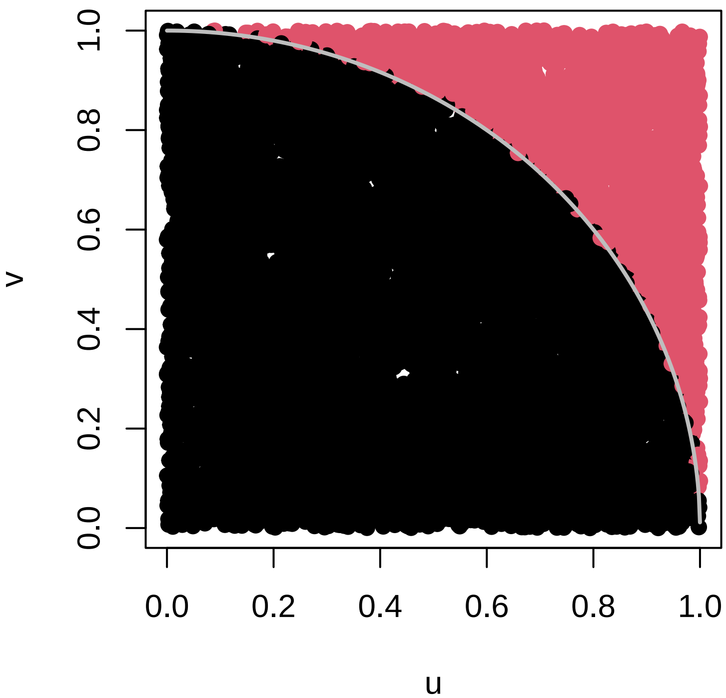
```r
# On défnit X_bar et par conséquent Y et on retourne la valeur de Y
  y = 4*mean(x)
  y
```

```
## [1] 3.128
```

```r
cat("En augmentant n on remarque qu'on peut prévoir une approximation précise que à 3 décimales près.")
```

```
## En augmentant n on remarque qu'on peut prévoir une approximation précise que à 3 décimales près.
```

```r
  par(mar=c(4,4,2,1))
  plot(u, v, col = 1 +  (u^2 + v^2 > 1), pch = 19)
  curve(sqrt(1-x^2), add=TRUE, col="grey", lwd=2, n=501)
```



Question 4

Soit $\alpha > 0$. À l'aide de l'inégalité de Chebyshev, chercher un entier n_0 tel que

$$\mathrm{P}(|Y_n - \pi| \leq \epsilon) \geq 1 - \alpha\,.$$

Solution. Démontrer sur le papier que la valeur suivante convient

$$n_0 = \frac{\pi(4-\pi)}{\alpha \times \epsilon^2}\,.$$

On veut montrer que :

$\mathbb{P}\left(|Y_n - \pi| \leq \varepsilon\right) \geq 1 - \alpha$

$\Leftrightarrow \quad \alpha \geq 1 - \mathbb{P}\left(|Y_n - \pi| \leq \varepsilon\right)$

$\Leftrightarrow \quad \mathbb{P}\left(|Y_n - \pi| > \varepsilon\right) \leq \alpha$

Or, d'après l'inégalité de Bienaymé–Tchebychev, on a :

$\mathbb{P}\left(|Y_n - \pi| > \varepsilon\right) \leq \frac{\mathrm{Var}(Y_n)}{\varepsilon^2} = \frac{\pi(4-\pi)}{n\varepsilon^2}$

Ainsi, toute valeur de n telle que $\frac{\pi(4-\pi)}{n\epsilon š} < \alpha$ vérifiera la condition.

Donc:

$n_0 = \frac{\pi(1-\pi)}{\alpha\,\varepsilon^2}$ convient.

Question 5

Pour $\alpha = 0.05$ et pour $\epsilon = 10^{-3}$, vérifier à l'aide d'une expérience si la valeur trouvée satisfait $|Y_{n_0} - \pi| < \epsilon$.

