Autour du paradoxe de simpson

Inès ACHIN

2022-10-17

##1. Importer les données

```
data = read.csv('C:/Users/iachin/Downloads/Subject6_smoking.csv', header =T)
head(data)
```

```
##   Smoker Status  Age
## 1    Yes  Alive 21.0
## 2    Yes  Alive 19.3
## 3     No   Dead 57.5
## 4     No  Alive 47.1
## 5    Yes  Alive 81.4
## 6     No  Alive 36.8
```

##2. Questions 1 #Représentez dans un tableau le nombre total de femmes vivantes et décédées sur la période en fonction de leur habitude de tabagisme.

```
T<- table(data$Smoker,data$Status)
addmargins(T)
```

```
##
##       Alive Dead  Sum
##   No    502  230  732
##   Yes   443  139  582
##   Sum   945  369 1314
```

```
T2 = table(data$Smoker,data$Status) ##Compilation en pourcentage
T2 <- prop.table(T,margin=1)*100
T2
```

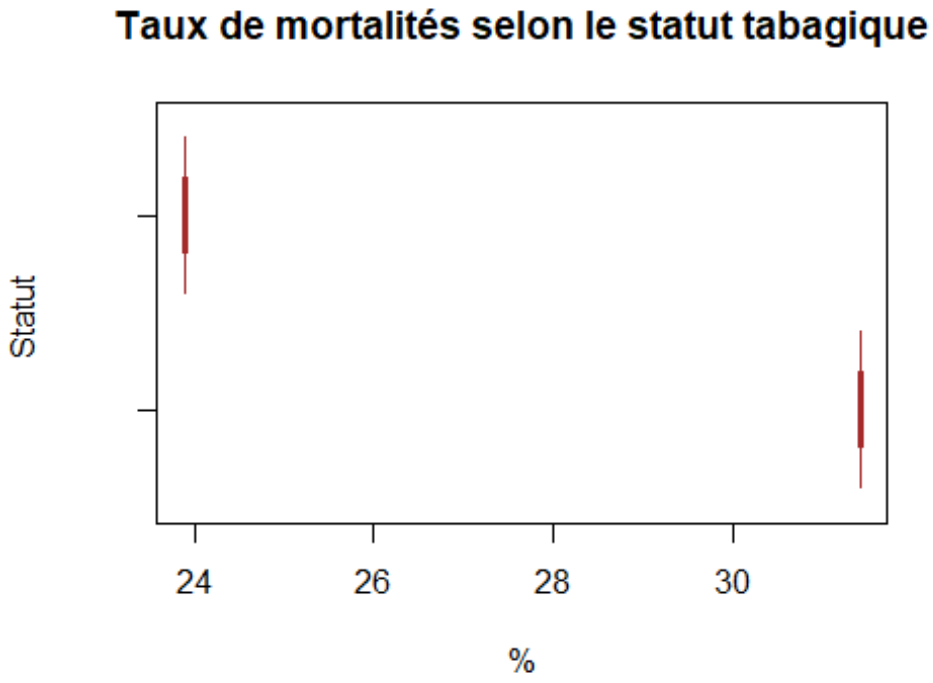
```
##
##          Alive     Dead
##   No  68.57923 31.42077
##   Yes 76.11684 23.88316
```

Le taux de mortalité chez les non fumeuses est de 31,4%. Le taux de mortalité chez les fumeuses est de 23,8%.

#Calculez dans chaque groupe (fumeuses / non fumeuses) le taux de mortalité (le rapport entre le nombre de femmes décédées dans un groupe et le nombre total de femmes dans ce groupe). Vous pourrez proposer une représentation graphique de ces données et calculer des intervalles de confiance si vous le souhaitez. En quoi ce résultat est-il surprenant ?

```r
taux_non_fumeurs<-T2[1,2]
taux_fumeurs<-T2[2,2]

p <-boxplot(taux_non_fumeurs,taux_fumeurs,
main = "Taux de mortalités selon le statut tabagique",
xlab = "%",
ylab = "Statut ",
col = "orange",
border = "brown",
horizontal = TRUE,
notch = TRUE
)
```



