Paradoxe de Simpson

EC30

2/1/2021

Intitulé de l'exercice : *En 1972-1974, à Whickham, une ville du nord-est de l'Angleterre, située à environ 6,5 kilomètres au sud-ouest de Newcastle upon Tyne, un sondage d'un sixième des électeurs a été effectué afin d'éclairer des travaux sur les maladies thyroïdiennes et cardiaques (Tunbridge et al. 1977). Une suite de cette étude a été menée vingt ans plus tard (Vanderpump et al. 1995). Certains des résultats avaient trait au tabagisme et cherchaient à savoir si les individus étaient toujours en vie lors de la seconde étude. Par simplicité, nous nous restreindrons aux femmes et parmi celles-ci aux 1314 qui ont été catégorisées comme "fumant actuellement" ou "n'ayant jamais fumé". Il y avait relativement peu de femmes dans le sondage initial ayant fumé et ayant arrêté depuis (162) et très peu pour lesquelles l'information n'était pas disponible (18). La survie à 20 ans a été déterminée pour l'ensemble des femmes du premier sondage.*

Question 1

Représentez dans un tableau le nombre total de femmes vivantes et décédées sur la période en fonction de leur habitude de tabagisme. Calculez dans chaque groupe le taux de mortalité. 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 ?

```
d<-read.csv("https://gitlab.inria.fr/learninglab/mooc-rr/mooc-rr-ressources/-/raw/master/module3/Practi
```

Pour cette exercice, nous ouvrons une base de données regroupant le statut tabagique des femmes interrogées, leur état (Alive/Dead) et leur âge.

Nous souhaitons étudier la variable vivante/décédée dans le groupe des fumeuses en comparaison avec le groupe des non fumeuses.

```
t<-table(d$Smoker, d$Status)
addmargins(t)
```

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

Nous avons donc créer un tableau de données croisées du statut "vivante ou décédée" en fonction du statut "fumeuse ou non fumeuse".

Nous allons à partir de là pouvoir calculer le taux de mortalité dans chaque groupe ("fumeuses"/"non fumeuses").

```
t<-table(d$Smoker, d$Status)
t2<-prop.table(t,margin=1)*100
t2
```

```
##
##          Alive     Dead
```

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

On a ainsi calculé la fréquence relative du statut “vivante” et du statut “décédée” pour chaque groupe (fumeuses/non fumeuses) qu’on a ensuite ramené en pourcentage.

Le taux de mortalité chez les non fumeuses est donc d’environ 31,42%.

Le taux de mortalité chez les fumeuses est quant à lui d’environ 23,88%.

```
binom.test(230,732)$conf.int
```

```
## [1] 0.2807031 0.3492176
## attr(,"conf.level")
## [1] 0.95
```

```
binom.test(139,582)$conf.int
```

```
## [1] 0.2047323 0.2756061
## attr(,"conf.level")
## [1] 0.95
```

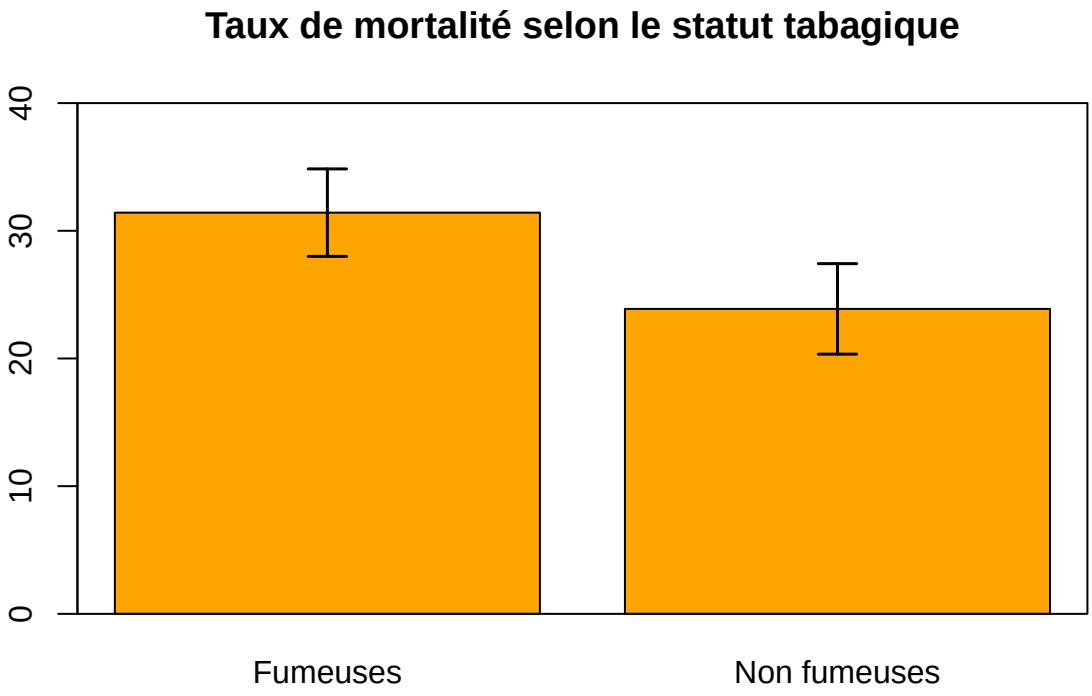
Le **taux de mortalité chez les non fumeuses qui est de 31.42077**, présente un **intervalle de confiance à 95 % qui vaut alors [28.07031; 34.92176]**.

