

Travaux Dirigés n°2

Représentation graphiques et Mesures de Tendances Centrales

Modules M1 - M2 | Durée : 1h30 | Documents autorisés : cours

Dr. Anicet EBOU, ediman.ebou@inphb.ci

Objectifs du TD

- Produire et interpréter des graphiques adaptés au type de variable
- Calculer et comparer les trois moyennes (arithmétique, géométrique, harmonique), la médiane et le mode sur des données réelles
- Choisir la mesure de tendance centrale la plus appropriée selon le contexte
- Écrire du code R propre, reproductible et commenté

Règles du jeu Vous travaillez dans un seul script `.R` que vous remettrez à la fin de la séance (fichier nommé `TD2_NOM_PRENOM.R`). Chaque ligne de code doit être **commentée** (`#`). Vous pouvez utiliser `?fonction` dans la console pour lire l'aide.

Données : plantation d'anacardiers à Korhogo

Un technicien agricole a relevé les mesures suivantes sur **40 anacardiers** (*Anacardium occidentale*) dans une plantation du nord de la Côte d'Ivoire, cinq ans après leur plantation.

Copiez ce bloc dans votre script et exécutez-le pour créer les données :

Listing 1 – Jeu de données - anacardiers de Korhogo

```

1 # Hauteur des arbres (en metres)
2 hauteur <- c(3.2, 4.1, 2.8, 5.0, 3.7, 4.4, 3.1, 6.2, 4.8, 3.5, 4.0,
3             2.6, 5.5, 3.9, 4.2, 3.3, 4.7, 5.1, 3.6, 4.9, 2.9, 4.5, 3.8,
4             5.3, 4.0, 3.4, 4.6, 3.0, 5.8, 4.3, 3.7, 4.1, 2.7, 4.8, 3.5, 5.0,
5             4.2, 3.9, 6.5, 4.4)
6
7 # Rendement en noix de cajou (kg/arbre/an)
8 rendement <- c(8.2, 12.5, 6.1, 15.0, 9.8, 11.3, 7.5, 18.4, 14.2,
9               9.0, 10.5, 5.8, 16.7, 10.2, 11.8, 8.7, 13.5, 15.3, 9.4, 14.8,
10              6.9, 13.1, 10.0, 16.1, 10.5, 8.9, 13.8, 7.2, 17.5, 12.0, 9.8,
11              11.2, 6.3, 14.5, 9.1, 15.2, 11.7, 10.4, 19.2, 12.8)
12
13 # Etat sanitaire (variable qualitative ordinale)
14 sante <- c("bon","bon","moyen","excellent","bon","bon","moyen","
15            excellent", "bon","moyen","bon","mauvais","excellent","bon","bon",
16            "moyen", "excellent","excellent","bon","bon","mauvais","bon","
17            bon", "excellent","bon","moyen","bon","mauvais","excellent","bon",
18            "bon","bon","mauvais","bon","moyen","excellent","bon","bon",
19            "excellent","bon")
20 sante <- factor(sante,
21               levels = c("mauvais","moyen","bon","excellent"),
22               ordered = TRUE)
23
24 # Taux de croissance annuel moyen (%)
25 taux_croissance <- c(12.5, 18.3, 10.2, 22.1, 15.4, 17.8, 11.6,
26                    25.3, 20.1, 14.2, 16.0, 9.8, 23.7, 15.9, 18.2, 13.1, 19.5, 21.4,
27                    14.8, 20.8, 10.9, 19.0, 15.2, 22.8, 16.0, 13.5, 19.8, 11.2,
28                    24.5, 17.3, 14.9, 17.1, 10.1, 20.3, 14.0, 21.7, 17.9, 15.8,
29                    26.0, 18.5)
30
31 # Assembler en data.frame
32 anacardiers <- data.frame(hauteur, rendement, sante, taux_
33                           croissance)

```

Vérification rapide. Après avoir exécuté le code, tapez `str(anacardiers)` dans la console. Vous devez voir 40 observations et 4 variables. Si vous obtenez une erreur, vérifiez que vous avez copié *tout* le bloc.

Exercice 1 : Explorer avant de calculer (25 min)

Contexte. Avant toute analyse, un bon statisticien *regarde* ses données.

