SI-M1-TD4

Guillaume Metzler

10/25/2021

Exercice 1

Commençons par charger les données

```
load('~/Desktop/tempsTV.Rdata')
data$sexe[data$sexe==1]="homme"
data$sexe[data$sexe==2]="femme"
```

Dans ce premier exercice, on se demande si les variables “periode” et “sexe” qui sont deux variables qualitatives ont une influence sur la variable “temps” (variable quantitative).

On va donc poser les hypothèses suivantes :

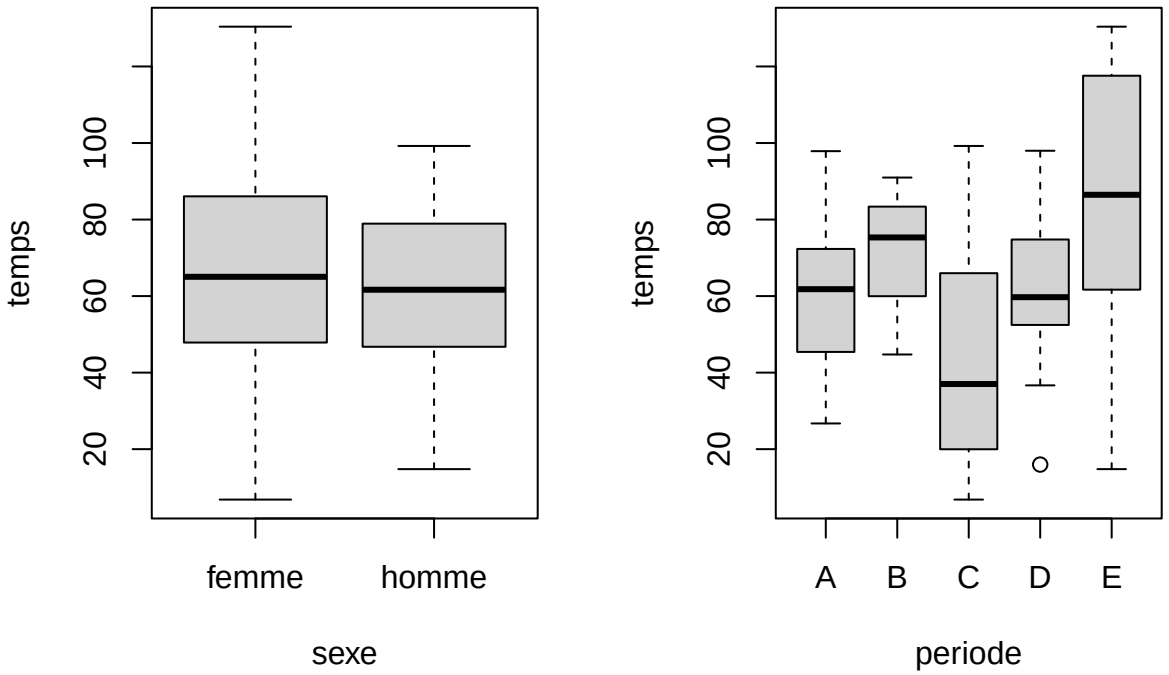
- H_0 : la variable "temps" est indépendante des deux autres facteurs v.s.
- H_1 : la variable "temps" est dépendante d’au moins un des autres facteurs.

La formulation gagnerait à être plus précise car la variable temps peu dépendre d’un seul des deux facteurs ou encore des deux facteurs.

Dans tous les cas, etant donné que nous devons étudier l’influence de deux facteurs sur une variable quantitative, nous devons effectuer une Analyse de Variance (ANOVA).

Regardons graphiquement si le sexe a une influence sur la variable temps

```
par(mfrow=c(1,2))
boxplot(temps~sexe,data=data)
boxplot(temps~periode,data=data)
```



Graphiquement, il semblerait que seule la variable "periode" ait une infleucne sur la variable temps. Il faudra donc vérifier cela à l'aide de notre ANOVA.

Remarque L'Analyse de Variance consiste à une étude de la variance de notre jeu de données afin de déterminer la part de la variance expliquée par les différents facteurs. Par exemple, pour une ANOVA à un facteur, la variance totale de notre jeu de données V_T peut s'exprimer à partir de la variance expliquée par le facteur A, V_A et une variance résiduelle V_R, *{i.e.} non expliquée par le facteur A :*

$$V_T = V_A + V_R$$

Pour une ANOVA à deux facteurs (disons A et B) il faut donc regarder la part de variance expliquée par le facteur A et la part de variance expliquée par le facteur B. Si on reprend notre formule précédente, nous pouvons alors décomposer notre variance résiduelle V_R *en* $V_R = V_B + \tilde{V_R}$*. Dit autrement, une part de la variance non expliquée précédemment l'est en fait par le facteur B.*

$$V_T = V_A + V_B + \tilde{V_R}.$$

Or l'importance d'un facteur est estimée en fonction du rapport entre la variance expliquée par ce facteur et la variance résiduelle du modèle. Cela suggère bien qu'il est important d'effectuer une ANOVA à deux facteurs et non deux ANOVA à un facteur.

Regardons cela avec notre test

```
res=aov(temps~sexe+periode,data=data)
print(anova(res))
```

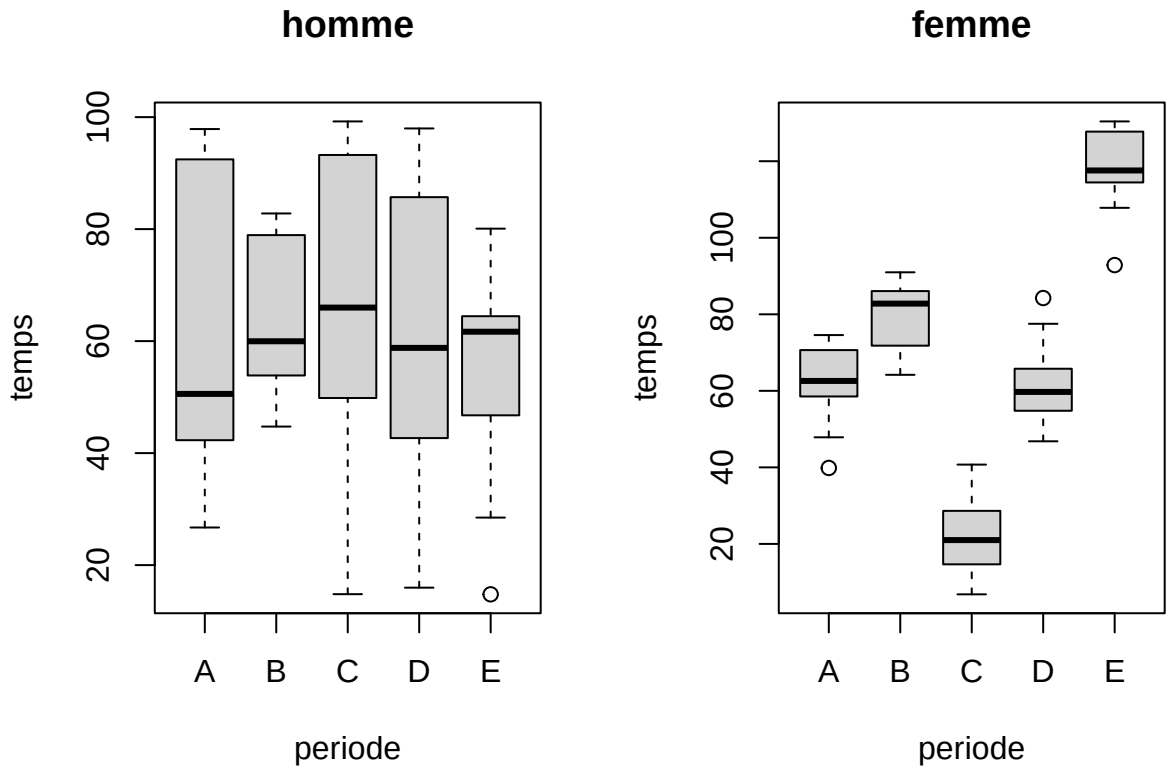
```
## Analysis of Variance Table
##
## Response: temps
##           Df Sum Sq Mean Sq F value    Pr(>F)
## sexe       1   1373  1372.7  2.2303    0.1387
## periode    4  19488  4872.1  7.9163 1.523e-05 ***
```

```
## Residuals 94  57852   615.5
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

La variable "periode" a donc une importance significative sur la variable "temps" mais ce n'est pas le cas de la variable "sexe".

Regardons maintenant ce qui se passe chez les hommes et chez les femmes

