Devoir Maison - Probabilités et statistiques pour l'ingénieur

Aïssatou Signaté

20/11/2021

Introduction :

Dans le cadre du cours de Probabilités et statistiques pour l'ingénieur dispensé en M1 informatique par Dr Vittorio Perduca , il a été demandé aux étudiants de réaliser un devoir d'analyse sur le jeu données *Titanic: Machine Learning from Disaster de Kaggle.*

Description des données

Question 1

Télécharger les données (dite d'apprentissage) titanic_train.Rdata disponibles à l'adresse

```
load('titanic_train.Rdata')
```

```
head(train)
```

```
##     PassengerId Survived Pclass                                          Name
## 707         707        1      2                 Kelly, Mrs. Florence "Fannie"
## 706         706        0      2 Morley, Mr. Henry Samuel ("Mr Henry Marshall")
## 566         566        0      3                            Davies, Mr. Alfred J
## 244         244        0      3                 Maenpaa, Mr. Matti Alexanteri
## 825         825        0      3                  Panula, Master. Urho Abraham
## 754         754        0      3                            Jonkoff, Mr. Lalio
##        Sex Age SibSp Parch            Ticket    Fare Cabin Embarked
## 707 female  45     0     0            223596 13.5000  <NA>        S
## 706   male  39     0     0            250655 26.0000  <NA>        S
## 566   male  24     2     0         A/4 48871 24.1500  <NA>        S
## 244   male  22     0     0 STON/O 2. 3101275  7.1250  <NA>        S
## 825   male   2     4     1           3101295 39.6875  <NA>        S
## 754   male  23     0     0            349204  7.8958  <NA>        S
```

Question 2

Explorer la structure des données:

• **donner le nombre d'observations (c'est à dire le nombre de lignes du data frame) et le nombre devariables (le nombre de colonnes)**

- **donner le nom des variables et dire si elles sont quantitatives ou qualitatives (indication: on utilisera les fonctions str(), names(), class(). . .)**

Les données que nous devons analyser sont issues du *Titanic: Machine Learning From Disaster de Keagle*. Ce data set regroupe les informations concernant les passagers du Titanic selon plusieurs critères : age, sexe, survie etc . Afin de réaliser ce devoir , nous travaillerons sur un data set réduit contenant 12 variables, ainsi que 594 observations(lignes), donc les informations sur 594 passagers

```
names(train)
```

```
##  [1] "PassengerId" "Survived"    "Pclass"      "Name"        "Sex"
##  [6] "Age"         "SibSp"       "Parch"       "Ticket"      "Fare"
## [11] "Cabin"       "Embarked"
```

```
str(train)
```

```
## 'data.frame':    594 obs. of  12 variables:
##  $ PassengerId: int  707 706 566 244 825 754 751 649 463 438 ...
##  $ Survived   : int  1 0 0 0 0 0 1 0 0 1 ...
##  $ Pclass     : int  2 2 3 3 3 3 2 3 1 2 ...
##  $ Name       : Factor w/ 891 levels "Abbing, Mr. Anthony",..: 439 561 205 505 635 420 859 872 285 69
##  $ Sex        : Factor w/ 2 levels "female","male": 1 2 2 2 2 2 1 2 2 1 ...
##  $ Age        : num  45 39 24 22 2 23 4 NA 47 24 ...
##  $ SibSp      : int  0 0 2 0 4 0 1 0 0 2 ...
##  $ Parch      : int  0 0 0 0 1 0 1 0 0 3 ...
##  $ Ticket     : Factor w/ 681 levels "110152","110413",..: 109 171 520 659 250 352 235 621 7 238 ...
##  $ Fare       : num  13.5 26 24.15 7.12 39.69 ...
##  $ Cabin      : Factor w/ 147 levels "A10","A14","A16",..: NA NA NA NA NA NA NA NA 134 NA ...
##  $ Embarked   : Factor w/ 3 levels "C","Q","S": 3 3 3 3 3 3 3 3 3 3 ...
```

