

Autour du paradoxe de Simpson

Arnaud Legrand

3/12/2020

Récupération des données

```
data_url = "https://gitlab.inria.fr/learninglab/mooc-rr/mooc-rr-ressources/-/raw/master/module3/Practicing%20Statistics/data/Smoking.csv"
data_filename = "smoking.csv";

if (!file.exists(data_filename)) {
  download.file(data_url, data_filename)
}
df = read.csv(data_filename);
```

Vérifions si tout à l'air en ordre. A priori, pas de valeurs manquantes, la lecture et le codage des données ont l'air s'être bien passé. On a pour chaque personne, l'indication de si elle fume ou pas, de si elle est décédée durant la période et son âge au début de la période. Les visualisations futures me permettront peut-être de repérer des valeurs étranges mais en attendant, ça ira bien.

```
summary(df)
```

```
##   Smoker      Status      Age
## No :732   Alive:945   Min.    :18.00
## Yes:582   Dead :369   1st Qu.:31.30
##                                     Median :44.80
##                                     Mean    :47.36
##                                     3rd Qu.:60.60
##                                     Max.    :89.90
```

```
str(df)
```

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

```
head(df)
```

```
##   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(df)
```

```
##   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
```

Calcul des effectifs et de la mortalité globale

Pour cela, nous utiliserons les paquets du tidyverse qui sont bien pratiques et très souples. Je vais aussi fixer une palette de couleur pour les différents niveaux des facteurs.

```
# I need Hmisc this to easily compute binary confidence intervals.  
# It has to be loaded before dplyr otherwise this results in some name conflicts.  
library(Hmisc)
```

```
## Loading required package: lattice  
## Loading required package: survival  
## Loading required package: Formula  
## Loading required package: ggplot2  
##  
## Attaching package: 'Hmisc'  
## The following objects are masked from 'package:base':  
##  
##      format.pval, units
```

```
library(dplyr)
```

```
##  
## Attaching package: 'dplyr'  
## The following objects are masked from 'package:Hmisc':  
##  
##      src, summarize  
## The following objects are masked from 'package:stats':  
##  
##      filter, lag  
## The following objects are masked from 'package:base':  
##  
##      intersect, setdiff, setequal, union  
library(ggplot2)  
mystyle = list(theme_classic(),scale_fill_manual(  
  values = c("Alive"="#377EB8", "Dead"= "#E41A1C", "No"="#4DAF4A", "Yes"= "#E41A1C")))
```

Commençons, comme demandé, par calculer les effectifs et la mortalité selon le tabagisme des individus.

```
df %>% group_by(Smoker,Status) %>% summarize(Number=n()) -> df_overview  
df_overview %>% tidyr::spread(Status,Number) %>%  
  mutate(Mortality = 100*Dead/(Alive+Dead)) -> df_mortality_overview  
df_mortality_overview
```

```
## # A tibble: 2 x 4  
## # Groups:   Smoker [2]  
##   Smoker Alive  Dead Mortality  
##   <fct> <int> <int>    <dbl>  
## 1 No      502   230     31.4  
## 2 Yes     443   139     23.9
```