Solution.

```
# On définit les constantes utilisées
alpha <- 0.05
epsilon = 1e-3
n0 <- (pi*(4-pi))/(alpha*(epsilon**2))

# On effectue une fois la simulation
 u <- runif(n0)
 v <- runif(n0)
 z <- 4*mean(u^2 + v^2 < 1)
 z
```

```
## [1] 3.141433
```

```
# verification
 abs(z-pi) < epsilon
```

```
## [1] TRUE
```

Question 6

On s'intéresse à la variable aléatoire

$$Z_n = \sqrt{n}(Y_n - \pi)/\sqrt{\pi(4-\pi)}\,.$$

Calculer la variance de Z_n. Converge-t-elle vers 0 ? On suppose $n = 100$. Est-ce que la loi de Z_n se rapproche d'une loi connue ?

$\mathbb{E}(Z_n) = 0$

$\mathbb{V}(Z_n) = 1$

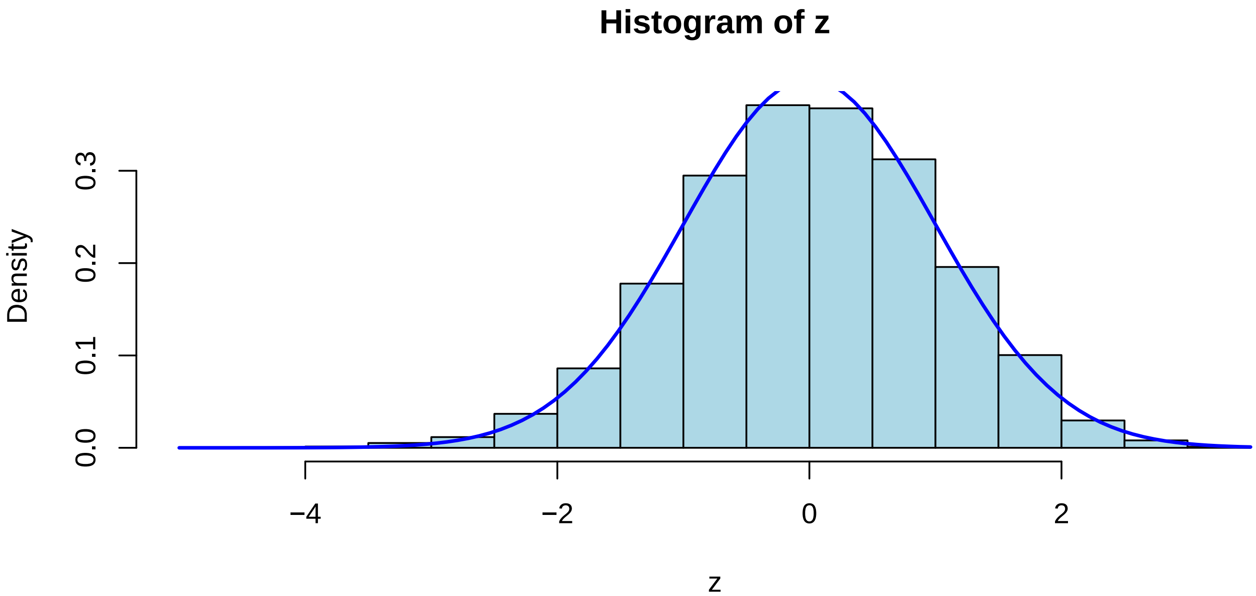
La variance ne converge pas vers 0.

D'après le théorème centrale limite, la loi de Z_n se rapproche d'une loi normale $\mathcal{N}$ (0,1)

Solution. Après avoir calculé la variance, commenter et compléter le code suivant.

```
# On effectue nsimu fois la simulation sur un échantillon de taille nsample
ry <- function(nsimu, nsample)
{
  ygen <- function()
  {
    u <- runif(nsample)
    v <- runif(nsample)
    4*mean(u^2 + v^2 < 1)
  }
  # On utilise replicate pour avoir une matrice pour toutes les simulations des échantillons.
```

```r
  return(replicate(nsimu, ygen()))
}

# On définit la VA Z
z <- (ry(10000, 100) - pi)*sqrt(100/(pi*(4-pi)))

# On trace un histographe des données empiriques et on superpose la fonction de densité de
#la loi normale N(0,1) par dessus.
# help(dnorm)
hist(z, prob = TRUE, col = "lightblue")
curve(dnorm(x), add=TRUE, col="blue", lwd=2)
```



Tendance vers la loi Normale

Exercice 3. Convergence vers la loi normale

Question 1

Soit $m \in \mathbb{R}$ et $\sigma \in \mathbb{R}_+$. Soit X une variable aléatoire de loi $N(m, \sigma^2)$. Montrer que la variable aléatoire

$$Z = \frac{X - m}{\sigma}$$

suit la loi normale $N(0,1)$.

