Sujet 6 : Autour du Paradoxe de Simpson

Clair Ch

20 avril 2020

Instructions

Voici les instructions du Sujet 6 : Autour du Paradoxe de Simpson

> 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.
>
> Les données sont disponibles dans ce fichier CSV. Vous trouverez sur chaque ligne si la personne fume ou non, si elle est vivante ou décédée au moment de la seconde étude, et son âge lors du premier sondage.
>
> Cet exercice peut être réalisé indifféremment en R ou en Python.
>
> Votre mission si vous l'acceptez :
>
> 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 (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 ?
> 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.
> 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).
> 4. Déposez votre étude dans FUN

Préparation des données

Téléchargement

Les données sont disponibles sur le Gitlab du MOOC "Recherche Reproductible : principes méthodologiques pour une science transparente" de l'Inria. On peut les récupérer au format .csv à cette adresse :

```
data_url<-"https://gitlab.inria.fr/learninglab/mooc-rr/mooc-rr-ressources/-/raw/master/
module3/Practical_session/Subject6_smoking.csv?inline=false"
```

Elles regroupent, par femme, les habitudes de tabagisme (fumeuse/non fumeuse), le statut lors de la deuxième étude de 1995 (vivante ou morte) et l'âge lors de la première étude (effectuée entre 1972 et 1974).

Pour nous protéger contre une éventuelle disparition ou modification du serveur du Gitlab du MOOC, nous faisons une copie locale de ce jeux de données que nous préservons avec notre analyse. Il est inutile et même risqué 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 <- "Subject6_smoking.csv"
if (!file.exists(data_file)) {
    download.file(data_url, data_file, method="auto")
}
```

Lecture

```
data<-read.csv(data_file)
```

Voyons voir à quoi ressemblent les données :

```
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
```

```
tail(data)
```

```
##      Smoker Status  Age
## 1309     No  Alive 42.1
## 1310    Yes  Alive 35.9
## 1311     No  Alive 22.3
## 1312    Yes   Dead 62.1
## 1313     No   Dead 88.6
## 1314     No  Alive 39.1
```

Nous avons les données pour 1314 femmes, avec dans l'ordre : si la personne fume ou non (colonne `Smoker` : `Yes` ou `No`), si elle est morte ou vivante au moment de la deuxième étude en 1995 (colonne `Status` : `Dead` ou `Alive`) et son âge lors de la première étude faite entre 1972 et 1974 (colonne `Age`).

Y a-t'il des points manquants dans nos données ?

```
na_records <- apply(data,1,function(x) any(is.na(x)))
data[na_records,]
```

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

Aucune donnée manquante, parfait !

Vérifions la classe de chaque colonne

```
class(data$Smoker)
```

```
## [1] "factor"
```

```
class(data$Status)
```

```
## [1] "factor"
```

```
class(data$Age)
```

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

Les données de la colonne `Age` sont bien de classe numeric, les autres colonnes de type factor.

Décès et tabagisme

Nous voulons étudier le taux de mortalité pour chaque groupe (fumeuses/non fumeuses)

Inspection des données

Faisons un tableau représentant le nombre de femmes par habitude de tabagisme (fumeuses/non fumeuses) et statut (vivantes/mortes)

```
t<-table(data$Smoker,data$Status)
t
```

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

Vérifions que la somme de chaque catégorie correspond bien au nombre total de femmes de l'étude

```
sum(t)==nrow(data)
```

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

Nous pouvons voir qu'il y a plus de femmes vivantes que de femmes mortes et plus de femmes qui ne fument pas :

```
alive<-sum(t[,1]) #femmes vivantes
alive
```

```
## [1] 945
```

```
dead<-sum(t[,2]) #femmes mortes
dead
```

```
## [1] 369
```

Il y a également plus de femmes non fumeuses que fumeuses :

```
nsmoke<-sum(t[1,]) #femmes non fumeuses
nsmoke
```

```
## [1] 732
```

```
smoke<-sum(t[2,]) #femmes fumeuses
smoke
```

```
## [1] 582
```

Calcul du taux de mortalité

Calculons dans chaque groupe (fumeuses/non fumeuses) le taux de mortalité (rapport entre le nombre de femmes décédées dans un groupe et le nombre total de femmes dans ce groupe)

taux de mortalité pour les femmes non fumeuses :

```
nsmoke_morta<-t[1,2]/nsmoke
nsmoke_morta
```

```
## [1] 0.3142077
```

taux de mortalité pour les femmes fumeuses :

```
smoke_morta<-t[2,2]/smoke
smoke_morta
```

```
## [1] 0.2388316
```

Les femmes non fumeuses présentent un taux de mortalité plus élevé !

De combien est-il plus élevé ?

