Compte rendu Travaux pratiques en langage R

**Pour charger les données :**

data = *read.csv*("psychology.csv", *TRUE*)  
*print*(data)

Le code permet de lire un fichier CSV nommé “psychology.csv” qui contient des données sur la psychologie des individus. Le paramètre TRUE (ou header = TRUE) indique que la première ligne du fichier contient les noms des variables. C’est-à-dire que dans le fichier csv à la 1er ligne on remarque : IBE, IP et qu’il est identifié comme un header. La fonction print (data) affiche le contenu du fichier dans la console.

**Question 1 :**

# ---- Question 1 ----  
# Statistiques descriptives pour l'IBE  
summary\_IBE = *summary*(data$IBE)  
mean\_IBE = *mean*(data$IBE)  
median\_IBE = *median*(data$IBE)  
sd\_IBE = *sd*(data$IBE)  
IQR\_IBE = *IQR*(data$IBE)  
  
# Affichage des résultats  
*cat*("Statistiques descriptives pour l'IBE:\n")  
*cat*("Sommaire:", summary\_IBE, "\n")  
*cat*("Moyenne:", mean\_IBE, "\n")  
*cat*("Médiane:", median\_IBE, "\n")  
*cat*("Écart-type:", sd\_IBE, "\n")  
*cat*("IQR:", IQR\_IBE, "\n")  
  
# Statistiques descriptives pour l'IP  
summary\_IP = *summary*(data$IP)  
mean\_IP = *mean*(data$IP)  
median\_IP = *median*(data$IP)  
sd\_IP = *sd*(data$IP)  
IQR\_IP = *IQR*(data$IP)  
  
# Affichage des résultats  
*cat*("\nStatistiques descriptives pour l'IP:\n")  
*cat*("Sommaire:", summary\_IP, "\n")  
*cat*("Moyenne:", mean\_IP, "\n")  
*cat*("Médiane:", median\_IP, "\n")  
*cat*("Écart-type:", sd\_IP, "\n")  
*cat*("IQR:", IQR\_IP, "\n")

Cette partie du code permet de calculer et d’afficher les statistiques descriptives pour les variables IBE et IP. La fonction summary(data$IBE) renvoie les valeurs minimale, maximale, moyenne, médiane et les quartiles de la variable. Les fonctions mean(data$IBE), median(data$IBE), sd(data$IBE) et IQR(data$IBE) renvoient respectivement la moyenne, la médiane, l’écart-type et l’intervalle interquartile de la variable. La fonction cat() permet d’afficher les résultats dans la console avec des sauts de ligne (\n) et des séparateurs (:) de la même manière qu’un print.Pour IP on utilise la même chose ce qui renvoie en console :

Statistiques descriptives pour l'IBE:

Sommaire: 1 24.25 50 48.29333 74.5 100

Moyenne: 48.29333

Médiane: 50

Écart-type: 29.28626

IQR: 50.25

Statistiques descriptives pour l'IP:

Sommaire: 0 4.09425 4.6505 4.475667 5.264 6.092

Moyenne: 4.475667

Médiane: 4.6505

Écart-type: 1.201006

IQR: 1.16975

**Question 2 :**

# ---- Question 2 ----  
# Créer un histogramme pour la variable IBE  
*hist*(data$IBE, main="Estimation de densité pour IBE", xlab="IBE", col="lightblue", border="black", probability=*TRUE*)  
  
# Créer un histogramme pour la variable IP  
*hist*(data$IP, main="Estimation de densité pour IP", xlab="IP", col="lightgreen", border="black", probability=*TRUE*)  
  
# Ajouter une légende  
*legend*("topright", legend="Densité", col="blue", lwd=2)

Cette partie du code permet de créer des histogrammes pour les variables IBE et IP, afin de visualiser la distribution des données. La fonction hist() prend en paramètre la variable à représente (soit IBE ou IP), le titre du graphique (main), le nom de l’axe des abscisses (xlab), la couleur des barres (col), la couleur des bordures (border) et le type de fréquence (probability=TRUE pour représenter la densité de probabilité plutôt que le nombre d'observations). Ainsi on a graphiquement pour IBE et IP avec les couleurs appliquées :

Une image contenant diagramme, ligne, Plan, Tracé

Description générée automatiquement

Une image contenant diagramme, capture d’écran, Tracé, ligne

Description générée automatiquement

**Question 3 :**

# ---- Question 3 ----  
  
# Diviser par deux la largeur des intervalles pour IBE  
new\_breaks\_ibe = *seq*(*min*(data$IBE), *max*(data$IBE), length.out = (*length*(*seq*(*min*(data$IBE), *max*(data$IBE))) - 1) \* 2 + 1)  
  
# Créer un histogramme pour la variable IBE avec des intervalles divisés par deux  
*hist*(data$IBE, breaks = new\_breaks\_ibe, main="Estimation de densité pour IBE (Intervalles divisés par deux)", xlab="IBE", col="lightblue", border="black", probability=*TRUE*)  
  
# Diviser par deux la largeur des intervalles pour IP  
new\_breaks\_ip = *seq*(*min*(data$IP), *max*(data$IP), length.out = (*length*(*seq*(*min*(data$IP), *max*(data$IP))) - 1) \* 2 + 1)  
  
# Créer un histogramme pour la variable IP avec des intervalles divisés par deux  
*hist*(data$IP, breaks = new\_breaks\_ip, main="Estimation de densité pour IP (Intervalles divisés par deux)", xlab="IP", col="lightgreen", border="black", probability=*TRUE*)

Ce code permet de créer des histogrammes avec des intervalles plus fins pour les variables IBE et IP, afin de mieux apprécier la forme de la distribution. La fonction seq() permet de créer une séquence de valeurs régulièrement espacées entre le minimum et le maximum de la variable, en précisant le nombre de valeurs (length.out). Ce nombre est calculé en multipliant par deux le nombre d’intervalles de l’histogramme précédent, moins un. Pour cela, il utilise une formule mathématique : *(nombre de valeurs - 1) \* 2 + 1)* qui calcule le nouveau nombre de valeurs à générer. Cette formule est basée sur le nombre de valeurs de l’histogramme précédent, qui correspond au nombre d’intervalles plus un. La formule soustrait un à ce nombre, le multiplie par deux, et ajoute un. Ainsi, elle obtient le double du nombre d’intervalles de l’histogramme précédent, plus un. Ce nouveau nombre de valeurs permet de créer des intervalles plus fins, qui divisent chaque intervalle de l’histogramme précédent en deux.