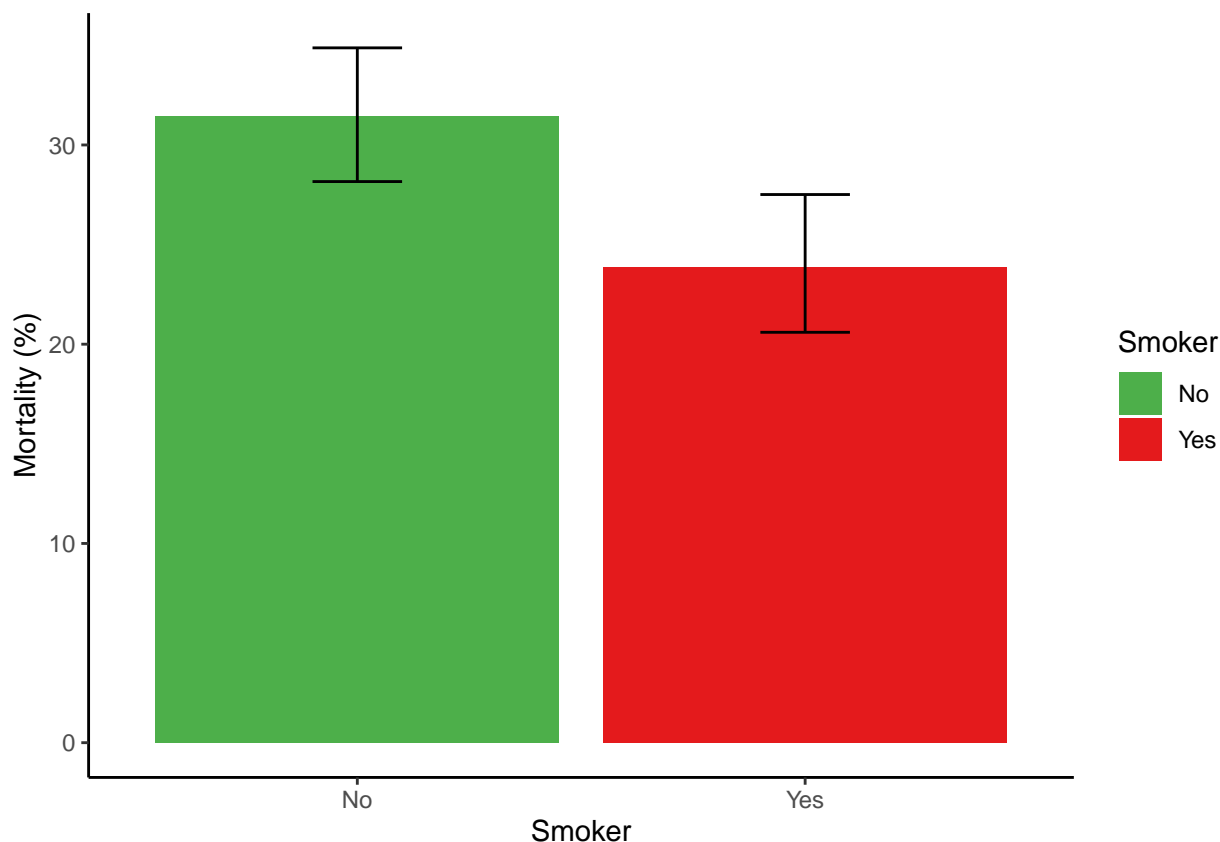
Tiens, oui, surprenant en effet que la mortalité soit supérieure chez les non fumeurs. Ça va un peu à l'encontre de ce à quoi on pourrait s'attendre. Calculons la confiance naïvement.

```
df_mortality_overview %>% group_by(Smoker) %>%
  do(data.frame(., Conf = Hmisc::binconf(.$Dead, .$Alive+.$Dead,
                                         alpha = 0.05, method = "wilson"))) ->
  df_mortality_overview
df_mortality_overview
```

```
## # A tibble: 2 x 7
## # Groups:   Smoker [2]
##   Smoker Alive Dead Mortality Conf.PointEst Conf.Lower Conf.Upper
##   <fct> <int> <int>     <dbl>      <dbl>      <dbl>      <dbl>
## 1 No      502  230     31.4      0.314      0.282      0.349
## 2 Yes     443  139     23.9      0.239      0.206      0.275
```

