Master 1 IMSD - semestre 2 - le 29 janvier 2022

1er Compte Rendu des Travaux pratiques (Statistiques inférentielles et Modélisation)

langage de programmation utilisé : R
saisie de ce documment sur : LaTeX

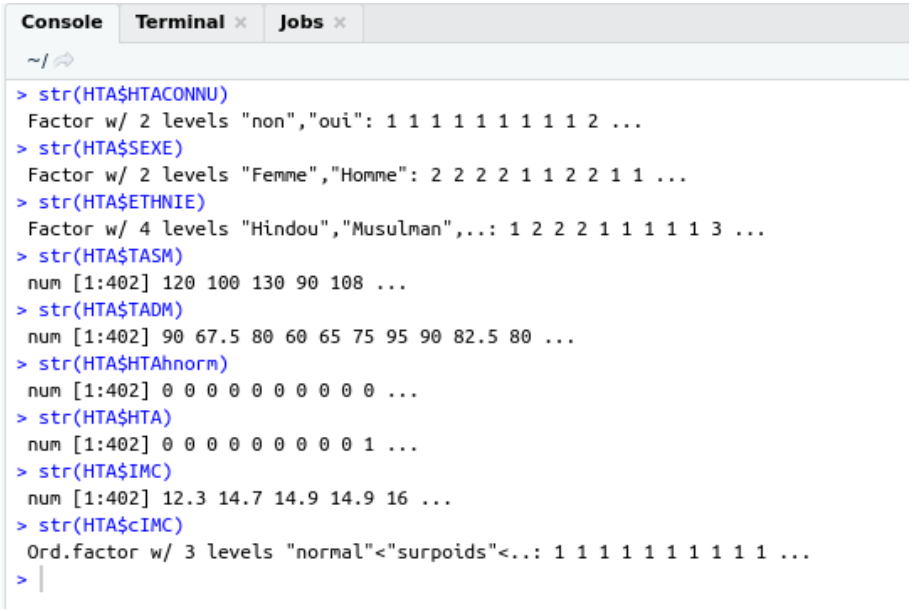
MOUDILA Marcel
MEUTCHOUNDJOU Thierry

1. Objectif du Tp

- Ouvrir un fichier de données
- Transformer les variables d'un fichier de données, en calculer de nouvelles
- Utiliser les commandes de R pour étudier le croisement de deux variables
- Donner les paramètres descriptifs du lien entre les deux variables
- Faire les tests correspondants
- Faire un test de comparaison de moyennes
- Faire un test du Chi-deux d'indépendance ou Fischer si besoin

2. Manipulation des variables

Ci-dessous la structure des variables manipulées :



- HTACONNU est qualitative avec 2 modalités : non, oui
- SEXE est qualitative avec 2 modalités : Femme, Homme
- ETHNIE est qualitative avec 4 modalités : Hindou, Musulman, Créole, Chinois
- cIMC est qualitative avec 3 modalités : normal, surpoids, obèse
- TASM, TADM, HTAnorm, et HTA sont quantitatives

3. Analyse conjointe de deux variables et tests correspondants

3.1 a)- Test du Chi-deux d'indépendance des variables ETHNIE3 et HTA

```
Console  Terminal  Jobs
~/
> ETHNIE3<-factor(ETHNIE3, labels=c('Hindou','Musulman','Creole'))
> t<-table(ETHNIE3,HTA$HTA[-which(HTA$ETHNIE== "Chinois")])
> addmargins(t)

ETHNIE3      0   1 Sum
  Hindou   161  64 225
  Musulman  58  19  77
  Creole    57  42  99
  Sum      276 125 401
> summary(t)
Number of cases in table: 401
Number of factors: 2
Test for independence of all factors:
        Chisq = 8.137, df = 2, p-value = 0.0171
>
```

- la p-value est égal à $0.0171 < 0.05$
- l'hypothèse nulle (les variables ETHNIE3 et HTA sont indépendants) est rejetée avec risque de première espèce $\alpha = 0.05$ de se tromper
- il y'a un lien entre l'hypertension et l'origine ethnique.

