```
######################################################################
###                    TP REVISION TESTS STATISTIQUES                ###
######################################################################

######################################################################
##          Exercice 1 : Application sur le jeu de données souris (moodle)          ##
######################################################################
```

Importer des données

```
# fonctions : read.table()->fichier .txt, ou read.csv()-> fichier .csv (excel)

# charger le tableau de données souris.txt (fichier texte)
souris<- read.table("souris.txt", header = TRUE, sep="\t", row.names = 1)
# header=TRUE --> pour dire que 1ere ligne= noms colonnes/variables
# sep= "\t"--> separateur colonnes=tabulation,
# row.name=1--> 1ere colonne = identifiant des individus échantillonnés
```

Toujours commencer par visualiser les données

```
# class(), str(), summary(), plot(), View()…

names(souris) # donne le nom des variables
#(NB: R remplace automatiquement les espaces par des points dans le nom des variables)
# pour la 5e variable, le "é" de activité est remplacé par un caractère spécial "Ã" --> R n'aime pas les caractères spéciaux, évitez les. On va donc changer le nom des variables.

# Renommez les variables comme suit : Masse, Taille, Sexe, Regime, Activite, Pelage
names(souris)<-c("Masse", "Taille","Sexe","Regime", "Activite", "Pelage")

# Pensez à transformer en facteur les variables qualitatives qui sont susceptibles d'être utilisées
# pour faire des comparaisons statistiques (qui définissent des groupes)
souris$Sexe<-as.factor(souris$Sexe)
souris$Pelage<-as.factor(souris$Pelage)

# vérifiez la présence de NA dans le data frame
which( is.na ( souris ) )

######################################################################
#                  A) Est ce que la taille des souris (quel que soit le sexe)          #
#       est conforme à une valeur moyenne de référence de 165mm?          #
######################################################################

#### quel test réaliser ?
# comparaison d'une moyenne observée avec une moyenne théorique (165mm)
# Quand la moyenne de la population est connue (μ) mais pas sa variance
# on peut employer la fonction t.test


#### quelles conditions d'application ?

### verification de la normalité

## a) test de Shapiro-Wilk
shapiro.test(souris$Taille)
# --> normalité OK
```

```
## b) graphiquement (normalité OK)
qqnorm(souris$Taille) ; qqline(souris$Taille) # Dessiner un « qqplot »
qqPlot(souris$Taille, col="brown",pch=19,main="QQplot Taille") # necessite de charger package “car”

histo<-hist(souris$Taille, prob=TRUE, breaks = seq(88,98, 1.02),
        include.lowest=TRUE, right = FALSE, col = "grey",
        main = "Répartition des tailles des souris", xlab = "taille",
        ylab="Densité",xlim = c(88,98))

abline(v=mean(souris$Taille), col="black", lwd=3, lty=4)
# Ajouter une ligne verticale correspondant à la moyenne, de couleur col= noir,
# lwd= epaisseur trait, lty=type trait (plein, tiret, pointillé...)
abline(v=median(souris$Taille), col="dark blue", lwd=3, lty=5)
# Ajouter une ligne verticale correspondant à la moyenne, de couleur col= noir,
# lwd= epaisseur trait, lty=type trait (plein, tiret, pointillé...)
curve(dnorm(x,mean(souris$Taille),sd(souris$Taille)),add=TRUE, lwd=3,col="red")
# ajouter une courbe de densité normale de couleur bleue sur votre graph
# (de meme mean et SD que votre echantillon)

#### quelle(s) conclusion(s) ?
t.test(souris$Taille, mu = 165)
# H1 = la moyenne observée de notre échantillon est différente de la moyenne théorique.

t.test(souris$Taille, mu = 165, alternative = "less") # si on a une idée a priori du sens de la différence
