Exercice paradoxe de Simpson

Hugues de Courson

La date du jour

Début: chargement des packages pouvant être utiles

```
library(epiDisplay)
library(epiR)
library(prettyR)
library(knitr)
library(kableExtra)
```

Puis l'import des données

Dans un premier temps, il s'agissait de télécharger le fichier dans mon dossier local, ensuite on commit cette action et avec un push il est envoyé sur gitlab.

Ensuite il s'agit de le charger dans R :

```
data <- read.csv("Subject6_smoking.csv")
View(data)
```

Vérification du format des données :

```
str(data)
```

```
## 'data.frame':    1314 obs. of  3 variables:
##  $ Smoker: Factor w/ 2 levels "No","Yes": 2 2 1 1 2 1 1 2 2 2 ...
##  $ Status: Factor w/ 2 levels "Alive","Dead": 1 1 2 1 1 1 1 2 1 1 ...
##  $ Age   : num  21 19.3 57.5 47.1 81.4 36.8 23.8 57.5 24.8 49.5 ...
```

Tout semble ok.

Mission 1

Creation du tableau :

L'objectif est de créer un tableau avec le nombre total de femme vivantes ou décédées sur la période en fonction des habitudes de tabagisme.

```
vivantes_fum <- sum(data$Status[data$Smoker=="Yes"]=="Alive")
vivantes_nonfum <- sum(data$Status[data$Smoker=="No"]=="Alive")

decedees_fum <- sum(data$Status[data$Smoker=="Yes"]=="Dead")
decedees_nonfum  <- sum(data$Status[data$Smoker=="No"]=="Dead")
tab_1 <- data.frame(Vivantes = c(vivantes_fum,vivantes_nonfum),
  Decedees = c(decedees_fum,decedees_nonfum), row.names = c("Fumeuses","Non fumeuses"))
kable(tab_1,align = 'l')
```

	Vivantes	Decedees
Fumeuses	443	139
Non fumeuses	502	230

Calcul des taux de mortalité :

Creation d'un tableau de contigence

```
mort_fum <- sum(data$Status[data$Smoker=="Yes"]=="Dead")
mort_nonfum <- sum(data$Status[data$Smoker=="No"]=="Dead")
vivant_fum <-sum(data$Status[data$Smoker=="Yes"]=="Alive")
vivant_nonfum <- sum(data$Status[data$Smoker=="No"]=="Alive")
nb_fum <- sum(data$Smoker=="Yes")
nb_nonfum <- sum(data$Smoker=="No")
nb_viv <- sum(data$Status=="Alive")
nb_deces <-sum(data$Status=="Dead")
tot <- sum(data$Status=="Alive"|data$Status=="Dead")

tab_2 <- data.frame(cbind(rbind(mort_fum, vivant_fum, nb_fum)),
                    rbind(mort_nonfum,vivant_nonfum, nb_nonfum),
                    rbind(nb_deces,nb_viv,tot))

row.names(tab_2) <- c("Decedees", "Vivantes","Total")
names(tab_2) <- c("Fumeuses", "Non Fumeuses", "Total")

kable(tab_2,align = "c")
```

	Fumeuses	Non Fumeuses	Total
Decedees	139	230	369
Vivantes	443	502	945
Total	582	732	1314