- 1.**PassengerId**: Numéro d'identification du passager
  ->Nominale représenté par un entier, qualitative
- 2.**Survived** : Survie du passager, 1 = Oui , 0 = Non
  –>Nominale, qualitative
- 3.**Pclass** : Classe du ticket, 1 = première, 2 = deuxieme, 3 = troisième
  –>Ordinale, qualitative
- 4.**Name** : Nom du passager
  –>Qualitative
- 5.**Sex** : Sexe du passager , homme ou femme
  –>Catégorique ,qualitaive
- 6.**Age** : Age en année du passager
  –>Quantitative, continue
- 7.**SibSp** : Nombre de frères et soeurs ou conjoint à bord
  –>Quantitative
- 8.**Parch** : Nombre d'enfants et de parents à bord
  –>Quantitative, discrète
- 9.**Ticket** : Numéro du tiket
  –>Nominal représenté par un entier, qualitative
- 10.**Fare**: Tarif passager
  –>Quantitative, continue
- 11.**Cabin**: Numéro de cabine
  –>Nominale représenté par un entier, qualitative
- 12.**Embarked**: Port d'empabarquement, 3 valeurs possibles : C = Cherbourg, Q = Queenstown, S = Southampton
  –> Nominale, qualitative.

Question 3

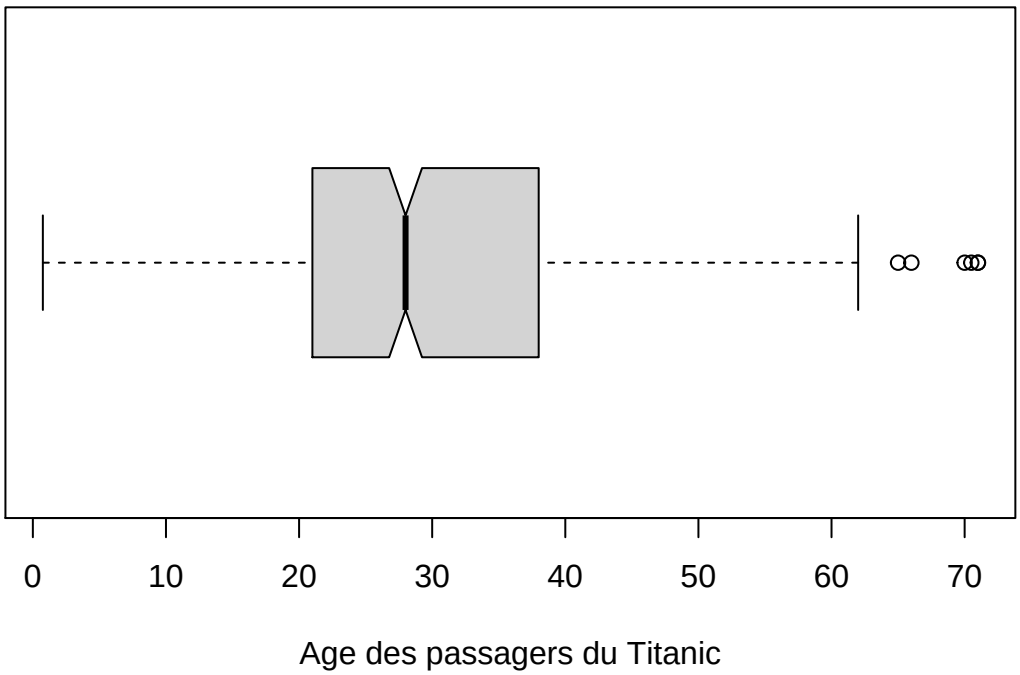
Pour cette question nous devons analyser 4 variable :

- **Surived/S**
- **Sex/Sx**
- **Pclass/P**
- **Age/A**

-Pour Age

Nous allons commencer avec la variable Age. C'est une variable quantitative et continue que nous pouvons représenter avec un boxplot

```
age <- na.omit(train$Age)
boxplot(age, xlab= "Age des passagers du Titanic",
        horizontal = TRUE,
        notch = TRUE)
```




```
median(age, na.rm= TRUE)
```

```
## [1] 28
```

L'age moyen des passagers du Titanic était de 29.58 ans. La médiane, qui vaut 28, est représentée par la ligne dans la boîte : la moitié des passagers avaient un âge inférieur à 28 ans et une autre moitié un âge supérieur à 28 ans.

La boîte correspond à l'étendue interquartile, indiquant la distance entre Q1(25%) et Q3(75%).
Une autre particulité de la boîte à moustache est d'indiquer si des données sont asymétriques ou non.
Dans notre cas, il y a une asymétrie à droite concernant la variable âge : la plupart des passagers étaient relativement jeunes.

