Sujet 6 : Autour du Paradoxe de Simpson

Louis Hognon

19/04/2020

Contents

Préparation des données

Les données **autour du Paradoxe de Simpson** se trouvent sont acessibles via le site de gitlab INRIA

L'URL de ce lien est :

```
data_url= "https://gitlab.inria.fr/learninglab/mooc-rr/mooc-rr-ressources/-/raw/master/module3/Practica
```

Pour **eviter une éventuelle disparition des données du serveur gitlab**, nous faisons une **copie locale de ce jeux de données** que nous préservons Il est inutile et même risquée de télécharger les données à chaque exécution, car dans le cas d'une panne nous pourrions remplacer nos données par un fichier défectueux. Pour cette raison, nous téléchargeons les données seulement si la copie locale n'existe pas.

```
data_file = "simpsons.csv"
if (!file.exists(data_file)) {
    download.file(data_url, data_file, method="auto")
}
```

Téléchargement des données

Nous téléchargons les données via la commande **read.csv**, appliquée à data_url et data_file, dans le cas d'une éventuelle disparition des données du serveur gitlab.

```
data_exo6 = read.csv(data_url)
data_exo6 = read.csv(data_file)
```

Description de la base de donnée

Avant de commencer les analyses sur la base de données, nous allons d'abord réaliser une description de celle-ci. Puis vérifier si certains paramètres : type de données, données manquantes, données abberantes,

Afficher la structure de la base de donnée

Avec la commande **str** nous allons vérifier le type de données dans chaque colonne ainsi que les premières valeurs.

```
str(data_exo6)
```

```
## '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 ...
```

On retrouve bien nos 3 colonnes:

- Smoker : si la personne fume ou non
- Status : si elle est vivante ou décédée au moment de la seconde étude
- Age : son âge lors du premier sondage

Pour rappel : *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.*

Vérifications : données abbérantes

On utilise la commande summary afin de voir s'il y a des données abbérantes dans chacune des colones. On précise la colone à inspecter grâce à la commande $.

```
summary(data_exo6$Smoker)
```

```
##  No Yes
## 732 582
```

```
summary(data_exo6$Status)
```

```
## Alive  Dead
##   945   369
```

```
summary(data_exo6$Age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   18.00   31.30   44.80   47.36   60.60   89.90
```

Il n'y a pas de données abbérantes, cad, des réponses autres que No ou Yes, Alive ou Dead et concernant l'âge pas d'âge < 0 et > 100 ans.

Vérificaitons : données manquantes

```
na_records = apply(data_exo6, 1, function (x) any(is.na(x)))
data_exo6[na_records,]
```

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

Il n'y a aucune données manquantes

Exercices

Partie 1

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

On utilise la fonction table avec les variables demandées : ici Smokers et status

```
tableau_mortalite= table(data_exo6$Smoker,data_exo6$Status)
tableau_mortalite
```

```
##
##       Alive Dead
##   No    502  230
##   Yes   443  139
```

Consignes :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 utilise la fonciton prop.table pour avoir les taux de mortalité dans chaque groupe. La fonction round permet d'avoir un arrondi du résultat

```
round(((prop.table((tableau_mortalite)))*100),2)
```

```
##
##       Alive  Dead
##   No  38.20 17.50
##   Yes 33.71 10.58
```

On obtient alors un pourcentage de mortalité de :

- 17.50 % : Groupe Non Fumeuse
- 10.58 % : groupe Fumeuse

Afin de ne pas utiliser manuellement les données écrites ci-dessus, nous allons les conserver dans des vecteurs

```
tableau = round(((prop.table(table(data_exo6$Smoker,data_exo6$Status)))*100),2)

#creation d'un data frame à partir du tableau

export_data_fram = as.data.frame(tableau)

# selection des colonnes et variables qui nous interesse et creation d'un nouveau data fram

new_data_frame = as.data.frame(export_data_fram[3:4,3])

# attribution des valeurs de mortalite pour les groupes à des vecteurs
```

```
mortalite_non_fum = new_data_frame[1,]

mortalite_fumeuse = new_data_frame[2,]
```