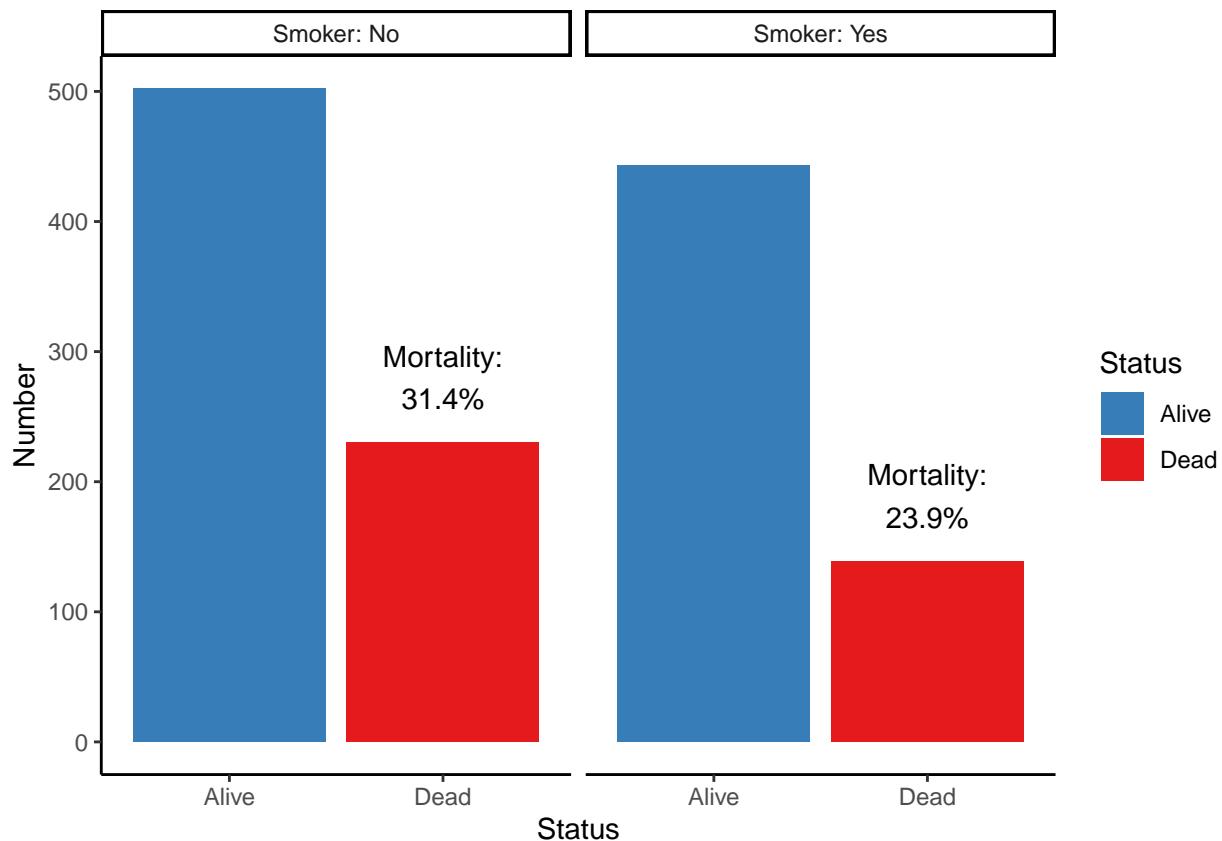
Une petite représentation graphique avec uniquement la mortalité.

```
ggplot(df_mortality_overview, aes(x=Smoker,y=Conf.PointEst*100)) + geom_bar(aes(fill=Smoker), stat = "identity") +
  geom_errorbar(aes(ymin=Conf.Lower*100,ymax=Conf.Upper*100),width=.2) +
  ylab("Mortality (%)") + mystyle
```



Une petite représentation graphique avec les effectifs:

```
ggplot(df_mortality_overview %>% select(Smoker, Alive, Dead) %>% tidyr::gather(key=Status,value=Number,
  aes(x=Status,y=Number)) + geom_bar(aes(fill=Status), stat = "identity", position=position_dodge(
  geom_text(data=df_mortality_overview, aes(x="Dead",y=Dead+50, label=paste0("Mortality:\n",round(Mor
  facet_wrap(~Smoker, labeller = label_both) +
  mystyle
```