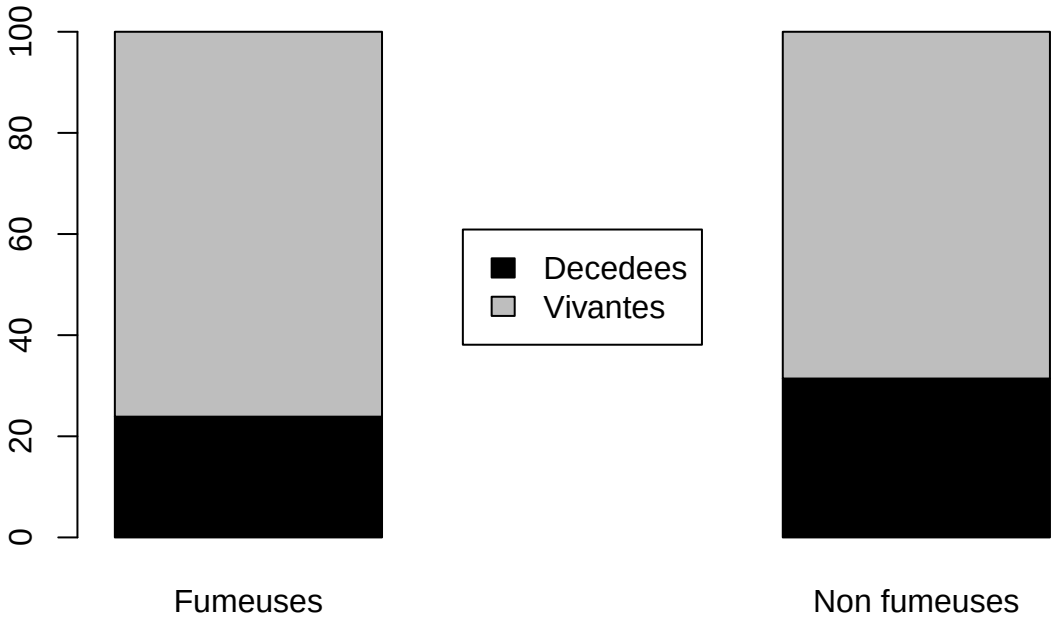
Representation graphique

Ceci passe par la création d'un tableau avec les pourcentages de mortalité

```
tab_3 <- data.frame(cbind(rbind((sum(data$Status[data$Smoker=="Yes"]=="Dead")/sum(data$Smoker=="Yes"))*
                                  (sum(data$Status[data$Smoker=="Yes"]=="Alive")/sum(data$Smoker=="Yes"))
                    rbind((sum(data$Status[data$Smoker=="No"]=="Dead")/sum(data$Smoker=="No"))*100,
                          (sum(data$Status[data$Smoker=="No"]=="Alive")/sum(data$Smoker=="No"))*100))

names(tab_3) <- c("Fumeuses", "Non fumeuses")
row.names(tab_3) <- c("Decedees", "Vivantes")

kable(tab_3, align = "c")
```

	Fumeuses	Non fumeuses
Decedees	23.88316	31.42076
Vivantes	76.11684	68.57923

```
barplot(as.matrix(tab_3), col = c("black","gray"), space = 1.5,width = 0.5)
legend("center",xpd=NA, legend = c("Decedees", "Vivantes"), fill = c("black","gray"))
```



Calcul des intervalles de confiance des proportions

```
#  creation de la fonction :

IC = function(x) {
  y <- x/100
  inf = y-(1.96*sqrt((y*(1-y))/n))
  sup = y+(1.96*sqrt((y*(1-y))/n))
  print(c(x,inf*100,sup*100))
}
```

Chez les fumeuses :

```
x <- tab_3[1,1]
n <- as.numeric(tab_2[3,1])

IC(x)
```

```
## [1] 23.88316 20.41914 27.34719
```

Chez les non fumeuses :

```
x <- tab_3[1,2]
n <- tab_2[3,2]

IC(x)
```

```
## [1] 31.42077 28.05793 34.78360
```