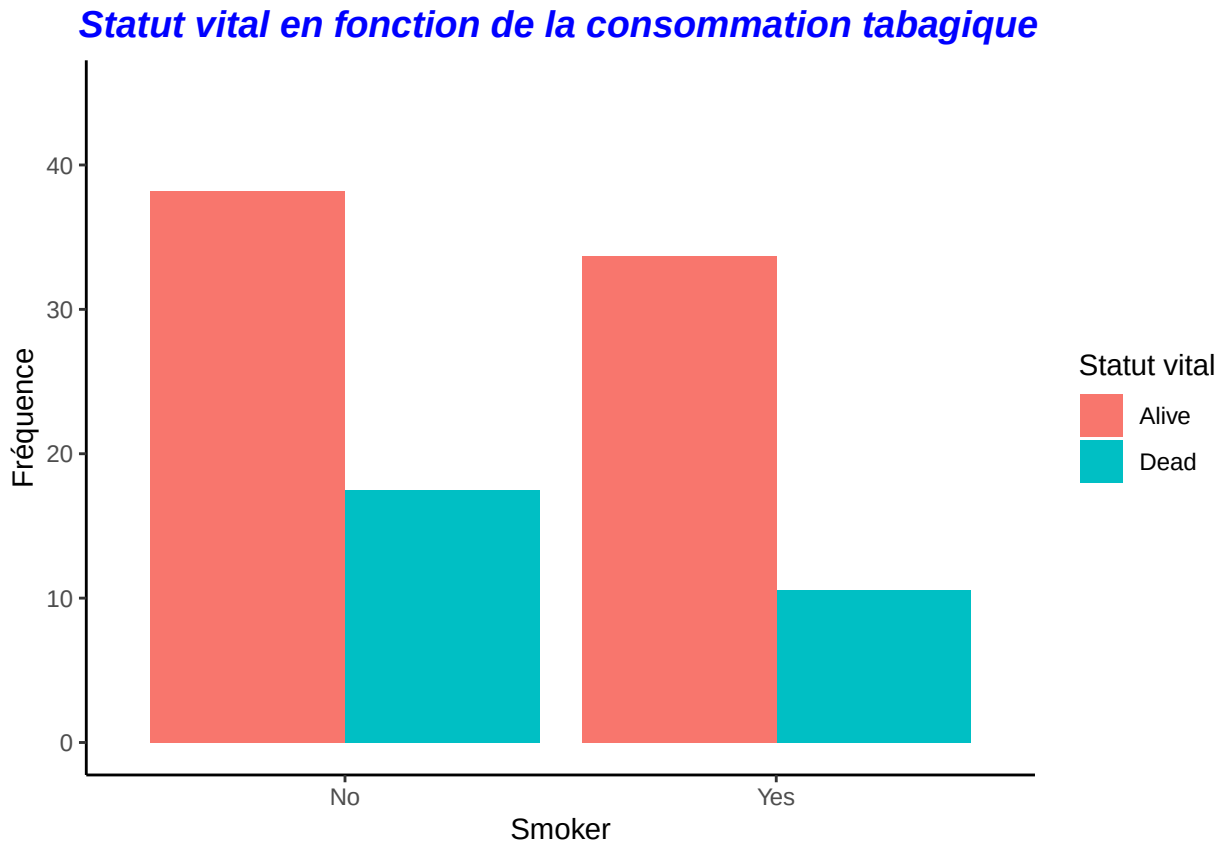
###**Consignes :** Vous pourrez proposer une représentation graphique de ces données et calculer des intervalles de confiance si vous le souhaitez.

Pour la représentation graphique nous allons procéder en 2 étapes :

- créer un data.frame, que l'on nomme figure de prop.table((tableau_mortalité)) avec la fonction as.data.frame
- utiliser le package ggplot2 pour faire une représentation graphique de ce data frame

```
figure = as.data.frame((prop.table((tableau_mortalite)))*100)


library(ggplot2)
ggplot(figure, aes(Var1,Freq,fill = Var2)) +
geom_bar(stat = "identity",position= "dodge") + labs(title="Statut vital en fonction de la consommation
        x ="Smoker", y = "Fréquence",fill = "Statut vital") +theme_classic()+ theme(plot.title = elemen
```



Pour le calcul des intervals de confiance nous allons avoir besoin d'utiliser le package binom.

```
library(binom)
```

```
## Warning: package 'binom' was built under R version 3.6.3
```

Puis on utilise la fonction binom.confint en precisant la proportion de mortalité dans le groupe fumeuse et ensuite non fumeuse en fonction de la population totale

```
# Tout d'abord je vais attribuer à la variable N le nombre total de femme dans la population

N = length(data_exo6$Age)

# On utilise la fonction binom.confint, et ici on choisit la methode exact

IC_mort_fume = binom.confint(mortalite_fumeuse, N, methods = "exact")

IC_mort_non_fum = binom.confint(mortalite_non_fum, N, methods = "exact")


IC_mort_fume
```

```
##   method     x    n       mean       lower      upper
## 1  exact 10.58 1314 0.00805175 0.003961841 0.01451961
```

```
IC_mort_non_fum
```

```
##   method    x    n       mean       lower     upper
## 1  exact 17.5 1314 0.01331811 0.007845613 0.0210993
```

Consignes : En quoi ce résultat est-il surprenant ?