Utilisation des catégories d'âges

On nous propose de définir des catégories d'âges. Allons-y:

```
df %>% mutate(Age_Cat = case_when(
  Age >=18 & Age <34 ~ "18-34",
  Age >=34 & Age <55 ~ "34-54",
  Age >=55 & Age <65 ~ "55-64",
  Age >=65 ~ "65+")) %>%
  mutate(Age_Cat = recode(as.factor(Age_Cat),
    "0" = "18-34", "1" = "34-54",
    "2" = "55-64", "3" = "65+")) -> df
```

Puis recalculons les effectifs et la mortalité comme précédemment:

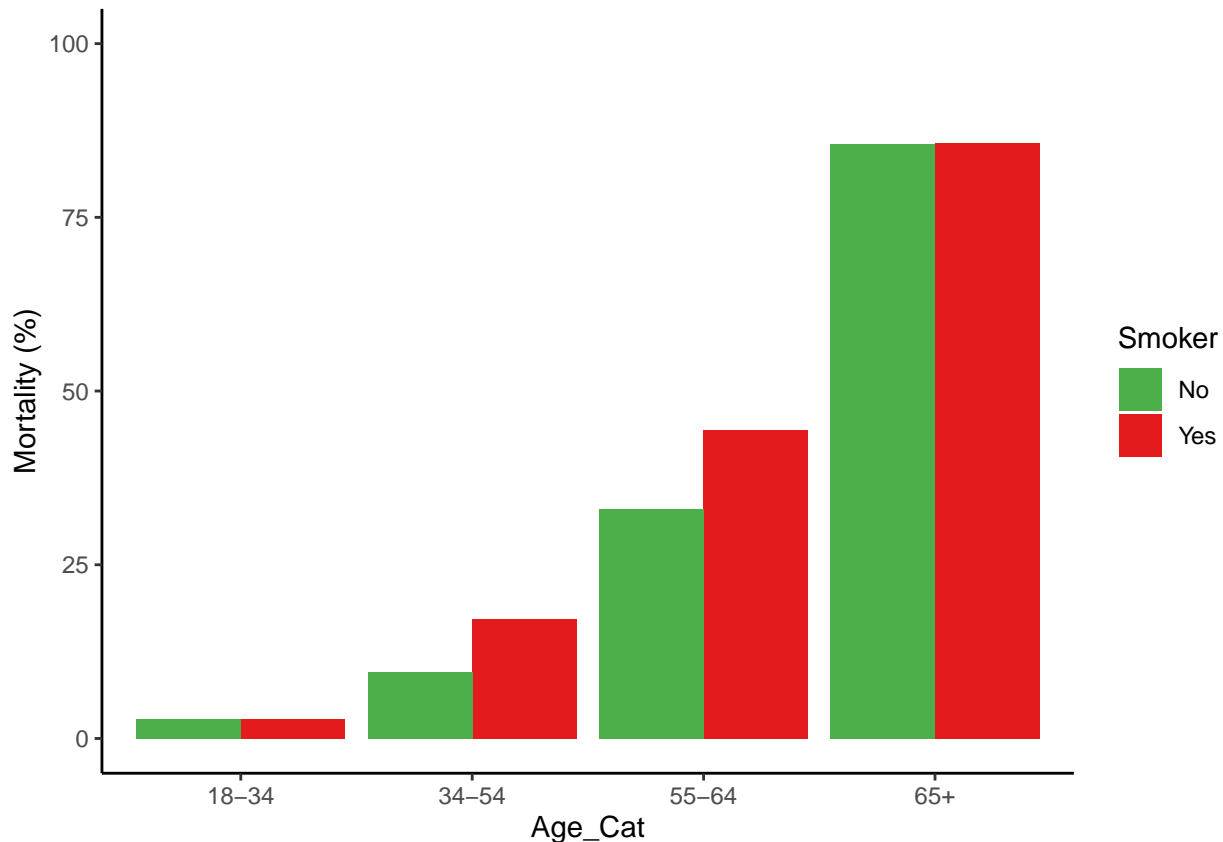
```
df %>% group_by(Smoker, Status, Age_Cat) %>% summarize(Number=n()) %>%
  tidyr::spread(Status, Number) %>% mutate(Mortality = 100*Dead/(Alive+Dead)) -> df_grouped
df_grouped %>% arrange(Age_Cat)
```

```
## # A tibble: 8 x 5
## # Groups:   Smoker [2]
##   Smoker Age_Cat Alive  Dead Mortality
##   <fct> <fct>   <int> <int>    <dbl>
## 1 No    18-34     213     6     2.74
## 2 Yes   18-34     174     5     2.79
## 3 No    34-54     180    19     9.55
## 4 Yes   34-54     198    41    17.2
## 5 No    55-64     81    40    33.1
```

```
## 6 Yes    55-64      64    51    44.3
## 7 No     65+      28   165    85.5
## 8 Yes    65+       7    42    85.7
```

