Analyse autour du paradoxe de Simon

Aurélie Zaros

Préparation des données

Le jeu de données a directement été téléchargé sur le lien donné. Le fichier est sous forme CSV.

```
data_file = "TP.csv"
```

Voici l'explication des colonnes donnée :

Nom de colonne	Libellé de colonne
`Smoker`	Fumeur (YES) ou non fumeur (NO)
`Status`	Vivant (ALIVE) ou mort (DEAD)
`Age`	Age

Téléchargement du fichier

```
data = read.csv(data_file)
```

Il ne manque aucunes données dans notre fichier :

```
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)
```

Les classes des données sont correctes :

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

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

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

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

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

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

Première question

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

Afin d'avoir une répresentation dans un tableau, on émet un tableau qui nous donne le nombre de femmes dans chaque condition.

```r
library(MASS)
tbl = table(data$Smoker, data$Status)
tbl
```

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

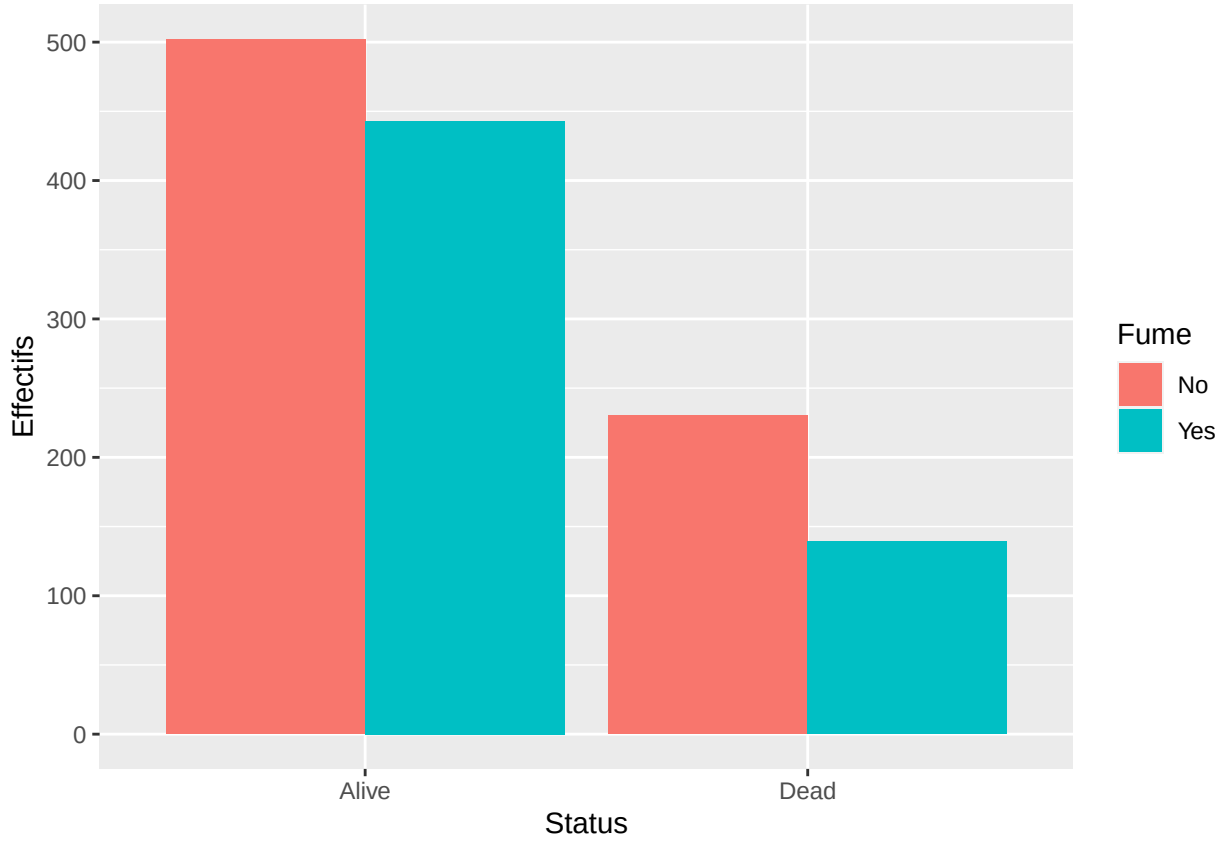
```r
library(gmodels)
```

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

```r
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 3.6.3
```

```r
ggplot(data) +
  aes(x = Status, fill = Smoker) +
  geom_bar(position = "dodge") +
  xlab("Status") +
  ylab("Effectifs") +
  labs(fill = "Fume")
```



Donc, dans notre échantillon, il y a 502 femmes non-fumeuses en vie, 443 femmes fumeuses en vie, 230 femmes non-fumeuses mortes et 139 femmes fumeuses mortes.

b. Calculez dans chaque groupe (fumeuses/ non fumeuses) le taux de mortalité (le rapport entre le nombre de femmes décédées dans un groupe avec le nombre total de femmes dans ce groupe).

