

TP 6 – SY02

Tests d'hypothèses

Corrigé

Pour ce TP, on utilisera des jeux de données disponibles sur Moodle sous forme d'un fichier `.data` et de jeux de données issus des bibliothèques (*library* en anglais) **MASS** et **isdals**. Pour les charger en mémoire, cliquer sur l'item **Packages** (en bas à droite de la fenêtre **RStudio**), les installer (si elles ne figurent pas dans la liste des bibliothèques installées) et les charger en les cochant dans la liste des bibliothèques disponibles ; une approche alternative consiste à exécuter les instructions suivantes :

```
| install.packages("bibliothèque")  
| library(bibliothèque)
```

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`.

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  $\alpha^* = 0.05$  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

```

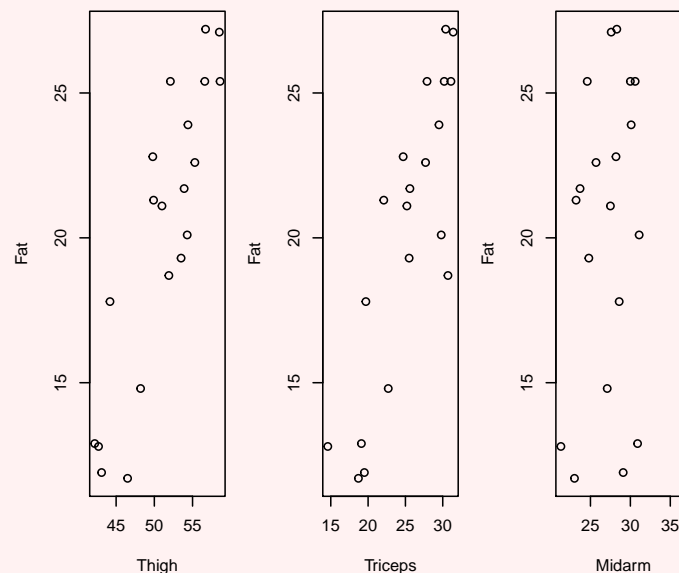
Régression linéaire

Le jeu de données "bodyfat" de la library "isdals" est accessible via les commandes `install.packages("isdals")`, `library(isdals)`, puis `data(bodyfat)` ; il contient les mesures de la masse grasseuse **Fat**, l'épaisseur du pli cutané du triceps **Triceps**, le tour de cuisse **Thigh** et le tour du bras **Midarm** de 20 femmes en bonne santé âgées entre 20 et 30 ans. La procédure pour déterminer le pourcentage de masse grasseuse d'un individu étant coûteuse et délicate, il est souhaitable de mettre au point un "bon modèle" qui explique la masse grasseuse en fonction de variables explicatives plus simple à mesurer, afin d'en déduire des prévisions fiables.

- ③ Tester la réalité d'une relation linéaire entre
1. la variable **Fat** et la variable explicative **Triceps**
 2. la variable **Fat** et la variable explicative **Thigh**
 3. la variable **Fat** et la variable explicative **Midarm**

On peut commencer par faire un plot de la variable à expliquer versus la variable explicative pour avoir une première idée graphique de la réalité de l'existence d'une relation linéaire entre ces variables.

```
install.packages("isdals")
Installing package into '/usr/local/lib/R/site-library'
(as 'lib' is unspecified)
library(isdals)
data(bodyfat)
par(mfrow=c(1,3))
plot(Fat~Thigh,data = bodyfat)
plot(Fat~Triceps,data = bodyfat)
plot(Fat~Midarm,data = bodyfat)
```



Après avoir stocké les objets renvoyés par la fonction `lm` des différentes régressions linéaires, la fonction `summary` d'argument l'un de ces objets, fournit un tableau dans lequel figure la p-value associée aux problèmes de test d'hypothèses

$$H_0 : a = 0 \quad \text{contre} \quad H_1 : a \neq 0$$

et

$$H_0 : b = 0 \quad \text{contre} \quad H_1 : b \neq 0,$$

où a et b sont respectivement la valeur de la droite de régression à l'origine et la valeur de la pente de la droite de régression.

```
regFTh<-lm(Fat~Thigh,data=bodyfat)
summary(regFTh)
Call:
lm(formula = Fat ~ Thigh, data = bodyfat)

Residuals:
    Min       1Q   Median       3Q      Max
-4.4949 -1.5671  0.1241  1.3362  4.4084

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -23.6345     5.6574  -4.178 0.000566 ***
Thigh         0.8565     0.1100   7.786 3.6e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.51 on 18 degrees of freedom
Multiple R-squared:  0.771,    Adjusted R-squared:  0.7583
F-statistic: 60.62 on 1 and 18 DF,  p-value: 3.6e-07
```

Pour la variable explicative "Thigh", les deux tests sont très significatifs et conduisent à rejeter très fortement H_0 (pour a et b).