```
par(mfrow=c(1,2))
datah=data[data$sexe=="homme",]
boxplot(temps~periode,data=datah,main='homme')
dataf=data[data$sexe=="femme",]
boxplot(temps~periode,data=dataf,main='femme')
```



*Remarquons que le comportement n'est pas du tout le même chez les hommes et chez les femmes. Le genre semble donc avoir un effet sur le lien entre les variables "temps" et "période". C'est ce que l'on appelle un {***effet d'intéraction entre les deux facteurs***} de notre ANOVA. Ainsi, dans une ANOVA à plusieurs facteurs, il est donc important de prendre en compte les effets d'intéractions entre les différents facteurs. Ce que l'on fait de la façon suivante :*

```
res=aov(temps~sexe+periode+sexe*periode,data=data)
print(anova(res))
```

```
## Analysis of Variance Table
##
## Response: temps
##               Df  Sum Sq Mean Sq F value    Pr(>F)
## sexe           1  1372.7  1372.7  4.1502   0.04457 *
## periode        4 19488.4  4872.1 14.7305 2.704e-09 ***
## sexe:periode   4 28085.0  7021.3 21.2284 2.359e-12 ***
```

```
## Residuals    90 29767.3   330.7
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

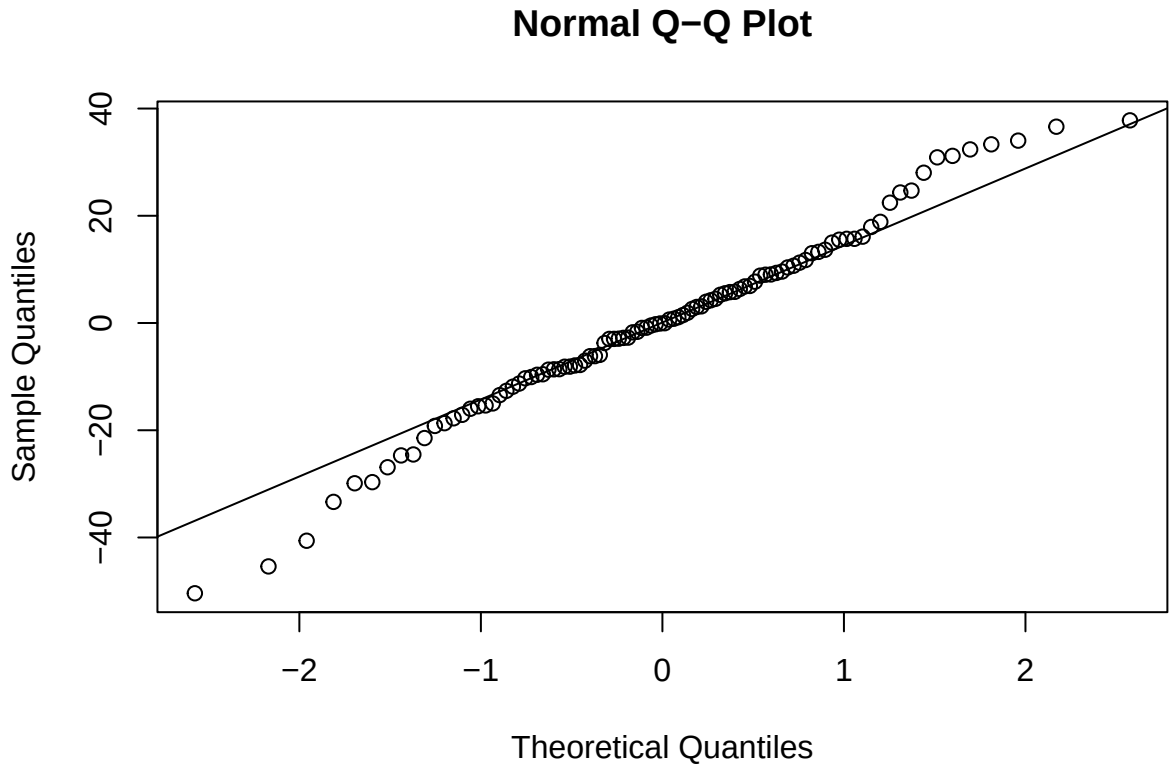
On constate que cette interaction est en effet significative.

Il reste tout de même une chose à vérifier ... avons-nous le droit de faire cette ANOVA ? Est-ce que toutes les conditions sont réunies ?

- **Ici, en ANOVA à deux facteurs, les échantillons doivent être gaussiens pour chque croisement des deux facteurs. Si on regarde la taille de notre échantillon, on remarque que ces derniers seront trop petits pour que l'on puisse utiliser un test de Shapiro qui ne sera alors pas suffisamment puissant. On se contentera donc, pour une ANOVA à deux facteurs, de vérifier uniquement la normalité des résidus.**
- **des variances homogènes pour chaque facteur.**

Commençons par regarder la normalité des résidus.

```
qqnorm(res$residuals)
qqline(res$residuals)
```




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

```
##
##  Shapiro-Wilk normality test
##
## data:  res$residuals
## W = 0.98482, p-value = 0.3081
```

On peut accepter cette hypothèse de normalité des données car la p-value est supérieure au rsique de première espèce $\alpha = 0.05$ Pour ce qui est de l'homogénéité des variances, on va la

tester avec un test de Bartlett pour chacun des deux facteurs.

```
bartlett.test(temps~periode,data=data)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  temps by periode
## Bartlett's K-squared = 19.856, df = 4, p-value = 0.0005331
```

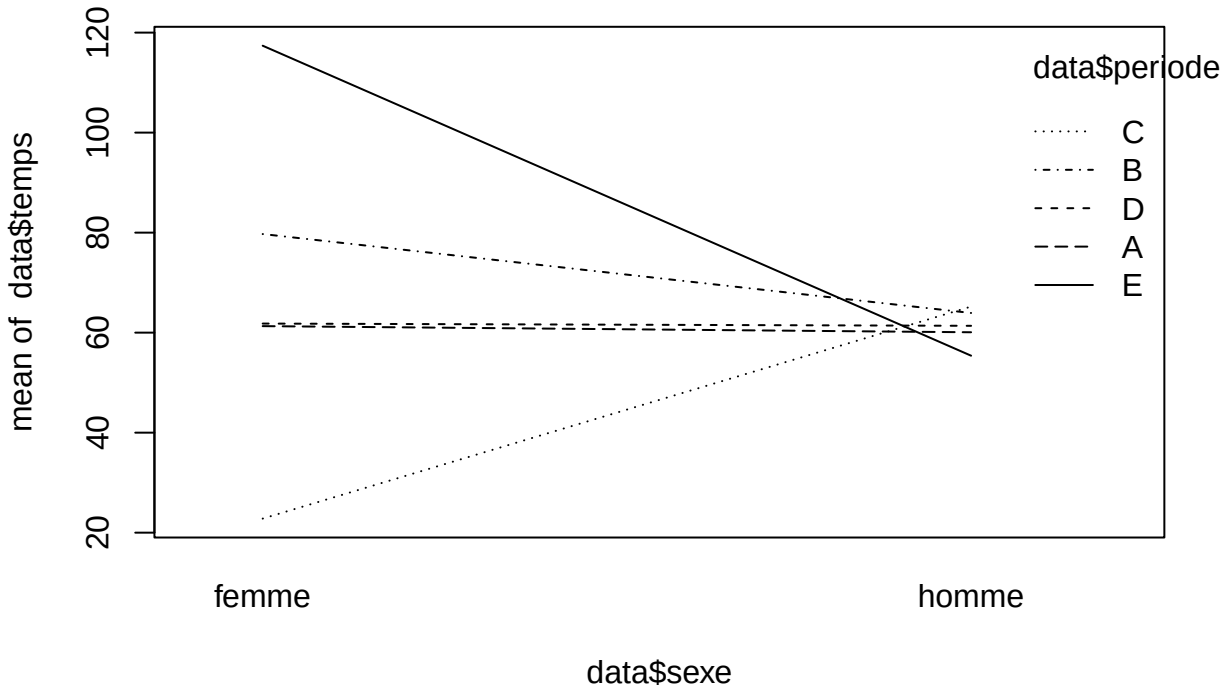
```
bartlett.test(temps~sexe,data=data)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  temps by sexe
## Bartlett's K-squared = 6.239, df = 1, p-value = 0.0125
```

L'homogénéité des variances n'est pas respectée ici. En effet, la p-value est ici bien inférieure à 0.05 pour les deux facteurs. Néanmoins, on considérera que l'ANOVA reste valable car elle est suffisamment robuste à la non homogénéité des variances lorsque les effectifs diffèrent peu d'une modalité à une autre, ce qui est le cas ici.
En outre, il n'existe pas d'alternative non paramétrique à l'ANOVA à deux facteurs.

