TP 6 – SY02
Tests d'hypothèses

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 ($\alpha^* = 0.1$, $\alpha^* = 0.05$)

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 ?

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 graisseuse `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 graisseuse d'un individu étant coûteuse et délicate, il est souhaitable de mettre au point un "bon modèle" qui explique la masse graisseuse 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`

2 Niveau de signification et fonction puissance

À partir de $X_1, \ldots, X_n$ 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).

(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 ?

(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 ?

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

(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.

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.