Avec summary() nous pouvons calculer les valeurs suivantes:

```
summary(age)
```

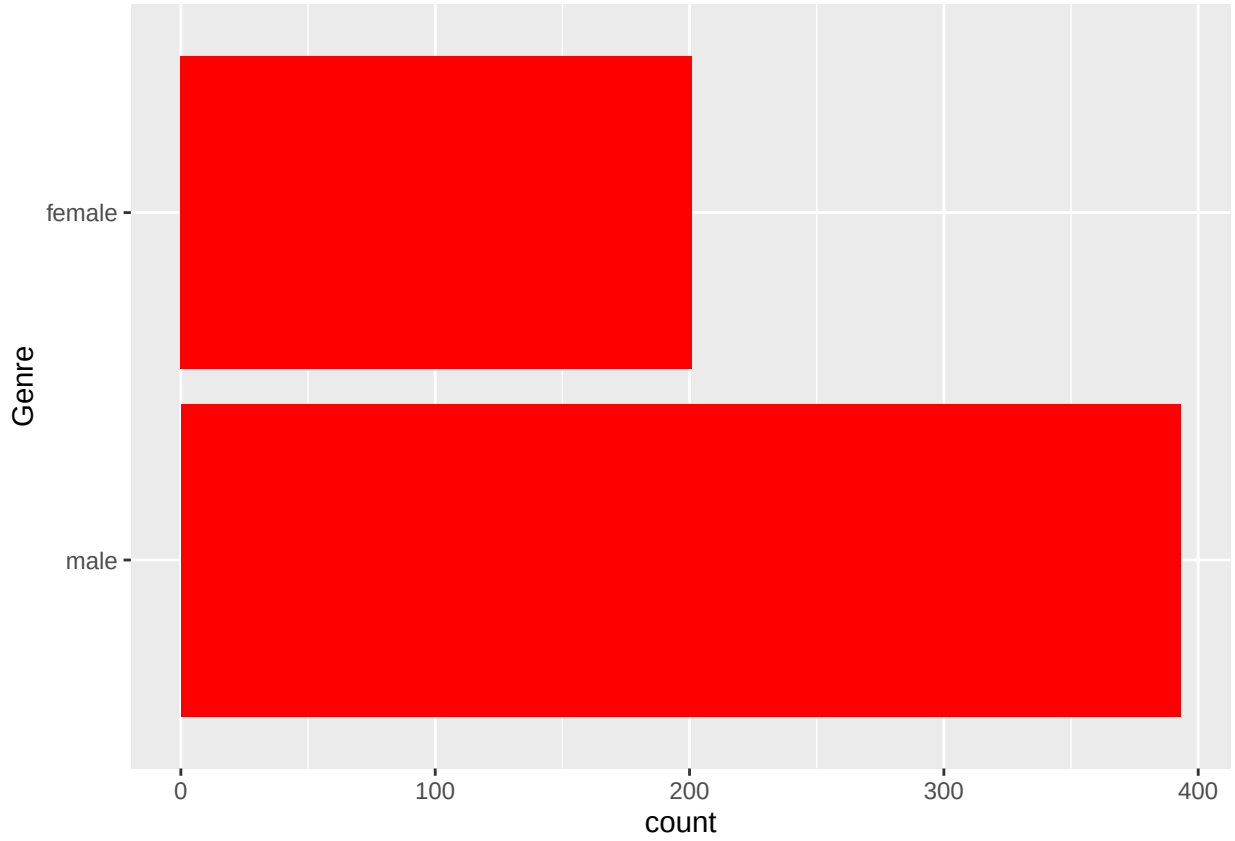
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.75   21.00   28.00   29.58   38.00   71.00
```

- Le passager le plus agé avait 71 ans
- Le passager le plus jeune avait 0.75 an
- L'age médian des passagers était de 28 ans
- En moeynne un passager avait 29.58 ans

-Pour Sex :

Nous pouvons à l'aide d'un histogramme voir la répartition des femmes et des hommes parmis les passagers.

```
Sx <- na.omit(train$Sex)
ggplot(train, aes(x=reorder(Sex, Sex, function(x)-length(x)))) +
geom_bar(fill='red') +  labs(x='Genre') + coord_flip()
```



Les passagers à bord du Titaic étaient en grande partie des hommes : 393 hommes et 201 femmes sur le data set train.

