TD de la 5° séance

Premiers exercices de prise en main des tests statistiques

1) Procédures de base

Ouvrir R
Ouvrir un éditeur de texte pour sauvegarder votre travail.
Vérifier le répertoire de travail avec la fonction getwd()

```
getwd()
```

Modifier si nécessaire le répertoire de travail

2) Test du Khi-deux pour comparer des effectifs (χ^2 d'homogénéité)

Soit le tableau de contingence A x B suivant à analyser

	A1	A2	A3
B1	13	24	20
B2	10	7	18

Le calcul du test du Khi-deux associé à ce tableau s'effectue de la manière suivante :

```
Var.a<-matrix(c(13,24,20,10,7,18), ncol=3, byrow=T)
Var.a
chisq.test(Var.a)
```

Vérifier si les effectifs théoriques permettent d'utiliser le test de khi-deux

```
chisq.test(Var.a)$expected
```

On peut aussi faire un test exact de fisher

```
fisher.test(Var.a)
```

3) χ^2 de conformité = comparer un pourcentage observé à une valeur théorique

Soit un échantillon de 131 personnes issu d'une population où le taux de prévalence de la syphilis est de 5%. Lors des sérologies nous trouvons 13 personnes avec une sérologie positive.
La proportion observée dans l'échantillon est-elle différente de celle de la population ?

```
binom.test(13,131,0.05)
```

Les effectifs théoriques sont-ils supérieur à 5 ?

```
Efftheor<-131*0.05
Efftheor
```

4) Test de comparaison de moyennes

Ce sont les mêmes fonctions - t.test (test paramétrique) et wilcox.test (test non paramétrique) qui permettent la comparaison entre deux groupes ; dans cet exemple on appliquera les deux tests aux données

```
a<-rnorm(10)
b<-rnorm(10,mean=1)
t.test(a,b)
wilcox.test(a,b)

Var.c<-c(a,b)
Var.d<-gl(2,10,20)
t.test(Var.c~Var.d)
wilcox.test(Var.c~Var.d)
```

5) Analyse de variance sur un facteur

Lorsque l'on est en présence d'un ensemble de k observations indépendantes (un seul facteur inter-sujets), on peut comparer leurs moyennes respectives à l'aide de la fonction aov() (ou selon un modèle linéaire général, avec la fonction lm()).

```
x<-rnorm(100)
a<-gl(4,25,100)
plot(x~a)

r<-aov(x~a)
anova(r)
pairwise.t.test(x,a)
t.test(x[a==1],x[a==2])
```

6) Régression linéaire

la démarche pour effectuer de la régression linéaire est la suivante :

```
A<-rnorm(100)
b<-2*a+rnorm(100)
plot(b~a)
r<-lm(b~a)
anova(r)
abline(r)
```

on peut aussi faire un test

```
cor.test(a,b)
cor(a,b)
```

Exercice à partir du fichier TD5_1

Comparaison d'effectifs ou de proportions

On veut comparer la répartition des sexes par groupe d'âge
On va d'abord visualiser le tableau de contingence correspondant à notre question

```
Tab.1<-table(Mydata$SEXE, Mydata$CLASSAGE)
Tab.1
```

On va chercher à voir les proportions en ligne mais avec un chiffre après la virgule

```
Prop.1<-round(prop.table(Tab.1,1),digits=3)*100
Prop.1
```

On peut visualiser les données à l'aide d'un diagramme en barres. On choisit de représenter les données par sexe sur une même fenêtre graphique

```
par(mfrow=c(2,1));
barplot(Prop.1[1,], main="Female");
barplot(Prop.1[2,], main="Male")
```

On peut aussi visualiser le profil colonne

```
par(mfrow=c(1,1));
Prop.2<-round(100*prop.table(Tab.1,2),digits=1)
barplot(Prop.2,beside=T)
```

On va réaliser maintenant un test de khi-deux

```
Result.1<-chisq.test(Tab.1)
Result.1
```

Vérifier les effectifs théoriques

```
chisq.test(Tab.1)$expected
```

ou

```
summary(Result.1)
Result.1$expected
```

Que concluez-vous?

