```
### A. Importation des données en R + initiation à R
speccom_groupeTP2_binome6 =
read.table("speccom_groupeTP2_binome6.txt",header=TRUE)
is.data.frame(speccom_groupeTP2_binome6)

## Question 1.Faire une introduction en decrivant la population concernée
par l'étude
## et donner les variables etudiees en precisant leur type.

# prasex : variable qualitative nominale
# praspe3 : variable qualitative nominale
# secteur : variable qualitative ordinale
# honormkf : variable quantitative continue
# rmore70 : variable quantitative continue

## Question 2. Effectuer les transformations necessaires des variables
qualitatives
## en leur donnant des labels appropries.
speccom_groupeTP2_binome6$prasex=factor(speccom_groupeTP2_binome6$prasex,labels=c("F"
speccom_groupeTP2_binome6$praspe3=factor(speccom_groupeTP2_binome6$praspe3,labels=c("
chirurgicale","specialite medicale","specialite mixte"))
speccom_groupeTP2_binome6$secteur=factor(speccom_groupeTP2_binome6$secteur,labels=c("
1","secteur 2"))

## Question 3. Comme les francs ne vous disent surement rien,
## transformer la variable honormkf en honormke pour l'avoir en C (taux
de change 1 C = 6,56 francs).
speccom_groupeTP2_binome6$honormkf = speccom_groupeTP2_binome6$honormkf /
6.56

## Question 4. Etude descriptive
## Commencer l'etude statistique par une etude descriptive de toutes
## les variables du fichier (on utilisera bien sur les variables
transformees).
## Pour chaque variable, decrire sa distribution a l'aide des outils
appropries
## en fonction de son type (graphiques et statistiques).

attach(speccom_groupeTP2_binome6)
speccom_groupeTP2_binome6$prasex
prasex

## a) Variables qualitatives

# tableau de distribution en effectifs de la variable prasex
table(prasex)
# tableau de distribution en fréquences
prop.table(table(prasex))
# beaucoup trop de décimales : arrondir l'affichage à 3 décimales avec la
fonction round
# et l'argument digits
round(prop.table(table(prasex)),digits=3)
# diagramme en secteurs
pie(prop.table(table(prasex)),main="Répartition des medecins par leurs
sexes")

# tableau de distribution en effectifs de la variable praspe3
table(praspe3)
```

```
# tableau de distribution en fréquences
prop.table(table(praspe3))
# beaucoup trop de décimales : arrondir l'affichage à 3 décimales avec la
fonction round
# et l'argument digits
round(prop.table(table(praspe3)),digits=3)
# diagramme en secteurs
pie(prop.table(table(praspe3)),main="Répartition des medecins par leurs
specialites")

# tableau de distribution en effectifs de la variable praspe3
table(secteur)
# tableau de distribution en fréquences
prop.table(table(secteur))
# beaucoup trop de décimales : arrondir l'affichage à 3 décimales avec la
fonction round
# et l'argument digits
round(prop.table(table(secteur)),digits=3)
# diagramme en colonnes en fréquences relatives
barplot(prop.table(table(secteur)),main="Répartition des medecins par
leurs sectuers",
        ylab="Fréquence relative", xlab="Secteur",col=c("cyan","red"))

## Commentaires :

# pour prasex :
# les hommes sont majoritaires

# pour praspe3 :
# les specialites chirurgicales sont les plus nombreux : ils représentent
61.4% % de tous les medecins.

# pour secteur :
# les secteurs 1 sont les plus nombreux
# et les secteurs 2 sont très peu nombreux : environ 17 % des medecins.

## b) Variables quantitatives

### 1 - résumés numériques (variables continues ou discrètes)

# honormkf
summary(honormkf) # minimum, Q1, médiane, moyenne, Q3, maximum

var(honormkf)    # variance
sd(honormkf)     # écart-type
sd(honormkf)/mean(honormkf)  # coefficient de variation : permet de
mesurer la dispersion

quantile(honormkf) # min, max et quartiles (par défaut)

# histogramme
hist(honormkf,freq=FALSE,xlab="honoraires totaux du medecin pour l'annee
1999 (en euro)",
     ylab="Densité de fréquence",main="Histogramme des honoraires")

# boîte à moustaches
boxplot(honormkf,xlab="honormkf",ylab="honoraires totaux du medecin pour
l'annee 1999 (en euro)",
```

```
        main="Boîte à moustaches du honormkf")

# rmore70
summary(rmore70) # minimum, Q1, médiane, moyenne, Q3, maximum

var(rmore70)    # variance
sd(rmore70)     # écart-type
sd(rmore70)/mean(rmore70)  # coefficient de variation : permet de mesurer
la dispersion

quantile(rmore70) # min, max et quartiles (par défaut)

# histogramme des patients  de plus de 70 ans
hist(rmore70,freq=FALSE,xlab="proportion de patients ages de plus de 70
ans dans la patientele du médecin",
     ylab="Densité de fréquence",main="Histogramme des patients de plus
de 70 ans")

# boîte à moustaches
boxplot(rmore70,xlab="rmore70",ylab="proportion de patients ages de plus
de 70 ans dans la patientele du médeci",
        main="Boîte à moustaches du rmore70")


## Commentaire :

# Pour honormkf :
# L'histogramme des honoraires nous montre une distribution très
asymétrique avec une sur-représentation
# des bas honaires et des valeurs extrêmes correspondant à des hauts
honaires,
# ce que nous confirme la boîte à moustaches. Plusieurs medecins
pratiquent des honoraires tres eleves.
