Module

Programmation Informatique Appliquée aux Sciences et Technologies

---

Chapitre III

Programmation Appliquée aux

Sciences Expérimentales (Écologie)

Par Mennour Hocine

Université de Mila

2025-2026

Table des Matières

Introduction

L'analyse statistique constitue un pilier fondamental de la recherche en écologie. Elle permet aux scientifiques de tester des hypothèses, d'identifier des patterns dans les données, et de tirer des conclusions fiables à partir d'observations souvent complexes et variables. Ce chapitre vous présente les principaux tests statistiques utilisés en écologie expérimentale, en vous guidant à travers leur application pratique avec le logiciel R.

Les tests statistiques se divisent en deux grandes catégories : les tests paramétriques, qui supposent que les données suivent une distribution spécifique (généralement normale), et les tests non paramétriques, qui ne font pas d'hypothèse sur la distribution des données. Le choix entre ces deux approches dépend des caractéristiques de vos données et de la validité des hypothèses sous-jacentes. Nous aborderons également les tests de corrélation pour étudier les relations entre variables, ainsi que la régression linéaire pour modéliser ces relations.

Tout au long de ce chapitre, nous continuerons à utiliser le jeu de données sur les chênes-lièges de la région de Mila, ainsi que des exemples complémentaires adaptés aux contextes écologiques. Chaque test sera présenté avec ses hypothèses, sa syntaxe R, l'interprétation des résultats et les conditions d'application.

1. Tests Paramétriques

Les tests paramétriques sont des méthodes statistiques qui supposent que les données échantillonnales proviennent de populations suivant une distribution connue, généralement la distribution normale. Ces tests sont plus puissants que leurs équivalents non paramétriques lorsque les hypothèses sont respectées, c'est-à-dire qu'ils ont une meilleure capacité à détecter une différence réelle lorsqu'elle existe. Cependant, leur validité dépend de conditions strictes qu'il convient de vérifier avant toute application.

1.0. Conditions d'application des tests paramétriques

Avant d'appliquer un test paramétrique, trois conditions fondamentales doivent être vérifiées. Premièrement, la normalité de la distribution : les données doivent suivre approximativement une distribution normale, ce qui peut être vérifié par le test de Shapiro-Wilk ou par l'examen visuel d'un histogramme. Deuxièmement, l'homogénéité des variances (homoscédasticité) : les variances des groupes comparés doivent être égales, vérifiable par le test de Levene ou le test de Bartlett. Troisièmement, l'indépendance des observations : chaque observation doit être indépendante des autres, ce qui relève davantage du plan expérimental que de l'analyse statistique.

Condition	Test de vérification	Alternative si non respectée
Normalité	`shapiro.test()`	Tests non paramétriques
Homoscédasticité	`leveneTest()`	Test de Welch ou non paramétrique
Indépendance	Plan expérimental	Tests pour données appariées

Tableau 1 : Conditions d'application des tests paramétriques

1.1. Le test t de Student

1.1.1. Principe et définition

Le test t de Student, développé par William Sealy Gosset en 1908 sous le pseudonyme de « Student », est l'un des tests statistiques les plus couramment utilisés en recherche. Il permet de comparer les moyennes de deux groupes et de déterminer si la différence observée est statistiquement significative ou si elle peut être attribuée au hasard. En écologie, ce test est fréquemment employé pour comparer des mesures entre deux sites, deux périodes, ou deux traitements expérimentaux.

Le test t repose sur le calcul d'une statistique t qui mesure l'écart entre les moyennes des deux groupes par rapport à la variabilité des données. Plus la valeur de t est élevée (en valeur absolue), plus la différence entre les groupes est importante par rapport à la variabilité intra-groupe. Cette statistique suit une distribution de Student à $(n_1 + n_2 - 2)$ degrés de liberté sous l'hypothèse nulle.

1.1.2. Types de test t

Il existe trois variantes principales du test t, chacune adaptée à une situation expérimentale particulière. Le test t pour échantillons indépendants compare les moyennes de deux groupes distincts et non apparentés, par exemple la hauteur des arbres dans deux forêts différentes. Le test t pour échantillons appariés compare deux séries de mesures effectuées sur les mêmes individus à deux moments différents ou dans deux conditions différentes, par exemple l'épaisseur du liège en 2022 et 2024 sur les mêmes arbres. Le test t à un échantillon compare la moyenne d'un seul groupe à une valeur théorique ou de référence.