On peut dans ce cas utiliser aussi le test exact de Fisher ou la correction de yates

```
Result.2<-fisher.test(Tab.1)
Result.2
Result.3<-chisq.test(Tab.1, correct=T)
Result.3
```

La table de données représente des données par individu on aurait pu construire la tableau de contingence avec la fonction xtabs()

```
Tab.2<-xtabs(~CLASSAGE+SEXE, data=Mydata)
Tab.2
chisq.test(Tab.2)
```

Faire le même exercice pour comparer la maladie en fonction du sexe et en fonction de la classe d'âge avec le test de Khi-deux et le test exact de Fisher.

Comparaison de deux moyennes

Nous allons comparer le poids moyen en fonction du sexe
Nous allons comparer graphiquement les deux sous populations

```
boxplot(POIDS~SEXE, ylab="Poids", xlab="Sexe",data=Mydata)
```

Quel est le poids moyen par sous population

```
by(Mydata$POIDS, Mydata$SEXE, summary)
```

Réaliser un test de normalité des données dans chacun des groupes

```
Select.male<-Mydata[,"SEXE"]=="M"
shapiro.test(Mydata[Select.male,"POIDS"])

Select.male<-Mydata[,"SEXE"]=="F"
shapiro.test(Mydata[Select.male,"POIDS"])
```

Si l'hypothèse de normalité est rejettée, le test d'égalité des moyennes peut être effectué à l'aide de test non-paramétriques (wilcoxon:wilcox.test(), ou Kruskal-Wallis:kruskal.test())

Réaliser ensuite un test d'égalité des variances

```
var.test(POIDS~SEXE,conf.level=.95,data=Mydata)
```

Comme les variances sont égales ont peut utiliser un test de student on utilise pour cela l'argument var.equal =TRUE (dans le cas contraire la fonction t.test() réalise un test de Welch)

```
t.test(POIDS~SEXE,alternative="two.sided", conf.level=.95,var.equal=T, data=Mydata)
```

Que concluez-vous?

Nous allons comparer l'IMC moyen en fonction du sexe
Nous allons comparer graphiquement les deux sous populations

```
boxplot(IMC~SEXE, ylab="IMC", xlab="Sexe",data=Mydata)
```

Quel est le poids moyen par sous population

```
by(Mydata$IMC, Mydata$SEXE, summary)
```

Réaliser un test de normalité des données dans chacun des groupes

```
Select.male<-Mydata[,"SEXE"]=="M"
shapiro.test(Mydata[Select.male,"IMC"])

Select.male<-Mydata[,"SEXE"]=="F"
```

```
shapiro.test(Mydata[Select.male,"IMC"])
```

Si l'hypothèse de normalité est rejettée, le test d'égalité des moyennes peut être effectué à l'aide de test non-paramétriques (wilcoxon:wilcox.test(), ou Kruskal-Wallis:kruskal.test())

Réaliser ensuite un test d'égalité des variances

```
var.test(IMC~SEXE,conf.level=.95,data=Mydata)
```

Comme les variances sont égales ont peut utiliser un test de student on utilise pour cela l'argument var.equal =TRUE (dans le cas contraire la fonction t.test() réalise un test de Welch)

```
t.test(POIDS~SEXE,alternative="two.sided", conf.level=.95,var.equal=T, data=Mydata)
```

Que concluez-vous?

Comparaison de deux variables quantitatives (régression linéaire simple)

On va comparer le poids en fonction de la taille
Représenter le nuage de points

```
plot(POIDS~TAILLE,data=Mydata,pch=15,col="red",cex=.5)
```

Réaliser la régression

```
Reglin<-lm(POIDS~TAILLE,data=Mydata)
anova(Reglin)
summary(Reglin)
abline(Reglin)
```

Comparaison de plusieurs moyennes (Analyse de variance à un facteur : ANOVA)

On va comparer l'IMC moyen en fonction des classes d'âge
Représenter graphiquement les données

```
plot(IMC~CLASSAGE,data=Mydata,pch=15,cex=.5)
```

Réaliser ensuite l'analyse

```
Regaov<-lm(IMC~CLASSAGE,data=Mydata)
anova(Regaov)
```

Exercice en autonomie

Utiliser les données TD5_2 créé lors du TD3.