Ces données parraissent surprennantes puisqu'elles sont contradictoires avec la littérature. Les femmes qui fumaîent semblent avoir un taux de mortalité amoindrie en comparaison avec celles qui ne fumaient pas.

Partie 2

Consignes : 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.

Nous allons créer une nouvelle variable que l'on nomme age_classe correspondant aux différentes catégories d'âges.

```
# D'abord recuperer l'âge maximal sur la variable Age

max = max(data_exo6$Age)

data_exo6$Age_classe = cut(data_exo6$Age, c(18,34.01,54,64,max),right = TRUE,include.lowest = TRUE,label

table(data_exo6$Age_classe)
```

```
##
## 18-34 ans 34-54 ans 55-64 ans  > 65 ans
##       400       436       236       242
```

```r
#On verifie qu'on obtient le même nombre d'individu : ici 1314 codée avec la variable N
length(data_exo6$Age_classe)== N
```

```
## [1] TRUE
```

Puis nous allons créer un nouveau tableau en rentrant la variable Age_classe

```r
new_tableau_mortalite= table(data_exo6$Smoker,data_exo6$Status,data_exo6$Age_classe)
new_tableau_mortalite
```

```
## , ,  = 18-34 ans
##
##
##       Alive Dead
##   No    213    6
##   Yes   176    5
##
## , ,  = 34-54 ans
##
##
##       Alive Dead
##   No    180   19
##   Yes   196   41
##
## , ,  = 55-64 ans
##
##
##       Alive Dead
##   No     81   40
##   Yes    64   51
##
## , ,  = > 65 ans
##
##
##       Alive Dead
##   No     28  165
##   Yes     7   42
```

Puis nous allons regarder le pourcentage de décès dans chacun des groupes

```r
round(((prop.table((new_tableau_mortalite)))*100),2)
```

```
## , ,  = 18-34 ans
##
##
##       Alive  Dead
##   No  16.21  0.46
##   Yes 13.39  0.38
##
## , ,  = 34-54 ans
##
```

```
##
##       Alive  Dead
##   No  13.70  1.45
##   Yes 14.92  3.12
##
## , ,  = 55-64 ans
##
##
##       Alive  Dead
##   No   6.16  3.04
##   Yes  4.87  3.88
##
## , ,  = > 65 ans
##
##
##       Alive  Dead
##   No   2.13 12.56
##   Yes  0.53  3.20
```

```
tab_mort_classe_age = as.data.frame((prop.table(new_tableau_mortalite))*100)

tab_mort_classe_age
```

```
##    Var1  Var2      Var3       Freq
## 1    No Alive 18-34 ans 16.2100457
## 2   Yes Alive 18-34 ans 13.3942161
## 3    No  Dead 18-34 ans  0.4566210
## 4   Yes  Dead 18-34 ans  0.3805175
## 5    No Alive 34-54 ans 13.6986301
## 6   Yes Alive 34-54 ans 14.9162861
## 7    No  Dead 34-54 ans  1.4459665
## 8   Yes  Dead 34-54 ans  3.1202435
## 9    No Alive 55-64 ans  6.1643836
## 10  Yes Alive 55-64 ans  4.8706240
## 11   No  Dead 55-64 ans  3.0441400
## 12  Yes  Dead 55-64 ans  3.8812785
## 13   No Alive  > 65 ans  2.1308980
## 14  Yes Alive  > 65 ans  0.5327245
## 15   No  Dead  > 65 ans 12.5570776
## 16  Yes  Dead  > 65 ans  3.1963470
```

Nous souhaitons disposer que du statut “Dead” pour cela nous allons créer un data frame en utilisant la fonction data.frame et utiliser la fonction filter, en utilisant avant le package dplyr pour selectionner dans la variable dead.

```
names(tab_mort_classe_age)
```

```
## [1] "Var1" "Var2" "Var3" "Freq"
```

```
names(tab_mort_classe_age)[2]= "Status"


library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
filter(tab_mort_classe_age, Status == "Dead")
```

```
##   Var1 Status      Var3       Freq
## 1   No   Dead 18-34 ans  0.4566210
## 2  Yes   Dead 18-34 ans  0.3805175
## 3   No   Dead 34-54 ans  1.4459665
## 4  Yes   Dead 34-54 ans  3.1202435
## 5   No   Dead 55-64 ans  3.0441400
## 6  Yes   Dead 55-64 ans  3.8812785
## 7   No   Dead  > 65 ans 12.5570776
## 8  Yes   Dead  > 65 ans  3.1963470
```

