

Autour du Paradoxe de Simpson

Jade

November 10, 2024

Contents

1	Importation des données	2
2	Vérification de la validité des données	3
3	Analyse des taux de mortalité selon les catégories fumeuses/non-fumeuses	4
3.1	Effectifs fumeuses/non-fumeuses	4
3.2	Effectif des femmes vivantes/mortes selon leur catégorie . . .	5
3.3	Taux de mortalité par catégorie	8
4	Analyse des taux de mortalité selon les catégories fumeuses/non-fumeuses avec la notion d'âge	9
4.1	Effectifs fumeuses/non-fumeuses par classe d'âge	9
4.2	Effectif des femmes vivantes/mortes par classe d'âge selon leur catégorie	12
4.3	Taux de mortalité par classe d'âge selon les catégories	23
5	Conclusion sur ces deux analyses	26

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.

1 Importation des données

Nous commençons par charger les données stockées dans le fichier csv (placé dans le même répertoire). Le chemin d'accès ci-dessous sera à modifier si vous souhaitez vérifier les opérations que j'ai pu faire avec le fichier.

```
data = read.csv('Subject6_smoking.csv')
head(data)
tail(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
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
```

Le jeu de données nous indique si la personne interrogée est une fumeuse ou non lors du premier sondage, si elle est en vie 20 ans plus tard, et puis son âge lors du premier sondage. On a également l'air de bien avoir les 1314 résultats attendus.

2 Vérification de la validité des données

Nous allons d'abord vérifier que chaque ligne du fichier est bien remplie

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

```
[1] Smoker Status Age
<0 lignes> (ou 'row.names' de longueur nulle)
```

Aucune des lignes n'est vide.

Nous allons maintenant vérifier le type des valeurs par colonnes, pour être sûr qu'elles soient du bon type et du même type (par colonnes)

```
class(data$Smoker)
class(data$Status)
class(data$Age)
```

```
[1] "character"
[1] "character"
[1] "numeric"
```

Tout semble bon.

Puis enfin nous allons vérifier que nous ayons bien le nombre attendu de réponses au sondage, c'est-à-dire 1314

```
number_rows = nrow(data)
number_rows
```

```
[1] 1314
```

Le compte est bon ! On peut maintenant s'attaquer au vif du sujet

3 Analyse des taux de mortalité selon les catégories fumeuses/non-fumeuses

3.1 Effectifs fumeuses/non-fumeuses

Nous allons commencer par séparer les données initiales dans deux tableaux différents : les fumeuses séparées des non-fumeuses

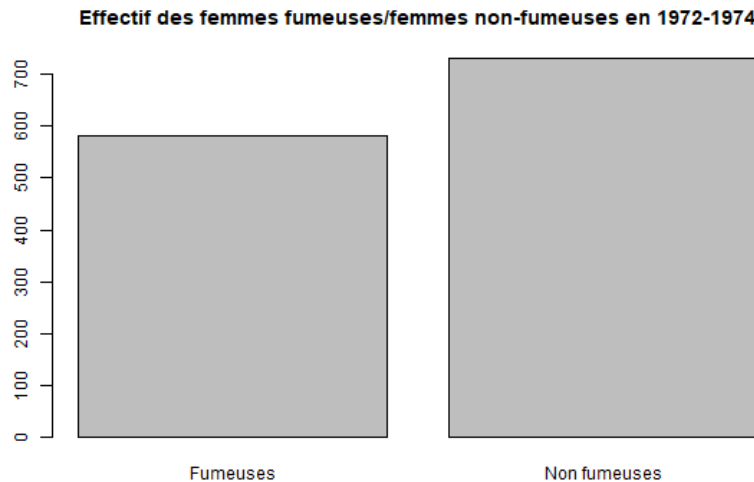
```
smokers <- subset(data, data$Smoker == 'Yes')
not_smokers <- subset(data, data$Smoker == 'No')
head(smokers)
head(not_smokers)
```

```
Smoker Status Age
1      Yes  Alive 21.0
2      Yes  Alive 19.3
5      Yes  Alive 81.4
8      Yes   Dead 57.5
9      Yes  Alive 24.8
10     Yes  Alive 49.5
Smoker Status Age
3      No   Dead 57.5
4      No  Alive 47.1
6      No  Alive 36.8
7      No  Alive 23.8
12     No   Dead 66.0
14     No  Alive 58.4
```

Les échantillons des deux tableaux semblent corrects.

Nous allons maintenant regarder les effectifs de manière graphique, par un diagramme en barre

```
x = c(nrow(smokers), nrow(not_smokers))
type = c("Fumeuses", "Non fumeuses")
barplot(x, names.arg=type, main="Effectif des femmes fumeuses/femmes non-fumeuses en 1972-1974")
```



On peut voir que l'effectif des femmes non-fumeuses est supérieur à celui des femmes fumeuses même si l'écart n'a pas l'air si important.

3.2 Effectif des femmes vivantes/mortes selon leur catégorie

Nous allons maintenant séparer les données réduites à nouveau dans deux tableaux différents : les fumeuses vivantes 20 ans plus tard séparées des fumeuses mortes (resp. non-fumeuses)

```
smokers_alive <- subset(smokers, smokers$Status == 'Alive')
smokers_dead <- subset(smokers, smokers$Status == 'Dead')
```

```
head(smokers_alive)
head(smokers_dead)
```

```
Smoker Status Age
1    Yes  Alive 21.0
2    Yes  Alive 19.3
5    Yes  Alive 81.4
9    Yes  Alive 24.8
10   Yes  Alive 49.5
11   Yes  Alive 30.0
Smoker Status Age
```

8	Yes	Dead	57.5
24	Yes	Dead	62.3
39	Yes	Dead	33.0
47	Yes	Dead	44.3
64	Yes	Dead	36.3
65	Yes	Dead	80.7

Les échantillons des deux tableaux semblent corrects.

```
not_smokers_alive <- subset(not_smokers, not_smokers$Status == 'Alive')
not_smokers_dead <- subset(not_smokers, not_smokers$Status == 'Dead')
```

```
head(not_smokers_dead)
head(not_smokers_alive)
```

