TP2

Tests Préliminaires et Compréhension des Données

Durée : 3 heures

Analyse statistique des données écologiques et forestières

Évaluation de la distribution, normalité et homogénéité des variances

Table des matières

Objectif principal

Comprendre la structure statistique des données écologiques et forestières avant d'appliquer des tests d'hypothèses. Vous apprendrez à évaluer la distribution, la normalité et l'homogénéité des variances.

Préparation

Ouvrez RStudio, définissez votre répertoire de travail et chargez les données avec le code suivant :

```
# Chargement des données
df <- read.csv2("Mila.csv")


# Chargement des packages nécessaires
# Si non installés : install.packages(c("ggplot2", "car"))
library(ggplot2)
library(car)
```

1. Types de distribution

Avant de faire des calculs, il est crucial de visualiser comment les valeurs d'une variable sont réparties. La forme de la distribution nous renseigne sur la nature des données et oriente le choix des tests statistiques appropriés.

1.1 Les principaux types de distributions

Distribution normale (Gaussienne) : En forme de cloche, symétrique. La majorité des valeurs se concentrent autour de la moyenne. C'est la distribution idéale pour de nombreux tests statistiques paramétriques.

Distributions asymétriques (skewed) : La "queue" de la distribution s'étire vers la gauche ou la droite. C'est souvent le cas pour des variables comme l'âge ou les revenus où les valeurs extrêmes étirent la distribution dans une direction.

Distribution bimodale : Présente deux "pics" distincts, ce qui suggère souvent la présence de deux sous-populations mélangées dans l'échantillon étudié.

1.2 Exercice Pratique : Visualisation

Nous allons observer la distribution de la hauteur_totale et du diametre_chene à l'aide de différentes méthodes graphiques.

```
# 1. Histogramme classique (Base R)
   hist(df$hauteur_totale, main = "Distribution de la hauteur totale",
        xlab = "Hauteur (m)", ylab = "Fréquence", col = "lightblue")

# 2. Courbe de densité (Base R)
   plot(density(df$diametre_chene), main = "Densité du diamètre des
   chênes",
        lwd = 2, col = "darkgreen")
```

2. Calcul de statistiques descriptives

Les statistiques descriptives résument l'information contenue dans vos variables quantitatives. Elles permettent de caractériser rapidement un jeu de données et d'identifier d'éventuelles anomalies.

2.1 Mesures de tendance centrale et de dispersion

Tendance centrale : Indique où se situe le "centre" des données. Les mesures principales sont la moyenne et la médiane. La médiane est particulièrement utile car elle est robuste aux valeurs extrêmes.

Dispersion : Indique à quel point les données sont étalées autour du centre. Les mesures principales incluent la variance, l'écart-type et l'écart interquartile (IQR).

2.2 Exercice Pratique : Calculs sur l'épaisseur du liège

```
# Sélection de notre variable d'intérêt
   liege_24 <- df$epaisseur_liege_2024

# Mesures de tendance centrale
   mean(liege_24)      # Moyenne
   median(liege_24)    # Médiane (robuste aux valeurs extrêmes)

# Mesures de dispersion
   var(liege_24)       # Variance
   sd(liege_24)        # Écart-type (Standard Deviation)
   IQR(liege_24)       # Écart interquartile (Q3 - Q1)
   quantile(liege_24)  # Les quartiles (0%, 25%, 50%, 75%, 100%)

# Astuce pour tout avoir d'un coup
   summary(df$epaisseur_liege_2024)
```

> *Note pour les étudiants : Observez la différence entre la moyenne et la médiane. Si elles sont très éloignées, votre distribution est probablement asymétrique !*

3. Normalité et Homogénéité : Concepts et Importance

La majorité des tests statistiques classiques (ANOVA, Test T, Régression linéaire) sont dits paramétriques. Ils reposent sur deux hypothèses fondamentales qu'il est essentiel de vérifier avant toute analyse.

