TD3 : Comparaison de deux échantillons paramétrique et non paramétrique

BIO5XX - BIOSTATISTIQUE L3

25 septembre 2016

1 Comparaisons de moyennes entre deux échantillons

1.1 importer le jeu de données peuplier

Après s'être placé dans le bon répertoire, charger les données peupliers

Exemple R - 1 :

Chargement des données peupliers et descriptions rapide de ces données .

```
peuplier<-read.table("peuplier.txt",header = TRUE)
colnames(peuplier)
```

```
## [1] "Site"       "Annee"      "Traitement" "Diametre"   "Hauteur"
## [6] "Poids"      "Age"
```

```
summary(peuplier)
```

```
##       Site           Annee         Traitement       Diametre
##  Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.030
##  1st Qu.:1.000   1st Qu.:1.000   1st Qu.:1.500   1st Qu.:3.675
##  Median :2.000   Median :2.000   Median :2.000   Median :5.200
##  Mean   :1.512   Mean   :1.515   Mean   :2.488   Mean   :4.909
##  3rd Qu.:2.000   3rd Qu.:2.000   3rd Qu.:3.000   3rd Qu.:6.235
##  Max.   :2.000   Max.   :2.000   Max.   :4.000   Max.   :8.260
##     Hauteur          Poids            Age
##  Min.   : 1.150   Min.   :0.010   Min.   :3.000
##  1st Qu.: 5.530   1st Qu.:0.635   1st Qu.:3.000
##  Median : 6.950   Median :1.680   Median :4.000
##  Mean   : 6.969   Mean   :2.117   Mean   :3.502
##  3rd Qu.: 8.785   3rd Qu.:3.470   3rd Qu.:4.000
##  Max.   :10.900   Max.   :6.930   Max.   :4.000
```

On va s'intéresser à la moyenne du poids des arbres dans le site 1 et dans le site 2. L'objectif du TD est de comparer les moyennes des poids des sites 1 et 2. Pour cela on peut commencer par calculer un intervalle de confiance pour les moyennes des poids dans chaque site. Puis on utilisera un test t de comparaison de 2 échantillons indépendants pour tester les hypothèses H_0 : la moyenne du poids des arbres dans le site 1 est égale à la moyenne des poids dans le site 2. H_1 - Les moyennes des poids dans les deux sites sont différentes.

1.2 Normalité des échantillons

1.2.1 Tester la normalité de la variable poids dans chaque site. Constater que cette hypothèse n'est pas respectée

.

En fait on va travailler dans chaque site sur les arbres âgés de 4 ans plantés l'année 1. Vérifier la normalité dans chaque site du poids des arbres âgés de 4 ans, planté l'année 1.

Rmq1. Vous utiliserez l'un des tests de normalité vu précedement.

Rmq2. Vous aurez sans doute besoin de préparer les données pour avoir les poids de chaque site dans 2 fichier séparés. Vous pouvez utiliser la commande subset Préparer les données

Exemple R - 2 :

Préparation des jeux de données .

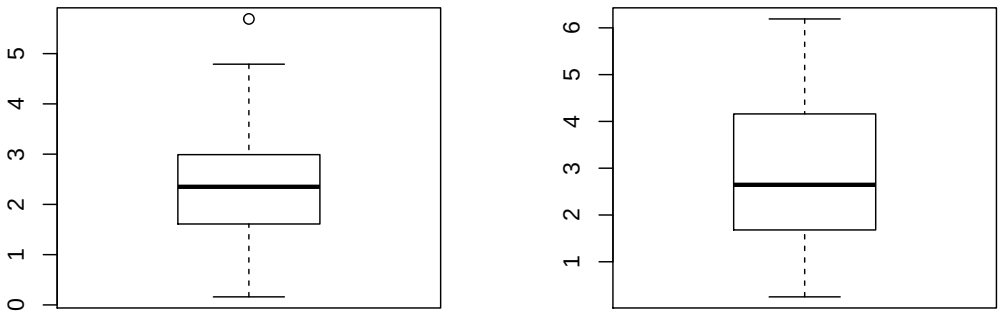
```
peuplier1<- peuplier[peuplier$Site==1 &
                     peuplier$Age==4  &
                     peuplier$Annee==1,]

peuplier2<- peuplier[peuplier$Site==2 &
                     peuplier$Age==4  &
                     peuplier$Annee==1,]
```