```
nsmoke_morta/smoke_morta
```

```
## [1] 1.315603
```

Arrondi au centième, les femmes non fumeuses présent donc un taux de mortalité 1.32 fois plus élevé que les femmes fumeuses dans cette étude. C'est surprenant car contre-intruitif !

On peut faire un test statistique pour avoir une idée si cette différence n'est pas due au hasard (test Chi-square de Pearson sur les proportions) :

```
chisq.test(t)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  t
## X-squared = 8.7515, df = 1, p-value = 0.003093
```

La p-value est en dessous de 0.05, il y a donc moins de 5% de chance que cette différence du taux de mortalité soit due au hasard. Mais est-elle vraiment due aux habitudes de tabagisme ?

Calcul de l'intervalle de confiance à 95% du taux de mortalité

Pour calculer les intervalles de confiance à 95% du taux de mortalité par catégorie (fumeuses/non fumeuses), nous allons suivre les instructions de l'article How to Determine the Confidence Interval for a Population Proportion du site dummies.com

La formule donnée pour calculé l'intervalle de confiance à 95% est la suivante :

$$1,96\sqrt{\frac{\beta(1-\beta)}{n}}$$

β représente ici le taux de mortalité des femmes fumeuses ou non fumeuses et n le nombre total de femmes fumeuses ou non fumeuses

Créons la fonction correspondante :

```
CI95<-function(morta,n){
  1.96*sqrt((morta*(1-morta))/n)
}
```

Calculons les intervalles de confiance pour les femmes non fumeuses et les femmes fumeuses :

```
nsmoke_CI<-CI95(nsmoke_morta,nsmoke)
smoke_CI<-CI95(smoke_morta,smoke)
```

Arrondi au centième, les femmes fumeuses ont un taux de mortalité de 0.24 ± 0.03

Arrondi au centième, les femmes non fumeuses ont un taux de mortalité de 0.31 ± 0.03

Représentation graphique

Pour faire les graphiques, nous utiliserons la librairie `ggplot2`. Ce code permet d'installer la libraire si nécessaire et de la loader.

```
if(!require(ggplot2)){
    install.packages("ggplot2")
    library(ggplot2)
}
```

```
## Loading required package: ggplot2
```

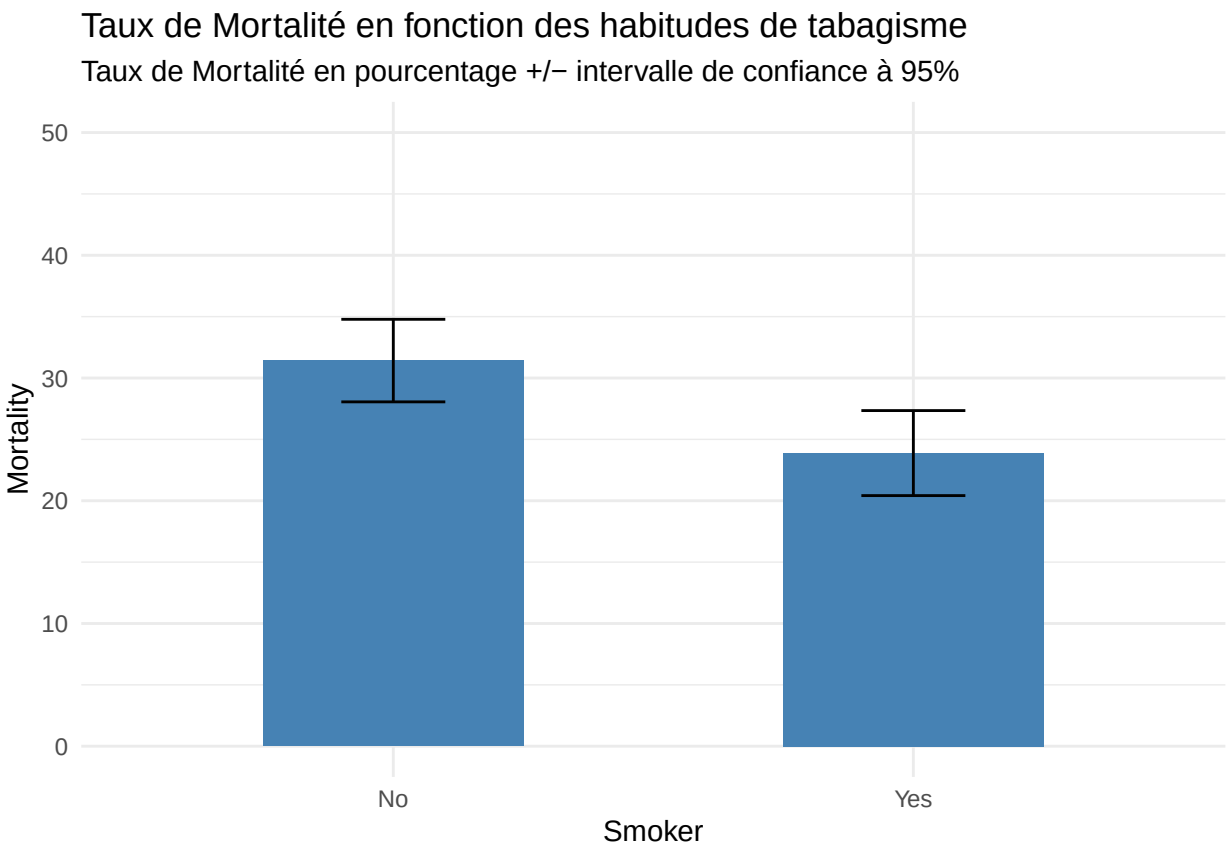
Pour faire le graphe, il faut d'abord mettre les données de taux de mortalité et d'intervalle de confiance dans une data frame, j'en profite pour les passer en pourcentage :

