

Autour du Paradoxe de Simpson

Laura

18 janvier 2021

Informations technique de l'ordinateur réalisant le code

```
sessionInfo()
```

```
## R version 4.0.3 (2020-10-10)
## Platform: x86_64-w64-mingw32/x64 (64-bit)
## Running under: Windows 10 x64 (build 19041)
##
## Matrix products: default
##
## locale:
## [1] LC_COLLATE=French_France.1252 LC_CTYPE=French_France.1252
## [3] LC_MONETARY=French_France.1252 LC_NUMERIC=C
## [5] LC_TIME=French_France.1252
##
## attached base packages:
## [1] stats      graphics  grDevices  utils      datasets  methods    base
##
## loaded via a namespace (and not attached):
## [1] compiler_4.0.3  magrittr_2.0.1  tools_4.0.3    htmltools_0.5.0
## [5] yaml_2.2.1      stringi_1.5.3   rmarkdown_2.6  knitr_1.30
## [9] stringr_1.4.0   xfun_0.20       digest_0.6.27  rlang_0.4.10
## [13] evaluate_0.14
```

```
devtools::session_info()
```

```
## - Session info -----
## setting value
## version R version 4.0.3 (2020-10-10)
## os      Windows 10 x64
## system  x86_64, mingw32
## ui      RTerm
## language (EN)
## collate French_France.1252
## ctype   French_France.1252
## tz      Europe/Paris
## date    2021-02-24
##
## - Packages -----
## package      * version date      lib source
## assertthat    0.2.1   2019-03-21 [1] CRAN (R 4.0.3)
```

```
## cachem      1.0.1    2021-01-21 [1] CRAN (R 4.0.3)
## callr       3.5.1    2020-10-13 [1] CRAN (R 4.0.3)
## cli         2.2.0    2020-11-20 [1] CRAN (R 4.0.3)
## crayon      1.3.4    2017-09-16 [1] CRAN (R 4.0.3)
## desc        1.2.0    2018-05-01 [1] CRAN (R 4.0.3)
## devtools    2.3.2    2020-09-18 [1] CRAN (R 4.0.3)
## digest      0.6.27   2020-10-24 [1] CRAN (R 4.0.3)
## ellipsis    0.3.1    2020-05-15 [1] CRAN (R 4.0.3)
## evaluate    0.14     2019-05-28 [1] CRAN (R 4.0.3)
## fansi       0.4.1    2020-01-08 [1] CRAN (R 4.0.3)
## fastmap     1.0.1    2019-10-08 [1] CRAN (R 4.0.3)
## fs          1.5.0    2020-07-31 [1] CRAN (R 4.0.3)
## glue        1.4.2    2020-08-27 [1] CRAN (R 4.0.3)
## htmltools   0.5.0    2020-06-16 [1] CRAN (R 4.0.3)
## knitr       1.30     2020-09-22 [1] CRAN (R 4.0.3)
## lifecycle   0.2.0    2020-03-06 [1] CRAN (R 4.0.3)
## magrittr    2.0.1    2020-11-17 [1] CRAN (R 4.0.3)
## memoise     2.0.0    2021-01-26 [1] CRAN (R 4.0.3)
## pkgbuild    1.2.0    2020-12-15 [1] CRAN (R 4.0.3)
## pkgload     1.1.0    2020-05-29 [1] CRAN (R 4.0.3)
## prettyunits 1.1.1    2020-01-24 [1] CRAN (R 4.0.3)
## processx    3.4.5    2020-11-30 [1] CRAN (R 4.0.3)
## ps          1.5.0    2020-12-05 [1] CRAN (R 4.0.3)
## purrr       0.3.4    2020-04-17 [1] CRAN (R 4.0.3)
## R6          2.5.0    2020-10-28 [1] CRAN (R 4.0.3)
## remotes     2.2.0    2020-07-21 [1] CRAN (R 4.0.3)
## rlang       0.4.10   2020-12-30 [1] CRAN (R 4.0.3)
## rmarkdown   2.6      2020-12-14 [1] CRAN (R 4.0.3)
## rprojroot   2.0.2    2020-11-15 [1] CRAN (R 4.0.3)
## sessioninfo 1.1.1    2018-11-05 [1] CRAN (R 4.0.3)
## stringi     1.5.3    2020-09-09 [1] CRAN (R 4.0.3)
## stringr     1.4.0    2019-02-10 [1] CRAN (R 4.0.3)
## testthat    3.0.1    2020-12-17 [1] CRAN (R 4.0.3)
## usethis     2.0.0    2020-12-10 [1] CRAN (R 4.0.3)
## withr       2.4.1    2021-01-26 [1] CRAN (R 4.0.3)
## xfun        0.20     2021-01-06 [1] CRAN (R 4.0.3)
## yaml        2.2.1    2020-02-01 [1] CRAN (R 4.0.3)
##
## [1] D:/Documents/R/win-library/4.0
## [2] C:/Program Files/R/R-4.0.3/library
```

