

# Travaux Dirigés N°2 MIO2

## Initiation à R pour l'Analyse de Données

Chargé de Cours: Dr Fatou Néné DIOP ; Chargé de TD/TP: Mr Abdoulaye Fall

(2024 / 2025)

### Objectif général :

S'initier à l'utilisation du logiciel R dans le cadre de l'analyse statistique : manipulation de données, représentations graphiques, tests statistiques et régression.

### 1 : Premiers pas avec R

#### Activités pratiques :

```
a <- 1:10
b <- rnorm(10)
plot(a, b, type = "b", col = "blue")
mean(b)
sd(b)
summary(b)
```

#### Questions :

1. Que fait la fonction `rnorm(10)` ?
2. Quelle est la différence entre `mean(b)` et `summary(b)` ?
3. Que représente le graphique généré par `plot(a, b)` ?
4. Que se passe-t-il si on écrit `plot(b, a)` au lieu de `plot(a, b)` ?

### 2 : Manipulation de données

#### Activités pratiques :

```
df <- read.table("data.txt", header = TRUE)
summary(df)
barplot(df$score, names.arg = df$nom)
```

#### Questions :

1. Que fait la fonction `read.table()` ?
2. Que signifie `header = TRUE` ?
3. Que représente l'argument `names.arg = df$nom` dans `barplot()` ?
4. Comment accéder uniquement à la colonne `age` de l'objet `df` ?

### 3 : Statistiques descriptives et graphiques

#### Activités pratiques :

```
x <- rnorm(100, mean = 50, sd = 10)
```

```
y <- runif(100, min = 30, max = 70)
hist(x)
boxplot(x)
hist(y, col = "lightgreen", breaks = 20)
```

### Questions :

1. Quelle est la différence entre une loi normale et une loi uniforme ?
2. Que représente la boîte à moustaches ? Qu'indiquent les traits ?
3. Qu'observez-vous si vous augmentez le nombre de classes dans `hist(x, breaks = 30)` ?
4. Quelle commande permettrait de superposer deux histogrammes ?

## 4 : Tests statistiques

### Activités pratiques :

```
a <- rnorm(15, mean = 100)
b <- rnorm(15, mean = 105)
t.test(a, b)

groupe <- gl(3, 10, labels = c("A", "B", "C"))
notes <- c(rnorm(10, 12), rnorm(10, 14), rnorm(10, 15))
anova_result <- aov(notes ~ groupe)
summary(anova_result)
```

### Questions :

1. Quelle est l'hypothèse nulle du test t ? Et de l'ANOVA ?
2. Que signifie un p-value  $< 0.05$  dans ce contexte ?
3. Quelle différence y a-t-il entre `t.test(a, b)` et `aov(notes ~ groupe)` ?
4. Peut-on utiliser un test t à la place de l'ANOVA dans l'exemple ci-dessus ? Pourquoi ?

## 5 : Régression linéaire

### Activités pratiques :

```
x <- rnorm(100, 10, 2)
y <- 3 * x + rnorm(100)
model <- lm(y ~ x)
summary(model)
plot(x, y)
abline(model, col = "red")
```

### Questions :

1. Quel est le rôle de la fonction `lm()` ?
2. Que représente `abline(model)` dans le graphique ?
3. Comment interpréter le coefficient de `x` dans `summary(model)` ?
4. Si tu voulais prédire une valeur de `y` pour `x = 12`, que devrais-tu faire ?

## 6. Exemples d'analyses de données

### 6 A.1 – Comparaison de deux médicaments sur le sommeil

**Données :** sommeil1.txt – scores de 10 patients ayant testé deux médicaments successivement.

1. Charger les données :
2. `sommeil <- read.table("sommeil1.txt", header = TRUE)`
3. `attach(sommeil)`
4. Afficher les statistiques descriptives :
5. `summary(M1)`
6. `summary(M2)`
7. Visualiser la relation :
8. `plot(M1, M2, xlim = c(0, 10), ylim = c(0, 10), col = 2)`
9. `abline(0, 1)`
10. Étudier la différence :
11. `stripchart(M2 - M1, method = "stack")`
12. `t.test(M2 - M1)`
13. `t.test(M1, M2, paired = TRUE)`
14. `detach(sommeil)`