```
df_morta<-data.frame(Smoker=c("No","Yes"), Mortality=c(nsmoke_morta*100,smoke_morta*100), CI95=c(nsmoke
df_morta
```

```
##   Smoker Mortality     CI95
## 1     No  31.42077 3.362832
## 2    Yes  23.88316 3.464024
```

Avec l'aide d'un tutoriel pour faire des barplots, faisons un graphe représentant le taux de mortalité en fonction des habitudes de tabagisme :

```
plot_morta<-ggplot(df_morta,aes(x=Smoker,y=Mortality))+
  geom_bar(stat="identity",fill="steelblue",width=0.5)+
  theme_minimal()+
  ylim(c(0,50))+ #limite de l'axe y
  geom_errorbar(aes(ymin=Mortality-CI95,ymax=Mortality+CI95),width=.2,position=position_dodge(.9))+ #ba
  labs(title="Taux de Mortalité en fonction des habitudes de tabagisme", subtitle="Taux de Mortalité en
plot_morta
```



Les femmes non fumeuses semblent avoir un taux de mortalité plus élevé que les femmes fumeuses. Si on considère ce résultat comme tel, on serait tenté de conclure que la cigarette est bonne pour la santé !

Décés et Tabagisme en fonction de la classe d'âge

Ajout de la tranche d'âge

Rajoutons une colonne `Tranche` dans le fichier `data`donnant la tranche d'âge (`18-34`: de 18 ans inclus à 35 ans non inclus ; `35-54`: de 35 ans inclus à 55 ans non inclus ; `55-64`: de 55 ans inclus à 65 non inclus ; `>65`: plus de 65 ans inclus), mettons le message `ERROR`si une tranche d'âge n'est pas trouvée

```
for(i in 1:nrow(data)){
    if(data$Age[i]>=65){data$Tranche[i]<-">65"
  }else if(data$Age[i]>=55){data$Tranche[i]<-"55-64"
  }else if(data$Age[i]>=35){data$Tranche[i]<-"35-54"
  }else if(data$Age[i]>=18){data$Tranche[i]<-"18-34"
  }else(data$Tranche[i]<-"ERROR")
}
```

Vérifions que ça ait bien marché :

```
head(data)
```

```
##   Smoker Status  Age Tranche
## 1    Yes  Alive 21.0   18-34
## 2    Yes  Alive 19.3   18-34
## 3     No   Dead 57.5   55-64
## 4     No  Alive 47.1   35-54
```

```
## 5    Yes  Alive 81.4     >65
## 6     No  Alive 36.8   35-54
```

```
tail(data)
```

```
##      Smoker Status  Age Tranche
## 1309     No  Alive 42.1   35-54
## 1310    Yes  Alive 35.9   35-54
## 1311     No  Alive 22.3   18-34
## 1312    Yes   Dead 62.1   55-64
## 1313     No   Dead 88.6     >65
## 1314     No  Alive 39.1   35-54
```

Est-ce qu'il y a eu des erreurs ?

```
sum(data$Tranche=="ERROR")
```

```
## [1] 0
```

Aucune erreur, tout va bien !

Calcul du taux de mortalité par tranche d'âge et tabagisme

Faisons une fonction permettant de calculer le nombre de femmes par tranche d'âge selon leur habitude de tabagisme (fumeuse/non fumeuse) et éventuellement leur statut (dead/alive)

```
comptage<-function(smoker,statut,tranche){
  if(statut=="NULL")
  {sum(data$Smoker==smoker&data$Tranche==tranche)
    }else{sum(data$Smoker==smoker&data$Status==statut&data$Tranche==tranche)}
}
```

Calculons le nombre de femmes totales et mortes par tranche d'âge et catégories

```
tranches<-c("18-34","35-54","55-64",">65")

nsmoke_age<-mapply(comptage,smoker="No",statut="NULL",tranche=tranches) #femmes non fumeuses
smoke_age<-mapply(comptage,smoker="Yes",statut="NULL",tranche=tranches) #femmes fumeuses
nsmoke_age_dead<-mapply(comptage,smoker="No",statut="Dead",tranche=tranches) #femmes non fumeuses morte
smoke_age_dead<-mapply(comptage,smoker="Yes",statut="Dead",tranche=tranches) #femmes fumeuses mortes

#nommer les tranches d'âge dans les vecteurs :
names(nsmoke_age)<-tranches
names(smoke_age)<-tranches
names(nsmoke_age_dead)<-tranches
names(smoke_age_dead)<-tranches
```

Vérifions que les nombres de femmes fumeuses et non fumeuses correspondent bien au nombre de femmes totales

```
sum(smoke_age,nsmoke_age)==nrow(data)
```

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

Vérifions également que le nombre de femmes mortes fumeuses et non fumeuses correspondent bien au nombre de femmes totales mortes :

```
sum(smoke_age_dead,nsmoke_age_dead)==nrow(data$Status=="Dead")
```

```
## logical(0)
```

Nous pouvons maintenant calculer le taux de mortalité pour les femmes non fumeuses par tranche d'âge

```
nsmoke_age_morta<-nsmoke_age_dead/nsmoke_age
nsmoke_age_morta
```

```
##      18-34      35-54      55-64        >65
## 0.02643172 0.09947644 0.33057851 0.85492228
```

Et pour les femmes fumeuses par tranche d'âge :