```
table(Sx)
```

```
## Sx
## female   male
##    201    393
```

-Pour Survived

Concernant la variable **Survived**, c'est une variable catégorique/binaire avec les valeurs 1 (le passager a survécu) ou 0 (le passager n'a pas survécu).

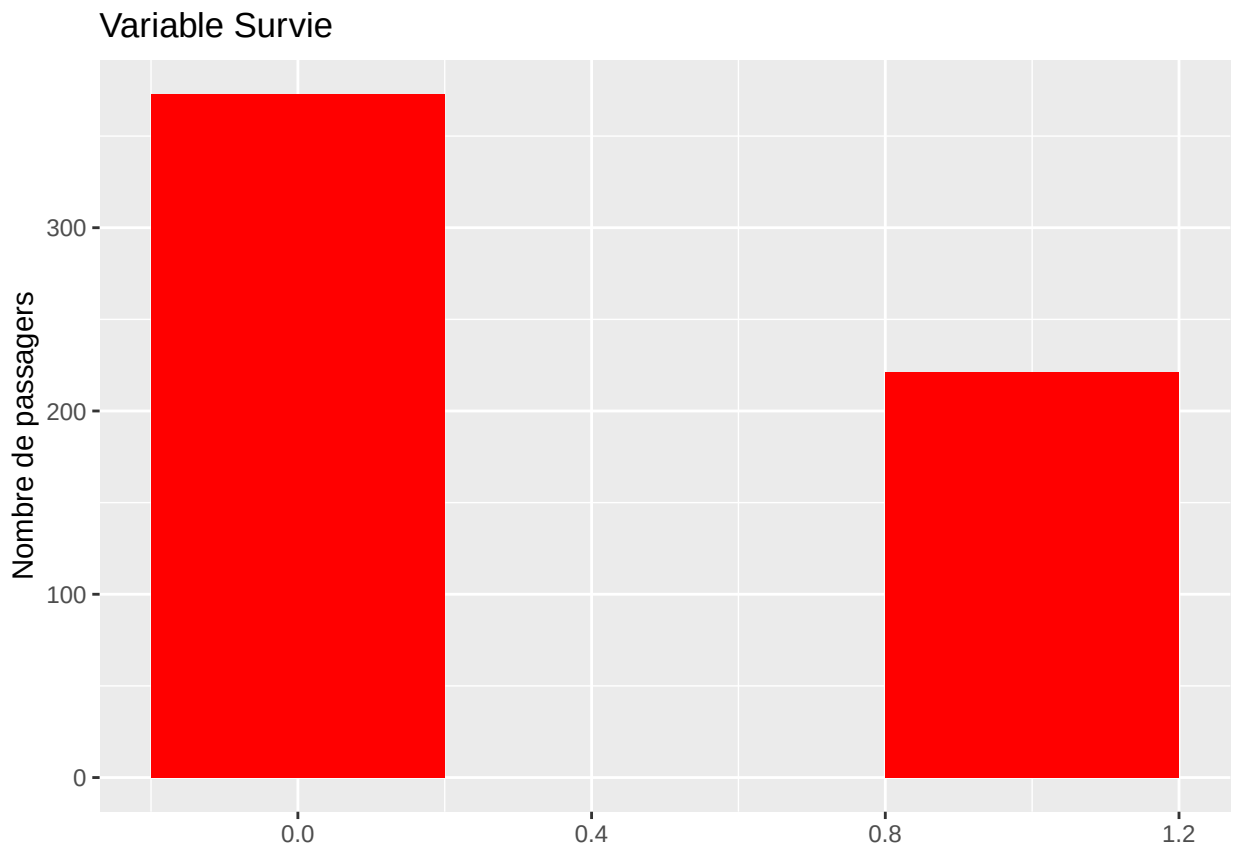
Nous pouvons utiliser la fonction summary() pour calculer une moyenne.

