Correction Examen 2021

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

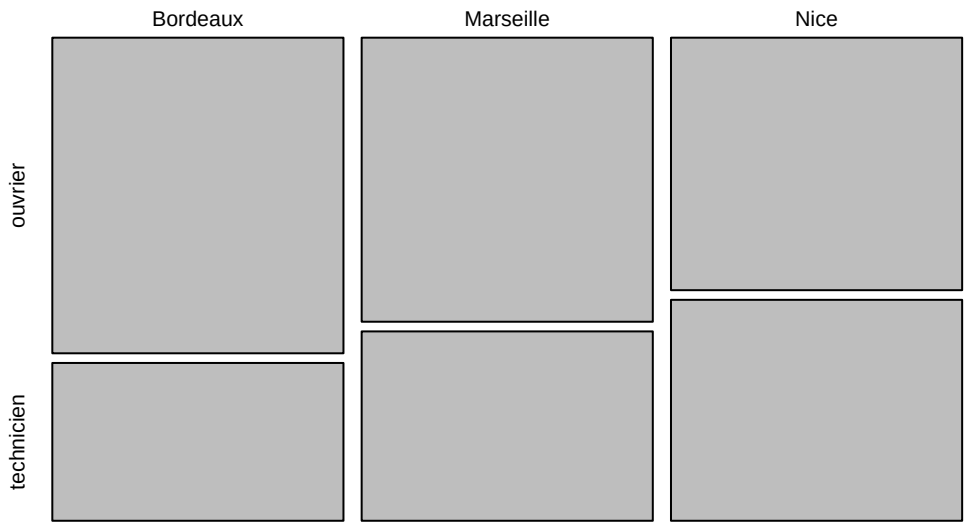
Commençons par charger les données

```
library(readxl)
data = read_excel("Data_Company.xlsx")

# On convertit les variables qualitatives en facteur.
data$Ville=as.factor(data$Ville)
data$Sexe=as.factor(data$Sexe)
data$Profession=as.factor(data$Profession)

# On transforme notre jeu de données afin de faire apparaître l'année
# comme étant uen nouvelle variable
newdata=rbind(data[,1:3],data[,1:3])
Revenus=c(data$Revenus_2019,data$Revenus_2020)
Annee=as.factor(rep(c(2019,2020),each=nrow(data)))
newdata=cbind(newdata,Revenus,Annee)
```

Question 1 : La répartition des professions est-elle la même dans chaque ville ?

On s'intéresse au lien entre deux variables qualitatives. Commençons par représenter les données.

```
plot(table(data$Ville,data$Profession))
```



On va chercher à étudier si la proportion d'ouvrier et de technicien est la même selon les différentes villes.

- H_0 : La proportion d'ouvrier-technicien est la même quelque soit la ville
- H_1 : Cette même proportion varie selon la ville.

On va donc effectuer un test du Khi-deux on vérifie les conditions sur les effectifs de la table de contingence puis on effectue notre test.

```
table(data$Ville,data$Profession)
```

```
##
##             ouvrier technicien
##   Bordeaux       10          5
##   Marseille       9          6
##   Nice            8          7
```

Tous les effectifs sont supérieurs ou égaux à 5.

```
chisq.test(table(data$Ville,data$Profession))
```

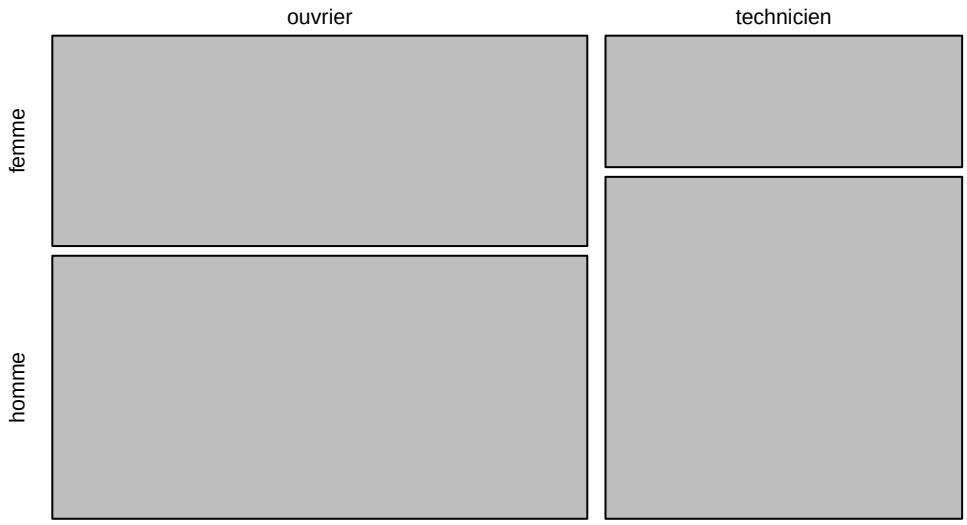
```
##
##  Pearson's Chi-squared test
##
## data:  table(data$Ville, data$Profession)
## X-squared = 0.55556, df = 2, p-value = 0.7575
```

La p-value n'est pas significative, on ne peut donc pas affirmer que la région ait un impact sur la proportion d'ouvrier/technicien dans cette entreprise.

