TP 6 – SY02
Tests d'hypothèses
Corrigé

Pour ce TP, on utilisera les jeux de données issus de la bibliothèque `MASS`. Pour les charger en mémoire, exécuter l'instruction suivante :

```
library(MASS)
```

En `R`, les fonctions réalisant des tests sont généralement de la forme `<mot clé>.test`. Par exemple, un test de Student est réalisé avec la fonction `t.test` et un test de Kolmogorov–Smirnov par la fonction `ks.test`.

1 Tests de conformité

Les tests de conformité testent la conformité d'un paramètre d'un échantillon à une valeur théorique.

Test sur l'espérance : test de Student

Le test de conformité de Student est un test portant sur l'espérance d'une loi gaussienne et s'effectue à l'aide de la fonction `t.test`. Une exécution typique est la suivante :

```
t.test(x, mu = mu0, alternative = "less")
```

où `x` est l'échantillon que l'on veut tester, `mu0` l'espérance de l'hypothèse simple H_0 et `alternative` la nature du test : bilatéral avec le mot clé `"two.sided"` (comportement par défaut), unilatéral inférieur avec le mot clé `"less"` et unilatéral supérieur avec le mot clé `"greater"`. Le niveau de signification peut être changé avec l'argument nommé `conf.level`.

(1) Le jeu de donnée stocké dans le fichier bottles.data contient des quantités effectives de liquide relevées dans 20 bouteilles de 500 ml.

En supposant l'échantillon gaussien, peut-on dire que la quantité de liquide est inférieure à 500 ml ? Tester pour différents niveaux de signification ($\alpha^* = 0.1$, $\alpha^* = 0.05$)

```
>bottles <- read.csv("data/bottles.data")
>t.test(bottles, mu = 500, alternative = "less")       # Niveau de signification α* = 0.05 par défaut

        One Sample t-test

data:  bottles
t = -1.5205, df = 19, p-value = 0.07243
alternative hypothesis: true mean is less than 500
95 percent confidence interval:
     -Inf 501.1569
sample estimates:
mean of x
 491.5705
>t.test(bottles, mu = 500, alternative = "less", conf.level = 0.9)

        One Sample t-test

data:  bottles
t = -1.5205, df = 19, p-value = 0.07243
alternative hypothesis: true mean is less than 500
90 percent confidence interval:
     -Inf 498.9315
sample estimates:
mean of x
 491.5705
```

Le degré de signification vaut 0.0724311. On rejette donc l'hypothèse d'un volume égal à 500 ml au niveau $\alpha^* = 0.1$ mais on ne peut pas rejeter cette hypothèse pour un niveau de signification $\alpha^* = 0.05$.

Test sur une proportion

Le test sur une proportion s'effectue avec la fonction `prop.test`. Elle s'utilise comme suit :

```
prop.test(x, n, p)
```

où `x` est le nombre d'expériences positives, `n` le nombre d'expériences total et `p` la proportion que l'on veut tester.

(2) Le jeu de données présent dans le fichier MM.data contient les effectifs de M&Ms de différentes couleurs issus de 30 sachets pour un total de 1713.