1.3 Tester l'hypothèse H_0 contre H_1 énoncée précédemment

1.3.1 Aspects graphiques

Une première comparaison graphique des échantillons peut être réalisée.

```
par(mfrow=c(1,2))
boxplot(peuplier1$Poids)
boxplot(peuplier2$Poids)
```



1.3.2 Test de la normalité des données

Le test de student demande que les données des deux échantillons suivent une loi normale. On teste cette précondition à l'aide du test de Shapiro

Exemple R - 3 :

Vérification de la normalité des échantillons à l'aide de la fonction *shapiro.test*

```
shapiro.test(peuplier1$Poids)

##
##  Shapiro-Wilk normality test
##
## data:  peuplier1$Poids
## W = 0.97875, p-value = 0.7331

shapiro.test(peuplier2$Poids)

##
##  Shapiro-Wilk normality test
##
## data:  peuplier2$Poids
## W = 0.96773, p-value = 0.4019
```

Pour l'échantillon *peuplier*1 la $p_{valeur} = 0.7330761$. Pour l'échantillon *peuplier*2 la $p_{valeur} = 0.4018532$. Les deux $p_{valeur} > 5\%$ on ne peut donc pas rejeter l'hypothèse nulle et l'on va considérer que les deux echantillions suivent une loi normale.

1.3.3 Comparaison des variances

Pour appliquer le test t de comparaison de 2 échantillons indépendants, il faut d'abord vérifier que les variances sont identiques. Il faut utiliser le test F de *Fisher*

— H_0 : égalité des variances
— H_1 : variances différentes

Exemple R - 4 :

Comparaison des variance par le test de extitFisher réalisable grâce à la fonction *var.test* .

```
var.test(peuplier1$Poids,peuplier2$Poids)

##
##  F test to compare two variances
##
## data:  peuplier1$Poids and peuplier2$Poids
## F = 0.6831, num df = 33, denom df = 33, p-value = 0.2787
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.3411556 1.3677661
## sample estimates:
## ratio of variances
##          0.6830966
```

$p = 0.2787266$, on ne rejete pas H_0 pour une une valeurs de $\alpha = 5\%$. On considère donc que les deux echantillons on la même variance.

On peut maintenant mettre en place le test de comparaison des moyennes.

Exemple R - 5 :

Comparaison des moyennes par le test *t* de extitstudent réalisable grâce à la fonction *t.test*

```
t.test(peuplier1$Poids,peuplier2$Poids,var.equal = TRUE)

##
##  Two Sample t-test
##
## data:  peuplier1$Poids and peuplier2$Poids
## t = -1.147, df = 66, p-value = 0.2555
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.0841985  0.2930221
## sample estimates:
## mean of x mean of y
##  2.415588  2.811176
```

Le test de student à une $p_{valeur} = 0.2555338$. Cette p_{valeur} est supérieure au risque $\alpha = 5\%$ habituellement pris. On peut donc rejeter l'hypothèse nulle H_0 et dire qu'il n'y a pas de différence entre les poids moyens des 2 sites avec un risque $\alpha = 5\%$ de se tromper.

Faire la même analyse pour la variable diamètre.

1.4 Comparaison de l'effet de deux drogues

Patients	Hyosciamine	Hyoscine
1	+0,7	+1,9
2	-1,6	+0,8
3	-0,2	+1,1
4	-1,2	+0,1
5	-0,1	-0,1
6	+3,4	+4,4
7	+3,7	+5,5
8	+0,8	+1,6
9	+0,0	+4,6
10	+2,0	+3,4

TABLE 1 – Augmentation (ou diminution) du temps de sommeil (en heures) en présence de la drogue par rapport au temps habituellement dormi.

Exemple R - 6 :

Les colonnes sont crées une à une puis jointes dans un *data.frame* par la fonction *data.frame* qui accepte autant d'arguments que l'on veut réunir de variables. La dernière commande permet juste de vérifier le résultat.

```
Hyosciamine=c(0.7,-1.6,-0.2,-1.2,-0.1,3.4,3.7,0.8,0,2)
Hyoscine=c(1.9,0.8,1.1,0.1,-0.1,4.4,5.5,1.6,4.6,3.4)
Hyosciamine
```

```
##  [1]  0.7 -1.6 -0.2 -1.2 -0.1  3.4  3.7  0.8  0.0  2.0
```

```
Hyoscine
```

```
##  [1]  1.9  0.8  1.1  0.1 -0.1  4.4  5.5  1.6  4.6  3.4
```

```
somnifere <- data.frame(Hyosciamine, Hyoscine)
somnifere
```

```
##    Hyosciamine Hyoscine
## 1          0.7      1.9
## 2         -1.6      0.8
## 3         -0.2      1.1
## 4         -1.2      0.1
## 5         -0.1     -0.1
## 6          3.4      4.4
## 7          3.7      5.5
## 8          0.8      1.6
## 9          0.0      4.6
## 10         2.0      3.4
```