La fonction hist() prend en paramètre le vecteur des intervalles (breaks) à utiliser pour créer l’histogramme. Les autres fonctions sont les mêmes que pour la question 2. Ce qui donne graphiquement : Une image contenant diagramme, ligne, Plan, texte

Description générée automatiquement

Une image contenant capture d’écran, texte, diagramme, ligne

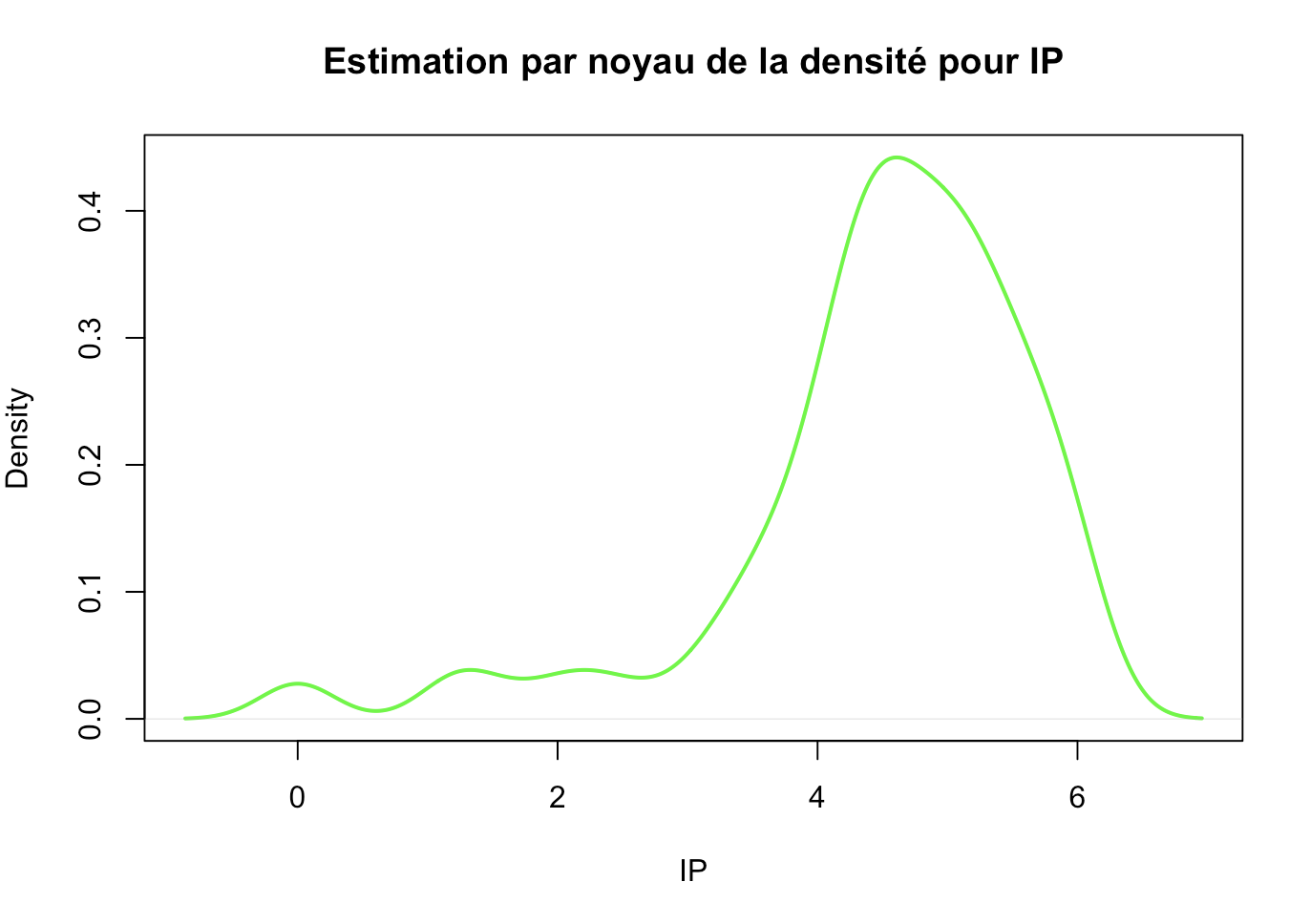
Description générée automatiquement

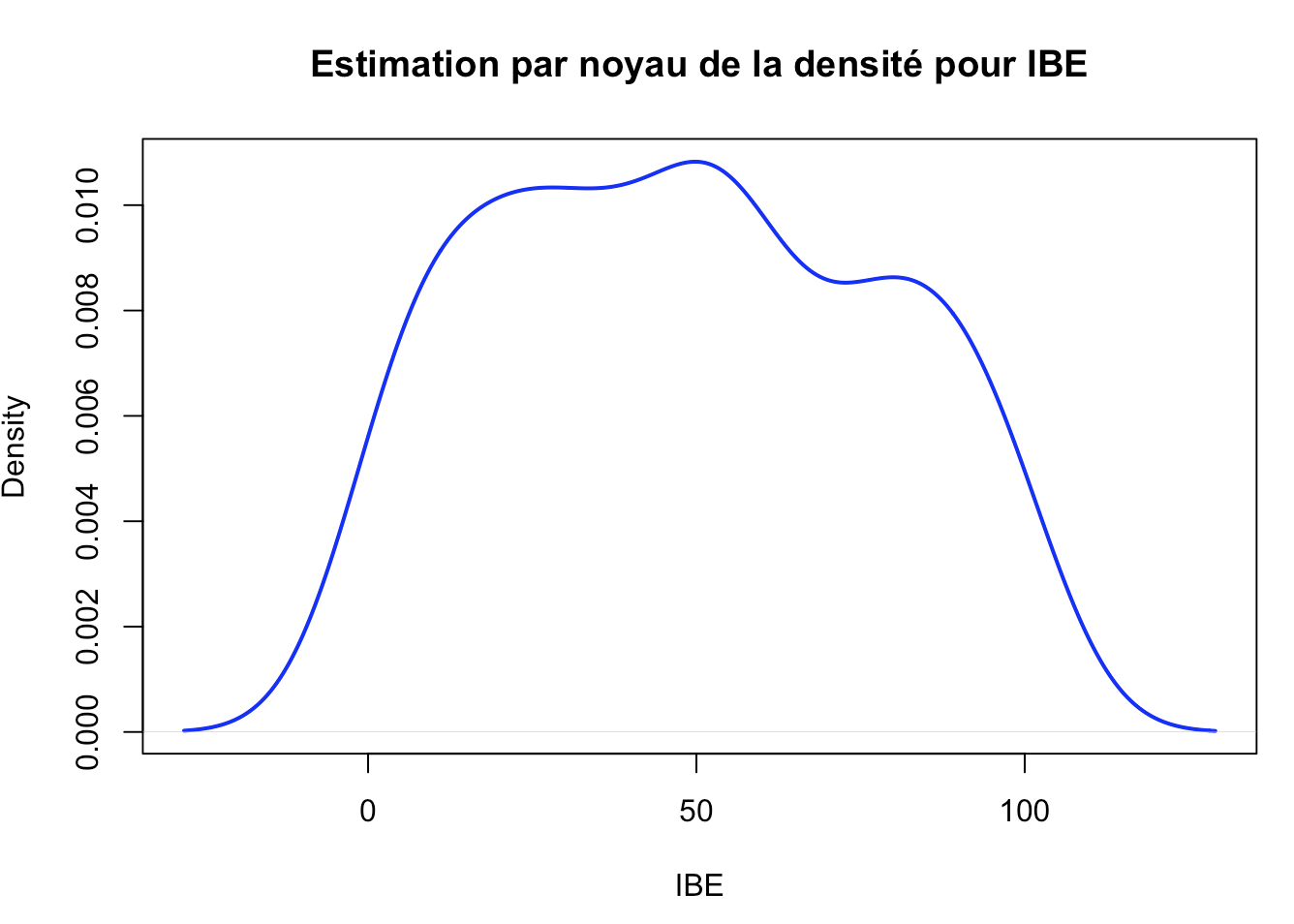
**Question 4 :**

# ---- Question 4 ----  
# Estimation par noyau de la densité pour IBE  
density\_ibe = *density*(data$IBE)  
  
# Tracer le KDE pour IBE  
*plot*(density\_ibe, main="Estimation par noyau de la densité pour IBE", xlab="IBE", col="blue", lwd=2)  
  
# Estimation par noyau de la densité pour IP  
density\_ip = *density*(data$IP)  
  
# Tracer le KDE pour IP  
*plot*(density\_ip, main="Estimation par noyau de la densité pour IP", xlab="IP", col="green", lwd=2)

Cette partie du code permet de tracer les courbes de densité estimées par noyau (KDE) pour les variables IBE et IP, sans utiliser d’histogramme. Cette méthode (KDE) sert à dessiner une ligne qui suit la forme des données. Elle donne plus d’importance aux données qui sont proches les unes des autres, et moins d’importance aux données qui sont loin les unes des autres. Ainsi, elle rend la ligne plus lisse.