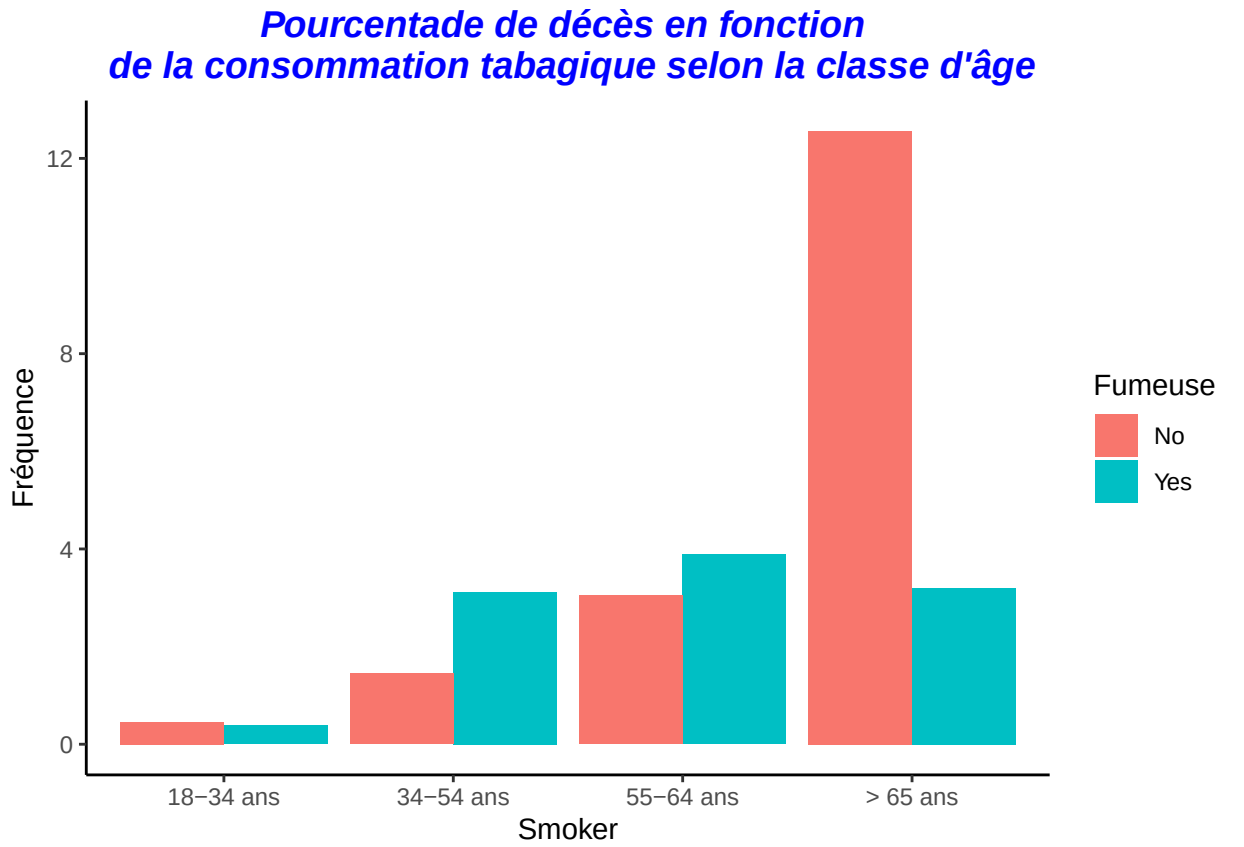
```
#Création du nouveau data_frame
new_tab_mort_classe_age = as.data.frame(filter(tab_mort_classe_age,Status == "Dead"))
```

Réalisation d'une figure de la mortalité selon l'âge et le statut tabagique

Pour cela nous utilison le package ggplot2

```
library(ggplot2)

ggplot(new_tab_mort_classe_age, aes(Var3, Freq, fill = Var1)) +
geom_bar(stat = "identity",position= "dodge") + labs(title="Pourcentade de décès en fonction \n de la c
        x ="Smoker", y = "Fréquence",fill = "Fumeuse")+  theme_classic() + theme(plot.title = element_t
```



###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 remarque que si l'on raissone par classe d'âge on obtient des différences de mortalité entre les groupes fumeuses et non fumeuses. Cependant, on observe que les femmes non fumeuses décèdent **3 fois plus** dans la tranche d'âge 64 et plus.