```
regFTr<-lm(Fat~Triceps,data=bodyfat)
summary(regFTr)
Call:
lm(formula = Fat ~ Triceps, data = bodyfat)

Residuals:
    Min       1Q   Median       3Q      Max
-6.1195 -2.1904  0.6735  1.9383  3.8523

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.4961     3.3192  -0.451  0.658
Triceps       0.8572     0.1288   6.656 3.02e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.82 on 18 degrees of freedom
Multiple R-squared:  0.7111,    Adjusted R-squared:  0.695
F-statistic: 44.3 on 1 and 18 DF,  p-value: 3.024e-06
```

Pour la variable explicative "Triceps", la constante a n'est pas significativement différente de 0 ; on modélise alors la relation entre 'FAT' et 'Triceps' en imposant une valeur de a nulle :

```
regFTr0<-lm(Fat~0+Triceps,data=bodyfat)
summary(regFTr0)
Call:
lm(formula = Fat ~ 0 + Triceps, data = bodyfat)

Residuals:
    Min       1Q   Median       3Q      Max
-5.8660 -2.6037  0.7245  1.9894  3.6156

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
Triceps  0.80020     0.02394   33.42  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.76 on 19 degrees of freedom
Multiple R-squared:  0.9833,    Adjusted R-squared:  0.9824
F-statistic: 1117 on 1 and 19 DF,  p-value: < 2.2e-16
```

Pour la variable explicative "Triceps", le test de nullité sur la pente deux tests est très significatif et conduit à rejeter très fortement $H_0 : b = 0$.

```
regFM<-lm(Fat~Midarm,data=bodyfat)
summary(regFM)
Call:
lm(formula = Fat ~ Midarm, data = bodyfat)

Residuals:
    Min       1Q   Median       3Q      Max
-8.590 -3.847  1.458  3.561  6.909

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  14.6868      9.0959   1.615   0.124
Midarm        0.1994      0.3266   0.611   0.549

Residual standard error: 5.193 on 18 degrees of freedom
Multiple R-squared:  0.02029,    Adjusted R-squared:  -0.03414
F-statistic: 0.3728 on 1 and 18 DF,  p-value: 0.5491
```

Pour la variable explicative "Midarm", le modèle linéaire ne semble pas pertinent. Mais lorsqu'on impose la droite des moindres carrés à valoir 0 à l'origine, le modèle linéaire entre "Fat" et "Midarm" est pertinent.

```
regFM0<-lm(Fat~0+Midarm,data=bodyfat)
summary(regFM0)
Call:
lm(formula = Fat ~ 0 + Midarm, data = bodyfat)

Residuals:
    Min       1Q   Median       3Q      Max
-9.425 -3.343  1.767  4.158  7.626

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
Midarm  0.72251      0.04342   16.64 8.76e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.408 on 19 degrees of freedom
Multiple R-squared:  0.9358,    Adjusted R-squared:  0.9324
F-statistic: 276.9 on 1 and 19 DF,  p-value: 8.763e-13
```

2 Niveau de signification et fonction puissance

À partir de X_1, \dots, X_n un n -échantillon iid de loi exponentielle de paramètre θ ($\theta > 0$) et au niveau de signification α^* , on s'intéresse au problème de test suivant :

$$H_0 : \theta = \theta_0 \quad \text{versus} \quad H_1 : \theta > \theta_0$$

- ④ Donner l'expression de la région critique approchée du test UPP de niveau de signification α^* (cf Ex 8.2).

$$W = \left\{ \bar{X} < \left(\frac{u_{\alpha^*}}{\sqrt{n}} + 1 \right) \frac{1}{\theta_0} \right\}$$

- ⑤ Le jeu de données `delai-data.data` contient des délais d'attente en jours pour obtenir un rendez-vous chez un ophtalmologiste. On propose de modéliser ces données par une loi exponentielle $\mathcal{E}(\theta)$. Pour le jeu de données `delai-data.data`,

1. peut-on dire au niveau $\alpha^* = 0.05$ que le délai d'attente est inférieur aux 151 jours d'attente moyenne de référence (soit environ 5 mois) ?
2. calculer la valeur approchée de la p-value du test ; cette valeur est-elle cohérente avec votre réponse à la question précédente ?