# la taille des souris de notre échantillon est significativement
# inferieure à la moyenne théorique

#########################################################
#                B) Est ce que la taille des males et femelles      #
#                  differe significativement ?                      #
#########################################################

#### Réalisez un graphique approprié ?
boxplot(souris$Taille, na.rm=TRUE) # boxplot des tailles de souris sans distinction de sexe (na.rm=T est non
necessaire ici car pas de valeurs manquantes, mais sachez que cet arument optionnel existe aussi pour les graphs)

## Faire un boxplot pour chaque sexe (plusieurs options)
boxplot(souris[souris$Sexe=="F","Taille"],souris[souris$Sexe=="M","Taille"],names=c("Femelles","Mâles"), col =
c("yellow","green"),ylab="Taille (mm)") # même chose mais en distinguant les sexes et personnalisant le
graphique
boxplot(souris$Taille~souris$Sexe, col = c("yellow","green"),ylab="Taille (mm)") # ou plus simplement Taille en
fonction sexe

# On peut aussi utiliser la fonction subset pour extraire du dataframe global, les variables par sexe:
Fem<-subset(souris, Sexe=="F") # extraction de TOUTES les variables des femelles
Mal<-subset(souris, Sexe=="M") # idem males
boxplot(Fem$Taille,Mal$Taille,names=c("Femelles","Mâles"), col = c("yellow","green"),ylab="Taille (mm)") # ici,
on utilise seulement variable Taille
#### quel test réaliser ?
# Comparaison des moyennes de deux échantillons indépendants (t.test?)

####  quelles conditions d'application ?

###  Vérification de la normalité
```

```
# a) test de Shapiro-Wilk (normalité OK)
shapiro.test(souris[souris$Sexe=="F","Taille"]) # sous échantillonnage des lignes "femelle" et de la colonne "taille" uniquement
shapiro.test(souris[souris$Sexe=="M","Taille"]) # idem pour lignes "males"

## b) Dessiner un « qqplot », histogramme et/ou kernel pour chaque échantillon
qqPlot(souris[souris$Sexe=="F","Taille"],ylab="Taille",main = "QQplot Femelles")
qqPlot(souris[souris$Sexe=="M","Taille"],ylab="Taille",main = "QQplot Males")
#ou
qqnorm(souris[souris$Sexe=="F","Taille"]);qqline(souris[souris$Sexe=="F","Taille"])
qqnorm(souris[souris$Sexe=="M","Taille"]);qqline(souris[souris$Sexe=="M","Taille"])


histo<-hist(souris[souris$Sexe=="F","Taille"], prob=TRUE, breaks = seq(90,98,0.5),
        include.lowest=TRUE, right = FALSE, col = "grey", main = "Répartition des tailles des femelles",
        xlab = "taille", ylab="Densité",xlim = c(90,98))
abline(v=mean(souris[souris$Sexe=="F","Taille"]), col="black", lwd=3, lty=4)
abline(v=median(souris[souris$Sexe=="F","Taille"]), col="dark blue", lwd=3, lty=5)
curve(dnorm(x,mean(souris[souris$Sexe=="F","Taille"]),sd(souris[souris$Sexe=="F","Taille"])),add=TRUE, lwd=3,col="red")

histo<-hist(souris[souris$Sexe=="M","Taille"], prob=TRUE, breaks = seq(88,96,1),
        include.lowest=TRUE, right = FALSE, col = "grey", main = "Répartition des tailles des males",
        xlab = "taille", ylab="Densité",xlim = c(88,96))
abline(v=mean(souris[souris$Sexe=="M","Taille"]), col="black", lwd=3, lty=4)
abline(v=median(souris[souris$Sexe=="M","Taille"]), col="dark blue", lwd=3, lty=5)
curve(dnorm(x,mean(souris[souris$Sexe=="M","Taille"]),sd(souris[souris$Sexe=="M","Taille"])),add=TRUE, lwd=3,col="blue")


kernel_taille_F <- density(souris[souris$Sexe=="F","Taille"])
plot(kernel_taille_F,main="kernel taille femelles",col="brown")

kernel_taille_M <- density(souris[souris$Sexe=="M","Taille"])
plot(kernel_taille_M,main="kernel taille males",col="orange")