Importer et vérifier les données

Aller récupérer les données sur le dépôt Gitlab

```
data_url="https://gitlab.inria.fr/learninglab/mooc-rr/mooc-rr-ressources/-/raw/master/
module3/Practical_session/Subject6_smoking.csv"
data=read.csv(data_url)
library(knitr)
kable(head(data))
```

Smoker	Status	Age
Yes	Alive	21.0
Yes	Alive	19.3
No	Dead	57.5
No	Alive	47.1
Yes	Alive	81.4
No	Alive	36.8

On cherche si il y a des données manquantes dans le fichier

```
lignes_na=apply(data,1,function(x) any(is.na(x)))
data[lignes_na,]
```

```
## [1] Smoker Status Age
## <0 rows> (or 0-length row.names)
```

Réponses aux exercices

1.Exercice 1

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

```
tab_exo1=as.data.frame(table(data$Smoker,data$Status))
colnames(tab_exo1)=c("Smoker","Status","Number_smoker_status")
library(knitr)
kable(tab_exo1)
```

Smoker	Status	Number_smoker_status
No	Alive	502
Yes	Alive	443
No	Dead	230
Yes	Dead	139

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)

- On récupère les valeurs du nombres total de fumeuses et non fumeuses

```
tab_smoker_exo1=as.data.frame(table(data$Smoker))
colnames(tab_smoker_exo1)=c("Smoker","Number_smoker")
library(knitr)
kable(tab_smoker_exo1)
```

Smoker	Number_smoker
No	732
Yes	582

- On ajoute ces valeurs au tableau précédent

```
tab_exo1=merge(tab_exo1,tab_smoker_exo1, by="Smoker", all.x=T)
library(knitr)
kable(tab_exo1)
```

Smoker	Status	Number_smoker_status	Number_smoker
No	Alive	502	732
No	Dead	230	732
Yes	Alive	443	582
Yes	Dead	139	582

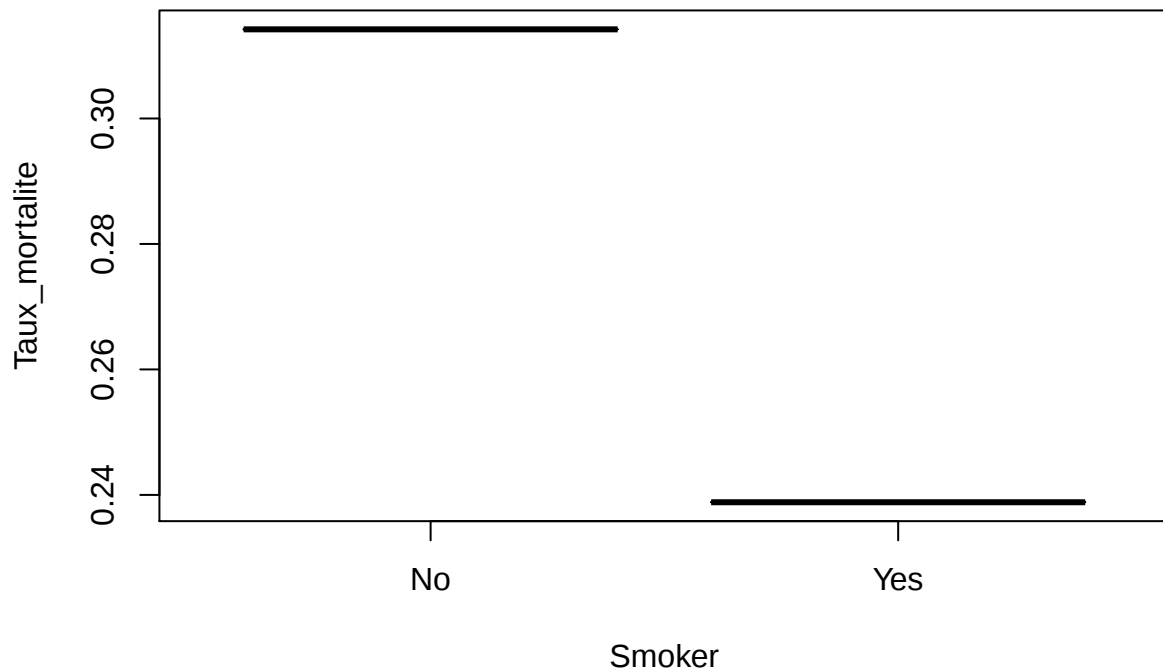
- On calcul le taux de mortalité (le rapport entre le nombre de femmes décédées dans un groupe et le nombre total de femmes dans ce groupe).

```
tab_exo1=subset(tab_exo1, Status=="Dead")
tab_exo1$Number_smoker_status=as.numeric(as.character(tab_exo1$Number_smoker_status))
tab_exo1$Number_smoker=as.numeric(as.character(tab_exo1$Number_smoker))
tab_exo1$Taux_mortalite= (tab_exo1$Number_smoker_status /tab_exo1$Number_smoker)
library(knitr)
kable(tab_exo1)
```

	Smoker	Status	Number_smoker_status	Number_smoker	Taux_mortalite
2	No	Dead	230	732	0.3142077
4	Yes	Dead	139	582	0.2388316