##3. Questions 2 #Reprenez la question 1 (effectifs et taux de mortalité) en rajoutant une nouvelle catégorie liée à la classe d'âge. On considérera par exemple les classes suivantes : 18-34 ans, 34-54 ans, 55-64 ans, plus de 65 ans. En quoi ce résultat est-il surprenant ? Arrivez-vous à expliquer ce paradoxe ? De même, vous pourrez proposer une représentation graphique de ces données pour étayer vos explications.

```r
df<-subset(data,Smoker=="Yes",select=c(Smoker,Status,Age))
ageclass <- cut(df$Age,c(18,34,54,64,100),right=FALSE, include.lowest=TRUE)
tab2<-table(ageclass,df$Status)
addmargins(tab2)
```

```
##
## ageclass   Alive Dead Sum
##   [18,34)    174    5 179
```

```
##   [34,54)    198   41 239
##   [54,64)     64   51 115
##   [64,100]     7   42  49
##   Sum        443  139 582
```

```
f<-prop.table(tab2,margin=1)*100#regroupe en classe d'âge

dnf<-subset(data,Smoker=="No",select=c(Smoker,Status,Age))#pour les non
fumeuses
ageclass <- cut(dnf$Age,c(18,34,54,64,100),right=FALSE, include.lowest=TRUE)
tab3<-table(ageclass,dnf$Status)
addmargins(tab3)
```

```
##
## ageclass   Alive Dead Sum
##   [18,34)    213    6 219
##   [34,54)    180   19 199
##   [54,64)     80   39 119
##   [64,100]    29  166 195
##   Sum        502  230 732
```

```
nf<-prop.table(tab3,margin=1)*100


maliste<-list(nf,f)
names(maliste)<-c("Non Fumeurs","Fumeurs")
maliste
```

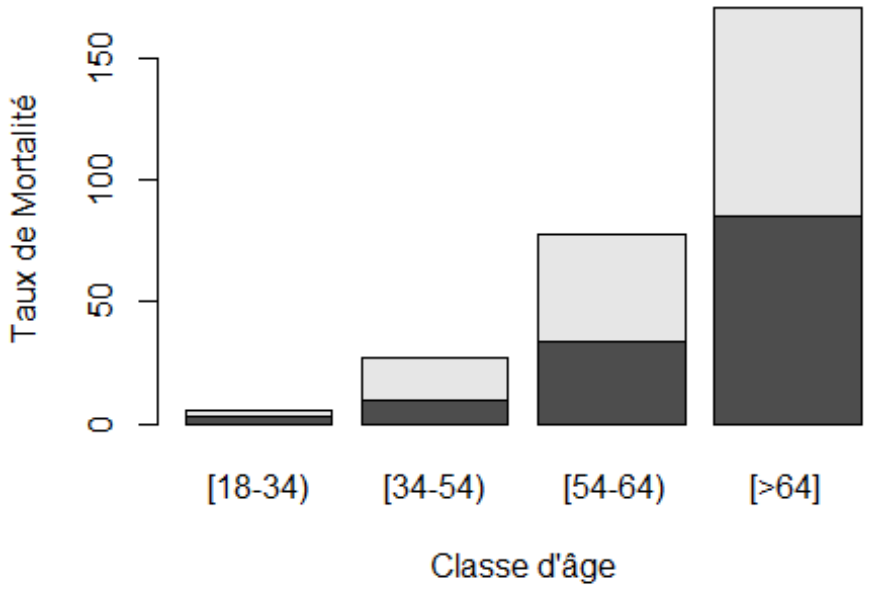
```
## $`Non Fumeurs`
##
## ageclass       Alive      Dead
##   [18,34)  97.260274  2.739726
##   [34,54)  90.452261  9.547739
##   [54,64)  67.226891 32.773109
##   [64,100] 14.871795 85.128205
##
## $Fumeurs
##
## ageclass       Alive      Dead
##   [18,34)  97.206704  2.793296
##   [34,54)  82.845188 17.154812
##   [54,64)  55.652174 44.347826
##   [64,100] 14.285714 85.714286
```

Le taux de mortalité est plus elevé chez les fumeuses, pour chaques classes d'âges.Ceci semble être contraire aux résultats de la question 1.

