Correction Examen 2018

Guillaume Metzler

11/30/2021

Pour chaque question, nous procéderons toujours de la même façon :

- on illustrera graphiquement le problème
- on précisera le cadre du test statistiques ainsi que les hypothèses nulle et alternative
- on effectuera le test statistique
- on conclut en répondant à la question posée

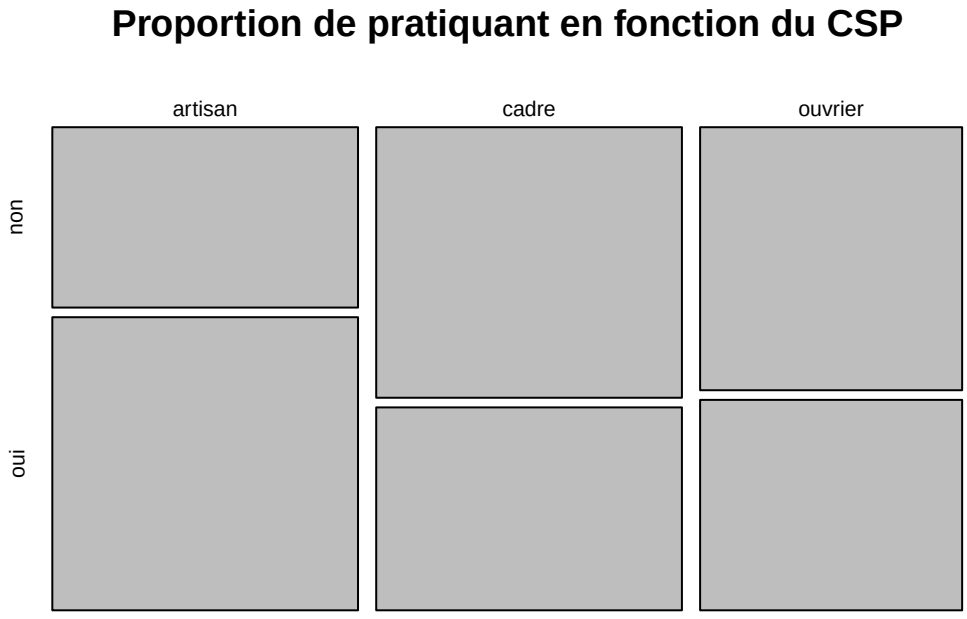
Commençons par charger les données

```
library(readr)
depense <- read.table("http://eric.univ-lyon2.fr/~jjacques/Download/DataSet/depense.txt",
                      header = TRUE)
depense <- data.frame(depense)
```

Question 1 : Le fait d'être pratiquant ou non dépend-il de la CSP du chef de famille ?

Représentons d'abord la situation à l'aide d'un graphique

```
plot(table(depense$csp,depense$pratiquant), main = "Proportion de pratiquant en fonction du CSP")
```



A priori, on voit bien que la proportion de pratiquant est la même selon la CSP du chef de famille. Confirmons cela à l'aide d'un

test du Khi-deux, où les hypothèses sont les suivantes :

- H_0 : Les deux caractères étudiés sont indépendants.
- H_1 : La CSM du chef de famille a un impact sur le fait d'être ou non pratiquant

On montre que les effectifs de notre table de contingence sont bien tous supérieurs à 5. Ce qui nous permet d'effectuer notre test du Khi-deux.

```
table(depense$csp,depense$pratiquant)
```

```
##
##           non oui
##   artisan   8  13
##   cadre    12   9
##   ouvrier  10   8
```

```
chisq.test(table(depense$csp,depense$pratiquant))
```