Vous pourrez proposer une représentation graphique de ces données et calculer des intervalles de confiance si vous le souhaitez.

```
plot(Taux_mortalite ~ Smoker, data=tab_exo1)
```

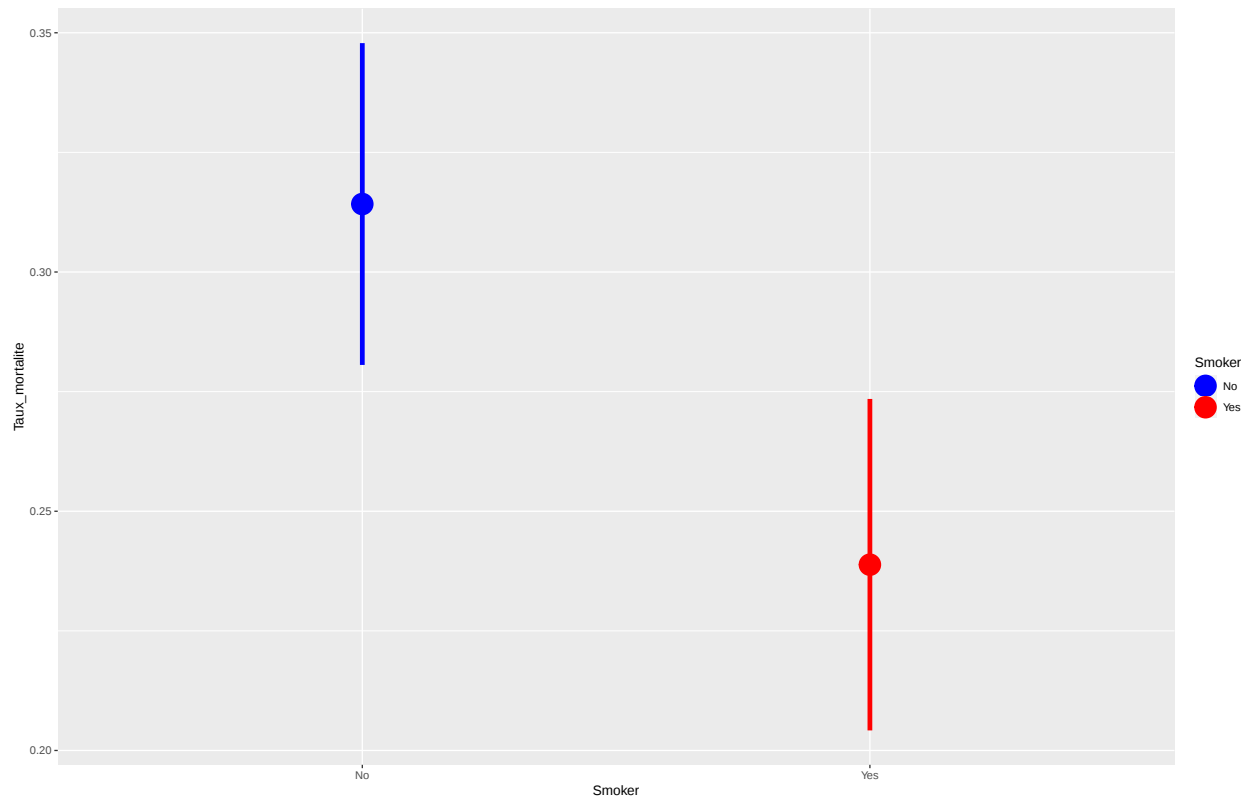


```
tab_exo1$ICdown=tab_exo1$Taux_mortalite-1.96*sqrt(tab_exo1$Taux_mortalite*
(1-tab_exo1$Taux_mortalite)/tab_exo1$Number_smoker)
tab_exo1$ICup=tab_exo1$Taux_mortalite+1.96*sqrt(tab_exo1$Taux_mortalite*
(1-tab_exo1$Taux_mortalite)/tab_exo1$Number_smoker)
library(knitr)
kable(tab_exo1)
```

	Smoker	Status	Number_smoker_status	Number_smoker	Taux_mortalite	ICdown	ICup
2	No	Dead	230	732	0.3142077	0.2805793	0.3478360
4	Yes	Dead	139	582	0.2388316	0.2041914	0.2734719

On a donc un taux de mortalité pour les fumeuses de 0.24 avec un IC [0.20;0.27] et pour les non fumeuse 0.31 avec un IC [0.28;0.35]

```
library(ggplot2)
library(ggthemes)
ggplot(tab_exo1, aes(x=Smoker, y=Taux_mortalite, group=Smoker, color=Smoker)) +
  geom_pointrange(aes(ymin=ICdown, ymax=ICup), lwd=1.8) +
  scale_color_manual(values=c("blue", "red"))
```



En quoi ce résultat est-il surprenant? On a un taux de mortalité plus important pour les femmes non fumeuses.