On peut regarder le graphique ci-dessous qui nous montre comment évolue les différentes moyennes en fonction des différents facteurs.

```
interaction.plot(data$sexe,data$periode,data$temps)
```



On peut aussi plus en détails en faisait des tests deux à deux entre les modalitésn au niveau des périodes. Pour palier aux problématiques de multiplicité des tests, on utilise des corrections comme la correction de Bonferronni

```
pairwise.t.test(data$temps,data$periode,p.adjust.method = "bonferroni")
```

```
##
```

```
##  Pairwise comparisons using t tests with pooled SD
##
## data:  data$temps and data$periode
##
##   A      B      C       D
## B 1.0000 -      -       -
## C 0.3731 0.0066 -       -
## D 1.0000 1.0000 0.2837  -
## E 0.0158 0.6834 5.7e-06 0.0225
##
## P value adjustment method: bonferroni
```

Les périodes B et C sont significativement différentes alors que A et C ne le sont pas.

Exercice 2

On se propose de travailler à nouveau sur le fichier GermanCredit afin de répondre à plusieurs questions. Cet exercice est très proche de ce qui vous attendra à l'examen.

```
library(readr)
data <- read_table2("http://eric.univ-lyon2.fr/~jjacques/Download/DataSet/GermanCredit.data",
    col_names = FALSE)
```

```
## Warning: `read_table2()` was deprecated in readr 2.0.0.
## Please use `read_table()` instead.
```

```
##
## -- Column specification --------------------------------------------------------
## cols(
##   .default = col_character(),
##   X2 = col_double(),
##   X5 = col_double(),
##   X8 = col_double(),
##   X11 = col_double(),
##   X13 = col_double(),
##   X16 = col_double(),
##   X18 = col_double(),
##   X21 = col_double()
## )
## i Use `spec()` for the full column specifications.
```

Question 1

On cherche à savoir si le sexe (variable quali) à une influence sur le montant emprunté (variable quanti) et plus précisément si les femmes empruntent un montant plus important que les hommes. On formule alors les hypothèses suivantes :

- H_0 : **le sexe n'a pas d'influence sur le montant emprunté :** $\mu_F = \mu_H$
- H_1 : **le sexe a une influence sur le montant emprunté et les femmes empruntent plus que les hommes :** $\mu_F > \mu_H$

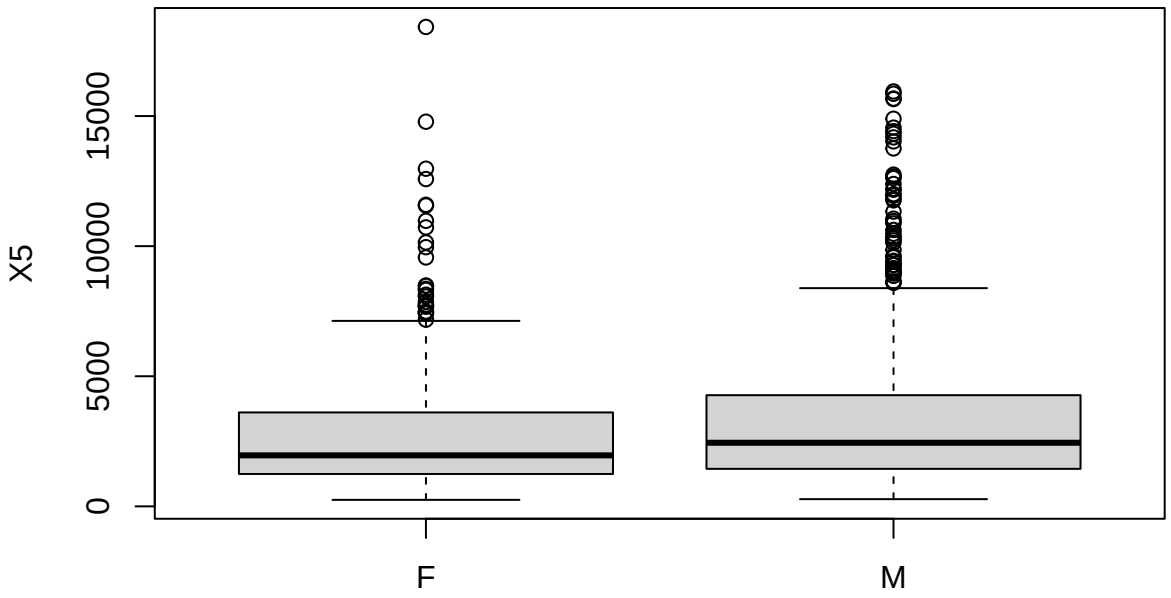
L'hypothèse alternative suggère de faire un test unilatéral, on verra selon l'ordre d'appararition des modalités, si c'est un test unilatéral supérieur ou inférieur. Commençons d'abord par une représentation graphique

```
# On recode la variable
```

```
sexe = as.vector(data$X9)
sexe[ sexe=="A92" | sexe=="A95" ]='F'
sexe[ sexe!="F" ]='M'
data=data.frame(data,sexe=as.factor(sexe))

# On fait nos graphes

boxplot(X5~sexe,data)
```



On va regarder la taille des échantillons pour voir s'il est nécessaire de faire ou non de vérifier le caractère gaussien des données des différents groupes

```
table(sexe)
```

```
## sexe
##   F   M
## 310 690
```

Nos échantillons sont suffisamment grands pour que cette vérification ne soit pas nécessaire.

Le boxplot suggère que les femmes empruntent moins que les hommes en moyenne. On va donc faire le test de Student suivant après avoir vérifier l'homogénéité des variances

```
var.test(X5~sexe,data)
```

```
##
##  F test to compare two variances
##
## data:  X5 by sexe
## F = 0.80548, num df = 309, denom df = 689, p-value = 0.02863
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.6683860 0.9774795
## sample estimates:
```

```
## ratio of variances
##          0.8054799
```

Les variances sont significativement différentes

```
t.test(X5~sexe,data,var.equal=FALSE, alternative = "less")
```

```
##
##  Welch Two Sample t-test
##
## data:  X5 by sexe
## t = -3.0904, df = 658.03, p-value = 0.001042
## alternative hypothesis: true difference in means between group F and group M is less than 0
## 95 percent confidence interval:
##       -Inf -266.3121
## sample estimates:
## mean in group F mean in group M
##        2877.774        3448.041
```

On peut donc bien affirmer que les femmes, en moyenne, empruntent un montant bien inférieur à celui des hommes.