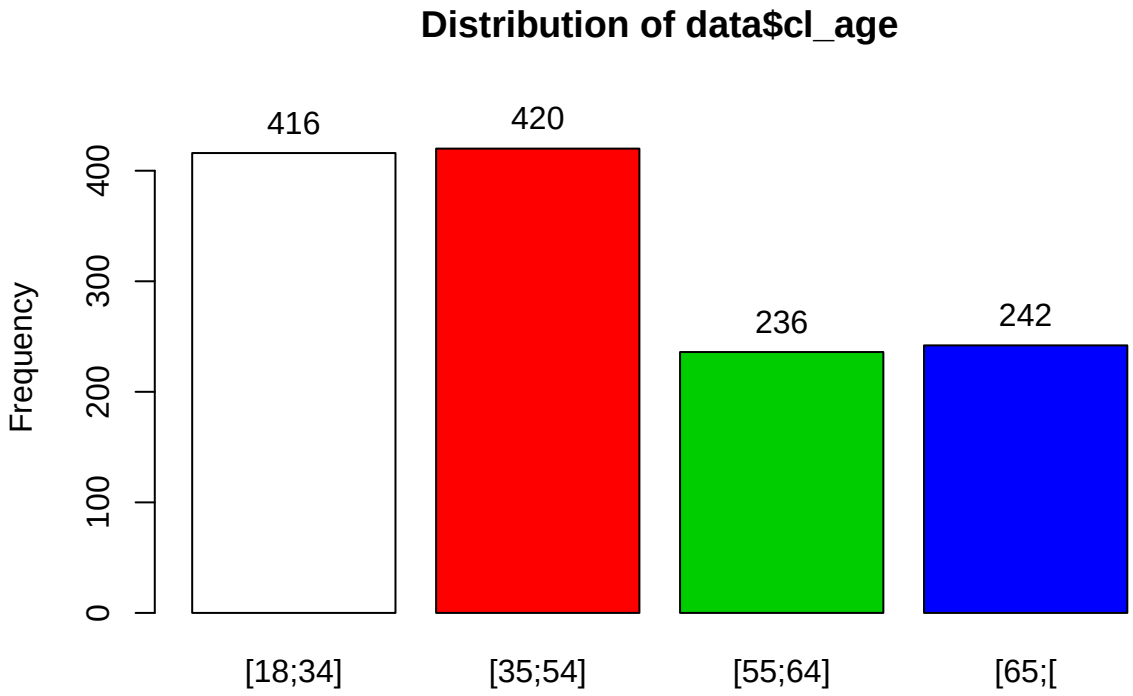
Au vu de l'ensemble des résultats, on fait le constat suivant : Il semblerait que le taux de mortalité est plus important chez les non-fumeuses... Ceci va à l'encontre des connaissances sur le tabac, on pourrait s'attendre à ce que les non fumeuses décèdent moins.

Mission 2

Il s'agit de reprendre les données, cette fois-ci en fonction de la classe d'âge

Première étape : création de la variable classe d'âge

```
data$cl_age <- as.factor(ifelse(data$Age>=18&data$Age<35,"[18;34]",
                                ifelse(data$Age>=35&data$Age<54,"[35;54]",
                                       ifelse(data$Age>=55&data$Age<65,"[55;64]","[65;["))))

tab1(data$cl_age)
```



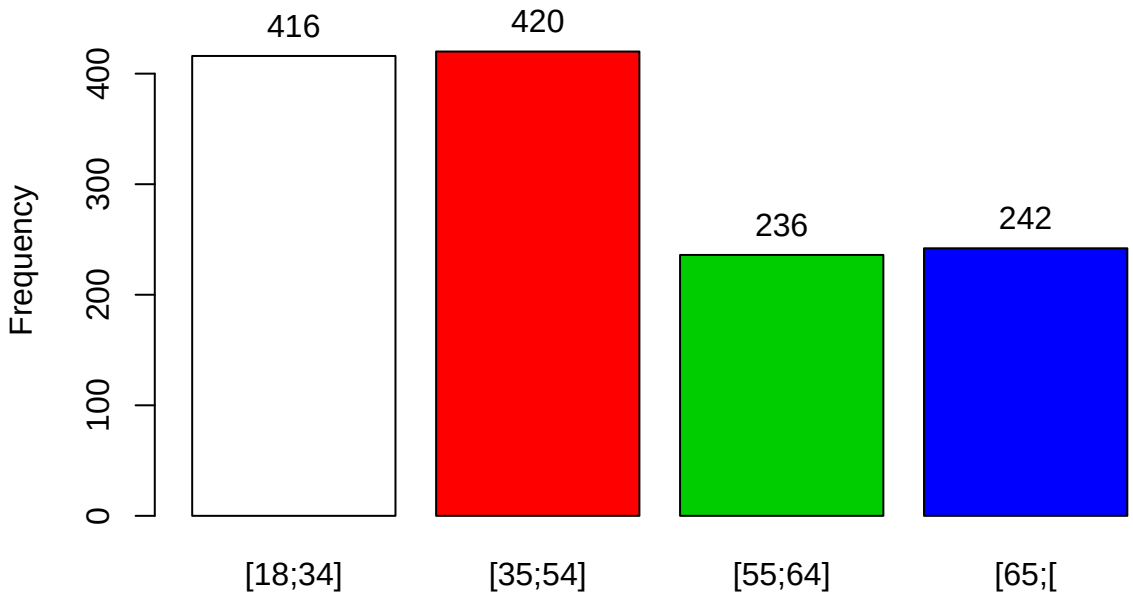

```
## data$cl_age :
##          Frequency Percent Cum. percent
## [18;34]        416    31.7         31.7
## [35;54]        420    32.0         63.6
## [55;64]        236    18.0         81.6
## [65;[          242    18.4        100.0
##   Total       1314   100.0        100.0
```