1.1.3. Hypothèses statistiques

Pour un test t bilatéral (two-sided), les hypothèses sont formulées comme suit :

Hypothèse nulle (H_0)	Hypothèse alternative (H_1)
$\mu_1 = \mu_2$ (les moyennes sont égales)	$\mu_1 \neq \mu_2$ (les moyennes sont différentes)

Tableau 2 : Hypothèses du test t bilatéral

1.1.4. Application avec R

La fonction **`t.test()`** dans R permet de réaliser les différents types de test t. Voici la syntaxe de base et des exemples concrets adaptés à notre jeu de données écologiques.

Test t pour échantillons indépendants

Ce test compare les moyennes de deux groupes indépendants. Dans notre exemple, nous allons comparer l'épaisseur du liège entre les arbres ayant subi des dégâts de feu et ceux qui n'en ont pas subi. Avant d'appliquer le test, nous devons vérifier les conditions d'application : normalité dans chaque groupe et égalité des variances.

```
# Étape 1 : Vérification de la normalité par groupe
# Groupe avec dégâts de feu
shapiro.test(df$epaisseur_liege_2024[df$degats_feu == TRUE])

# Groupe sans dégâts de feu
shapiro.test(df$epaisseur_liege_2024[df$degats_feu == FALSE])

# Étape 2 : Test t pour échantillons indépendants
# Formule 1 : avec variances égales (var.equal = TRUE)
t.test(epaisseur_liege_2024 ~ degats_feu, data = df, var.equal =
TRUE)

# Formule 2 : test de Welch (variances inégales, par défaut)
t.test(epaisseur_liege_2024 ~ degats_feu, data = df)
```

Explication de la syntaxe : L'expression `epaisseur_liege_2024 ~ degats_feu` utilise la notation formule de R, où la variable à gauche du tilde (~) est la variable dépendante (réponse), et la variable à droite est la variable de groupement (facteur). Le paramètre `var.equal = TRUE` indique que l'on suppose l'égalité des variances. Par défaut (sans ce paramètre ou avec `var.equal = FALSE`), R effectue le test de Welch qui ne suppose pas l'égalité des variances.

Test t pour échantillons appariés

Le test t apparié est utilisé lorsque les deux séries de mesures proviennent des mêmes individus. Dans notre jeu de données, nous pouvons comparer l'épaisseur du liège mesurée en 2022 et en 2024 sur les mêmes arbres. Ce type de test est plus puissant que le test pour échantillons indépendants car il élimine la variabilité inter-individuelle.

```
# Test t pour échantillons appariés
# Comparaison épaisseur liège 2022 vs 2024 (mêmes arbres)
t.test(df$epaisseur_liege_2022, df$epaisseur_liege_2024, paired =
TRUE)
```

Le paramètre `paired = TRUE` indique qu'il s'agit d'un test apparié. Les deux vecteurs doivent avoir la même longueur et les observations doivent être dans le même ordre (la première valeur du premier vecteur correspond à la première valeur du deuxième vecteur pour le même individu).

1.1.5. Interprétation des résultats

La fonction `t.test()` retourne plusieurs informations essentielles pour l'interprétation. La statistique t indique l'importance de la différence relative à la variabilité. Les degrés de liberté (df) déterminent la forme de la distribution t de référence. La p-value représente la probabilité d'observer une différence au moins aussi importante que celle mesurée, si l'hypothèse nulle était vraie. L'intervalle de confiance à 95% donne une estimation de la différence réelle entre les moyennes. Enfin, les moyennes des deux groupes permettent de connaître le sens de la différence.

Règle de décision : Si la p-value est inférieure au seuil de significativité α (généralement 0,05), on rejette l'hypothèse nulle et on conclut qu'il existe une différence significative entre les deux moyennes. Si la p-value est supérieure à α, on ne peut pas rejeter l'hypothèse nulle ; cela ne signifie pas que les moyennes sont égales, mais plutôt qu'il n'y a pas suffisamment de preuves pour conclure à une différence.