La fonction density() permet de calculer la densité par noyau pour une variable. La fonction plot() permet de tracer la courbe de densité, en précisant le titre du graphique (main), le nom de l’axe des abscisses (xlab), la couleur (col) et l’épaisseur (lwd) de la ligne.

****

****

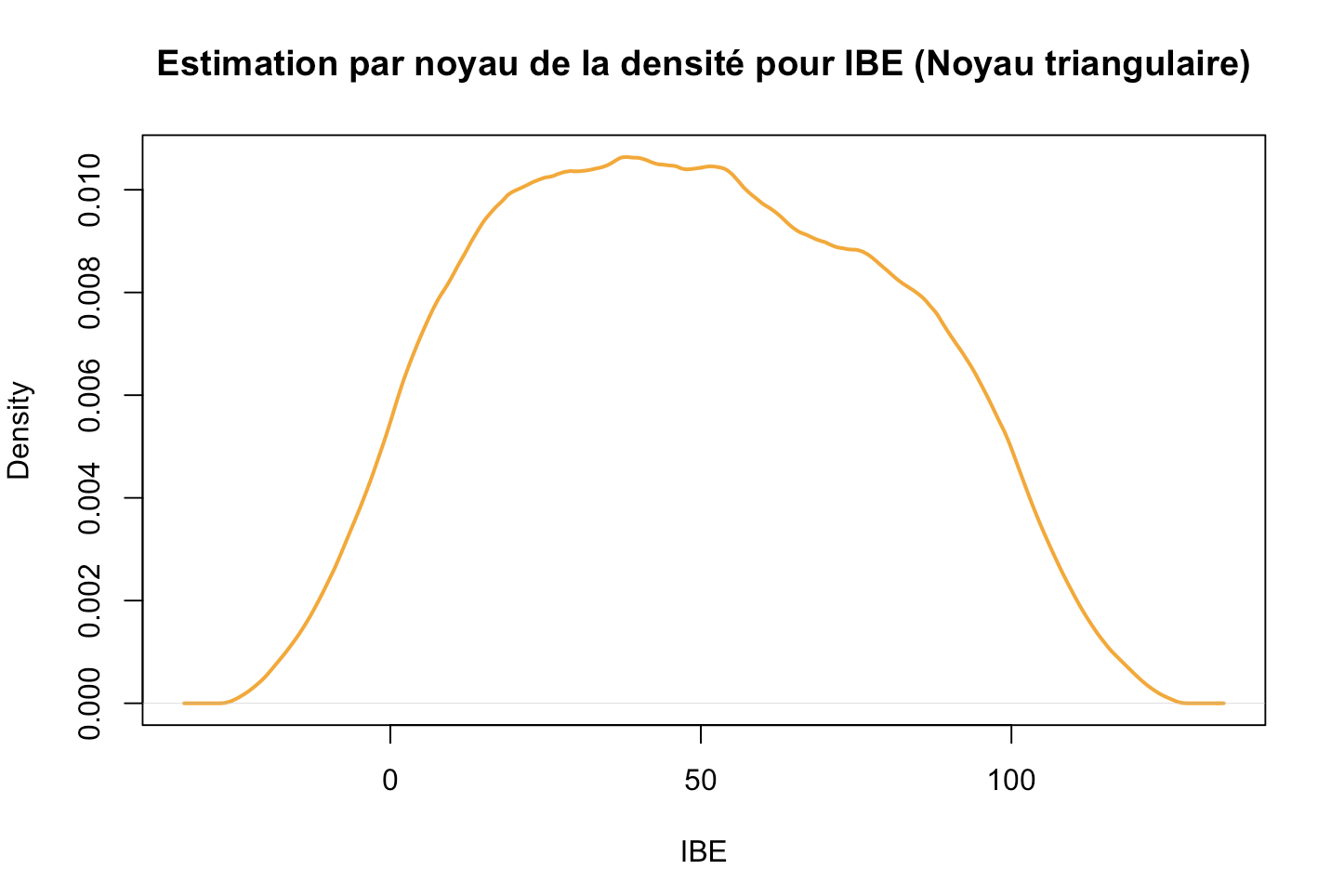
**Question 5 :**

# ---- Question 5 ----  
  
# Estimation par noyau de la densité pour IBE avec noyau triangulaire et validation croisée biaisée  
density\_ibe\_triangular = *density*(data$IBE, kernel = "triangular", bw = "nrd")  
  
# Tracer le KDE pour IBE avec noyau triangulaire  
*plot*(density\_ibe\_triangular, main="Estimation par noyau de la densité pour IBE (Noyau triangulaire)", xlab="IBE", col="orange", lwd=2)  
  
# Estimation par noyau de la densité pour IP avec noyau triangulaire et validation croisée biaisée  
density\_ip\_triangular = *density*(data$IP, kernel = "triangular", bw = "nrd")  
  
# Tracer le KDE pour IP avec noyau triangulaire  
*plot*(density\_ip\_triangular, main="Estimation par noyau de la densité pour IP (Noyau triangulaire)", xlab="IP", col="purple", lwd=2)

Cette partie du code permet d’estimer la densité par noyau (KDE) pour les variables IBE et IP, en utilisant un noyau triangulaire et une validation croisée biaisée.

L’utilisation d’un noyau triangulaire permet de dessiner une ligne qui suit la forme des données, en donnant plus ou moins d’importance aux données selon leur proximité. Il utilise une fonction qui ressemble à un triangle pour faire cela. On choisit le paramètre de lissage, en utilisant la validation croisée biaisée, indiqué dans la question. Celle-ci permet de vérifier si la méthode du noyau marche bien, en comparant les prédictions du modèle aux données réelles.

La fonction density() permet de spécifier le type de noyau à utiliser avec le paramètre kernel, ici triangulaire, et la largeur de bande (bw). La fonction plot() permet de tracer la courbe de densité, en précisant le titre du graphique (main), le nom de l’axe des abscisses (xlab), la couleur (col) et l’épaisseur (lwd) de la ligne. Le code est répété pour les deux variables IBE et IP, en changeant la couleur de la courbe.



