CY-Tech	Statistiques	Mathématiques
2020 - 2021	**TD 4 (Corrigé)**	ING2-GI

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 $n = 100$ médicaments à la sortie de chacune des deux chaînes de fabrication.
On a trouvé un dosage moyen de 10.75 mg pour la première chaîne et 10.70 mg pour la deuxième. On sait par ailleurs que l'écart-type des chaînes de production est le même et est égal à 0.2 mg.
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 1 produit des médicaments contenant davantage de substance A que la chaîne 2.

Correction 1. -

Notons X le dosage de la substance A dans un médicament de la chaîne 1, Y le dosage de la substance A dans un médicament de la chaîne 2, $\mu_1 = E(X)$ et $\mu_2 = E(Y)$.
On sait que $Var(X) = Var(Y) = \sigma^2 = (0.2)^2$.
Nous partons de 2 échantillons indépendants de X et Y de taille $n = 100$.
Le test à construire a les caractéristiques suivantes :

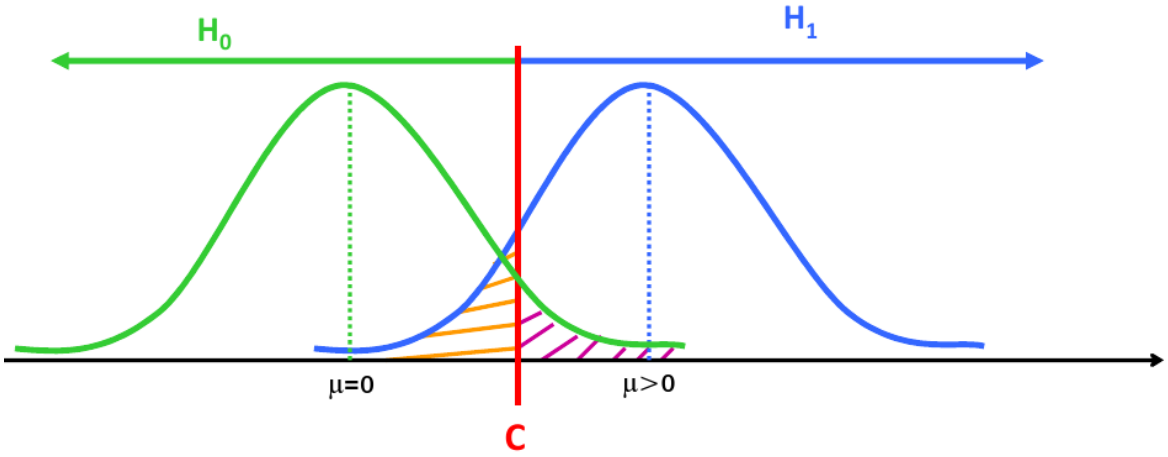
(a) Hypothèses : $\begin{cases} (H_0) & \mu_1 = \mu_2 \\ (H_1) & \mu_1 > \mu_2 \end{cases} \iff \begin{cases} (H_0) & \mu_1 - \mu_2 = 0 \\ (H_1) & \mu_1 - \mu_2 > 0 \end{cases}$

(b) Variable de décision : $D = \overline{X} - \overline{Y}$, avec $\overline{X}$ moyenne empirique, estimateur usuel de μ_1 et $\overline{Y}$ moyenne empirique, estimateur usuel de μ_2.

Comme $\overline{X} \sim \mathcal{N}\left(\mu_1, \frac{\sigma^2}{n}\right)$ et $\overline{Y} \sim \mathcal{N}\left(\mu_2, \frac{\sigma^2}{n}\right)$, la loi suivie par D est : $\mathcal{N}\left(\mu_1 - \mu_2, \frac{2\sigma^2}{n}\right)$

La statistique utile sous (H_0) sera donc : $Z = \frac{D}{\sigma\sqrt{2}}\sqrt{n} \sim \mathcal{N}(0,1)$

(c) Région critique :



Les hypothèses en jeu $\mu_D = 0$ contre $\mu_D > 0$ impliquent que la région critique est de la forme : $W = \{D > C\}$.

$$\alpha = 0.01 = P_{H_0}(D > C) = P_{H_0}\left(\frac{D-0}{\sigma\sqrt{2}}\sqrt{n} > \frac{C}{\sigma\sqrt{2}}\sqrt{n}\right) = P\left(Z > \frac{C}{\sigma\sqrt{2}}\sqrt{n}\right)$$