Cela pourrait s'expliquer par une moyenne d'âge plus élevée

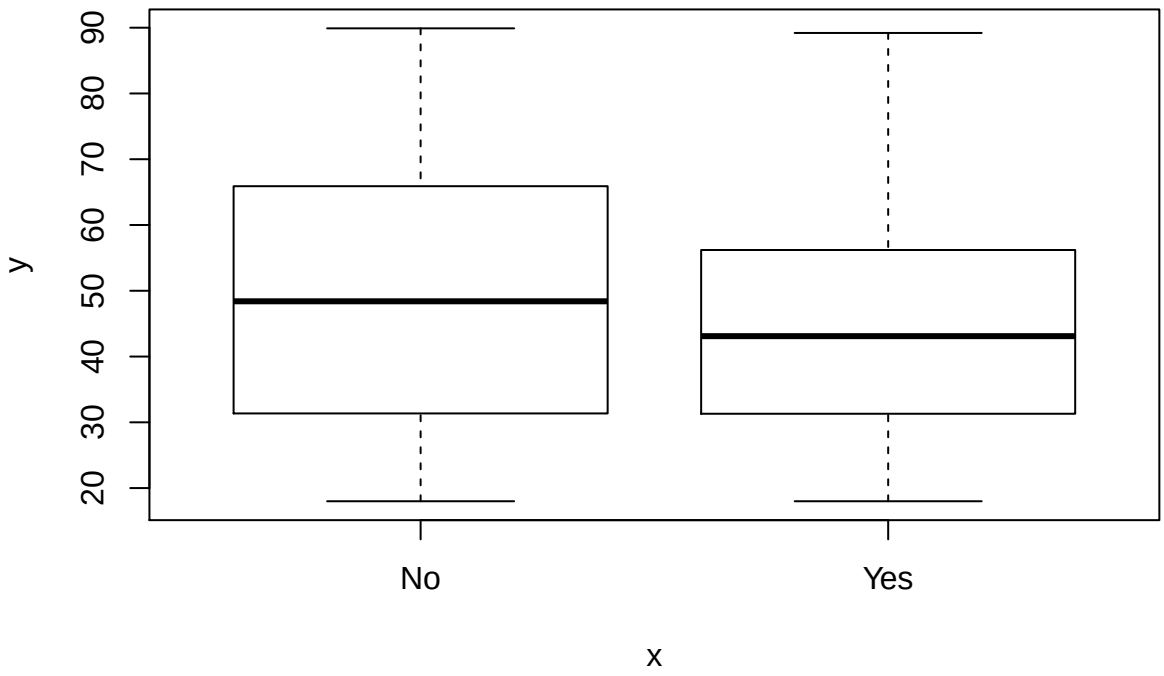
Nous pouvons le vérifier en :

- calculant la moyenne d'âge dans le groupe fumeuse et non fumeuse avec la fonction tapply
- realiser un boxplot avec la fonction plot
- réaliser un test t de student en supposant que la normalité et la variance sont respectée

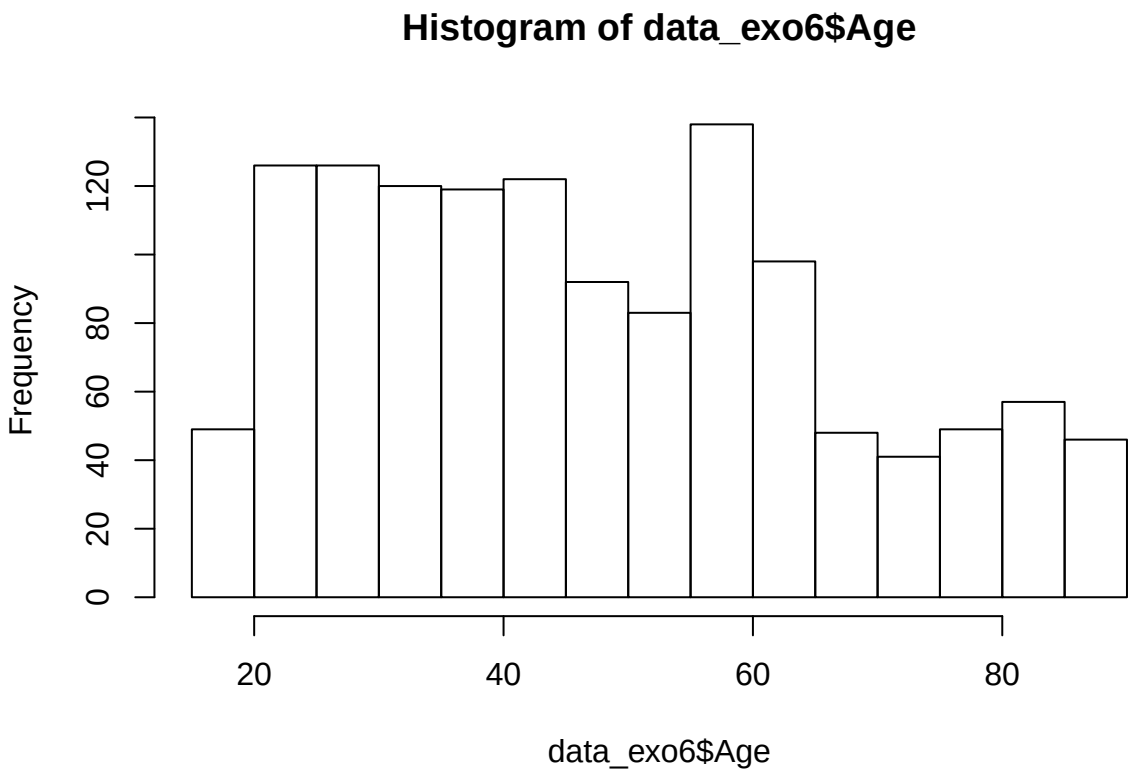
```
tapply(data_exo6$Age,data_exo6$Smoker, mean)
```