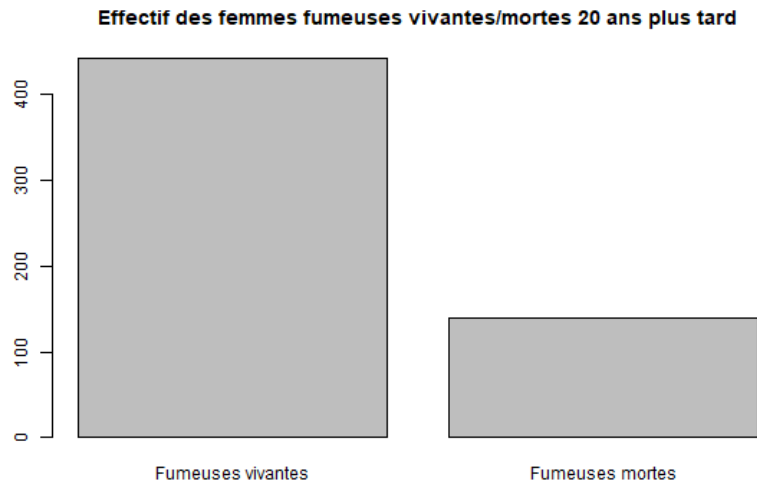
Smoker	Status	Age
3	No	Dead 57.5
12	No	Dead 66.0
15	No	Dead 60.6
21	No	Dead 73.2
29	No	Dead 36.9
42	No	Dead 69.7

Smoker	Status	Age
4	No	Alive 47.1
6	No	Alive 36.8
7	No	Alive 23.8
14	No	Alive 58.4
16	No	Alive 25.1
17	No	Alive 43.5

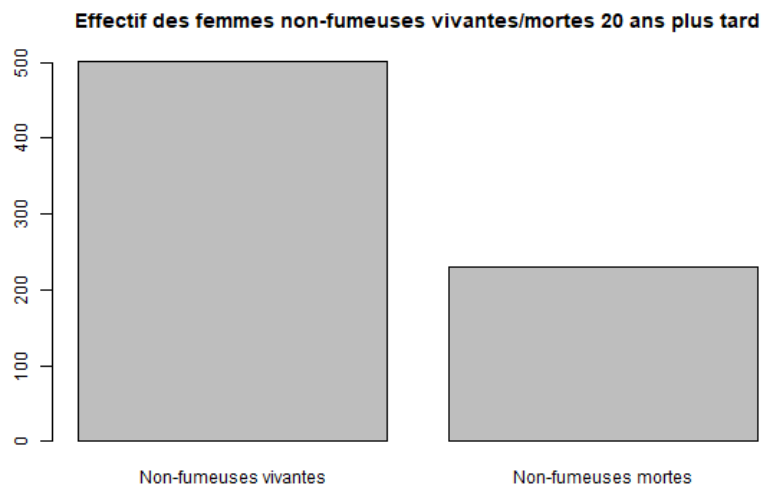
De même pour les non-fumeuses.

Nous allons maintenant de nouveau regarder les effectifs de manière graphique, par un diagramme en barre pour les deux catégories

```
x2 = c(nrow(smokers_alive), nrow(smokers_dead))
type2 = c("Fumeuses vivantes", "Fumeuses mortes")
barplot(x2, names.arg = type2, main = "Effectif des femmes fumeuses vivantes/mortes
20 ans plus tard")
```



```
x3 = c(nrow(not_smokers_alive),nrow(not_smokers_dead))
type3 = c("Non-fumeuses vivantes", "Non-fumeuses mortes")
barplot(x3,names.arg=type3,main="Effectif des femmes non-fumeuses vivantes/mortes
20 ans plus tard")
```



On remarque déjà visuellement qu'il y a plus de femmes non-fumeuses mortes 20 ans plus tard que de femmes fumeuses, mais on ne peut rien

conclure puisque nous avons vu qu'il y avait plus de femmes non-fumeuses dans l'ensemble de départ que de fumeuses.

3.3 Taux de mortalité par catégorie

Il ne reste plus qu'à calculer le taux de mortalité pour chaque groupe :

```
eff_dead_smokers = nrow(smokers_dead)
eff_smokers = nrow(smokers)
taux_mortalite_smokers = eff_dead_smokers/eff_smokers

eff_dead_smokers
eff_smokers
taux_mortalite_smokers

[1] 139
[1] 582
[1] 0.2388316

eff_dead_not_smokers = nrow(not_smokers_dead)
eff_not_smokers = nrow(not_smokers)
taux_mortalite_not_smokers = eff_dead_not_smokers/eff_not_smokers

eff_dead_not_smokers
eff_not_smokers
taux_mortalite_not_smokers

[1] 230
[1] 732
[1] 0.3142077
```

Le taux de mortalité est plus élevé chez les femmes non-fumeuses. Le résultat peut paraître surprenant dû aux problèmes de santé liés au tabagisme. On pourrait s'attendre à ce que les résultats soient orientés vers le groupe des fumeuses.

4 Analyse des taux de mortalité selon les catégories fumeuses/non-fumeuses avec la notion d'âge

Nous allons poursuivre l'analyse des données précédentes en les séparants par classe d'âge : de 18 à 34 ans (exclu), de 34 à 54 ans (exclu), de 54 à 65 ans (exclu) et plus de 65 ans.