Le **taux de mortalité chez les fumeuses qui est de 23.88316**, présente un **intervalle de confiance à 95 % qui vaut alors [20.47323; 27.56061]**.

Notre résultat montre un taux de mortalité plus élevé chez les non fumeuses ce qui semble à priori surprenant étant donné que le tabac est un facteur de risque de pathologies cardio-vasculaires et on pourrait ainsi se dire que le taux de mortalité devrait être plus élevé dans ce groupe.

On peut représenter cette surprenante différence du taux de mortalité en fonction de l’habitude tabagique du sujet par un diagramme sur lequel on ajoute les intervalles de confiance pour le taux de mortalité de chaque groupe.

```
taux_non_fumeurs<-t2[1,2]
taux_fumeurs<-t2[2,2]
mortalité<-c(taux_non_fumeurs,taux_fumeurs)
noms_barres <- c("Fumeuses","Non fumeuses")
intervalles=c(100*((binom.test(230,732)$conf.int[2]-binom.test(230,732)$conf.int[1])/2),100*((binom.tes
bp<-barplot(mortalité,col="orange",ylim=c(0,40),names.arg=noms_barres, main="Taux de mortalité selon le
arrows(bp,mortalité-intervalles,bp,mortalité+intervalles,lwd=1.5,angle=90,length=0.1,code=3)
```



Question 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, 54-64 ans, plus de 64 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.

On s'intéresse d'abord au groupe des fumeuses.

```
df<-subset(d,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
```

On l'isole et on regroupe les âges en classe d'âge afin d'étudier leur statut vivant ou mort.

```
f<-prop.table(tab2,margin=1)*100
```

On a ensuite calculé les fréquences des évènements "Alive" et "Dead" pour chaque classe d'âge.

Maintenant intéressons nous au groupe des non fumeuses.

```
dnf<-subset(d,Smoker=="No",select=c(Smoker,Status,Age))
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
```

On fait de même en affichant le tableau statut "Alive/Dead" en fonction de la classe d'âge. Et on calcule la fréquence de chaque évènement par classe d'âge chez les non fumeuses.