```
delai <- read.table("data/delai-data.data", header = TRUE)$delai
theta0 <- 1 / 151
n <- length(delai)
(mean(delai) < ((qnorm(0.05) / sqrt(n)) + 1) * (1 / theta0))
[1] TRUE
```

On est dans la région critique du test, on rejette donc l'hypothèse nulle au niveau $\alpha^* = 0.05$.
On en déduit la p-value approchée :

```
pnorm((mean(delai) - 1 / theta0) / (1 / (theta0 * sqrt(n))))
[1] 0.04672944
```

La p-value étant inférieure à 0.05, il est cohérent d'avoir rejeter H_0 au niveau 0.05.

- ⑥ Au niveau de signification $\alpha^* = 0.05$, créer une fonction `puiss_emp`, qui prend en argument θ_0 , θ , et un entier n , qui génère un n -échantillon iid de loi exponentielle θ et qui renvoie **TRUE** si l'échantillon réalisé appartient à la région critique du test. Que représente cette fonction ?

```
puiss_emp <- function(theta, theta0, n) {
  x <- rexp(n, theta)
  return(mean(x) < ((qnorm(0.05) / sqrt(n)) + 1) * (1 / theta0))
}
```

La fonction `puiss_emp` fournit une valeur approchée de puissance calculée en θ .

- ⑦ Utiliser la fonction précédente pour illustrer le niveau de signification $\alpha^* = 0.05$ du test.

Remarquons que la fonction puissance calculée en θ_0 coïncide avec le risque de première espèce du test. Pour n une taille quelconque d'échantillon, la proportion des échantillons qui tombent dans la région critique approche le niveau de signification du test. Il suffit alors de prendre la moyenne empirique de la réplication d'un certain nombre de fois de la fonction `puiss_emp` calculée en (θ_0, θ_0, n)

```
theta0 <- 1 / 151
mean(replicate(100, puiss_emp(theta0, theta0, 100)))
[1] 0.04
```

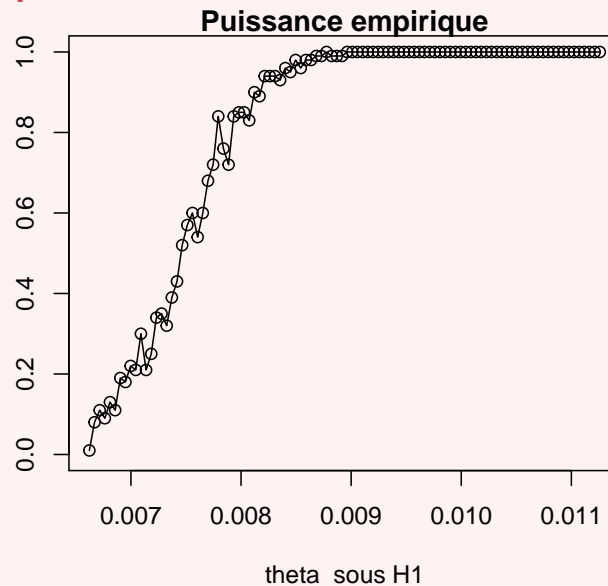
- ⑧ Toujours au niveau $\alpha^* = 0.05$, représenter graphiquement les variations de la fonction `puiss_emp` en fonction de θ dans l'alternative ; qu'observez-vous ?
Indication : créer notamment un vecteur de valeurs de θ dans l'alternative dont la première composante est θ_0 et utiliser la fonction `sapply` vue en TP3, Question 18.

```
theta_alt <- seq(theta0, theta0 * (1 + 0.70), length.out = 100)
```

```

puissance_em_rep <- function(theta0, theta, n, k) {
  sim <- function() {
    x <- rexp(n, theta)
    c <- qnorm(0.05) / (theta0 * sqrt(n)) + 1 / theta0
    return((mean(x) <= c))
  }
  mean(replicate(k, sim()))
}
n <- length(delai)
puis_emp_vec <- sapply(theta_alt, function(theta) puissance_em_rep(theta0, theta, n, 100))
plot(theta_alt, puis_emp_vec, type = "o", ylab = "", xlab = "theta sous H1", xlim = c(theta0,
  ↪ theta_alt[length(theta_alt)]), ylim = c(0, 1))
title("Puissance empirique")

```



On observe que la fonction puissance est croissante en fonction l'écart de $\theta - \theta_0$ et qu'elle atteint rapidement la valeur 1.

3 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 différence de la durée de sommeil en heures par rapport à un groupe de contrôle pour le médicament numéro 1. De la même manière, les dix dernières lignes correspondent aux résultats pour le médicament numéro 2. 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 :

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

Il s'agit d'un test de Student unilatéral.

On vérifie que les données sont distribuées normalement et on utilise ensuite un test de Student.

```
x1 <- sleep$extra[sleep$group == 1]
x2 <- sleep$extra[sleep$group == 2]

shapiro.test(x1)
      Shapiro-Wilk normality test

data:  x1
W = 0.92581, p-value = 0.4079
shapiro.test(x2)
      Shapiro-Wilk normality test

data:  x2
W = 0.9193, p-value = 0.3511
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.