### Vérification de l'homoscedasticité

## a) Test de comparaison des variances
# Si la distribution est normale alors on peut vérifier l'homoscédasticité (égalité des variances).
# Pour cela il faut comparer les variances de deux échantillons :
# faire un test d'égalité des variances sur la taille des souris en fonction de leur sexe en
# comparant les variances des Males et des Femelles
var.test(souris$Taille ~ souris$Sexe)
# si la p-value < 0.05 alors hétéroscedasticité, si elle est supérieure à 0.05 alors on ne peut pas rejeter H0 = hypothèse d'homoscedasticité.
# Homoscédasticité OK

## b) On peut également comparer directement les écarts-types
# On considère que si l'un des écarts types est au moins 1.5 fois plus grand que l'autre alors on rejette
# l'homoscédasticité.
varTaille_Sx<-aggregate( souris$Taille, by = souris["Sexe"], FUN = sd)
varTaille_Sx
varTaille_Sx[2,2]/varTaille_Sx[1,2] # 1.2 (avec écart type le plus grand au numérateur)→ Homoscédasticité OK
```

```
# alternative pour mettre automatiquement au numérateur le plus grand des écart types
comp_ET<-max(varTaille_Sx[2,2],varTaille_Sx[1,2])/min(varTaille_Sx[2,2],varTaille_Sx[1,2])
comp_ET  # 1.2 → Homoscédasticité OK


### Après avoir vérifié et validé les conditions de normalité et d'homoscédasticité → t.test() pour échantillons indépendants (les individus sont non appariés)
t.test(souris$Taille~souris$Sexe, var.equal=TRUE) # test bilateral si on n'a pas d'a priori sur le sens de la différence

t.test(souris$Taille~souris$Sexe, var.equal=TRUE, alternative="greater") # unilateral si on a une idée a priori du sens de la différence (ex ici Taille Femelles > Males)
# NB : dans la colonne "sexe", la modalité Femelle apparait en premier -avant modalité Mâle-, on test donc ici Femelles > Males. Si on veut tester l'hypothèse inverse (Femelles < Males), il faut indiquer alternative="less".

####  quelle(s) conclusion(s) ?
# test bilatéral : la différence est hautement significative
# test unilatéral : la taille des femelles est hautement significativement plus grande que celle des males

##############################################################################################
#               C) Est ce que la masse des souris avant et après une modification de leur régime      #
#               alimentaire durant 1 mois (conditions contrôlées) diffère significativement ?         #
##############################################################################################

####  Réalisez un graphique approprié ?
## Faire un boxplot de la masse avant vs après régime
boxplot(souris$Masse,souris$Regime) # version basique (a gauche variable n°1 ,a droite variable n°2)
# version personnalisée :
boxplot(souris$Masse, souris$Regime, col=c("#00FFCC", "#3366FF"), ylab="Masse (g)", xlab="Avant vs Après régime", main="Souris soumises à un changement de régime alimentaire") # on peut utiliser un code hexadecimale pour les couleurs "#xxxxxx"  modifiable directement sur google)

####  quel test réaliser ?
# Comparaison de moyennes de deux échantillons appariés (t.test apparié ?)

####  quelles conditions d'application ?
# vérification de la normalité sur les différences (les di), on commence par créer un vecteur di :
Avant<- souris$Masse
Après<- souris$Regime
di<-Avant-Après
di

### vérification normalité (OK)
shapiro.test(di)

hist(di, breaks = seq(-1,6,0.7),prob=TRUE,include.lowest=TRUE, right = FALSE, col = "grey")
curve(dnorm(x,mean(di),sd(di)),add=TRUE, lwd=3,col="blue")

qqPlot(di, pch=19, col="black", main="QQplot di") # Dessiner un « qqplot »

# Si on vérifie la condition de normalité → t.test() pour échantillons appariés :
t.test(Après,Avant, paired=TRUE) # si on veut tester simplement une différence de masse avant/après
t.test(Après, Avant, paired=TRUE, alternative="less") # si on veut tester masse Après < masse Avant

####  quelle(s) conclusion(s) ?
# Bilatéral -> différence significative
```