Afin de connaitre le taux de mortalité dans chaque groupe de femme (fumeuse ou non), nous faisons une analyse avec un khi² :

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

khi_SmokerStatus <- chisq.test(SmokerStatus,
                               correct = FALSE)

library(gmodels)
library(ggplot2)
CrossTable(SmokerStatus,
           fisher = TRUE,
           chisq = TRUE,
           expected = TRUE,
           sresid = TRUE,
           format = 'SPSS')
```

```
##
##    Cell Contents
## |-------------------------|
```

```
## |                   Count |
## |         Expected Values |
## | Chi-square contribution |
## |             Row Percent |
## |          Column Percent |
## |           Total Percent |
## |            Std Residual |
## |-------------------------|
##
## Total Observations in Table:  1314
##
##              |
##              |    Alive  |     Dead  | Row Total |
## -------------|-----------|-----------|-----------|
##           No |      502  |      230  |      732  |
##              |  526.438  |  205.562  |           |
##              |    1.134  |    2.905  |           |
##              |   68.579% |   31.421% |   55.708% |
##              |   53.122% |   62.331% |           |
##              |   38.204% |   17.504% |           |
##              |   -1.065  |    1.705  |           |
## -------------|-----------|-----------|-----------|
##          Yes |      443  |      139  |      582  |
##              |  418.562  |  163.438  |           |
##              |    1.427  |    3.654  |           |
##              |   76.117% |   23.883% |   44.292% |
##              |   46.878% |   37.669% |           |
##              |   33.714% |   10.578% |           |
##              |    1.195  |   -1.912  |           |
## -------------|-----------|-----------|-----------|
## Column Total |      945  |      369  |     1314  |
##              |   71.918% |   28.082% |           |
## -------------|-----------|-----------|-----------|
##
##
## Statistics for All Table Factors
##
##
## Pearson's Chi-squared test
## ------------------------------------------------------------
## Chi^2 =  9.120903     d.f. =  1     p =  0.002527052
##
## Pearson's Chi-squared test with Yates' continuity correction
## ------------------------------------------------------------
## Chi^2 =  8.7515     d.f. =  1     p =  0.003093475
##
##
## Fisher's Exact Test for Count Data
## ------------------------------------------------------------
## Sample estimate odds ratio:  0.6850392
##
## Alternative hypothesis: true odds ratio is not equal to 1
## p =  0.002988803
## 95% confidence interval:  0.5307485 0.8822128
```