Il faudra créer dans un premier temps une variable GRIPPE qui détermine si l'individu est suspect d'infection par un virus respiratoire, cette variable est codée (1: si l'individu à une fièvre accompagnée d'une toux et 0 dans tous les autres cas)

```
Mydata$GRIPPE<-ifelse(Mydata$TEMP>=38 & Mydata$TOUX==1,1,0)
```

Faire une description de la variable GRIPPE
summary(as.factor(Mydata$GRIPPE)) ou table(Mydata$GRIPPE)
prop.table(table(Mydata$GRIPPE))

Décrire (tableau de contingence, pourcentage et graphique) la variable GRIPPE en fonction du SEXE et faire le test de comparaison de ces deux variables. Dans le graphique vous affecterez une couleur différente pour les données des hommes et pour les données des femmes.

```
A<-table(Mydata$SEXE,Mydata$GRIPPE)
B<-prop.table(A,1)
Conting<-rbind(A[1,],B[1,],A[2,],B[2,])
dimnames(Conting)<-list("Sexe"=c("Male","%","Female","%"),"Grippe"=c("Non","Oui"))
Conting

A2<-table(Mydata$GRIPPE,Mydata$SEXE)
barplot(A2,xaxt="n",ylab="Effectifs",ylim=c(0,2500))
axis(1,c(0.7,1.8),labels=c("Male","Female"))
chisq.test(A2)
```

Décrire (tableau de contingence et graphique) la variable GRIPPE en fonction du QUARTIER et faire le test de comparaison de ces deux variables.

```
A<-table(Mydata$GRIPPE,Mydata$QUARTIER)
B<-prop.table(A,2)
Conting<-rbind(A[1,],B[1,],A[2,],B[2,])
dimnames(Conting)<-list("Grippe"=c("Non","%","Oui","%"),"Quartier"=c("Grand
Yoff","Medina","Ouakam","Pikine"))
Conting

barplot(A,xaxt="n",ylab="Effectifs",ylim=c(0,1500))
axis(1,c(0.7,1.9,3.1,4.3),labels=c("Grand Yoff","Medina","Ouakam","Pikine"))
chisq.test(A)
```

Décrire (tableau de contingence et graphique) la variable GRIPPE en fonction de l'année et faire le test de comparaison de ces deux variables. Dans le graphique vous nommerez l'axe des abscisses par 'Année' et l'axe des ordonnées par 'Effectif'.

On veut savoir si la valeur moyenne de la fièvre est différente chez les patients avec un accès à plasmodium falciparum. Décrire les données de fièvre par groupe et faire le test de comparaison qui convient en vérifiant que les conditions d'utilisation du test sont bien respectées.

L'âge moyen des consultants varie-t-il en fonction du sexe ?

La température moyenne des patients présentant un accès à Plasmodium falciparum est-elle différente en fonction des quartiers?

```
Mydata1<-subset(Mydata,Mydata$PLFACIP==1)
by(Mydata1$TEMP,Mydata1$QUARTIER,summary)
bartlett.test(Mydata1$TEMP~Mydata1$QUARTIER)
```

Continuer maintenant à analyser les différentes données du fichier, en vous posant vos propres questions.

Pour les passionnés

Cinq sportifs ont couru un 1500m et un 5000m. Leurs temps sont donnés dans le tableau suivant :

	Coureur 1	Coureur 2	Coureur 3	Coureur 4	Coureur 5
1500 m	3'58"17	4'05"48	4'12"97	4'08"29	4'00"12
5000 m	14'58"12	14'47"08	15'37"85	13'57"70	14'48"34

Laquelle des deux courses a les temps les plus homogènes ?