2.Exercice 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. Dans la consigne le 34 est pris 2 fois, j'ai fais le choix de ne le prendre que dans la première classe

- On définit les différentes classes d'âge

```
data$classe_age=NA
data$classe_age[data$Age>=18 & data$Age<35]="18-34"
data$classe_age[data$Age>=35 & data$Age<54]="35-54"
data$classe_age[data$Age>=55 & data$Age<64]="55-64"
data$classe_age[data$Age>=65 ]="65 et +"
data$classe_age=as.factor(data$classe_age)
levels(data$classe_age)
```

```
## [1] "18-34" "35-54" "55-64" "65 et +"
```

- On représente dans un tableau le nombre total de femmes vivantes et décédées sur la période en fonction de leur habitude de tabagisme et de leur classe d'âge

```
tab_exo2=as.data.frame(table(data$Smoker,data$Status,data$classe_age))
colnames(tab_exo2)=c("Smoker","Status","Class_Age","Number_smoker_status_classage")
library(knitr)
kable(tab_exo2)
```

Smoker	Status	Class_Age	Number_smoker_status_classage
No	Alive	18-34	221
Yes	Alive	18-34	182
No	Dead	18-34	6
Yes	Dead	18-34	7
No	Alive	35-54	172
Yes	Alive	35-54	190
No	Dead	35-54	19
Yes	Dead	35-54	39
No	Alive	55-64	80
Yes	Alive	55-64	64
No	Dead	55-64	39
Yes	Dead	55-64	51
No	Alive	65 et +	28
Yes	Alive	65 et +	7
No	Dead	65 et +	165
Yes	Dead	65 et +	42

- On récupère les valeurs du nombre total de fumeuses et non fumeuses par classe d'âge

```
tab_smoker_classage_exo2=as.data.frame(table(data$Smoker,data$classe_age))
colnames(tab_smoker_classage_exo2)=c("Smoker","Class_Age","Number_smoker_classage")
library(knitr)
kable(tab_smoker_classage_exo2)
```

Smoker	Class_Age	Number_smoker_classage
No	18-34	227
Yes	18-34	189
No	35-54	191
Yes	35-54	229
No	55-64	119
Yes	55-64	115
No	65 et +	193
Yes	65 et +	49

- On ajoute ces valeurs au tableau précédent

```
tab_exo2=merge(tab_exo2,tab_smoker_classage_exo2, by=c("Smoker", "Class_Age"), all.x=T)
library(knitr)
kable(tab_exo2)
```

Smoker	Class_Age	Status	Number_smoker_status_classage	Number_smoker_classage
No	18-34	Alive	221	227
No	18-34	Dead	6	227
No	35-54	Alive	172	191
No	35-54	Dead	19	191
No	55-64	Alive	80	119
No	55-64	Dead	39	119
No	65 et +	Alive	28	193
No	65 et +	Dead	165	193
Yes	18-34	Alive	182	189
Yes	18-34	Dead	7	189
Yes	35-54	Alive	190	229
Yes	35-54	Dead	39	229
Yes	55-64	Alive	64	115
Yes	55-64	Dead	51	115
Yes	65 et +	Alive	7	49
Yes	65 et +	Dead	42	49

- On calcul le taux de mortalité (le rapport entre le nombre de femmes décédées dans un groupe et le nombre total de femmes dans ce groupe).

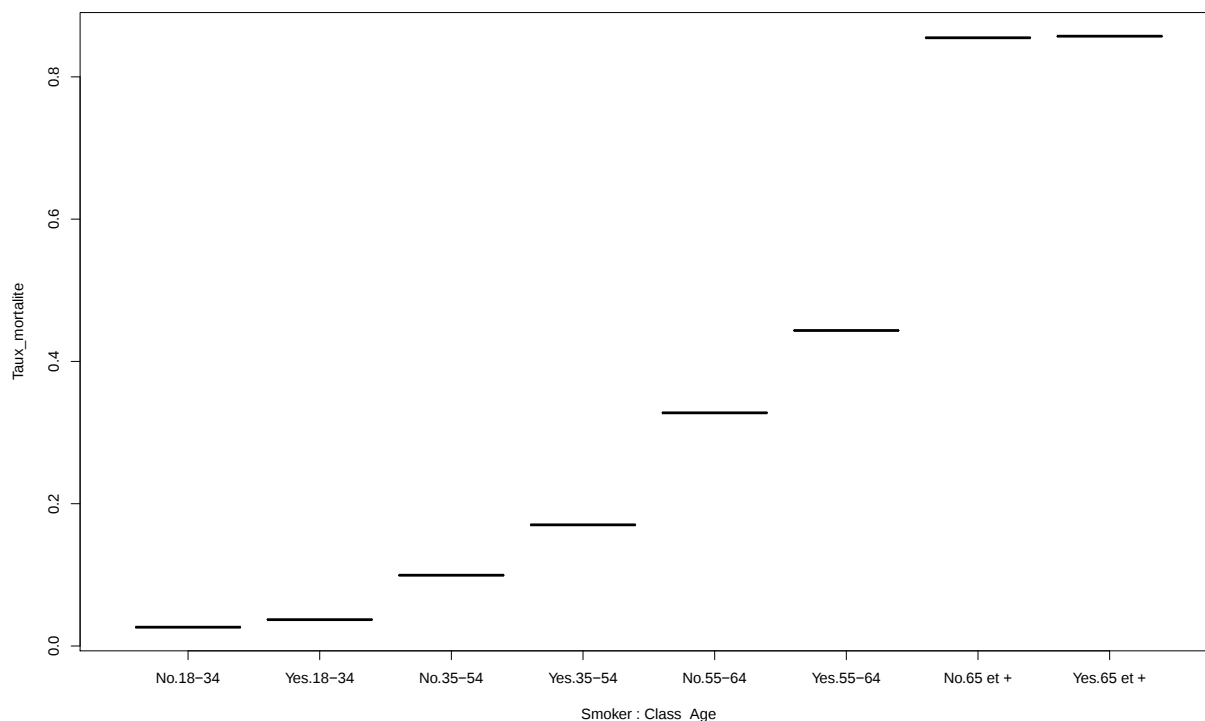
```
tab_exo2=subset(tab_exo2, Status=="Dead")
tab_exo2$Number_smoker_status_classage=as.numeric(as.character
(tab_exo2$Number_smoker_status_classage))
tab_exo2$Number_smoker_classage=as.numeric(as.character
(tab_exo2$Number_smoker_classage))
tab_exo2$Taux_mortalite= tab_exo2$Number_smoker_status_classage /
tab_exo2$Number_smoker_classage
library(knitr)
kable(tab_exo2)
```

