


TP 7 – SY02

XXX



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 le moodle sous forme d'un fichier ".data" et de jeux de données issus des bibliothèques ("library") **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) et les charger en les cochant dans la liste des bibliothèques ; une approche alternative consiste à exécuter les instructions suivantes :

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

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 $\alpha^* = 0.05$?

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 $p = 0.5$.

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.

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 ?

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 ?

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 suppose que les délais suivent 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 .

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. Lors du TP6, nous avons étudié 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`

⑦ Peut-on accepter l’hypothèse de normalité des modèles de régression linéaire qui ont été validés au TP6 ?

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

- ⑧ Définir le `data.frame` regroupant les données de la table précédente.
- ⑨ Faire un test d’indépendance du χ^2 avec la fonction `chisq.test`. Que peut-on en conclure ?
- ⑩ La fonction `chisq.test` renvoie une liste qui contient les informations calculées pour le test. Stocker le résultat du test dans la variable `ct`.
- ⑪ Que représente les tables `ct$observed` et `ct$expected` ?
- ⑫ À l’aide de ces deux tables, retrouver la statistique d^2 .

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`.

- ⑬ L’effet de la vitamine C est-il significatif ?