Question 2 : Y a-t'il un lien entre genre et profession

On étudie à nouveau deux variables qualitatives. On va donc procéder comme à la question précédente. A nouveau, on commence par regarder nos données.

```
plot(table(data$Profession,data$Sexe))
```



- H_0 : La proportion homme/femme est la même quelque soit la profession
- H_1 : La profession a un impact sur la proportion homme/femme

On regarde nos effectifs afin de vérifier que l'on peut faire notre test du Khi-deux.

```
table(data$Sexe,data$Profession)
```

```
##
##         ouvrier technicien
##   femme      12          5
##   homme      15         13
```

Nos effectifs sont supérieurs à 5, on peut donc effectuer notre test du Khi-deux.

```
chisq.test(data$Sexe,data$Profession)
```

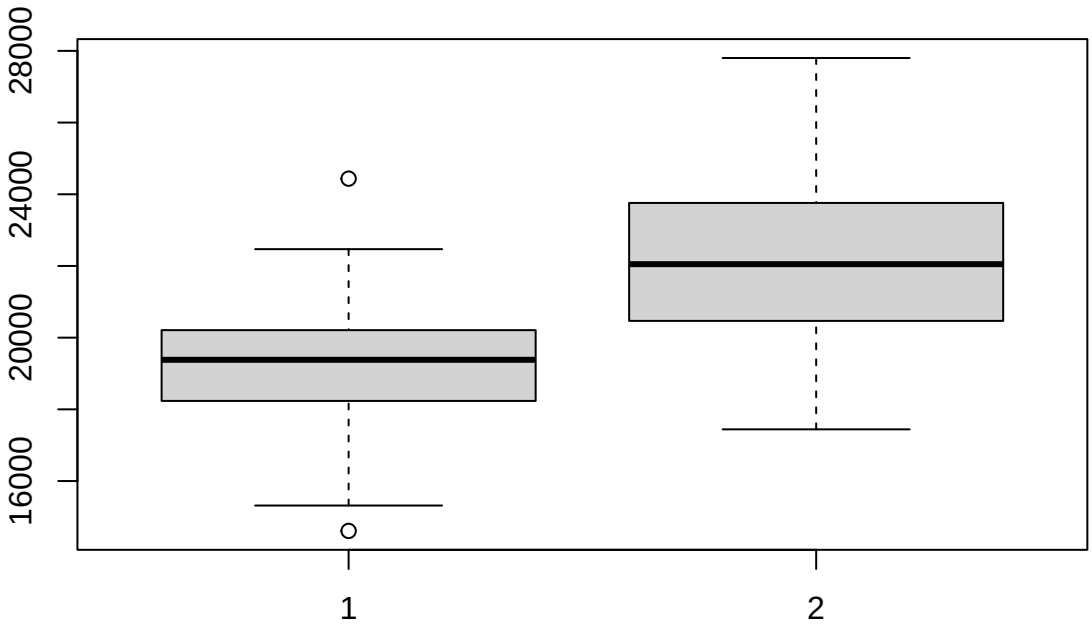
```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  data$Sexe and data$Profession
## X-squared = 0.6657, df = 1, p-value = 0.4146
```

A nouveau, on ne peut pas dire que le sexe ait un impact sur la profession exercée.

Question 3 : Les revenus ont-ils augmenté entre 2019 et 2020 ?

On représente nos données à l'aide d'un boxplot.

```
boxplot(data$Revenus_2019,data$Revenus_2020)
```



On va maintenant étudier si la différence observée sur le graphe est significative ou non. On doit donc comparer des moyennes.

- H_0 : Les revenus restent stables
- H_1 : Les revenus ont augmenté entre 2019 et 2020

Regardons d'abord nos données

```
# Taille des échantillons

dim(data)[1]
```

```
## [1] 45
```

Nous avons deux échantillons de taille supérieure à 30, on va donc effectuer un test de Student en prenant soin de préciser que les échantillons sont appariés.

```
t.test(data$Revenus_2019,data$Revenus_2020,alternative='less',paired=TRUE)
```

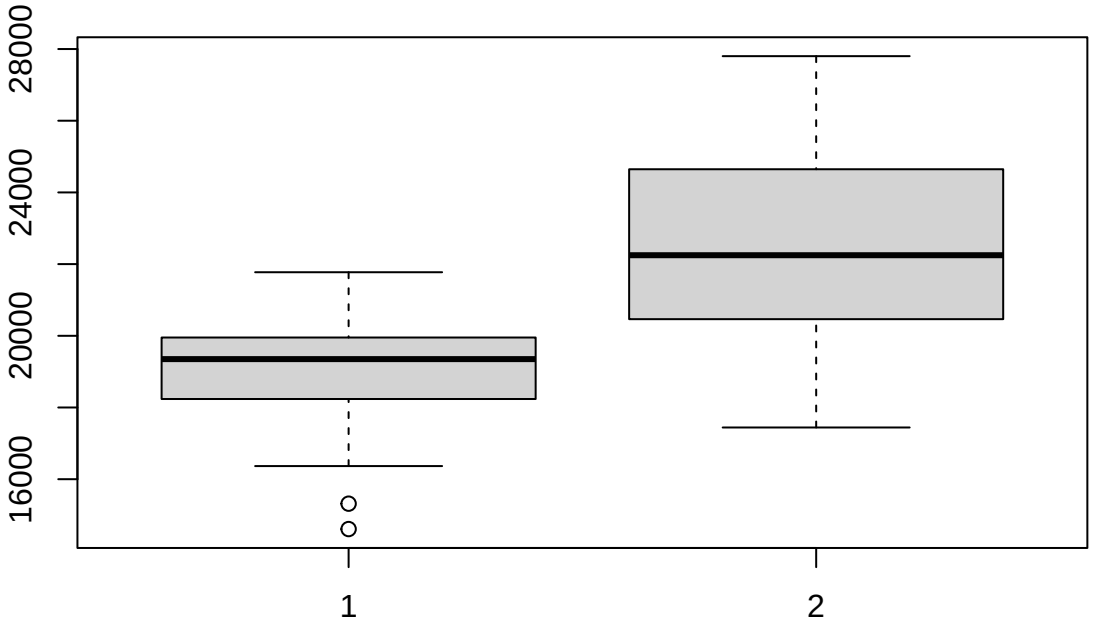
```
##
##  Paired t-test
##
## data:  data$Revenus_2019 and data$Revenus_2020
## t = -6.8838, df = 44, p-value = 8.463e-09
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##       -Inf -2369.102
## sample estimates:
## mean of the differences
##               -3134.089
```

La p-value est inférieure à 0.05, on peut donc affirmer que le salaire a bien augmenté entre les années 2019 et 2020.