	Smoker	Class_Age	Status	Number_smoker_status_classage	Number_smoker_classage	Taux_mortalite
2	No	18-34	Dead	6	227	0.0264317
4	No	35-54	Dead	19	191	0.0994764
6	No	55-64	Dead	39	119	0.3277311
8	No	65 et +	Dead	165	193	0.8549223

	Smoker	Class_Age	Status	Number_smoker_status_classage	Number_smoker_classage	Taux_mortalite
10	Yes	18-34	Dead	7	189	0.0370370
12	Yes	35-54	Dead	39	229	0.1703057
14	Yes	55-64	Dead	51	115	0.4434783
16	Yes	65 et +	Dead	42	49	0.8571429

Vous pourrez proposer une représentation graphique de ces données et calculer des intervalles de confiance si vous le souhaitez.

```
boxplot(Taux_mortalite ~ Smoker*Class_Age, data=tab_exo2)
```



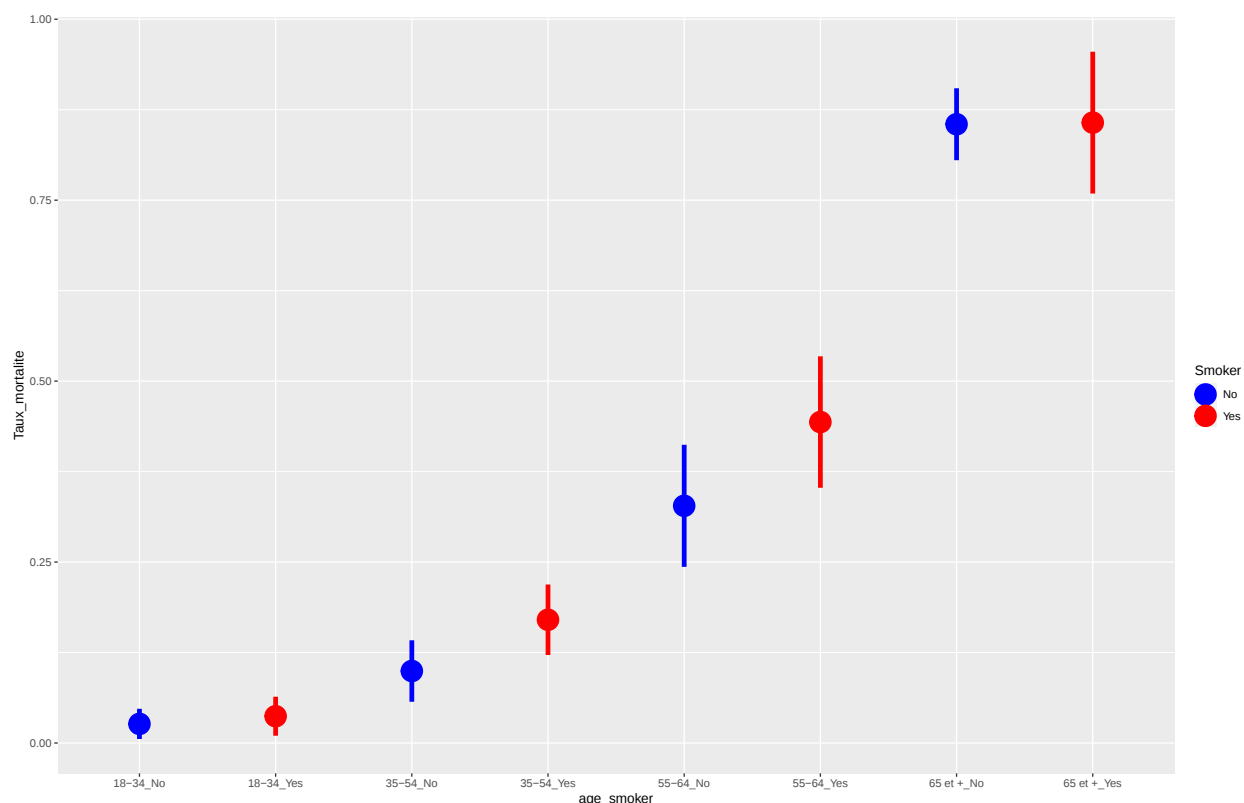
```
tab_exo2$ICdown=tab_exo2$Taux_mortalite-1.96*sqrt(tab_exo2$Taux_mortalite*
(1-tab_exo2$Taux_mortalite)/tab_exo2$Number_smoker_classage)
tab_exo2$ICup=tab_exo2$Taux_mortalite+1.96*sqrt(tab_exo2$Taux_mortalite*
(1-tab_exo2$Taux_mortalite)/tab_exo2$Number_smoker_classage)
library(knitr)
kable(tab_exo2)
```