Une image contenant texte, Tracé, diagramme, ligne

Description générée automatiquement

**Question 6 :**

# ---- Question 6 ----  
  
# Afficher le nuage de points entre les deux variables IBE et IP  
*plot*(data$IBE, data$IP, main="Nuage de points entre IBE et IP", xlab="IBE", ylab="IP", col=*c*("blue", "green"), pch=16)  
  
# Ajouter une régression linéaire pour illustrer la tendance  
*abline*(*lm*(IP ~ IBE, data=data), col="red", lwd=2)  
  
# Ajouter une légende  
*legend*("topright", legend=*c*("IBE", "IP", "Régression linéaire"), col=*c*("blue", "green", "red"), lwd=2)  
  
# Calculer et afficher la corrélation  
correlation\_coefficient = *cor*(data$IBE, data$IP)  
*cat*("Coefficient de corrélation entre IBE et IP:", correlation\_coefficient, "\n")

Cette partie du code permet d’afficher le nuage de points entre les deux variables IBE et IP, pour visualiser la relation entre elles. On a utilisé une régression linéaire pour visualiser la tendance générale entre les variables IBE et IP, en traçant une droite qui passe au plus près des points. Ainsi cela permet d’obtenir le coefficient de corrélation, qui mesure le degré de relation linéaire entre les deux variables.

La fonction plot() permet de tracer le nuage de points, en précisant les coordonnées des points (data$IBE et data$IP), le titre du graphique (main), le nom des axes (xlab et ylab), la couleur (col) et la forme (pch) des points. La fonction abline() permet d’ajouter une régression linéaire pour illustrer la tendance, en utilisant la fonction lm() pour ajuster le modèle linéaire (IP ~ IBE). La fonction legend() permet d’ajouter une légende, en précisant la position (topright), le texte (legend), la couleur (col) et l’épaisseur (lwd) des lignes. Les fonctions cor() et cat() permettent de calculer et d’afficher le coefficient de corrélation entre IBE et IP, ainsi qu’un commentaire sur la relation entre les variables. Ce qui renvoie en console :

Coefficient de corrélation entre IBE et IP: 0.7235905

Le coefficient de corrélation trouvé, permet de mesure le degré de relation linéaire entre les deux variables. Mais pour décrire la relation entre ces deux variables, il faut d’abord regarder le graphique et interpréter le coefficient de corrélation.

Une image contenant texte, ligne, diagramme, capture d’écran

Description générée automatiquement

Le nuage de points entre les deux variables IBE et IBP montre que la relation entre ces deux variables est positive, mais non linéaire. La droite de régression linéaire rouge passe au plus près des points, mais elle ne les touche pas tous. Le coefficient de corrélation est de 0.75 environ, ce qui indique une relation modérée entre les deux variables. Cela suggère que la relation n’est pas bien captée par une droite et qu’une transformation des données pourrait être nécessaire. C’est pourquoi, dans la question 7, on nous propose de faire une transformation des données permettant de linéariser la relation entre les deux variables.

**Question 7 :**

On utilise une transformation logarithmique des variables, qui permet de rendre la relation plus linéaire et plus adaptée à la régression linéaire.

# ---- Question 7 ----  
  
# Filtrer les valeurs positives avant la transformation logarithmique  
positive\_values = data$IP > 0  
transformed\_IBE = *log*(data$IBE[positive\_values])  
transformed\_IP = *log*(data$IP[positive\_values])  
  
# Afficher le nuage de points avec les données transformées  
*plot*(transformed\_IBE, transformed\_IP, main="Nuage de points entre IBE et IP (Transformées logarithmiques)", xlab="log(IBE)", ylab="log(IP)", col=*c*("purple", "orange"), pch=16)  
  
# Ajouter une régression linéaire pour illustrer la tendance  
*abline*(*lm*(transformed\_IP ~ transformed\_IBE), col="red", lwd=2)  
  
# Ajouter une légende avec position manuelle  
*legend*("topright", legend=*c*("log(IBE)", "log(IP)", "Régression linéaire"), col=*c*("purple", "orange", "red"), lwd=2)  
  
# Calculer et afficher la corrélation pour les données transformées  
correlation\_coefficient\_transformed = *cor*(transformed\_IBE, transformed\_IP)  
*cat*("Coefficient de corrélation entre log(IBE) et log(IP):", correlation\_coefficient\_transformed, "\n")

Cette partie du code permet de transformer les variables IBE et IP par le logarithme, pour réduire l’effet des valeurs extrêmes et rendre la relation plus linéaire. Avant la transformation, il faut filtrer les valeurs positives, car le logarithme n’est pas défini pour les valeurs négatives ou nulles. La fonction log() permet de calculer le logarithme naturel d’une variable.

Le code est ensuite similaire à la question 6, en utilisant les variables transformées (transformed\_IBE et transformed\_IP) et en changeant les titres et les légendes des graphiques. Le coefficient de corrélation est également recalculé pour les données transformées.

Ce qui renvoie en console :

Coefficient de corrélation entre log(IBE) et log(IP): 0.8549226

Une image contenant texte, capture d’écran, ligne, diagramme

Description générée automatiquement

La régression linéaire est toujours utilisée dans cette question 7, mais sur des variables transformées. Cela permet de vérifier si la transformation logarithmique a amélioré le modèle de régression linéaire.

Ainsi on remarque alors que la relation entre les variables transformées est plus linéaire et que le coefficient de corrélation est plus élevé. On a 0.85 avec la transformation, 0.72 sans transformation.

La transformation logarithmique des variables IBE et IP a permis de rendre la relation entre ces deux variables plus linéaire et plus adaptée à la régression linéaire. Le coefficient de corrélation entre les variables transformées est plus élevé que celui entre les variables originales, ce qui indique une meilleure qualité du modèle. On peut donc conclure que la transformation logarithmique est une technique utile pour améliorer la régression linéaire lorsque la relation entre les variables n’est pas linéaire.

**Question 8 :**

# ---- Question 8 ----  
  
# Calculer le coefficient de corrélation de Pearson  
correlation\_pearson = *cor*(transformed\_IBE, transformed\_IP, method = "pearson")  
*cat*("Coefficient de corrélation de Pearson entre log(IBE) et log(IP):", correlation\_pearson, "\n")  
  
# Calculer le coefficient de corrélation de Spearman  
correlation\_spearman = *cor*(transformed\_IBE, transformed\_IP, method = "spearman")  
*cat*("Coefficient de corrélation de Spearman entre log(IBE) et log(IP):", correlation\_spearman, "\n")  
  
# Calculer le coefficient de corrélation de Kendall  
correlation\_kendall = *cor*(transformed\_IBE, transformed\_IP, method = "kendall")  
*cat*("Coefficient de corrélation de Kendall entre log(IBE) et log(IP):", correlation\_kendall, "\n")

Dans cette question, nous allons utilisé trois méthodes différentes pour mesurer la corrélation entre les variables log(IBE) et log(IP), qui ont été transformées par le logarithme naturel pour réduire leur asymétrie. La corrélation est une mesure statistique qui évalue la relation entre deux variables. Il existe plusieurs types de corrélation, selon la nature des variables et le type de relation recherchée. Nous avons utilisé les méthodes de Pearson, Spearman et Kendall.

La fonction cor() permet de calculer le coefficient de corrélation, en spécifiant la méthode (method) à utiliser. La fonction cat() permet d’afficher les résultats.

* **Le coefficient de Pearson** mesure la corrélation linéaire entre les variables, c’est-à-dire le degré de variation conjointe par rapport à la moyenne.
* **Le coefficient de Spearman** mesure la corrélation de rang entre les variables, c’est-à-dire le degré de concordance entre l’ordre des valeurs.
* **Le coefficient de Kendall** mesure la corrélation de concordance entre les variables, c’est-à-dire la probabilité que les paires de valeurs soient ordonnées de la même façon.