```
S <- train$Survived
summary(S)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.0000  0.0000  0.3721  1.0000  1.0000
```

Ici, la moyenne représente le taux de survie, il est de 37.21% (221/594). Nous pouvons également faire un barplot pour visualiser cette variable

```
train %>%
  ggplot(aes(x = Survived)) +
  geom_bar(width = 0.4, fill="red") +
  labs(title = "Variable Survie",col = "red", x = NULL, y = "Nombre de passagers")
```



La majorité des paassagers présent sur le data set n'ont pas survécu au naufrage : 373 personnes sont décédés et 221 ont survécu.

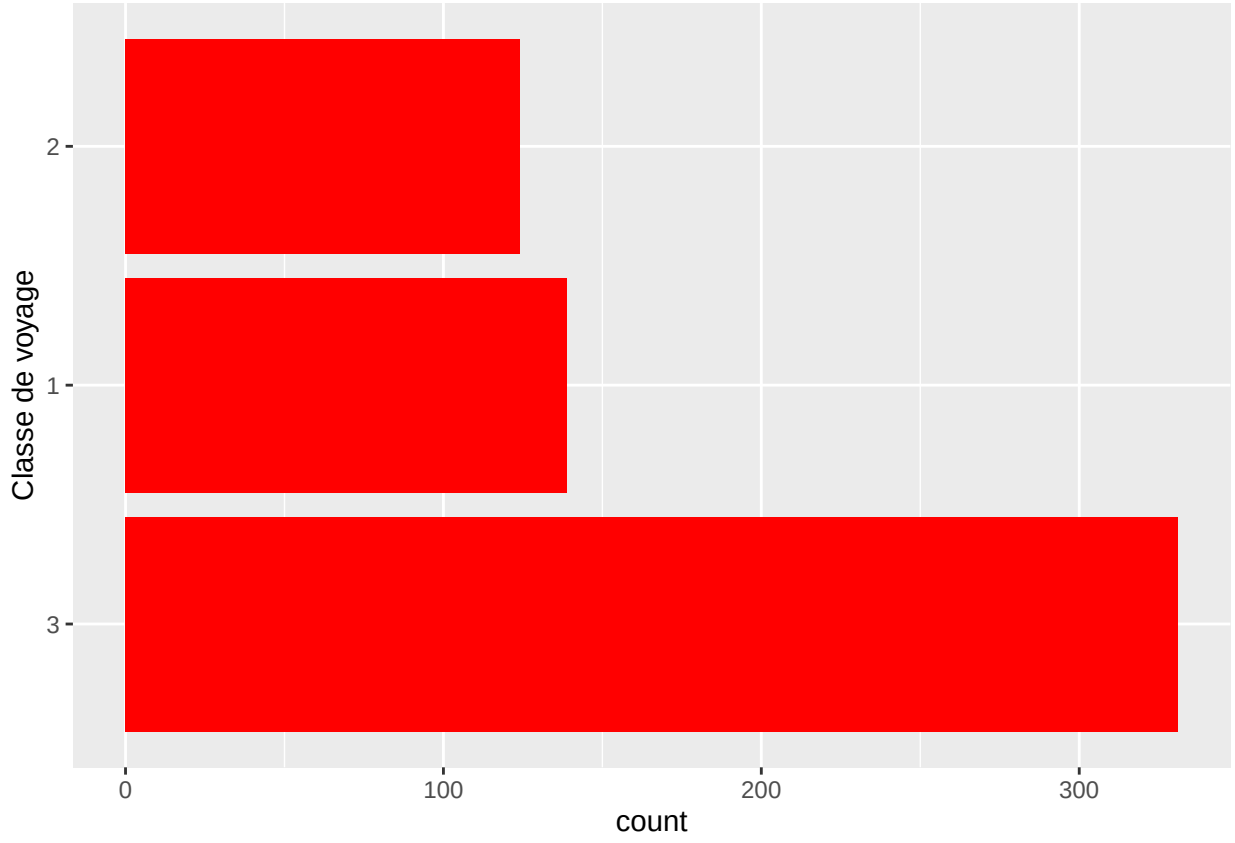
```
table(S)
```

```
## S
##   0   1
## 373 221
```

-Pour Pclass :

Enfin, concernant la classe de voyage ,nous pouvons réaliser un box plot

```
P <- na.omit(train$Pclass)
ggplot(train, aes(x=reorder(Pclass, Pclass, function(x)-length(x)))) +
geom_bar(fill='red') +  labs(x='Classe de voyage') + coord_flip()
```




```
table(P)
```

```
## P
##   1   2   3
## 139 124 331
```

La majorité des passagers voyageaient en 3eme classe : 331.
Les classes 1 et 2 étaient composées respectivement de 139 et 124 passagers.

Question 4

Construire une nouvelle variable cAge qui catégorises Age à l'aide de la fonction cut() (consulter l'aide!). On considérera les catégories d'ages par tranches de 20 ans, allant de 0 à 80 ans: (0, 20], (20, 40], (40, 60] et (60, 80] ans. Décrire cette nouvelle variable, dénotée cA dans la suite de ce document

