

Sujet 6 : Autour du Paradoxe de Simpson

Martin DAVY

14 décembre 2021

Contexte

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

Les données sont directement accessibles depuis le gitlab de la formation.

Si les données ne sont pas présentes sur l'ordinateur elles sont automatiquement téléchargées

```
data_nom = "Subject6_smoking.csv"
git_url = "https://gitlab.inria.fr/learninglab/mooc-rr/mooc-rr-ressources/-/raw/master/"
download_path = "module3/Practical_session/Subject6_smoking.csv?inline=false"
data_url = paste0(git_url, download_path)
```

```
# Le fichier existe ?
if(!file.exists(data_nom)) {
  # préciser method = "auto" sinon une colonne NA est rajoutée
  download.file(data_url, data_nom, method = "auto")
}
```

Lecture des données en précisant le paramètre header = TRUE pour conserver le nom des colonnes

```
data = read.csv(data_nom, header = TRUE)
```

Inspection des 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
```

Structure des données

Nous pouvons constater que le tableau est composé de 3 colonnes comme décrit ci-dessous:

Nom de colonne	Libellé de colonne
Smoke	La personne fume (<i>Yes</i>) ou non (<i>No</i>)
Status	La personne est toujours vivante (<i>Alive</i>) ou non (<i>Dead</i>)
Age	Âge de la personne

Vérifions si le tableau contient des valeurs NA

```
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)
```

On peut constater qu'il n'y a aucune valeur NA dans les données.

Analyses

Tabagisme et taux de mortalité

Identification des lignes en fonction des données

```
fumeuses      = which(data$Smoker == "Yes")
non_fumeuses  = which(data$Smoker == "No")

vivantes      = which(data$Status == "Alive")
mortes        = which(data$Status == "Dead")
```

Identification des différents groupes

```
fumeuses_vivantes      = intersect(fumeuses, vivantes)
fumeuses_mortes        = intersect(fumeuses, mortes)
non_fumeuses_vivantes  = intersect(non_fumeuses, vivantes)
non_fumeuses_mortes    = intersect(non_fumeuses, mortes)
```

Création d'un tableau montrant le nombre de personnes vivantes et morte pour chaque groupe (fumeuses et non fumeuses)

```
tab = rbind(c(length(non_fumeuses_mortes), length(non_fumeuses_vivantes)),
            c(length(fumeuses_mortes),      length(fumeuses_vivantes)))
colnames(tab) = c("mortes", "vivantes")
rownames(tab) = c("non_fumeuses", "fumeuses")
tab
```

```
##           mortes vivantes
## non_fumeuses    230    502
## fumeuses        139    443
```

Calcul pour 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).

```
taux_mortalite_non_fumeuses = tab["non_fumeuses", "mortes"] / sum(tab["non_fumeuses",])
taux_mortalite_fumeuses     = tab["fumeuses", "mortes"]      / sum(tab["fumeuses",])
```

```
## Taux de mortalité pour les non fumeuses : 0.31
## Taux de mortalité pour les fumeuses      : 0.24
```

Nous pouvons constater avec surprise que le taux de mortalité chez les non fumeuses est **plus fort** que chez les fumeuses !

L'âge peut-il expliquer ce résultat surprenant ?

Inspection de la colonne âge

```
summary(data$Age)
```

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

Création de catégorie d'âge

```

a_18_34ans = which(18 <= data$Age & data$Age <= 34)
a_34_54ans = which(34 < data$Age & data$Age <= 54)
a_55_64ans = which(54 < data$Age & data$Age <= 64)
a_plus_65ans = which(64 < data$Age)

nom_tranches_ages = c("18-34ans", "34-54ans", "55-64ans", "plus de 65ans")
tranches_ages = list(a_18_34ans, a_34_54ans, a_55_64ans, a_plus_65ans)

```

Calcul des taux de mortalité pour chaque catégories (fumeuses , non fumeuses) et pour chaque tranche d'âge

```

# taux de mortalité fumeuses
tm = NULL
nm = NULL
status = NULL
tas = NULL
for(i in 1:length(tranches_ages))
{
  ta = tranches_ages[[i]]

  # identification des catégories pour chaque tranche d'age
  fumeuses_vivantes_ta = intersect(fumeuses_vivantes, ta)
  fumeuses_mortes_ta = intersect(fumeuses_mortes, ta)
  non_fumeuses_vivantes_ta = intersect(non_fumeuses_vivantes, ta)
  non_fumeuses_mortes_ta = intersect(non_fumeuses_mortes, ta)

  # taux de mortalité fumeuses pour la tranche d'age ta
  nb_f_mortes = length(fumeuses_mortes_ta)
  nb_f = length(fumeuses_vivantes_ta) + length(fumeuses_mortes_ta)
  tmf_ta = nb_f_mortes / nb_f

  # taux de mortalité non fumeuses pour la tranche d'age ta
  nb_nf_mortes = length(non_fumeuses_mortes_ta)
  nb_nf = length(non_fumeuses_vivantes_ta) + length(non_fumeuses_mortes_ta)
  tmnf_ta = nb_nf_mortes / nb_nf