```
smoke_age_morta<-smoke_age_dead/smoke_age
smoke_age_morta
```

```
##      18-34      35-54      55-64        >65
## 0.03703704 0.17030568 0.44347826 0.85714286
```

Mettons les taux de mortalité en pourcentage au centième près selon les tranches d'âges et habitude de tabagisme dans une data frame pour pouvoir les comparer :

```
df_morta_age<-data.frame(Age=tranches,"Mortality non Smoker"=round(nsmoke_age_morta*100,digit=2),"Morta
df_morta_age
```

```
##         Age Mortality.non.Smoker Mortality.Smoker
## 18-34 18-34                 2.64             3.70
## 35-54 35-54                 9.95            17.03
## 55-64 55-64                33.06            44.35
## >65     >65                85.49            85.71
```

On peut voir tout d'abord que de façon générale, le taux de mortalité est plus important pour les catégories d'âges les plus "avancées", ce qui est un résultat attendu. Ici la mortalité des femmes qui fument semble plus importante que celle des femmes qui ne fument pas dans toutes les tranches d'âge (à part celle des femmes de plus de 65 ans). On a donc une différence quand on regarde le taux de mortalité total et le taux de mortalité par tranche d'âge. Le taux de mortalité plus élevé chez les femmes non fumeuses n'était peut-être pas du au tabagisme...

Calcul de l'intervalle de confiance à 95%

De la même manière que pour le taux de mortalité global par tranche d'âge, nous calculons l'intervalle de confiance à 95% en prenant en compte la tranche d'âge :

```
nsmoke_age_IC<-mapply(CI95, morta=nsmoke_age_morta,n=nsmoke_age)
smoke_age_IC<-mapply(CI95,morta=smoke_age_morta,n=nsmoke_age)
```

Graphe

Mettons au propres les données pour faire le graphe: la 1ère colonne repésente l'habitude de tabagisme, la deuxième la tranche d'âge, la troisième le taux de mortalité (en pourcentage au centième près), la quatrième l'intervalle de confiance à 95% (en pourcentage au centième près)

```
df_morta_age2<-data.frame(Smoker=c(rep("No",4),rep("Yes",4)),Age=rep(c("18-34","35-54","55-64",">65"),2
df_morta_age2$CI95<-round(df_morta_age2$CI95*100,digits=2)
df_morta_age2$Mortality<-round(df_morta_age2$Mortality*100,digits=2)
df_morta_age2
```

```
##   Smoker   Age Mortality CI95
## 1     No 18-34      2.64 2.09
## 2     No 35-54      9.95 4.24
## 3     No 55-64     33.06 8.38
## 4     No   >65     85.49 4.97
```

```
## 5    Yes 18-34      3.70 2.46
## 6    Yes 35-54     17.03 5.33
## 7    Yes 55-64     44.35 8.85
## 8    Yes   >65     85.71 4.94
```

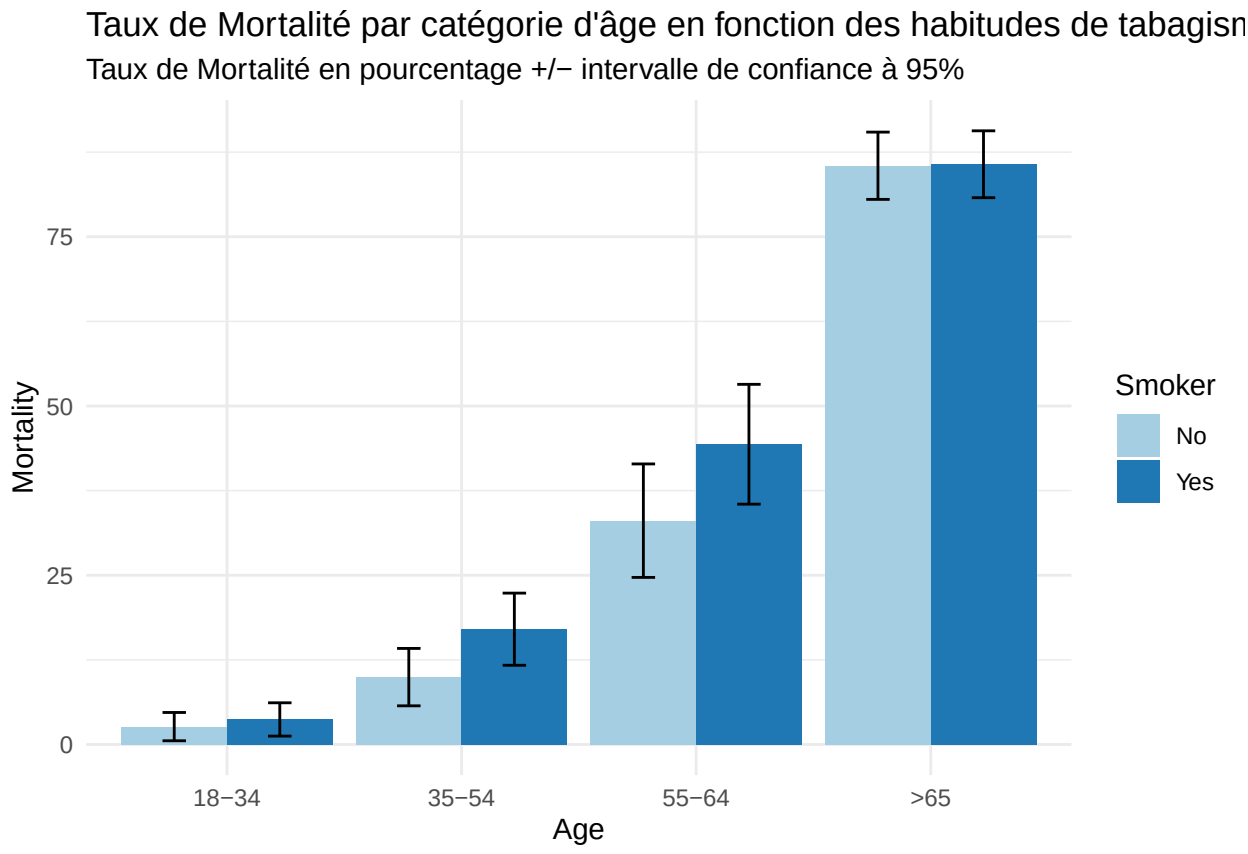
Pour que les tranches d'âges soient dans le bon ordre dans le graphe, nous ordonnons les niveaux de la colonne `Age`