# l'honoraire moyen que pratique un medecin de notre echantillion est de
165,65 euros
# et le median est de 132,41 euros.
# Le coefficient de variation, égal à 0.78, révèle une forte dispersion
des honormkf.

# Pour rmore70 :
# L'histogramme des patients de plus de 70 ans nous montre une
distribution très asymétrique avec une sur-représentation
# On a tres peu de medecins qui ont une grande proportion de patiens de
plus de 70 ans.
# En moyenne, dans notre echantillon, les medecins ont 16.66 % de patiens
de plus de 70 ans
# et le median est de 14.08 %
# Le coefficient de variation, égal à 0.76, révèle une forte dispersion
des rmore70.

## Question 5. Differences hommes/femmes

# (a) Le choix de la specialite par le medecin est-il lie a son sexe?

## Liaison prasex/praspe3

# i) Tableau de contingence du couple (prasex, praspe3)
tableauPP = table(prasex, praspe3)
```

```
tableauPP
# ce sont les effectifs observés (n_ij)

# on ajoute les marges au tableau de contingence avec la fonction
addmargins
addmargins(tableauPP)

# ii) Profils-lignes et profils-colonnes

# tableau des profils-lignes
round(prop.table(tableauPP,1),digits=3)
# 59,9% des femmes sont specialites chirurgicales
# 62,2% des hommes sont specialites chirurgicales

# iii) Représentations graphiques

# profils-lignes
barplot(t(prop.table(tableauPP,
1)),beside=TRUE,col=c("blue","green","red"),
        main="Distributions conditionnelles du specialites
        sachant le sexe",xlab= "sexe",ylab="Fréquence relative",
        legend.text=TRUE)
## Commentaires :
# On constate que quelque soit le sexe, la specialite medicale est la
moins representée
# et la specialite chirurgicale est la plus representée.


# iv) Test du Khi-deux d'indépendance
# H0 : indépendance du sexe et de la specialite
# H1 : pas indépendance (= les variables sont liées)
# conditions de validité
round(chisq.test(tableauPP)$expected,digits=2)
# tous les effectifs attendus sous H0 (= e_ij) dépassent 5 donc le test
du khi-deux
# d'indépendance est valide

khi2PP=chisq.test(tableauPP)
khi2PP
# conclusion : la p-valeur est inférieure à 5 % donc on rejette H0.
# Il y a donc une liaison entre le sexe et la specialite.
# On va donc préciser cette liaison par l'étude des contributions.

# Contributions au Khi2
round(khi2PP$residuals^2,digits=2)
# 3 forte contributions : (F,specialite medicale), (F, specialite mixte)
et (M, specialite medicale)

## Commentaire :
# On constate que la majorite des femmes ont pour specialite la chirurgie
medicale comme leurs confreres
# cepandant on voit que l'ecart d'effectif qu'il y a entre les femmes qui
ont pour specialite <medicale> et mixte
# est tres eleve comparé aux homologues hommes.

## Conclusion :
# Le choix de la specialite par le medecin est lie a son sexe.
```

```
# (b) Les medecins specialistes masculins gagnent-ils mieux leur vie que
leurs consœurs?

## Liason honormkf/prasex

# i) Boîtes à moustaches juxtaposées
boxplot(honormkf~prasex,main="Boîtes à moustaches juxtaposées honormkf en
fonction du prasex")

# ii)
# Moyennes par groupe
moycond=tapply(honormkf,prasex,mean)
moycond

# Variances par groupe
tapply(honormkf,prasex,var)

points(moycond,col="red",pch="*",cex=2)

# Commentaires pour les deux questions (i et ii)
# on constate que la moyenne des honoraires des hommes est beaucoup plus
grande que celle des femmes.