La table de $\mathcal{N}(0,1)$ donne : $\dfrac{C}{\sigma\sqrt{2}}\sqrt{n} = 2.33 \implies C = 2.33\dfrac{\sigma\sqrt{2}}{\sqrt{n}} = 0.066$

Le seuil est donc $C = 0.066$

(d) Application à notre échantillon :

Sur notre échantillon, nous avons $d = \overline{x} - \overline{y} = 10.75 - 10.70 = 0.05$

$d < C$, on ne peut pas rejeter (H_0).

Conclusion : il n'y a pas de diffrérence significative de dosage de la substance A entre les deux chaînes de production.

Exercice 2. -

Un sondage effectué auprès de $n = 2000$ personnes indique que 19% d'entre elles connaissent la marque de lessive Omopaic.

Après une campagne publicitaire, un sondage analogue auprès de $n = 1000$ personnes montre que 230 d'entre elles connaissent cette marque.

Peut-on considérer que la campagne a été efficace ?

Correction 2. -

(a) Hypothèses :

Soient

• p_1 la proportion de personnes connaissant la marque avant la campagne publicitaire,

• p_2 le taux de personnes connaissant la marque après cette campagne.

Si la campagne a été efficace, on devrait avoir : $p_2 > p_1$.

On teste donc les hypothèses :

$$\begin{cases}(H_0) & p_1 = p_2\\ (H_1) & p_1 < p_2\end{cases} \iff \begin{cases}(H_0) & p_1 - p_2 = 0\\ (H_1) & p_1 - p_2 < 0\end{cases}$$

(b) Variable de décision : $D = F_{n1} - F_{n2}$, avec F_{n1} moyenne empirique, estimateur usuel de p_1 et F_{n2} moyenne empirique, estimateur usuel de p_2.

La taille des échantillons ($n \geqslant 1000$) autorise l'utilisation du TCL.

Comme $F_{n1} \sim \mathcal{N}\left(p_1, \dfrac{p_1(1-p_1)}{n}\right)$ et $F_{n2} \sim \mathcal{N}\left(p_2, \dfrac{p_2(1-p_2)}{n}\right)$,

la loi suivie par D est : $\mathcal{N}\left(p_1 - p_2, \sigma^2\right)$ avec $\sigma^2 = \dfrac{p_1(1-p_1)}{n_1} + \dfrac{p_2(1-p_2)}{n_2}$

La statistique utile sous (H_0) sera donc : $Z = \dfrac{D}{\sigma} \sim \mathcal{N}(0,1)$

Dans ce cas, $p = p_1 = p_2$ sera remplacé par sa meilleure estimation :

$$\hat{p} = \frac{n_1 f_{n1} + n_2 f_{n2}}{n_1 + n_2} = \frac{2000 \times 0.19 + 1000 \times 0.23}{3000} = 0.203$$

Et donc : $\sigma \simeq \sqrt{\dfrac{\hat{p}(1-\hat{p})}{n_1} + \dfrac{\hat{p}(1-\hat{p})}{n_2}} = \sqrt{\hat{p}(1-\hat{p})\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}$

(c) Région critique :

Nous allons fixer $\alpha = 5\%$.

Les hypothèses en jeu $\mu_D = 0$ contre $\mu_D < 0$ impliquent que la région critique est de la forme : $W = \{D < C\}$.

$$\alpha = 0.05 = P_{H_0}(D < C) = P_{H_0}\left(\frac{D-0}{\sigma} < \frac{C}{\sigma}\right) = P\left(Z < \frac{C}{\sigma}\right)$$

La table de $\mathcal{N}(0,1)$ donne : $\dfrac{C}{\sigma} = -1.645 \implies C = -1.645\sigma = -0.026$

Le seuil est donc $C = -0.026$

(d) Application à notre échantillon :

Sur notre échantillon, nous avons $d = f_{n1} - f_{n2} = 0.19 - 0.23 = -0.04$
$d < C$, on doit rejeter (H_0) et valider (H_1).
Conclusion : la campagne publicitaire a été efficace et a augmenté de manière significative la proportion de personnes connaissant la marque.