1.2. L'Analyse de la Variance (ANOVA)

1.2.1. Principe et définition

L'ANOVA (ANalysis Of VAriance) est une extension du test t permettant de comparer simultanément les moyennes de trois groupes ou plus. Développée par Ronald Fisher dans les années 1920, cette méthode est fondamentale en écologie expérimentale où l'on cherche souvent à comparer plusieurs traitements, plusieurs sites, ou plusieurs périodes. L'ANOVA permet de répondre à la question : existe-t-il au moins une différence significative parmi les moyennes des groupes étudiés ?

Le principe de l'ANOVA repose sur la décomposition de la variance totale des données en deux composantes : la variance inter-groupes (due aux différences entre les moyennes des groupes) et la variance intra-groupes (due à la variabilité individuelle au sein de chaque groupe). Le rapport de ces deux variances, appelé statistique F, permet de tester si la variance inter-groupes est significativement supérieure à ce que l'on attendrait par hasard.

1.2.2. Types d'ANOVA

Il existe plusieurs types d'ANOVA adaptés aux différents plans expérimentaux. L'ANOVA à un facteur (one-way ANOVA) compare les moyennes de plusieurs groupes définis par un seul facteur, par exemple comparer la croissance des plantes selon trois types de sol. L'ANOVA à

deux facteurs (two-way ANOVA) étudie l'effet de deux facteurs simultanément ainsi que leur interaction, par exemple l'effet du type de sol et de l'irrigation sur la croissance. L'ANOVA à mesures répétées est utilisée lorsque les mêmes sujets sont mesurés plusieurs fois dans le temps.

1.2.3. Hypothèses statistiques

Hypothèse nulle (H_0)	Hypothèse alternative (H_1)
$\mu_1 = \mu_2 = \mu_3 = ... = \mu_k$ (toutes les moyennes sont égales)	Au moins une moyenne diffère des autres

Tableau 3 : Hypothèses de l'ANOVA à un facteur

1.2.4. Application avec R : ANOVA à un facteur

Pour illustrer l'ANOVA à un facteur, nous utiliserons la variable `foret` qui indique la forêt d'origine de chaque arbre. Nous testerons si la production de glands diffère significativement entre les forêts. Notez que notre exemple nécessite d'ajouter des forêts supplémentaires au jeu de données pour avoir plus de deux groupes.

```
# Étape 1 : Vérification des conditions d'application
# Test d'homogénéité des variances (test de Levene)
library(car)
leveneTest(production_glands ~ foret, data = df)

# Étape 2 : Réalisation de l'ANOVA
# Méthode 1 : avec aov()
modele_anova <- aov(production_glands ~ foret, data = df)
summary(modele_anova)

# Méthode 2 : avec lm() (plus flexible)
modele_lm <- lm(production_glands ~ foret, data = df)
anova(modele_lm)
```

1.2.5. Tests post-hoc après ANOVA

Lorsque l'ANOVA révèle une différence significative (p-value < 0,05), elle indique qu'au moins une moyenne diffère des autres, mais ne précise pas lesquelles. Les tests post-hoc (ou tests a posteriori) permettent d'identifier les paires de groupes qui diffèrent significativement, tout en contrôlant le risque d'erreur globale lié aux comparaisons multiples.

Le test de Tukey HSD (Honestly Significant Difference) est l'un des plus utilisés. Il compare toutes les paires de moyennes tout en maintenant le taux d'erreur de type I à 5% pour l'ensemble des comparaisons. C'est un test équilibré, ni trop conservateur ni trop libéral, adapté à la plupart des situations.

```
# Test post-hoc de Tukey HSD
TukeyHSD(modele_anova)

# Alternative : test de Dunnett (comparaison à un groupe témoin)
library(multcomp)
summary(glht(modele_anova, linfct = mcp(foret = "Dunnett")))
```

1.2.6. Interprétation des résultats

Le tableau de l'ANOVA contient plusieurs colonnes essentielles. La somme des carrés (Sum Sq) mesure la variabilité attribuable à chaque source (facteur ou résidus). Les degrés de liberté (Df) représentent le nombre de valeurs indépendantes dans le calcul. La moyenne des carrés (Mean Sq) est obtenue en divisant la somme des carrés par les degrés de liberté. La statistique F est le rapport entre la moyenne des carrés du facteur et celle des résidus. La p-value indique la significativité du facteur.