3.1 b)- Test du Chi-deux d'indépendance des variables SEXE et HTA

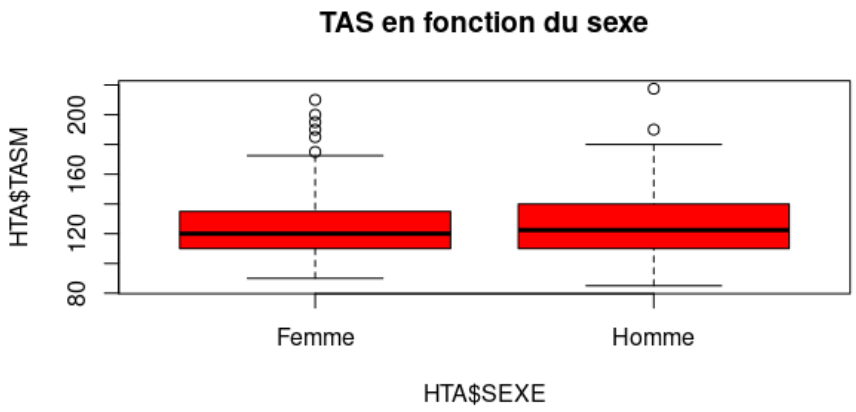
```
Console  Terminal  Jobs
~/
> t1 <-table(HTA$SEXE,HTA$HTA)
> round(prop.table(t1)*100,1)

           0    1
  Femme 42.0 17.9
  Homme 26.9 13.2
> summary(t1)
Number of cases in table: 402
Number of factors: 2
Test for independence of all factors:
        Chisq = 0.4173, df = 1, p-value = 0.5183
>
```

- le pourcentage d'hypertendus est de 17.9 % chez les femmes et de 13.2 % chez les hommes.
- il ne semble pas avoir de lien entre l'hypertension et le sexe.

- la p-value est égal à $0.51 > 0.05$, on ne rejette pas l'hypothèse nulle (les variables SEXE et HTA sont indépendantes

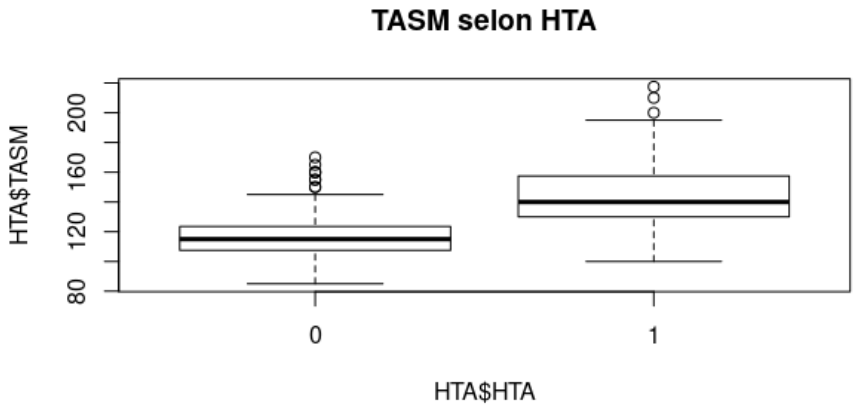
3.2 a)- Test de comparaison de moyenne de TASM chez les Femmes et chez les Hommes

```
Console  Terminal  Jobs
~/
> table(HTA$SEXE) # pour les effectifs des femmes et des hommes

Femme Homme
  241   161
> boxplot(HTA$TASM~HTA$SEXE,col="red",main= "TAS en fonction du sexe") # description
> t.test(HTA$TASM~HTA$SEXE)

        Welch Two Sample t-test

data:  HTA$TASM by HTA$SEXE
t = -1.7374, df = 349.59, p-value = 0.08321
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -8.1387805  0.5040804
sample estimates:
mean in group Femme mean in group Homme
           123.2199            127.0373

> sd(HTA$TASM[which(HTA$SEXE=='Femme')]) # ecart-type
[1] 21.94642
> sd(HTA$TASM[which(HTA$SEXE=='Homme')])
[1] 21.34263
>
```



- La TAS moyenne chez les femmes ($n = 241$) est de 123.22 ($sd = 21.95$).
- La TAS chez les hommes ($n = 161$) est de 127.04 ($sd = 21.34$)
- La différence n'est pas significative (p-value $= 0.08 > 0.05$, on ne rejette pas l'hypothèse nulle, on est sujette au risque de deuxième espèce)