Exercice 3. -

Un archéologue utilise deux isotopes différents pour dater $n = 130$ objets.
Pour chacun d'entre eux, il calcule la différence d_i des dates avec les deux isotopes. Ces 130 différences ont pour moyenne $\overline{d} = 53$ ans et un écart-type $s_d = 680$ ans.

1. Quel test doit-on effectuer pour comparer les deux méthodes de datation ?
2. Effectuer ce test et conclure.

Correction 3. -

1- Il s'agit ici de comparer deux échantillons appariés. Chacun des 130 objets est daté par les deux isotopes. Si X_i est la date donnée par le premier isotope pour le i-ème objet et Y_i la date donnée par le deuxième. Ces deux dates ne sont évidemment pas indépendantes.
Notons $\mu_X = E(X)$ et $\mu_Y = E(Y)$.
On fait alors intervenir les différences : $D_i = X_i - Y_i$,
avec $\mu_D = E(D) = \mu_X - \mu_Y$
leur moyenne empirique : $\overline{D} = \dfrac{1}{n}\sum_{i=1}^{n} D_i$,
ainsi que la variance empirique corrigée : $S_D = \dfrac{1}{n-1}\sum_{i=1}^{n}(D_i - \overline{D})^2$

2-
(a) Hypothèses : $\begin{cases}(H_0) & \mu_X = \mu_Y \\ (H_1) & \mu_X \neq \mu_Y\end{cases} \iff \begin{cases}(H_0) & \mu_X - \mu_Y = 0 \\ (H_1) & \mu_X - \mu_Y \neq 0\end{cases} \iff \begin{cases}(H_0) & \mu_D = 0 \\ (H_1) & \mu_D \neq 0\end{cases}$
(b) Variable de décision : $\overline{D}$, moyenne empirique, estimateur usuel de μ_D.
La statistique utile sous (H_0) sera donc : $T = \dfrac{\overline{D}}{S_D}\sqrt{n} \sim T_{n-1}$: loi de Student à $(n-1)$ d.d.l.
(c) Région critique :
Nous allons fixer $\alpha = 5\%$.
Les hypothèses en jeu $\mu_D = 0$ contre $\mu_D \neq 0$ indiquent un test bilatéral.
La région critique est de la forme : $W = \{\overline{D} < C_1 \quad ou \quad \overline{D} > C_2\}$.
Comme $\mu_D = 0$, les deux seuils sont symétriques par rapport à 0, c à d : $C_1 = -C_2 = -C$.
On peut donc écrire : $W = \{|\overline{D}| > C\}$.
La loi de Student est également symétrique par rapport à 0.
$$\alpha = 0.05 = P_{H_0}(|\overline{D}| > C) = P_{H_0}\left(|\frac{\overline{D}}{S_D}\sqrt{n}| > \frac{C}{S_D}\sqrt{n}\right) = P\left(|T| > \frac{C}{S_D}\sqrt{n}\right)$$
La table de Student à 99 d.d.l se confond avec celle de $\mathcal{N}(0,1)$, ce qui donne :
$\dfrac{C}{S_D}\sqrt{n} = 1.96 \implies C = 1.96 \times \dfrac{S_D}{\sqrt{n}} \simeq 117$ ans.
Le seuil est donc $C = 117$ (un écart de ± 117 années).
(d) Application à notre échantillon :
Sur notre échantillon, nous avons $\overline{d} = 53$ ans.
$|d| < C$, on ne peut pas rejeter (H_0)
Conclusion : il n'y a pas de différence significative entre les deux méthodes de datation.

Exercice 4. -

On va utiliser R pour comparer les échantillons des fichiers Ech1, Ech2 et Ech3 deux à deux afin de savoir s'ils sont issus de la même population.
La commande principale est : **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 si la moyenne est un indicateur pertinent sur les données traitées (par de valeurs extrêmes).
Faire un test bilatéral avec $\alpha = 5\%$? Que remarque-t-on ?

Correction 4. -

Ech1=read.table("Ech1.txt",header=T)
Ech2=read.table("Ech2.txt",header=T)
Ech3=read.table("Ech3.txt",header=T)
t.test(Ech1,Ech2,alternative="two.sided")
t.test(Ech2,Ech3,alternative="two.sided")
t.test(Ech1,Ech3,alternative="two.sided")
p-valeur du test bilatéral de Student :

- Ech1 et Ech2 : $51,16\% \implies$ même moyenne.
- Ech2 et Ech3 : $14,65\% \implies$ même moyenne.
- Ech1 et Ech3 : $3,96\% \implies$ pas même moyenne.