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

```
plot(data_exo6$Smoker,data_exo6$Age)
```




```
# Verfication normalité distribution et normalité de la variance
hist(data_exo6$Age)
```




```
by(data_exo6$Age,data_exo6$Smoker, sd, na.rm = TRUE)
```

```
## data_exo6$Smoker: No
## [1] 20.89829
## ------------------------------------------------------------
## data_exo6$Smoker: Yes
## [1] 16.21789
```

```
t.test(data_exo6$Age~data_exo6$Smoker, var.equal=TRUE)
```

```
##
##  Two Sample t-test
##
## data:  data_exo6$Age by data_exo6$Smoker
## t = 5.2647, df = 1312, p-value = 1.639e-07
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  3.479442 7.612734
## sample estimates:
##  mean in group No mean in group Yes
##          49.81585           44.26976
```

On peut compléter ces résultats en calculant la moyenne d'âge dans chacun des groupes dans la tranche d'âge 65 ans et +.

Pour cela nous allons :

- utiliser le package dplyr pour utiliser la fonction filter
- creer un nouveau data frame contenant seulement les individu > 65 ans
- calculer la moyenne dans chacun des groupes

```
library(dplyr)

verification_age = as.data.frame(filter(data_exo6, Age_classe== "> 65 ans"))

by(verification_age$Age,verification_age$Smoker,mean)
```

```
## verification_age$Smoker: No
## [1] 77.68756
## ------------------------------------------------------------
## verification_age$Smoker: Yes
## [1] 77.6551
```

Il y a une différence significative d'âge entre les deux groupes. Le groupe non fumeuse est significativement plus âgée. De plus il y a un pourcentage de femme non fumeuse importante et par conséquent cela renforce les chiffres de mortalité à la fin. Plus de femme âgée et non fumeuse en comparaison avec peu de femme fumeuse et moins jeunes. Toutefois on n'observe pas de différence dans la tranche d'âge 65 ans et plus.

Partie 3

Conignes : 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 commence par créer une nouvelle variable Death avec la fonction recode Alive = 0 Dead = 1 et qui devra être numérique

```
library(questionr)
```

```
## Warning: package 'questionr' was built under R version 3.6.3
```

```
library(dplyr)
data_exo6$Death <- recode(data_exo6$Status,
                "Alive" = "0",
                "Dead" = "1")
data_exo6$Death <- as.numeric(as.character(data_exo6$Death))


class(data_exo6$Death)
```

```
## [1] "numeric"
```

Realisation de la regression logistique avec la fonction glm

```r
#on va utiliser utiliser une regression logistique multiple avec fonction glm

# on precise binomial car on se trouve dans une regression logistique

mod4= glm(Death~Age*Smoker, data= data_exo6, family="binomial")
summary(mod4)
```

```
##
## Call:
## glm(formula = Death ~ Age * Smoker, family = "binomial", data = data_exo6)
##
## 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)   -6.795507   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
```

```r
# Age * Smokers : permet de voir l'interraction entre les variables age et Smokers
```

Dans le cadre d'un modèle logistique, habituelement on ne présente pas les coefficients du modèle mais leur valeur exponentielle, cette dernière correspondant en effet à des odds ratio

Pour cela on utilise le package **questionr** avec la fonction odds.ratio()

```r
library(questionr)
odds.ratio(mod4)
```