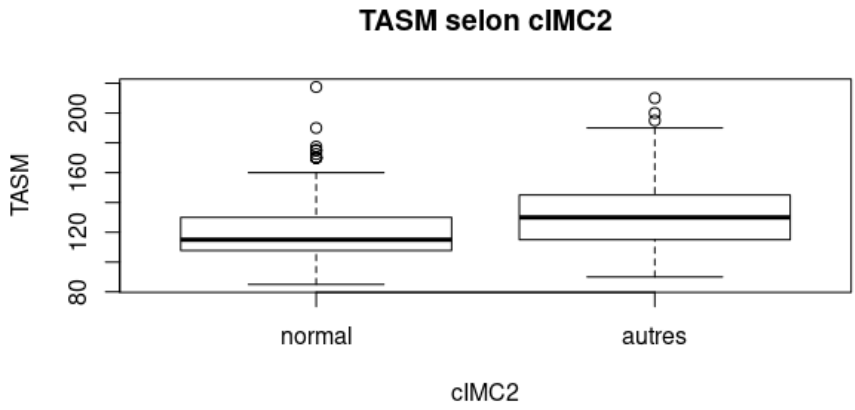
3.2 b)- Test de comparaison de moyenne de TASM chez les hypertendus et les non-hypertendus




```
Console   Terminal ×   Jobs ×
~/
> table(HTA$HTA)

  0   1
277 125
>
> boxplot(HTA$TASM~HTA$HTA,main="TASM selon HTA")
>
> t.test(HTA$TASM~HTA$HTA)

        Welch Two Sample t-test

data:  HTA$TASM by HTA$HTA
t = -12.909, df = 169.17, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -32.84707 -24.13314
sample estimates:
mean in group 0 mean in group 1
       115.8899        144.3800

> sd(HTA$TASM[which(HTA$HTA==0)])
[1] 14.21559
> sd(HTA$TASM[which(HTA$HTA==1)])
[1] 22.75318
>
```

- La TASM moyenne chez les hypertendus (n=125) est de 144.38 (sd= 22.75).
- La TASM moyenne chez les non-hypertendus(n=277) est de 115.90 (sd= 14.21)
- La différence est significative (p-value $= 2.2 \times 10^{-16} < 0.05$, on rejette l'hypothèse nulle)

3.2 c)- Test de comparaison de moyenne de TASM chez les IMC normal et les autres (surpoids + obèse)




```
Console   Terminal ×   Jobs ×
~/
> table(cIMC2)
cIMC2
normal autres
   260    142
> boxplot(TASM~cIMC2, main= "TASM selon cIMC2")
> t.test(HTA$TASM~cIMC2)

        Welch Two Sample t-test

data:  HTA$TASM by cIMC2
t = -5.3041, df = 256.8, p-value = 2.444e-07
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -16.620725  -7.620716
sample estimates:
mean in group normal mean in group autres
            120.4673             132.5880

>
```

- La TASM moyenne chez les personnes qui ont IMC "surpoids" ou "obèse" , n = 142, est de 132.59.
- La TASM moyenne chez les personnes qui ont IMC "normal", n = 260, est de 120.47.
- Le test est significatif car on rejette l'hypothèse nulle (p-value = $2.44.10^{-7} < 0.05$)

3.3 a)- Test de comparaison de deux moyennes appariées de TAS1 et TAS2

```
Console   Terminal   Jobs
~/
> mean(HTA$TAS1)
[1] 129.0274
> mean(HTA$TAS2)
[1] 125.408
> t.test(HTA$TAS1, HTA$TAS2, paired=TRUE)

        Paired t-test

data:  HTA$TAS1 and HTA$TAS2
t = 8.1037, df = 401, p-value = 6.507e-15
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 2.741368 4.497438
sample estimates:
mean of the differences
               3.619403

>
```

- La moyenne de TAS1 est plus élevée que la moyenne de TAS2.
- la p-value est égal à $6.50.10^{-15} < 0.05$, on rejette l'hypothèse nulle.

3.3 b)- Test de comparaison de deux moyennes appariées de TAD2 et TAD3

```
Console   Terminal   Jobs
~/
> mean(HTA$TAD2)
[1] 83.23881
> mean(HTA$TAD3)
[1] 82.75622
> L <- HTA$TAD2 - HTA$TAD3
> t.test(L)

        One Sample t-test

data:  L
t = 1.956, df = 401, p-value = 0.05116
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 -0.002436442  0.967610572
sample estimates:
mean of x
0.4825871

>
```

- La moyenne de TAD2 n'est pas significativement plus élevée que la moyenne de TAD3.
- la p-value est égal à $0.05116 > 0.05$, on ne pas rejette l'hypothèse nulle.