Question 4 : Les revenus des techniciens ont-ils augmenté entre 2019 et 2020 ?

Regardons graphiquement l'évolution des revenus

```r
# Préparation des données

tech19=data$Revenus_2019[data$Profession=="technicien"]
tech20=data$Revenus_2020[data$Profession=="technicien"]

# Représentation graphique

boxplot(tech19,tech20)
```



On va procéder au test suivant :

- H_0 : Les revenus des techniciens restent stables
- H_1 : Les revenus des techniciens ont augmenté entre 2019 et 2020

On va regarder notre échantillon différence et commencer par voir si ce dernier est gaussien.

```r
table(data$Profession)
```

```
##
##    ouvrier technicien
##         27         18
```

En effet, notre échantillon n'est que de taille 18, il va donc falloir faire un test de Shapiro. On va directement s'intéresser à l'échantillon différence.

```r
shapiro.test(tech19-tech20)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  tech19 - tech20
## W = 0.97052, p-value = 0.8074
```

Notre échantillon est gaussien, on va donc pouvoir procéder à un test de Student en prenant soin de préciser que les données sont appariées

```r
t.test(tech19,tech20,alternative='less',paired=TRUE)
```

```
##
##  Paired t-test
##
## data:  tech19 and tech20
## t = -4.2684, df = 17, p-value = 0.0002596
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##       -Inf -2161.917
## sample estimates:
## mean of the differences
##               -3649.167
```

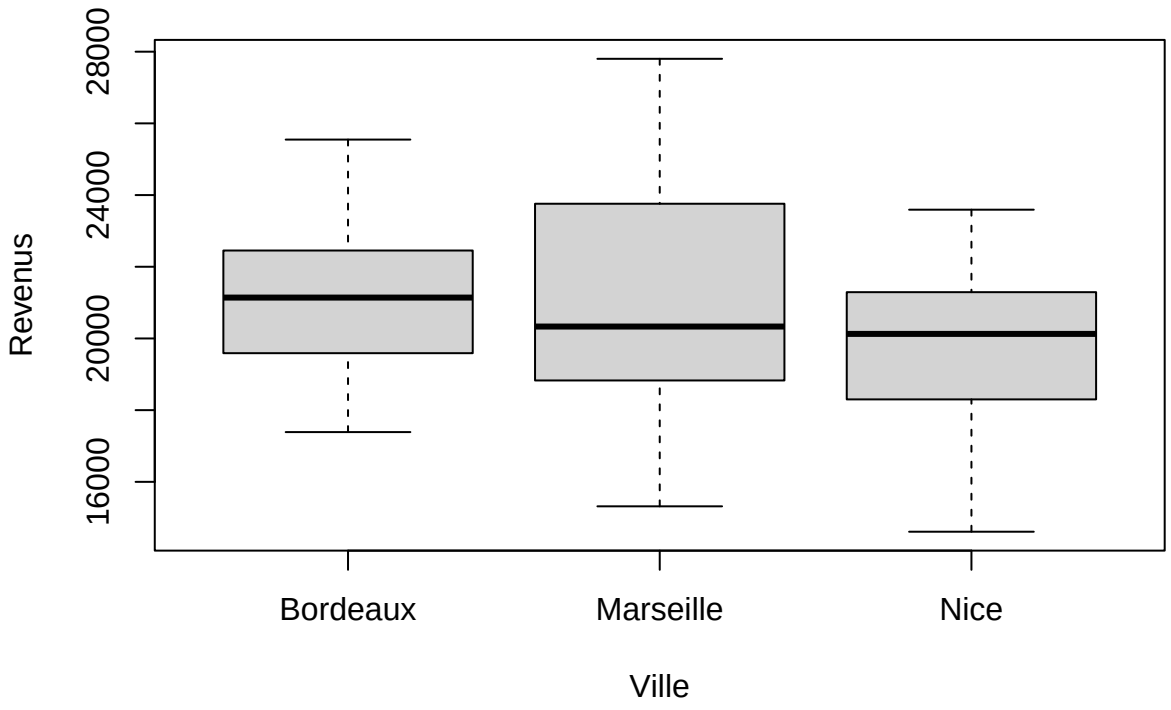
```
# On aurait aussi pu écrire
# t.test(tech19-tech20,alternative='less')
```

Le salaire des techniciens a significativement augmenté entre les années 2019 et 2020.

Question 5 : La ville influe-t-elle sur le salaire, toutes années confondue ?

Regardons la distribution des salaires en fonction de la ville étudiée

```
boxplot(Revenus~Ville,data=newdata)
```



La variable qualitative prenant plus que 2 modalités, on doit donc procéder à une Analyse de Variance.