	Smoker	Class_Age	Status	Number_smoker_status_classage	Number_smoker_classage	Taux_mortalite	ICdown	ICup
2	No	18-34	Dead	6	227	0.0264317	0.0055633	0.0473001
4	No	35-54	Dead	19	191	0.0994764	0.0570294	0.1419234

	Smoker	Class_Age	Status	Number_smoker_status	Number_smoker_class	Taux_mortalite	ICdown	ICup
6	No	55-64	Dead	39	119	0.3277311	0.2433951	0.4120671
8	No	65 et +	Dead	165	193	0.8549223	0.8052355	0.9046091
10	Yes	18-34	Dead	7	189	0.0370370	0.0101125	0.0639616
12	Yes	35-54	Dead	39	229	0.1703057	0.1216188	0.2189926
14	Yes	55-64	Dead	51	115	0.4434783	0.3526786	0.5342780
16	Yes	65 et +	Dead	42	49	0.8571429	0.7591633	0.9551224

En quoi ce résultat est-il surprenant? Dans ce cas là, on voit que l'on a une tendance à plus fort taux de mortalité pour les femmes fumeuses dans la classe d'âge 2 et une tendance similaire pour les femmes de la classe d'âge 1. Cependant les interval de confiance sont chevauchant

```
tab_exo2$age_smoker=as.factor(paste(tab_exo2$Class_Age,tab_exo2$Smoker, sep="_"))
library(ggplot2)
library(ggthemes)
ggplot(tab_exo2, aes(x=age_smoker, y=Taux_mortalite, group=Smoker, color=Smoker)) +
  geom_pointrange(aes(ymin=ICdown, ymax=ICup), lwd=1.8) +
  scale_color_manual(values=c("blue", "red"))
```



Arrivez-vous à expliquer ce paradoxe? De même, vous pourrez proposer une représentation graphique de ces données pour étayer vos explications.

On peut voir qu'il y a beaucoup plus de femmes de la classe d'âge 3 dans les non fumeuses, donc une mortalité qui semble plus importante pour les non fumeuse car cette classe d'âge et celle ou il y a un plus fort taux de mortalité

```

tab_smoker_classage_exo2=as.data.frame(table(data$Smoker,data$classe_age))
colnames(tab_smoker_classage_exo2)=c("Smoker","Class_Age","Number_smoker_classage")
library(knitr)
kable(tab_smoker_classage_exo2)

```

Smoker	Class_Age	Number_smoker_classage
No	18-34	227
Yes	18-34	189
No	35-54	191
Yes	35-54	229
No	55-64	119
Yes	55-64	115
No	65 et +	193
Yes	65 et +	49

