TP 7 – SY02
Test : comparaison – adéquation – indépendance
Corrigé

Les questions/sections marquées par un sont des questions qui sont prévues pour être traitées en autonomie en dehors de la séance de TP.

Pour ce TP, on utilisera des jeux de données disponibles sur Moodle sous forme d'un fichier `.data` et de jeux de données issus de la bibliothèque (*library* en anglais) `MASS`. 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 Kolmogorov–Smirnov est réalisé par la fonction `ks.test` et un test de Shapiro-Wilks par la fonction `shapiro.test`.

1 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.

(1) 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 mean difference is not equal to 0
95 percent confidence interval:
  6.121954 25.704713
sample estimates:
mean difference
       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.

(2) 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` de la library `MASS` contient deux vecteurs mesurant l'usure de chaussures de marque A et B.

(3) 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`.

(4) 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.

2 Tests d'adéquation

Adéquation à une loi gaussienne

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

(5) 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 rejetée, l'échantillon ne suit donc pas une loi normale.

Adéquation à une loi exponentielle

Le fichier de données delai-data.data contient des délais d'attente en jours pour un rendez-vous chez un ophtalmologiste. On veut tester l'adéquation à une loi exponentielle.

(6) 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)
        Asymptotic 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 $\lambda = 0.0074848$.

Le test précédent présente l'inconvénient de nécessiter l'estimation d'un paramètre sans que cela soit pris en compte dans le calcul du degré de signification. On se propose donc de faire un test du χ^2 d'adéquation. Avant d'appliquer le test, nous créons d'abord des boites ainsi que les probabilités théoriques d'appartenance correspondantes.

(7) À l'aide de la fonction `quantile`, créer un vecteur de séparation entre les boites. On prendra garde à inclure le minimum et le maximum dans le vecteur de séparation.

```
(breaks = quantile(delai, seq(0, 1, .1)))
         0%         10%         20%         30%         40%         50%
  0.8154854  14.4562398  33.1079843  58.1793627  80.7285491 113.2493353
        60%         70%         80%         90%        100%
133.8543697 163.1484028 188.4162916 280.6983557 553.3578055
```

(8) Utiliser les fonctions `cut` et `table` pour créer un tableau d'effectifs des boites repérées par les séparations créées précédemment.

```
(x = table(cut(delai, breaks=breaks, include.lowest = TRUE)))
[0.815,14.5]  (14.5,33.1]  (33.1,58.2]  (58.2,80.7]   (80.7,113]    (113,134]
          22           21           21           21           21           21
   (134,163]    (163,188]    (188,281]    (281,553]
          21           21           21           22
```

Il faut utiliser `include.lowest` ici pour inclure dans une boite la valeur minimale. On peut vérifier que toutes les valeurs ont été affectées à une boite avec le code suivant :

```
sum(x) == length(delai)
[1] TRUE
```

(9) Calculer les probabilités théoriques d'appartenance aux boites d'après le vecteur de séparation.

On peut le faire manuellement en appelant de manière répétée la fonction `pexp` sur les séparations comme suit

```
pexp(breaks[2], 1 / mean(delai))
      10%
0.1025539
pexp(breaks[3], 1 / mean(delai)) - pexp(breaks[2], 1 / mean(delai))
      20%
0.1169356
pexp(breaks[4], 1 / mean(delai)) - pexp(breaks[3], 1 / mean(delai))
      30%
0.1335446
```

ou avec le code automatisé suivant :

```
(p = diff(c(0, pexp(breaks, 1 / mean(delai))[2:(length(breaks) - 1)], 1)))
       10%        20%        30%        40%        50%        60%        70%
0.10255394 0.11693556 0.13354459 0.10047529 0.11807015 0.06123016 0.07229515
       80%        90%
0.05081554 0.12174251 0.12233711
```

(10) Utiliser la fonction `chisq.test` pour réaliser le test d'adéquation des effectifs des boites par rapport au probabilités théoriques.

```
chisq.test(x, p = p)
        Chi-squared test for given probabilities

data:  x
X-squared = 21.381, df = 9, p-value = 0.01106
```

(11) La fonction `chisq.test` n'a pas pu prendre en compte le fait qu'on a du estimer un paramètre. Calculer le nouveau degré de signification en ajustant le nombre de degré de liberté.

```
stat <- chisq.test(x, p = p)$statistic
1 - pchisq(stat, df = length(x) - 1 - 1)
 X-squared
0.00620206
```

3 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

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

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

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

```r
chisq.test(glace)
	Pearson's Chi-squared test

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

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

```r
(ct <- chisq.test(glace))
	Pearson's Chi-squared test

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

(15) 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.

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

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

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

4 Cas d'études

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`.

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

Les résultats sont compilés dans un tableau de contingence. Il s'agit donc de déterminer si les deux caractères qualitatifs `Placebo`, `VitC` d'une part et `Cold`, `NoCold` d'autre part sont indépendants.

	Cold	NoCold
Placebo	335	76
VitC	302	105

```
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
```

L'hypothèse H_0 du test d'indépendance est rejetée au niveau $\alpha^* = 0.05$. Les deux caractères ne sont donc pas indépendants. En revanche, pour un niveau de signification plus faible $\alpha^* = 0.01$, on garde l'hypothèse d'indépendance.