- H_0 : Le salaire ne dépend pas de la ville étudiée
- H_1 : Le salaire dépend de la ville étudiée

On va effectuer notre analyse de variance et on va regarder la normalité des résidus.

```
res <- aov(Revenus~Ville,data=newdata)
summary(res)
```

```
##             Df    Sum Sq  Mean Sq F value Pr(>F)
## Ville        2  48048728 24024364   3.335 0.0402 *
```

```
## Residuals   87 626639845  7202757
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

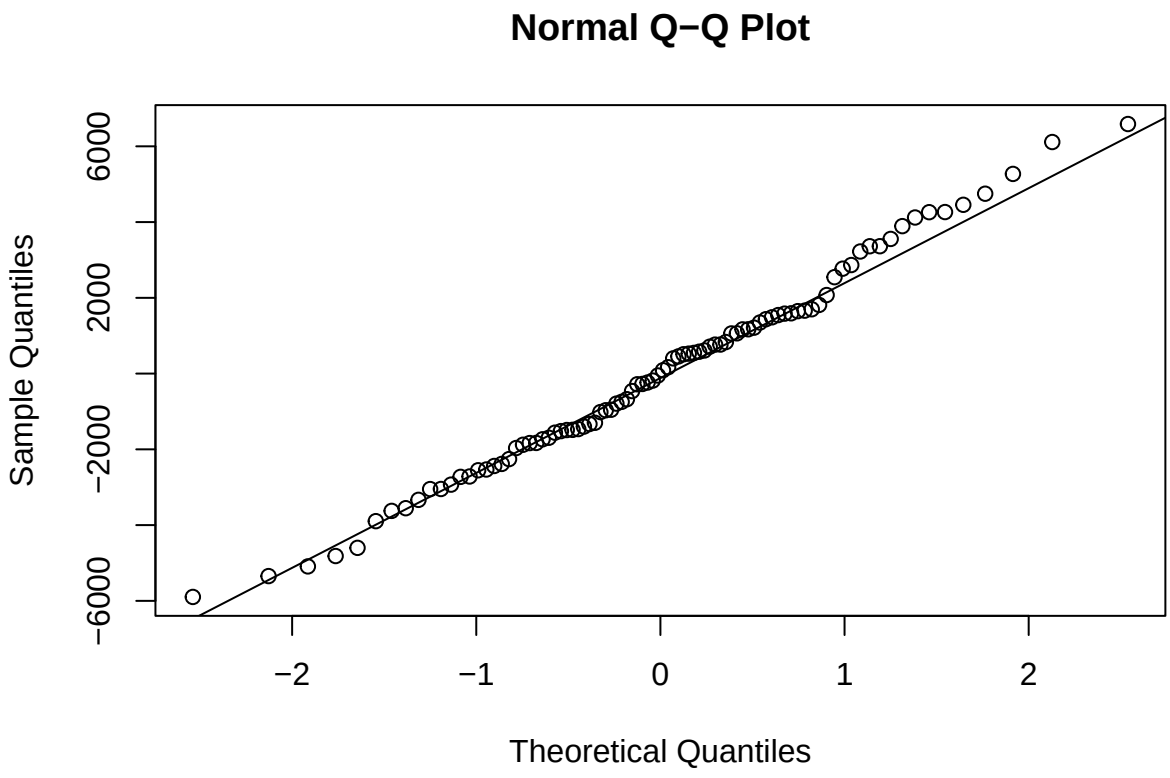
Nous avons 30 exemples dans chaque groupe, il n'est donc pas nécessaire de vérifier le caractère gaussien des différents échantillons. On va vérifier l'homogénéité des variances avec un test de Bartlett.

```
bartlett.test(Revenus~Ville,data=newdata)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  Revenus by Ville
## Bartlett's K-squared = 7.8113, df = 2, p-value = 0.02013
```

Les variances entre les différents groupes ne sont donc pas homogènes. Mais l'ANOVA reste robuste même si cette hypothèse n'est pas vérifiée, du moment que les effectifs entre les différents groupes sont semblables. Etudions nos résidus

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




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

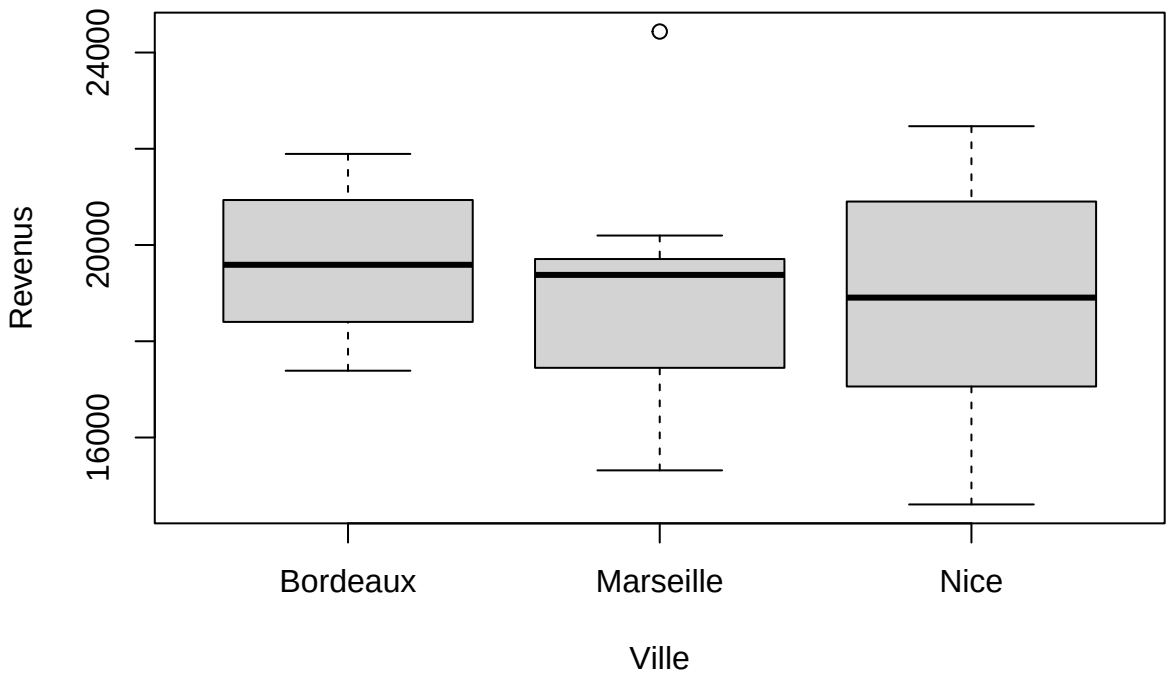
```
##
##  Shapiro-Wilk normality test
##
## data:  res$residuals
## W = 0.99134, p-value = 0.8248
```

Les résidus sont normaux, on peut donc conclure que la facteur ville a bien un impact sur la moyenne des salaires observée.

Question 6 : Les salaires en 2019 étaient-ils similaires quelque soit la ville ?

La question se traite de la même façon que la question précédente, on fera cependant attention que, cette fois-ci, nos échantillons sont deux fois plus petits. Il faudrait docn vérifier que ces derniers soient gaussiens pour normalement pouvoir faire l'ANOVA.