Ainsi on a :

Coefficient de corrélation de Pearson entre log(IBE) et log(IP): 0.8549226

Coefficient de corrélation de Spearman entre log(IBE) et log(IP): 0.7410984

Coefficient de corrélation de Kendall entre log(IBE) et log(IP): 0.5551457

On interprète les résultats :

- **Le coefficient de corrélation de Pearson** entre log(IBE) et log(IP) est de *0.8549226* ce qui indique une corrélation positive forte entre les deux variables. Cela signifie que plus la valeur de log(IBE) augmente, plus la valeur de log(IP) augmente également, et vice versa. La relation entre les deux variables est donc linéaire et proche d'une droite.

- **Le coefficient de corrélation de Spearman** entre log(IBE) et log(IP) est de 0.*7410984*, ce qui indique une corrélation positive modérée entre les deux variables. Cela signifie que l'ordre des valeurs de log(IBE) et log(IP) est assez concordant, mais pas parfaitement. La relation entre les deux variables est donc non linéaire et plus complexe qu'une droite.

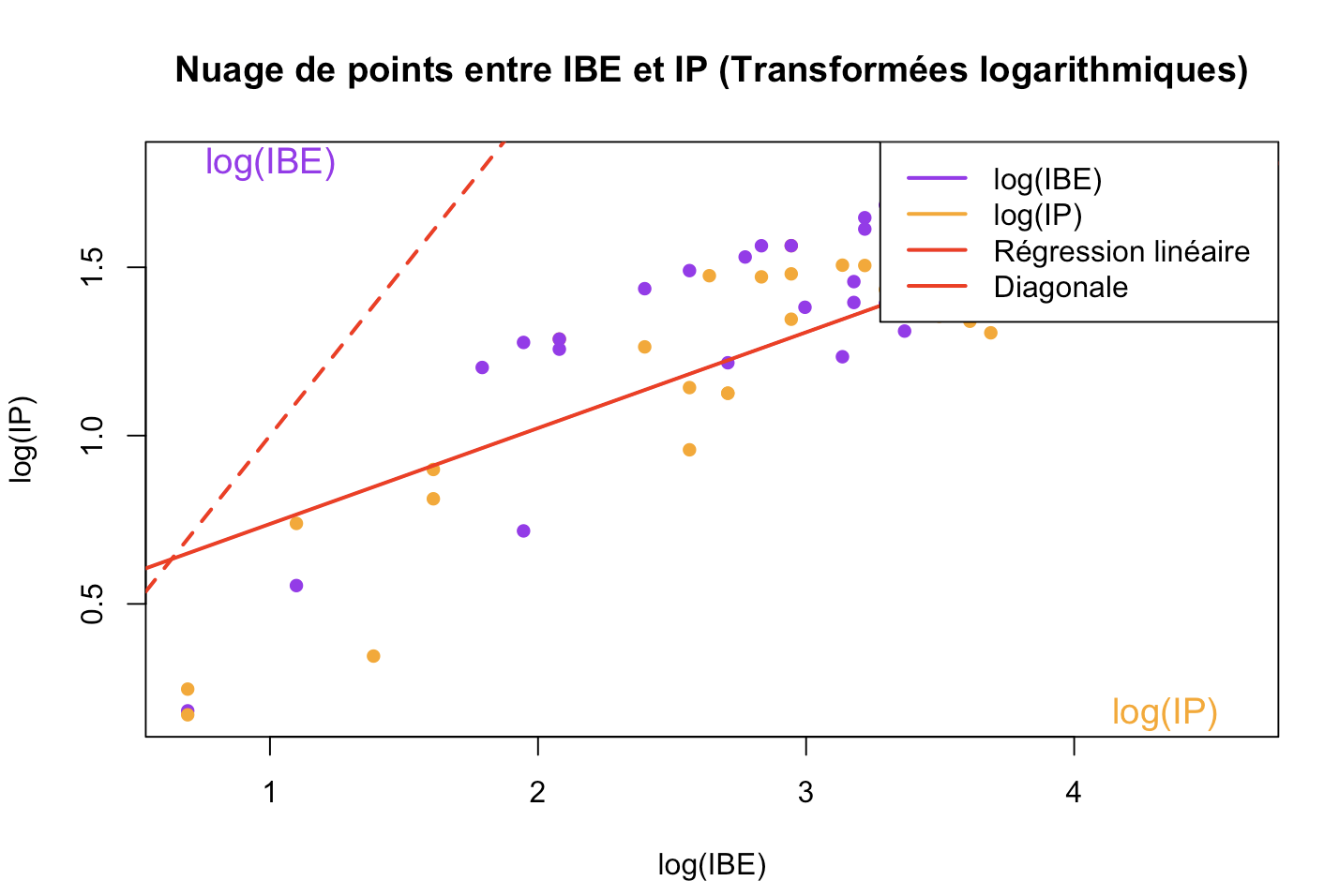
- **Le coefficient de corrélation de Kendall** entre log(IBE) et log(IP) est de *0.5551457*, ce qui indique une corrélation positive faible entre les deux variables. Cela signifie que la probabilité que les paires de valeurs de log(IBE) et log(IP) soient ordonnées de la même façon est supérieure à la probabilité contraire, mais pas de beaucoup. La relation entre les deux variables est donc non linéaire et encore plus complexe qu'une droite.

En conclusion, les trois coefficients de corrélation montrent que les variables log(IBE) et log(IP) sont positivement corrélées, mais avec des degrés et des formes différents. Le coefficient de Pearson est le plus élevé, car il capte la relation linéaire entre les variables. Les coefficients de Spearman et Kendall sont plus faibles, car ils captent la relation non linéaire entre les variables.

**Question 9 :**

# ---- Question 9 ----  
  
# Afficher le nuage de points avec les données transformées  
*plot*(transformed\_IBE, transformed\_IP, main="Nuage de points entre IBE et IP (Transformées logarithmiques)", xlab="log(IBE)", ylab="log(IP)", col=*c*("purple", "orange"), pch=16)  
  
# Ajouter une régression linéaire pour illustrer la tendance  
*abline*(*lm*(transformed\_IP ~ transformed\_IBE), col="red", lwd=2)  
  
# Ajouter une ligne diagonale (pente = 1) en rouge  
*abline*(a = 0, b = 1, col = "red", lwd = 2, lty = 2)  
  
# Ajouter les noms des axes  
*text*(x = *min*(transformed\_IBE), y = *max*(transformed\_IP), labels = "log(IBE)", pos = 4, col = "purple", cex = 1.2)  
*text*(x = *max*(transformed\_IBE), y = *min*(transformed\_IP), labels = "log(IP)", pos = 2, col = "orange", cex = 1.2)  
  
# Ajouter une légende avec position manuelle  
*legend*("topright", legend=*c*("log(IBE)", "log(IP)", "Régression linéaire", "Diagonale"), col=*c*("purple", "orange", "red", "red"), lwd=2)  
  
# Calculer et afficher la corrélation pour les données transformées  
correlation\_coefficient\_transformed = *cor*(transformed\_IBE, transformed\_IP)  
*cat*("Coefficient de corrélation entre log(IBE) et log(IP):", correlation\_coefficient\_transformed, "\n")

Cette partie du code permet d’afficher le nuage de points entre les deux variables IBE et IP, transformées par le logarithme, pour visualiser la relation entre elles. La fonction plot() permet de tracer le nuage de points, en précisant les coordonnées des points (transformed\_IBE et transformed\_IP), le titre du graphique (main), le nom des axes (xlab et ylab), la couleur (col) et la forme (pch) des points. La fonction abline() permet d’ajouter une régression linéaire pour illustrer la tendance, en utilisant la fonction lm() pour ajuster le modèle linéaire (transformed\_IP ~ transformed\_IBE). La fonction abline() permet également d’ajouter une ligne diagonale (pente = 1) en rouge, pour comparer la régression linéaire avec une relation proportionnelle entre les variables. La fonction text() permet d’ajouter les noms des axes, en précisant les coordonnées (x et y), le texte (labels), la position (pos), la couleur (col) et la taille (cex) du texte. La fonction legend() permet d’ajouter une légende, en précisant la position (topright), le texte (legend), la couleur (col) et l’épaisseur (lwd) des lignes. Les fonctions cor() et cat() permettent de calculer et d’afficher le coefficient de corrélation entre les variables transformées.