Deuxième étape on refait les tableaux de contingence mais cette fois-ci de manière plus optimisée

Pour cela on utilise ce lien pour connaître la méthode. Et ce lien pour la réalisation de beaux tableaux.

```
tab_5 <- ftable(data[,c(1,2,4)])
tab_5 <- round(prop.table(tab_5,2)*100,1)
tab_5 <- rbind(tab_5, tab1(data$cl_age)$output.table[,1])
```

```
## Warning in rbind(tab_5, tab1(data$cl_age)$output.table[, 1]): number of columns
## of result is not a multiple of vector length (arg 2)
```



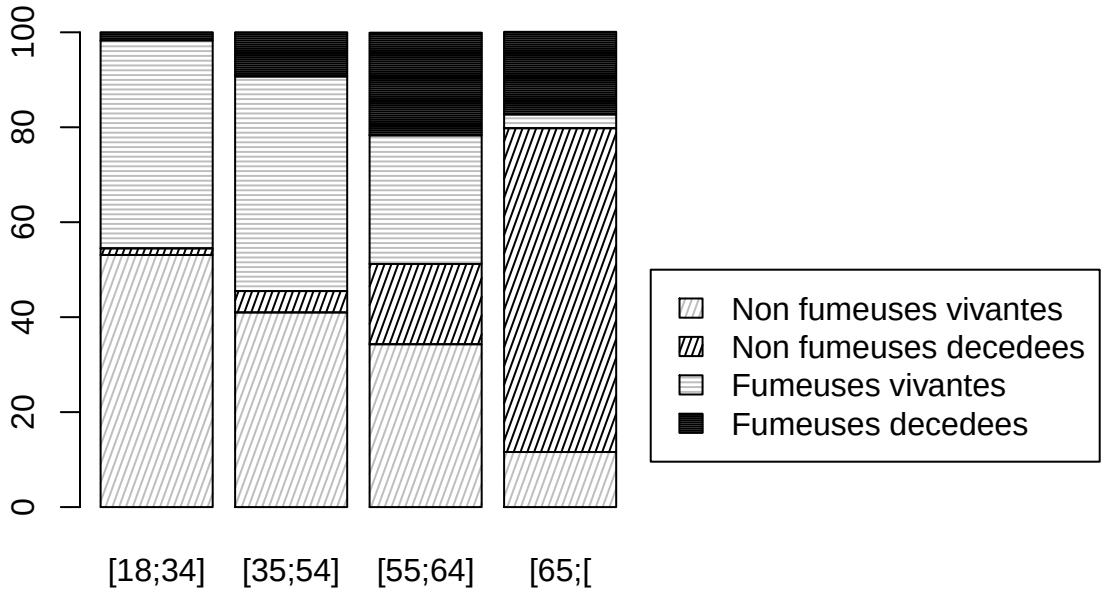

```
tab_5 <- as.table(tab_5)
row.names(tab_5) <- c("Non fumeuses vivantes ", "Non fumeuses decedees",
                      "Fumeuses vivantes", "Fumeuses decedees" ,"Effectifs")
colnames(tab_5) <- c("[18;34]", "[35;54]", "[55;64]", "[65;[")

kable(tab_5,"latex", align = "c") %>% kable_styling(latex_options = "striped",
                                                    stripe_index = c(1,2))
```

	[18;34]	[35;54]	[55;64]	[65;[
Non fumeuses vivantes	53.1	41.0	34.3	11.6
Non fumeuses decedees	1.4	4.5	16.9	68.2
Fumeuses vivantes	43.8	45.2	27.1	2.9
Fumeuses decedees	1.7	9.3	21.6	17.4
Effectifs	416.0	420.0	236.0	242.0