```
boxplot(Revenus~Ville,data=newdata[newdata$Annee=="2019",])
```



- H_0 : Les revenus de 2019 sont indépendants de la ville considérée
- H_1 : Les revenus de 2019 varient selon la ville étudiée

```
res=aov(Revenus~Ville,data=newdata[newdata$Annee=="2019",])
summary(res)
```

```
##             Df    Sum Sq Mean Sq F value Pr(>F)
## Ville        2   6812116 3406058     0.8  0.456
## Residuals   42 178921910 4260045
```

On commence par regarder si l'hypothèse d'homoscédasticité est vérifiée.

```
bartlett.test(Revenus~Ville,data=newdata[newdata$Annee=="2019",])
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  Revenus by Ville
## Bartlett's K-squared = 2.7366, df = 2, p-value = 0.2545
```

C'est bien le cas comme l'indique la p-value de notre test de Bartlett. Regardons le caractère gaussien de nos échantillons puis de nos résidus.

```
shapiro.test(newdata[(newdata$Annee=="2019")&(newdata$Ville=="Marseille"),"Revenus"])
```

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

```
## data:  newdata[(newdata$Annee == "2019") & (newdata$Ville == "Marseille"), "Revenus"]
## W = 0.889, p-value = 0.06476
```

```
shapiro.test(newdata[(newdata$Annee=="2019")&(newdata$Ville=="Nice"),"Revenus"])
```

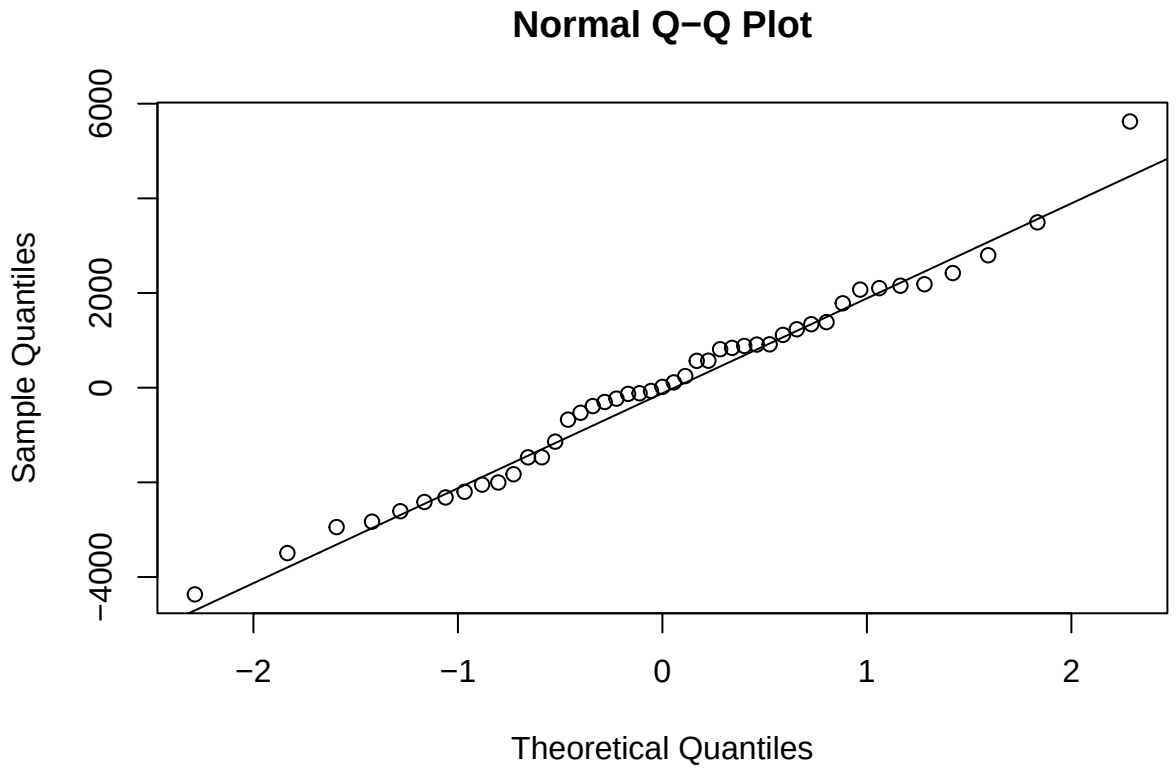
```
##
##  Shapiro-Wilk normality test
##
## data:  newdata[(newdata$Annee == "2019") & (newdata$Ville == "Nice"), "Revenus"]
## W = 0.96365, p-value = 0.7556
```

```
shapiro.test(newdata[(newdata$Annee=="2019")&(newdata$Ville=="Bordeaux"),"Revenus"])
```