  # -> pour le taux de mortalité
  # stockage de ces taux
  tm = c(tm, tmf_ta, tmnf_ta)

  # stockage des status
  # le premier taux était pour les fumeuses de la tranche d'age donnée et le
  # second pour les non fumeuses
  status = c(status, "fumeuses", "non fumeuses")

  # stockage tranche d'age
  # deux fois la même valeur car on a calculé deux taux de mortalité
  tas = c(tas, nom_tranches_ages[i], nom_tranches_ages[i])

  # pour le nombre de mort
  nm = c(nm, length(fumeuses_mortes_ta), length(non_fumeuses_mortes_ta))
}

# data frame pour le taux de mortalité

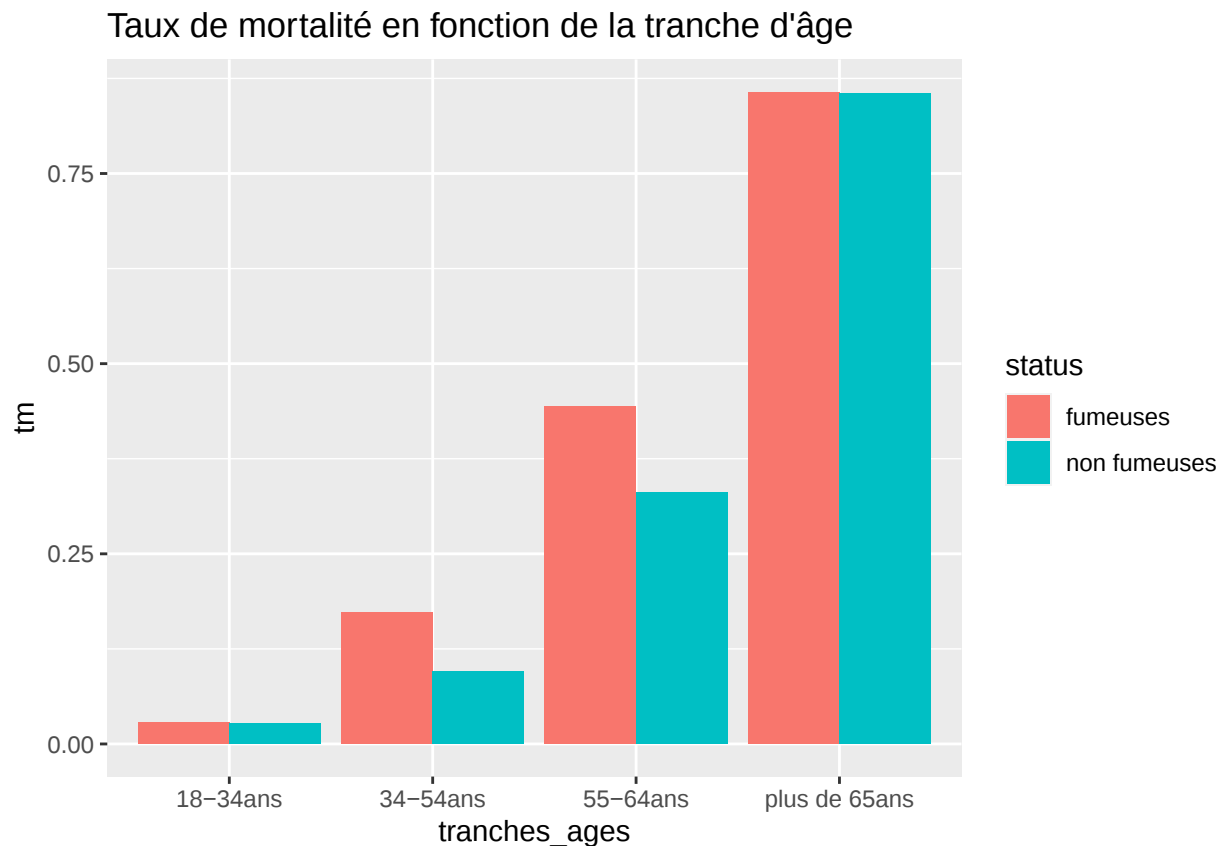
```

```
mortalite_tranche_age = NULL
mortalite_tranche_age = data.frame(tm = tm, quantite = nm, status = status,
                                   tranches_ages = tas)
```

Affichage des taux de mortalité dans un graphe