```
cAge<- train$Age
cAge<-cut(cAge, breaks=c(0,20,40,60,80))
head(cAge)
```

```
## [1] (40,60] (20,40] (20,40] (20,40] (0,20]  (20,40]
## Levels: (0,20] (20,40] (40,60] (60,80]
```

Liens entre les variables

Question 5

En utilisant les statistiques descriptives et/ou les graphiques les plus appropriés, décrire le lien entre

- **Sx et S**
- **P et S**
- **A et S**
- **cA et S.**

Par la suite nous ne considérerons pas la variable A, préférant travailler avec cA.

Relation entre variable Sex et Survie:

Ce sont deux variables discrètes et catégoriques. En analysant ces deux variables ensemble , nous pouvons, par exemple savoir si la consigne "enfants et femmes d'abord" a bien été respectée: si c'est le cas , les femmes seront plus présentes parmis les survivants . Nous pouvons faire une table de contingence.

```
tally(~ Sex + Survived, data = train, margins = T)
```

```
##         Survived
## Sex        0   1 Total
##   female  49 152   201
##   male   324  69   393
##   Total  373 221   594
```

Parmis les 594 individus, 201 étaient des femmes et 393 des hommes.Les femmes sont plus nombreuses parmis les survivants (152 femmes , 69 hommes).

```
(152/201)*100 #femme
```

```
## [1] 75.62189
```

```r
(69/393)*100#homme
```

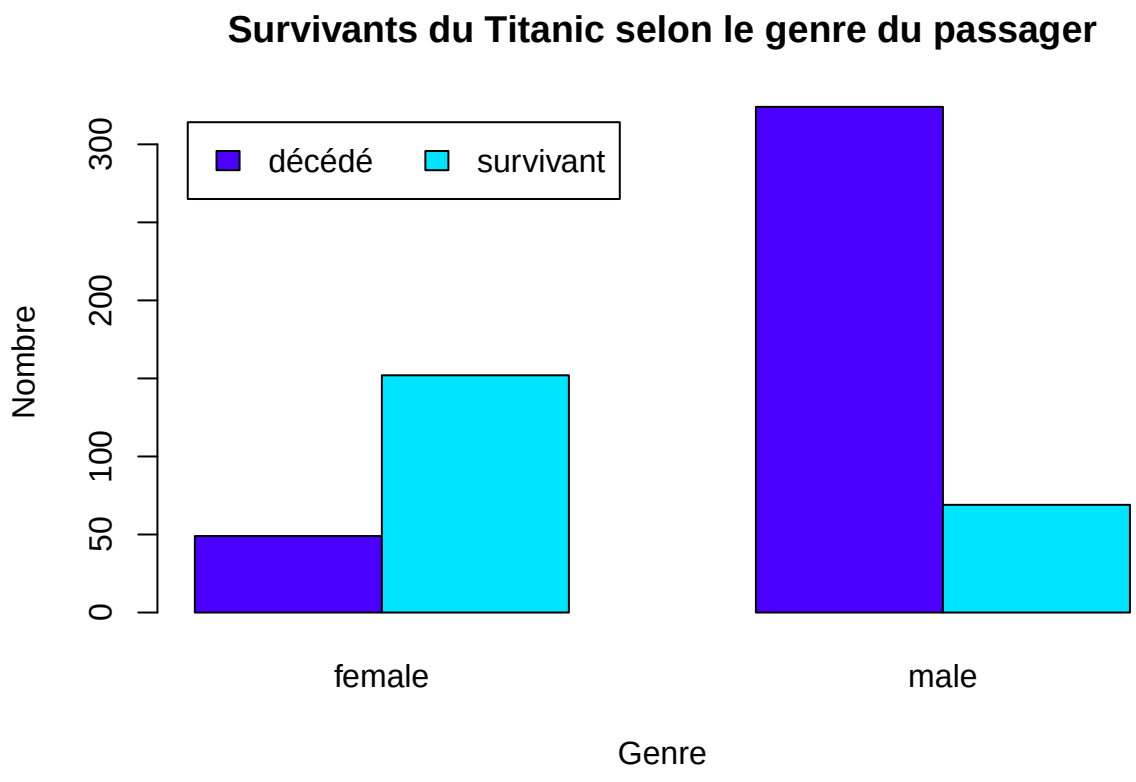
```
## [1] 17.55725
```

- 76% des femmes ont survécu.
- 18% des hommes ont survécu.

Nous pouvons illustrer cela à l'aide d'un barplot.