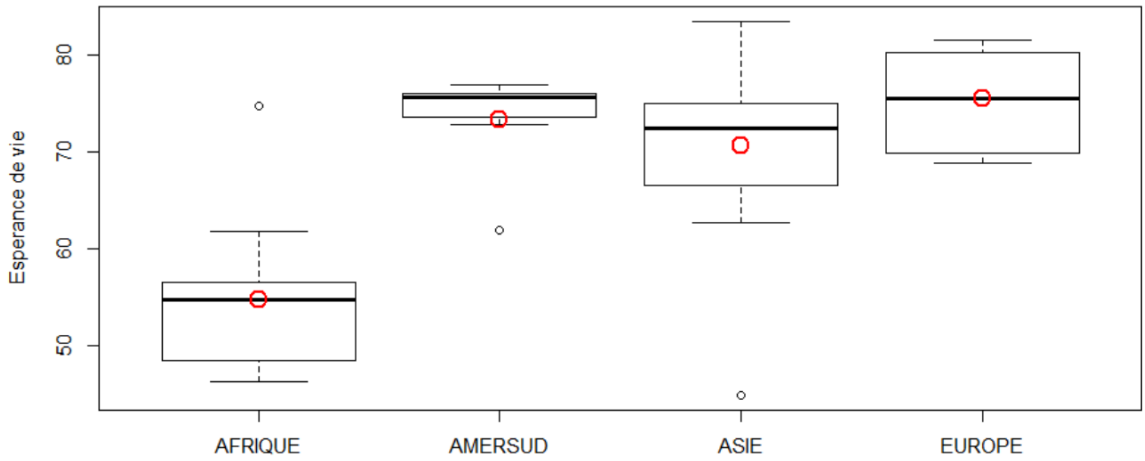
On note que les conclusions des tests 2 à 2 ne sont pas "transitives".
La différence entre premier et deuxième échantillon peut être non significative, tout comme celle entre le deuxième et le troisième. Cela ne peut empêcher la différence entre le premier et le troisième d'être significative.

Exercice 5. -

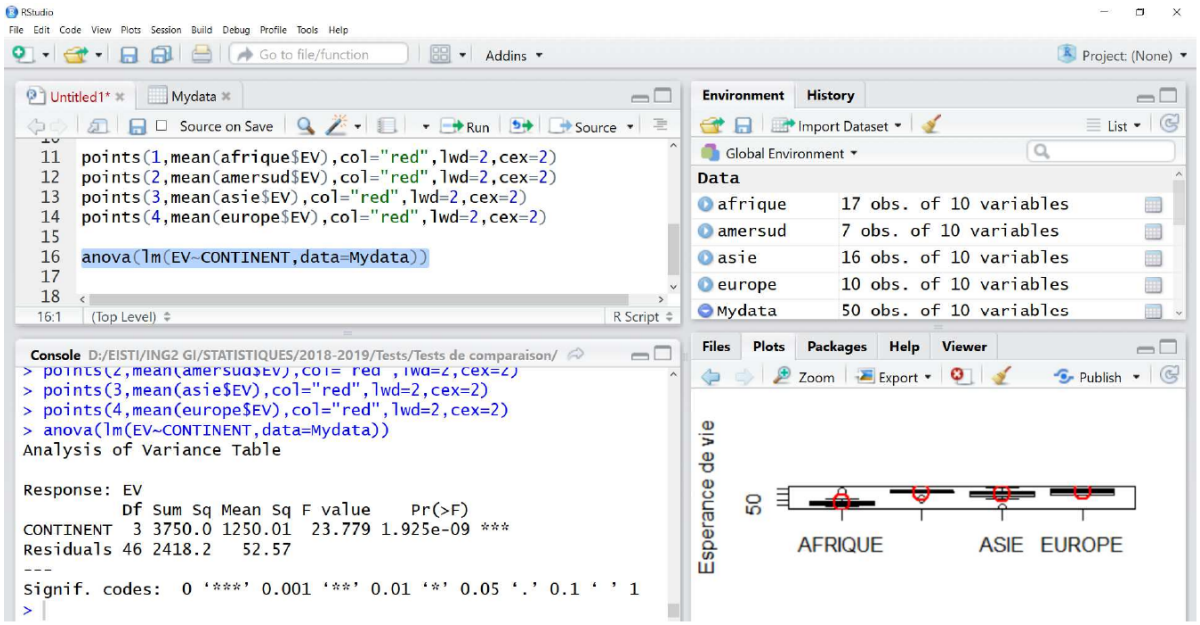
1. Le fichier **Demographie.csv** contient l'espérance de vie de 50 pays classés par continent et choisis aléatoirement. L'objectif est d'utiliser R afin de déterminer si l'espérance de vie dépend du continent. On utilise pour cela l'instruction ***anova(lm(EV~CONTINENT,data=Mydata))***
   (a) Tracer les boxplot de l'espérance de vie suivant les continents.
   (b) Justifier la possibilité d'effectuer un test de l'ANOVA.
   (c) Conclure à partir des résultats de l'ANOVA.
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. Quelle aurait été la conclusion si on avait effectué une ANOVA ?

