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 :

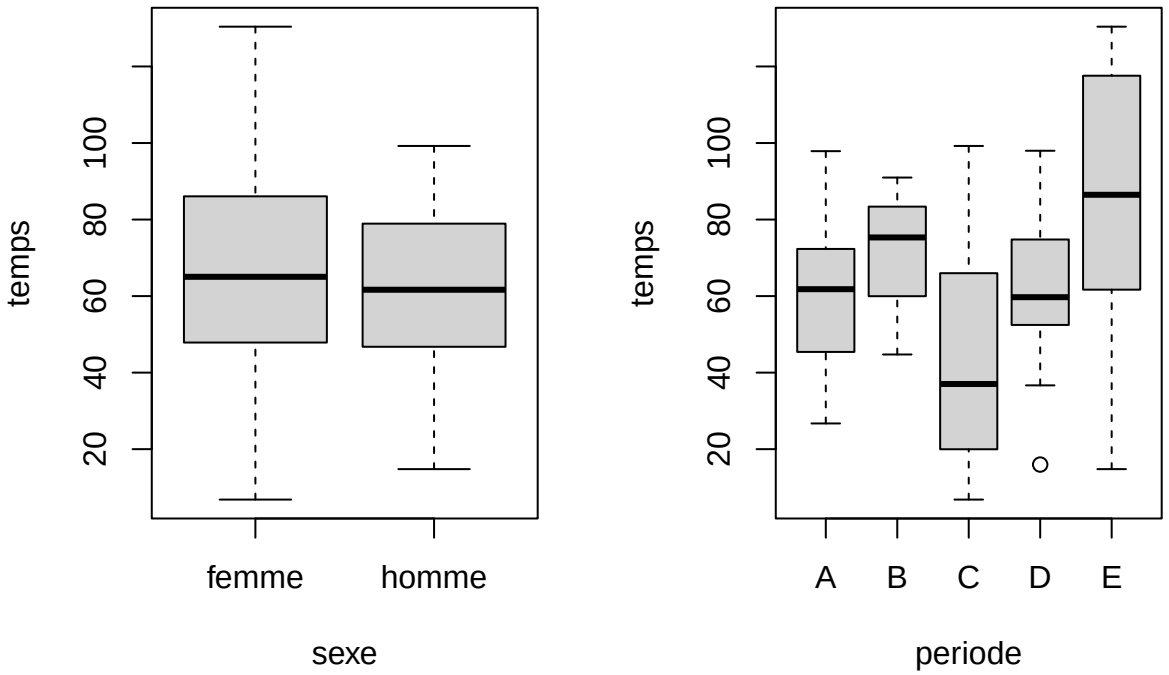
- H_0 : la variable "temps" est indépendante des deux autres facteurs v.s.
- H_1 : la variable "temps" est dépendante d’au moins un des autres facteurs.

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))
```

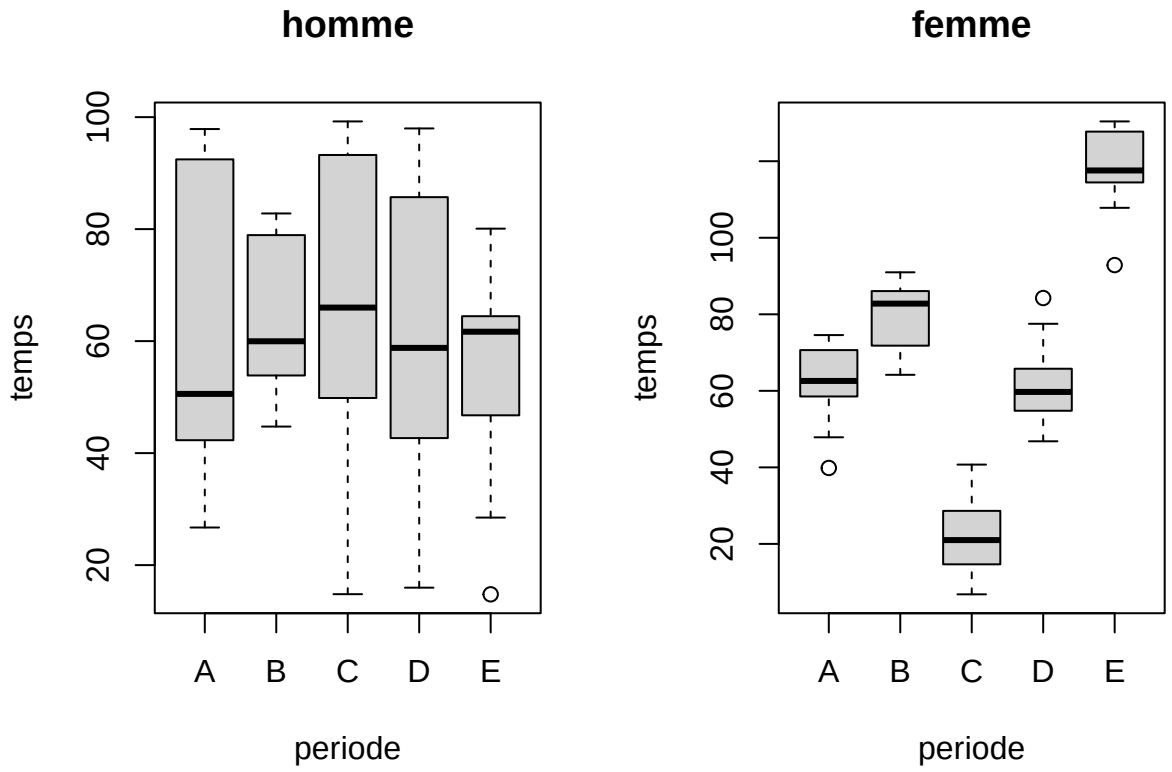
```