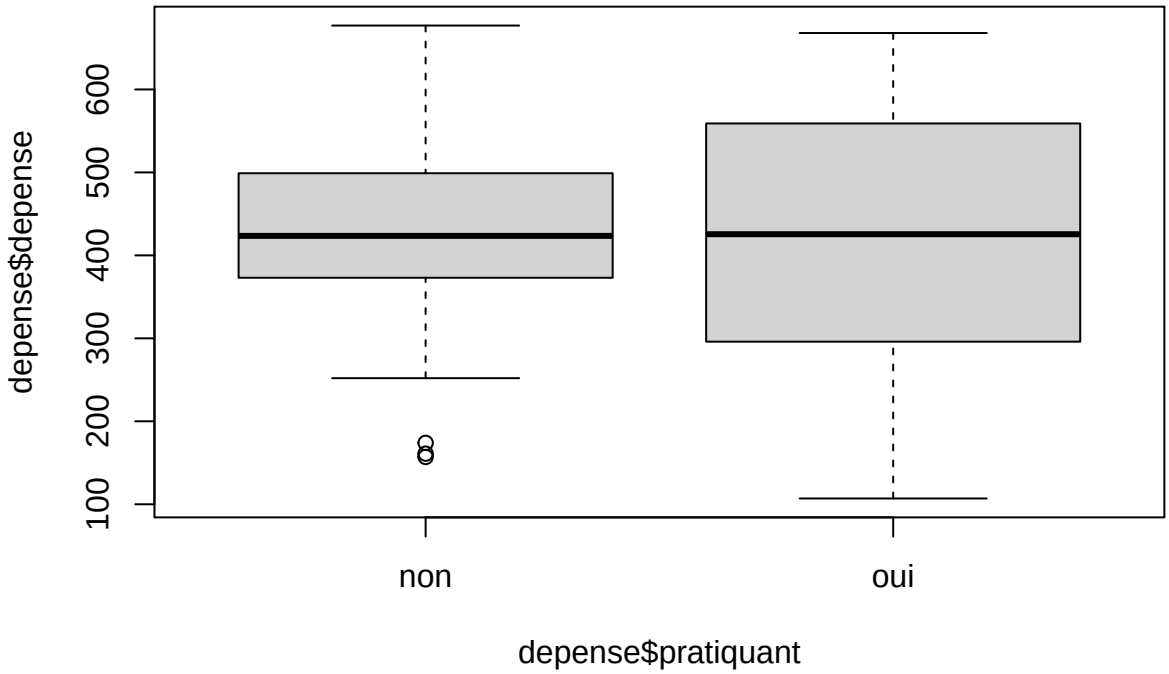
```
##
##  Pearson's Chi-squared test
##
## data:  table(depense$csp, depense$pratiquant)
## X-squared = 1.8413, df = 2, p-value = 0.3983
```

Notre p-value n'est pas significative ici, on ne peut pas donc pas affirmer que la CSP a un impact sur le fait que la famille soit ou non pratiquante.

Question 2 : Les familles pratiquantes dépensent-elles plus à Noël ?

On va représenter les dépenses des deux groupes sur un même graphique

```
boxplot(depense$depense~depense$pratiquant)
```



A nouveau, il n'y a pas de différence marquée en entre les deux groupes, il faudra confirmer tout cela avec un test sur la moyenne.

- H_0 : Les dépenses sont identiques que la famille soit ou non pratiquante
- H_1 : Les dépenses sont plus importantes chez les familles pratiquantes.

```
table(depense$pratiquant)
```

```
##
## non oui
##  30  30
```

Etant données les tailles de nos échantillons (>30), on va pouvoir faire un test de comparaison des moyennes pour deux échantillons indépendants en utilisant la loi de Student.
On prendra soin de vérifier si les variances sont ou non égales entre les deux groupes.

```
var.test(depense$depense~depense$pratiquant)
```

```
##
##  F test to compare two variances
##
## data:  depense$depense by depense$pratiquant
## F = 0.63311, num df = 29, denom df = 29, p-value = 0.2243
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.3013359 1.3301518
## sample estimates:
## ratio of variances
##          0.6331054
```

Les variances sont homogènes entre les deux groupes, ce que l'on précisera pour la réalisation de notre test de Student

```
t.test(depense$depense~depense$pratiquant, var.equal=TRUE, alternative = "less")
```

```
##
##  Two Sample t-test
##
## data:  depense$depense by depense$pratiquant
## t = -0.11993, df = 58, p-value = 0.4525
## alternative hypothesis: true difference in means between group non and group oui is less than 0
## 95 percent confidence interval:
##      -Inf 55.63363
## sample estimates:
## mean in group non mean in group oui
##             413.8             418.1
```

Nous ne pouvons donc pas affirmer que les familles pratiquantes dépensent plus à Noël que les familles non pratiquantes.