4.1 Effectifs fumeuses/non-fumeuses par classe d'âge

Les étapes restent similaires à ce qu'on a pu faire précédemment : On va commencer par séparer le tableau des fumeuses selon les classes d'âge fixées :

```
smokers_18_34 <- subset(smokers, smokers$Age >= 18.0 &
smokers$Age < 34.0)
smokers_34_54 <- subset(smokers, smokers$Age >= 34.0 &
smokers$Age < 54.0)
smokers_54_64 <- subset(smokers, smokers$Age >= 54.0 &
smokers$Age < 65.0)
smokers_65 <- subset(smokers, smokers$Age >= 65.0)

head(smokers_18_34)
head(smokers_34_54)
head(smokers_54_64)
head(smokers_65)
```

```
Smoker Status Age
1      Yes  Alive 21.0
2      Yes  Alive 19.3
9      Yes  Alive 24.8
11     Yes  Alive 30.0
38     Yes  Alive 29.5
39     Yes   Dead 33.0
Smoker Status Age
10     Yes  Alive 49.5
13     Yes  Alive 49.2
22     Yes  Alive 38.3
31     Yes  Alive 34.6
32     Yes  Alive 51.9
33     Yes  Alive 49.9
```

```

Smoker Status Age
8      Yes   Dead 57.5
24     Yes   Dead 62.3
27     Yes   Alive 59.2
61     Yes   Alive 58.1
84     Yes   Alive 58.3
91     Yes   Alive 56.1
Smoker Status Age
5       Yes   Alive 81.4
20      Yes   Alive 65.7
65      Yes   Dead 80.7
113     Yes   Dead 66.5
130     Yes   Dead 87.8
137     Yes   Dead 71.7

```

Les sorties semblent cohérentes.

Les mêmes opérations sur l'ensemble des non fumeuses :

```

not_smokers_18_34 <- subset(not_smokers, not_smokers$Age >= 18.0 &
not_smokers$Age < 34.0)
not_smokers_34_54 <- subset(not_smokers, not_smokers$Age >= 34.0 &
not_smokers$Age < 54.0)
not_smokers_54_64 <- subset(not_smokers, not_smokers$Age >= 54.0 &
not_smokers$Age < 65.0)
not_smokers_65 <- subset(not_smokers, not_smokers$Age >= 65.0)

head(not_smokers_18_34)
head(not_smokers_34_54)
head(not_smokers_54_64)
head(not_smokers_65)

```

```

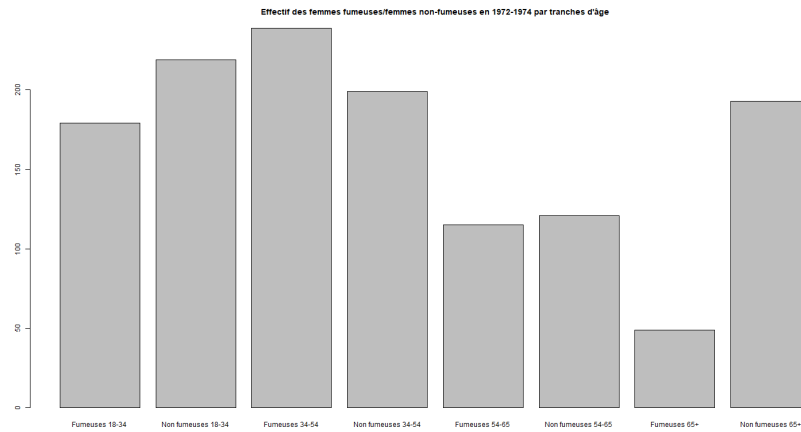
Smoker Status Age
7       No   Alive 23.8
16      No   Alive 25.1
18      No   Alive 27.1
23      No   Alive 33.4
25      No   Alive 18.0
28      No   Alive 25.8
Smoker Status Age
4       No   Alive 47.1

```

6	No	Alive	36.8
17	No	Alive	43.5
29	No	Dead	36.9
52	No	Alive	45.0
58	No	Alive	51.2
Smoker Status Age			
3	No	Dead	57.5
14	No	Alive	58.4
15	No	Dead	60.6
19	No	Alive	58.3
26	No	Alive	56.2
35	No	Alive	56.9
Smoker Status Age			
12	No	Dead	66.0
21	No	Dead	73.2
42	No	Dead	69.7
44	No	Dead	75.8
46	No	Dead	83.0
51	No	Alive	82.8

Nous allons maintenant regarder les effectifs de manière graphique, par un diagramme en barre

```
x_age_72 = c(nrow(smokers_18_34),nrow(not_smokers_18_34),
nrow(smokers_34_54),nrow(not_smokers_34_54),
nrow(smokers_54_64),nrow(not_smokers_54_64),nrow(smokers_65),nrow(not_smokers_65))
type_age_72 = c("Fumeuses 18-34", "Non fumeuses 18-34","Fumeuses 34-54",
"Non fumeuses 34-54","Fumeuses 54-65",
"Non fumeuses 54-65","Fumeuses 65+", "Non fumeuses 65+")
barplot(x_age_72,names.arg=type_age_72,main="Effectif des femmes fumeuses/femmes
non-fumeuses en 1972-1974 par tranches d'âge")
```



On observe déjà qu'il a un plus grand nombre de non-fumeuses de la classe d'âge 65+ que de fumeuses, ce qui pourrait expliquer les taux de mortalité obtenus précédemment.

4.2 Effectif des femmes vivantes/mortes par classe d'âge selon leur catégorie

Comme fait précédemment, nous allons maintenant réduire les échantillons selon si les personnes sont vivantes ou non lors du second sondage :

```
smokers_18_34_alive <- subset(smokers_18_34, smokers_18_34$Status == "Alive")
smokers_34_54_alive <- subset(smokers_34_54, smokers_34_54$Status == "Alive")
smokers_54_64_alive <- subset(smokers_54_64, smokers_54_64$Status == "Alive")
smokers_65_alive <- subset(smokers_65, smokers_65$Status == "Alive")
```

```
head(smokers_18_34_alive)
head(smokers_34_54_alive)
head(smokers_54_64_alive)
head(smokers_65_alive)
```

```
Smoker Status Age
1    Yes  Alive 21.0
2    Yes  Alive 19.3
9    Yes  Alive 24.8
11   Yes  Alive 30.0
38   Yes  Alive 29.5
```

```

50    Yes  Alive 22.1
Smoker Status  Age
10    Yes  Alive 49.5
13    Yes  Alive 49.2
22    Yes  Alive 38.3
31    Yes  Alive 34.6
32    Yes  Alive 51.9
33    Yes  Alive 49.9
Smoker Status  Age
27    Yes  Alive 59.2
61    Yes  Alive 58.1
84    Yes  Alive 58.3
91    Yes  Alive 56.1
136   Yes  Alive 63.6
178   Yes  Alive 56.8
Smoker Status  Age
5     Yes  Alive 81.4
20    Yes  Alive 65.7
255   Yes  Alive 72.1
525   Yes  Alive 74.1
873   Yes  Alive 71.5
966   Yes  Alive 73.8