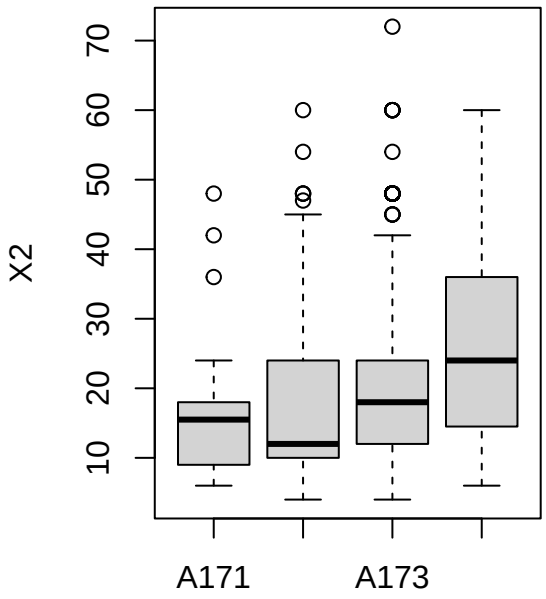
Question 2

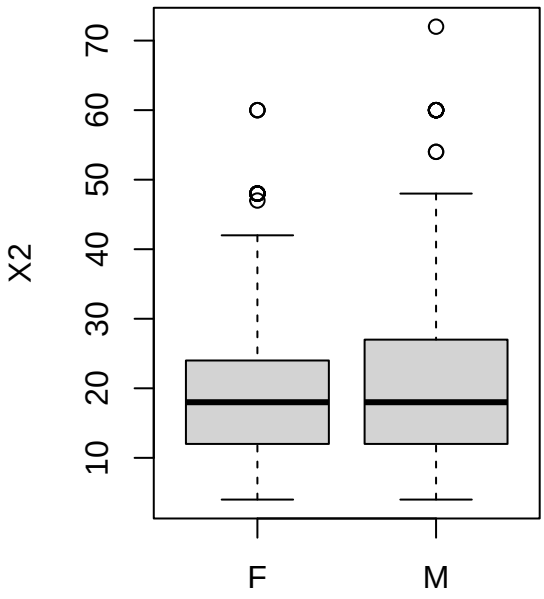
On cherche à savoir si l'emploi (variable quali) et le sexe (variable quali) ont une influence sur la durée de l'emprunt (variable quanti). On formule alors les hypothèses suivantes :

- H_0 : **les deux facteurs n'ont pas d'influence sur la durée de l'emprunt v.s.**
- H_1 : **les facteurs ont une infleunce sur la durée de l'emprunt**

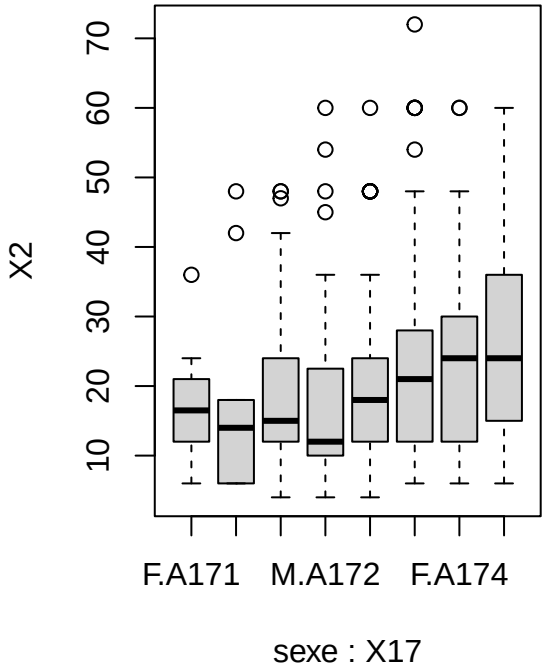
On va à nouveau réaliser une ANOVA à deux facteur dans ce cas là. Mais regardons d'abord nos graphiques

```
par(mfrow=c(1,2))
boxplot(X2~X17,data)
boxplot(X2~sexe,data)
```






```
boxplot(X2~sexe+X17,data)
```



Faisons maintenant notre test

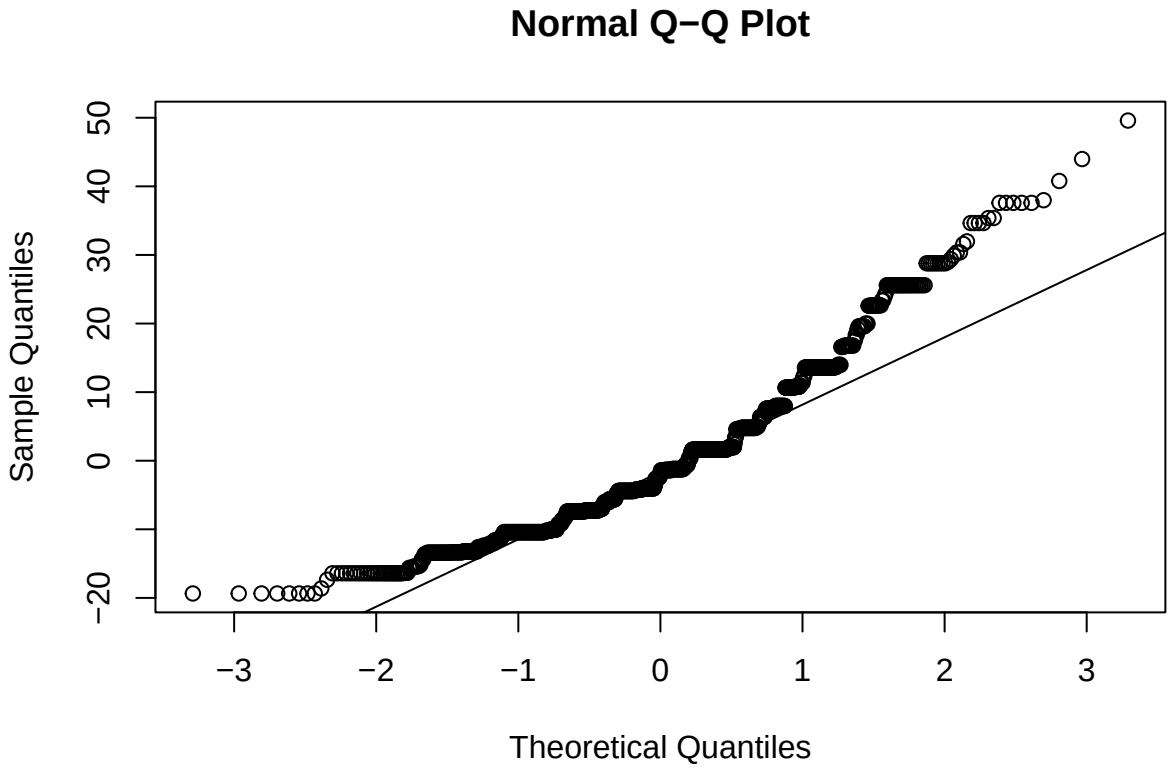
```
modele=aov(X2~sexe*X17,data=data)
summary(modele)
```

```
##              Df Sum Sq Mean Sq F value   Pr(>F)
## sexe          1    963   963.3   6.985  0.00835 **
## X17           3   6701  2233.5  16.195 2.78e-10 ***
## sexe:X17      3    798   266.1   1.930  0.12309
## Residuals   992 136807   137.9
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

On peut donc affirmer que le sexe et l'emploi ont une influence sur la durée de l'emprunt mais pas l'intéraction de ces deux facteurs. Il nous faut maintenant regarder si les conditions d'application sont réunies. Nous avons dit que dans le cas d'une ANOVA à deux facteurs, on va uniquement étudier la normalité des résidus.