Est-ce qu'une couleur est sur- ou sous-représentée ?

```
>mm <- read.csv("data/MM.data")
>prop.test(mm[1,1], 1713, p = 1/6)

        1-sample proportions test with continuity correction

data:  mm[1, 1] out of 1713, null probability 1/6
X-squared = 0.016813, df = 1, p-value = 0.8968
alternative hypothesis: true p is not equal to 0.1666667
95 percent confidence interval:
 0.1508840 0.1868777
sample estimates:
        p
0.1681261
```

```
>mm <- read.csv("data/MM.data")
>prop.test(mm[1,2], 1713, p = 1/6)

        1-sample proportions test with continuity correction

data:  mm[1, 2] out of 1713, null probability 1/6
X-squared = 16.682, df = 1, p-value = 4.419e-05
alternative hypothesis: true p is not equal to 0.1666667
95 percent confidence interval:
 0.1142414 0.1466409
sample estimates:
        p
0.1295972
```

```
>mm <- read.csv("data/MM.data")
>prop.test(mm[1,3], 1713, p = 1/6)

        1-sample proportions test with continuity correction

data:  mm[1, 3] out of 1713, null probability 1/6
X-squared = 19.435, df = 1, p-value = 1.041e-05
alternative hypothesis: true p is not equal to 0.1666667
95 percent confidence interval:
 0.1114824 0.1435758
sample estimates:
        p
0.1266783
```

```
>mm <- read.csv("data/MM.data")
>prop.test(mm[1,4], 1713, p = 1/6)

        1-sample proportions test with continuity correction

data:  mm[1, 4] out of 1713, null probability 1/6
X-squared = 31.087, df = 1, p-value = 2.468e-08
alternative hypothesis: true p is not equal to 0.1666667
95 percent confidence interval:
 0.1015732 0.1325184
sample estimates:
        p
0.1161705
```

```
>mm <- read.csv("data/MM.data")
>prop.test(mm[1,5], 1713, p = 1/6)

        1-sample proportions test with continuity correction

data:  mm[1, 5] out of 1713, null probability 1/6
X-squared = 67.793, df = 1, p-value < 2.2e-16
alternative hypothesis: true p is not equal to 0.1666667
95 percent confidence interval:
 0.2211520 0.2622182
sample estimates:
        p
0.2410975
```

```
>mm <- read.csv("data/MM.data")
>prop.test(mm[1,6], 1713, p = 1/6)

        1-sample proportions test with continuity correction

data:  mm[1, 6] out of 1713, null probability 1/6
X-squared = 32.549, df = 1, p-value = 1.162e-08
alternative hypothesis: true p is not equal to 0.1666667
95 percent confidence interval:
 0.1991274 0.2388127
sample estimates:
        p
0.2183304
```

2 Tests d'homogénéité

Tests sur des échantillons appariés

La fonction `t.test` permet également de tester deux échantillons appariés en spécifiant l'argument `paired = TRUE`.

Le jeu de données `immer` présent dans la bibliothèque `MASS` contient les rendements de plantations d'orge en différents lieux lors de deux années successives. On souhaite tester si le rendement a été différent d'une année sur l'autre.

(3) Faites un test de Student apparié sur les deux rendements. Que peut-on en conclure au niveau de signification $\alpha^* = 0.05$?

```
>t.test(immer$Y1, immer$Y2, paired = TRUE)

        Paired t-test

data:  immer$Y1 and immer$Y2
t = 3.324, df = 29, p-value = 0.002413
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  6.121954 25.704713
sample estimates:
mean of the differences
               15.91333
```

Le degré de signification est plus petit que $\alpha^* = 0.05$. On rejette donc l'hypothèse H_0 : les deux rendements ne sont pas les mêmes pour les deux années successives.

Le test de Student apparié suppose que la différence des deux échantillons suit une loi gaussienne. Lorsque ça n'est pas le cas, on peut faire un test du signe.

(4) Faire un test du signe sur les deux échantillons précédents. Pour cela :

1. Créer un vecteur de booléen qui indique si la différence entre les deux échantillons est négative et compter le nombre de ces différences négatives.
2. Utiliser la fonction `prop.test` pour tester si la proportion vaut $p = 0.5$.

```
>sign <- immer$Y1 < immer$Y2
>nsuccess <- length(sign[sign])
>n <- length(sign)
>prop.test(nsuccess, n, p = 0.5)

        1-sample proportions test with continuity correction

data:  nsuccess out of n, null probability 0.5
X-squared = 9.6333, df = 1, p-value = 0.001911
alternative hypothesis: true p is not equal to 0.5
95 percent confidence interval:
 0.08404764 0.39130738
sample estimates:
  p
0.2
```

On rejette également l'hypothèse H_0, sans supposer que la différence est gaussienne.