Correction 5. -

1-
(a) Commandes à utiliser :
boxplot(EV~CONTINENT,data=Mydata,ylab="Esperance de vie")
afrique=subset(Mydata,CONTINENT=="AFRIQUE")
amersud=subset(Mydata,CONTINENT=="AMERSUD")
asie=subset(Mydata,CONTINENT=="ASIE")
europe=subset(Mydata,CONTINENT=="EUROPE")
points(1,mean(afrique$EV),col="red",lwd=2,cex=2)
points(2,mean(amersud$EV),col="red",lwd=2,cex=2)
points(3,mean(asie$EV),col="red",lwd=2,cex=2)
points(4,mean(europe$EV),col="red",lwd=2,cex=2)

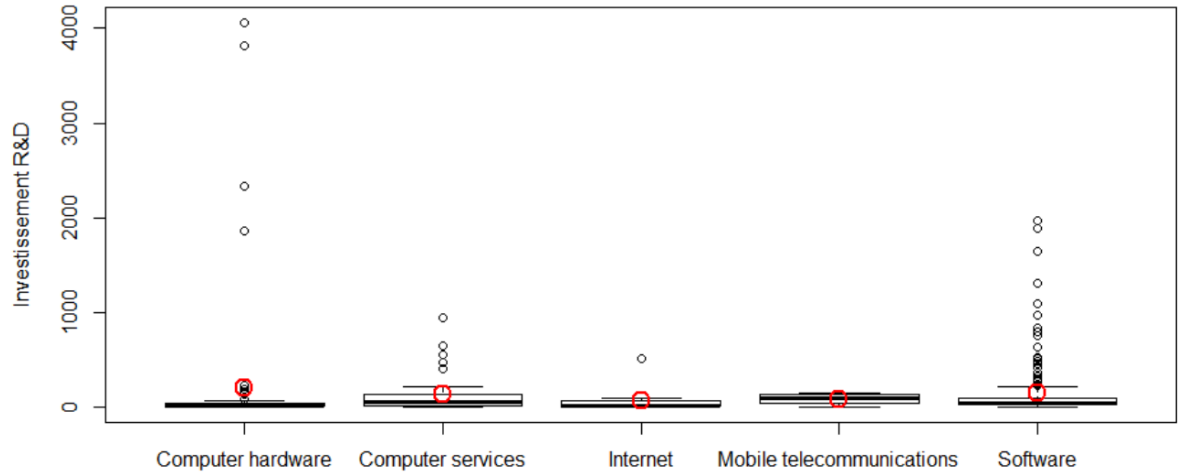


(b) Il semble que le continent ait un impact sur l'espérance de vie notamment avec l'Afrique. Les distributions de l'EV sur chaque continent semblent symétriques sans trop de valeurs extrêmes et de variances égales, excepté pour l'Amérique du Sud. Mais ceci peut s'expliquer par le peu de pays de ce continent dans l'échantillon. On valide donc la pertinence de test de l'ANOVA.
AFRIQUE : 17, AMERSUD : 7, ASIE : 16, EUROPE : 10
Commande à utiliser :
anova(lm(EV˜CONTINENT,data=Mydata))