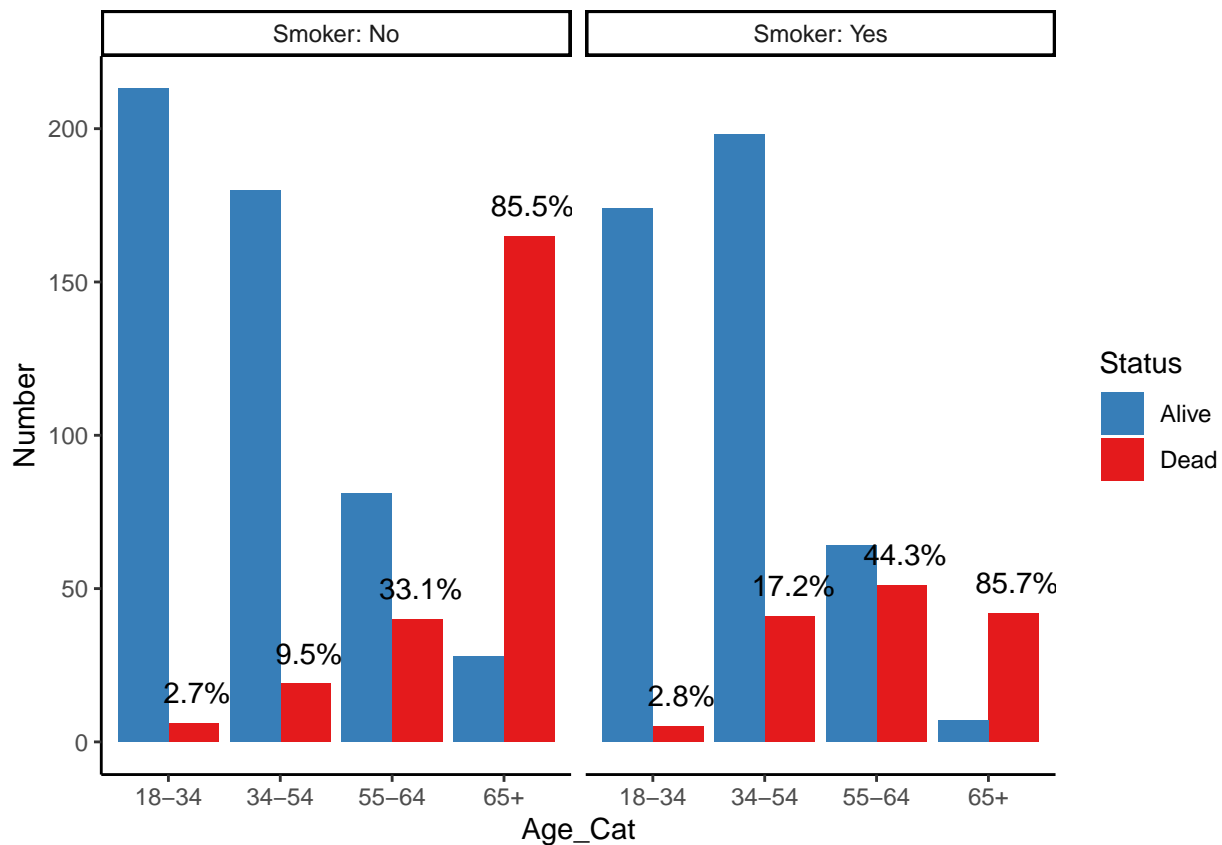
Une petite représentation graphique de la mortalité. Effectivement il est paradoxal que la mortalité soit globalement supérieure pour les non fumeurs mais que pour **chaque** catégorie d'âge, elle soit inférieure!

```
ggplot(df_grouped, aes(x=Age_Cat,y=Mortality,fill=Smoker)) +
  geom_bar(stat="identity", position=position_dodge()) +
  ylab("Mortality (%)") + ylim(0,100) + mystyle
```