#Représentation graphique

```
t_non_fumeurs=c(2.739726,9.547739,33.61345,85.12821)
t_fumeurs=c(2.793296,17.154812,44.34783,85.71429)
```

```
classe_age=c("[18-34)","[34-54)"," [54-64)","[>64]")
mortalité_age=c(t_non_fumeurs,t_fumeurs)
mortalité_age=matrix(mortalité_age,nc=4,nr=2,byrow=T)
colnames(mortalité_age)=classe_age
barplot(mortalité_age,xlab="Classe d'âge", ylab="Taux de Mortalité")
```



Dans le groupes non fumeuses, il y a plus de personnes agés que dans les fumeuses.

4. Question 3

Afin d'éviter un biais induit par des regroupements en tranches d'âges arbitraires et non régulières, il est envisageable d'essayer de réaliser une régression logistique. Si on introduit une variable Death valant 1 ou 0 pour indiquer si l'individu est décédé durant la période de 20 ans, on peut étudier le modèle Death ~ Age pour étudier la probabilité de décès en fonction de l'âge selon que l'on considère le groupe des fumeuses ou des non fumeuses. Ces régressions vous permettent-elles de conclure sur la nocivité du tabagisme ? Vous pourrez proposer une représentation graphique de ces régressions (en n'omettant pas les régions de confiance).

```
death=FALSE
death[data$Status=="Dead"] <- 1
death[data$Status=="Alive"] <-0
Age <- data$Age


df<-subset(data,Smoker=="Yes",select=c(Smoker,Status,Age))
death1=FALSE
```

```r
death1[df$Status=="Dead"]<-1
death1[df$Status=="Alive"]<-0
reg4<-glm(death1~Age,data=df,family=binomial(logit))
summary(reg4)
```

```
##
## Call:
## glm(formula = death1 ~ Age, family = binomial(logit), data = df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.0745  -0.6464  -0.3756  -0.2013   2.6560
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -5.508106   0.466221  -11.81   <2e-16 ***
## Age          0.088977   0.008721   10.20   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 639.89  on 581  degrees of freedom
## Residual deviance: 480.41  on 580  degrees of freedom
## AIC: 484.41
##
## Number of Fisher Scoring iterations: 5
```

```r
exp(cbind(coef(reg4), confint(reg4)))
```

```
## Attente de la réalisation du profilage...
```

```
##                               2.5 %      97.5 %
## (Intercept) 0.004053779 0.001549786 0.009681186
## Age         1.093054975 1.075312303 1.112796977
```

```r
dnf<-subset(data,Smoker=="No",select=c(Smoker,Status,Age))
death2=FALSE
death2[dnf$Status=="Dead"]<-1
death2[dnf$Status=="Alive"]<-0
reg5<-glm(death2~Age,data=dnf,family=binomial(logit))
summary(reg5)
```

```
##
## Call:
## glm(formula = death2 ~ Age, family = binomial(logit), data = dnf)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4019  -0.5179  -0.2003   0.4728   3.0457
##
```

```
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -6.795507   0.479430  -14.17   <2e-16 ***
## Age          0.107275   0.007806   13.74   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 911.23  on 731  degrees of freedom
## Residual deviance: 519.08  on 730  degrees of freedom
## AIC: 523.08
##
## Number of Fisher Scoring iterations: 6
```

```
exp(cbind(coef(reg5), confint(reg5)))
```

```
## Attente de la réalisation du profilage...
```

```
##                                2.5 %      97.5 %
## (Intercept) 0.001118791 0.0004158016 0.002733484
## Age         1.113240705 1.0971189020 1.131283975
```

Dans le groupe des non fumeur, effet de l'âge sur la mortalité.