(c) La p-valeur du test est très petite (1.95×10^{-9}) donc on rejette l'hypothèse (H_0) disant qu'il y a égalité des moyennes. Le continent a donc un impact significatif sur l'espérance de vie.
2-
(a) Commandes à utiliser :
Mydata=read.table("Industries.csv",header=T,sep=";",dec=",")
boxplot(R.D.investment˜Industry.classfication,data=Mydata,ylab="Investissement R&D")
hardware=subset(Mydata,Industry.classfication=="Computer hardware")
services=subset(Mydata,Industry.classfication=="Computer services")
internet=subset(Mydata,Industry.classfication=="Internet")
telecom=subset(Mydata,Industry.classfication=="Mobile telecommunications")
software=subset(Mydata,Industry.classfication=="Software")

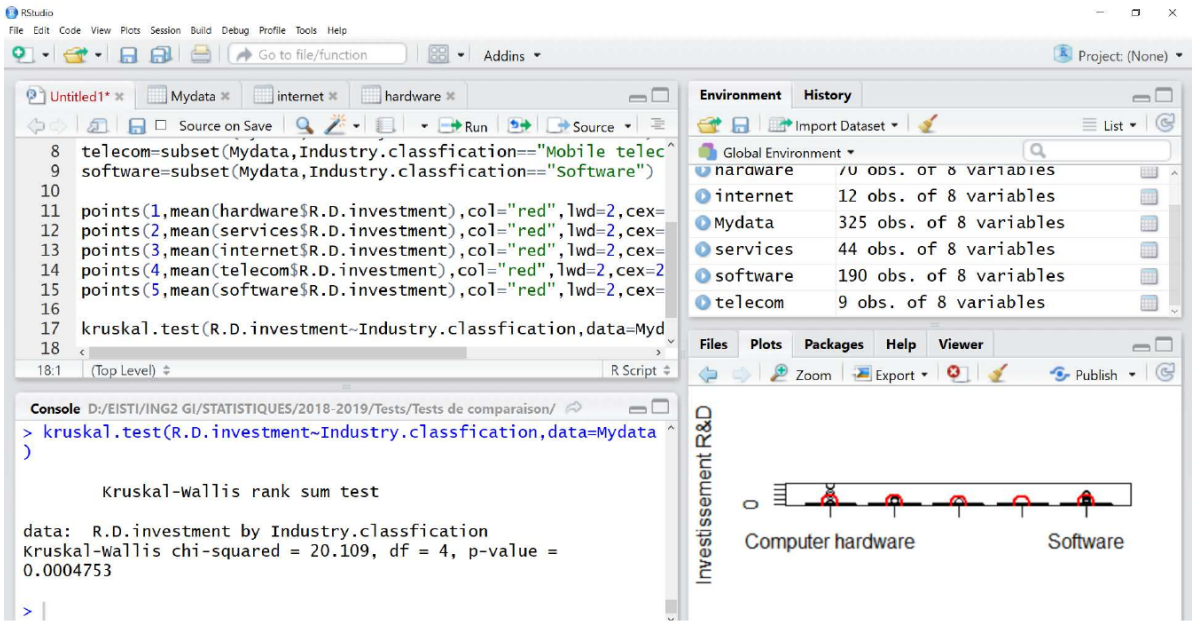
points(1,mean(hardware$R.D.investment),col=”red”,lwd=2,cex=2)
points(2,mean(services$R.D.investment),col=”red”,lwd=2,cex=2)
points(3,mean(internet$R.D.investment),col=”red”,lwd=2,cex=2)
points(4,mean(telecom$R.D.investment),col=”red”,lwd=2,cex=2)
points(5,mean(software$R.D.investment),col=”red”,lwd=2,cex=2)



Les distributions présentent des valeurs extrêmes. Les moyennes ne sont pas représentatives des échantillons. On ne va donc pas faire le test de l'ANOVA mais celui de Kruskal-Wallis.

(b)

kruskal.test(R.D.investment~Industry.classfication,data=Mydata)



La p-valeur du test est très petite (0.0004753) donc on rejette l'hypothèse (H_0) et on admet que le type d'industries a un impact significatif sur l'investissement R&D.

Si on avait effectué une ANOVA, on aurait obtenu une p-valeur de 0.7312 ce qui nous aurait amené à la conclusion inverse.