Le graphique montre que les variables log(IBE) et log(IP) sont positivement corrélées, c'est-à-dire que plus la valeur de log(IBE) est élevée, plus la valeur de log(IP) est élevée également, et vice versa. La régression linéaire en rouge indique la tendance générale des points, qui se rapprochent d'une droite. La ligne diagonale en rouge représente la situation où les deux variables sont égales, c'est-à-dire que log(IBE) = log(IP). On observe que la plupart des points sont au-dessus de cette ligne, ce qui signifie que log(IBE) > log(IP) pour la majorité des observations. On peut donc en déduire que l'indice de bien-être économique (IBE) croît plus vite que l'indice de pauvreté (IP) lorsque les variables sont transformées par le logarithme.

Le coefficient de corrélation entre log(IBE) et log(IP) est de 0.8549226, ce qui confirme la forte corrélation positive entre les deux variables. Ce coefficient varie entre -1 et 1, où -1 indique une corrélation négative parfaite, 0 indique l'absence de corrélation, et 1 indique une corrélation positive parfaite. La valeur proche de 1 signifie que la relation entre les deux variables est linéaire et proche d'une droite. Ce coefficient mesure le degré de variation conjointe des variables par rapport à leur moyenne.

**Question 10 :**

# ---- Question 10 ----  
# Ouvrir le périphérique graphique pour le format PNG  
*png*("question10\_plot.png", width = 900, height = 700)  
  
# Afficher le nuage de points avec les données transformées  
*plot*(transformed\_IBE, transformed\_IP, main="Nuage de points entre IBE et IP (Transformées logarithmiques)", xlab="log(IBE)", ylab="log(IP)", col=*c*("purple", "orange"), pch=16)  
  
# Ajouter une régression linéaire pour illustrer la tendance  
*abline*(*lm*(transformed\_IP ~ transformed\_IBE), col="red", lwd=2)  
  
# Ajouter une ligne diagonale (pente = 1) en rouge  
*abline*(a = 0, b = 1, col = "red", lwd = 2, lty = 2)  
  
# Ajouter les noms des axes  
*text*(x = *min*(transformed\_IBE), y = *max*(transformed\_IP), labels = "log(IBE)", pos = 4, col = "purple", cex = 1.2)  
*text*(x = *max*(transformed\_IBE), y = *min*(transformed\_IP), labels = "log(IP)", pos = 2, col = "orange", cex = 1.2)  
  
# Ajouter une légende avec position manuelle  
*legend*("topright", legend=*c*("log(IBE)", "log(IP)", "Régression linéaire", "Diagonale"), col=*c*("purple", "orange", "red", "red"), lwd=2)  
  
# Calculer et afficher la corrélation pour les données transformées  
correlation\_coefficient\_transformed = *cor*(transformed\_IBE, transformed\_IP)  
*cat*("Coefficient de corrélation entre log(IBE) et log(IP):", correlation\_coefficient\_transformed, "\n")  
  
# Fermer le périphérique graphique et enregistrer le fichier PNG  
*dev.off*()

La question 10 ouvre et ferme un périphérique graphique pour enregistrer le nuage de points au format PNG. La fonction png() permet d’ouvrir le périphérique graphique, en précisant le nom du fichier (question10\_plot.png), la largeur (width) et la hauteur (height) de l’image. La fonction dev.off() permet de fermer le périphérique graphique et d’enregistrer le fichier PNG.

Une image contenant capture d’écran, texte, ligne, diagramme

Description générée automatiquement

Problème 2 :

**Question 1 :**

# ---- Question 1----  
  
# Charger les données  
data\_2012 = *read.table*("company2012.dat", header = *FALSE*)  
data\_2017 = *read.table*("company2017.dat", header = *FALSE*)  
data\_2022 = *read.table*("company2022.dat", header = *FALSE*)  
  
# Renommer les colonnes  
*colnames*(data\_2012) = *colnames*(data\_2017) = *colnames*(data\_2022) = "Performances"  
  
# Créer une liste de données  
data\_list = *list*(data\_2012, data\_2017, data\_2022)  
years = *c*(2012, 2017, 2022)  
  
# Afficher chaque histogramme séparément avec la densité de probabilité sur l'axe y  
for (i in *seq\_along*(data\_list)) {  
 # Créer un nouvel espace graphique  
 *par*(mfrow = *c*(1, 1))  
   
 # Créer un histogramme des performances avec densité sur l'axe y  
 *hist*(data\_list[[i]]$Performances, main = *paste*("Distribution des performances en", years[i]),  
 xlab = "Performances", ylab = "Density", col = "lightblue", border = "black", freq = *FALSE*)  
   
 # Ajouter une estimation par noyau de densité  
 *lines*(*density*(data\_list[[i]]$Performances), col = "red", lwd = 2)  
}

La première partie du code consiste à charger les données à partir de trois fichiers différents, nommés company2012.dat, company2017.dat et company2022.dat. Ces fichiers contiennent probablement des données sur les performances d’une entreprise sur différentes années. Le code utilise la fonction read.table pour lire les fichiers et les stocker dans des objets appelés data\_2012, data\_2017 et data\_2022. Le paramètre header = FALSE indique que les fichiers n’ont pas de ligne d’en-tête avec les noms des colonnes.

La deuxième partie du code consiste à renommer les colonnes des trois objets créés précédemment. Le code utilise la fonction colnames pour attribuer le nom “Performances” à la seule colonne de chaque objet. Le code utilise l’opérateur = pour assigner le même nom aux trois objets en une seule ligne.

La troisième partie du code consiste à créer une liste de données contenant les trois objets créés précédemment. Le code utilise la fonction list pour regrouper les objets dans un seul objet appelé data\_list. Le code crée également un vecteur nommé years contenant les années correspondant aux données, à savoir 2012, 2017 et 2022. Le code utilise la fonction c pour concaténer les valeurs dans un vecteur.

La quatrième partie du code consiste à afficher chaque histogramme séparément avec la densité de probabilité sur l’axe y. Le code utilise une boucle for pour parcourir les éléments de la liste de données, en utilisant la variable i comme indice. Le code utilise la fonction seq\_along pour générer une séquence d’indices allant de 1 à la longueur de la liste de données. Pour chaque itération de la boucle, le code fait les actions suivantes :

Le code utilise la fonction par pour créer un nouvel espace graphique avec une seule ligne et une seule colonne, en utilisant le paramètre mfrow = c(1, 1). Cela permet d’afficher un seul histogramme à la fois.