```
##
##  Shapiro-Wilk normality test
##
## data:  newdata[(newdata$Annee == "2019") & (newdata$Ville == "Bordeaux"), "Revenus"]
## W = 0.93607, p-value = 0.3355
```

Les échantillons étudiés sont gaussiens, il ne nous reste plus qu'à vérifier le caractère gaussien des résidus.

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




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

```
##
##  Shapiro-Wilk normality test
##
## data:  res$residuals
## W = 0.98518, p-value = 0.8264
```

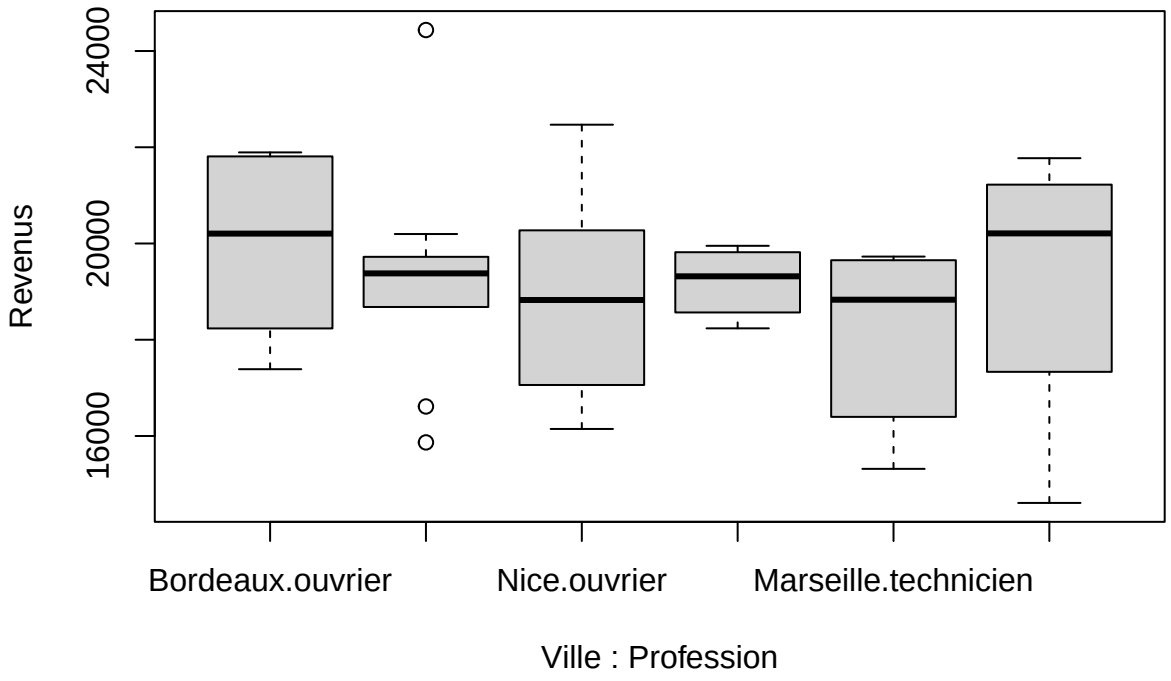
On peut donc prendre en compte les résultats de notre ANOVA qui indiquent que la ville n'a pas d'impact

sur les revenus observés en 2019.

Question 7 : En 2019, le salaire dépendait-il de la profession et de la ville ?

On nous demande à nouveau d'étudier les revenus des salariés mais en fonction de deux facteurs cette fois-ci : profession et ville. Représentons d'abord nos données

```
boxplot(Revenus~Ville*Profession,data=newdata[newdata$Annee=="2019",])
```



- H_0 : Les revenus de 2019 ne dépendent pas des deux facteurs étudiés
- H_1 : Les revenus en 2019 dépendent d'au moins un de ces deux facteurs

On va donc procéder à une ANOVA à deux facteurs, il ne faudra pas oublier de prendre en compte l'intéraction entre les deux facteurs Ville:Profession.

```
res=aov(Revenus~Ville*Profession,data=newdata[newdata$Annee=="2019",])
summary(res)
```

```
##                  Df    Sum Sq Mean Sq F value Pr(>F)
## Ville             2   6812116 3406058   0.773  0.469
## Profession        1   3241571 3241571   0.735  0.396
## Ville:Profession  2   3751893 1875947   0.426  0.656
## Residuals        39 171928446 4408422
```

On procède aux mêmes vérifications que lors de la question précédente, en commençant par l'homogénéité des variances.