```
##
## Alternative hypothesis: true odds ratio is less than 1
## p =  0.001492048
## 95% confidence interval:  0 0.848289
##
## Alternative hypothesis: true odds ratio is greater than 1
## p =  0.9990132
## 95% confidence interval:  0.5524562 Inf
##
##
##
##        Minimum expected frequency: 163.4384
```

Afin de connaitre le taux de mortalité dans chaque groupe de femme, nous nous basons sur le pourcentage brute (Row Percent). Nous avons donc un taux de mortalité de 31% chez les femmes non fumeuses et de 23% chez les fumeuses. Le taux de mortalité est supérieur chez les femmes non-fumeuses.

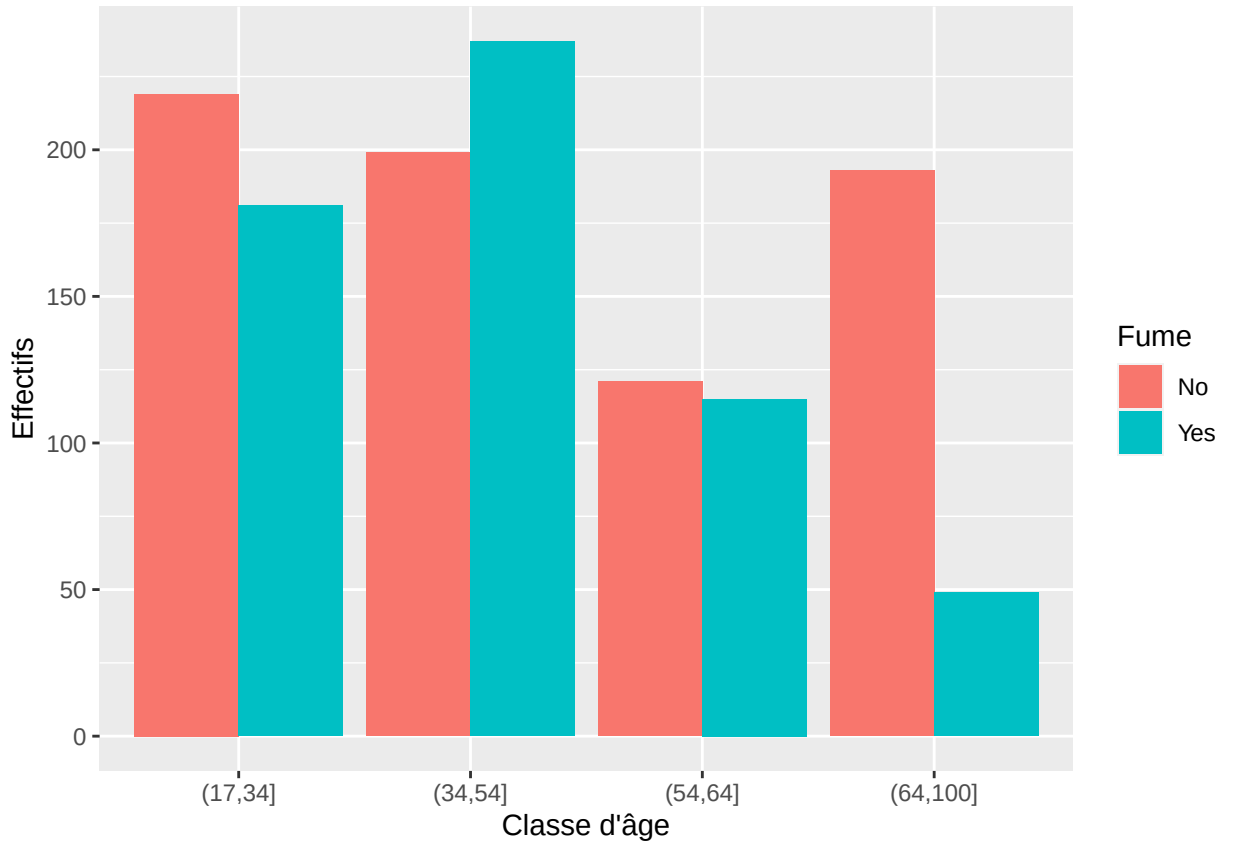
Deuxième question

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.

```
data$Classe <- cut(data$Age, c(17, 34, 54 , 64 , 100))
table(data$Classe)
```

```
##
##  (17,34]  (34,54]  (54,64] (64,100]
##      400      436      236      242
```

```
ggplot(data) +
  aes(x = Classe, fill = Smoker) +
  geom_bar(position = "dodge") +
  xlab("Classe d'âge") +
  ylab("Effectifs") +
  labs(fill = "Fume")
```



Le paradoxe de Simpson est un paradoxe statistique dans lequel un phénomène observé de plusieurs groupes semble s'inverser lorsque les groupes sont combinés. Un des raisons possibles qu'un paradoxe de Simpson s'applique à des données est la présence d'un **facteur de confusion** : il s'agit d'une *variable qui va avoir une influence à la fois sur la cause observée et l'effet observé.*

Troisème question

Afin d'éviter un biais induis 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 de conclure ou pas sur la nocivité du tabagisme ? Vous pourrez proposer une représentation graphique de ces régressions (en n'omettant pas les régions de confiance).

D'abord on recode la variable Status en Death :

```
library(forcats)
```

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

```
data$Death_binomiale <- fct_recode(data$Status,
                "1" = "Alive",
                "0" = "Dead")
table(data$Death_binomiale)
```

```
##
##   1   0
## 945 369
```

Notre variable a bien été recodée. Dans notre échantillons, 945 femmes sont en vie contre 369 sont décédées.

Ensuite, on analyse 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 :