Exemple R - 7 :

Calcul des moyennes des deux échantillons

```
mean(Hyosciamine)
```

```
## [1] 0.75
```

```
mean(Hyoscine)
```

```
## [1] 2.33
```

La moyenne obtenue pour l'Hyosciamine semble plus faible que la moyenne obtenue pour l'Hyoscine. Tester l'augmentation ou la diminution du sommeil provoquée par l'une de ces drogues revient donc à tester que la première moyenne est effectivement plus faible que la seconde.

Il s'agit donc d'un test de *student* entre deux échantillons appariés. Dans ce cas l'important avant de réaliser ce test est de vérifier que la différence de temps de sommeil entre les deux somnifères pour chaque patients suit à peu près une loi normale.

Exemple R - 8 :

R permet de réaliser des opérations simultanément sur tous les éléments d'un vecteur. La normalité est testée par la fonction *shapiro.test*

```
Hyosciamine-Hyoscine

##  [1] -1.2 -2.4 -1.3 -1.3  0.0 -1.0 -1.8 -0.8 -4.6 -1.4

shapiro.test(Hyosciamine-Hyoscine)

##
##  Shapiro-Wilk normality test
##
## data:  Hyosciamine - Hyoscine
## W = 0.82987, p-value = 0.03334
```

Dans notre cas la $p_{value} = 0.0333416$ est inférieure au risque $\alpha = 5\%$ pris habituellement sans en être très éloigné. Nous décidons de continuer malgré tout l'analyse en prenant désormais un risque de $\alpha = 3\%$.

Exemple R - 9 :

L'égalité entre les deux moyennes est testée par la fonction *t.test* en speécifiant le paramêtre *paired* = *TRUE*

```
t.test(Hyosciamine, Hyoscine, alternative='less', paired=TRUE)

##
##  Paired t-test
##
## data:  Hyosciamine and Hyoscine
## t = -4.0621, df = 9, p-value = 0.001416
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##        -Inf -0.8669947
## sample estimates:
## mean of the differences
##                   -1.58
```

Le test de student à une $p_{valeur} = 0.0014164$. Cette p_{valeur} est inférieure au risque $\alpha = 3\%$ que nous avons décidé de prendre. Nous pouvons donc rejeter l'hypothèse nulle H_0 et considerer que l'effet des deux drogues est différent.

2 Statistique non paramétrique

Les tests non paramétriques seront utiles si la taille des échantillons est petite et si les hypothèses de normalités ou d'égalité des variances ne sont pas vérifiées.

2.1 Comparaison de 2 régimes alimentaires

Deux groupes de 10 lapins chacun soumis à un régime enrichi en cholestérol ont été soumis à 2 traitements différents (pour lutter contre le cholestérol). Les résultats sont-ils différents entre les deux régimes ? Cholestérolémie observée en dg/l

X	23	15	28	26	13	8	21	25	24	29
Y	18	22	33	34	19	12	27	32	31	30

TABLE 2 – Cholestérolémie observée

H0 : les deux régimes n'affectent pas la cholestérolémie.
H1 : les deux régimes affectent la cholestérolémie
Vous allez utiliser un test de comparaison d'échantillons indépendants non paramétrique. C'est le test de wilcoxon.

Exemple R - 10 :

Test de Wilcoxon : Les deux drogues n'ont pas d'effets différents.

```
x<-c(23,15,28,26,13,8,21,25,24,29)
y<-c(18,22,33,34,19,12,27,32,31,30)
wilcox.test(x,y)

##
##  Wilcoxon rank sum test
##
## data:  x and y
## W = 31, p-value = 0.1655
## alternative hypothesis: true location shift is not equal to 0
```

2.2 Comparaison de l'effet de deux drogues

Le test de normalité des données pour la comparaison des deux somnifères n'était pas vraiment concluant. Refaite ce test en utilisant un test non paramétrique. Attention, il s'agissait d'un test sur des données appariées. Regardez bien le manuel de la fonction *wilcox.text*

Rédigez cet exercice pour le compte rendu final de TP