Règle de décision : Si la p-value associée au facteur est inférieure à 0,05, on conclut qu'au moins une des moyennes diffère significativement des autres. Les tests post-hoc permettent alors d'identifier les différences spécifiques entre paires de groupes.

2. Tests Non Paramétriques

Les tests non paramétriques constituent une alternative aux tests paramétriques lorsque les conditions de normalité et d'homoscédasticité ne sont pas respectées. Ces tests ne font pas d'hypothèse sur la distribution sous-jacente des données et travaillent généralement sur les rangs des observations plutôt que sur les valeurs brutes. Bien que légèrement moins puissants que les tests paramétriques lorsque les conditions sont remplies, ils sont plus robustes face aux violations des hypothèses et sont particulièrement adaptés aux petits échantillons ou aux données comportant des valeurs extrêmes.

2.1. Le test de Mann-Whitney

2.1.1. Principe et définition

Le test de Mann-Whitney, également connu sous le nom de test U de Mann-Whitney ou test de Wilcoxon rank-sum, est l'équivalent non paramétrique du test t pour échantillons indépendants. Proposé par Henry Mann et Donald Whitney en 1947, ce test compare les distributions de deux groupes en utilisant les rangs des observations plutôt que les valeurs numériques. Il teste si les observations d'un groupe ont tendance à être systématiquement plus grandes ou plus petites que celles de l'autre groupe.

Le principe du test est le suivant : toutes les observations des deux groupes sont combinées et classées par ordre croissant. On attribue un rang à chaque observation, puis on calcule la somme des rangs pour chaque groupe. Si les deux groupes proviennent de populations identiques, les rangs devraient être répartis de manière similaire entre les groupes. Une différence importante dans la distribution des rangs suggère une différence entre les populations.

2.1.2. Hypothèses statistiques

Hypothèse nulle (H_0)	Hypothèse alternative (H_1)
Les deux populations ont la même distribution	Les distributions diffèrent (tendance centrale ou forme)

Tableau 4 : Hypothèses du test de Mann-Whitney

2.1.3. Application avec R

La fonction **`wilcox.test()`** permet de réaliser le test de Mann-Whitney. Par défaut, cette fonction effectue le test de Mann-Whitney pour échantillons indépendants. Le test de Wilcoxon pour échantillons appariés est obtenu en ajoutant le paramètre `paired = TRUE`.

```
# Test de Mann-Whitney pour échantillons indépendants
# Comparaison épaisseur liège selon dégâts de feu
wilcox.test(epaisseur_liege_2024 ~ degats_feu, data = df)

# Formule alternative avec deux vecteurs séparés
groupe_feu <- df$epaisseur_liege_2024[df$degats_feu == TRUE]
groupe_sans_feu <- df$epaisseur_liege_2024[df$degats_feu == FALSE]
wilcox.test(groupe_feu, groupe_sans_feu)
```

Remarque importante : En cas d'ex-aequo (valeurs identiques dans les données), R affiche un avertissement et calcule une p-value approximative. Pour des résultats plus précis, vous pouvez ajouter le paramètre `exact = TRUE` ou `correct = FALSE` selon les cas.

2.2. Le test de Kruskal-Wallis

2.2.1. Principe et définition

Le test de Kruskal-Wallis, développé par William Kruskal et Allen Wallis en 1952, est l'équivalent non paramétrique de l'ANOVA à un facteur. Il permet de comparer trois groupes ou plus sans supposer la normalité des données. Comme le test de Mann-Whitney, il travaille sur les rangs des observations. En écologie, ce test est particulièrement utile pour comparer des indices de biodiversité, des abondances d'espèces, ou des mesures de qualité environnementale entre plusieurs sites ou traitements.

Le test calcule une statistique H qui suit approximativement une distribution du chi-carré sous l'hypothèse nulle. Cette statistique compare les sommes des rangs observées dans chaque groupe à ce que l'on attendrait si toutes les observations provenaient de la même population. Plus la valeur de H est élevée, plus les groupes diffèrent dans leur distribution.

