Tests de randomisation et bootstrap

Données

Ce laboratoire utilise la base de données Portal, qui contient des données de suivi à long terme de plusieurs espèces de rongeurs sur un site d'étude en Arizona.

> Ernest, M., Brown, J., Valone, T. and White, E.P. (2018) *Portal Project Teaching Database*. https://figshare.com/articles/Portal_Project_Teaching_Database/1314459.

Le tableau de données `portal_surveys.csv` contient une rangée par individu capturé. Les variables incluent la date (jour, mois, année), le numéro de parcelle, le code d'espèce, le sexe, la longueur de patte arrière et le poids des individus.

```
surveys <- read.csv("../donnees/portal_surveys.csv", stringsAsFactors = FALSE)
str(surveys)
```

```
## 'data.frame':    35549 obs. of  9 variables:
##  $ record_id      : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ month          : int  7 7 7 7 7 7 7 7 7 7 ...
##  $ day            : int  16 16 16 16 16 16 16 16 16 16 ...
##  $ year           : int  1977 1977 1977 1977 1977 1977 1977 1977 1977 1977 ...
##  $ plot_id        : int  2 3 2 7 3 1 2 1 1 6 ...
##  $ species_id     : chr  "NL" "NL" "DM" "DM" ...
##  $ sex            : chr  "M" "M" "F" "M" ...
##  $ hindfoot_length: int  32 33 37 36 35 14 NA 37 34 20 ...
##  $ weight         : int  NA NA NA NA NA NA NA NA NA NA ...
```

Le tableau de données `portal_plots.csv` indique le type de traitement appliqué à chaque parcelle. Les traitements visent à exclure différents types de rongeurs: "Control" = aucune clôture, pas d'exclusion; "Rodent Exclosure" = clôture, tous les rongeurs exclus; "Krat Exclosure" = clôture avec porte laissant passer les petits rongeurs, pas pas les rats-kangourous. Ces traitements ont été assignés aléatoirement après délimitation des parcelles.

```
plots <- read.csv("../donnees/portal_plots.csv", stringsAsFactors = FALSE)
str(plots)
```

```
## 'data.frame':    24 obs. of  2 variables:
##  $ plot_id  : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ plot_type: chr  "Spectab exclosure" "Control" "Long-term Krat Exclosure" "Control" ...
```

1. Tests de randomisation

a) Tout d'abord, nous devons préparer les données pour l'analyse:

- Dans le tableau `surveys`, conservez uniquement les observations de l'année 2002 où le poids n'est pas manquant. *Rappel*: La fonction `is.na(x)` vérifie si `x` est une valeur manquante.
- Pour simplifier les données, nous allons grouper les traitements autres que "Control" et "Rodent Exclosure" sous le nom "Krat Exclosure". Voici l'instruction pour effectuer cette transformation.

```
plots$plot_type[!(plots$plot_type %in% c("Control", "Rodent Exclosure"))] <- "Krat Exclosure"
```

- Finalement, joignez les tableaux `surveys` et `plots` pour connaître les traitements des parcelles liés à chaque observation. Vous pouvez utiliser la fonction `merge` dans R ou la fonction `inner_join`, qui requiet le package *dplyr*. Nommez le tableau résultant `surveys_plots`.

Ensuite, visualisez la distribution du poids (`weight`, en grammes) des individus selon le traitement `plot_type`. Vous pouvez utiliser des boîtes à moustaches, par exemple. D'après ce graphique, pour quelle raison serait-il utile d'appliquer une méthode non-paramétrique pour comparer les effets de ces traitements?

b) Nous utiliserons un test de randomisation basé sur l'ANOVA pour déterminer si la masse des individus capturés varie selon le traitement. Pour ce faire, nous écrirons une fonction qui randomise les types de traitement dans le tableau de données `plots`, avant de joindre ce nouveau tableau à `surveys` et d'exécuter l'ANOVA.

- Pourquoi procéder de cette façon, plutôt que de simplement randomiser la colonne `plot_type` dans le tableau de données déjà combiné en (a)? (Pour répondre à cette question, pensez à la justification du test de randomisation dans le contexte de ce plan d'expérience.)

c) Créez la fonction décrite en (b), qui effectue une randomisation de `plots`, joint ce tableau à `surveys`, exécute une ANOVA du poids des individus en fonction du traitement, puis retourne la valeur F. Déterminez la distribution de cette statistique pour l'hypothèse nulle avec 4999 permutations. Quelle est la valeur p pour la valeur F observée si les traitements n'ont aucun effet sur la masse des individus capturés?

d) Effectuez un test de permutation semblable à c) pour l'hypothèse nulle selon laquelle la médiane du poids est la même pour les traitements "Control" et "Krat Exclosure". Quelle est l'écart-type de la statistique du test sous l'hypothèse nulle?

e) Quelle est la valeur p pour le test en d)? La différence est-elle significative avec un seuil $\alpha = 0.01$?

2. Bootstrap

a) Utilisez la méthode du bootstrap avec 10 000 réplicats pour calculer la différence de la médiane du poids des individus capturés, entre les traitements "Krat Exclosure" et "Control". Effectuez une correction du biais et rapportez la différence corrigée avec son erreur-type.

b) Calculez l'intervalle de confiance à 99% pour la différence estimée en a).

c) L'intervalle de confiance obtenu en b) est-il cohérent avec le résultat du test en 1.e)? Est-ce que le bootstrap représente bien le processus d'échantillonnage pour ce problème?