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

```
##                      OR     2.5 %  97.5 %       p
## (Intercept)   0.0011188 0.0004158  0.0027 < 2e-16 ***
## Age           1.1132407 1.0971189  1.1313 < 2e-16 ***
## SmokerYes     3.6233570 0.9748334 13.5390 0.05419 .
## Age:SmokerYes 0.9818676 0.9596182  1.0048 0.11792
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Pour réaliser une représentation graphique des regression logistiques, nous allons d'abord utiliser le package **forestmodel** et la fonction forest_model. Cela permet d'afficher une représentation graphique visuelle et tabulaire

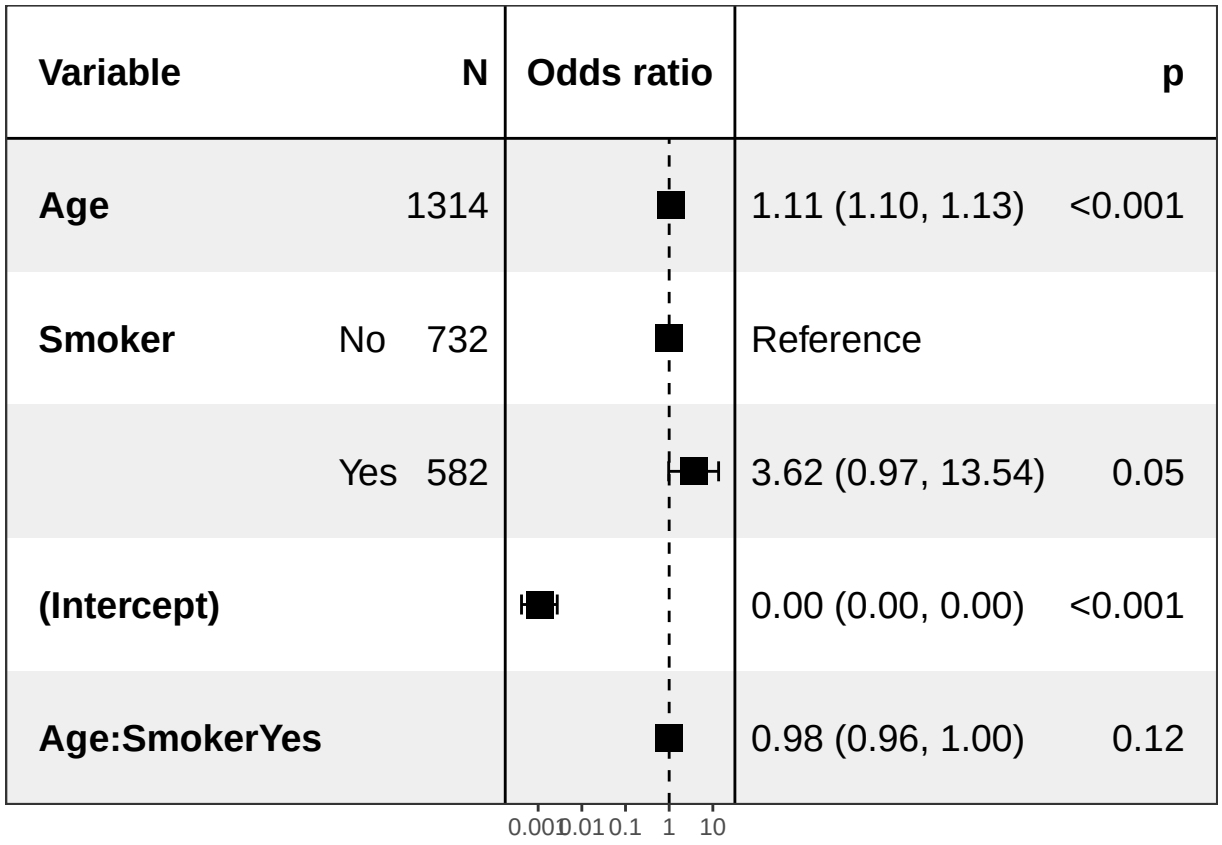
```
# la fonction forest_model va permettre de representer un tableau d'ODDS ratio du modèle de regression
# On doit d'abord regarder l'interraction, si elle n'est pas significative on n'accorde pas d'important
library(forestmodel)
```

```
## Warning: package 'forestmodel' was built under R version 3.6.3
```

```
forest_model(mod4)
```

```
## Warning in recalculate_width_panels(panel_positions, mapped_text =
## mapped_text, : Unable to resize forest panel to be smaller than its heading;
## consider a smaller text size
```

```
## Warning: Ignoring unknown aesthetics: x
```



Pour réaliser une représentation graphique de la régression logistique en fonction du statut tabagique on peut utiliser la fonction ggeffects