2.2.2. Hypothèses statistiques

Hypothèse nulle (H_0)	Hypothèse alternative (H_1)
Toutes les populations ont la même distribution (mêmes médianes)	Au moins une population diffère des autres

Tableau 5 : Hypothèses du test de Kruskal-Wallis

2.2.3. Application avec R

```
# Test de Kruskal-Wallis
# Comparaison production glands entre forêts
kruskal.test(production_glands ~ foret, data = df)
```

2.2.4. Tests post-hoc après Kruskal-Wallis

Lorsque le test de Kruskal-Wallis est significatif, il est nécessaire d'effectuer des comparaisons par paires pour identifier les groupes qui diffèrent. Plusieurs méthodes existent, dont le test de Dunn et le test de Nemenyi. Ces tests ajustent les p-values pour tenir compte des comparaisons multiples.

```
# Test post-hoc de Dunn (package Dunn.test)
install.packages("dunn.test")  # Installation si nécessaire
library(dunn.test)
dunn.test(df$production_glands, df$foret, method = "bonferroni")

# Alternative : Test de Nemenyi (package PMCMRplus)
install.packages("PMCMRplus")
library(PMCMRplus)
kwAllPairsNemenyiTest(df$production_glands, df$foret)
```

3. Tests de Corrélation

Les tests de corrélation permettent d'évaluer l'existence et l'intensité d'une relation linéaire entre deux variables quantitatives. En écologie, ces tests sont essentiels pour comprendre les associations entre variables environnementales, comme la relation entre la pluviosité et la croissance des plantes, ou entre la température et la phénologie des espèces. Deux coefficients de corrélation sont couramment utilisés : le coefficient de corrélation de Pearson pour les

données suivant une distribution normale, et le coefficient de corrélation de Spearman pour les données ne suivant pas cette condition.

3.1. Le coefficient de corrélation de Spearman

3.1.1. Principe et définition

Le coefficient de corrélation de Spearman, noté ρ (rho) ou r_s, mesure la force et la direction d'une relation monotone entre deux variables. Développé par Charles Spearman en 1904, ce coefficient non paramétrique travaille sur les rangs des observations plutôt que sur les valeurs brutes. Il est particulièrement adapté lorsque les données ne suivent pas une distribution normale, contiennent des valeurs extrêmes, ou lorsque la relation entre les variables est monotone mais non nécessairement linéaire.

Le coefficient de Spearman varie de -1 à +1. Une valeur de +1 indique une relation monotone croissante parfaite (quand X augmente, Y augmente toujours). Une valeur de -1 indique une relation monotone décroissante parfaite (quand X augmente, Y diminue toujours). Une valeur proche de 0 indique l'absence de relation monotone. Contrairement à Pearson, Spearman peut détecter des relations non linéaires tant qu'elles sont monotones.

3.1.2. Interprétation du coefficient

Valeur de |ρ|	Force de la relation	Interprétation
0,00 - 0,19	Très faible	Relation négligeable
0,20 - 0,39	Faible	Relation légère
0,40 - 0,59	Modérée	Relation substantielle
0,60 - 0,79	Forte	Relation marquée
0,80 - 1,00	Très forte	Relation très marquée

Tableau 6 : Interprétation du coefficient de corrélation de Spearman

3.1.3. Application avec R

```
# Test de corrélation de Spearman
# Relation entre altitude et épaisseur du liège
cor.test(df$altitude, df$epaisseur_liege_2024, method = "spearman")

# Calcul du coefficient de corrélation seul
cor(df$altitude, df$epaisseur_liege_2024, method = "spearman")
```

3.2. Le coefficient de corrélation de Pearson

3.2.1. Principe et définition

Le coefficient de corrélation de Pearson, noté r, mesure la force et la direction d'une relation linéaire entre deux variables quantitatives. Développé par Karl Pearson à la fin du XIXe siècle, c'est le coefficient de corrélation le plus couramment utilisé. Contrairement à Spearman qui détecte les relations monotones, Pearson est spécifiquement conçu pour détecter les relations linéaires et nécessite que les données suivent une distribution normale.