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

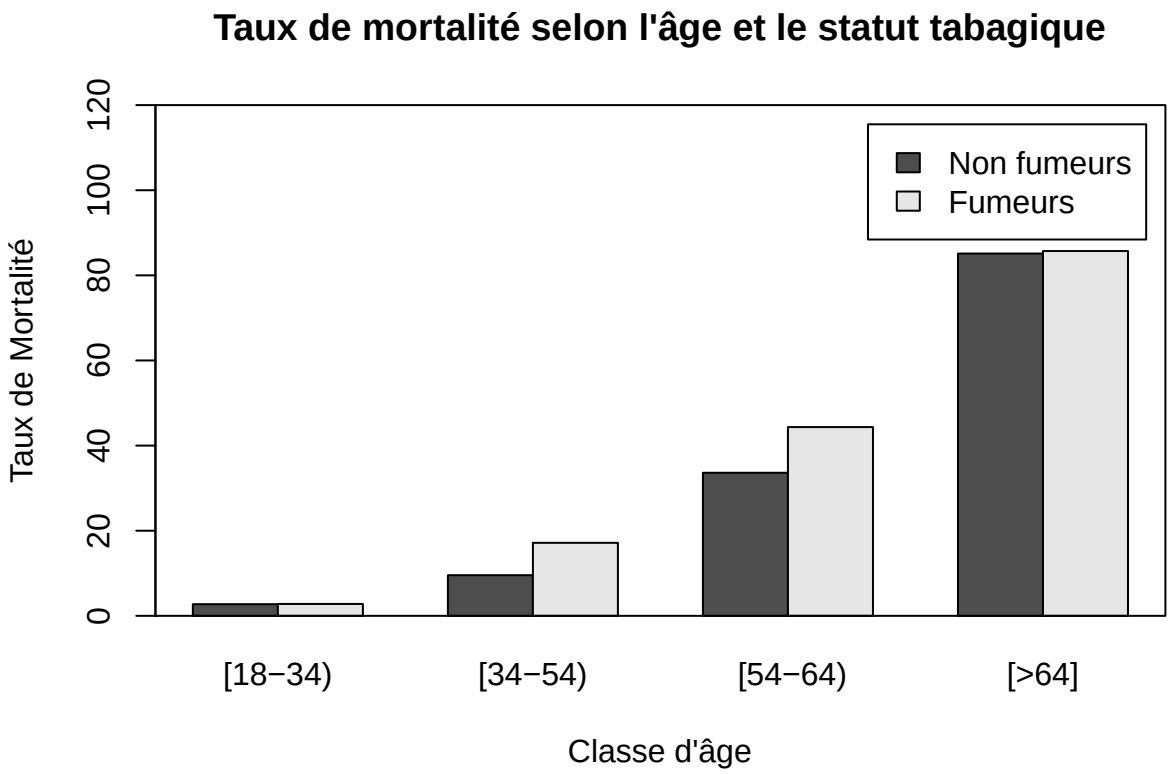
```
maliste<-list(nf,f)
names(maliste)<-c("Non Fumeurs","Fumeurs")
maliste
```

```
## $`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
```

Ainsi, bien que dans la réponse à la première question, on trouvait un taux de mortalité plus élevé chez les non fumeuses, en s'intéressant à une autre donnée : l'âge, on voit désormais que **le taux de mortalité est plus élevé dans chaque classe d'âge dans le groupe des fumeuses** par rapport à celui des non fumeuses. Ces résultats semblent surprenant car a priori en opposition avec ceux de la question 1. En fait, il faut étudier tous les paramètres pour arriver à la bonne conclusion.

On va représenter ce résultat par un digramme en barres accolées.

```
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,beside=T,xlab="Classe d'âge", ylab="Taux de Mortalité", legend.text=c("Non fumeurs
```



Dans le groupe des non fumeuses on voit que la proportion de personnes dites “âgées” est bien plus importante que la proportion des personnes dites “âgées” dans le groupe des fumeuses. Les personnes âgées étant celles avec un taux de mortalité le plus important dans la population générale, il est normal que le taux de mortalité du groupe des fumeuses ait été impacté par cette différence de proportion. Le taux de mortalité calculé à la question 1 ne dépendait pas que du statut fumeuse ou non fumeuse.

Cela se vérifie en comparant pour chaque groupe la description statistique de la variable âge.

```
tapply(d$Age,d$Smoker,mean)
```

```
##       No      Yes
## 49.81585 44.26976
```

```
t.test(Age~Smoker,data=d)
```

```
##
##  Welch Two Sample t-test
##
## data:  Age by Smoker
## t = 5.4161, df = 1311.3, p-value = 7.237e-08
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  3.537244 7.554931
## sample estimates:
##  mean in group No mean in group Yes
##          49.81585           44.26976
```

On voit alors que la moyenne d’âge est significativement plus élevée chez les non fumeurs .

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).

On introduit donc la variable death qui vaut 1 lorsque le sujet est mort et 0 lorsqu'il est vivant.

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

A partir de là on étudie la probabilité de décès en fonction de l'âge chez les fumeuses puis chez les non fumeuses.

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

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

```
## Waiting for profiling to be done...
```

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

```r
dnf<-subset(d,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
```

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

```
## Waiting for profiling to be done...
```

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

Dans le groupe des non fumeurs et des fumeurs, on voit que l'âge a une influence significative sur la mortalité, ce qui est logique le taux de mortalité augmente avec le veillissement.

```r
OR <-c(1.093054975,1.113240705)
IC2.5=c(1.075312303,1.0971189020)
IC97.5=c(1.112796977,1.131283975)
z<-data.frame(OR,IC2.5,IC97.5)
type=c("OR","IC 2.5%","IC 97.5%")
type2=c("Fumeurs","Non Fumeurs")
colnames(z)=type
rownames(z)=type2
z
```

```
##                   OR  IC 2.5% IC 97.5%
## Fumeurs     1.093055 1.075312 1.112797
## Non Fumeurs 1.113241 1.097119 1.131284
```