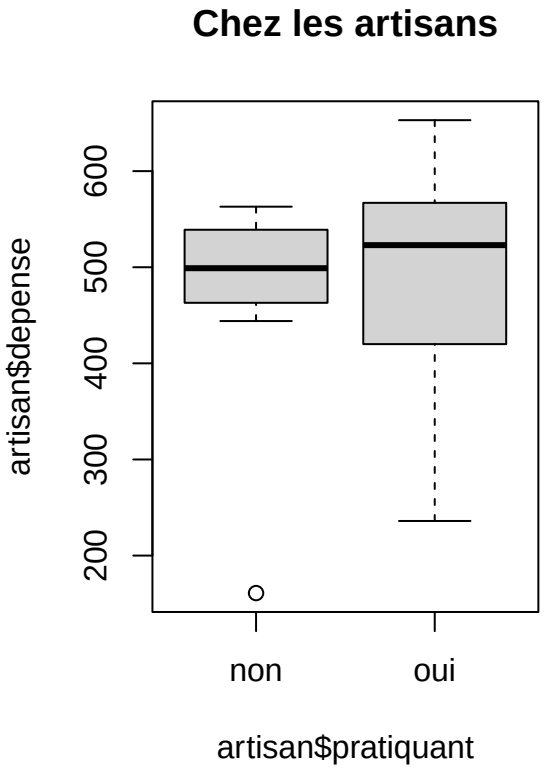
Question 3 : Chez les artisans, les familles pratiquantes dépensent-elles plus à Noël ? Quid chez les ouvrier ?

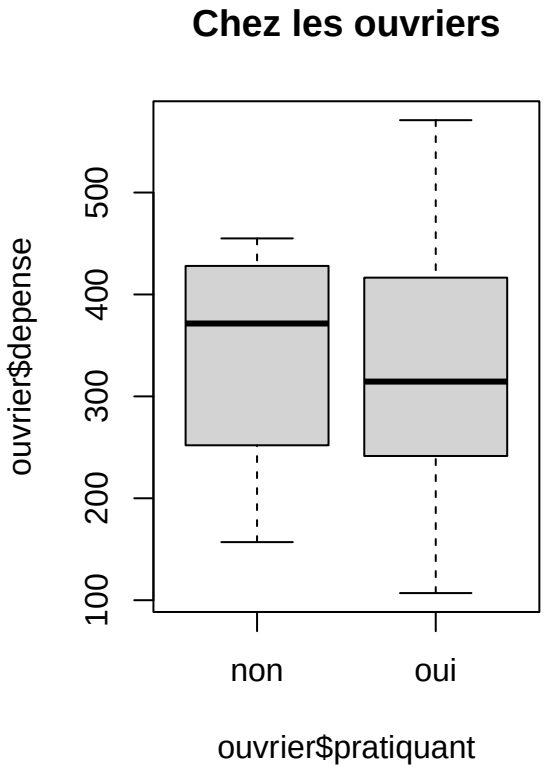
On commence par extraire les informations relatives aux deux populations étudiées et on représente les dépenses deux deux groupes étudiés à l'aide de deux boxplots.

```
artisan <- depense[depense$csp=="artisan",]
ouvrier <- depense[depense$csp=="ouvrier",]

par(mfrow = c(1,2))
```

```
boxplot(artisan$depense~artisan$pratiquant, main = "Chez les artisans")
boxplot(ouvrier$depense~ouvrier$pratiquant, main = "Chez les ouvriers")
```





- H_0 : Chez les artisans (resp. chez les ouvriers), les dépenses sont identiques que la famille soit ou non pratiquante
- H_1 : Chez les artisans (resp. chez les ouvriers), les dépenses sont plus importantes chez les familles pratiquantes.

Dans les deux cas, nous devons procéder à un test de comparaison des moyennes. Il reste à savoir qu'elle sera le test le plus approprié.