# c'est vrai aussi pour les médianes et les dispersions (visible sur les
variances mais aussi sur
# les intervalles interquartiles des boîtes à moustaches). On voit en
particulier que la médiane
# des honoraires des femmes est inférieure au 1er quartile des honoraires
des hommes, ce qui signifie que
# la moitié des femmes ont des honoraires qui n'atteignent pas les
honoraires que touchent
# 75 % des hommes. Les deux distributions présentent des valeurs extrèmes
(de tres hauts honorairies)

# iii) Test statistique
# on veut tester H_0 : mu_F=mu_H contre H_1 : mu_F différent de mu_H
# on va utiliser un test paramétrique
# (test de comparaison des moyennes de Student si les
# variances sont égales et test de Welch sinon)
# Conditions de validité : loi gaussienne
# pour chaque échantillon ou
# n_F et n_H grand (supérieurs à 30 pour fixer les idées)
# ce sont des salaires donc pas gaussien mais n_F et n_H assez grands
# on peut le vérifier par
table(prasex)
tapply(honormkf,prasex,hist,freq=FALSE)
# on voit que les distributions ne sont pas gaussiennes donc il faut
avoir
# des effectifs suffisants pour que le test soit valide

# procédure en deux étapes

# Etape 1 : test de comparaison des variances de Fisher
# H0 : les variances des honoraires sont égales chez les hommes et chez
les femmes
# contre H1 : les variances des honoraires sont différentes chez les
hommes
# et chez les femmes
var.test(honormkf~prasex)
```

```
# Comme la p-valeur qui vaut 2,2 10^(-16) est inférieure à alpha (0.05),
# on rejette l'égalité des variances et conclut que les variances des
honoraires
# dans les deux groupes sont différentes.

# Etape 2 : on doit donc faire le test de Welch, avec variances inégales
t.test(honormkf~prasex,var.equal=FALSE)
# on rejette l'égalité des moyennes, et à l'aide des "moycond" calculées
plus haut,
# on conclut que les honoraires des hommes sont en moyenne
# significativement plus élevés que ceux des femmes
# (203.04 euros pour les medecins hommes versus 97.37 euros pour leurs
homologues femmes)

## Conclusion : Les medecins specialistes masculins gagnen mieux leur vie
que leurs consœurs.

## Question 6. Influence du secteur tarifaire

# (a) La repartition par secteur est-elle la meme dans les trois types de
specialites?

## Liason secteur/praspe3

# i) Tableau de contingence du couple (secteur, praspe3)
tableauSP = table(secteur, praspe3)
tableauSP
# ce sont les effectifs observés (n_ij)

# on ajoute les marges au tableau de contingence avec la fonction
addmargins
addmargins(tableauSP)

# ii) profils-colonnes

# tableau des profils-colonnes
round(prop.table(tableauSP,2),digits=3)
# 94,1% des specialites chirurgicales sont dans le secteur 1
# 67,9% des specialites medicales sont dans le secteur 2
# 77,4% des specialites mixtes dans le secteur 1

# iii) Représentations graphiques

# profils-colonnes
barplot(prop.table(tableauSP,
2),beside=TRUE,col=c("magenta","blue"),ylab="Fréquence relative",
        main="Distributions conditionnelles du secteur sachant le
specialite",
        legend.text=TRUE)

## Commentaire:
# On constate que pour la specialite chirurgicale et la specialite mixte,
le secteur 1 est majoritaire,
# et pour la specialite medicale, le secteur 2 est majoritaire.

# iv) Test du Khi-deux d'indépendance
# H0 : indépendance du  et de la specialite
# H1 : pas indépendance (= les variables sont liées)
```

```
# conditions de validité
round(chisq.test(tableauSP)$expected,digits=2)
# tous les effectifs attendus sous H0 (= e_ij) dépassent 5 donc le test
du khi-deux
# d'indépendance est valide

khi2SP=chisq.test(tableauSP)
khi2SP
# conclusion : la p-valeur est inférieure à 5 % donc on rejette H0.
# Il y a donc une liaison entre le secteur et la specialite.
# On va donc préciser cette liaison par l'étude des contributions.

# Contributions au Khi
round(khi2SP$residuals^2,digits=2)
# 3 fortes contributions : (secteur 2,specialite medicale), (Secteur 2,
specialite chirurgicale) et (secteur 1, specialite medicale)

## Commentaire :
# On constate que les medecins ayant pour specialite la specialite
medicale et chirurgicale et
# faisant partie du secteur 2 contribuent beaucoup au khi 2 et dans une
moindre mesure,
# les medecins du secteur 1 ayant partie pour specialite medicale
contribue aussi au khi2.