```
# Unilatéral -> différence significative (Après<Avant)


################################################################################
#                D) Est ce qu'en moyenne, il y a une différence de taille, masse et/ou          #
#                Activite entre les souris présentant une couleur de pelage différente?          #
################################################################################

####  Réalisez les graphiques appropriés ?
# Avec la fonction boxplot, représenter chaque variable quantitative (Taille, Masse, Regime, Activite) en fonction
# de la couleur de pelage des souris :
boxplot(souris$Taille~souris$Pelage, col=c("yellow", "brown", "orange"), ylab="Taille")
boxplot(souris$Masse~souris$Pelage, col=c("yellow", "brown", "orange"), ylab="Masse")
boxplot(souris$Regime~souris$Pelage, col=c("yellow", "brown", "orange"), ylab="Masse post-régime")
boxplot(souris$Activite~souris$Pelage, col=c("yellow", "brown", "orange"), ylab="Activite")

####  quel test réaliser ?
# ANOVA  1 facteur (?):

aov_Taille<-aov(souris$Taille~souris$Pelage)
summary(aov_Taille) # NS (Non Significatif)

aov_Masse<-aov(souris$Masse~souris$Pelage)
summary(aov_Masse) # NS

aov_Regime<-aov(souris$Regime~souris$Pelage)
summary(aov_Regime) # NS

aov_Activite<-aov(souris$Activite~souris$Pelage)
summary(aov_Activite)
# Il y a une différence très significative entre au moins 2 groupes de souris (couleur pelage)

####  quelles conditions d'application ?

### Vérifier si les résidus du modèle suivent une distribution normale
shapiro.test(residuals(aov_Activite)) # normalité OK
plot(aov_Activite,2,pch=19)
hist(residuals(aov_Activite))
kernel_residus <- density(residuals(aov_Activite))
plot(kernel_residus,main="kernel residus",col="dark blue")

### Vérification de l'homoscedasticité :
bartlett.test(souris$Activite~souris$Pelage) # OK

# Si les conditions ne sont pas satisfaites--> Kruskal-Wallis.
# Si elles sont satisfaites on peut interpreter le test ANOVA et chercher quels échantillons diffèrent des autres
# avec un test post hoc de comparaisons multiples 2 à 2 (Tukey HSD):
TukeyHSD(aov_Activite) # toutes les p-values sont <0.05 (donc significatives)
# et aucun intervalle de confiance n'inclus 0

####  quelle(s) conclusion(s) ?
# L'Activite des souris en fonction de leur couleur de pelage
# est très significativement différente, pour tous les groupes comparés.
```

```
######################################################################
#                E) Est ce que la distribution de la couleur du pelage              #
#                des souris varie selon le sexe des individus ?                     #
######################################################################

####  quel test réaliser ?
# 25 males et 25 femelles = 2 échantillons --> Chi² homogeneite
Souris_PelSx<-chisq.test(souris$Pelage,souris$Sexe)
Souris_PelSx

####  quelles conditions d'application ?
# pas d'effectif attendu <5 (ou <20% d'effectifs attendus <5, mais sans 0)
Souris_PelSx$expected
Souris_PelSx$expected<=5

addmargins(Souris_PelSx$observed) # pour obtenir le TDC
difference<-Souris_PelSx$observed-Souris_PelSx$expected

####  quelle(s) conclusion(s) ?
# pas de différence significative de la couleur du pelage selon le sexe


###################################
##          Exercice 2 : Alliage          ##
###################################

# on recode la variable de dureté :
# tf=1, f=2, m=3, F=4, TF=5
Haute_T<-c(1, 2, 2, 3, 4, 1, 2)
Basse_T<-c(2, 3, 3, 5, 1, 4, 4, 5, 5)
wilcox.test(Haute_T,Basse_T)
# différence non significative au seuil alpha = 0.05, mais la différence serait significative au seuil alpha = 0.1.

# correction du test version « à la main » :
```