Faisons une autre visualisation, plus explicite sur les effectifs et permettant de commencer à comprendre d'où peut venir le problème. En gros, pas de différence de mortalité significative pour les très jeunes et les très vieux mais une différence très importante pour les gens d'âge moyen. Par contre très peu de vieux fumeurs (de vieilles fumeuses en l'occurrence). Fumer tue effectivement (on meurt plus tôt) mais les effectifs trompent les ratios.

```
ggplot(df %>% group_by(Smoker,Status,Age_Cat) %>% summarize(Number=n()),
  aes(x=Age_Cat,y=Number)) +
  geom_bar(aes(fill=Status), stat = "identity",position=position_dodge()) +
  geom_text(data=df_grouped, aes(x=Age_Cat,y=Dead+10,
                                label=paste0("      ",round(Mortality,1),"%")) +
  facet_wrap(~Smoker, labeller = label_both) +
  mystyle
```



Régression logistique

Tentons une régression logistique qui ne devrait pas être impactée par ces catégorisations arbitraires d'âge.

```
summary(glm(data=df, Status ~ Age * Smoker, family=binomial(link='logit')))
```

```
##
## Call:
## glm(formula = Status ~ Age * Smoker, family = binomial(link = "logit"),
##      data = df)
##
## 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
```

Ah, on voit un petit effet mais ce n'est pas vraiment significatif (pour Smoker... évidemment que Age est significatif, plus on vieillit plus on a de chances de mourrir!).

Essayons à tout hasard une régression pour chaque catégorie on pourra voir où les courbes de régression s'intersectent en prenant en compte la confiance

```
summary(glm(data=df %>% filter(Smoker=="Yes"),
            Status ~ Age, family=binomial(link='logit')))
```

```
##
## Call:
## glm(formula = Status ~ Age, family = binomial(link = "logit"),
##      data = df %>% filter(Smoker == "Yes"))
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -2.0745  -0.6464  -0.3756  -0.2013   2.6560
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -5.508106   0.466221  -11.81  <2e-16 ***
## Age          0.088977   0.008721   10.20  <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: 639.89 on 581 degrees of freedom
## Residual deviance: 480.41 on 580 degrees of freedom
## AIC: 484.41
##
## Number of Fisher Scoring iterations: 5
```