```
qqnorm(modele$residuals)
qqline(modele$residuals)
```

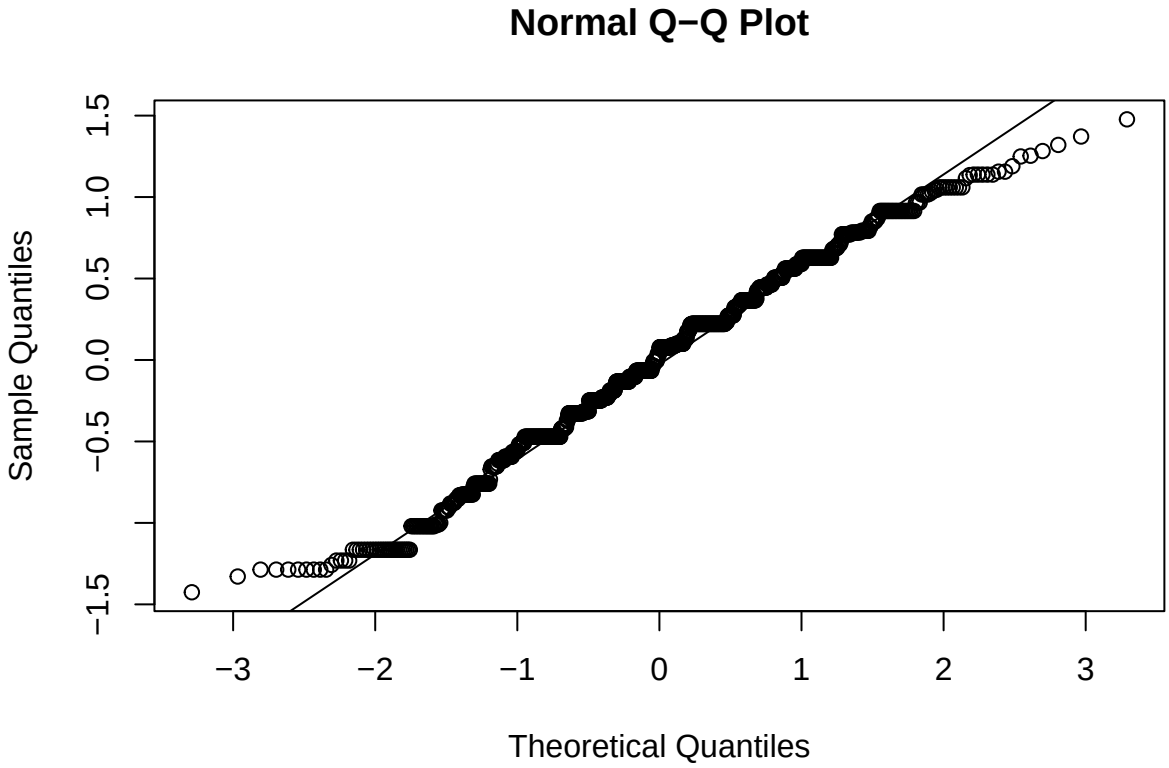


A priori, les résidus sont loins de se retrouver sur la droite de Henry ... on va essayer de regarder ce qui se pase si l'on fait une ANOVA sur le log de la durée.

```
modele=aov(log(X2)~sexe*X17,data=data)
summary(modele)
```

```
##              Df Sum Sq Mean Sq F value   Pr(>F)
## sexe          1    1.7   1.678   5.242   0.0223 *
## X17           3   17.2   5.750  17.959 2.35e-11 ***
## sexe:X17      3    2.1   0.713   2.228   0.0834 .
## Residuals   992  317.6   0.320
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
qqnorm(modele$residuals)
qqline(modele$residuals)
```




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

```
##
##  Shapiro-Wilk normality test
##
## data:  (modele$residuals)
## W = 0.99031, p-value = 3.692e-06
```

Nos résidus ne sont toujours pas gaussiens dans ce cas là ... enfin d'après le test de Shapiro, or ce dernier est un test extrêmement puissant ! Il notamment très strict en présence de très grands échantillons, ce qui est le cas ici. On va donc supposer que nos résidus sont gaussiens. Il faudrait encore tester l'homogénéité des variances avec un test de Bartlett sur chacun des facteurs. (on ne le fera pas ici)

On pourrait aussi regarder l'effet de chaque variable seul (qui sera atténué vu que la variance résiduelle sera plus importante). On effectuera un test de Student pour l'effet du genre car les effectifs sont grands

```
table(data$sexe)
```

```
##
##   F   M
## 310 690
```

```
t.test(X2~sexe,data)
```

```
##
##  Welch Two Sample t-test
##
## data:  X2 by sexe
## t = -2.6996, df = 664.47, p-value = 0.007118
## alternative hypothesis: true difference in means between group F and group M is not equal to 0
```

```
## 95 percent confidence interval:
##  -3.6656903 -0.5786295
## sample estimates:
## mean in group F mean in group M
##        19.43871        21.56087
```

Ce test confirme que le sexe a bien une influence sur la durée de l'emprunt. On peut maintenant faire un test de Kruskall-Wallis pour regarder si le facteur "emploi" a une influence sur la durée (en effet les conditions de l'ANOVA ne sont pas réunies)

```
kruskal.test(X2~X17,data)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  X2 by X17
## Kruskal-Wallis chi-squared = 52.95, df = 3, p-value = 1.879e-11
```

La situation professionnelle a donc bien une influence significative sur la durée de l'emprunt.

Question 3

On se propose ensuite d'étudier si les variables "montant du crédit" (variable quanti) et "durée de l'emprunt" (variable quanti) sont des variables gaussiennes. On formule alors les hypothèses suivantes :

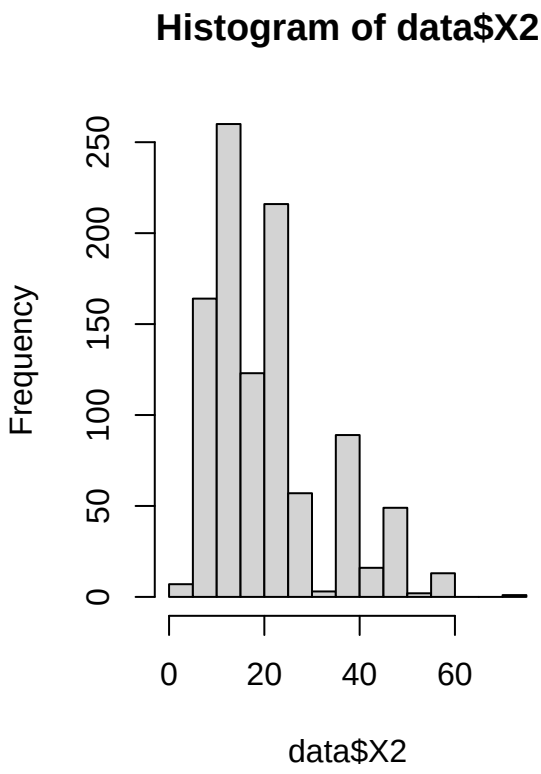
- H_0 **: le montant du crédit (resp. la durée de l'emprunt) suit une distribution gaussienne v.s.**
- H_1 **: le montant du crédit (resp. la durée de l'emprunt) ne suit pas une distribution gaussienne**

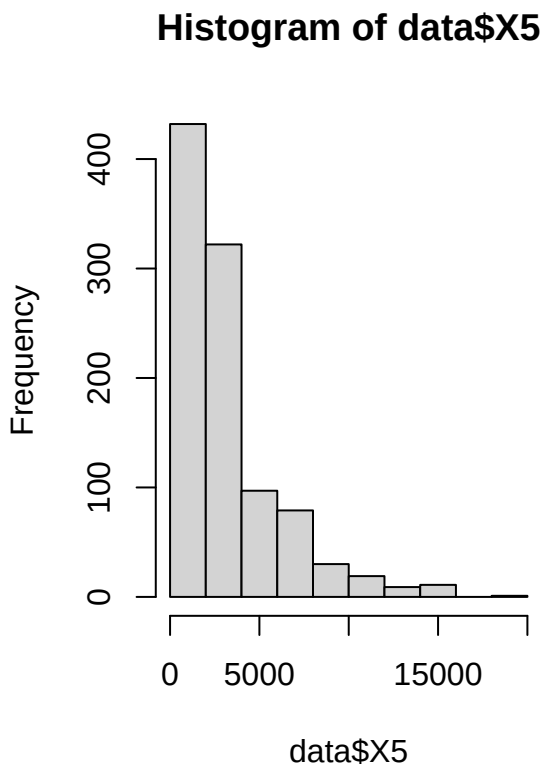
On peut commencer par faire un histogramme de ces deux variables

```
par(mfrow=c(1,2))
hist(data$X2)
shapiro.test(data$X2)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  data$X2
## W = 0.89979, p-value < 2.2e-16
```