4. Analyse du fichier

4.1 - Description des variables IMC, cIMC, et ETHNIE

```
Console  Terminal  Jobs
~/
> summary(HTA$IMC)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  12.27   20.59   23.40   23.90   26.57   37.59
> table(HTA$cIMC)

  normal surpoids    obese
     260       94       48
> table(HTA$ETHNIE)

  Hindou Musulman   Créole  Chinois
     225       77       99        1
> tab1 <- table(HTA$cIMC)
> tab2 <- table(HTA$ETHNIE)
> round(prop.table(tab1)*100,1)

  normal surpoids    obese
    64.7     23.4     11.9
> round(prop.table(tab2)*100,1)

  Hindou Musulman   Créole  Chinois
    56.0     19.2     24.6      0.2
>
```

- Les valeurs de l'IMC sont comprises entre 12.27 (minimum) et 37.59 (maximum), la moyenne de l'IMC est de 23.90.

- On a 260 personnes (64.7 %) qui ont le statut pondéral **normal**, 94 personnes (23.4 %) qui sont en **surpoids**, et 48 personnes (11.9 %) qui sont **obèses**.

- En ce qui concerne l'ETHNIE, les Hindous (n=225, pourcentage = 56 %) sont plus représentés suivis des Créoles (n= 99, pourcentage = 24.6 %), et des Musulmans (n= 77, pourcentage = 19.2 %). L'ETHNIE chinois est sous représenté (n= 1, pourcentage = 0.2 %)

4.2 - test d'indépendance de cIMC et SEXE

```
Console  Terminal  Jobs
~/
> tab3 <- table(HTA$cIMC,HTA$SEXE)
> addmargins(tab3)

          Femme Homme Sum
  normal    137   123 260
  surpoids   65    29  94
  obese      39     9  48
  Sum       241   161 402
> round(prop.table(tab3)*100,1)

          Femme Homme
  normal   34.1  30.6
  surpoids 16.2   7.2
  obese     9.7   2.2
> summary(tab3)
Number of cases in table: 402
Number of factors: 2
Test for independence of all factors:
        Chisq = 18.087, df = 2, p-value = 0.0001182
>
```

- On a 34.1 % des femmes qui ont le statut pondéral normal vs 30.6 % des hommes.
- On a 16.2 % des femmes en surpoids vs 7.2 % des hommes.
- Enfin on a 9.7 % des des femmes obèses vs 2.2 % des hommes obèses.

il semble avoir un lien significatif entre le statut pondéral et le sexe.

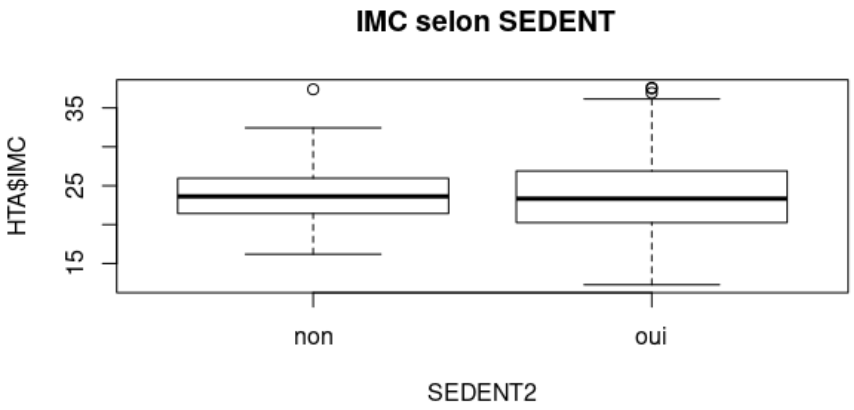
la p-value est $0.0001182 < 0.05$, on rejette l'hypothèse nulle : **le statut pondéral est lié au sexe**. On est sujette au risque de première espèce.