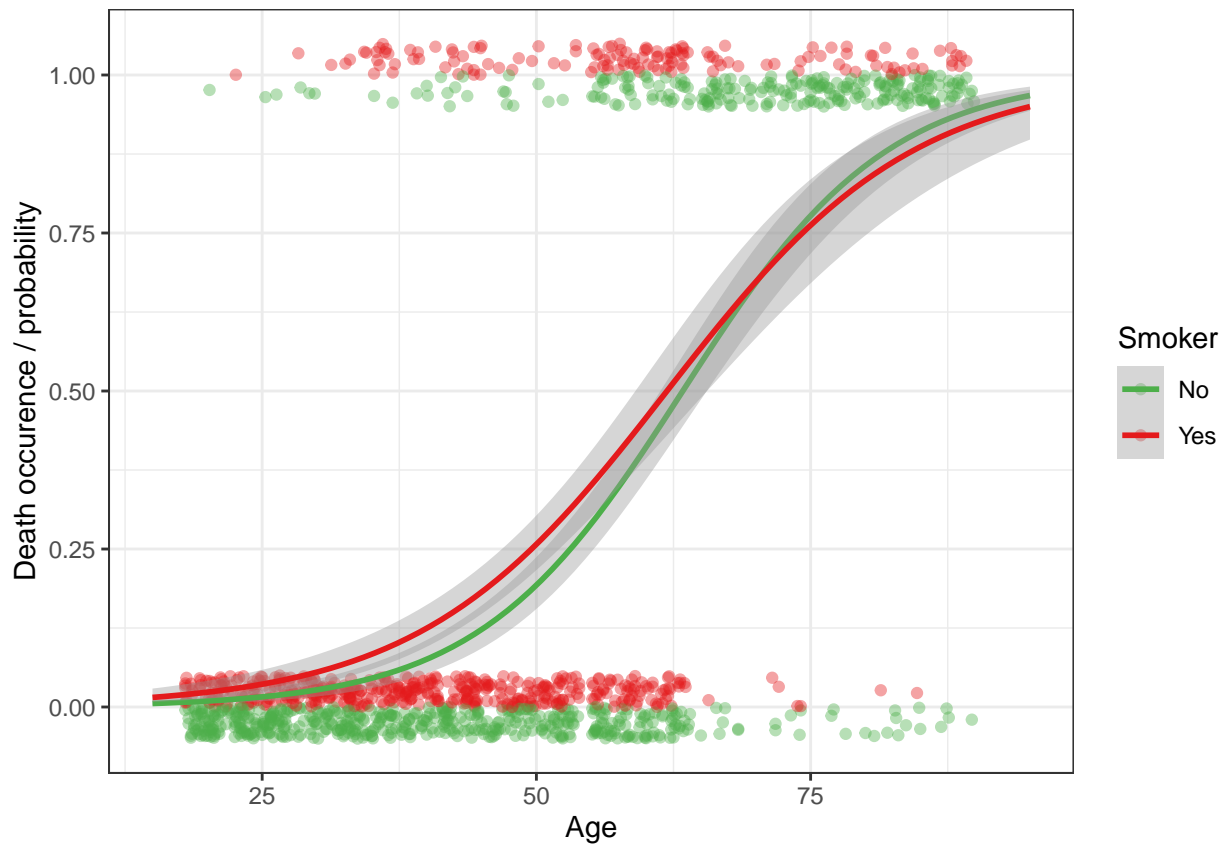
```
summary(glm(data=df %>% filter(Smoker=="No"),
            Status ~ Age, family=binomial(link='logit')))
```

```
##
## Call:
## glm(formula = Status ~ Age, family = binomial(link = "logit"),
##      data = df %>% filter(Smoker == "No"))
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -2.4019  -0.5179  -0.2003   0.4728   3.0457
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -6.795507   0.479430  -14.17  <2e-16 ***
## Age          0.107275   0.007806   13.74  <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: 911.23 on 731 degrees of freedom
## Residual deviance: 519.08 on 730 degrees of freedom
## AIC: 523.08
##
## Number of Fisher Scoring iterations: 6
```

Bon, pas facile à lire mais graphiquement, ça devrait être plus explicite. Pour éviter l'*overplotting*, je met un peu de jitter vertical sur mes points (manuellement plutôt qu'avec `geom_jitter` afin de séparer les *Smoker*).

```
df %>% mutate(Status_num=ifelse(Status=="Dead",1,0),
               y = Status_num + (as.numeric(as.factor(Smoker))-2)*0.05 +
                 0.05*runif(n())) -> df_raw
ggplot(df_raw, aes(x=Age, y=Status_num, color=Smoker)) +
  geom_point(alpha=.4, aes(y=y)) +
  geom_smooth(method="glm", method.args = list(family = "binomial"),fullrange = TRUE) +
  theme_bw() + xlim(15,95) + ylab("Death occurence / probability") +
  scale_color_manual(values = c("No"="#4DAF4A", "Yes"= "#E41A1C"))
```



On voit nettement le décalage entre les deux courbes même si les régions de confiances s'interceptent. Avec des effectifs plus importants, la séparation serait nette. Le point important n'est pas la mortalité après 20 ans mais l'âge auquel on meurt...