Dureté	tf	tf	tf	f	f	f	f	m	m	m	F	F	F	TF	TF	TF
Temp	H	H	B	H	H	H	B	H	B	B	H	B	B	B	B	B
Score B			1				3,5		5,5	5,5		6,5	6,5	7	7	7
Score H	0,5	0,5		1,5	1,5	1,5		3			5					

Somme Score B	49,5		verification :	U1 + U2	=	n1 x n2			Uobs = 13,5		
Somme Score H	13,5			63		63			Uthéo = 12		

```
###################################
##          Exercice 3 : Prairie          ##
###################################

Prairie_1<-c(19.8, 20.6, 27.0, 29.5, 29.9)
Prairie_2<-c(15.9, 19.8, 20.9, 22.5, 26.3)
```

on nous dit que la nomalité de la variable de rendement est respectées, donc on part sur une comparaison de moyennes pour 2 echantillons indépendants → test de Student (avec ou sans correction de Welch ? ou wilcoxon ?)

```
# il faut vérifier la condition d'homogénéité des variances :
var.test(Prairie_1,Prairie_2) # OK

# on peut donc effectuer le test de Student pour echantillon indépendants et variances homogènes :
t.test(Prairie_1,Prairie_2, var.equal = T) # var.equal=T car homognénéité des variances respectée.
# au seuil de 5%, le test est non significatif ; on ne peut pas conclure à une difference de rendement entre les 2 prairies.


# pour l'entrainement, vérifions quand même la condition de normalité à partir des données des échantillons :
shapiro.test(Prairie_1) # OK
shapiro.test(Prairie_2) # OK

qqPlot(Prairie_1) # les points (observations) ne s'alignent pas très bien avec la droite. Ils tombent dans l'IC à 95% car celui-ci est très large en raison du faible effectif de l'échantillon.
qqPlot(Prairie_2) # même remarque + 2 points borderline

# la normalité de la distribution de la variable rendement dans les 2 echantillons est difficile à établir sur de si petits effectifs. Par sécurité, il vaut mieux utiliser un test non paramétrique :  test de Wicoxon-Mann-Whitney avec la fonction wilcox.test()

wilcox.test(Prairie_1,Prairie_2)
# meme conclusion, pas de différence significative.
# les 2 tests convergent en conclusion.



##################################
##        Exercice 4 : Apport en Fer       ##
##################################

omni<-c(11.2,7.6,8.4,9.8,10.9,11,12.2,8.1,10,9.8)
vege<-c(9.3,10.5,11.7,12.4,8.9,10.8,14.2,10,9.5,9.7)

# test de Wilcoxon pour echantillons appariés car on compare 10 individus statistiques avant/après traitement

wilcox.test(omni,vege,paired=T, alternative = "greater") # test unilateral pour voir si l'adoption de ce regime alimentaire vegetarien conduit à une carence en fer (cad [fer] omni –avant-  superieur à [fer] vegetarien –après-)
# difference non significative

# pour info :
wilcox.test(omni,vege,paired=T) # test bilateral, pour tester une simple difference entre les 2 dates
# non significatif également


##################################
##           Exercice 5 : Peupliers          ##
##################################

Aout<-c(8.1,10,16.5,13.6,9.5,8.3,18.3,13.3,7.9,8.1,8.9,12.6,13.4)
Novembre<-c(11.2,16.3,15.3,15.6,10.5,15.5,12.7,11.1,19.9,20.4,14.2,12.7,36.8)

# 13 clones mesurés à 2 dates différentes → test de Student apparié ?
di<-Aout-Novembre
```

shapiro.test(di) # verification de la normalité des di = OK

t.test(Aout, Novembre, paired = T) # test bilatéral
il existe une différence significative de concentration en Aluminium dans le bois entre les 2 dates

# si on avait suspecté une accumulation d'Aluminium dans le bois entre ces 2 dates, il aurait fallu faire un test unilatéral :
t.test(Aout, Novembre, paired = T, alternative = "less")
# la différence de concentration d'Aluminium est significativement plus grande en Novembre qu'en Aout

##################################
Exercice 6 : Pigments
##################################

on a une VD qualitative ordinale à 4 niveaux et une VI qualitative à 3 modalités definissant les groupes à comparer. On ne peut utiliser qu'un test de Kruskal-Wallis car la VD est ordinale.

on recode la variable de concentration en pigments (VD):
abs=0, f=1, m=2, F=3

C_pigments<-c(0,2,1,3,0,2,1,3,2,0,0,3,0,1,0,1,3,1,2,0,3,3,2)
substrat<-c(rep("Argile",7),rep("Sable",9),rep("Calcaire",7))

Pigm_subst<-data.frame(C_pigments,substrat)
View(Pigm_subst)

kruskal.test(Pigm_subst$C_pigments,Pigm_subst$substrat)
# il n'y a pas de différence significative de concentration en pigment entre les différents échantillons.

correction du test version « à la main » :

classement	1	2	3	4	5	6	7	8	9	10	11	12
Pigmt	abs	abs	abs	abs	abs	abs	abs	f	f	f	f	f
Subst	Arg	Arg	Sab	Sab	Sab	Sab	Calc	Arg	Arg	Sab	Sab	Calc
Rang	4	4	4	4	4	4	4	10	10	10	10	10

classement (suite)	13	14	15	16	17	18	19	20	21	22	23
Pigmt (suite)	m	m	m	m	m	F	F	F	F	F	F
Subst (suite)	Arg	Arg	Sab	Calc	Calc	Arg	Sab	Sab	Calc	Calc	Calc
Rang (suite)	15	15	15	15	15	20,5	20,5	20,5	20,5	20,5	20,5

Somme rangs :	Arg =	78,5	n= 7	Hobs =	2,1479	ddl = K-1 = 2	
	Sab =	92	n= 9	correction =	0,9353	Hthéo = 5,991	
	Calc =	105,5	n= 7	Hobs_corr =	2,2965		
			N = 23				