Solution. La solution peut être rédigée sur une feuille de papier libre. On reportera les arguments principaux.

Z est à valeurs dans $\mathbb{R}$ et est une variable aléatoire continue.

La fonction de répartition de Z est celle d'une loi normale N(0,1).

En effet, $F_Z(z) = F_X(m + \sigma z)$ puis en dérivant on obtient la fonction de densité de Z en utilisant le théorème fondamental de l'analyse.

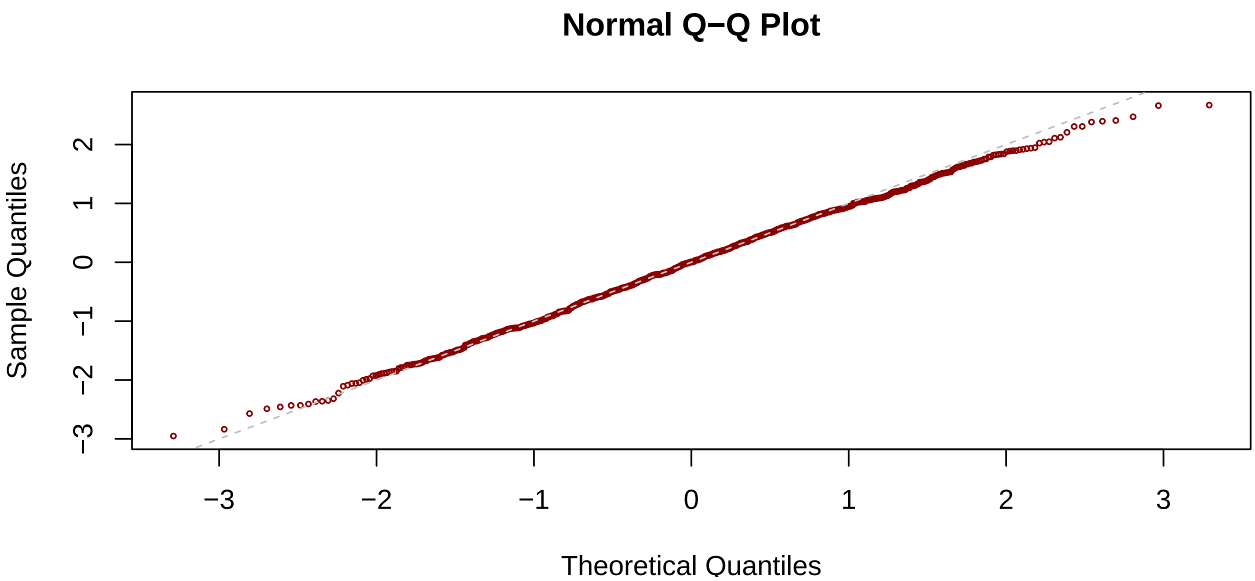
On peut alors identifier la fonction de densité de Z comme celle d'une loi normale N(0,1)

Question 2

À l'aide de simulations, vérifier graphiquement que la transformation obéit à la loi attendue dans la question précédente.

Solution. Compléter le code suivant utilisant un graphe quantile-quantile. Pour une probabilité donnée, un quantile correspond à l'inverse de la fonction de répartition. Comparer les quantiles de deux lois revient à comparer leurs fonctions de répartition. La fonction `qqnorm` compare les quantiles empiriques d'un échantillon aux quantiles de la loi normale.

```
n_rep = 1000
m = 10
sigma = 2
# simulation de la loi normale
x = rnorm(n_rep, m = m, sd = sigma)

# definition de l'échantillon z
z = (x-m)/sigma

# adequation graphique des quantiles empiriques et theoriques N(0,1)
qqnorm(z, cex = .4, col = "red4")
#tracer la droite y=x pour vérifier l'adéquation graphique.
abline(0,1, col = "grey", lty = 2)
```



Question 3

Soit $X_1, \dots, X_n$ des variables aléatoires indépendantes de loi normale $N(m, \sigma^2)$. Déterminer la loi de la variable

$$Z_n = \sqrt{n}\frac{\overline{X}_n - m}{\sigma}\,.$$

Vérifier le résultat à l'aide de 1000 répétitions de la simulation d'un échantillon contenant $n = 20$ variables de moyenne $m = 10$ et d'écart-type $\sigma = 2$.