```
reg <- glm(Age ~ Death_binomiale+Smoker,
  data = data)
summary(reg)
```

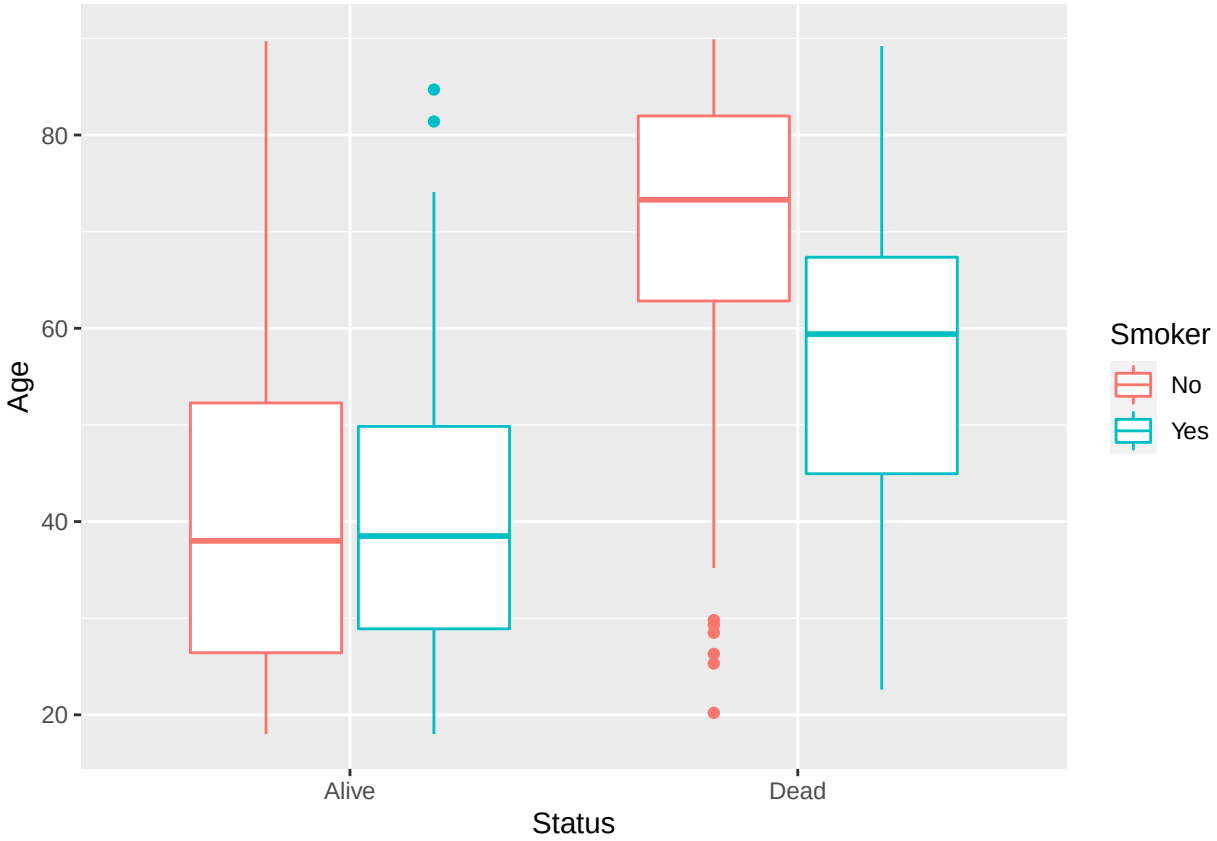
```
##
## Call:
## glm(formula = Age ~ Death_binomiale + Smoker, data = data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -47.312  -11.310   -0.661   11.564   47.992
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)        41.7081     0.6276  66.461  < 2e-16 ***
## Death_binomiale0   25.8037     0.9265  27.849  < 2e-16 ***
## SmokerYes          -3.6011     0.8383  -4.296 1.87e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 226.2407)
##
##     Null deviance: 482043  on 1313  degrees of freedom
## Residual deviance: 296602  on 1311  degrees of freedom
## AIC: 10858
##
## Number of Fisher Scoring iterations: 2
```

```
exp(coefficients(reg))
```

```
##      (Intercept) Death_binomiale0        SmokerYes
##     1.299017e+18     1.608376e+11     2.729341e-02
```

Conclusion :

```
options(digits = 2, scipen = 999)
ggplot(data, aes(x=Status, y=Age, color=Smoker)) +
  geom_boxplot() +
  theme(legend.position = "right")
```



Il semble qu'il n'y ait pas d'impact du tabagisme sur la longévité.