```
bartlett.test(Revenus~Ville,data=newdata[newdata$Annee=="2019",])
```

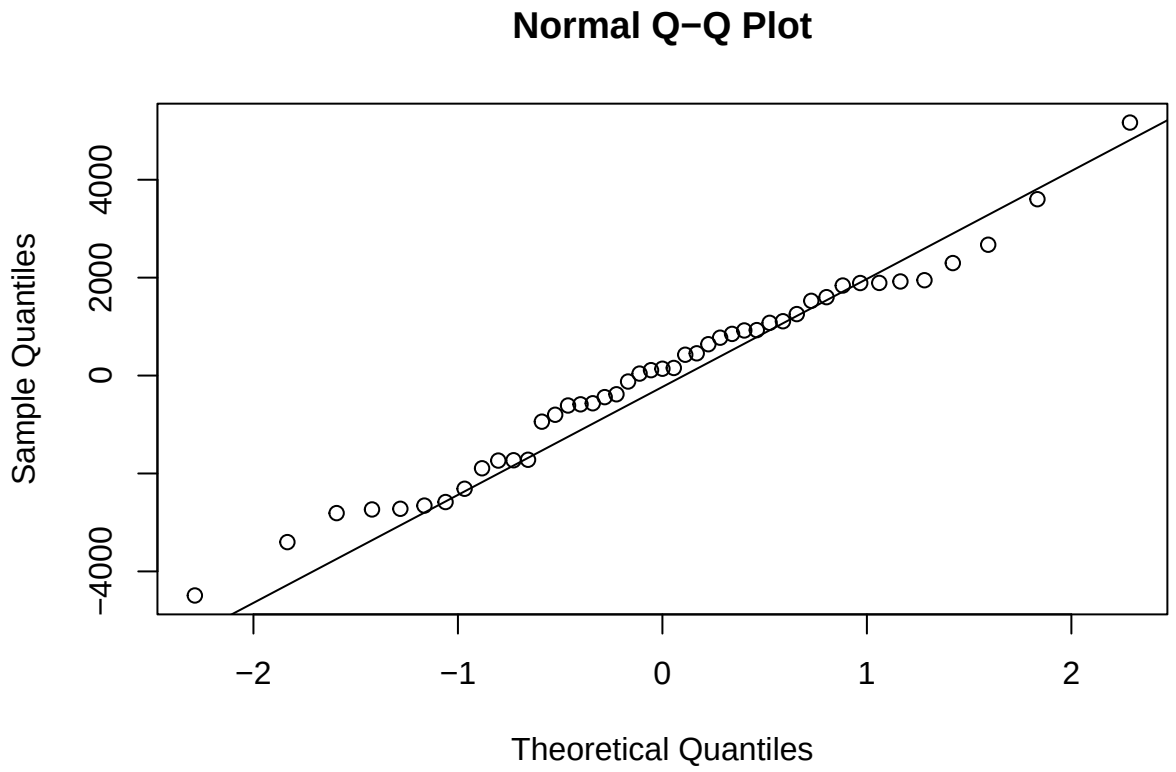
```
##
##  Bartlett test of homogeneity of variances
##
## data:  Revenus by Ville
## Bartlett's K-squared = 2.7366, df = 2, p-value = 0.2545
```

```
bartlett.test(Revenus~Profession,data=newdata[newdata$Annee=="2019",])
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  Revenus by Profession
## Bartlett's K-squared = 0.014858, df = 1, p-value = 0.903
```

Les variances entre les deux groupes étudiés sont bien homogènes. Il faudra aussi tester la normalité des donnés, mais c'est ici ridicule vu la taille des différents échantillons, notre test de Shapiro n'est pas assez puissant avec de si petits échantillons. On va donc se contenter de vérifier la normalité des résidus afin de savoir si nous validons ou non les résultats de notre ANOVA.

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




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

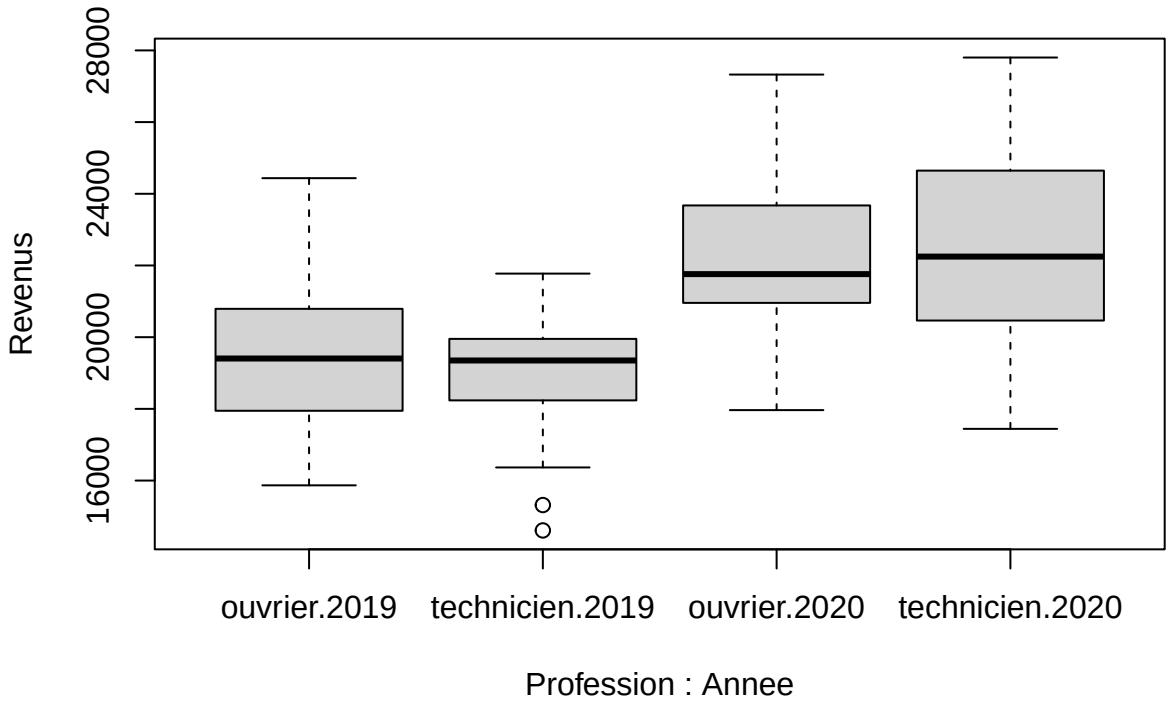
```
##
##  Shapiro-Wilk normality test
##
## data:  res$residuals
## W = 0.98359, p-value = 0.7653
```

Nos résidus sont bien normalement distribués. On peut donc conclure que les facteurs villes et profession n'ont pas d'impact sur les revenus observés en 2019.

Question 8 : La profession et l'année influent-ils sur le salaire ?

Ici, nous procédons de la même façon qu'à la question précédente, on modifie simplement les facteurs étudiés. Commençons par notre représentation graphique.