Et maintenant on trace les barplots

Super lien pour les paramètres graphiques. Et autre lien pour les diagrammes en barre.

```
par(mar = c(5,5,5,13))
barplot(tab_5[c(1:4),], col = c("gray","black","gray","black"), density = c(30,30,40,100),
        angle = c(70,70,0,0))
xmin <- par("usr")[1]
xmax <- par("usr")[2]
ymin <- par("usr")[3]
ymax <- par("usr")[4]
par(xpd=TRUE)
lambda <- 0.025
legend(((1 + lambda) * par("usr")[2] - lambda * par("usr")[1]),50,
       legend = c("Non fumeuses vivantes", "Non fumeuses decedees",
                  "Fumeuses vivantes", "Fumeuses decedees"),
       fill = c("gray","black","gray","black"),density = c(30,30,40,100),angle = (c(70,70,0,0)))
```



NB : dans le code précédent :

```
par(mar = c(5,5,5,15))
```

On définit les paramètres de marge, dans l'ordre c(bas,gauche,haut,droite)

```
xmin <- par("usr")[1]
xmax <- par("usr")[2]
ymin <- par("usr")[3]
ymax <- par("usr")[4]
```

On récupère les valeurs min et max du plot précédent

Et maintenant on peut obtenir les intervalles de confiance

Pour le premier groupe d'age non fumeuses decedees

```
x <- tab_5[2,1]
n <- tab_5[5,1]

IC1<- round(IC(x),1)
```

```
## [1] 1.4000000 0.2709534 2.5290466
```

```
IC1 <- paste(c(IC1[1],"[",IC1[2],"-",IC1[3],"]"),collapse = "")
```

Pour le premier groupe d'age fumeuses decedees

```
x <- tab_5[4,1]
n <- tab_5[5,1]

IC2 <- round(IC(x),1)
```

```
## [1] 1.7000000 0.4577454 2.9422546
```

```
IC2 <- paste(c(IC2[1],"[",IC2[2],"-",IC2[3],"]"), collapse = "")
```

Pour le deuxième groupe d'age non fum

```
x <- tab_5[2,2]
n <- tab_5[5,2]

IC3 <- round(IC(x),1)
```

```
## [1] 4.500000 2.517381 6.482619
```

```
IC3 <- paste(c(IC3[1],"[",IC3[2],"-",IC3[3],"]"),collapse = "")
```

Pour le deuxième groupe d'age fum

```
x <- tab_5[4,2]
n <- tab_5[5,2]

IC4 <- round(IC(x),1)
```

```
## [1]  9.300000  6.522356 12.077644
```

```
IC4 <- paste(c(IC4[1],"[",IC4[2],"-",IC4[3],"]"),collapse = "")
```

Pour le troisieme groupe d'age non fum

```
x <- tab_5[2,3]
n <- tab_5[5,3]

IC5 <- round(IC(x),1)
```

```
## [1] 16.90000 12.11872 21.68128
```

```
IC5 <- paste(c(IC5[1],"[",IC5[2],"-",IC5[3],"]"),collapse = "")
```

Pour le troisieme groupe d'age fum

```
x <- tab_5[4,3]
n <- tab_5[5,3]

IC6 <- round(IC(x),1)
```

```
## [1] 21.60000 16.34969 26.85031
```

```
IC6 <- paste(c(IC6[1],"[",IC6[2],"-",IC6[3],"]"),collapse = "")
```

Pour le quatrieme groupe d'age non fum

```
x <- tab_5[2,4]
n <- tab_5[5,4]
```

```
IC7<- round(IC(x),1)
```

```
## [1] 68.20000 62.33249 74.06751
```

```
IC7 <- paste(c(IC7[1],"[",IC7[2],"-",IC7[3],"]"),collapse = "")
```

Pour le quatrieme groupe d'age fum

```
x <- tab_5[4,4]
n <- tab_5[5,4]

IC8<- round(IC(x),1)
```

```
## [1] 17.40000 12.62346 22.17654
```

```
IC8 <- paste(c(IC8[1],"[",IC8[2],"-",IC8[3],"]"), collapse = "")
```

Sous forme d'un tableau :

```
tab_6 <- cbind(rbind(IC1,IC2), rbind(IC3,IC4), rbind(IC5,IC6), rbind(IC7,IC8))
colnames(tab_6) <- colnames(tab_5)
row.names(tab_6) <- c("Non Fumeuses", "Fumeuses")
kable(tab_6,align = "c")
```

