Randomization tests and bootstrap

This assignment must be submitted before **February 4th at 5pm** on Moodle.

Data

This lab uses the Portal database, which contains long-term monitoring data for several rodent species at a study site in 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.

The portal_surveys.csv dataset contains one row per captured individual. Variables include the capture date (day, month, year), plot number, species code, sex, hindfoot length and weight of individuals.

```
surveys <- read.csv("../donnees/portal_surveys.csv")
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 ...
```

The portal_plots.csv dataset indicates the type of treatment applied to each plot. The treatments are designed to exclude different types of rodents: "Control" = no fence, no exclusion; "Rodent Exclusion" = fence, all rodents excluded; "Krat Exclusion" = fence with a gate for small rodents, but not for kangaroo rats. These treatments were randomly assigned after setting up the plots.

```
plots <- read.csv("../donnees/portal_plots.csv")
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. Randomization tests

a) First, we must prepare the data for analysis:

- In the `surveys` table, keep only the observations from the year 2002 where the weight is not missing. *Reminder*: The function `is.na(x)` checks if `x` is a missing value.
- To simplify the data, we will group treatments other than "Control" and "Rodent Exclosure" under the name "Krat Exclosure". Here is the statement to perform this transformation.

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

- Finally, join the `surveys` and `plots` data frames to find out which plot treatment is related to each observation. You can use the `merge` function in R or the `inner_join` function, which requires the *dplyr* package. Name the resulting data frame `surveys_plots`.

Next, view the distribution of the weight (in grams) of the individuals according to the treatment `plot_type`. You can use boxplots, for example. From this graph, why would it be useful to use a non-parametric method to compare the effects of these treatments?

b) We will use a randomization test based on ANOVA to determine if the weight of captured individuals varies with the treatment. To do this, we will write a function that randomizes the types of treatment in the `plots` data table, before joining this new table to `surveys` and running the ANOVA.

- Why do this, rather than simply randomizing the `plot_type` column in the combined data frame produced by a)? (To answer this question, consider the rationale for randomization testing in the context of this experimental design).

c) Create the function described in (b), which performs a randomization of `plot_types`, joins this table to `surveys`, performs an ANOVA of the weight of individuals as a function of treatment, and then returns the value F. Determine the distribution of this statistic for the null hypothesis with 4999 permutations. What is the p value for the observed F value if the treatments have no effect on the mass of individuals captured?

d) Perform a randomized test similar to c) for the null hypothesis that the median weight is the same for the "Control" and "Krat Exclosure" treatments. What is the standard deviation of the test statistic under the null hypothesis?

e) What is the p value for the test in d)? Is the difference significant with a threshold $\alpha = 0.01$?

2. Bootstrap

a) Use the bootstrap method with 10,000 replicates to calculate the difference in the median weight of individuals caught between the "Krat Exclosure" and "Control" treatments. Perform a bias correction and report the corrected difference with its standard error.

b) Calculate the 99% confidence interval for the difference estimated in a).

c) Is the confidence interval obtained in b) consistent with the test result in 1.e)? Does the bootstrap accurately represent the sampling process for this problem?