TP1 : Tests d'hypothèses

Exercice 1 : Lecture de l'article

Il nous est demandé de lire l'article [Elmunzer et al., 2012].

Ordre de lecture adopté : *Abstract* puis *Figures* pour se faire une idée générale. Ici la conclusion est présentée dans l'*Abstract*.

Nous relevons ici l'ensemble des tests statistiques réalisés :

- **Test n°1 (partie 1) :**
  - Nom : test exact de Fisher bilatéral.
  - Hypothèse de test nulle : Soient p_{i_1} (resp. p_{p_1}) la probabilité associée à l'apparition d'une pancréatite post-ERCP. L'hypothèse de test nulle est la suivante : "L'indométacine n'a pas d'effet, le taux de pancréatite reste à 10% dans les deux groupes". Cela se traduit mathématiquement par $H_0 : p_{i_1} = p_{p_1} = 0.1$.
  - Données utilisées : Quantitatives discrètes.

	Pancréatite	Pas de pancréatite	Total
Anti-inflammatoire	47 (sous H_0) \ 27 (sous H_1)	427 (sous H_0)	474
Placebo	47 (sous H_0 et H_1)	427 (sous H_0 et H_1)	474
Total	94 (sous H_0)	854 (sous H_0)	948

  - Justification : Ce test est justifié lors de cas d'un tableau de contingence 2x2 car le nombre de degrés de liberté est toujours égal à 1. C'est le test de référence pour comparer deux proportions quand les effectifs sont modérés.
    Dans l'article, il est dit que la p-value est significative (0.041) pour rejeter H_0 et donc conclure que l'anti-inflammatoire est effectif.

- **Test n°1 (partie 2) :**
  - Nom : Test exact de Fisher bilatéral.
  - Hypothèse de test nulle : Nous faisons la même hyptohèse mais pour un critère différent. Cela donne : Soient p_{i_2} (resp. p_{p_2}) la probabilité associée à l'apparition d'une pancréatite **modérée à grave** post-ERCP. L'hypothèse de test nulle est la suivante : "L'indométacine n'a pas d'effet, le taux de pancréatite **modérée à grave** reste la même dans les deux groupes". Cela se traduit mathématiquement par $H_0 : p_{i_2} = p_{p_2}$.
  - Données utilisées : même configuration que pour la partie 1.
- **Test n°2 :**
  - Nom : Test de Kruskal-Wallis (d'égalité des populations par les rangs). Ici on l'applique à la durée d'intervention.
  - Hypothèse de test nulle : H_0 : Les distributions de durée d'hospitalisation sont identiques dans les deux groupes, ce qui donne formellement : $F_i(x) = F_p(x), $ $\forall x$.
  - Données utilisées : Durées d'hospitalisation (données quantitatives et continues) dans chaque groupe (anti-inflammatoire contre placebo).
  - Justification : La distribution est **asymétrique** ("skewed"), ce qui suggère l'usage d'un test non-paramétrique plutôt qu'un test T de Student par exemple.

Exercice 2 : Stéatose hépatique

Nous reprenons les données du TP0 sur la stéatose hépatique.

```
library(readr)
meta_data <- read_delim("../TP0/Steatosis_phenotype.csv",
delim = ";", escape_double = FALSE, trim_ws = TRUE)
```

```
Rows: 96 Columns: 24
-- Column specification --------------------------------------------------------
Delimiter: ";"
chr  (2): Group, Type
dbl (22): ID, Steatosis, Fibrose, Hemolyse, Foci.surface, Biliruline tot, PA...

i Use `spec()` to retrieve the full column specification for this data.
i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

Soit la question :

Observe-t'on une différence significative entre les poids (le taux de Cholestérol ici) moyens des souris contrôle et des souris qui subissent une diète "high-fat, high-carbohydrate" ?

Nous proposons ici d'effectuer un test de permutation classique. Plusieurs hypothèses permettent de justifier ce test. Tout d'abord, l'ordre de grandeur du nombres d'individus N est bas : 96 souris observées. Aussi, la forme de la question induit un test d'hypothèse à deux variables.