```
hist(data$X5)
```






```
shapiro.test(data$X5)
```

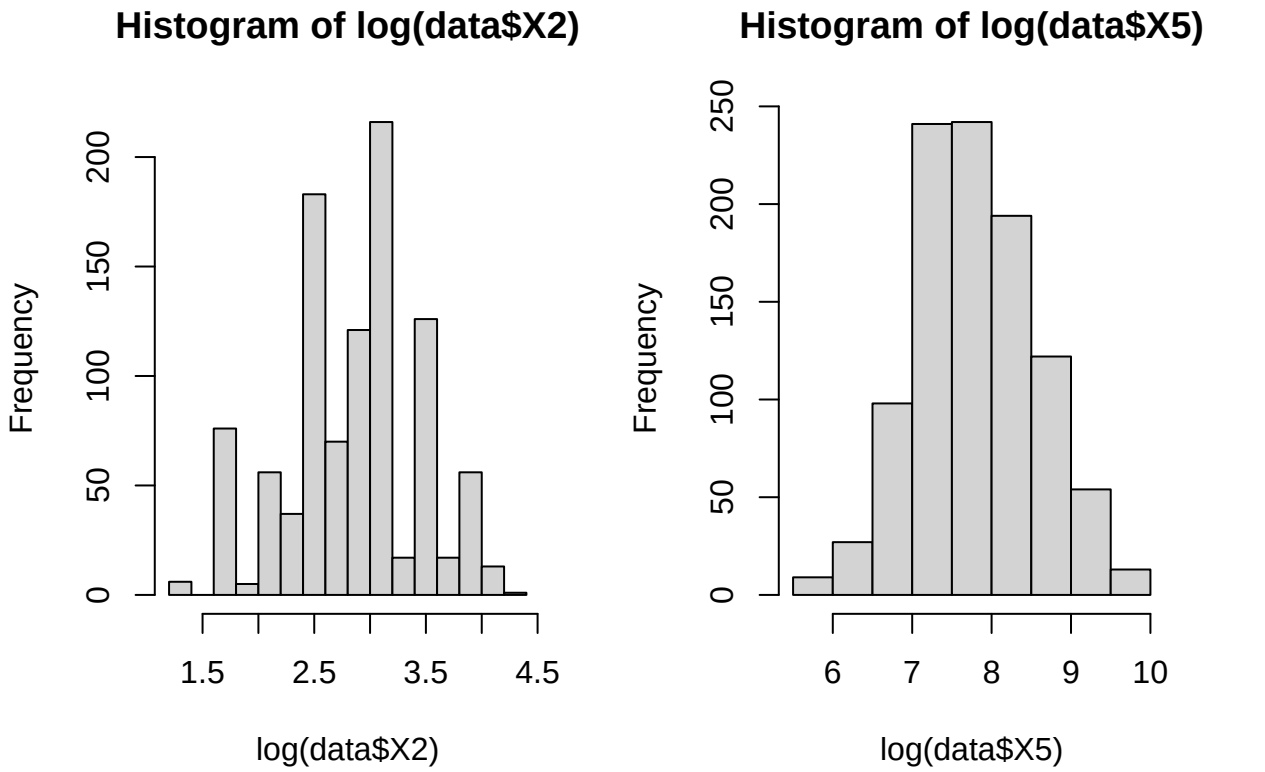
```
##
##  Shapiro-Wilk normality test
##
## data:  data$X5
## W = 0.7934, p-value < 2.2e-16
```

Les graphiques ainsi que les tests efectués montrent que clairement que les distributions de ces deux variables ne sont pas gaussiennes ! Regardons ce qu'il se passe si on considère le log de ces deux deux variables, comme nous l'avons fait plus tôt.

```
par(mfrow=c(1,2))
hist(log(data$X2))
shapiro.test(log(data$X2))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  log(data$X2)
## W = 0.9731, p-value = 1.179e-12
```

```
hist(log(data$X5))
```




```
shapiro.test(log(data$X5))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  log(data$X5)
## W = 0.99304, p-value = 0.0001242
```

Les histoigrammes sont d'avantage symétriques, mais le test conduit à nouveau au rejet de l'hypothèse nulle. Comme précédemment rappellons que le test de Shapiro est très strict en présence de grands échantillons, les résultats peuvent donc être nuancés et une étude graphique plus poussée pourrait nous permettre de trancher.

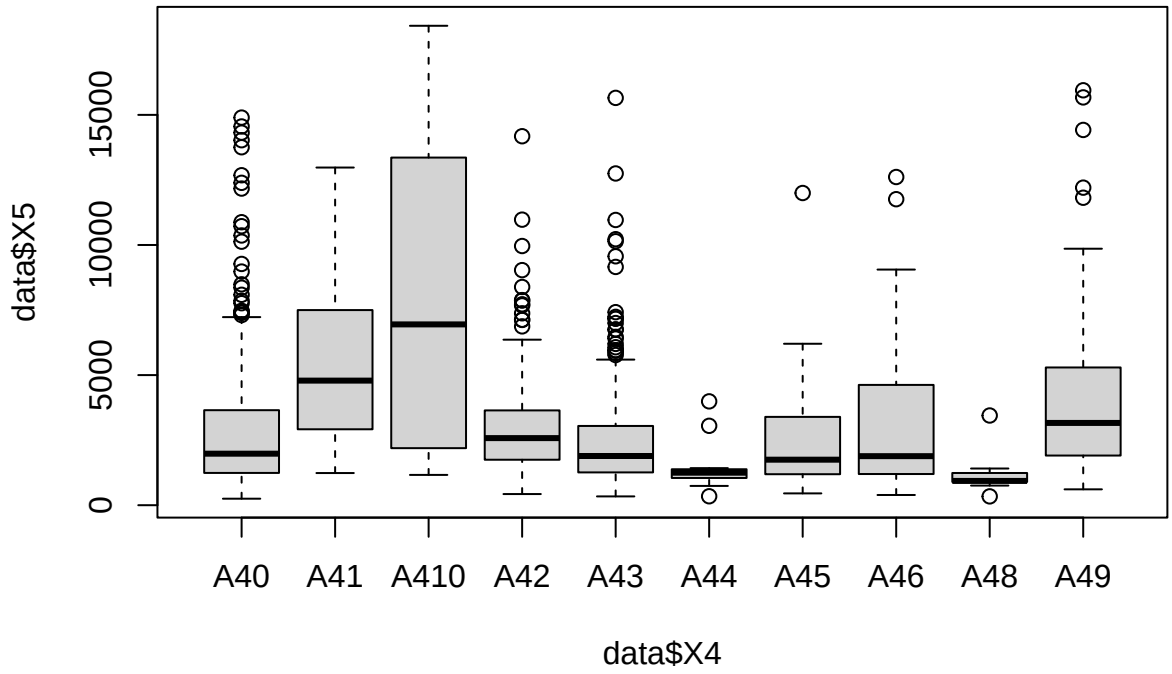
Question 4

On cherche à savoir si le montant du crédit (variable quanti) est lié au but du crédit (variable quali). On formule alors les hypothèses suivantes :

- H_0 : **le facteur "but du crédit" n'a pas d'influence sur le "montant du crédit" (les moyennes des différents groupes sont égales)**
- H_1 : **le facteur "but du crédit" a une influence sur le "montant du crédit" (au moins une moyenne est différente des autres)**

Petite représentation graphique du problème

```
boxplot(data$X5~data$X4)
```




```
modele=aov(data$X5~data$X4)
```

A prior le facteur "but du crédit" aura bien une infleunce sur le montant emprunté. Confirmons cela à l'aide d'une analyse de variance (le but prenant plus que deux modalités)

Regardons nos échantillons de plus près :