## Conclusion :
# La repartition par secteur n'est pas la meme dans les trois types de
specialites.


# (b) Le secteur tarifaire auquel il appartient a-t-il un effet sur les
honoraires du medecin?

## Liason secteur/honoraire

# i) Boîtes à moustaches juxtaposées
boxplot(honormkf~secteur,main="Boîtes à moustaches juxtaposées honormkf
en fonction du secteur")

# ii)
# Moyennes par groupe
moycond=tapply(honormkf,secteur,mean)
moycond

# Variances par groupe
tapply(honormkf,secteur,var)

points(moycond,col="red",pch="*",cex=2)

# Commentaires pour les deux questions (i et ii)
# on constate que la moyenne des honoraires dans le secteur 2 est plus
grande que celle dans le secteur 1.
# c'est vrai aussi pour les médianes et les dispersions.
# On voit en particulier que la médiane des honoraires dans le secteur 1
est inférieure au 1er quartile des honoraires dans le secteur 2,
# ce qui signifie que la moitié du secteur 1 ont des honoraires qui
n'atteignent pas les honoraires que touchent
```

```
# 75 % du secteur 2. La distribution pour le secteur 1 présente des
valeurs extrèmes (de tres hauts honorairies)

# iii) Test statistique
# on veut tester H_0 : mu_F=mu_H contre H_1 : mu_F différent de mu_H
# on va utiliser un test paramétrique
# (test de comparaison des moyennes de Student si les
# variances sont égales et test de Welch sinon)
# Conditions de validité : loi gaussienne
# pour chaque échantillon ou
# n_F et n_H grand (supérieurs à 30 pour fixer les idées)
# ce sont des salaires donc pas gaussien mais n_F et n_H assez grands
# on peut le vérifier par
table(secteur)
tapply(honormkf,secteur,hist,freq=FALSE)
# on voit que les distributions ne sont pas gaussiennes donc il faut
avoir
# des effectifs suffisants pour que le test soit valide

# procédure en deux étapes

# Etape 1 : test de comparaison des variances de Fisher
# H0 : les variances des honoraires sont égales chez les hommes et chez
les femmes
# contre H1 : les variances des honoraires sont différentes chez les
hommes
# et chez les femmes
var.test(honormkf~secteur)
# Comme la p-valeur qui vaut 2,2 10^(-16) est inférieure à alpha (0.05),
# on rejette l'égalité des variances et conclut que les variances des
honoraires
# dans les deux groupes sont différentes.

# Etape 2 : on doit donc faire le test de Welch, avec variances inégales
t.test(honormkf~secteur,var.equal=FALSE)
# on rejette l'égalité des moyennes, et à l'aide des "moycond" calculées
plus haut,
# on conclut que les honoraires dans le secteur 2 sont en moyenne
# significativement plus élevés que ceux dans le secteur 1
# (207,35 euros pour les medecins dans le secteur 2 versus 156,84 euros
pour leurs homologues dans le secteur 1)

## Conclusion :
# Le secteur tarifaire auquel le medecin appartient influe sur les
honoraires du medecin :
# Les medecins dans le secteur 2 sont plus susceptible de gagner mieux
que leurs homologues du secteur 1.

# Question 7. Effet des autres variables sur les honoraires

# (a) Les honoraires d'un medecin dependent-ils de la proportion de
personnes agees dans sa patientele ?
## liason rmore70/honoraires
# i) Nuage de points
plot(rmore70,honormkf,main="Nuage de points", xlab="Proportion de
patients de plus de 70 ans",
     ylab="honorraires des medecins en euros (1999)")
# le nuage semble reparti autour d'une droite tres faiblement croissante
```

```
# cepandant le nuage de point semble dispersé

# ii) Coefficient de corrélation linéaire empirique (arrondi à deux
décimales)
round(cor(rmore70,honormkf),digits=2)
# le coefficient est plutôt faible

# iii) Test statistique
# H0 : absence de liaison linéaire
# entre la proportion de patients de plus de 70 ans
# et les honoraires des medecins
# qu'on écrit aussi : rho = 0

# H1 : liaison linéaire
# qu'on peut aussi écrire : rho différent de 0

# conditions de validité : variables gaussiennes
# pas trop vrai ici mais test valide quand même
# car n = 500 grand
# nom : test du coefficient de corrélation linéaire
cor.test(rmore70,honormkf)

#Conclusion
# la p-valeur (3.557e-12) du test est inférieure au niveau de 5 % donc on
rejette H0.
