

TP n°1

stoehr@ceremade.dauphine.fr

Objectifs Ce TP doit permettre d'apprendre à exécuter et à conserver des commandes dans un script, à importer un fichier de données, à examiner son contenu et à produire des représentations graphiques simples. La dernière partie illustre sur des données historiques, puis par simulation, les notions relatives à un test statistique.

Bien démarrer et organiser son travail Le dossier distribué est organisé comme un projet R. Il s'agit de la bonne structure pour travailler! Toutes les commandes seront alors exécutées relativement à la racine du projet. Un projet utilise un chemin relatif qui rend le projet facilement transportable d'un ordinateur à un autre.

1. Après avoir décompressé le dossier, ouvrir le fichier `Statistical_Models.Rproj` avec RStudio. Il ne faudra ni déplacer le fichier `mais.txt`, ni utiliser la fonction `setwd`.
2. Créer dans le dossier R un script nommé `TP1.R`. Les commandes importantes doivent être conservées dans un script plutôt qu'exécutées uniquement dans la console, afin que l'analyse puisse être reproduite. Une ligne, ou une sélection de lignes, est exécutée avec `Ctrl + Entrée` sous Windows et Linux, ou `Cmd + Entrée` sous macOS.
3. S'il n'est pas déjà disponible, installer le package `ggplot2`

```
install.packages("ggplot2")
```

4. Charger le package `ggplot2`

```
library(ggplot2)
```

Exercice 1.

1. Exécuter les commandes suivantes depuis le script. Que renvoient-elles?

```
getwd()  
list.files()  
list.files("data")
```

2. Créer le vecteur suivant. La flèche `<-` réalise une affectation. Utiliser les fonctions `length`, `mean`, `sd`, `median`, `min` et `max` pour décrire `x`.

```
x <- c(12, 15, 9, 18, 11, 14)
```

3. Comparer les résultats des commandes suivantes et expliquer le rôle des crochets.

```
x[1]  
x[2:4]  
x[x >= 14]
```

4. Consulter l'aide de la fonction `mean` à l'aide de la commande `?mean`. Repérer l'argument permettant d'ignorer d'éventuelles valeurs manquantes. Enregistrer le script.

Exercice 2. Le fichier `mais.txt` décrit 288 parcelles expérimentales. Pour chacune d'elles, on dispose d'un identifiant, d'un marqueur génétique, d'une variété, du rendement de la parcelle ainsi que de différentes caractéristiques des grains et des plants.

1. Importer les données à l'aide de la commande suivante. Expliquer le rôle des arguments `header` et `sep`.

```
mais <- read.table("data/mais.txt", header = TRUE, sep = "\t")
```

Utiliser ensuite `class`, `dim`, `names`, `str` et `summary` pour examiner l'objet obtenu. Vérifier le nombre d'observations, le nombre de variables et la présence éventuelle de valeurs manquantes.

2. Les variables `Marqueur` et `Variete` sont qualitatives. Les convertir en facteurs à l'aide de la fonction `factor`, puis utiliser de nouveau `str`. Calculer les effectifs des différentes variétés.
3. Afficher les colonnes `id`, `Variete` et `Rendement` pour les cinq premières parcelles. Sélectionner ensuite les parcelles de variété `Tropical`. Calculer la moyenne, l'écart-type et les quartiles du rendement sur l'ensemble des parcelles.

4. (a) Réaliser en R de base :

1. un histogramme du rendement ;
2. un diagramme en barres des effectifs des variétés ;
3. des boîtes à moustaches du rendement selon la variété ;
4. un nuage de points représentant le nombre de feuilles en fonction de la floraison.

Donner à chaque graphique un titre et des légendes d'axes informatives.

- (b) Décrire brièvement les principales informations apportées par chaque représentation et calculer la corrélation empirique entre `Floraison` et `Feuilles`.

5. (*Bonus*) Reproduire les quatre graphiques précédents avec le package `ggplot2`.

Exercice 3. Le jeu de données `morley`, fourni avec R, contient 100 mesures réalisées par Albert A. Michelson en 1879 afin d'estimer la vitesse de la lumière. Les mesures sont réparties en cinq expériences de vingt essais. La variable `Speed` donne la vitesse mesurée, après soustraction de 299 000 km/s. La valeur de référence actuelle utilisée dans cet exercice est $\mu_0 = 299\,792,458$ km/s.