## Analysis of Variance Table
##
## Response: temps
##           Df Sum Sq Mean Sq F value    Pr(>F)
## sexe       1   1373  1372.7  2.2303    0.1387
## periode    4  19488  4872.1  7.9163 1.523e-05 ***
```

```
## 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
```

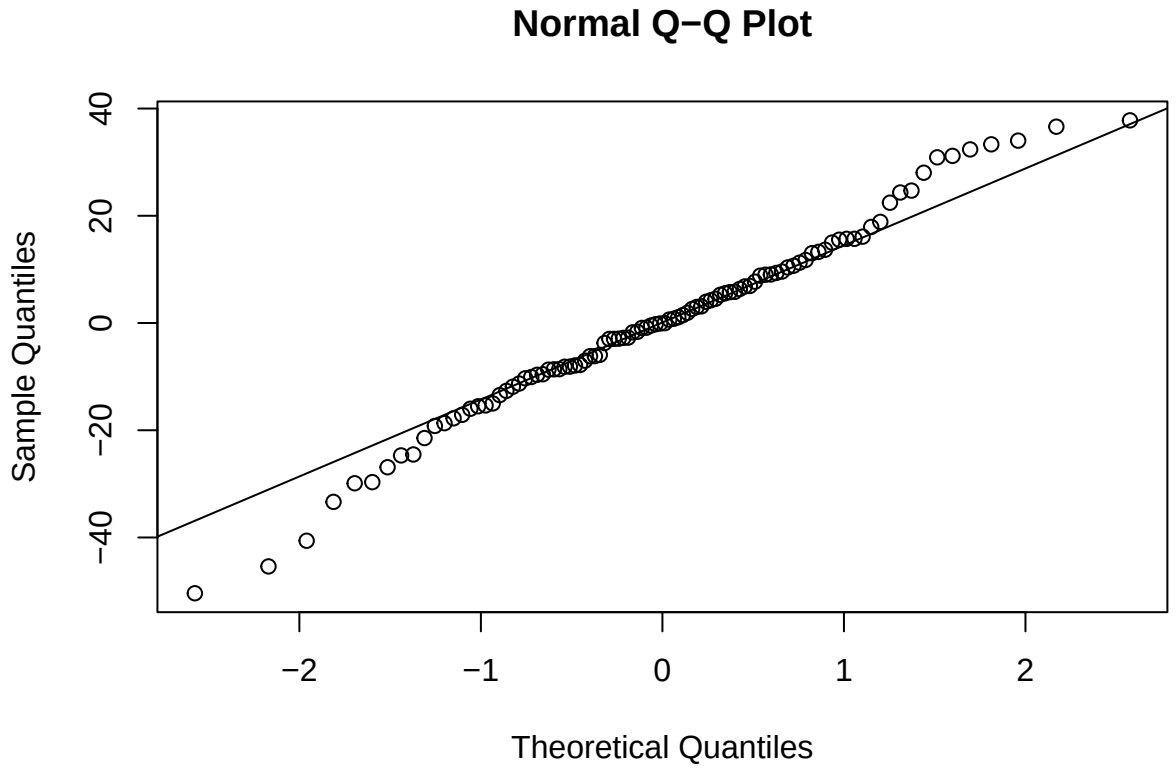
On constate que cette interaction est en effet significative.