```

```

smokers_18_34_dead <- subset(smokers_18_34, smokers_18_34$Status == "Dead")
smokers_34_54_dead <- subset(smokers_34_54, smokers_34_54$Status == "Dead")
smokers_54_64_dead <- subset(smokers_54_64, smokers_54_64$Status == "Dead")
smokers_65_dead <- subset(smokers_65, smokers_65$Status == "Dead")

```

```

head(smokers_18_34_dead)
head(smokers_34_54_dead)
head(smokers_54_64_dead)
head(smokers_65_dead)

```

```

Smoker Status  Age
39      Yes    Dead 33.0
828     Yes    Dead 22.6
973     Yes    Dead 28.3
1017    Yes    Dead 32.6
1115    Yes    Dead 31.3

```

Smoker	Status	Age
47	Yes	Dead 44.3
64	Yes	Dead 36.3
88	Yes	Dead 53.6
133	Yes	Dead 35.7
140	Yes	Dead 40.8
172	Yes	Dead 48.4
Smoker Status Age		
8	Yes	Dead 57.5
24	Yes	Dead 62.3
98	Yes	Dead 55.5
102	Yes	Dead 61.0
110	Yes	Dead 62.8
123	Yes	Dead 63.8
Smoker Status Age		
65	Yes	Dead 80.7
113	Yes	Dead 66.5
130	Yes	Dead 87.8
137	Yes	Dead 71.7
191	Yes	Dead 78.3
200	Yes	Dead 68.4

Les résultats semblent bon.

Nous allons faire de même pour les ensembles des non-fumeuses:

```
not_smokers_18_34_alive <- subset(not_smokers_18_34,
not_smokers_18_34$Status == "Alive")
not_smokers_34_54_alive <- subset(not_smokers_34_54,
not_smokers_34_54$Status == "Alive")
not_smokers_54_64_alive <- subset(not_smokers_54_64,
not_smokers_54_64$Status == "Alive")
not_smokers_65_alive <- subset(not_smokers_65, not_smokers_65$Status == "Alive")

head(not_smokers_18_34_alive)
head(not_smokers_34_54_alive)
head(not_smokers_54_64_alive)
head(not_smokers_65_alive)
```

Smoker	Status	Age
--------	--------	-----

7	No	Alive	23.8
16	No	Alive	25.1
18	No	Alive	27.1
23	No	Alive	33.4
25	No	Alive	18.0
28	No	Alive	25.8

Smoker	Status	Age
--------	--------	-----

4	No	Alive	47.1
6	No	Alive	36.8
17	No	Alive	43.5
52	No	Alive	45.0
58	No	Alive	51.2
60	No	Alive	41.9

Smoker	Status	Age
--------	--------	-----

14	No	Alive	58.4
19	No	Alive	58.3
26	No	Alive	56.2
35	No	Alive	56.9
74	No	Alive	62.4
75	No	Alive	62.5

Smoker	Status	Age
--------	--------	-----

51	No	Alive	82.8
109	No	Alive	83.7
139	No	Alive	82.0
160	No	Alive	67.2
173	No	Alive	82.7
188	No	Alive	78.2

```
not_smokers_18_34_dead <- subset(not_smokers_18_34,
not_smokers_18_34$Status == "Dead")
not_smokers_34_54_dead <- subset(not_smokers_34_54,
not_smokers_34_54$Status == "Dead")
not_smokers_54_64_dead <- subset(not_smokers_54_64,
not_smokers_54_64$Status == "Dead")
not_smokers_65_dead <- subset(not_smokers_65, not_smokers_65$Status == "Dead")
```

```
head(not_smokers_18_34_dead)
head(not_smokers_34_54_dead)
head(not_smokers_54_64_dead)
head(not_smokers_65_dead)
```

Smoker	Status	Age
147	No	Dead 26.3
516	No	Dead 25.3
565	No	Dead 29.8
628	No	Dead 29.3
675	No	Dead 20.2
1256	No	Dead 28.5

Smoker	Status	Age
29	No	Dead 36.9
215	No	Dead 35.2
299	No	Dead 52.4
309	No	Dead 47.9
344	No	Dead 47.0
608	No	Dead 47.2

Smoker	Status	Age
3	No	Dead 57.5
15	No	Dead 60.6
71	No	Dead 58.1
86	No	Dead 55.9
135	No	Dead 62.3
150	No	Dead 58.3

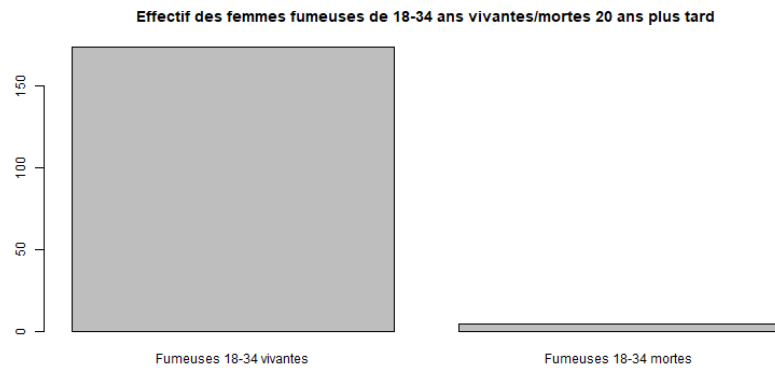
Smoker	Status	Age
12	No	Dead 66.0
21	No	Dead 73.2
42	No	Dead 69.7
44	No	Dead 75.8
46	No	Dead 83.0
53	No	Dead 73.3

Les résultats restent satisfaisants.