Comparaison de deux variances

La liste `shoes` contient deux vecteurs mesurant l'usure de chaussures de marque A et B.

(5) À l'aide la fonction `var.test`, tester si la variance de l'usure est la même pour les deux types de chaussures.

```
>var.test(shoes$A, shoes$B)

        F test to compare two variances

data:  shoes$A and shoes$B
F = 0.94739, num df = 9, denom df = 9, p-value = 0.9372
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.2353191 3.8142000
sample estimates:
ratio of variances
         0.9473933
```

Comparaison de deux espérances

On souhaite à présent tester si l'usure moyenne des deux marques est la même. On sait déjà d'après la question précédente que les variances sont les mêmes.

Pour comparer les espérances, on utilise encore la fonction `t.test` avec les deux échantillons en spécifiant en plus que les variances des deux échantillons sont supposées les mêmes avec le paramètre `var.equal = TRUE`.

(6) Faites un test d'égalité de l'usure sur les deux marques. Que peut-on en conclure ?

```
>t.test(shoes$A, shoes$B, var.equal = TRUE)

        Two Sample t-test

data:  shoes$A and shoes$B
t = -0.36891, df = 18, p-value = 0.7165
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -2.744924  1.924924
sample estimates:
mean of x mean of y
    10.63     11.04
```

L'usure n'est pas significativement différente.

3 Tests d'adéquation–indépendance

Tests d'adéquation

Le jeu de données `galaxies` regroupe les vitesses calculées de 82 galaxies. On souhaite tester la normalité de ces données.

(7) Faire un test de normalité à l'aide de la fonction `shapiro.test`. La distribution peut-elle être considérée comme issue d'une loi normale ?

```
>shapiro.test(galaxies)

        Shapiro-Wilk normality test

data:  galaxies
W = 0.87177, p-value = 7.302e-07
```

Le degré de signification vaut 7.3016095×10^{-7}. L'hypothèse de normalité H_0 est rejettée, l'échantillon ne suit donc pas une loi normale.

Le fichier de données `delai-data.data` contient des délais d'attente en jours pour un rendez-vous chez un ophtalmologiste. On suppose que les délais suivent une loi exponentielle.

(8) Sachant que l'espérance d'une loi exponentielle de paramètre λ vaut $1/\lambda$, estimer le paramètre λ puis effectuer un test de Kolmogorov–Smirnov avec la fonction `ks.test` pour tester si l'échantillon est bien issu d'une loi exponentielle de paramètre λ.

```
>delai <- read.table("data/delai-data.data", header = TRUE)$delai
>(lambda <- 1/mean(delai))
[1] 0.007484814
>ks.test(delai, "pexp", lambda)

        One-sample Kolmogorov-Smirnov test

data:  delai
D = 0.091389, p-value = 0.05795
alternative hypothesis: two-sided
```

Le degré de signification vaut 0.0579526. On accepte donc l'hypothèse H_0 pour le niveau de signification $\alpha^* = 0.05$. L'échantillon suit bien une loi exponentielle de paramètre $\lambda = 0.0074848$.

Tests d'indépendance

On souhaite tester l'indépendance du choix d'un parfum de glace par rapport au caractère homme–femme. Pour cela, on dispose du tableau de contingence suivant :

	chocolat	vanille	fraise
homme	100	120	60
femme	350	200	90

(9) Définir le `data.frame` regroupant les données de la table précédente.

```
>glace <- data.frame(chocolat = c(100, 350), vanille = c(120, 200), fraise = c(60, 90), row.names =
↪  c("homme", "femme"))
```

(10) Faire un test d'indépendance du χ^2 avec la fonction `chisq.test`. Que peut-on en conclure ?

```
>chisq.test(glace)

        Pearson's Chi-squared test

data:  glace
X-squared = 28.362, df = 2, p-value = 6.938e-07
```

(11) La fonction `chisq.test` renvoie une liste qui contient les informations calculées pour le test. Stocker le résultat du test dans la variable `ct`.

```
>(ct <- chisq.test(glace))

        Pearson's Chi-squared test

data:  glace
X-squared = 28.362, df = 2, p-value = 6.938e-07
```

(12) Que représente les tables `ct$observed` et `ct$expected` ?