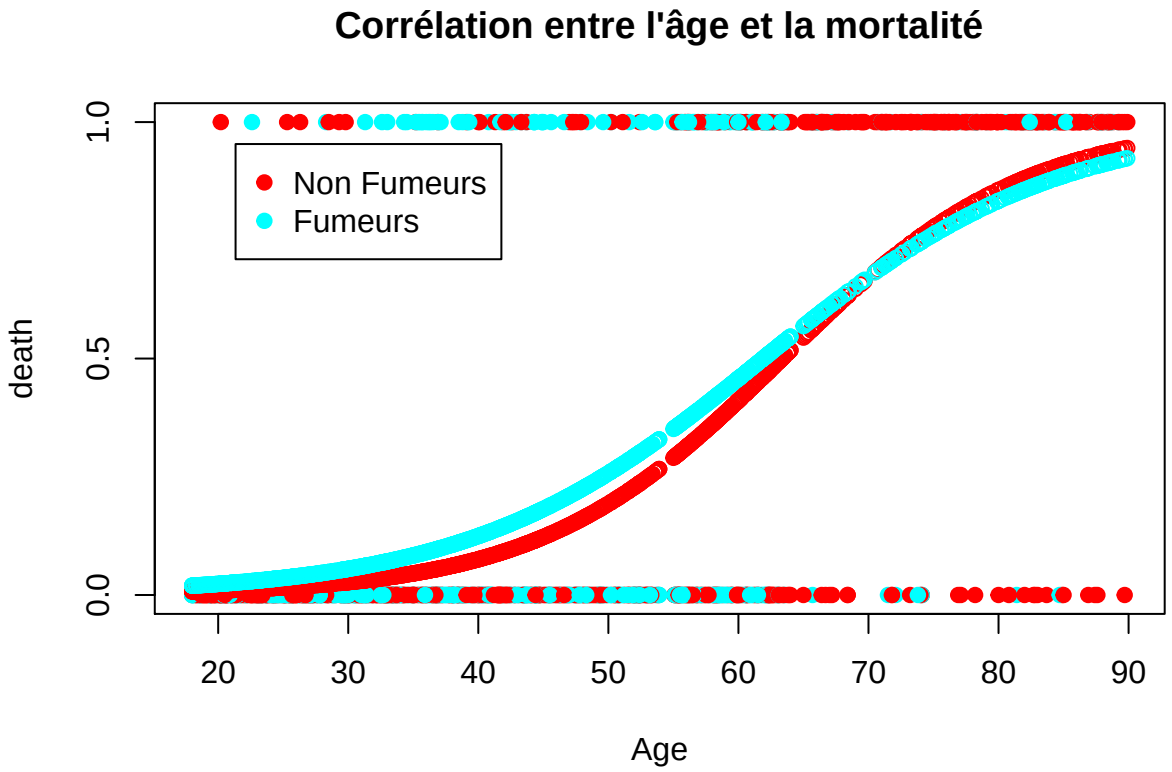
Voici la comparaison des corrélations entre la mortalité et l'âge chez les fumeuses et les non fumeuses avec les intervalles de confiance. **L'intervalle de confiance à 95 % chez les fumeuses vaut alors [1.075312 ;**

1.112797] L'intervalle de confiance à 95 % chez les non fumeuses vaut alors [1.097119 ; 1.131284]

```
Smokerfac<-factor(d$Smoker)
mescouleurs<-rainbow(length(levels(Smokerfac)))
logit_ypredit1= -5.508106 + 0.088977*Age
ypredit1 = exp(logit_ypredit1)/(1+ exp(logit_ypredit1))

logit_ypredit2= -6.795507+ 0.107275*Age
ypredit2 = exp(logit_ypredit2)/(1+ exp(logit_ypredit2))

plot(Age, death, pch = 19, yaxp=c(0,1,2),col = mescouleurs[Smokerfac],main="Corrélation entre l'âge et l
legend("topleft", inset=0.08, pch=19, legend=c("Non Fumeurs", "Fumeurs"), col=c("red", "#00FFFF"))
points(Age,ypredit2,col="red")
points(Age,ypredit1,col="#00FFFF")
```



Grâce à cette représentation graphique et au calcul des intervalles de confiance, on voit que la mortalité chez les fumeurs est plus importante jusqu'à environ 70ans. Puis elle est légèrement inférieure à la mortalité des non fumeurs chez les plus de 70ans.

Ainsi on a tendance à conclure à une nocivité du tabac puisque en comparaison avec le groupe des non fumeuses, les fumeuses ont une probabilité de décès plus importante pour des âges relativement peu élevés.