```
library("ggplot2")
```

```
# Graphe pour le taux de mortalité
ggplot(mortalite_tranche_age,
       aes(x = tranches_ages,
           y = tm,
           fill = status)) +
  geom_bar(stat = "identity",
           position = "dodge") +
  labs(title = "Taux de mortalité en fonction de la tranche d'âge")
```

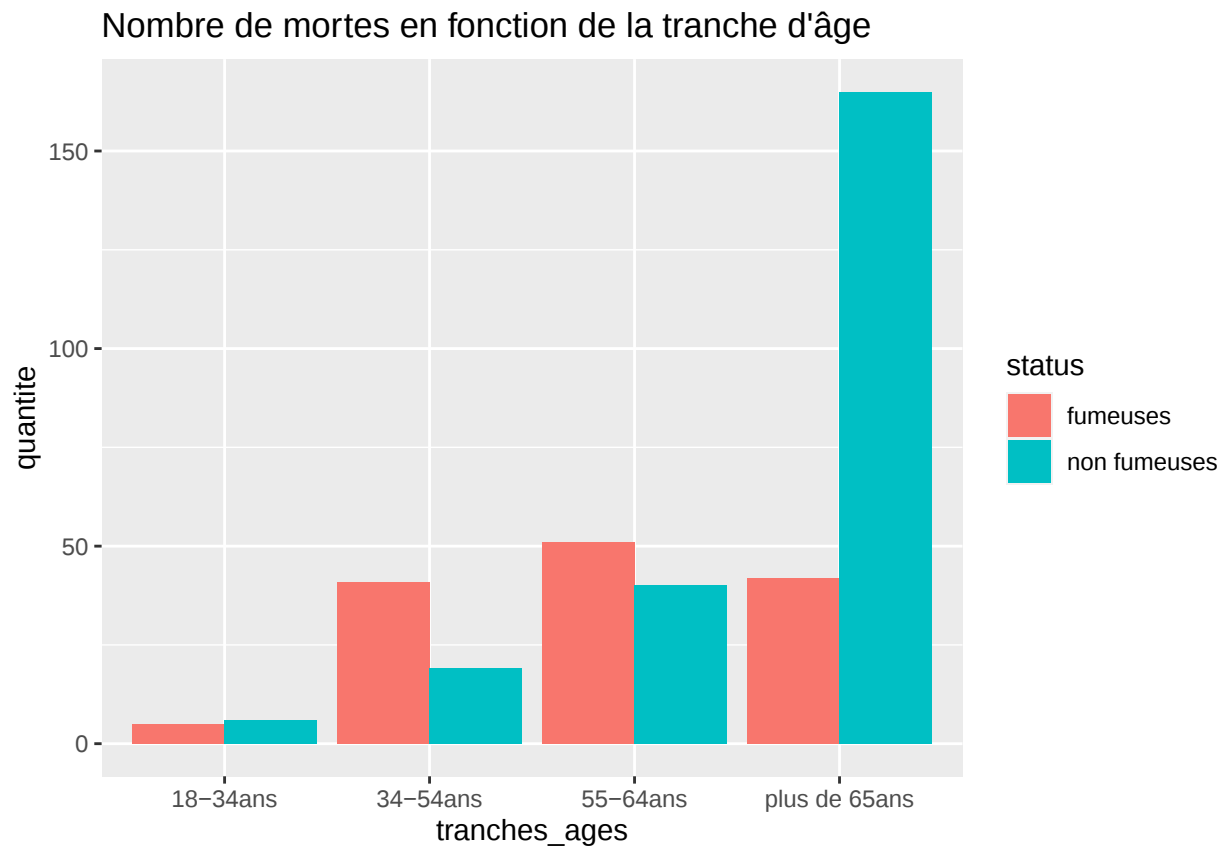


Ce graphe permet de voir que le taux de mortalité sont similaires entre fumeuses et non fumeuses pour la tranche d'âge la plus jeune, où ce taux est très faible, et pour la tranche d'âge la plus âgée avec un taux très élevé.

Dans les tranches intermédiaire le taux de mortalité est plus élevé chez les fumeuses. Il y avait donc un biais causé par l'âge dans la première analyse.

Affichage des nombres de mortes dans un graphe

```
# Graphe pour le nombre de personne mortes
ggplot(mortalite_tranche_age,
      aes(x = tranches_ages,
          y = quantite,
          fill = status)) +
  geom_bar(stat = "identity",
          position = "dodge") +
  labs(title = "Nombre de mortes en fonction de la tranche d'âge")
```



En regardant le nombre de mortes en fonction de l'âge, on peut supposer que la sur-représentation des personnes non-fumeuses mortes à potentiellement induit le biais dû à l'âge.

L'impact de l'âge de sur la mortalité

Création d'une nouvelle colonne

```
data$Death = as.numeric(data$Status == "Dead")
```

Réalisation d'une régression logistique en utilisant la fonction glm et en précisant que la distribution des erreurs suit une loi Binomiale

```
reg_log = glm(Death ~ Age + Smoker, family = binomial(link = logit), data = data)
```

```
summary(reg_log)
```

```
##
## Call:
## glm(formula = Death ~ Age + Smoker, family = binomial(link = logit),
##      data = data)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -2.3129  -0.5947  -0.2830   0.4570   2.9490
##
## 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
```

La régression logistique permet rejeter l'hypothèse affirmant que l'âge n'a pas d'effet sur la variable *Death* car la p-value est inférieure à 0.05 (<2e-16).

Il est cependant impossible de conclure quant à la nocivité du tabagisme étant donnée que la variable *Smoker* n'a pas un effet significatif sur la variable *Death* car la p-value est supérieure à 0.05 (0.0912).