### Questions

1. Quelle est la moyenne des scores sous M1 et M2 ?
2. Quelle représentation graphique illustre mieux la comparaison ?

### 6. A.2 – Comparaison de méthodes pédagogiques

**Données :** pedago.txt – méthode (pedago) et note.

1. Charger les données :
2. `a <- read.table("pedago.txt")`
3. `attach(a)`
4. Résumé graphique :
5. `boxplot(notes ~ pedago)`
6. `stripchart(notes ~ pedago, method="stack", vertical=TRUE)`
7. Moyennes par groupe :
8. `tapply(notes, pedago, mean)`
9. `tapply(notes, pedago, sd)`
10. `barplot(t(tapply(notes, pedago, mean)))`
11. Analyse de variance + post hoc :
12. `m <- aov(notes ~ pedago)`
13. `summary(m)`
14. `TukeyHSD(m)`
15. `plot(TukeyHSD(m))`

### Questions

1. Quelle méthode donne les meilleures performances ?
2. Que montre l'ANOVA ?
3. Y a-t-il une différence significative entre certaines méthodes ?

## 6. A.3 – Effets croisés de 2 facteurs sur une performance

**Données :** neglig4.txt

1. Charger les données :
2. `d <- read.table("neglige4.txt")`
3. `x <- d$V1`
4. `a <- gl(2, 12, 24)`
5. `b <- gl(2, 6, 24)`
6. Moyennes croisées :
7. `tapply(x, list(a = a, b = b), mean)`
8. Tracer les interactions :
9. `interaction.plot(a, b, x)`
10. Analyse de variance :
11. `l <- aov(x ~ a * b)`
12. `summary(l)`
13. `model.tables(l, se = TRUE)`

### Questions

1. Y a-t-il un effet principal de a ? de b ?
2. L'interaction entre a et b est-elle significative ?
3. Quelles combinaisons ont les meilleures performances ?

## 6. A.4 – Corrélation entre perceptions familiales

**Données :** family.txt – variables : FATH, MOTH, INST, GIRL

1. Charger les données :
2. `fam <- read.table("family.txt", header = TRUE)`
3. `attach(fam)`
4. Corrélations :
5. `data <- as.matrix(fam[, -1])`
6. `pairs(data, panel = panel.smooth)`
7. `cor(data)`
8. Tests de corrélation individuels :
9. `cor.test(FATH, GIRL)`
10. `cor.test(MOTH, GIRL)`
11. `cor.test(INST, GIRL)`
12. Régression multiple :
13. `l <- lm(GIRL ~ FATH + MOTH + INST)`
14. `summary(l)`
15. `detach(fam)`

### Questions

1. Quelles sont les variables les plus corrélées ?
2. Quelle variable prédit le mieux le jugement sur GIRL ?
3. L'ajustement global du modèle est-il bon ?

## 6.A.5 – Analyse d'un tableau de contingence

**Données :** io.txt – variables Sexe, Matière

1. Charger les données :
2. `a <- read.table("io.txt", header = TRUE)`
3. `attach(a)`
4. Tableau de contingence :
5. `table(Sexe, Matière)`
6. Test du  $\chi^2$  :
7. `chisq.test(table(Sexe, Matière))`

## Questions

1. Existe-t-il une dépendance entre le sexe et la matière ?
2. Quel est le degré de liberté du test ?
3. Que signifie une p-value  $< 0.05$  dans ce contexte ?

## 7: Régression linéaire

Jeu de données intégré mtcars

(Données techniques sur 32 voitures : mpg = consommation, cyl = nombre de cylindres, hp = puissance, etc.)

### 1. Charger les données et explorer

```
data(mtcars)
head(mtcars)
str(mtcars)
```

### 2. Régression linéaire simple : mpg ~ hp

```
model1 <- lm(mpg ~ hp, data = mtcars)
summary(model1)
plot(mtcars$hp, mtcars$mpg, pch=19, col="blue")
abline(model1, col="red", lwd=2)
```

### 3. Régression linéaire multiple : mpg ~ hp + wt + cyl

```
model2 <- lm(mpg ~ hp + wt + cyl, data = mtcars)
summary(model2)
```

### 4. Diagnostiques graphiques

```
par(mfrow = c(2, 2))
plot(model2)
```

### 5. Comparer les modèles

```
AIC(model1, model2)
```

## Questions

1. Quelle est l'équation estimée de la régression simple ?

2. Le coefficient de hp est-il significatif ?
3. Quel modèle explique mieux mpg (consommation) ?
4. Quelle variable a l'impact négatif le plus fort sur mpg ?
5. Y a-t-il des valeurs influentes ou aberrantes ?