Solution. On rappelle que la somme de variables aléatoires indépendantes de loi normale suit une loi normale.

$E(Z_n) = 0$

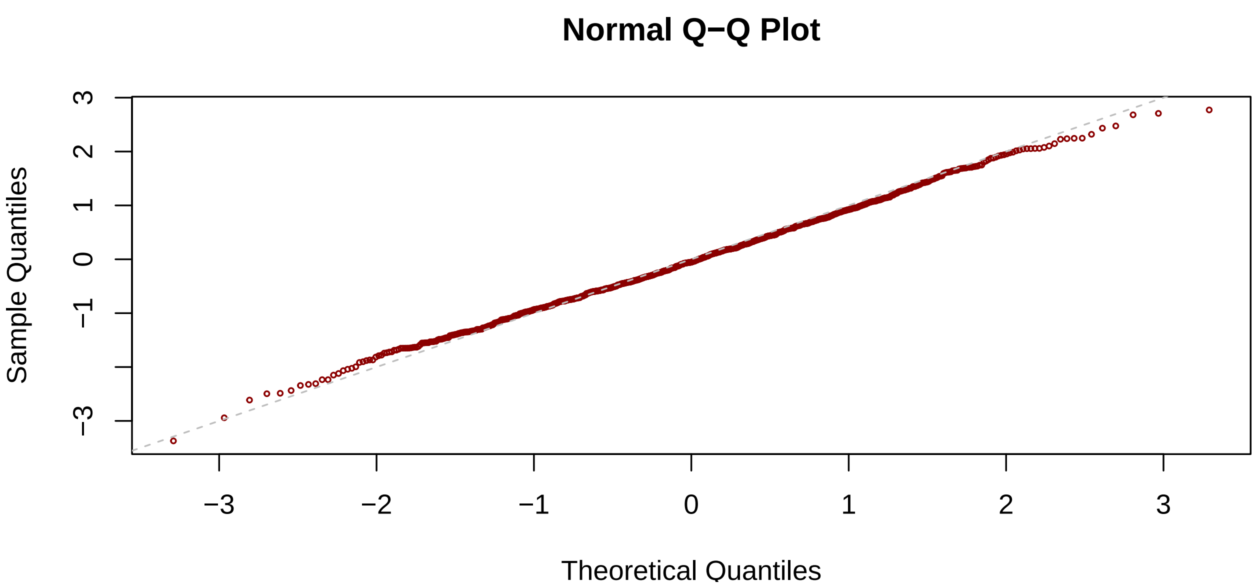
$V(Z_n) = 1$

Puisque Z_n est une combianaison linéaire de variables aléatoires indépendantes toutes de loi normale, alors Z_n est aussi de loi normale et on obtient ses paramètres grâce au calcul de l'espérance et de la variance.

Ainsi, Z_n est de loi normale N(0,1)

La vérification expérimentale est la suivante (à compléter).

```
n_rep = 1000
n = 20
m = 10
sigma = 2


# Création d'une matrice de dimensions 1000 lignes, 20 colonnes
# X est une matrice contenant 10000 échantillons de taille 20 de la loi N(m,sigma^2).

X = matrix(rnorm(n*n_rep, m = m, sd = sigma), nrow = n_rep)

# Créer le vecteur z

z = (apply(X, MARGIN = 1, mean) - m)*sqrt(n)/sigma

# adequation graphique des quantiles N(0,1)
qqnorm(z, cex = .4, col = "red4")
abline(0,1, col = "grey", lty = 2)
```



Question 4

On observe l'échantillon **x** généré par le code suivant.

```
set.seed(12412)
n = 40
m = 10
sigma = 2

# simulation de la loi normale
```

```r
x = rnorm(n, m = m, sd = sigma)
```

On suppose que les paramètres de simulation m et σ sont inconnus (bien que visibles dans le code).

Proposer des estimations des paramètres m et σ, c'est à dire une manière d'estimer m et σ à partir de l'échantillon $\mathbf{x}$. Calculer ces valeurs pour les observations.