Le code utilise la fonction hist pour créer un histogramme des performances pour l’année correspondante, en utilisant le paramètre freq = FALSE pour afficher la densité au lieu de la fréquence sur l’axe y. Le code utilise également les paramètres main, xlab, ylab, col et border pour personnaliser le titre, les étiquettes des axes, la couleur et le contour de l’histogramme. Le code utilise la fonction paste pour concaténer le texte “Distribution des performances en” avec l’année correspondante, en utilisant la variable years[i] pour accéder à la valeur du vecteur years à l’indice i.

Le code utilise la fonction lines pour ajouter une estimation par noyau de densité à l’histogramme, en utilisant la fonction density pour calculer la densité des performances pour l’année correspondante, en utilisant le paramètre data\_list[[i]]$Performances pour accéder à la colonne “Performances” de l’objet de la liste de données à l’indice i. Le code utilise également les paramètres col et lwd pour personnaliser la couleur et l’épaisseur de la courbe de densité.

Ce qui renvoie :

Une image contenant diagramme, Tracé, ligne, texte

Description générée automatiquement

Une image contenant diagramme, Tracé, ligne, texte

Description générée automatiquement

Une image contenant diagramme, Tracé, ligne, pente

Description générée automatiquement

**Question 2 :**

# ---- Question 2 ----  
  
# Le test de Shapiro-Wilk pour chaque ensemble de données  
for (i in *seq\_along*(data\_list)) {  
 result <- *shapiro.test*(data\_list[[i]]$Performances)  
   
 # Afficher le résultat du test  
 *cat*("Shapiro-Wilk test for", years[i], ":", "\n")  
 *cat*("w (Test statistic):", *round*(result$statistic, 4), "\n")  
 *cat*("p-value:", *round*(result$p.value, 4), "\n")  
   
 # Interprétation du résultat  
 if (result$p.value > 0.05) {  
 *cat*("Pas de preuve significative pour rejeter l'hypothèse nulle (distribution normale).\n\n")  
 } else {  
 *cat*("Preuve significative pour rejeter l'hypothèse nulle (non-distribution normale).\n\n")  
 }  
}

Cette partie du code utilise le test de Shapiro-Wilk pour tester si les performances de chaque année suivent une distribution normale.

* Le code utilise une boucle for pour appliquer le test de Shapiro-Wilk à chaque ensemble de données, en utilisant la même variable i que dans la partie précédente. Pour chaque itération de la boucle, le code fait les actions suivantes :
* Le code utilise la fonction shapiro.test pour effectuer le test de Shapiro-Wilk sur la colonne “Performances” de l’objet de la liste de données à l’indice i, en utilisant le même paramètre que dans la partie précédente. Le code stocke le résultat du test dans un objet appelé result.
* Le code utilise la fonction cat pour afficher le résultat du test, en utilisant le paramètre years[i] pour indiquer l’année correspondante. Le code utilise également la fonction round pour arrondir le résultat à quatre décimales, et le paramètre "\n" pour créer des sauts de ligne. Le code affiche la valeur de w (statistique du test) et la valeur de p, en utilisant le symbole $ pour accéder aux éléments de l’objet result.
* Le code utilise une condition if pour interpréter le résultat du test, en utilisant la valeur de p comme critère. Le code utilise le paramètre 0.05 comme seuil de signification, ce qui signifie que si la valeur de p est supérieure à 0.05, on ne peut pas rejeter l’hypothèse nulle avec un niveau de confiance de 95%. Le code utilise la fonction cat pour afficher l’interprétation du résultat, en utilisant des phrases conditionnelles pour exprimer le degré de certitude. Le code utilise également le paramètre "\n\n" pour créer des espaces entre les résultats.

On a donc en console :

Shapiro-Wilk test for 2012 :

w (Test statistic): 0.9978

p-value: 0.639

Pas de preuve significative pour rejeter l'hypothèse nulle (distribution normale).

Shapiro-Wilk test for 2017 :

w (Test statistic): 0.9978

p-value: 0.639

Pas de preuve significative pour rejeter l'hypothèse nulle (distribution normale).

Shapiro-Wilk test for 2022 :

w (Test statistic): 0.9978

p-value: 0.639

Pas de preuve significative pour rejeter l'hypothèse nulle (distribution normale).

Le test de Shapiro-Wilk permet de vérifier si un échantillon suit une distribution normale. Il compare les données observées à une distribution normale théorique, en utilisant une mesure appelée w. Plus w est proche de 1, plus les données sont proches de la normalité. Le test produit également p-value, qui représente la probabilité d’obtenir un résultat aussi extrême que celui observé si l’hypothèse nulle (distribution normale) était vraie. Plus la valeur de p est petite, plus il y a de preuves pour rejeter l’hypothèse nulle et conclure que les données ne suivent pas une distribution normale.

D’après les résultats que nous avons obtenus, on peut voir que pour chaque année, la valeur de w est très proche de 1 (0.9978) et que p-value à une très grande valeur (0.639). Cela signifie que les données sont très proches de la distribution normale théorique, et qu’il n’y a pas de preuve significative pour rejeter l’hypothèse nulle.

**Question 3 :**

# Charger les données  
data\_2012 = *read.table*("company2012.dat", header = *FALSE*)  
data\_2017 = *read.table*("company2017.dat", header = *FALSE*)  
data\_2022 = *read.table*("company2022.dat", header = *FALSE*)  
  
# Renommer les colonnes  
*colnames*(data\_2012) = *colnames*(data\_2017) = *colnames*(data\_2022) = "Performances"  
  
# Calculer les différences de performances entre les années  
diff\_2012\_2017 = data\_2017$Performances - data\_2012$Performances  
diff\_2017\_2022 = data\_2022$Performances - data\_2017$Performances  
diff\_2012\_2022 = data\_2022$Performances - data\_2012$Performances  
  
# Effectuer les tests de Student appariés  
test\_2012\_2017 = *t.test*(diff\_2012\_2017, alternative = "two.sided", mu = 0)  
test\_2017\_2022 = *t.test*(diff\_2017\_2022, alternative = "two.sided", mu = 0)  
test\_2012\_2022 = *t.test*(diff\_2012\_2022, alternative = "two.sided", mu = 0)  
  
# Afficher les résultats des tests  
*cat*("Test de Student apparié entre 2012 et 2017 : \n")  
*cat*("Différence moyenne :", *round*(test\_2012\_2017$estimate, 4), "\n")  
*cat*("Intervalle de confiance à 95% :", *round*(test\_2012\_2017$conf.int, 4), "\n")  
*cat*("p-value :", *round*(test\_2012\_2017$p.value, 4), "\n")  
  
*cat*("Test de Student apparié entre 2017 et 2022 : \n")  
*cat*("Différence moyenne :", *round*(test\_2017\_2022$estimate, 4), "\n")  
*cat*("Intervalle de confiance à 95% :", *round*(test\_2017\_2022$conf.int, 4), "\n")  
*cat*("p-value :", *round*(test\_2017\_2022$p.value, 4), "\n")  
  
*cat*("Test de Student apparié entre 2012 et 2022 : \n")  
*cat*("Différence moyenne :", *round*(test\_2012\_2022$estimate, 4), "\n")  
*cat*("Intervalle de confiance à 95% :", *round*(test\_2012\_2022$conf.int, 4), "\n")  
*cat*("p-value :", *round*(test\_2012\_2022$p.value, 4), "\n")