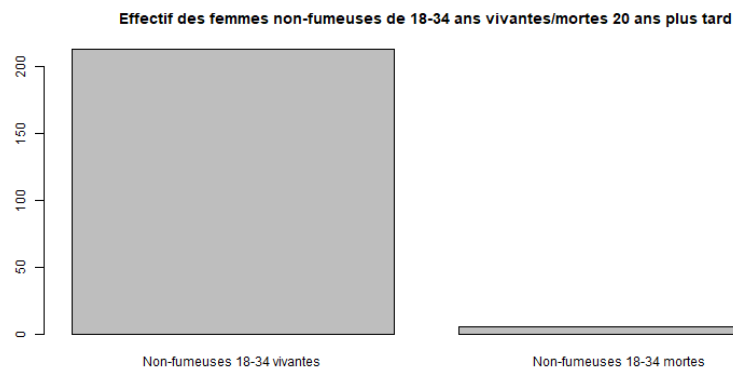
Regardons maintenant tout ça sur des graphiques :

Pour les 18-34 ans :

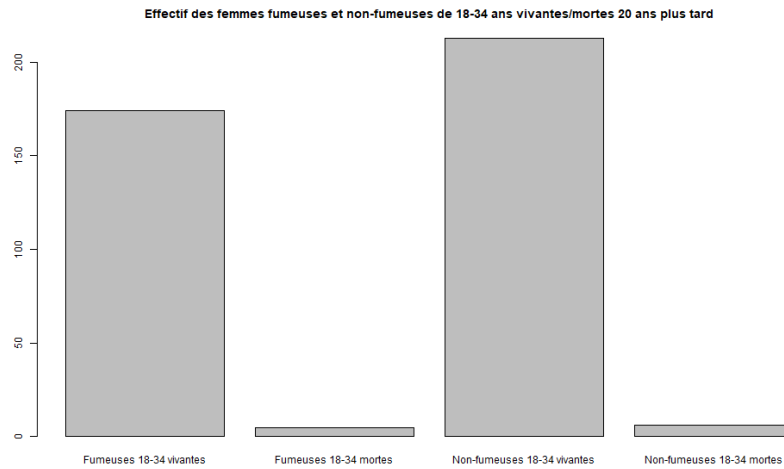
```
x_age_92 = c(nrow(smokers_18_34_alive),nrow(smokers_18_34_dead))
type_age_92 = c("Fumeuses 18-34 vivantes", "Fumeuses 18-34 mortes")
barplot(x_age_92,names.arg=type_age_92,main="Effectif des femmes fumeuses de
18-34 ans vivantes/mortes 20 ans plus tard")
```

```
x_age_92 = c(nrow(not_smokers_18_34_alive),nrow(not_smokers_18_34_dead))
type_age_92 = c("Non-fumeuses 18-34 vivantes", "Non-fumeuses 18-34 mortes")
barplot(x_age_92,names.arg=type_age_92,main="Effectif des femmes non-fumeuses de
18-34 ans vivantes/mortes 20 ans plus tard")
```



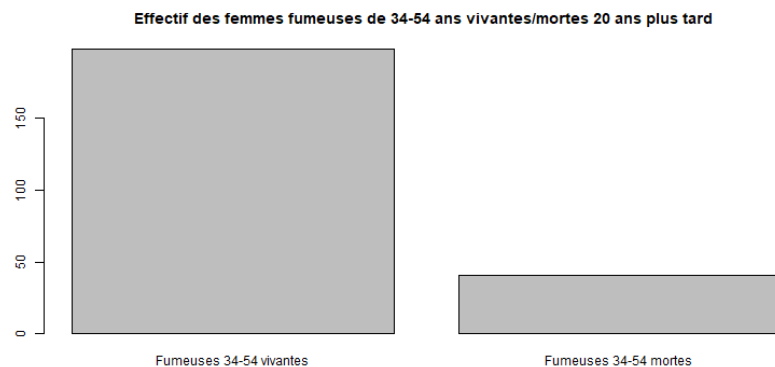
```
x_age_92 = c(nrow(smokers_18_34_alive),nrow(smokers_18_34_dead),
nrow(not_smokers_18_34_alive),nrow(not_smokers_18_34_dead))
type_age_92 = c("Fumeuses 18-34 vivantes", "Fumeuses 18-34 mortes",
"Non-fumeuses 18-34 vivantes", "Non-fumeuses 18-34 mortes")
barplot(x_age_92,names.arg=type_age_92,main="Effectif des femmes fumeuses et
non-fumeuses de 18-34 ans vivantes/mortes 20 ans plus tard")
```



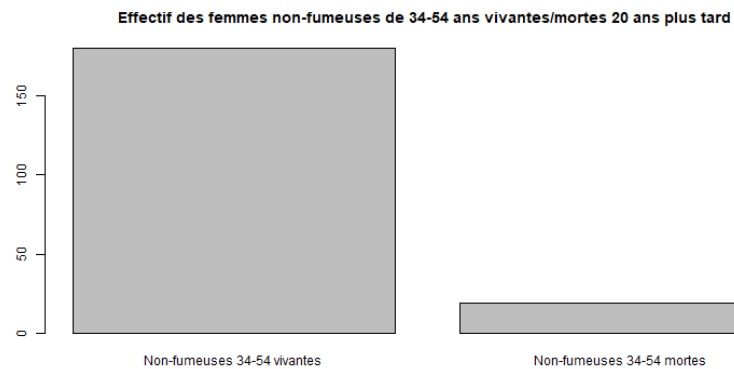
L'effectif de femmes mortes entre celles fumeuses et non-fumeuses à l'air d'être équivalent. On a en revanche un plus grand effectif de femmes non-fumeuses vivantes. Rappelons-nous que dans les ensembles de départ (1972-1974), nous avions également un plus grand effectif de non-fumeuses vivantes. On peut donc conjecturer que le taux de mortalité sera plus élevé pour les fumeuses cette fois.