```
df_morta_age2$Age<-ordered(df_morta_age2$Age, levels=c("18-34","35-54","55-64",">65"))
```

Faisons maintenant le graphe, toujours grâce à la libraire `ggplot2` :

```
if(!require(ggplot2)){
    install.packages("ggplot2")
    library(ggplot2)
}
```

```
plot_morta_age<-ggplot(df_morta_age2,aes(x=Age,y=Mortality,fill=Smoker))+
  geom_bar(stat="identity",position="dodge")+
  scale_fill_brewer(palette="Paired")+
  theme_minimal()+
  geom_errorbar(aes(ymin=Mortality-CI95, ymax=Mortality+CI95), width=.2, position=position_dodge(.9))+
  labs(title="Taux de Mortalité par catégorie d'âge en fonction des habitudes de tabagisme", subtitle="
plot_morta_age
```



Ici il apparait que la mortalité est plus importante pour les femmes fumeuses que non fumeuses, la cigarette ne semble au final pas bon pour la santé !

Paradoxe

Le paradoxe entre le premier résultat (si on regarde le taux de mortalité de façon globale, il est plus élevé pour les femmes non fumeuses) et ce dernier (si on classe les femmes par catégories d'âges, on a un taux de mortalité plus élevé pour les femmes fumeuses) pourrait provenir du fait que les catégories d'âges ne sont pas représentées de manière égales entre les femmes fumeuses et non fumeuses : les femmes "jeunes" auraient plus tendance à fumer, et au contraire les femmes "âgées" auraient tendance à moins fumer et seraient plus représentées dans le groupe des non fumeuses. Ainsi le taux de mortalité "naturelle" (de vieillesse) est globalement plus important pour le groupe des non fumeuses.

Nous pouvons vérifier cette hypothèse en analysant la distribution des âges par habitude de tabagisme :

```
summary(data$Age[data$Smoker=="No"])
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   18.00   31.38   48.40   49.82   65.85   89.90
```

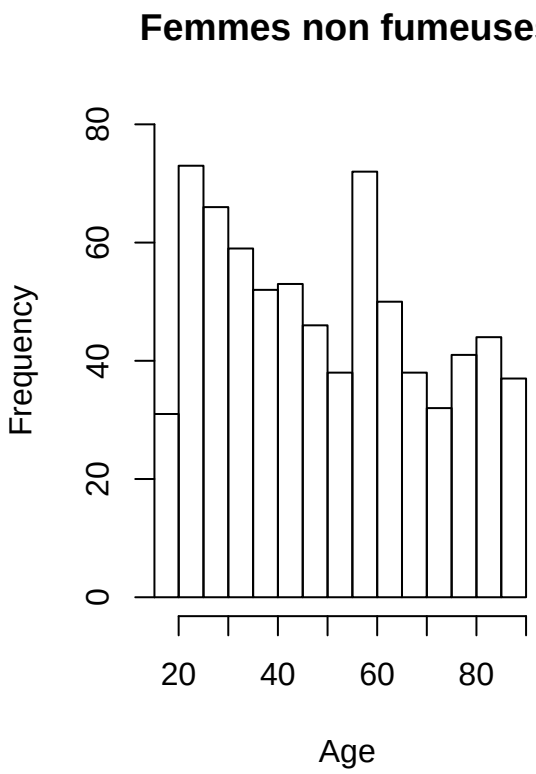
```
summary(data$Age[data$Smoker=="Yes"])
```

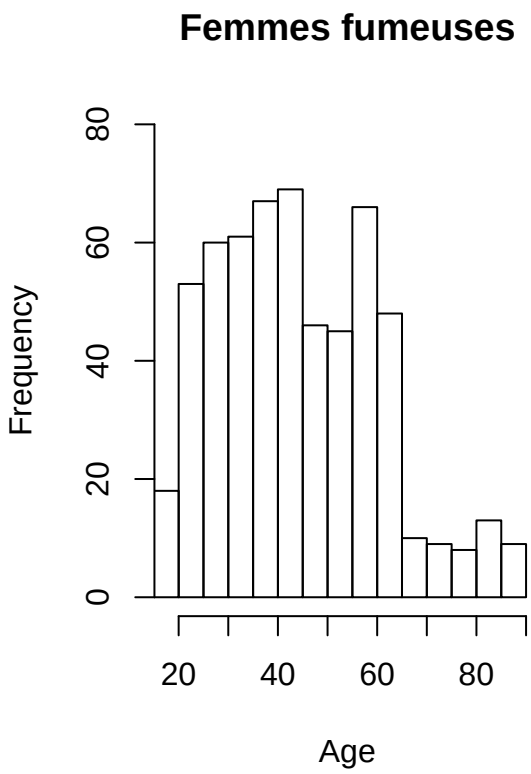
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   18.00   31.30   43.10   44.27   56.17   89.20
```