1.1. Utilisez `summary(anacardiers)` et complétez sur une feuille le tableau suivant :

	Hauteur (m)	Rendement (kg)	Taux de croissance (%)
Minimum			
Maximum			
Médiane			
Moyenne			

1.2. Combien d'arbres sont dans chaque catégorie d'état sanitaire? Utilisez la fonction `table()` sur la variable `sante`. Quelle est la catégorie **modale**?

1.3. Produisez un **histogramme** de la hauteur avec **8 classes**. Ajoutez un titre et des étiquettes d'axes en français. *Indice* : `hist(x, breaks = 8, main = "...", xlab = "...", col = "...")`

1.4. Produisez un **diagramme en barres** de l'état sanitaire, en respectant l'ordre des niveaux (*mauvais* → *excellent*).
Indice : la variable `sante` est déjà ordonnée ; `barplot(table(sante))`.

1.5. **Question d'interprétation** : En regardant les deux graphiques produits, quelle première impression avez-vous sur la santé de cette plantation? Répondez en **deux phrases** dans un commentaire R (#).

Exercice 2 : Trois moyennes, trois contextes (30 min)

Contexte. La *plantation voisine* n'a que 5 arbres témoins, mesurés plus précisément. Ces 5 arbres serviront à illustrer les différences entre les trois moyennes.

Listing 2 – Arbres témoins (5 individus)

```

1 # Hauteurs des 5 arbres temoins (m)
2 h_temoins <- c(2.1, 3.4, 4.8, 5.2, 8.1)
3
4 # Taux de croissance annuel moyen sur 5 ans (%)
5 # Ex : un arbre qui grandit de 12% la 1re annee, puis 15%, etc.
6 taux_temoins <- c(12.5, 15.8, 18.3, 14.1, 22.6)
7
8 # Vitesse de trajet entre deux parcelles (km/h) pour 3 passages
9 # (utile pour planifier le travail du technicien)
10 vitesse_passages <- c(20, 30, 60)

```

2.1. Calculez la **moyenne arithmétique** des hauteurs `h_temoins` avec `mean()`. Vérifiez à la main en utilisant `sum(h_temoins) / length(h_temoins)`. Les deux résultats doivent être identiques.

2.2. Calculez la **moyenne géométrique** des taux de croissance `taux_temoins`. R ne dispose pas d'une fonction directe : utilisez la formule $G = \exp\left(\frac{1}{n} \sum_{i=1}^n \ln(x_i)\right)$.

```
1 # Moyenne geometrique, formule via logarithme
2 G <- exp(mean(log(taux_temoins)))
```

- Comparez G avec `mean(taux_temoins)`. Laquelle est plus petite ? Pourquoi est-ce toujours le cas ?
- **Interprétation** : si un arbre croît selon ces taux pendant 5 ans, à quel taux constant aurait-il le même résultat final ?

2.3. Calculez la **moyenne harmonique** des vitesses `vitesse_passages`. Utilisez la formule $H = n / \sum_{i=1}^n (1/x_i)$.

```
1 # Moyenne harmonique
2 H <- length(vitesse_passages) / sum(1 / vitesse_passages)
```

- Le technicien parcourt une distance de 30 km à chacun des 3 passages. Calculez le **temps total** pour les 3 aller-retours, puis la vitesse moyenne réelle. Cette vitesse est-elle égale à H , à G ou à `mean(vitesse_passages)` ?
- Expliquez en une phrase pourquoi la moyenne arithmétique serait ici trompeuse.

2.4. Bilan comparatif. Remplissez ce tableau dans votre script sous forme de commentaires, puis justifiez chaque choix de mesure :

Situation	Mesure à utiliser	Pourquoi ?
Hauteur moyenne des 40 anacardiens		
Taux de croissance moyen sur 5 ans		
Vitesse moyenne du technicien		

Exercice 3 : Médiane, mode et distributions asymétriques (35 min)

Contexte. Deux parcelles supplémentaires ont été inventoriées dans la même région. Les rendements sont très différents selon la parcelle.