La table `ct$observed` est la table des données observées. La table `ct$expected` est la table des effectifs théoriques si on suppose que les deux caractères sont indépendants.

(13) À l'aide de ces deux tables, retrouver la statistique d^2.

```
>sum((ct$expected - glace)^2/ct$expected)
[1] 28.3621
```

On retrouve bien la statistique calculée par `chisq.test`.

4 Cas d'études

Effet d'un médicament soporifique

On souhaite étudier l'effet sur la durée de sommeil de deux médicaments soporifiques. Pour cela, on mesure la durée de sommeil de dix patients après qu'ils aient pris l'un des deux médicaments. Dans les données `sleep`, incluses dans `R`, les dix premières lignes de la première colonne correspondent à la variance de la durée de sommeil en heures par rapport à un groupe de contrôle pour le médicament numéro 1. On supposera dans toute la suite de l'exercice que ces observations sont issues d'une loi $\mathcal{N}(\mu, 4)$. De la même manière, les dix dernières lignes correspondent aux résultats pour le médicament numéro 2 qu'on supposera issues d'une loi $\mathcal{N}(\mu', 4)$. On souhaite déterminer si ces médicaments ont effectivement un effet sur la durée de sommeil, plus précisément si la durée de sommeil est prolongée par la prise de ces médicaments. Formuler le problème sous la forme d'un test d'hypothèses et répondre à la question posée.

On peut formuler le problème de la manière suivante :

$$\begin{cases} H_0 : \mu = 0 & \text{(pas d'effet)} \\ H_1 : \mu > 0 & \text{(prolongation de la durée de sommeil)} \end{cases}.$$

On sait qu'il existe un test UPP et que le degré de signification vaut $\hat{\alpha} = \mathbb{P}_{H_0}\left(\overline{X} \geq \overline{x}\right)$.

```
>x1 <- sleep$extra[sleep$group == 1]
>x2 <- sleep$extra[sleep$group == 2]
>t.test(x1, mu = 0, alternative = "greater")

        One Sample t-test

data:  x1
t = 1.3257, df = 9, p-value = 0.1088
alternative hypothesis: true mean is greater than 0
95 percent confidence interval:
 -0.2870553        Inf
sample estimates:
mean of x
     0.75
>t.test(x2, mu = 0, alternative = "greater")

        One Sample t-test

data:  x2
t = 3.6799, df = 9, p-value = 0.002538
alternative hypothesis: true mean is greater than 0
95 percent confidence interval:
 1.169334      Inf
sample estimates:
mean of x
     2.33
```

Pour le premier médicament, on ne peut pas rejeter H_0, on en conclut que ce médicament n'a pas d'effet significatif sur la durée de sommeil. Pour le second médicament, par contre, on rejette très fortement H_0. Le deuxième médicament a un effet très significatif sur la durée de sommeil.

Rhume et vitamine C

Un groupe de 407 volontaires a reçu des doses de 1000 mg de vitamine C tous les jours durant la saison froide et 411 ont reçu un placebo. Les résultats des personnes ayant attrapés un rhume durant cette période sont compilés dans le fichier cold.data.

(14) L'effet de la vitamine C est-il significatif ?

```
>cold <- read.csv("data/cold.data", row.names = 1)
>chisq.test(cold)

        Pearson's Chi-squared test with Yates' continuity correction

data:  cold
X-squared = 5.9196, df = 1, p-value = 0.01497
```

Le test d'indépendance échoue pour $\alpha^* = 0.05$.