On voit que l'étendu des âges est sensiblement la même entre les femmes non fumeuses et fumeuses (de 18 à environ 89-90 ans). Par contre l'âge médian est plus élevé chez les femmes non fumeuses que non fumeuses (48.40 ans contre 43.10 ans). C'est en concordance avec notre théorie.

Comparons maintenant les histogrammes (distribution des âges) pour avoir une meilleure idée de la distribution des âges :

```
par(mfrow=c(1,2)) #pour avoir les histogrammes côte à côte
hist(data$Age[data$Smoker=="No"],main="Femmes non fumeuses",xlim=c(18,90),ylim=c(0,80),xlab="Age")
hist(data$Age[data$Smoker=="Yes"], main="Femmes fumeuses",xlim=c(18,90),ylim=c(0,80),xlab="Age")
```





En comparant ces histogrammes, on voit bien que la distribution des âges n'est pas du tout la même entre les femmes fumeuses et non fumeuses : il y a très peu de femmes "âgées" dans les femmes fumeuses. Ainsi beaucoup moins de femmes meurent de mort "naturelle" (vieillesse) dans le groupe des femmes fumeuses, ce qui abaisse le taux de mortalité global. Pour savoir si la cigarette a en effet un effet sur le taux de mortalité, il faut prendre en compte cette variable âge pour s'affranchir du taux de mortalité du à la vieillesse.

Régression logistique

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.

La régression logistique permet d'évaluer et caractériser les relations entre deux variables, ici l'âge et la mortalité. On peut trouver une explication dans ce document du MOOC "Recherche reproductible" et une mise en pratique dans celui-ci

Création de la variable Death

Ajoutons une variable `Death`prenant la valeur de 0 si la personne est décédée et 1 si la personne est vivante

```
for(i in 1:nrow(data)){
  if(data$Status[i]=="Dead")
  {data$Death[i]<-0}
  else(data$Death[i]<-1)
}
```

On vérifie que ça a bien fonctionné

```r
head(data)
```

```
##   Smoker Status  Age Tranche Death
## 1    Yes  Alive 21.0   18-34     1
## 2    Yes  Alive 19.3   18-34     1
## 3     No   Dead 57.5   55-64     0
## 4     No  Alive 47.1   35-54     1
## 5    Yes  Alive 81.4     >65     1
## 6     No  Alive 36.8   35-54     1
```

Calcul de la régression logistique

Faisons une régression logistique de la mortalité en fonction de l'âge selon que la personne fume ou non :

```r
logistic_reg<-glm(Death~Age+Smoker,data=data,family=binomial)
summary(logistic_reg)
```

```
##
## Call:
## glm(formula = Death ~ Age + Smoker, family = binomial, data = data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.9490  -0.4570   0.2830   0.5947   2.3129
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  6.351874   0.360121  17.638   <2e-16 ***
## Age         -0.099837   0.005774 -17.291   <2e-16 ***
## SmokerYes   -0.278654   0.164981  -1.689   0.0912 .
## ---
## 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: 1001.9  on 1311  degrees of freedom
## AIC: 1007.9
##
## Number of Fisher Scoring iterations: 5
```

On voit qu'il y a un effet négatif de l'âge sur la mortalité (estimate de -0.099+/-0.005) qui est signicatif ($p<0.05$). Il y a aussi un effet négatif de l'habitude de tabagisme (estimate de -0.27+/-0.14) mais qui n'est pas significatif ($p>0.05$).

Cette régression logistique ne montre aucune évidence d'un effet du tabagisme sur la mortalité dans ce jeu de données.