3.1 Les deux hypothèses fondamentales

1. La Normalité : Les données (ou les résidus du modèle) suivent une distribution normale. Cette hypothèse assure la validité des estimations et des tests d'hypothèses.
2. L'Homogénéité des variances (Homoscédasticité) : La dispersion des données est similaire entre les différents groupes comparés. Cette hypothèse garantit la comparabilité des groupes.

3.2 Conséquences du non-respect des hypothèses

Si vous appliquez un test paramétrique sur des données qui ne respectent pas ces hypothèses, vous risquez des erreurs importantes dans vos conclusions. Vous pourriez conclure à tort qu'il y a une différence significative (Faux positif ou erreur de Type I), ou au contraire manquer une différence réelle (Faux négatif ou erreur de Type II). Dans les deux cas, le test perd sa validité et vos conclusions peuvent être erronées.

4. Choix entre tests paramétriques et non paramétriques

Comment savoir quel test appliquer pour répondre à une question écologique ? Voici la règle d'or pour guider votre choix méthodologique.

4.1 Tableau de décision

Hypothèses validées ?	Type de test à choisir	Exemple de test
Normal + Homogène	Paramétrique (Plus puissant)	Test T de Student, ANOVA
Non normal OU Non homogène	Non paramétrique (Basé sur les rangs)	Mann-Whitney, Kruskal-Wallis

Tableau 1 : Guide de choix entre tests paramétriques et non paramétriques

4.2 Exemple concret en écologie

Question : La production de glands est-elle différente selon que l'arbre a subi des dégâts de feu (degats_feu) ?

Démarche : On vérifie d'abord la normalité de production_glands dans chaque groupe (Vrai/Faux), puis on teste l'homogénéité des variances entre les deux groupes. Ces

vérifications préliminaires orienteront le choix vers un test paramétrique ou non paramétrique.

5. Application des tests préliminaires

C'est ici que nous vérifions mathématiquement nos hypothèses à l'aide de la valeur p (p-value). La p-value représente la probabilité d'obtenir les résultats observés si l'hypothèse nulle est vraie.

> **Règle générale : Si p-value < 0.05, on rejette l'hypothèse nulle du test.**

5.1 Test de Normalité : Shapiro-Wilk

Hypothèse Nulle (H0) : La distribution est normale.

Interprétation : Si p > 0.05, la distribution est considérée normale (on ne rejette pas H0). Si p < 0.05, la distribution s'écarte significativement de la normalité.

```
# Testons la normalité de la hauteur des arbres
   shapiro.test(df$hauteur_totale)
```

5.2 Test d'Homogénéité des variances : Test de Levene

Hypothèse Nulle (H0) : Les variances sont égales (homogènes) entre les groupes.

Interprétation : Si p > 0.05, les variances sont homogènes (on ne rejette pas H0). Si p < 0.05, les variances diffèrent significativement entre les groupes.

```
# La variance de l'accroissement du liège est-elle la même selon la forêt ?
# Nécessite le package 'car'
   leveneTest(accroissement_liege ~ as.factor(foret), data = df)
```

6. Exercice de synthèse

Cet exercice vous permet de mettre en pratique l'ensemble des concepts abordés dans ce TP. Il constitue une vérification complète des hypothèses statistiques.

1. Extrayez la variable **production_glands.**
2. **Tracez son histogramme.**
3. **Effectuez le test de Shapiro-Wilk. La variable est-elle normale ?**
4. Effectuez un **test de Levene** pour voir si la variance de la **production_glands** est **homogène selon les degats_feu.**
5. Conclusion : **Quel type de test (paramétrique ou non) devriez-vous utiliser pour comparer la production de glands entre les arbres brûlés et non brûlés ?**

> *Conseil méthodologique : Documentez systématiquement vos vérifications d'hypothèses dans vos rapports d'analyse. Cette rigueur méthodologique renforce la crédibilité de vos conclusions et permet à d'autres chercheurs de reproduire votre démarche.*