Solution.

```r
# estimateur m
mean(x)
```

```
## [1] 10.06391
```

```r
# estimateur sigma
sqrt(var(x))
```

```
## [1] 2.297215
```

```r
#on peut aussi utiliser sd(x)
```

Intervalle de confiance

Exercice 3 – suite. Convergence vers la loi normale

Question 5

À partir des observations $\mathbf{x}$, définir un intervalle $[I_{\text{inf}}(\mathbf{x}), I_{\text{sup}}(\mathbf{x})]$ tel que

$$\mathrm{P}(m \in [I_{\text{inf}}(\mathbf{x}), I_{\text{sup}}(\mathbf{x})]) = 0.95.$$

Cet intervalle est appelé *intervalle de confiance* au seuil 95% pour m. Calculer un intervalle de confiance approché pour la moyenne des observations simulées.

Solution. Soit

$$Z_n = \sqrt{n}\frac{\overline{X}_n - m}{\sigma}\,.$$

Pour trouver l'intervalle demandé, on part de l'égalité suivante issue du Théorème Central Limite

$$\mathrm{P}(\Phi^{-1}(0.05/2) \leq Z_n \leq \Phi^{-1}(1-0.05/2)) = 0.95\,.$$

où Φ^{-1} est la fonction quantile de la loi normale centrée réduite (`qnorm`).

Le code à compléter est donné par

Méthode: On remplace Z_n par son expression dans l'égalité ci-dessous puis on isole m et on obtient que $P(\overline{X_n} - \frac{\sigma}{\sqrt{n}}\Phi^{-1}(0.975)/ < m < \overline{X_n} + \frac{\sigma}{\sqrt{n}}\Phi^{-1}(0.975)) = 0.95$.

```r
# calcul de l'intervalle
I_inf = mean(x) - sigma/sqrt(n)*qnorm(1-0.05/2)
I_sup = mean(x) + sigma/sqrt(n)*qnorm(1-0.05/2)

IC = c(I_inf, I_sup)
names(IC) = c("borne inf.", "borne sup.")
round(IC, 2)
```

```
## borne inf. borne sup.
##       9.44      10.68
```

Question 6

Répéter 10000 fois l'expérience de simulation précédente. Vérifier que le paramètre m se trouve dans l'intervalle proposé dans environ 95% des simulations.

Solution. Commenter et compléter le code suivant.

```
n = 40
m = 10
sigma = 2

# initialisation du test booléen
boo_ic = rep(0, 10000)

for (i in 1:10000){

 # simulation d'un échantillon N(m, sigma) de taille n
 x = rnorm(n,m=m,sd=sigma)

 # calcul de l'intervalle
I_inf = mean(x) - (sigma/sqrt(n))*qnorm(1-0.05/2)
I_sup = mean(x) + (sigma/sqrt(n))*qnorm(1-0.05/2)

 # test d'appartenance de m à l'intervalle
 boo = (I_inf < m) & (m < I_sup)

 boo_ic[i] = boo
}
```

La proportion des simulations où paramètre m se trouve dans l'intervalle proposé est

```
round(mean(boo_ic), 3)
```

```
## [1] 0.946
```

Analyse du résultat: cette proportion est très proche de 0.95, donc m se trouve bien dans l'intervalle proposé dans environ 95% des simulations.

Remarque. Que se passe-t-il lorsqu'on remplace la valeur de sigma par son estimation empirique `sd(x)` ?

On obtient presque le même résultat puisque $sd(x)$ est une estimation de σ.

Cependant, de manière plus précise on s'aperçoit que lorsqu'on remplace σ par l'estimation empirique $sd(x)$ en gardant qnorm(1-0.05/2), on ne prend pas en compte l'incertitude additionnelle due à cette approximation. Alors la proportion de fois où l'intervalle contient m descend légèrement sous 95%

```
n = 40
m = 10
sigma = sd(x)

# initialisation du test booléen
boo_ic = rep(0, 10000)

for (i in 1:10000){

 # simulation d'un échantillon N(m, sigma) de taille n
```

```r
 x = rnorm(n,m=m,sd=sigma)

 # calcul de l'intervalle
I_inf = mean(x) - (sigma/sqrt(n))*qnorm(1-0.05/2)
I_sup = mean(x) + (sigma/sqrt(n))*qnorm(1-0.05/2)

 # test d'appartenance de m à l'intervalle
 boo = (I_inf < m) & (m < I_sup)

 boo_ic[i] = boo
}
round(mean(boo_ic),3)
```

```
## [1] 0.951
```