Graphes

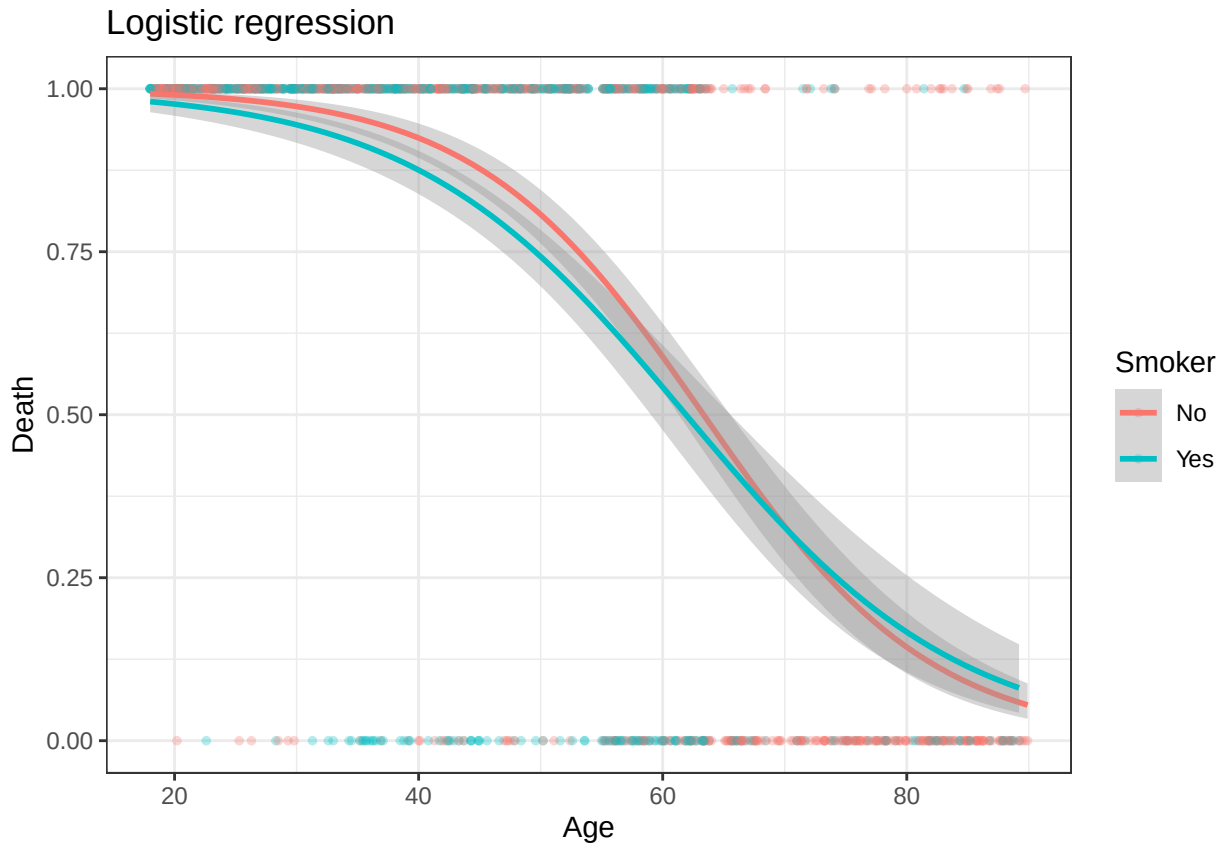
Pour faire le graphe, nous utiliserons la librairie `ggplot2`

```r
if(!require(ggplot2)){
    install.packages("ggplot2")
    library(ggplot2)
}
```

Faisons le graphe de la régression logistique de la mortalité (Death vaut 0 is la personne est décédée et 1 si la personne est morte) en fonction des habitudes de tabagisme et de l'âge, avec en grisé l'intervalle de confiance à 95%

```
ggplot(data,aes(x=Age,y=Death,color=Smoker)) + geom_point(alpha=.3,size=1) +
  theme_bw() +
  geom_smooth(method = "glm",
  method.args = list(family = "binomial"))+
    labs(title="Logistic regression")
```

```
## `geom_smooth()` using formula 'y ~ x'
```



Sur ce graphe, il ne semble pas globalement que le fait de fumer augmente ou non la probabilité d'être décedée 20 ans plus tard. Cependant la courbe des femmes non fumeuses semble faire un coude et être légèrement au dessus de celles des fumeuses pour un âge entre 25 et 65 ans. Re-faire une analyse sur cette tranche d'âge serait intéressant. Le tabagisme n'a peut-être aucun effet dans nos données sur le femmes les plus jeunes et les plus âgées (les plus âgées vont mourir de vieillesse, et peut-être que les femmes les plus jeunes ne fument pas depuis assez longtemps pour que ça ait un effet évident ou qu'elles sont plus protégées d'un effet néfaste du tabac).

Régression logistique sur les femmes entre 25 et 65 ans

Pour cette partie nous utiliserons la librarie 'dplyr

```
if(!require(dplyr)){
    install.packages("dplyr")
    library(dplyr)
}
```

```
## Loading required package: 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
```

Créons un deuxième jeu de données ne comprenant que les femmes entre 25 et 65 ans.

```
data2<-filter(data,Age>25&Age<65)
```

Faisons une régression logistique sur ces données :

```
logistic_reg2<-glm(Death~Age+Smoker,data=data2,family=binomial)
summary(logistic_reg2)
```