Pour les 34-54 ans :

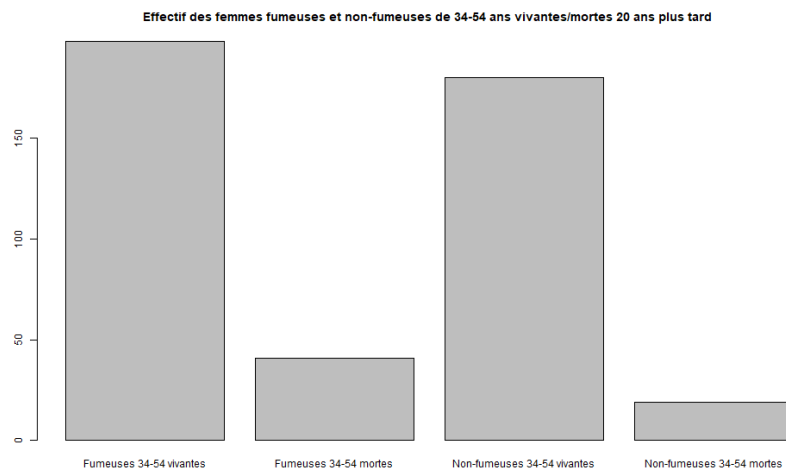
```
x_age_92 = c(nrow(smokers_34_54_alive),nrow(smokers_34_54_dead))
type_age_92 = c("Fumeuses 34-54 vivantes", "Fumeuses 34-54 mortes")
barplot(x_age_92,names.arg=type_age_92,main="Effectif des femmes fumeuses de 34-54 ans
vivantes/mortes 20 ans plus tard ")
```



```
x_age_92 = c(nrow(not_smokers_34_54_alive),nrow(not_smokers_34_54_dead))
type_age_92 = c("Non-fumeuses 34-54 vivantes", "Non-fumeuses 34-54 mortes")
barplot(x_age_92,names.arg=type_age_92,main="Effectif des femmes non-fumeuses de
34-54 ans vivantes/mortes 20 ans plus tard ")
```



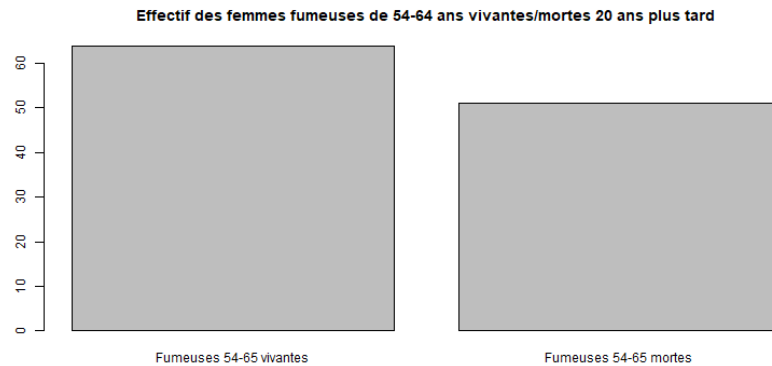
```
x_age_92 = c(nrow(smokers_34_54_alive),nrow(smokers_34_54_dead),
nrow(not_smokers_34_54_alive),nrow(not_smokers_34_54_dead))
type_age_92 = c("Fumeuses 34-54 vivantes", "Fumeuses 34-54 mortes",
"Non-fumeuses 34-54 vivantes", "Non-fumeuses 34-54 mortes")
barplot(x_age_92,names.arg=type_age_92,main="Effectif des femmes fumeuses et
non-fumeuses de 34-54 ans vivantes/mortes 20 ans plus tard ")
```



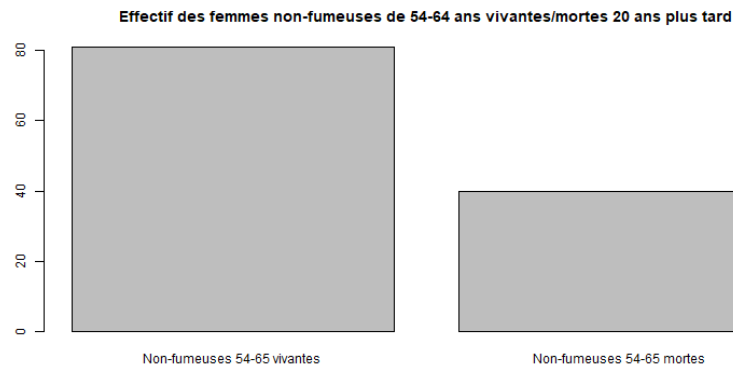
Il y a plus de fumeuses mortes que de non-fumeuses mortes, mais il reste plus de fumeuses vivantes que de non-fumeuses vivantes. Dans les ensembles de départ (1972-1974), il y avait plus de fumeuses vivantes que de non-fumeuses vivantes. Nous ne pouvons pas encore nous prononcer sur les taux de mortalité.

Pour les 54-65 ans :