# Charger les données  
data\_2012 = *read.table*("company2012.dat", header = *FALSE*)  
data\_2017 = *read.table*("company2017.dat", header = *FALSE*)  
data\_2022 = *read.table*("company2022.dat", header = *FALSE*)  
  
# Renommer les colonnes  
*colnames*(data\_2012) = *colnames*(data\_2017) = *colnames*(data\_2022) = "Performances"  
  
# Calculer les différences de performances entre les années  
diff\_2012\_2017 = data\_2017$Performances - data\_2012$Performances  
diff\_2017\_2022 = data\_2022$Performances - data\_2017$Performances  
diff\_2012\_2022 = data\_2022$Performances - data\_2012$Performances  
  
# Effectuer les tests de Student appariés  
test\_2012\_2017 = *t.test*(diff\_2012\_2017, alternative = "two.sided", mu = 0)  
test\_2017\_2022 = *t.test*(diff\_2017\_2022, alternative = "two.sided", mu = 0)  
test\_2012\_2022 = *t.test*(diff\_2012\_2022, alternative = "two.sided", mu = 0)  
  
# Afficher les résultats des tests  
*cat*("Test de Student apparié entre 2012 et 2017 : \n")  
*cat*("Différence moyenne :", *round*(test\_2012\_2017$estimate, 4), "\n")  
*cat*("Intervalle de confiance à 95% :", *round*(test\_2012\_2017$conf.int, 4), "\n")  
*cat*("p-value :", *round*(test\_2012\_2017$p.value, 4), "\n")  
  
*cat*("Test de Student apparié entre 2017 et 2022 : \n")  
*cat*("Différence moyenne :", *round*(test\_2017\_2022$estimate, 4), "\n")  
*cat*("Intervalle de confiance à 95% :", *round*(test\_2017\_2022$conf.int, 4), "\n")  
*cat*("p-value :", *round*(test\_2017\_2022$p.value, 4), "\n")  
  
*cat*("Test de Student apparié entre 2012 et 2022 : \n")  
*cat*("Différence moyenne :", *round*(test\_2012\_2022$estimate, 4), "\n")  
*cat*("Intervalle de confiance à 95% :", *round*(test\_2012\_2022$conf.int, 4), "\n")  
*cat*("p-value :", *round*(test\_2012\_2022$p.value, 4), "\n")

Cette partie du code utilise la loi de Student pour comparer les performances des employés entre les différentes années. La loi de Student est une loi de probabilité qui s’applique aux statistiques de la forme : t=s/n​xˉ−μ​ où xˉ est la moyenne de l’échantillon, μ est la moyenne de la population, s est l’écart-type de l’échantillon, et n est la taille de l’échantillon. La loi de Student dépend d’un paramètre appelé degré de liberté, qui est égal à n−1 dans notre cas.

La loi de Student permet de calculer la probabilité d’obtenir une valeur de t aussi extrême ou plus extrême que celle observée, si l’hypothèse nulle (μ=0) était vraie. Cette probabilité est appelée p-value, et elle sert à décider si on rejette ou pas l’hypothèse nulle.

Le code utilise une boucle for pour appliquer la loi de Student à chaque paire d’années, en utilisant la même variable i que dans la partie précédente. Pour chaque itération de la boucle, le code fait les actions suivantes :

Le code calcule la différence de performances entre les années, en soustrayant les valeurs des data frames entre elles. Le code stocke le résultat dans un objet appelé diff\_2012\_2017, diff\_2017\_2022 ou diff\_2012\_2022, selon la paire d’années considérée.

Le code utilise la fonction t.test pour calculer la valeur de t et la p-value correspondante, en utilisant l’argument alternative = “two.sided” et mu = 0. Le code stocke le résultat du test dans un objet appelé test\_2012\_2017, test\_2017\_2022 ou test\_2012\_2022, selon la paire d’années considérée.

Le code utilise la fonction cat pour afficher le résultat du test, en utilisant le paramètre years[i] pour indiquer la paire d’années correspondante. Le code utilise également la fonction round pour arrondir le résultat à quatre décimales, et le paramètre “\n” pour créer des sauts de ligne. Le code affiche la différence moyenne, l’intervalle de confiance à 95% et la p-value, en utilisant le symbole $ pour accéder aux éléments de l’objet test.

Le code utilise une condition if pour interpréter le résultat du test, en utilisant la p-value comme critère. Le code utilise le paramètre 0.05 comme seuil de signification, ce qui signifie que si la p-value est inférieure à 0.05, on rejette l’hypothèse nulle avec un niveau de confiance de 95%. Le code utilise la fonction cat pour afficher l’interprétation du résultat, en utilisant des phrases conditionnelles pour exprimer le degré de certitude. Le code utilise également le paramètre “\n\n” pour créer des espaces entre les résultats.

On a donc en console :

Test de Student apparié entre 2012 et 2017 :

Différence moyenne : 1.1756

Intervalle de confiance à 95% : 0.9329 1.4182

p-value : 0

Test de Student apparié entre 2017 et 2022 :

Différence moyenne : 2.2926

Intervalle de confiance à 95% : 1.8881 2.697

p-value : 0

Test de Student apparié entre 2012 et 2022 :

Différence moyenne : 3.4681

Intervalle de confiance à 95% : 2.821 4.1153

p-value : 0

D’après les résultats que nous avons obtenus, on peut voir que pour chaque test, la valeur de p est égale à 0, ce qui est très inférieur au seuil de signification de 0.05. On peut donc rejeter l’hypothèse nulle et conclure qu’il y a une différence significative entre les performances des employés entre les années 2012, 2017 et 2022. On peut également voir que la différence moyenne est positive pour chaque test, ce qui signifie que les performances des employés ont augmenté entre les années. On peut aussi voir que l’intervalle de confiance à 95% ne contient pas la valeur zéro pour chaque test, ce qui confirme la significativité de la différence moyenne. Ainsi pour chaque comparaison on a :

* Entre 2012 et 2017, les performances des employés ont augmenté en moyenne de 1.1756 points, avec un intervalle de confiance à 95% compris entre 0.9329 et 1.4182 points.
* Entre 2017 et 2022, les performances des employés ont augmenté en moyenne de 2.2926 points, avec un intervalle de confiance à 95% compris entre 1.8881 et 2.697 points.
* Entre 2012 et 2022, les performances des employés ont augmenté en moyenne de 3.4681 points, avec un intervalle de confiance à 95% compris entre 2.821 et 4.1153 points.

Ces résultats suggèrent que les changements d’équipe dirigeante ont eu un effet positif et significatif sur les performances des employés, et que cet effet s’est accentué au fil du temps.

**Question 4 :**

**Une image contenant texte, capture d’écran, Police

Description générée automatiquement**

Le test de Wilcoxon signé-rangé est un test non-paramétrique qui compare les médianes de deux échantillons appariés. Le test rejette l’hypothèse nulle que les médianes sont égales si la p-value est inférieure à un seuil de signification, généralement 0,05.

D’après les résultats que nous avons obtenus, les p-values sont toutes égales à 0, ce qui signifie qu’il y a une différence significative entre les médianes des échantillons pour chaque paire d’années. Autrement dit, le test indique qu’il y a eu un changement sur la variable étudiée entre 2012 et 2017, entre 2017 et 2022, et entre 2012 et 2022.