4.3 a)- test d'indépendance cIMC et SEDENT

```
Console  Terminal  Jobs
~/
> SEDENT2 <- HTA$SEDENT
> # marche plus de 5 miles par semaine (1,oui), sinon (0,non)
> SEDENT2 <- factor(SEDENT2,levels = c(0,1),labels = c("non","oui"))
> tab4 <- table(HTA$cIMC,SEDENT2)
> addmargins(tab4)
          SEDENT2
           non oui Sum
  normal    56 204 260
  surpoids  26  68  94
  obese      9  39  48
  Sum       91 311 402
> round(prop.table(tab4)*100,0)
          SEDENT2
           non oui
  normal    14  51
  surpoids   6  17
  obese      2  10
> summary(tab4)
Number of cases in table: 402
Number of factors: 2
Test for independence of all factors:
        Chisq = 1.9473, df = 2, p-value = 0.3777
>
```

- La p- value est $0.377 > 0.05$, on ne rejette pas l'hypothèse nulle **(le statut pondéral est indépendant du nombre de miles de marche par semaine)**

4.3 b)- Test de comparaison de moyenne IMC entre modalités de SEDENT



```
Console  Terminal  Jobs
~/
> table(SEDENT2)
SEDENT2
non oui
 91 311
> boxplot(HTA$IMC~SEDENT2,main =" IMC selon SEDENT")
> t.test(HTA$IMC~SEDENT2)

        Welch Two Sample t-test

data:  HTA$IMC by SEDENT2
t = 0.39923, df = 177.3, p-value = 0.6902
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.7619027  1.1483528
sample estimates:
mean in group non mean in group oui
         24.04622          23.85299

> sd(HTA$IMC[which(SEDENT2=="non")])
[1] 3.843472
> sd(HTA$IMC[which(SEDENT2=="oui")])
[1] 4.729267
>
```

- L'IMC moyen des personnes qui marchent plus de 5 miles par semaine (n = 311) est de 23.85 (sd= 4.72)
- L'IMC moyen des personnes qui ne marchent pas plus de 5 miles par semaine (n = 91) est de 24.04
- Le test n'est pas significatif, on ne rejette pas l'hypothèse nulle (p-value = 0.6902 > 0.05)
- L'IMC est le même chez les deux modalités de SEDENT. On est sujette au risque de deuxième espèce

4.4 a)- Intervalle de confiance de proportion d'HTAhnorm

```
Console  Terminal  Jobs
~/
> sum (HTA$HTAhnorm == 1)
[1] 45
> prop.test(45,402,0.11)

        1-sample proportions test with continuity correction

data:  45 out of 402, null probability 0.11
X-squared = 0.0019921, df = 1, p-value = 0.9644
alternative hypothesis: true p is not equal to 0.11
95 percent confidence interval:
 0.08363439 0.14788546
sample estimates:
        p
0.1119403

>
```

- 45 personnes ont HTAhnorm=1
- la p-value est 0.9644 > 0.05, on ne rejette pas l'hypothèse nulle (HTAhnorm suit la loi binomiale(45,0.11)) avec 0.11 une valeur estimée de la proportion p
- l'intervalle de confiance de p au niveau 95% est $[0.0836, 0.1478]$

4.4 b)- Intervalle de IMC

d'après 4.1 , on a le minimum et la maximum de l'IMC, donc l'intervalle de l'IMC est $[min(IMC), max(IMC)] = [12.27, 37.59]$

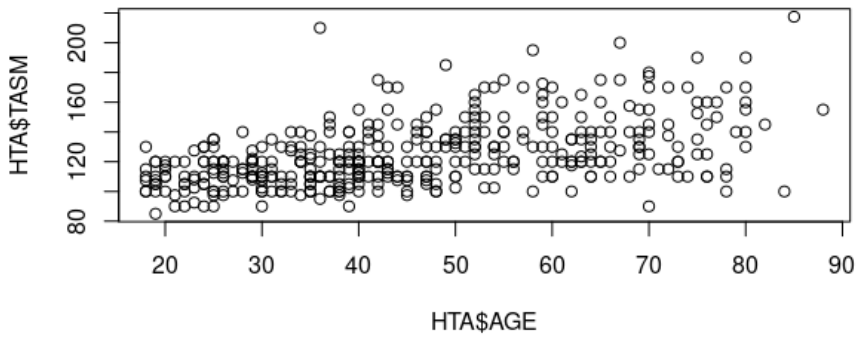
4.5 - Test d'indépendance de HTA et SEDENT

```
Console   Terminal ×   Jobs ×
~/
> tab5 <- table(HTA$HTA,HTA$SEDENT)
> addmargins(tab5)

       0   1 Sum
  0   58 219 277
  1   33  92 125
  Sum 91 311 402
> round(prop.table(tab5)*100,0)

    0  1
  0 14 54
  1  8 23
> summary(tab5)
Number of cases in table: 402
Number of factors: 2
Test for independence of all factors:
        Chisq = 1.467, df = 1, p-value = 0.2258
>
```

- 54 % des personnes qui marchent plus de 5 milles par semaine sont non-hypertendus , 23 % sont hypertendus.
- 14 % des personnes qui ne marchent ne marchent pas plus de 5 milles par semaine sont non-hypertendus, 8% sont hypertendus.
- La p-value est $0.2258 > 0.05$, on ne rejette pas l'hypothèse nulle(HTA et SEDENT sont indépendants)
- **il n'y a pas de lien entre l'hypertension et le nombre de miles de marche par semaine**.

