TD N°4 : Comparaison d'échantillons

L'objectif de ce TD est d'étudier si deux ou plusieurs échantillons sont issus d'une même population ou bien s'il y a une différence significative entre eux. Nous utiliserons les tests paramétriques et non paramétriques pour comparer 2 échantillons ou k>2 échantillons simultanément.

Exercice 1

Une entreprise fabrique un médicament sur deux chaînes de production. On s'intéresse aux variations de la quantité d'une certaine substance A contenue dans chaque médicament. On a contrôlé le dosage de la substance A avec un échantillon de 100 médicaments à la sortie de chacune des deux chaînes de fabrication. On a trouvé un dosage moyen de 10.75mg pour la première chaîne et 10.70mg pour la deuxième. On sait par ailleurs que l'écart-type des chaînes de production est le même et est égal à 0.2mg.

Construire un test à 1% permettant de savoir si la différence des moyennes observées est due à des fluctuations de l'échantillonnage ou bien si la chaîne de fabrication n°1 produit des médicaments contenant davantage de substance A que la chaîne n°2.

Exercice 2

Un sondage effectué auprès de 2000 personnes indique que 19% d'entre elles connaissent la marque de lessive Omopaic. Après une campagne publicitaire, un sondage analogue auprès de 1000 personnes montre que 230 d'entre elles connaissent cette marque. Peut-on considérer que la campagne a été efficace ?

Exercice 3

On reprend les échantillons 1, 2 et 3 de l'exercice 6 du TD 3. On va utiliser R pour déterminer les échantillons deux à deux afin de savoir s'ils sont issus de la même population.

```
T.TEST(data1 ;data2;…)
```

Il faut préciser si le test est unilatéral avec l'argument `alternative = "less"` ou `"greater"` ou bilatéral avec l'argument `alternative = "two.sided"`. On rappelle que ce test n'est valable que si les échantillons sont gaussiens ou bien suffisamment grands pour appliquer le TCL, et s'il la moyenne est un indicateur pertinent sur les données traitées (par de valeurs extrêmes).

Faites un test bilatéral. Quelles sont vos conclusions si on considère α=5%? Que remarquez-vous ?

Exercice 4

Le jeu de données EmpruntsBordeaux.xls correspond aux détails des lignes d'emprunts remboursés depuis 2002 pour le budget principal de Bordeaux. L'échantillon possède 643 individus qui sont ici des emprunts.

Le jeu de données a été créé en 2002 sur l'open-data de la ville de Bordeaux et il est mis à jour chaque année (http://opendata.bordeaux.fr/content/emprunts-depuis-2002).

1) Dans un premier temps, nous allons chercher à déterminer si le type d'emprunt (Variable ou Fixe) a un impact sur le taux.
   (a) Tracer les boxplot des taux en fonction du type d'emprunts. Peut-on valider l'utilisation du test de Student ?
   (b) Mettre en œuvre le test et conclure.
2) Nous cherchons maintenant à savoir si le type d'emprunt a un impact sur le montant de l'échéance.
   (c) Tracer les boxplot du montant des échéances en fonction du type d'emprunts. Peut-on valider l'utilisation du test de Student ?
   (d) Mettre en œuvre le test approprié et conclure.

Exercice 5

Une enquête sur la consommation annuelle des ménages est réalisée par l'INSEE régulièrement. Ces ménages sont répartis en 5 grandes catégories suivant leur localisation :

C1 : ménages en zone rurale,

C2 : ménages résidant dans une unité urbaine inférieure à 20000 habitants,

C3 : ménages résidant dans une unité urbaine comprise entre 20000 habitants et 100000 habitants,

C4 : ménages résidant dans une unité urbaine supérieure à 100000 habitants autre que l'agglomération parisienne,

C5 : ménages résidant dans l'agglomération parisienne.

Un groupement commercial s'intéresse particulièrement à la consommation annuelle des produits contenus dans la nomenclature 17 de l'INSEE c'est-à-dire, la consommation annuelle en mouton, agneau et chevreau et il souhaite savoir s'il y a un effet "localisation" sur la consommation annuelle moyenne des ménages pour ces produits. Le groupement commercial interroge 5 ménages par catégories. Les résultats en euro sont :

C1	C2	C3	C4	C5
56	47	55	61	69
66	50	51	62	71
54	55	59	54	55
61	46	54	54	62
56	56	59	62	53

On suppose que, pour tout i∈ {1,...,5}, la consommation annuelle d'un ménage en euro de catégorie C_i peut être modélisée par une var X_i suivant la loi normale $N(\mu_i,\sigma^2)$, avec μ_i et σ inconnus.

1. On donne S^2_T= 908,64 et S^2_R= 556,40. Dresser le tableau ANOVA.
2. Effectuez, au risque 5%, le test ANOVA :

   H_0:μ_1=...=μ_5

   H_1: "il existe au moins 2 moyennes différentes"

   Interprétez le résultat. Que pensez-vous des hypothèses de validité du test ?
3. Peut-on affirmer, au risque 5%, que $\mu_3 \neq \mu_4$? $\mu_2 \neq \mu_5$?

	Moyenne	Variance
C1	58,6	23,8
C2	50,8	20,7
C3	55,6	11,8
C4	58,6	17,8
C5	62	65

Résumé numérique

Exercice 6

1) Le fichier Demographie.csv contient l'espérance de vie de 50 pays classés par continent et choisis aléatoirement. L'objectif est d'utiliser R afin de déterminer si l'espérance de vie dépend du continent. On utilise pour cela l'instruction

```
anova(lm(EV~CONTINENT,data=Mydata))
```

   (a) Tracer les boxplot de l'espérance de vie suivant les continents.

   (b) Quelle est votre conclusion au test de l'ANOVA?

   (c) Les conditions d'application du test sont-elles vérifiées ?

2) Le fichier Industries.csv répertorie un certain nombre de caractéristiques (Capital, CA, Continent, …) de 275 entreprises du numériques en 2005. L'objectif est de déterminer si l'investissement en R&D dépend du type d'industries.

   (a) Pourquoi ne peut-on pas utiliser le test de l'ANOVA ? Quel test semble plus approprié ?

   (b) Effectuez ce test avec R pour conclure.