Listing 3 – Rendements sur deux autres parcelles

```

1 # Parcelle A: rendements relativement homogenes (kg/arbre/an)
2 parcA <- c(9.5, 10.2, 10.8, 9.1, 10.5, 11.0, 9.8, 10.3, 10.6, 9.7)
3
4 # Parcelle B: un arbre exceptionnel tire les rendements vers
5 # le haut
6 parcB <- c(9.5, 10.2, 10.8, 9.1, 10.5, 11.0, 9.8, 10.3, 10.6, 38.4)

```

3.1. Calculez la **moyenne** et la **médiane** pour `parcA` et pour `parcB`. Organisez vos résultats dans un commentaire R sous cette forme :

```

1 # Parcelle A : moyenne = ?   mediane = ?
2 # Parcelle B : moyenne = ?   mediane = ?

```

3.2. La valeur 38,4 correspond à un arbre greffé avec une variété améliorée.

- De combien la moyenne change-t-elle entre `parcA` et `parcB` ?
- De combien la médiane change-t-elle ?
- Qu'est-ce que cela vous apprend sur la **robustesse** de chaque mesure face aux valeurs extrêmes ?

3.3. Produisez des **boîtes à moustaches** (*box plots*) côte à côte pour comparer les deux parcelles. *Indice :*

```

1 boxplot(
2   parcA, parcB,
3   names = c("Parcelle A", "Parcelle B"),
4   ylab = "Rendement (kg/arbre/an)",
5   main = "Comparaison des deux parcelles",
6   col = c("lightgreen", "lightyellow")
7 )

```

- La valeur 38,4 apparaît-elle sur le graphique ? Sous quel symbole ? Qu'appelle-t-on ce type de point ?
- La boîte de la parcelle B est-elle symétrique ? Que cela indique-t-il sur la forme de la distribution ?

3.4. Sur la variable `sante` du jeu de données principal, déterminez le **mode** (catégorie la plus fréquente). R n'a pas de fonction `mode()` pour les statistiques (elle fait autre chose !). Utilisez `table()` et `which.max()` :

```

1 freq_sante <- table(anacardiers$sante)
2 mode_sante <- names(which.max(freq_sante))

```

Interprétez : que dit ce mode sur l'état général de la plantation ?

3.5. Défi final : à ne traiter que si vous avez terminé les autres questions.

Construisez un **graphique composite** (deux panneaux côte à côte) montrant l'histogramme du rendement de la parcelle principale (40 arbres) et la boîte à moustaches du même rendement, avec une ligne verticale rouge pour la moyenne et une ligne bleue pour la médiane sur l'histogramme.

```
1 par(mfrow = c(1, 2))    # 2 panneaux cote a cote
2
3 # Panneau 1 : histogramme du rendement
4 hist(
5   anacardiers$rendement, breaks = 8,
6   col = "lightgreen", border = "white",
7   main = "Distribution du rendement",
8   xlab = "Rendement (kg/arbre/an)"
9 )
10 abline(v = mean(anacardiers$rendement), col = "red", lwd = 2)
11 abline(v = median(anacardiers$rendement), col = "blue", lwd = 2)
12 legend(
13   "topright",
14   legend = c("Moyenne", "Mediane"),
15   col = c("red", "blue"), lwd = 2
16 )
17
18 # Panneau 2 : boite a moustaches
19 boxplot(
20   anacardiers$rendement,
21   col = "lightyellow",
22   main = "Boite a moustaches",
23   ylab = "Rendement (kg/arbre/an)"
24 )
25
26 par(mfrow = c(1, 1))    # retour a un seul panneau
```

- La moyenne et la médiane se superposent-elles sur l'histogramme ?
- La distribution est-elle symétrique, étalée à gauche ou à droite ?
- Quelle mesure recommanderiez-vous au technicien pour résumer le rendement typique de sa plantation ? Justifiez.

À remettre à la fin de la séance

Fichier à remettre : TD2_NOM_PRENOM.R

Ce script doit contenir :

- Un en-tête commenté (noms, date, TD2)
- Tout le code qui produit vos réponses, commenté
- Les réponses aux questions d'interprétation **en commentaires R** (lignes commençant par #)
- Des graphiques exportables (ils doivent s'afficher quand on lance le script)

Un script qui ne s'exécute pas sans erreur ne sera pas noté.