```
# Rappels sur la composition des données

total_samples <- nrow(meta_data)
cat("Nombre total de souris :", total_samples, "\n")
```

```
Nombre total de souris : 96
```

```
# Nombre de groupes de traitement
nb_groups <- length(unique(meta_data$Group))
cat("Nombre de groupes de traitement :", nb_groups, "\n")
```

```
Nombre de groupes de traitement : 4
```

```
samples_per_group <- table(meta_data$Group)
cat("Nombre de souris par groupe :", samples_per_group, "\n")
```

```
Nombre de souris par groupe : 23 23 25 25
```

```
library(dplyr)
```

```

Attaching package: 'dplyr'

The following objects are masked from 'package:stats':

    filter, lag
```

```
The following objects are masked from 'package:base':

    intersect, setdiff, setequal, union
```

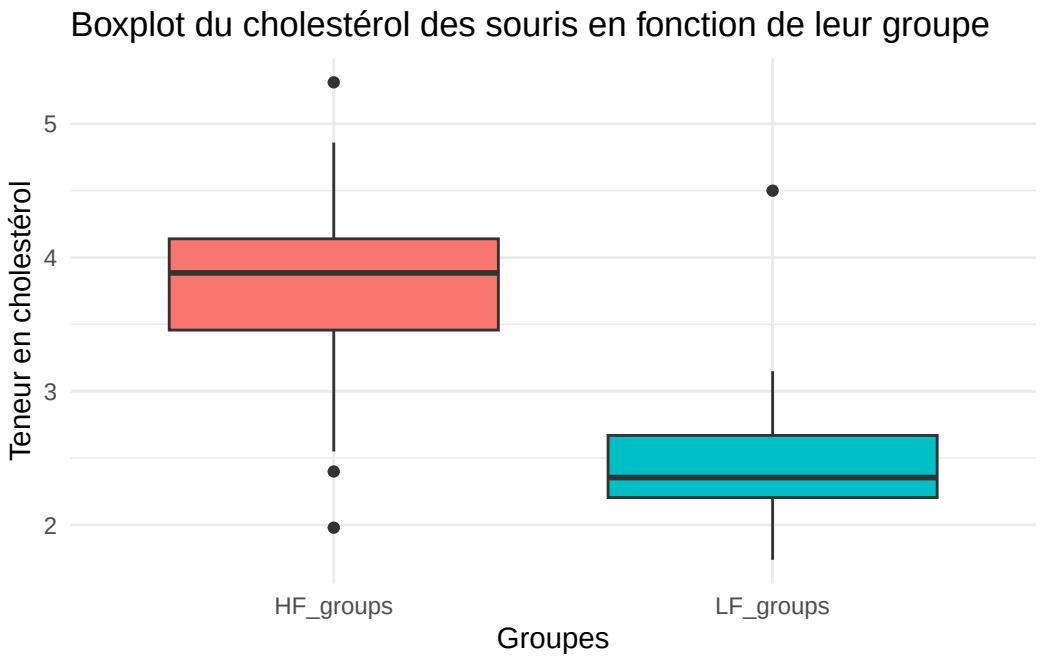
```r
library(ggplot2)

meta_data$Group_2 <- case_when(
  meta_data$Group %in% c("HF", "HF+IRON") ~ "HF_groups",
  meta_data$Group %in% c("CON", "IRON") ~ "LF_groups"
)

# Visualisation des données avec ggplot2 :

ggplot(meta_data, aes(x = Group_2, y = Cholesterol, fill = Group_2)) +
geom_boxplot() +
labs(title = "Boxplot du cholestérol des souris en fonction de leur groupe",
x = "Groupes",
y = "Teneur en cholestérol") +
theme_minimal() +
theme(legend.position = "none")
```

```
Warning: Removed 8 rows containing non-finite outside the scale range
(`stat_boxplot()`).
```



Test de permutation :

```r
# Test de permutation sur le cholesterol

# fonction que calcule la statistique de test
calculate_test_stat <- function(data, groups) {
  group1_mean <- mean(data[groups == "HF_groups"], na.rm = TRUE)
  group2_mean <- mean(data[groups == "LF_groups"], na.rm = TRUE)
  return(abs(group1_mean - group2_mean))
}

# fonction du test de permutation pour la variable Cholesterol
permutation_test <- function(cholesterol_data, groups, n_permutations = 1000) {
  # Statistique observée
  observed_stat <- calculate_test_stat(cholesterol_data, groups)

  # Permutations
  permuted_stats <- replicate(n_permutations, {
    # on mélange les groupes de manière aléatoire :
    shuffled_groups <- sample(groups)
    calculate_test_stat(cholesterol_data, shuffled_groups)
  })

  # on calcule la p-value
  p_value <- (sum(permuted_stats >= observed_stat) + 1) / (n_permutations + 1)

  return(list(
    observed_stat = observed_stat,
    p_value = p_value,
    permuted_stats = permuted_stats
  ))
}

# Test sur la variable Cholesterol
result <- permutation_test(meta_data$Cholesterol, meta_data$Group_2, n_permutations = 1000)

cat("=== RÉSULTATS DU TEST DE PERMUTATION SUR CHOLESTEROL ===\n")
```

```
=== RÉSULTATS DU TEST DE PERMUTATION SUR CHOLESTEROL ===
```

```r
cat("Statistique observée (différence des moyennes) :", round(result$observed_stat, 4), "\n")
```