Chez les artisans

On regarde les effectifs

```
table(artisan$pratiquant)
```

```
##
## non oui
##   8  13
```

Comme l'effectif de l'un des groupes est inférieur à 10 on va directement procéder à un test de comparaison des moyennes non paramétriques, le test de Wilcoxon.

```
wilcox.test(artisan$depense~artisan$pratiquant, alternative = "less")
```

```
##
##  Wilcoxon rank sum exact test
##
## data:  artisan$depense by artisan$pratiquant
## W = 44, p-value = 0.2975
```

```
## alternative hypothesis: true location shift is less than 0
```

A priori les dépenses sont, en moyenne, identiques chez les ouvriers que la famille soit ou non pratiquante.

chez les ouvriers

On regarde les effectifs

```
table(ouvrier$pratiquant)
```

```
##
## non oui
##  10   8
```

Comme l'effectif de l'un des groupes est inférieur à 10 on va directement procéder à un test de comparaison des moyennes non paramétriques, le test de Wilcoxon.

```
wilcox.test(ouvrier$depense~ouvrier$pratiquant, alternative = "less")
```

```
## Warning in wilcox.test.default(x = c(429L, 334L, 252L, 270L, 174L, 419L, :
## cannot compute exact p-value with ties
```

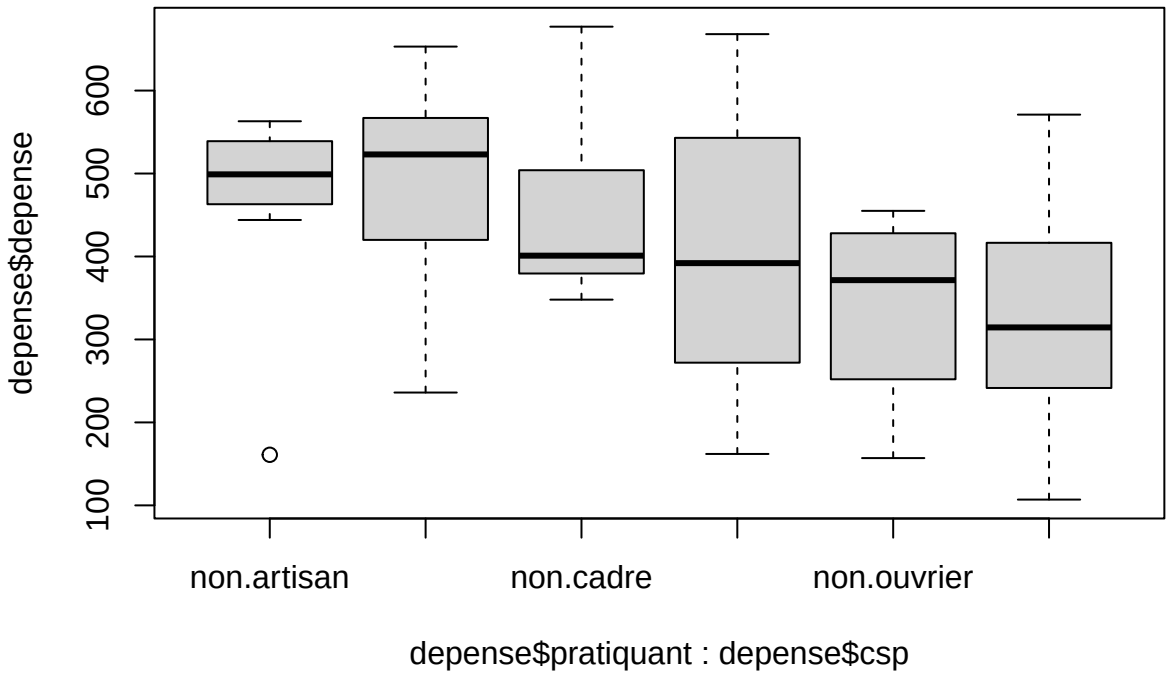
```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  ouvrier$depense by ouvrier$pratiquant
## W = 43.5, p-value = 0.6389
## alternative hypothesis: true location shift is less than 0
```

A priori les dépenses sont, en moyenne, identiques chez les ouvriers que la famille soit ou non pratiquante.

Question 4 : Analyser l'influence de la CSP du chef de famille et du fait d'être ou non pratiquant sur le montant des dépenses.

On va à nouveau représenter la distribution des dépenses au sein des différents groupes.

```
boxplot(depense$depense~depense$pratiquant+depense$csp)
```



- H_0 : Les dépenses sont indépendantes des facteurs étudiés.
- H_1 : Les facteurs étudiés ont un impact sur le montant des dépenses.

On va donc devoir procéder à une Analyse de Variance à deux facteurs. On va commencer par les vérifications standards, à savoir, l'homogénéité des variances entre les deux groupes et le caractère gaussien des des différents groupes.