```
boxplot(Revenus~Profession*Annee,data=newdata)
```



- H_0 : Les revenus sont indépendants de l'année ou encore de la profession
- H_1 : Les revenus dépendent d'au moins l'un des deux facteurs

La question se traitant comme la précédente, on ne détaille pas les étapes.

```
res=aov(Revenus~Profession*Annee,data=newdata)
summary(res)
```

```
##                  Df    Sum Sq   Mean Sq F value   Pr(>F)
## Profession        1    722704    722704   0.138    0.711
## Annee             1 221006546 221006546  42.333 4.89e-09 ***
## Profession:Annee  1   3979577   3979577   0.762    0.385
## Residuals        86 448979746   5220695
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
bartlett.test(Revenus~Profession,data=newdata)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  Revenus by Profession
## Bartlett's K-squared = 0.96919, df = 1, p-value = 0.3249
```

```
bartlett.test(Revenus~Annee,data=newdata)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  Revenus by Annee
## Bartlett's K-squared = 1.4527, df = 1, p-value = 0.2281
```

On va rajouter la vérification du carctère gaussien des différents échantillons étudiés.

```
shapiro.test(newdata[(newdata$Annee=="2019")&(newdata$Profession=="ouvrier"),"Revenus"])
```

```
##
##  Shapiro-Wilk normality test
##
## data:  newdata[(newdata$Annee == "2019") & (newdata$Profession == "ouvrier"), "Revenus"]
## W = 0.97958, p-value = 0.8528
```

```
shapiro.test(newdata[(newdata$Annee=="2019")&(newdata$Profession=="technicien"),"Revenus"])
```

```
##
##  Shapiro-Wilk normality test
##
## data:  newdata[(newdata$Annee == "2019") & (newdata$Profession == "technicien"), "Revenus"]
## W = 0.93891, p-value = 0.2775
```

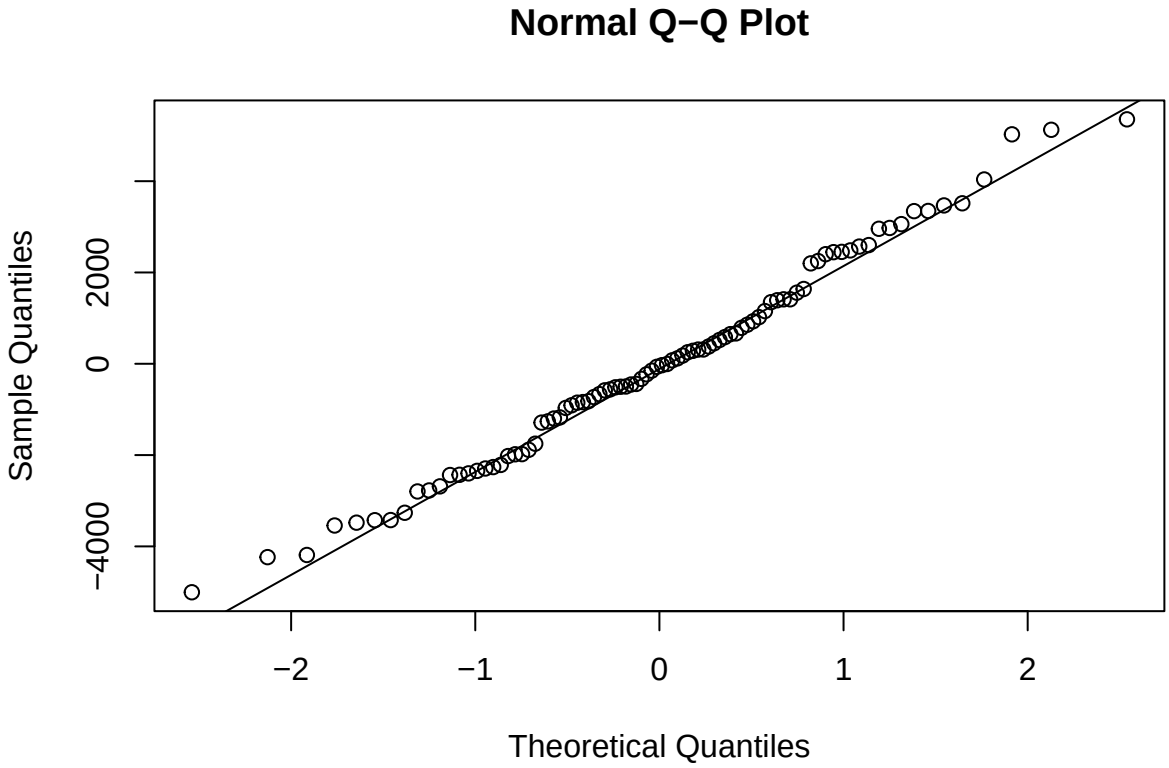
```
shapiro.test(newdata[(newdata$Annee=="2020")&(newdata$Profession=="ouvrier"),"Revenus"])
```

```
##
##  Shapiro-Wilk normality test
##
## data:  newdata[(newdata$Annee == "2020") & (newdata$Profession == "ouvrier"), "Revenus"]
## W = 0.97318, p-value = 0.6872
```

```
shapiro.test(newdata[(newdata$Annee=="2020")&(newdata$Profession=="technicien"),"Revenus"])
```

```
##
##  Shapiro-Wilk normality test
##
## data:  newdata[(newdata$Annee == "2020") & (newdata$Profession == "technicien"), "Revenus"]
## W = 0.97287, p-value = 0.8496
```

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




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

```
##
##  Shapiro-Wilk normality test
##
## data:  res$residuals
## W = 0.98883, p-value = 0.6449
```

Ici seule l'année semble avoir un impact sur les revenus.