1. Charger les données à l'aide la commande

```
data(morley)
```

Examiner les avec les outils de l'Exercice 2. Transformer `Expt` en facteur et créer la variable `Vitesse` exprimée en km/s. Calculer la moyenne et l'écart-type des mesures pour chaque expérience.

Test de Student. On modélise provisoirement les mesures comme des variables indépendantes de loi normale $\mathcal{N}(\mu, \sigma^2)$, où μ et σ^2 sont inconnus. On souhaite tester

$$\mathcal{H}_0 : \mu = 299\,792,458 \quad \text{contre} \quad \mathcal{H}_1 : \mu \neq 299\,792,458.$$

Sous l'hypothèse nulle et lorsque les observations sont indépendantes de loi normale, la statistique de test

$$T_n = \frac{\sqrt{n}(\bar{Y}_n - \mu_0)}{S_n}, \quad S_n^2 = \frac{1}{n-1} \sum_{i=1}^n (Y_i - \bar{Y}_n)^2,$$

suit une loi de Student à $n - 1$ degrés de liberté. Pour un niveau $\alpha \in]0, 1[$, la zone de rejet du test bilatéral est

$$\mathcal{R}_n = \left\{ |T_n| > q_{1-\alpha/2}^{t(n-1)} \right\},$$

où $q_{1-\alpha/2}^{t(n-1)}$ désigne le quantile d'ordre $1 - \alpha/2$ de la loi de Student à $n - 1$ degrés de liberté.

2. Effectuer le test de Student bilatéral avec `t.test`, au niveau 5%. Identifier dans la sortie la statistique de test, le nombre de degrés de liberté, la p -valeur, l'intervalle de confiance et l'estimation de μ . Quelle décision prend-on ? Expliquer pourquoi cette conclusion doit être interprétée avec prudence.

Illustration sur données simulées. On étudie maintenant le comportement du test dans le cadre contrôlé de données synthétiques.

3. Prendre $n = 20$, $\sigma = 80$ km/s et simuler un échantillon sous $\mathcal{H}_0$ en utilisant la *seed* 2026. Calculer à la main la statistique et sa p -valeur bilatérale, puis vérifier les résultats avec `t.test`.
4. (a) Considérer les tailles d'échantillon $n \in \{10, 20, 50, 100\}$ et la grille de valeurs

$$\mathcal{G} = \{\mu_0 - 80, \mu_0 - 40, \mu_0 - 20, \mu_0, \mu_0 + 20, \mu_0 + 40, \mu_0 + 80\}.$$

Pour chaque n et chaque $\mu \in \mathcal{G}$, estimer la probabilité de rejet du test de Student bilatéral au niveau $\alpha = 5\%$ en simulant $B = 1\,000$ échantillons de taille n suivant la loi $\mathcal{N}(\mu, \sigma^2)$, avec $\sigma = 80$ km/s.

- (b) Pour $\mu = \mu_0$, comparer cette probabilité au niveau nominal du test. Pour $\mu \neq \mu_0$, les probabilités obtenues fournissent des estimations de la fonction puissance

$$\mu \mapsto \pi_n(\mu) = \mathbb{P}_{\mu, \sigma}(\text{rejeter } \mathcal{H}_0), \quad \mu \neq \mu_0,$$

avec $\sigma = 80$ km/s. Représenter les résultats en fonction de $\mu - \mu_0$, en traçant une courbe pour chaque valeur de n . Commenter l'influence de la taille de l'échantillon et de l'écart entre μ et μ_0 .

Pour aller plus loin.

5. Reproduire aussi fidèlement que possible le graphique suivant en R. La fonction `stripchart` permet d'ajouter des points légèrement décalés horizontalement et la fonction `adjustcolor` de modifier leur transparence. Reproduire ensuite le même graphique avec `ggplot2`. On pourra utiliser `geom_boxplot`, `geom_point`, `position_jitter` et `geom_hline`.

Mesures de Michelson en 1879