Caractère Gaussien

```
table(depense$csp,depense$pratiquant)
```

```
##
##           non oui
##   artisan   8  13
##   cadre    12   9
##   ouvrier  10   8
```

Nos échantillons sont trop petits, on va donc écarter l'hypothèse gaussienne. Bien que cela soit une hypothèse importante de l'ANOVA, on va surtout regarder si nos résidus sont gaussiens avant de procéder à un éventuel usage de Kruskal.

Homoscédasticité

```
bartlett.test(depense$depense~depense$csp)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  depense$depense by depense$csp
## Bartlett's K-squared = 0.073103, df = 2, p-value = 0.9641
```

```
bartlett.test(depense$depense~depense$pratiquant)
```

```
##
```

```
##  Bartlett test of homogeneity of variances
##
## data:  depense$depense by depense$pratiquant
## Bartlett's K-squared = 1.4765, df = 1, p-value = 0.2243
```

Les résultats de ces deux tests nous montrent que les variances sont bien homogènes entre les différents groupes.

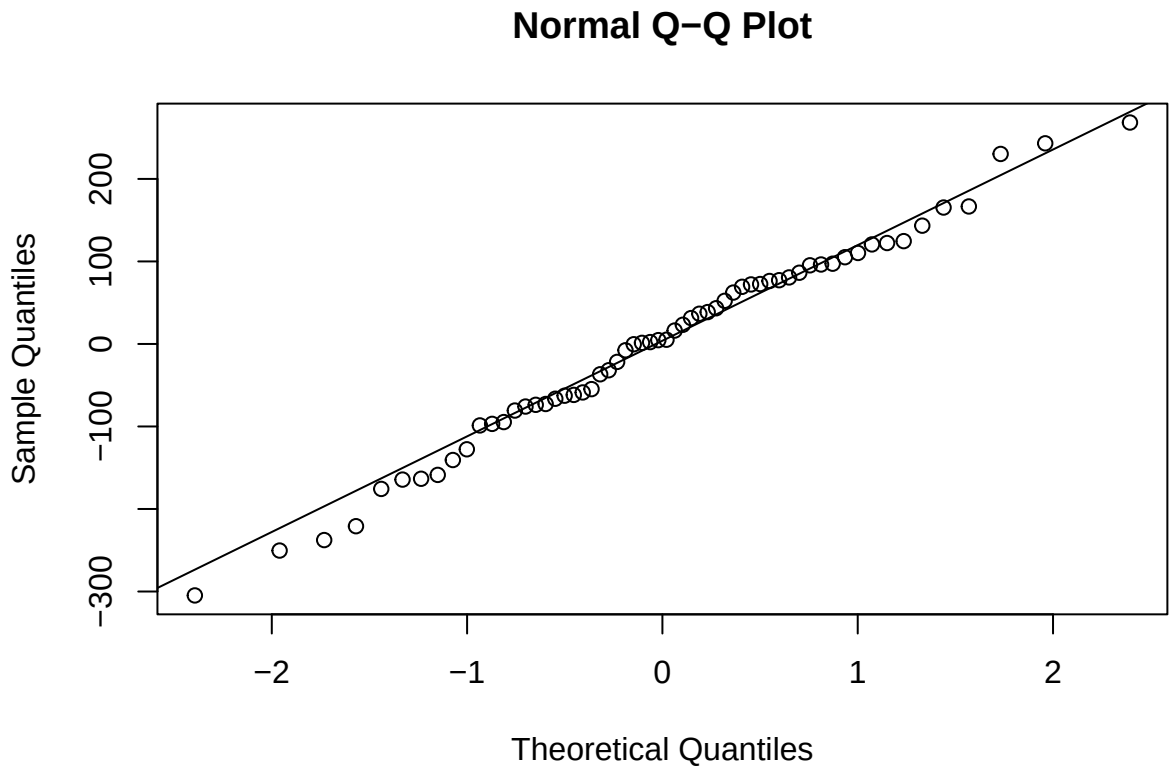
ANOVA

```
my_model <- aov(depense~csp*pratiquant,data=depense)
summary(my_model)
```

```
##                Df Sum Sq Mean Sq F value  Pr(>F)
## csp             2 215854  107927   6.554 0.00283 **
## pratiquant      1   1792    1792   0.109 0.74280
## csp:pratiquant  2  11821    5911   0.359 0.70008
## Residuals      54 889270   16468
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

D'après les résultats de l'ANOVA, seul le facteur csp a un impact sur les dépenses. Vérifions que notre ANOVA est bien valide en regardant la normalité des résidus