4.6 a)- Test d'indépendance de AGE et TASM : test de nullité de coefficient de corrélation de Pearson

Car le test du Chi-deux semblait incorrecte

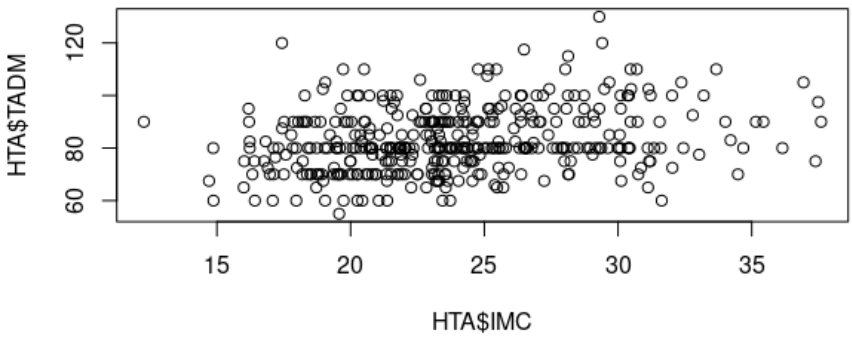


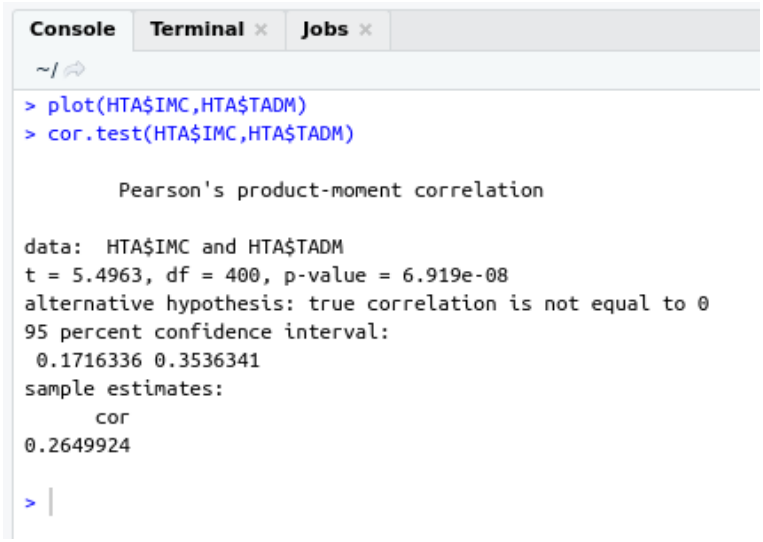
```
Console   Terminal ×   Jobs ×
~/
> summary(HTA$AGE)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  18.00   32.00   43.00   45.65   59.00   88.00
> summary(HTA$TASM)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   85.0   110.0   120.0   124.7   135.0   217.5
> plot(HTA$AGE,HTA$TASM)
> cor.test(HTA$AGE,HTA$TASM)

        Pearson's product-moment correlation

data:  HTA$AGE and HTA$TASM
t = 11.602, df = 400, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.4248147 0.5715315
sample estimates:
      cor
0.5017733

>
```

- La p-value est $2,2.10^{-16} < 0.05$, l'hypothèse nulle (le coefficient de corrélation est 0, i.e, AGE et TASM sont indépendants) est rejetée
- **il y'a un lien significatif entre l'âge et tasm**

4.6 b)- Test d'indépendance de IMC et TADM : test de nullité de coefficient de corrélation de Pearson

Car le test du Chi-deux semblait incorrecte





- La p-value est $6,9.10^{-08} < 0.05$, l'hypothèse nulle (le coefficient de corrélation est 0, i.e, IMC et TADM sont indépendants) est rejetée
- **il y'a un lien significatif entre l'indice de masse corporelle et tadm**