3.Exercice 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 $\text{Death} \sim \text{Age}$ 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 une variable Death valant 1 ou 0 pour indiquer si l'individu est décédé durant la période de 20 ans

```

data$Death=NA
data$Death[data$Status=="Alive"]=0
data$Death[data$Status=="Dead"]=1
library(knitr)
kable(head(data))

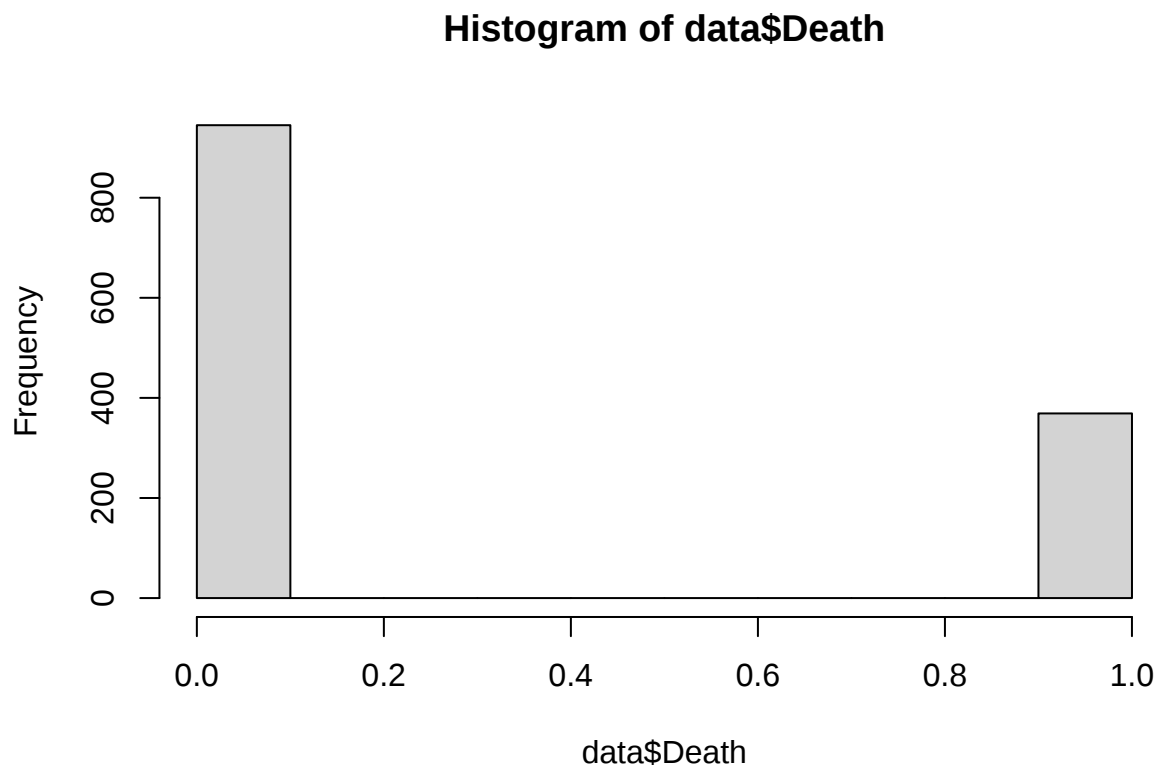
```

Smoker	Status	Age	classe_age	Death
Yes	Alive	21.0	18-34	0
Yes	Alive	19.3	18-34	0
No	Dead	57.5	55-64	1
No	Alive	47.1	35-54	0
Yes	Alive	81.4	65 et +	0
No	Alive	36.8	35-54	0

- On étudie le modèle $\text{Death} \sim \text{Age}$ pour étudier la probabilité de décès en fonction de l'âge selon que l'on considère le groupe des fumeuses ou des non fumeuses.

Pour cela on utilise un modèle linéaire généralisé avec une distribution binomiale, comprenant l'interaction des effets Age et Smoker

```
hist(data$Death)
```



```
m1=glm(Death ~ Age * Smoker, data=data, family=binomial(logit))
summary(m1)
```

```
##
## Call:
## glm(formula = Death ~ Age * Smoker, family = binomial(logit),
##      data = data)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -2.4019  -0.6010  -0.2854   0.4339   3.0457
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -6.79507    0.479341 -14.177  <2e-16 ***
## Age           0.107275    0.007805  13.745  <2e-16 ***
## SmokerYes     1.287401    0.668678   1.925   0.0542 .
##
```

```
## Age:SmokerYes -0.018299  0.011703 -1.564  0.1179
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
## Null deviance: 1560.32 on 1313 degrees of freedom
## Residual deviance: 999.49 on 1310 degrees of freedom
## AIC: 1007.5
##
## Number of Fisher Scoring iterations: 5
```