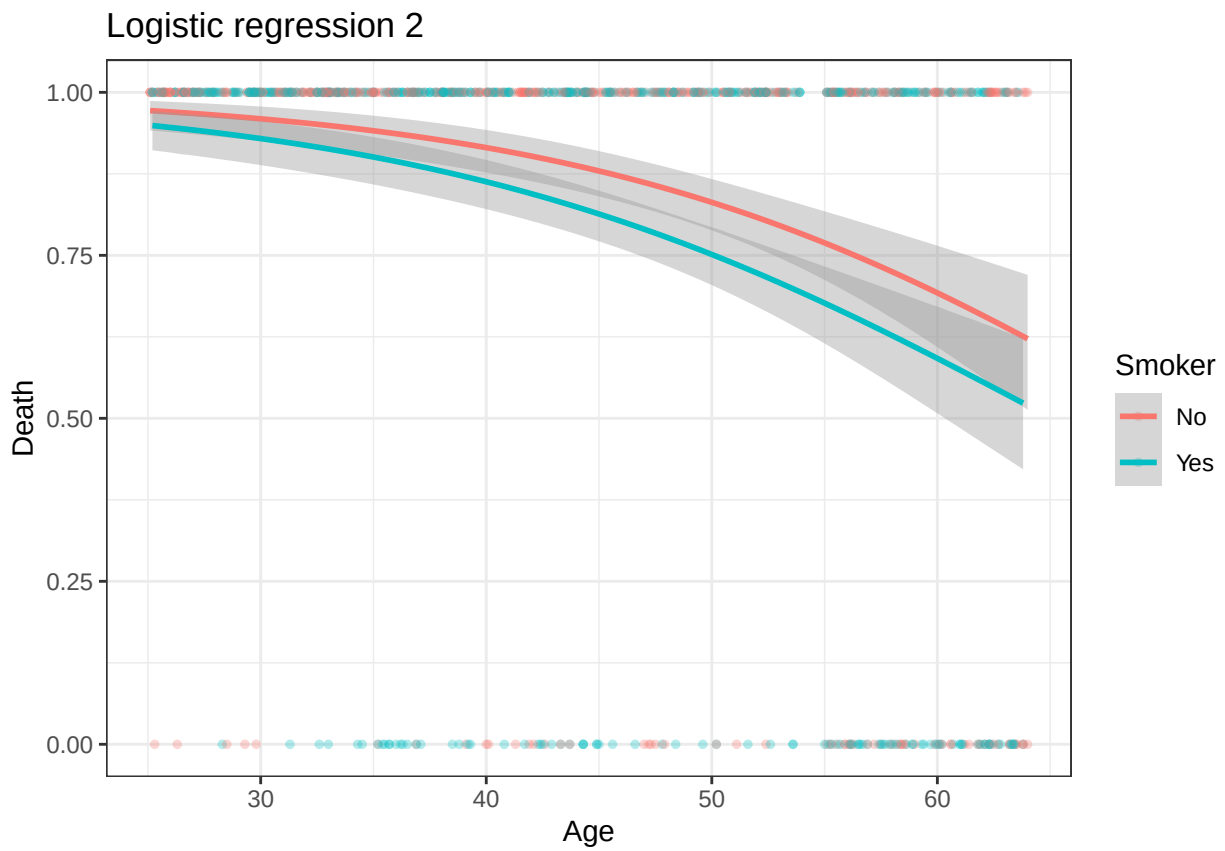
```
##
## Call:
## glm(formula = Death ~ Age + Smoker, family = binomial, data = data2)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.6404   0.3035   0.4508   0.6721   1.1444
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  5.363157   0.476938  11.245  < 2e-16 ***
## Age         -0.075432   0.008929  -8.448  < 2e-16 ***
## SmokerYes   -0.487531   0.186798  -2.610  0.00906 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 841.24  on 896  degrees of freedom
## Residual deviance: 751.42  on 894  degrees of freedom
## AIC: 757.42
##
## Number of Fisher Scoring iterations: 5
```

Ici à la fois l'âge et l'habitude de tabagisme ont un effet négatif significatif. Il semble donc que le fait de fumer affecte la probabilité de décéder pour les femmes qui avaient entre 25 et 65 ans.

Faisons le nouveau graphe :

```
ggplot(data2,aes(x=Age,y=Death,color=Smoker)) + geom_point(alpha=.3,size=1) +
  theme_bw() +
  geom_smooth(method = "glm",
  method.args = list(family = "binomial"))+
    labs(title="Logistic regression 2")
```

```
## `geom_smooth()` using formula 'y ~ x'
```



Sur ce nouveau graphe il apparait que le fait de fumer augmente légèrement la probabilité d'être décédée 20 ans plus tard pour les personnes qui avaient entre 25 et 65 ans.

Conclusion

Les femmes fumeuses qui avaient entre 25 et 65 ans au moment de la première étude semblent avoir une probabilité plus forte de mourir que les femmes non fumeuses. Le fait de fumer augmenterait donc la probabilité de mourir dans les 20 ans si on est d'âge "jeune-moyen". Avec ce jeu de données, il semble donc que le fait de fumer pour les femmes "âgées" ou "très jeunes" n'ait pas d'incidence sur la probabilité de mourir 20 ans plus tard. On peut faire l'interprétation que si la femme est âgée, 20 ans plus tard elle va très probalement mourir qu'elle fume ou non (pour les fumeuses à cause du tabagisme et/ou vieillesse et pour les non fumeuses de vieillesse), et pour les plus jeunes soit elles ne fument pas depuis assez longtemps pour que ça ait une incidence sur leur probabilité de mourir, soit elles sont "protégées"" par leur "jeunesse". Il serait intéressant de faire une étude à intervalles de temps plus rapproché (par exemple tous les 5 ans), pour limiter l'effet de la mort naturelle sur les personnes les plus âgées et en prenant en compte depuis combien de temps la personne fume pour voir si cela a un effet sur la probabilité de mourir.

Le tabagisme semble donc avoir un effet néfaste sur la mortalité qui est dépendant de l'âge.