Estimation de paramètres et tests d'hypothèse

18 septembre 2019

Dans ce laboratoire, vous appliquerez les concepts vus lors des deux derniers cours.

1. Caractéristiques des fleurs de trois espèces d'iris

Pour cet exercice, nous utiliserons le célèbre jeu de données des iris d'Edgar Anderson, qui contient différentes mesures (en cm) prises sur 50 fleurs de 3 espèces d'iris. Ce tableau de données est déjà chargé dans R sous le nom `iris`.

```
data(iris)
head(iris)
```

```
##   Sepal.Length Sepal.Width Petal.Length Petal.Width Species
## 1          5.1         3.5          1.4         0.2  setosa
## 2          4.9         3.0          1.4         0.2  setosa
## 3          4.7         3.2          1.3         0.2  setosa
## 4          4.6         3.1          1.5         0.2  setosa
## 5          5.0         3.6          1.4         0.2  setosa
## 6          5.4         3.9          1.7         0.4  setosa
```

a) Nous allons d'abord visualiser une partie des données. À l'aide de *ggplot2*, produisez un nuage de points reliant la longueur et la largeur des sépales (`Sepal.Length` et `Sepal.Width`), en différenciant les points de chaque espèce par couleur.

Comment procéderiez-vous pour calculer la moyenne de `Sepal.Width` pour l'espèce *setosa* et son intervalle de confiance à 95%?

b) De quelles quantités avez-vous besoin pour ce calcul?

c) Déterminez la moyenne de `Sepal.Width` pour l'espèce *setosa* et son erreur-type.

d) Durant le cours sur les distributions statistiques, nous avons vu les fonctions `rnorm`, `dnorm`, `pnorm` et `qnorm` qui permettent de calculer des valeurs à partir de la distribution normale. Des fonctions similaires existent pour la distribution t (`rt`, `dt`, `pt`, `qt`).

Quelle fonction nous permet d'obtenir la valeur de la variable t pour une probabilité cumulative p donnée?

Utilisons la fonction `qt(p, df)` pour déterminer l'intervalle correspondant à 95% de la probabilité. Quelles valeurs de p (la probabilité cumulative) utiliser? Quel nombre de degrés de liberté (df) en fonction de la taille de l'échantillon n?

e) Calculez l'intervalle de confiance à 95% pour la moyenne calculée en (c). Comment interprétez-vous cet intervalle?

2. Choix des méthodes d'échantillonnage

Vous souhaitez déterminer la moyenne et l'écart-type du taux de croissance des semis de bouleau jaune dans une région donnée. Votre unité d'échantillonnage est un quadrat de 1 m^2 où la croissance annuelle des semis est mesurée. Quelle méthode d'échantilonnage préconiserez-vous pour le placement de ces quadrats selon les différents scénarios présentés, et pourquoi? Vous pouvez choisir un échantilonnage stratifié, par grappe, systématique, ou adaptatif.

a) Les peuplements de cette espèce dans la région sont relativement semblables mais très éloignés l'un de l'autre.

b) La croissance pourrait être fortement influencée par le gradient de température nord-sud à l'échelle considérée.

c) L'espèce se retrouve dans plusieurs types de peuplement distincts couvrant des proportions différentes du territoire (ex.: 70% type A, 25% type B, 5% type C).

d) C'est une espèce rare dans la région et sa distibution est peu connue.

3. Simulation d'échantillonnage stratifié

Pour cet exercice, nous comparerons l'échantillonnage simple et stratifié à partir d'échantillons simulés du tableau de données `iris`.

a) Créez un graphique de boîtes à moustaches montrant la distribution des 50 longueurs de pétales (*Petal.Length*) pour chaque espèce.

b) Créez deux fonctions `iris_alea` et `iris_strat`. La première fonction choisit 30 observations au hasard d'`iris`, puis retourne la moyenne de *Petal.Length* pour ces observations. La deuxième choisit 10 observations au hasard de chacune des trois espèces, plus retourne la moyenne de *Petal.Length* pour les 30 observations.

Notes:

- La fonction `sample(x, size)` dans R simule l'échantillonnage d'un nombre de valeurs égal à *size* parmi celles du vecteur *x*.
- Vous pouvez écrire ces fonctions sans arguments (parenthèses vides après `function`), comme dans l'exemple ci-dessous.

```
iris_alea <- function() {
    # Insérer code de la fonction ici
}
```

c) Générez un vecteur 1000 résultats de chaque fonction avec `replicate`, comme suit:

```
rep_alea <- replicate(1000, iris_alea())
rep_strat <- replicate(1000, iris_strat())
```

Notez qu'il est important d'inclure les parenthèses vides pour appeler la fonction.

Calculez l'erreur-type de chaque moyenne (à partir des écarts-type de `rep_alea` et `rep_strat`). Avant de faire le calcul, pouvez-vous deviner quelle méthode sera la plus précise? Pour quelle raison?

4. Concentration d'ozone dans trois jardins

Pour cet exercice, nous utiliserons le tableau de données `gardens.csv`, qui provient du manuel *Statistics: An Introduction Using R* de Michael Crawley. Ces données représentent les concentrations d'ozone (en parties par 100 millions ou pphm) mesurées dans trois jardins (A, B et C) lors de différentes journées.

```
gardens <- read.csv("gardens.csv")
```

a) Observez la distribution des mesures d'ozone dans chaque jardin et indiquez la moyenne. Quel type de graphique pourriez-vous utiliser?

b) Calculez la moyenne et l'écart-type de la concentration d'ozone dans chaque jardin. Que remarquez-vous? Est-ce que la moyenne est une bonne indicatrice de la valeur "typique" dans chaque jardin?

Note: La fonction `tapply(X, INDEX, FUN)` applique la fonction *FUN* à *X* pour chaque valeur du facteur *INDEX*. Donc la moyenne de la concentration d'ozone par jardin peut être calculée avec `tapply(gardens$Ozone, gardens$Garden, mean)`.

c) À partir de ces données, testez l'hypothèse nulle selon laquelle les jardins A et B reçoivent la même concentration d'ozone en moyenne. Quel est votre estimé de la différence entre les moyennes et son intervalle de confiance à 99%? Est-ce que ce test donne une bonne idée de la différence entre les deux jardins? Expliquez votre réponse.

d) Répétez l'exercice précédent pour l'hypothèse nulle selon laquelle les jardins A et C reçoivent la même concentration d'ozone en moyenne. Commentez sur la différence entre ce résultat et le résultat précédent.