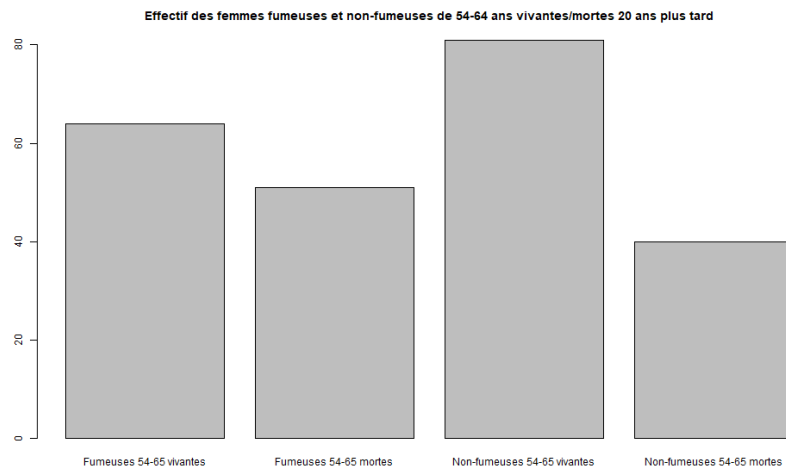
```
x_age_92 = c(nrow(smokers_54_64_alive),nrow(smokers_54_64_dead))
type_age_92 = c("Fumeuses 54-65 vivantes", "Fumeuses 54-65 mortes")
barplot(x_age_92,names.arg=type_age_92,main="Effectif des femmes fumeuses
de 54-64 ans vivantes/mortes 20 ans plus tard")
```



```
x_age_92 = c(nrow(not_smokers_54_64_alive),nrow(not_smokers_54_64_dead))
type_age_92 = c("Non-fumeuses 54-65 vivantes", "Non-fumeuses 54-65 mortes")
barplot(x_age_92,names.arg=type_age_92,main="Effectif des femmes non-fumeuses
de 54-64 ans vivantes/mortes 20 ans plus tard")
```



```
x_age_92 = c(nrow(smokers_54_64_alive),nrow(smokers_54_64_dead),
nrow(not_smokers_54_64_alive),nrow(not_smokers_54_64_dead))
type_age_92 = c("Fumeuses 54-65 vivantes", "Fumeuses 54-65 mortes",
"Non-fumeuses 54-65 vivantes", "Non-fumeuses 54-65 mortes")
barplot(x_age_92,names.arg=type_age_92,main="Effectif des femmes fumeuses
et non-fumeuses de 54-64 ans vivantes/mortes 20 ans plus tard")
```

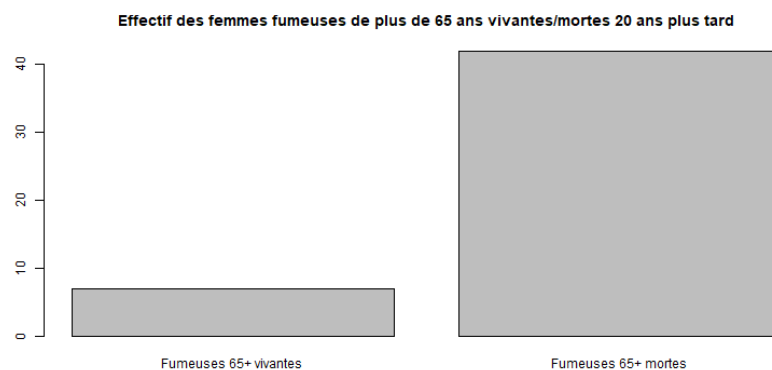


Il y a plus de fumeuses mortes que de non-fumeuses mortes, et il y a moins de fumeuses vivantes que de non-fumeuses vivantes, sachant que dans les ensembles de départ (1972-1974), il y avait plus de non-fumeuses vivantes que de fumeuses vivantes. On peut donc deviner que le taux de mortalité

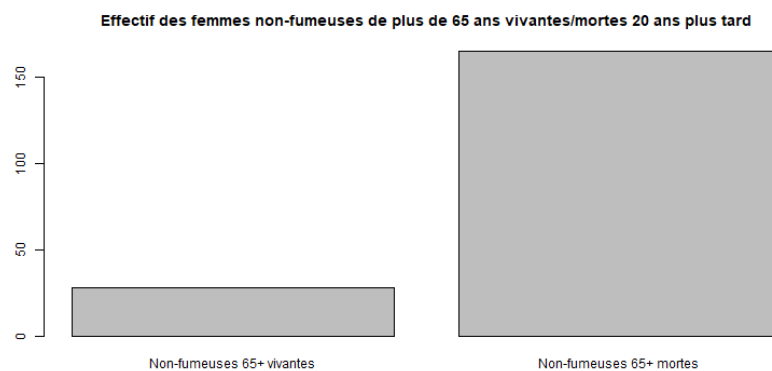
sera plus élevé pour les fumeuses.

Pour les plus de 65 ans :

```
x_age_92 = c(nrow(smokers_65_alive),nrow(smokers_65_dead))
type_age_92 = c("Fumeuses 65+ vivantes", "Fumeuses 65+ mortes")
barplot(x_age_92,names.arg=type_age_92,main="Effectif des femmes fumeuses de plus
de 65 ans vivantes/mortes 20 ans plus tard")
```

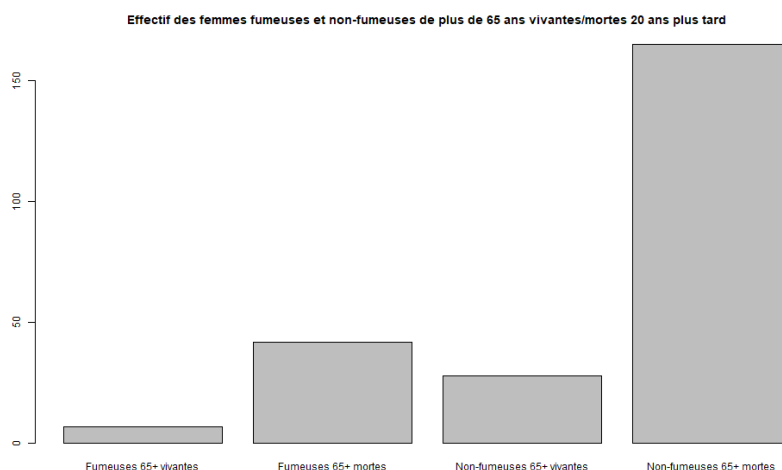


```
x_age_92 = c(nrow(not_smokers_65_alive),nrow(not_smokers_65_dead))
type_age_92 = c("Non-fumeuses 65+ vivantes", "Non-fumeuses 65+ mortes")
barplot(x_age_92,names.arg=type_age_92,main="Effectif des femmes non-fumeuses de plus
de 65 ans vivantes/mortes 20 ans plus tard")
```



```
x_age_92 = c(nrow(smokers_65_alive),nrow(smokers_65_dead),
```

```
nrow(not_smokers_65_alive),nrow(not_smokers_65_dead))
type_age_92 = c("Fumeuses 65+ vivantes", "Fumeuses 65+ mortes",
"Non-fumeuses 65+ vivantes", "Non-fumeuses 65+ mortes")
barplot(x_age_92,names.arg=type_age_92,main="Effectif des femmes fumeuses
et non-fumeuses de plus de 65 ans vivantes/mortes 20 ans plus tard")
```



Il y a bien plus de non-fumeuses mortes que de fumeuses mortes, mais il y a plus de non-fumeuses vivantes que de fumeuses vivantes. L'ensemble de départ (1972-1974) est encore plus important ici puisque l'écart d'effectif entre les fumeuses/non-fumeuses est flagrant : environ 150 personnes d'écart, mais nous nous prononcerons pas encore sur le taux de mortalité.