	[18;34]	[35;54]	[55;64]	[65;[
Non Fumeuses	1.4[0.3-2.5]	4.5[2.5-6.5]	16.9[12.1-21.7]	68.2[62.3-74.1]
Fumeuses	1.7[0.5-2.9]	9.3[6.5-12.1]	21.6[16.3-26.9]	17.4[12.6-22.2]

Une fois de plus à la vue de ces résultats on est étonnés ! En effet, autant c'est cohérent chez les non fumeurs mais chez les fumeurs il y a une discordance car ce ne sont pas les personnes les plus agées qui décèdent le plus. Il y a probablement une interaction entre l'âge et le fait d'être fumeur, où alors les catégories sont mal faites...

Mission 3

Il s'agit maintenant de faire une régression logistique pour utiliser l'âge en continu et ainsi ne pas perdre d'information.

On commence par bien coder la variable décès

```
data$deces <- as.factor(ifelse(data$Status=="Alive",0,1))
```

Ensuite on fait la regression logistique, d'abord univariable.

Pour le tabac :

```
regunita <- glm(deces~Smoker, family = binomial, data = data)

# pour obtenir les OR

logistic.display(regunita)
```

```
##
## Logistic regression predicting deces : 1 vs 0
##
##                    OR(95%CI)         P(Wald's test) P(LR-test)
```

```
## Smoker: Yes vs No 0.68 (0.54,0.88)  0.003          0.002
##
## Log-likelihood = -775.5584
## No. of observations = 1314
## AIC value = 1555.1167
```

Pour l'âge :

```
reguniage <- glm(deces~Age, family = binomial(), data = data)

logistic.display(reguniage)
```

```
##
## Logistic regression predicting deces : 1 vs 0
##
##                   OR(95%CI)        P(Wald's test) P(LR-test)
## Age (cont. var.) 1.1 (1.09,1.11)  < 0.001        < 0.001
##
## Log-likelihood = -502.3928
## No. of observations = 1314
## AIC value = 1008.7856
```

La subtilité pour l'âge est qu'il faut vérifier l'hypothèse de linéarité du Logit car variable quantitative et hypothèse sous-jacente d'un modèle de régression logistique. Pour cela on introduit un polynome fractionnaire dans le modèle de régression.

```
library(mfp)

reguniage2 <- mfp(deces~fp(Age, df =4, select = 1, scale = T),
family = binomial, data = data)

summary(reguniage2)
```

```
##
## Call:
## glm(formula = deces ~ I((Age/100)^1), family = binomial, data = data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.3335  -0.5897  -0.2848   0.4551   2.8803
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)     -6.1045     0.3214  -18.99   <2e-16 ***
## I((Age/100)^1)   9.7651     0.5555   17.58   <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: 1560.3  on 1313  degrees of freedom
## Residual deviance: 1004.8  on 1312  degrees of freedom
## AIC: 1008.8
##
```

```
## Number of Fisher Scoring iterations: 5
```

En lisant le summary, le meilleur polynome est de degrès 1 donc l'hypothèse de linéarité est respectée.

Modèle multivariable

```
regtot <- glm(deces~Age+Smoker, family = binomial, data = data)

logistic.display(regtot)
```

```
##
## Logistic regression predicting deces : 1 vs 0
##
##                    crude OR(95%CI)   adj. OR(95%CI)    P(Wald's test) P(LR-test)
## Age (cont. var.)  1.1 (1.09,1.11)   1.1 (1.09,1.12)   < 0.001        < 0.001
##
## Smoker: Yes vs No 0.68 (0.54,0.88)  1.32 (0.96,1.83)  0.091          0.09
##
## Log-likelihood = -500.954
## No. of observations = 1314
## AIC value = 1007.908
```

Et là on constate une inversion de l'odds ratio associé au tabac lorsque l'on ajuste sur l'âge... Soit l'âge est un facteur confondant sur la relation entre le fait d'être fumeur et le risque de décés, soit il s'agit d'un modificateur de l'effet.

Dans tous les cas l'interprétation brute du de la relation entre le fait de fumer et l'âge n'est pas juste.