# On conclut donc qu'il y a une liaison linéaire positive (car le
coefficient
# de corrélation linéaire empirique de Pearson est 0.304 > 0)
# entre la proportion de patients de plus de 70 ans et les honorraires
des medecins.
# Donc, sans surprise, plus on a les patients de plus de 70 ans, meilleur
sont les honoraires des medecins.

# (b) Le type de specialite a-t-il une influence sur les honoraires des
medecins?
# Si oui, dire quelles specialites different significativement en termes
d'honoraires.

## liason specialite/honoraires

# i) Boîtes à moustaches juxtaposées
boxplot(honormkf~praspe3,xlab="specialite des medecins",
        main="Boîtes à moustaches juxtaposées du honoraires en fonction
du specialite")

# ii)
# Moyennes par groupe
moycond2=tapply(honormkf,praspe3,mean)
moycond2
# on rajoute les moyennes conditionnelles aux bàm juxtaposées
points(moycond2,col="red",pch="*",cex=2)

# Variances par groupe
tapply(honormkf,praspe3,var)
# la variance empirique est beaucoup plus importante dans le groupe de
specialite chirurgicale
# donc plus grande dispersion dans ce groupe
```

```
# Commentaires
# On voit trois boîtes à moustaches très différentes. La médiane des
specialites medicale est bien
# au dessus de celles des deux autres specialites et cela est aussi vrai
pour les moyennes.
# Les medecins de la specialite medicale et les medecins de la specialite
mixte ont des
# honoraires médians et moyens assez simimaires (6220 et 8741 euros
annuels respectivement).
# La dispersion des honoraires des specialites chirurgicales est aussi
beaucoup plus grande.
# La dispersion des honoraires des specialites medicales et des
specialites mixtes est très faible.

# iii) Test statistique : test de l'analyse de variance car la variable
qualitative a 3 modalités
# H0 : mu_1=mu_2=mu_3
# contre H1 : au moins deux moyennes diffèrent

# Condition de validité : loi gaussienne pour chaque échantillon ou
# effectifs par groupe assez grands (tous doivent être supérieurs à 30)
table(praspe3)
par(mfrow=c(1,3))
tapply(honormkf,praspe3,hist,freq=FALSE)
par(mfrow=c(1,1))

# on n'a donc pas la normalité pour les trois groupes d'après les
histogrammes
# mais les effectifs sont élevés
# on peut donc considérer qu'on peut approcher les moyennes empiriques
par groupe par
# des lois gaussiennes

# 2 tests possibles pour comparer les moyennes : le test classique de
l'analyse
# de variance ou sa version modifiée en cas de variances inégales
# pour choisir quel test faire :
# on teste l'homoscédasticité = homogénéité des variances par groupe
# test de Brown-Forsythe : H0 : sigma_1^2=sigma_2^2=sigma_3^2
library(car)    # nécessite le package "car"
leveneTest(honormkf,praspe3)
# on rejette l'égalité des variances (p-valeur est inférieure à 0.05)


# il faut donc utiliser la variante de l'analyse de la variance pour
variances inégales
oneway.test(honormkf~praspe3,var.equal=FALSE)
# p < 0.05 donc on rejette H0 et on conclut que les moyennes diffèrent
globalement

# si on conclut à  des moyennes différentes (= rejeter HO), alors on
continue pour savoir quelles
# moyennes diffèrent deux à deux (sinon, on a fini)
# donc là, on continue ave les tests de comparaison multiples
# les tests vus en cours estiment un sigma2 commun (car hypothèse
d'homoscédasticité).
# c'est l'argument pool.sd=TRUE
```

```
# si variances inégales (c'est le cas ici), il vaut mieux prendre une
version
# "non pooled" avec pool.sd=FALSE

# test LSD (pas d'ajustement du risque alpha)
pairwise.t.test(honormkf,praspe3,p.adjust.method= "none",pool.sd=FALSE)

# on conclut ici que les trois populations sont différentes deux à deux
car toutes les
# p-valeurs sont inférieures à 0.05
# et en s'aidant des valeurs des moyennes empiriques, on peut conclure à
l'ordre suivant :
# mu_E < mu_A < mu_M
# c'est-à-dire que les salaires des managers sont significativement plus
élevés que ceux des
# agents de sécurité qui sont significativement plus élevés que ceux des
employés de bureau.

# Question 8. Conclusion de l'etude
# Faire une courte synthese (eventuellement accompagne e d'un tableau
recapitulatif)
# montrant ce que vous avez compris des relations entre les differentes
variables
# et en particulier des facteurs qui influent sur les honoraires.
```