4.3 Taux de mortalité par classe d'âge selon les catégories

De nouveau, il ne reste plus qu'à calculer le taux de mortalité pour chaque groupe :

Les 18-34 ans :

```
eff_dead_smokers_18_34 = nrow(smokers_18_34_dead)
eff_smokers_18_34 = nrow(smokers_18_34)
taux_mortalite_smokers_18_34 = eff_dead_smokers_18_34/eff_smokers_18_34

eff_dead_not_smokers_18_34 = nrow(not_smokers_18_34_dead)
```

```

eff_not_smokers_18_34 = nrow(not_smokers_18_34)
taux_mortalite_not_smokers_18_34 = eff_dead_not_smokers_18_34/eff_not_smokers_18_34

eff_dead_smokers_18_34
eff_smokers_18_34
taux_mortalite_smokers_18_34

eff_dead_not_smokers_18_34
eff_not_smokers_18_34
taux_mortalite_not_smokers_18_34

[1] 5
[1] 179
[1] 0.02793296
[1] 6
[1] 219
[1] 0.02739726

```

Le taux de mortalité chez les fumeuses de 18 à 34 ans est plus élevé que celui des non-fumeuses (quoique les valeurs restent proches). Cela correspond à notre conjecture.

Les 34-54 ans :

```

eff_dead_smokers_34_54 = nrow(smokers_34_54_dead)
eff_smokers_34_54 = nrow(smokers_34_54)
taux_mortalite_smokers_34_54 = eff_dead_smokers_34_54/eff_smokers_34_54

eff_dead_not_smokers_34_54 = nrow(not_smokers_34_54_dead)
eff_not_smokers_34_54 = nrow(not_smokers_34_54)
taux_mortalite_not_smokers_34_54 = eff_dead_not_smokers_34_54/eff_not_smokers_34_54

eff_dead_smokers_34_54
eff_smokers_34_54
taux_mortalite_smokers_34_54

eff_dead_not_smokers_34_54
eff_not_smokers_34_54
taux_mortalite_not_smokers_34_54

```



```
[1] 41
[1] 239
[1] 0.1715481
[1] 19
[1] 199
[1] 0.09547739
```

Le taux de mortalité chez les fumeuses de 34 à 54 ans est plus élevé que celui des non fumeuses.

Les 54-65 ans :

```
eff_dead_smokers_54_64 = nrow(smokers_54_64_dead)
eff_smokers_54_64 = nrow(smokers_54_64)
taux_mortalite_smokers_54_64 = eff_dead_smokers_54_64/eff_smokers_54_64

eff_dead_not_smokers_54_64 = nrow(not_smokers_54_64_dead)
eff_not_smokers_54_64 = nrow(not_smokers_54_64)
taux_mortalite_not_smokers_54_64 = eff_dead_not_smokers_54_64/eff_not_smokers_54_64

eff_dead_smokers_54_64
eff_smokers_54_64
taux_mortalite_smokers_54_64

eff_dead_not_smokers_54_64
eff_not_smokers_54_64
taux_mortalite_not_smokers_54_64

[1] 51
[1] 115
[1] 0.4434783
[1] 40
[1] 121
[1] 0.3305785
```

De même pour les 54 à 65 ans, le taux est plus élevé pour les fumeuses. Notre conjecture est validée.

Les 65+ ans :

```

eff_dead_smokers_65 = nrow(smokers_65_dead)
eff_smokers_65 = nrow(smokers_65)
taux_mortalite_smokers_65 = eff_dead_smokers_65/eff_smokers_65

eff_dead_not_smokers_65 = nrow(not_smokers_65_dead)
eff_not_smokers_65 = nrow(not_smokers_65)
taux_mortalite_not_smokers_65 = eff_dead_not_smokers_65/eff_not_smokers_65


eff_dead_smokers_65
eff_smokers_65
taux_mortalite_smokers_65

eff_dead_not_smokers_65
eff_not_smokers_65
taux_mortalite_not_smokers_65


[1] 42
[1] 49
[1] 0.8571429
[1] 165
[1] 193
[1] 0.8549223

```

Et on retrouve la même conclusion pour les plus de 65 ans: la taux de mortalité est plus élevé chez les fumeuses.

5 Conclusion sur ces deux analyses

Contrairement à notre première analyse des données, la deuxième semble s'orienter vers le fait que les fumeuses ont un plus fort taux de mortalité que les non-fumeuses, ce qui confirme probablement ce que chacun aurait pensé intuitivement.

Les paramètres qu'on prend en compte influent sûrement beaucoup sur nos résultats finaux et surtout, la manière dont les données sont réparties selon les paramètres pris en compte.

De plus les taux que nous avons pu trouvés sont souvent assez proches, donc si on modifie les ensembles de départ, on peut sûrement vite avoir des

analyses paradoxales.

La décision de faire une analyse selon un paramètre doit avoir un sens: dans notre situation, je ne saurais pas dire si le choix de l'âge était pertinent ou non mais par exemple, si on avait conduit toutes nos opérations selon la couleur de cheveux des femmes interrogées, on aurait eu d'autres résultats qui n'auraient pas eu le mérite d'être considérés puisqu'il n'y a aucun lien entre ce paramètre et le tabagisme/les morts liées au tabagisme.