```
library(effects)
```

```
## Warning: package 'effects' was built under R version 3.6.3
```

```
## Loading required package: carData

## lattice theme set by effectsTheme()
## See ?effectsTheme for details.
```

```
library(ggeffects)
```

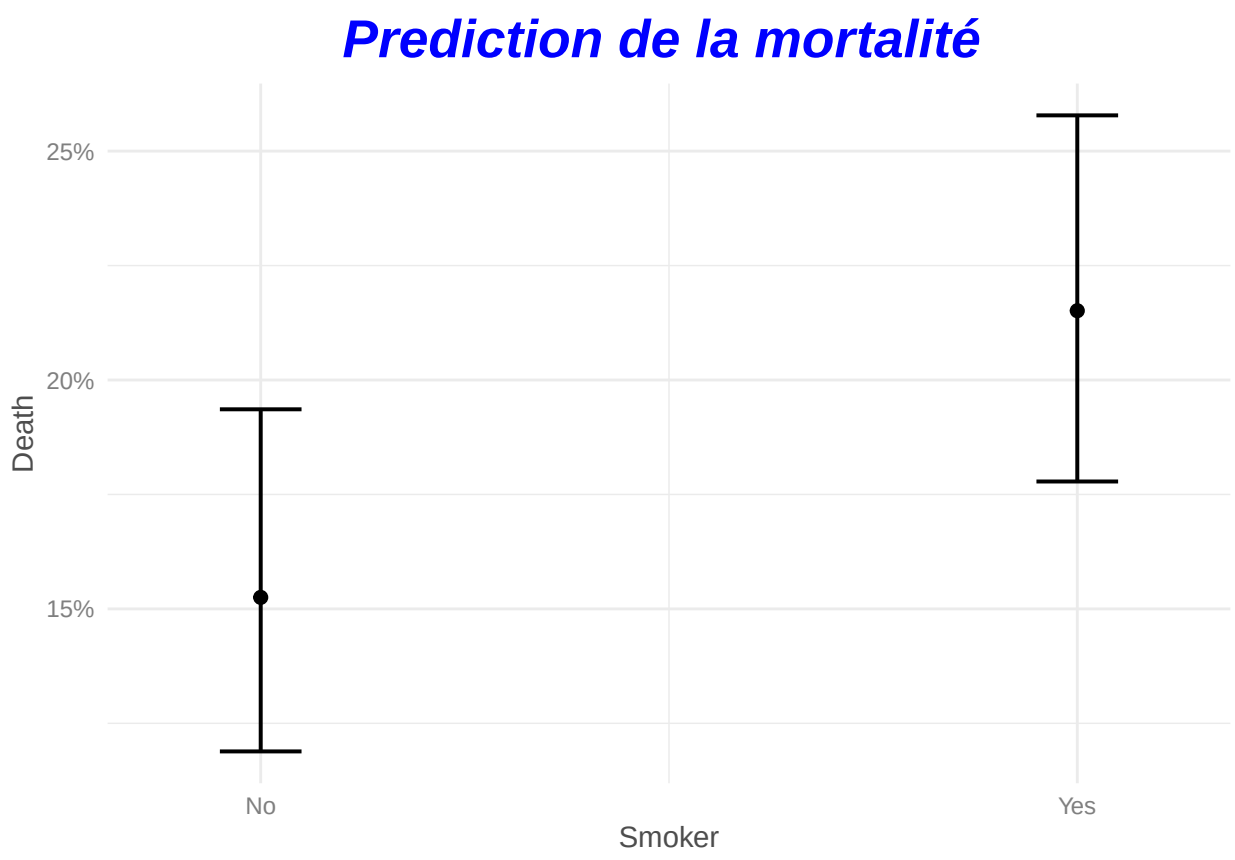
```
## Warning: package 'ggeffects' was built under R version 3.6.3
```

```
#Pour afficher un tableau
ggeffect(mod4, "Smokers")
```

```
## `Smokers` was not found in model terms. Maybe misspelled?
## Can't compute marginal effects, 'effects::Effect()' returned an error.
##
## Reason: the following predictor is not in the model: Smokers
## You may try 'ggpredict()' or 'ggemmeans()'.

## NULL
```

```
#Pour afficher un graphique
plot(ggeffect(mod4, "Smoker")) + labs(title = "Prediction de la mortalité ")  + theme(plot.title = eleme
```



A partir de ce dernier graphique on peut induire que les femmes fumeuses ont plus de risque de décès que les femmes non fumeuses.

En conclusion :

Ces résultats ne permettent pas de conclure sur la nocivité du tabac. En effet, dans un premier temps il apparaissait que les femmes fumeuses avaient un taux de mortalité amoindrie en comparaison avec les non fumeuses.Cela pouvait amener à première vue à dire que les femmes fumeuses ont un taux de mortalité amoindrie

Néanmoins, cet effet sur l'ensemble de la population semble s'inverser lorsqu'on réalise des catégories sur les tranches d'âges. De plus en réalisant ces tranches d'âges, on peut vérifier la moyenne d'âge dans les groupes fumeuses et non fumeuses. On observe alors une différence significative. Les femmes non fumeuses sont plus âgées en moyennes de façon significative comparée aux femmes fumeuses.

Enfin en réalisant une régression logistique, aucun résultat significatif apparaît entre la mortalité et l'interaction entre l'âge et le statut tabagique.

Ce résultats sont liés à des éléments qui ne sont pas pris en compte (comme la présence de variables non indépendantes ou de différences d'effectifs entre les groupes, et la manière d'observer et d'analyser les variables)

Le youtubeur et bloggeur Science Etonnante l'explique parfaitement dans son article.