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
```

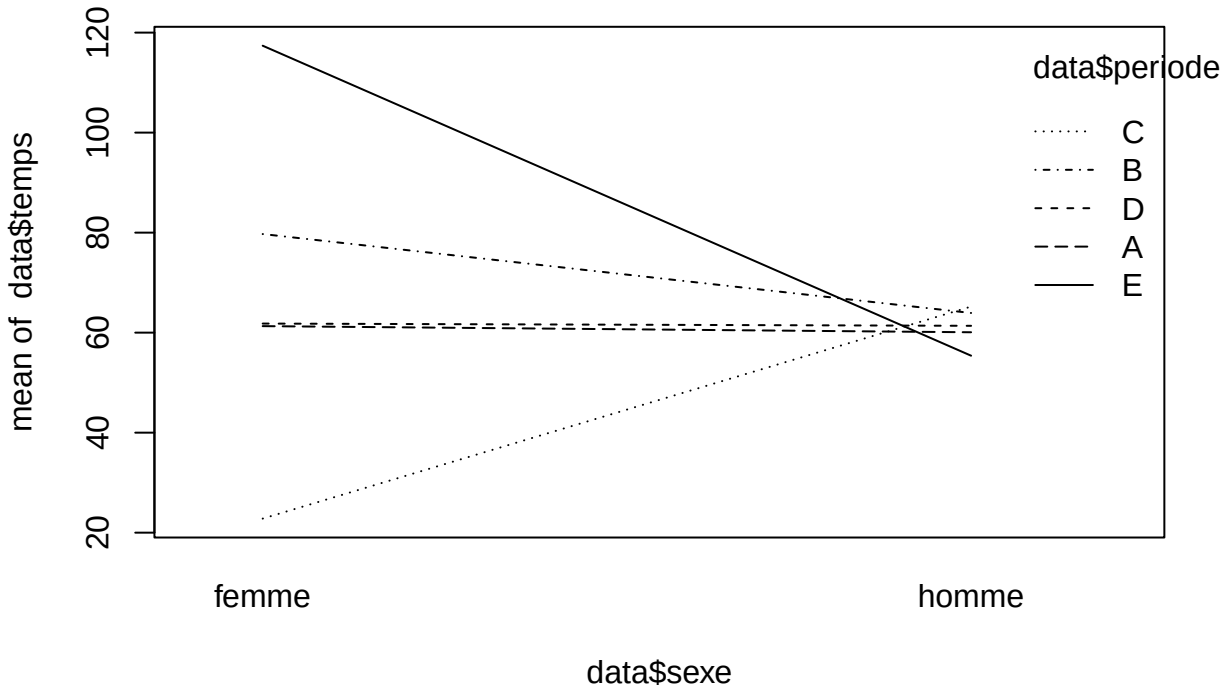
```
bartlett.test(temps~sexe,data=data)
```

```
##
##  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.

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

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 :
- H_1 :

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 :
- H_1 :

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 :

- H_1 :

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 :
- H_1 :

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 :
- H_1 :