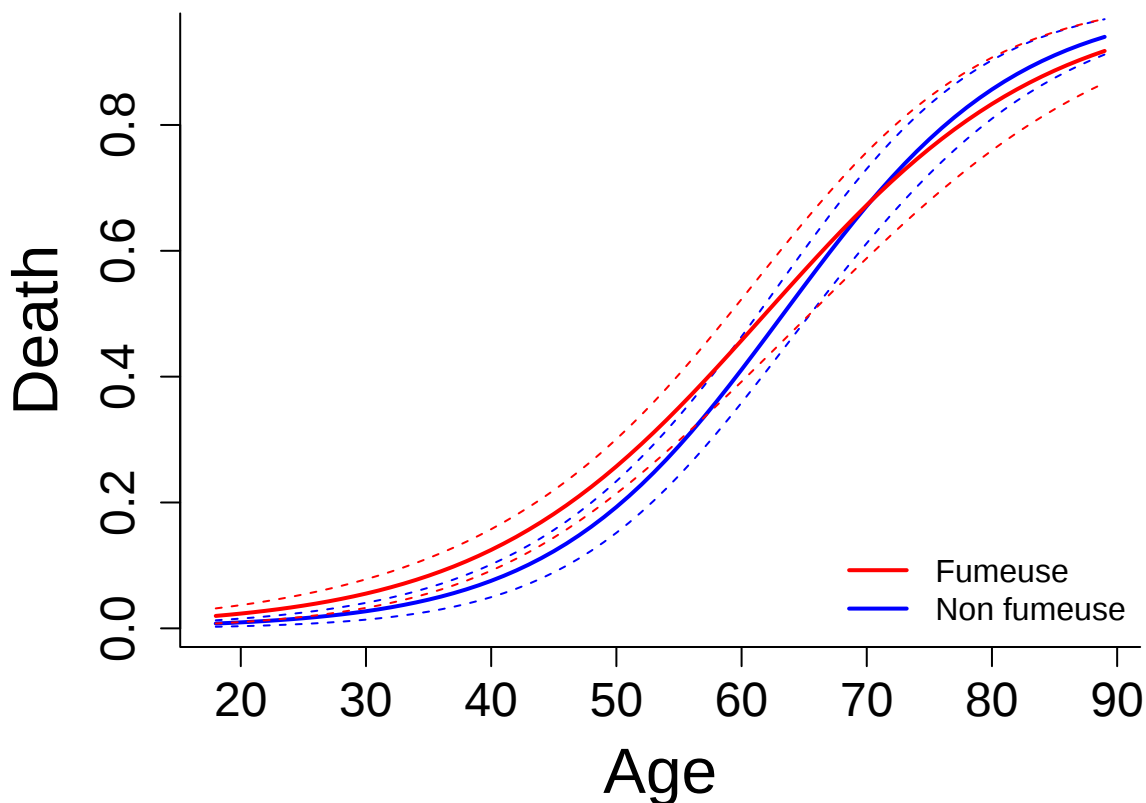
- On plot les prédictions de ce modèle en indiquant les intervals de confiance à 95%

```
data$Smoker=as.factor(data$Smoker)
grid = expand.grid( Smoker=levels(data$Smoker),Age=seq(min(data$Age), max(data$Age),1))
library(AICcmodavg)
pred <- predict(m1,grid, se.fit=T,type="response")
grid$y=pred$fit
grid$low=pred$fit-1.96*pred$se.fit
grid$up=pred$fit+1.96*pred$se.fit
head(grid)
```

```
##   Smoker Age      y      low      up
## 1     No  18 0.007656220 0.002530764 0.01278168
## 2     Yes  18 0.019713741 0.007729517 0.03169796
## 3     No  19 0.008515833 0.002941982 0.01408968
## 4     Yes  19 0.021508746 0.008791997 0.03422549
## 5     No  20 0.009471039 0.003413296 0.01552878
## 6     Yes  20 0.023463280 0.009981550 0.03694501
```

```
Nonfumeuse=subset(grid, Smoker=="No")
Fumeuse=subset(grid, Smoker=="Yes")
par(mar=c(5,5.5,1,2))
plot(y~Age,xlab="Age",ylab="Death" ,
     data=Nonfumeuse,axes=F, cex.lab=2, col="blue",type="l",lwd=2)
lines(low~Age, data=Nonfumeuse, col="blue",lwd=1,lty=2)
lines(up~Age, data=Nonfumeuse, col="blue",lwd=1,lty=2)

lines(y~Age, data=Fumeuse, col="red",lwd=2)
lines(low~Age, data=Fumeuse, col="red",lwd=1,lty=2)
lines(up~Age, data=Fumeuse, col="red",lwd=1,lty=2)
box(bty="l")
axis(1,cex.axis=1.5)
axis(2,cex.axis=1.5)
legend("bottomright",bty="n", legend=c("Fumeuse","Non fumeuse"),
     col=c("red","blue"),lwd=2,cex=1)
```



Ces régressions vous permettent-elles de conclure sur la nocivité du tabagisme ?

Non, on voit que d'une manière générale la probabilité de mourir augmente avec l'âge, mais on trouve de faibles différences entre les courbes des fumeuses et des non fumeuses, avec des intervalles de confiance très chevauchant bien qu'il y ait une tendance à ce que les femmes dans la tranche d'âge autour de 40 ans qui sont fumeuses aient un taux de mortalité légèrement supérieur aux femmes non fumeuses

Pour confirmer cela on peut faire une sélection de modèle sur le critère AICc, en comparant le modèle avec l'effet smoker et celui sans:

```
m0=glm(Death ~ Age , data=data, family=binomial)
m1=glm(Death ~ Age + Smoker, data=data, family=binomial)
m2=glm(Death ~ Age * Smoker, data=data, family=binomial)
library(AICcmodavg)
Cands <- list(m0,m1,m2)
Model.names <- c("Age","Age+smoker","Age*smoker")
tab_aic2=aictab(cand.set = Cands, modnames = Model.names, sort = TRUE, second.ord = T)
library(knitr)
kable(tab_aic2)
```

	Modnames	K	AICc	Delta_AICc	ModelLik	AICcWt	LL	Cum.Wt
3	Age*smoker	4	1007.520	0.0000000	1.0000000	0.4264440	-499.7448	0.4264440
2	Age+smoker	3	1007.926	0.4060959	0.8162391	0.3480802	-500.9540	0.7745242
1	Age	2	1008.795	1.2745365	0.5287348	0.2254758	-502.3928	1.0000000

On voit bien que malgré que le modèle avec les effet Age x smoker soit celui avec un plus faible AIC, il y a un faible soutien au modèle comprenant l'interaction Age*smoker, avec un delta AICc < 2 par rapport au modèle uniquement avec Age, il est donc difficile de conclure sur la nocivité du tabagisme. Si l'on utilise le principe de parcimonie on retiendrait le modèle le plus simple avec un delta AICc < 2 , donc le modèle uniquement avec âge dans ce cas là.