TD 4 Stats AEB

1 Exercice 1

Pour recenser les caractéristiques de poissons vivants dans une rivière on procède à plusieurs séries de mesures. Une première série de mesures est réalisée et les résultats (tailles en cm) stockés dans le fichier exo_1.txt.

1.1 Question 1

- Charger le fichier dans une variable var1

var1<-read.table("exo_1.txt",header=T) # header est optionnel

- En utilisant les différentes commandes vues en cours (graphiques ou non), explorer les données et présentez vos premières conclusions.

dim(var1) # permet de connaitre la taille de l'échantillon (300)
summary(var1) # permet de connaitre les tendances centrales et dispersions de la VA et son nom (size)
hist(var1[,1],breaks=100) # permet de visualiser la distribution des valeurs.
L'histogramme de distribution montre nettement 3 pics et il semble donc que les données soient très hétérogènes concernant les prises de mesures. Il est envisageable que plusieurs sortes de poissons aient été mélangés lors des prises de mesure (au moins 3)

1.2 Question 2

Une nouvelle série de mesures est donc réalisée dans plusieurs cours d'eau. Les données sont rassemblées dans exo_2.txt et concernent cette fois-ci une seule espèce animale.

- Chargez les données dans une variable var2

var2<-read.table("exo_2.txt",header=T)

- Quelle commande simple pouvez-vous utiliser d'emblée pour avoir un visuel global et comparatif des prélèvements dans les 3 rivières ?

boxplot(var2)

- Que pouvez-vous en conclure ?

montre que si les populations des rivières 1 et 2 semblent présenter des distributions assez homogènes entre elles et ce n'est pas le cas pour celle de la rivière 3.

- Quelle hypothèse pouvez-vous alors poser et comment la vérifier ? (détaillez bien les conditions d'application des méthodes utilisées et travaillez au risque $\alpha = 5\%$)

Il semble qu'il faille tester si les populations des 3 rivières sont similaires. Pour cela il est envisageable d'appliquer une ANOVA sur les 3 séries. Cette ANOVA nécessite:

des tailles d'échantillon assez grandes (250 c'est OK)

une homogénéité des variances (à tester mais non vu en cours)

une normalité des distributions (un contrôle visuel est alors effectué et semble probant)

hist(var2$river_1,breaks=10)

hist(var2$river_2,breaks=10)

hist(var2$river_3,breaks=10)

Pour appliquer le test tel qu'on l'a vu précédemment il faut agencer différemment les données. Il faut d'abord aligner les données par facteur (ici le numéro de rivière)

aa<-data.frame(c(var2$river_1,var2$river_2,var2$river_3))

aa<-cbind(c(rep("r1",250),rep("r2",250),rep("r3",250)),aa)

colnames(aa)<-c("river","size")

On peut ensuite appliquer l'ANOVA:

aov1<-aov(aa$size~aa$river)

summary(aov1)

Il apparait ici que l'ANOVA met en évidence qu'au risque 5% il est possible d'écarter H0. Un posthoc peut dont être appliqué:

taov1<-TukeyHSD(aov1)

plot(taov1)

Il semble au risque 5% que la taille moyenne observée pour les poissons dans la rivière 3 soit supérieure à celle observée dans les 2 autres.

1.3 Question 3

De nouvelles mesures (100 prélèvements) sont effectuées dans 2 des 3 rivières (1 et 3) à la recherche de la présence de micro-organismes de 2 variétés. Les résultats sont précisés dans le tableau suivant:

	rivière 1	rivière 3
souche 1 présente	52	60
souche 1 absente	48	40

	rivière 1	rivière 3
souche 2 présente	55	72
souche 2 absente	45	28

- Saisir les données dans 2 variables: acs1 et acs2

acs1<-matrix(c(52,60,48,40),nrow=2)

acs2<-matrix(c(55,72,45,28),nrow=2)

- Existe-t-il une relation entre la présence d'une souche et la rivière où elle est prélevée ? (risque $\alpha = 5\%$)

Cela revient à tester l'indépendance des variables: rivière et présence de la souche. Il est possible de réaliser des chi2 d'indépendance pour chaque tableau.

cs1<-chisq.test(acs1)
cs1$p.value
[1] 0.3186893
cs1<-chisq.test(acs1)
cs2$p.value
[1] 0.01877237

Au risque 5% on peut rejet H0 pour la présence de la souche 2 en fonction de la rivière où a lieu le prélèvement.

2 Exercice 2

Soit 3 échantillons de plantes traitées avec des engrais différents. Les résultats sur la floraison sont présentés dans le tableau suivant:

	Engrais A	Engrais B	Engrais C
Fleuri	34	63	12
Non fleuri	73	16	12

- Créez le tableau en ajoutant les noms de lignes et de colonne

matrice<-matrix(c(34,73,63,16,12,12),nrow=2)
colnames(matrice)<-c("Engrais A","Engrais B","Engrais C")
rownames(matrice)<-c("Fleuri","Non fleuri")

- Vous cherchez à mettre en évidence une relation entre le type d'engrais et la présence de floraison. Comment procédez-vous ?

On réalise un test d'indépendance du chi2
cs1<-chisq.test(matrice)
cs1$p.value [1] 7.840597e-10
On peut donc conclure au risque $\alpha = 5\%$ qu'il existe une relation entre la floraison et le type d'engrais utilisé.

3 Exercice 3

Les bretons, les normands et le beurre salé

	Breton	Non breton
Consomme du beurre salé	10	3
Consomme du beurre doux	2	15

Le type de beurre consommé est-il dépendant de la région d'origine du consommateur ?

matrice<-rbind(c(10,3),c(2,15))
cs1<-chisq.test(matrice)
cs1$p.value
[1] 0.001221099
cs1$expected
[,1] [,2]
[1,] 5.2 7.8
[2,] 6.8 10.2
Au risque 5% on peut affirmer qu'il existe bien une relation entre la région du consommateur et le type de beurre consommé.

	Normand	Non normand
Consomme du beurre salé	10	1
Consomme du beurre doux	2	15

Le type de beurre consommé est-il dépendant de la région d'origine du consommateur ?

matrice<-rbind(c(10,1),c(2,15))
cs1<-chisq.test(matrice)
> cs1$expected
[,1] [,2]
[1,] **4.714286** 6.285714
[2,] 7.285714 9.714286
On ne peut pas conclure avec ces données car l'un des effectifs théoriques est inférieur à 5.