```
qqnorm(my_model$residuals)
qqline(my_model$residuals)
```




```
shapiro.test(my_model$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
```

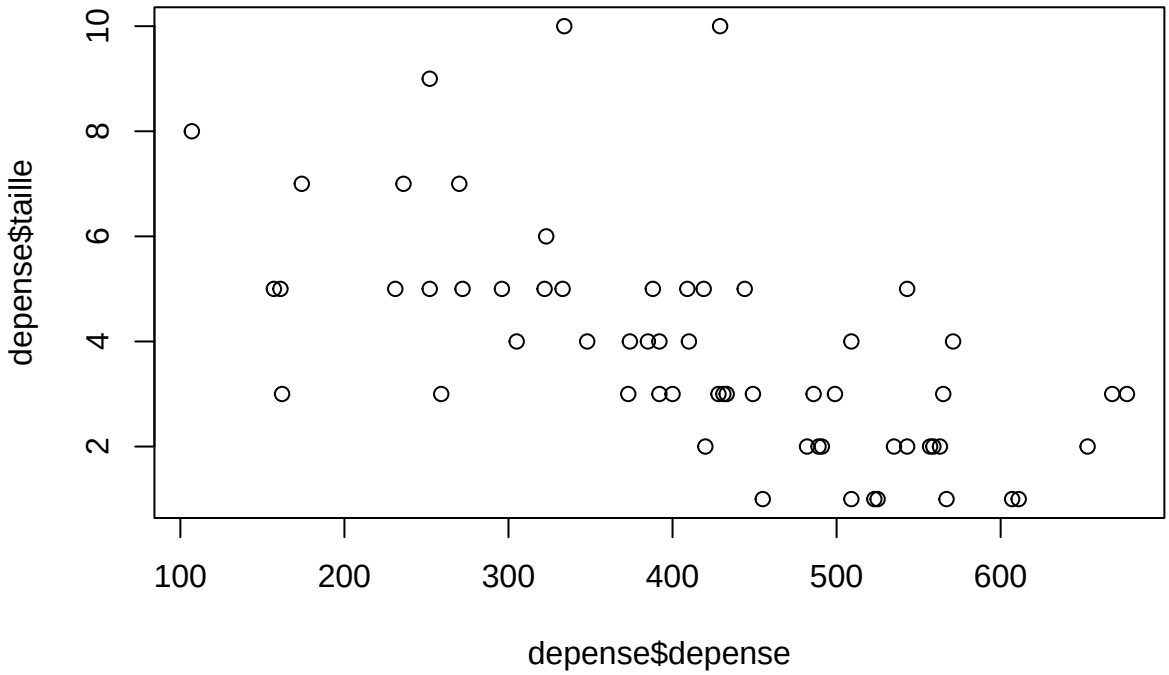
```
## data:  my_model$residuals
## W = 0.98762, p-value = 0.8032
```

Nos résidus sont biens gaussiens on peut donc valider les résultats de notre ANOVA.

Question 5 : Le montant des dépenses dépend-il de la taille de la famille ?

Représentons le montant des dépenses en fonction de la taille de la famille.

```
plot(depense$depense,depense$taille)
```



On a une tendance linéire qui semble plutôt négative si on regarde ce graphe, on va essayer de confirmer cela en procédant à un test de corrélation pour savoir si le lien entre les deux variables étudiées est significatif.

- H_0 : Le montant des dépenses dépend de la taille de la famille
- H_1 : Le montant des dépenses ne dépend pas de la taille de la famlle

```
cor.test(depense$depense,depense$taille)
```

```
##
##  Pearson's product-moment correlation
##
## data:  depense$depense and depense$taille
## t = -6.0512, df = 58, p-value = 1.122e-07
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.7565151 -0.4372365
## sample estimates:
##        cor
## -0.6220932
```

La p-value indique que la corrélation entre les deux variables est significative. De plus cette corrélation est bien négative comme l'indique la valeur du coefficient de corrélation.