```
Statistique observée (différence des moyennes) : 1.3136
```

```r
cat("P-value :", round(result$p_value, 4), "\n")
```

```
P-value : 0.001
```

```r
cat("\n=== STATISTIQUES DESCRIPTIVES ===\n")
```

```

=== STATISTIQUES DESCRIPTIVES ===
```

```r
hf_cholesterol <- meta_data$Cholesterol[meta_data$Group_2 == "HF_groups"]
lf_cholesterol <- meta_data$Cholesterol[meta_data$Group_2 == "LF_groups"]

cat("HF_groups - Moyenne :", round(mean(hf_cholesterol, na.rm = TRUE), 4),
    ", Écart-type :", round(sd(hf_cholesterol, na.rm = TRUE), 4), "\n")
```

```
HF_groups - Moyenne : 3.78 , Écart-type : 0.6566
```

```r
cat("LF_groups - Moyenne :", round(mean(lf_cholesterol, na.rm = TRUE), 4),
    ", Écart-type :", round(sd(lf_cholesterol, na.rm = TRUE), 4), "\n")
```

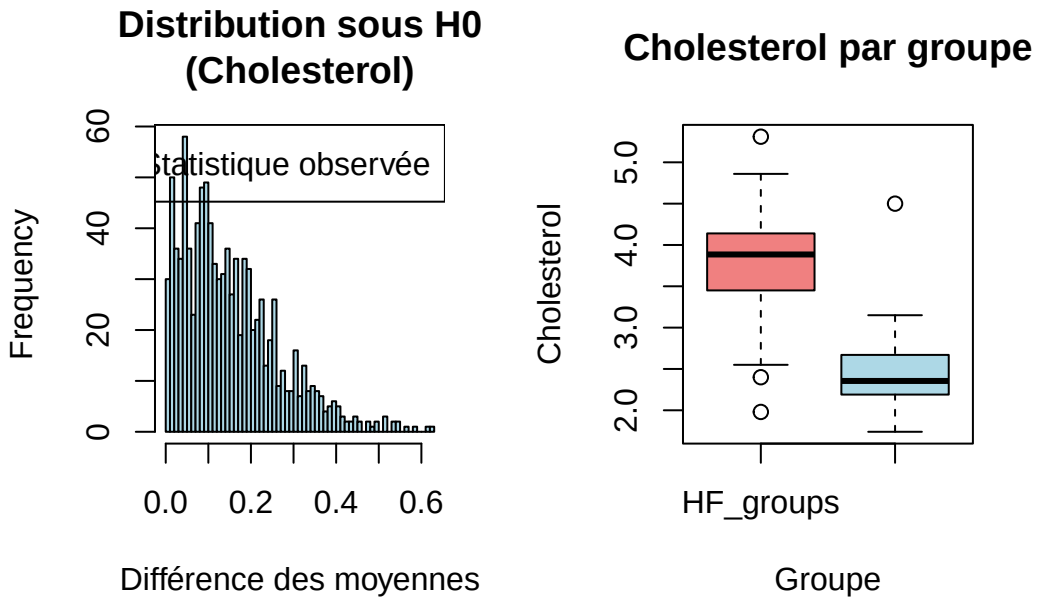
```
LF_groups - Moyenne : 2.4664 , Écart-type : 0.4375
```

```r
# Visualisation des résultats
par(mfrow = c(1, 2))

# Histogramme des permutations
hist(result$permuted_stats, breaks = 50,
     main = "Distribution sous H0\n(Cholesterol)",
     xlab = "Différence des moyennes",
     col = "lightblue")
abline(v = result$observed_stat, col = "red", lwd = 10)
legend("topright", "Statistique observée", col = "red", lty = 1, lwd = 2)

# Boxplot par groupe
```

```r
boxplot(meta_data$Cholesterol ~ meta_data$Group_2,
        main = "Cholesterol par groupe",
        xlab = "Groupe",
        ylab = "Cholesterol",
        col = c("lightcoral", "lightblue"))
```




```r
par(mfrow = c(1, 1))  # Reset layout
```

La statistique observée n'est même pas dans le graphe car elle vaut 1.3...

Exercice 3 - prévention de la pancréatite suite à une cholangiopancréatographie rétrograde endoscopique (CRPE)

On choisit un test décrit dans l'exercice 1. Ici, nous effectuons le test de Fisher bilatéral (partie 1).

```r
load("medicaldata/data/indo_rct.rda")
rx = indo_rct$rx
outcome = indo_rct$outcome

fisher.test(rx, outcome)
```

```
    Fisher's Exact Test for Count Data

data:  rx and outcome
p-value = 0.005339
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 0.2891364 0.8302797
sample estimates:
odds ratio
 0.4946083
```