Le coefficient de Pearson représente le rapport entre la covariance des deux variables et le produit de leurs écarts-types. Il varie également de -1 à +1 avec les mêmes interprétations que Spearman concernant le signe. Cependant, il est sensible aux valeurs extrêmes (outliers) qui peuvent fausser considérablement les résultats. Une valeur de r proche de 0 ne signifie pas nécessairement qu'il n'y a pas de relation, mais plutôt qu'il n'y a pas de relation linéaire.

3.2.2. Conditions d'application

Le coefficient de corrélation de Pearson nécessite que plusieurs conditions soient respectées pour produire des résultats valides. Les deux variables doivent être quantitatives et mesurées sur une échelle continue ou discrète. La relation entre les variables doit être linéaire, ce qui peut être vérifié visuellement avec un nuage de points. Les deux variables doivent suivre approximativement une distribution normale, vérifiable avec le test de Shapiro-Wilk. Les observations doivent être indépendantes et les valeurs extrêmes doivent être identifiées et traitées car elles peuvent fortement influencer le coefficient.

3.2.3. Application avec R

```
# Vérification de la normalité des variables
shapiro.test(df$temperature_moy)
shapiro.test(df$epaisseur_liege_2024)


# Test de corrélation de Pearson
# Relation entre température et épaisseur du liège
cor.test(df$temperature_moy, df$epaisseur_liege_2024, method =
"pearson")


# Visualisation avec un nuage de points
plot(df$temperature_moy, df$epaisseur_liege_2024, main = "Corrélation
température-épaisseur", xlab = "Température (°C)", ylab = "Épaisseur
(mm)", pch = 19, col = "blue")
```

3.2.4. Matrice de corrélation

Lorsque vous avez plusieurs variables quantitatives, il peut être utile de calculer une matrice de corrélation qui affiche les coefficients de corrélation pour toutes les paires de variables. Cette matrice permet d'identifier rapidement les relations les plus fortes dans votre jeu de données.

```
# Sélection des variables numériques
vars_num <- df[, c("altitude", "pluie_annuelle", "temperature_moy",
                   "ph_sol", "epaisseur_liege_2024")]"

# Matrice de corrélation de Pearson
cor(vars_num)

# Visualisation avec un graphique (optionnel)
install.packages("corrplot")
library(corrplot)
corrplot(cor(vars_num), method = "circle")
```

4. La Régression Linéaire

La régression linéaire est une méthode statistique qui permet de modéliser la relation entre une variable dépendante (aussi appelée variable réponse ou expliquée) et une ou plusieurs variables indépendantes (aussi appelées variables explicatives ou prédicteurs). Contrairement à la corrélation qui mesure simplement l'association entre variables, la régression permet de quantifier cette relation sous forme d'une équation mathématique et de faire des prédictions.

4.1. Principe de la régression linéaire simple

La régression linéaire simple modélise la relation entre une variable dépendante Y et une seule variable indépendante X. Le modèle cherche la droite qui passe au plus près de tous les points du nuage de points, en minimisant la somme des carrés des écarts verticaux entre les points observés et la droite. Cette droite est appelée droite de régression ou droite des moindres carrés.

L'équation de la droite de régression s'écrit : $Y = \beta_0 + \beta_1 X + \varepsilon$, où Y est la variable dépendante, X est la variable indépendante, β_0 est l'ordonnée à l'origine (intercept), β_1 est la pente de la droite, et ε représente l'erreur aléatoire (résidus). La pente β_1 indique de combien Y change en moyenne lorsque X augmente d'une unité. L'intercept β_0 représente la valeur prédite de Y lorsque $X = 0$.

4.2. Hypothèses du modèle

Pour que les résultats de la régression linéaire soient valides et interprétables, plusieurs hypothèses doivent être vérifiées. La linéarité suppose que la relation entre X et Y est linéaire. L'indépendance des résidus signifie que les erreurs ne sont pas corrélées entre elles. L'homoscédasticité implique que la variance des résidus est constante pour toutes les valeurs de X. La normalité des résidus suppose que les erreurs suivent une distribution normale. Enfin,

l'absence de valeurs extrêmes influentes est nécessaire car certains points peuvent déformer considérablement la droite de régression.