```
table (data$X4)
```

```
##
##  A40  A41 A410  A42  A43  A44  A45  A46  A48  A49
##  234  103   12  181  280   12   22   50    9   97
```

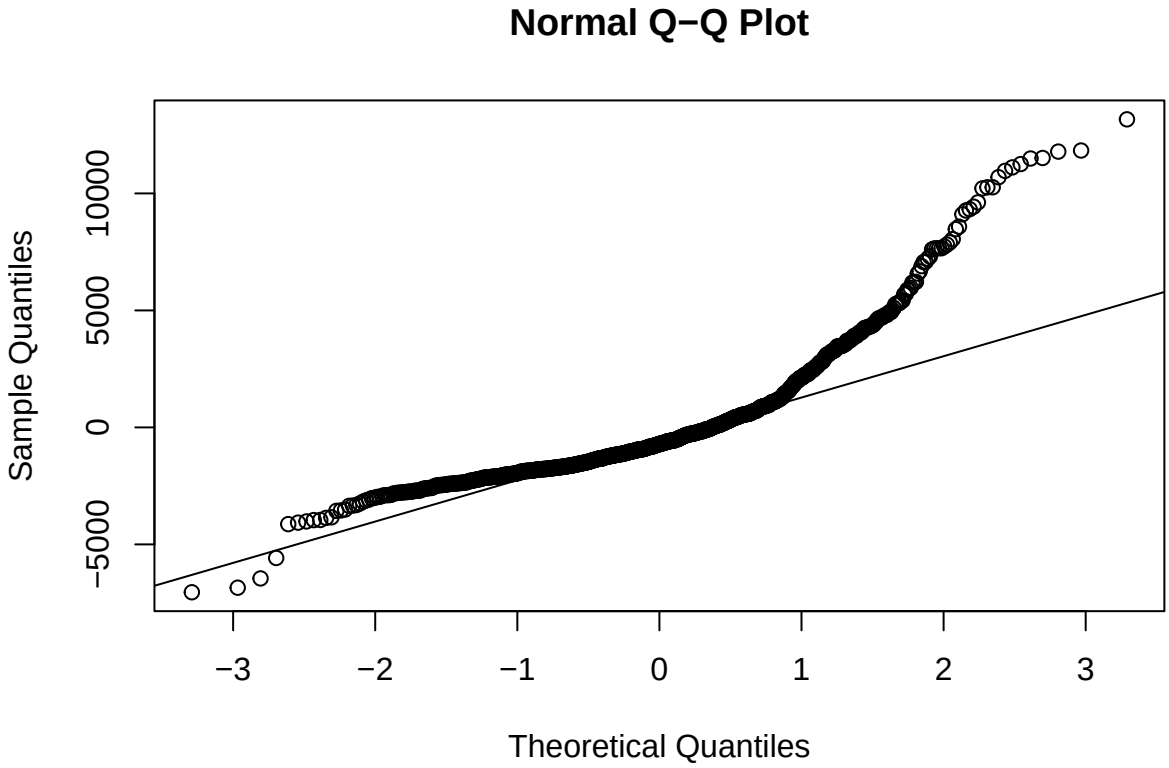
Certains échantilllons sontr très petits ce qui peut fortement nuir à l'hypothèse de normalité des échantillons. Regardons les résidus dans ce cas.

```
modele = aov(X5~X4,data)
summary(modele)
```

```
##              Df    Sum Sq   Mean Sq F value Pr(>F)
## X4            9 1.095e+09 121703799   17.55 <2e-16 ***
## Residuals   990 6.865e+09   6933880
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

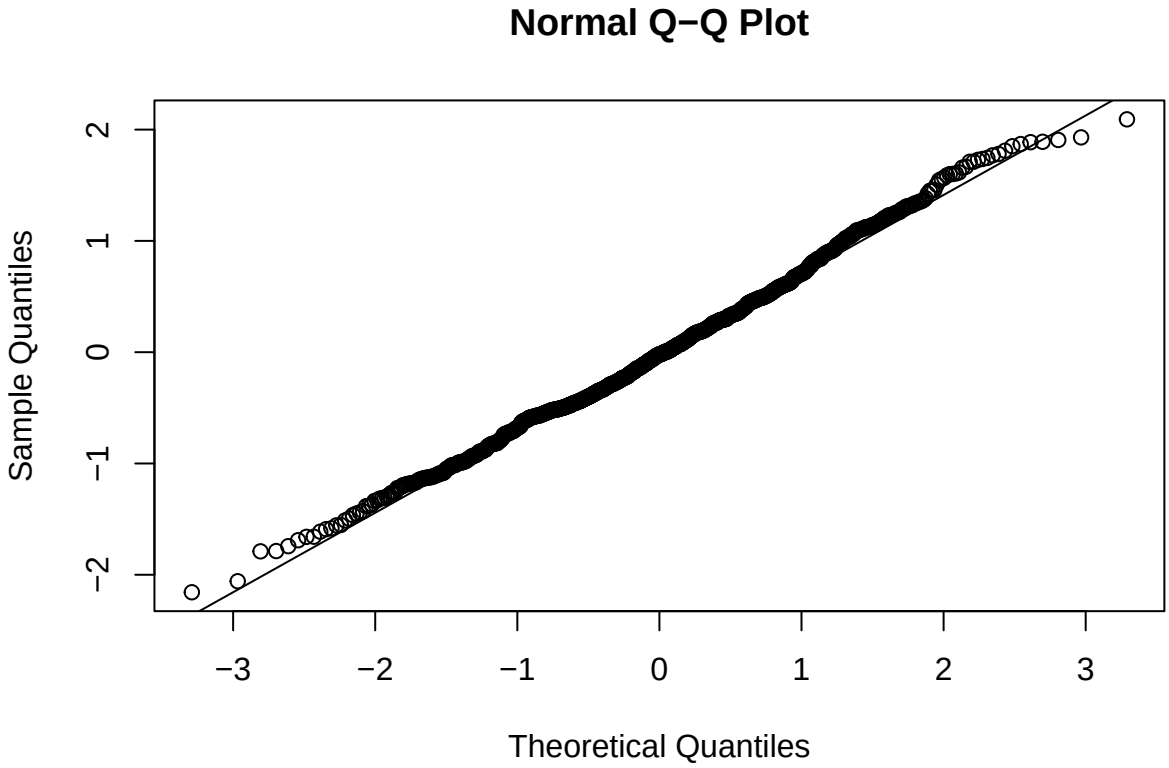
Le facteur but du crédit semble avoir une influence, mais est-ce bien le cas, regardons si le modèle est valide en regardant les résidus.

```
qqnorm(modele$residuals)
qqline(modele$residuals)
```



Inutile de faire un test de Shapiro pour voir que les résidus ne sont pas gaussiens. Regardons à nouveau avec le log des données pour voir si cela change les choses.

```
modele=aov(log(data$X5)~data$X4)
qqnorm(modele$residuals)
qqline(modele$residuals)
```




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

```
##
##  Shapiro-Wilk normality test
##
## data:  modele$residuals
## W = 0.99564, p-value = 0.006155
```

Nos résidus sont gaussiens visuellement même si un test de shapiro semble dire de la contraire (on se rappelle que ce modèle est très robuste ainsi si quelques données dévient un peu de la normalité, cette hypothèse sera automatiquement rejetée).

Malheureusement nos échantillons ne sont pas gaussiens, nous n'avons pas d'autres choix que de faire une alternative non paramétrique de l'ANOVA, i.e. un test de Kruskall-Wallis.

```
modele=kruskal.test(data$X5~data$X4)
modele
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  data$X5 by data$X4
## Kruskal-Wallis chi-squared = 148.92, df = 9, p-value < 2.2e-16
```

Ce test conduit au rejet de l'hypothèse d'indépendance entre les deux variables.