4.3. Application avec R

```
# Régression linéaire simple
# Modélisation : épaisseur du liège en fonction du diamètre
modele <- lm(epaisseur_liege_2024 ~ diametre_chene, data = df)

# Affichage des résultats
summary(modele)

# Coefficients du modèle
coef(modele)
```

4.3.1. Visualisation de la droite de régression

```
# Nuage de points avec droite de régression
plot(df$diametre_chene, df$epaisseur_liege_2024,
     main = "Régression : Épaisseur vs Diamètre",
     xlab = "Diamètre du chêne (cm)",
     ylab = "Épaisseur du liège (mm)",
     pch = 19, col = "steelblue")

# Ajout de la droite de régression
abline(modele, col = "red", lwd = 2)
```

4.3.2. Diagnostic du modèle

La vérification des hypothèses du modèle est essentielle pour valider les résultats de la régression. R fournit des graphiques de diagnostic automatiques qui permettent d'évaluer la linéarité, l'homoscédasticité, la normalité des résidus et l'absence de valeurs influentes.

```
# Graphiques de diagnostic (4 graphiques)
par(mfrow = c(2, 2))  # Affichage en 2x2
plot(modele)
```

4.4. Interprétation des résultats

Le résumé du modèle (summary) fournit plusieurs informations clés. Les coefficients estimés (Estimate) donnent les valeurs de l'intercept et de la pente. L'erreur standard (Std. Error) mesure la précision de l'estimation des coefficients. La statistique t et sa p-value testent si chaque coefficient est significativement différent de zéro. Le coefficient de détermination R^2 indique la proportion de variance expliquée par le modèle. Le R^2 ajusté prend en compte le nombre de

prédicteurs et est plus adapté pour comparer des modèles. La statistique F et sa p-value testent la significativité globale du modèle.

Le R^2 varie de 0 à 1. Un R^2 de 0,75 signifie que 75% de la variabilité de la variable dépendante est expliquée par le modèle. Un R^2 élevé ne garantit pas que le modèle est approprié ; il faut toujours vérifier les hypothèses et examiner les résidus.

4.5. Prédictions avec le modèle

L'un des principaux avantages de la régression est la possibilité de faire des prédictions pour de nouvelles valeurs de la variable indépendante. La fonction `predict()` permet d'obtenir les valeurs prédites ainsi que des intervalles de confiance.

```
# Prédiction pour de nouvelles valeurs
nouveaux_diametres <- data.frame(diametre_chene = c(40, 50, 60, 70))

# Prédictions avec intervalle de confiance
predictions <- predict(modele, newdata = nouveaux_diametres,
                       interval = "confidence")

# Affichage des prédictions
print(predictions)
```

Résumé et Guide de Choix des Tests

Ce chapitre vous a présenté les principaux tests statistiques utilisés en écologie expérimentale. Le choix du test approprié dépend des caractéristiques de vos données et de votre question de recherche. Le tableau suivant récapitule les situations dans lesquelles utiliser chaque test.

Objectif	Données normales	Données non normales	Fonction R
Comparer 2 groupes	Test t	Mann-Whitney	`t.test()`, `wilcox.test()`
Comparer 3+ groupes	ANOVA + Tukey	Kruskal-Wallis + Dunn	`aov()`, `kruskal.test()`
Corrélation 2 variables	Pearson	Spearman	`cor.test()`
Modéliser relation	Régression linéaire	Régression non param.	`lm()`

Tableau 7 : Guide de choix des tests statistiques

Workflow recommandé pour l'analyse statistique

1. Définissez clairement votre question de recherche et vos hypothèses.
2. Explorez vos données avec des statistiques descriptives et des visualisations.
3. Vérifiez les conditions d'application des tests (normalité, homoscédasticité).
4. Choisissez le test approprié selon vos données et votre question.
5. Appliquez le test et interprétez les résultats avec prudence.
6. Effectuez les tests post-hoc si nécessaire.
7. Rapportez vos résultats de manière complète et transparente.

En suivant cette méthodologie rigoureuse, vous serez en mesure de mener des analyses statistiques solides et reproductibles en écologie expérimentale. N'oubliez jamais que la statistique est un outil au service de la biologie : l'interprétation écologique de vos résultats est aussi importante que l'analyse statistique elle-même.