Question 5

On cherche à savoir si le montant emprunté est différent selon notre situation personnelle en terme de logement (propriétaire, locataire, ...). On formule alors les hypothèses suivantes :

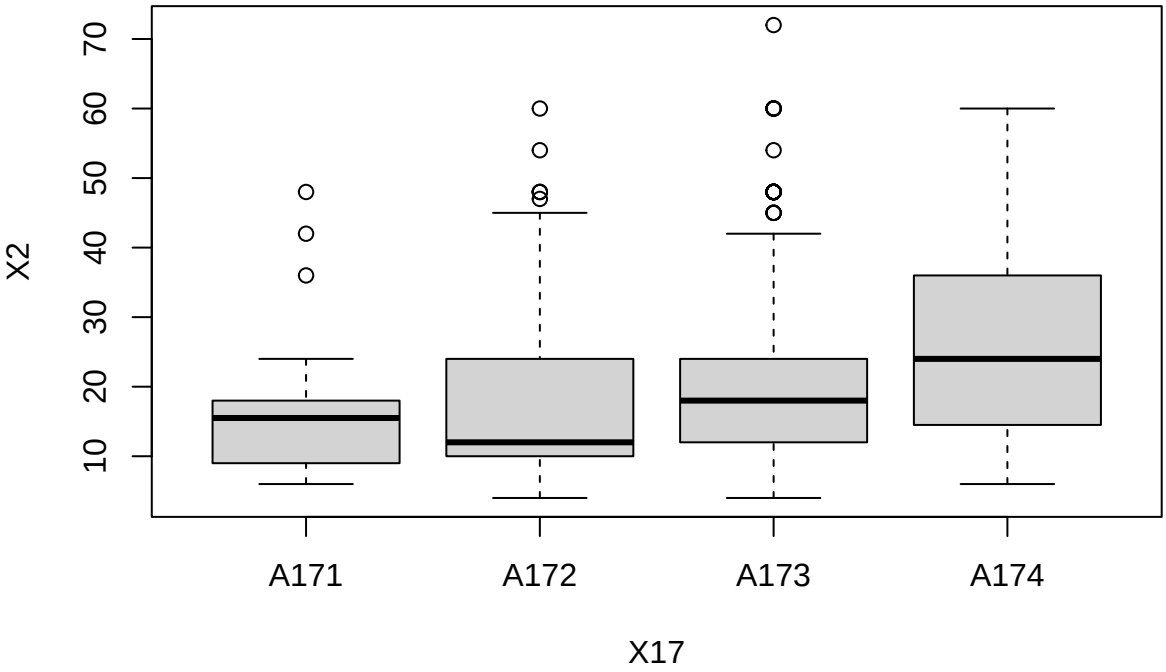
- H_0 : **le montant emprunté ne dépend pas de notre situation personnelle**

- H_1 : **le montant emprunté dépend de notre situation personnelle**

La variable qualitative prenant plus que deux modalités, nous devrons faire une Analyse de Variance ou son alternative non paramétrique si les conditions de l'ANOVA ne sont pas vérifées.

Commençons par illuster graphique les données à l'aide d'un boxplot et remarquons que le facteur étudié a un bien impact sur le montant emprunté.

```
boxplot(X2~X17,data)
```



Commençons par regarder l'homogénéité des variances à l'aide d'un test de Bartlett

```
bartlett.test(X2~X17,data)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  X2 by X17
## Bartlett's K-squared = 15.772, df = 3, p-value = 0.001263
```

On remarque que l'hypothèse d'homoscédasticité n'est pas vérifiée ici. Mais on ne soucit pas réellement de cette hypothèse là en pratique surtout quand nos effectifs sont grands et semblables.

On peut maintenant regarder si nos échantillons sont gaussiens ou non.

```
table(data$X17)
```

```
##
## A171 A172 A173 A174
##   22  200  630  148
```

Il nous faut donc vérifier si le montant de l'emprunt des individus du groupe "A171" sont normalement distribués.

```
shapiro.test(data$X2[data$X17=="A171"])
```

```
##
##  Shapiro-Wilk normality test
##
## data:  data$X2[data$X17 == "A171"]
## W = 0.83354, p-value = 0.001761
```

Ce qui n'est pas le cas, nous n'avons donc pas d'autre choix que de faire un test non paramétrique de Kruskal Wallis pour vérifier si les variables étudiantes sont indépendantes ou non.

```
modele = kruskal.test(X2~X17,data)
modele
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  X2 by X17
## Kruskal-Wallis chi-squared = 52.95, df = 3, p-value = 1.879e-11
```

On remarque que nos variables sont à nouveau liées, i.e. on peut dire que le facteur situation personnelle en terme de logement à un impact sur le montant emprunté.

Question 6

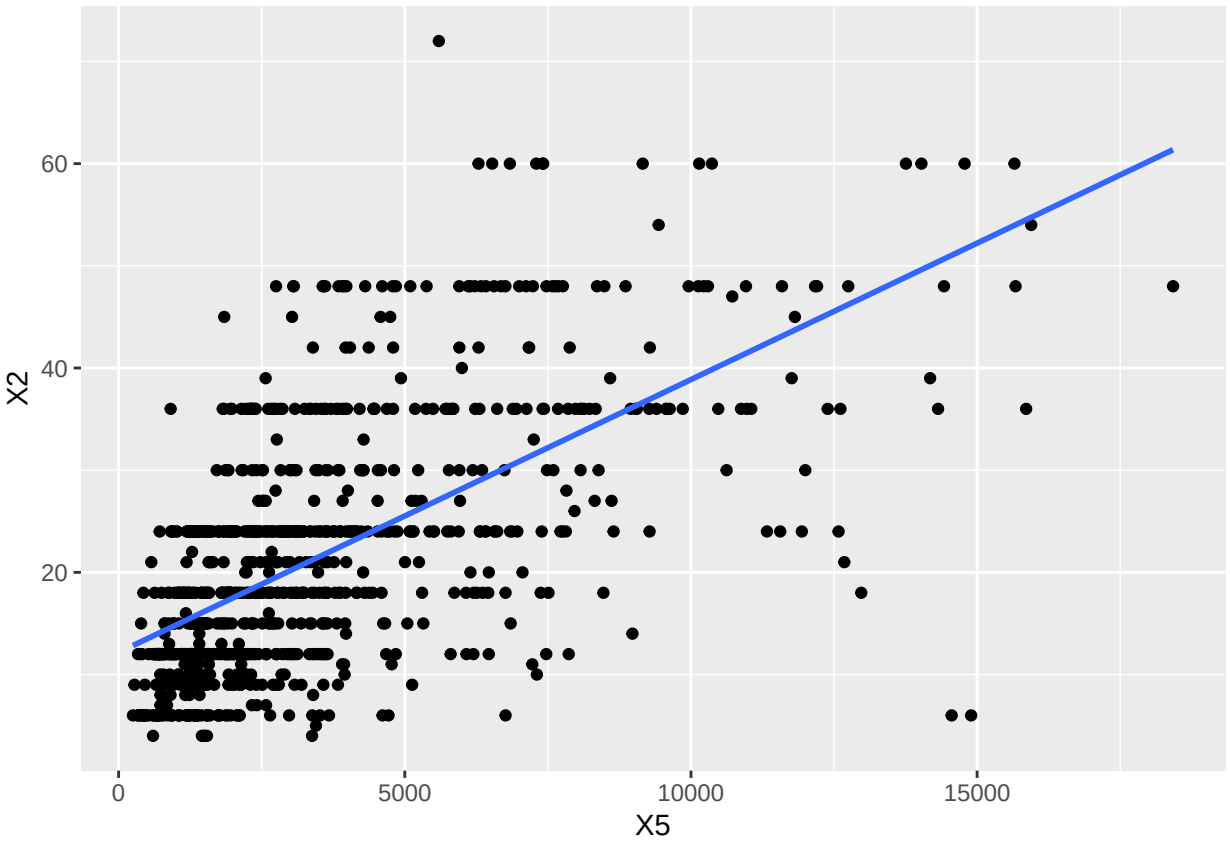
Enfin, on souhaite savoir si le montant du crédit est lié à la durée de ce dernier. On formule alors les hypothèses suivantes :

- H_0 : **le montant du crédit est lié à la durée du crédit, il y a une dépendance linéaire entre les deux facteurs**
- H_1 : **le montant du crédit n'est pas lié à la durée du crédit. Il n'y a pas de dépendance linéaire entre les deux variables.**

On va commencer par observer si cette dépendance linéaire existe ou non.

```
library(ggplot2)
ggplot(data, aes(x=X5, y=X2)) +
geom_point() +
geom_smooth(method=lm,se=FALSE)
```

```
## `geom_smooth()` using formula 'y ~ x'
```



Est-ce que cette relation est significative ou non ? On va faire un test de corrélation pour vérifier cela.

```
cor.test(data$X5,data$X2)
```

```
##
##  Pearson's product-moment correlation
##
## data:  data$X5 and data$X2
## t = 25.292, df = 998, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.5856832 0.6613533
## sample estimates:
##       cor
## 0.6249842
```

La relation linéaire entre les deux variables est significative.