```r
counts = table(train$Survived,train$Sex)
barplot(counts,
        main = "Survivants du Titanic selon le genre du passager",
        xlab = "Genre",
        ylab = "Nombre",
        col = topo.colors(2),
        beside = TRUE)

legend("topleft",
       inset =.03,
       legend = c("décédé", "survivant"),
       fill = topo.colors(2),
       horiz = TRUE)
```

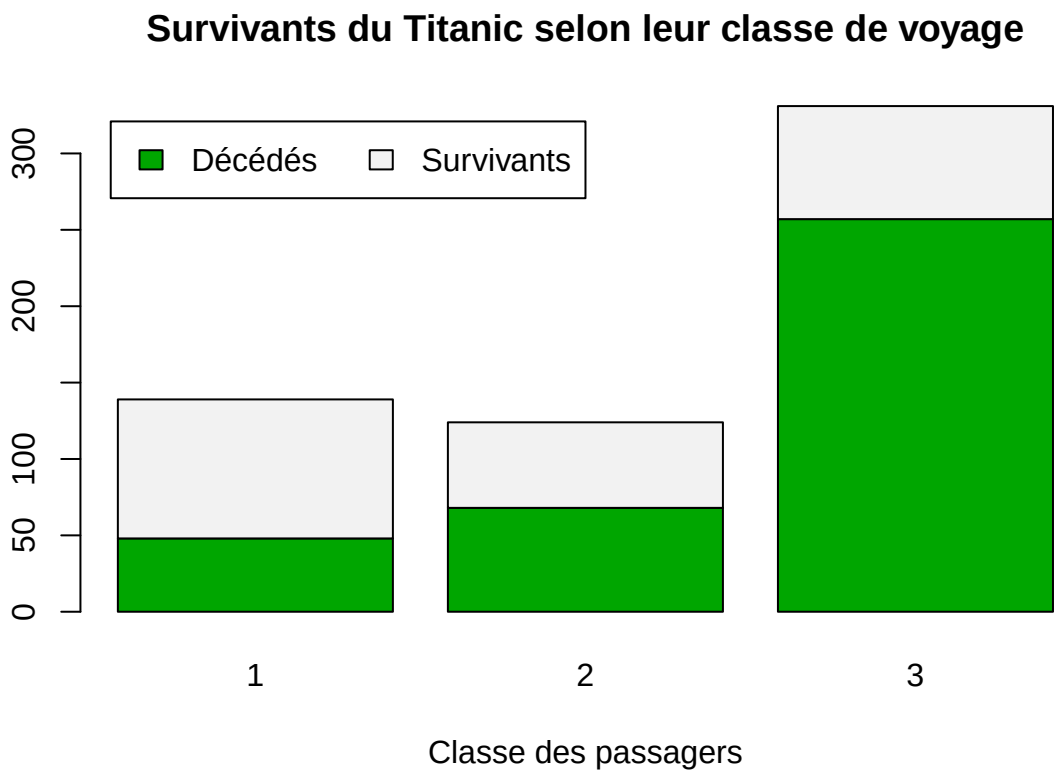


Relation entre variables P et S

Nous allons voir le lien entre la classe occupée par le passager et la variable survie

```r
counts = table(train$Survived,train$Pclass)
#construction du barplot
barplot(counts,
        main = "Survivants du Titanic selon leur classe de voyage",
        xlab = "Classe des passagers",
        col = terrain.colors(2))


legend("topleft",
       inset = .03,
       legend = c("Décédés", "Survivants"),
       fill = terrain.colors(2),
       horiz = TRUE)
```



Nous pouvons voir que les passagers en classe 3 avaient une plus grande probabilité de mourrir lors du naufrage. Ils représentent la plus grande catégorie de décés, ce qui peut sembler cohérent puisqu'ils étaient les plus nombreux sur le bateau .

Relation entre A et S :

Ici, nous avons recatégorisé les valeurs de la variable Survived afin de faciliter la compréhension du graphique:

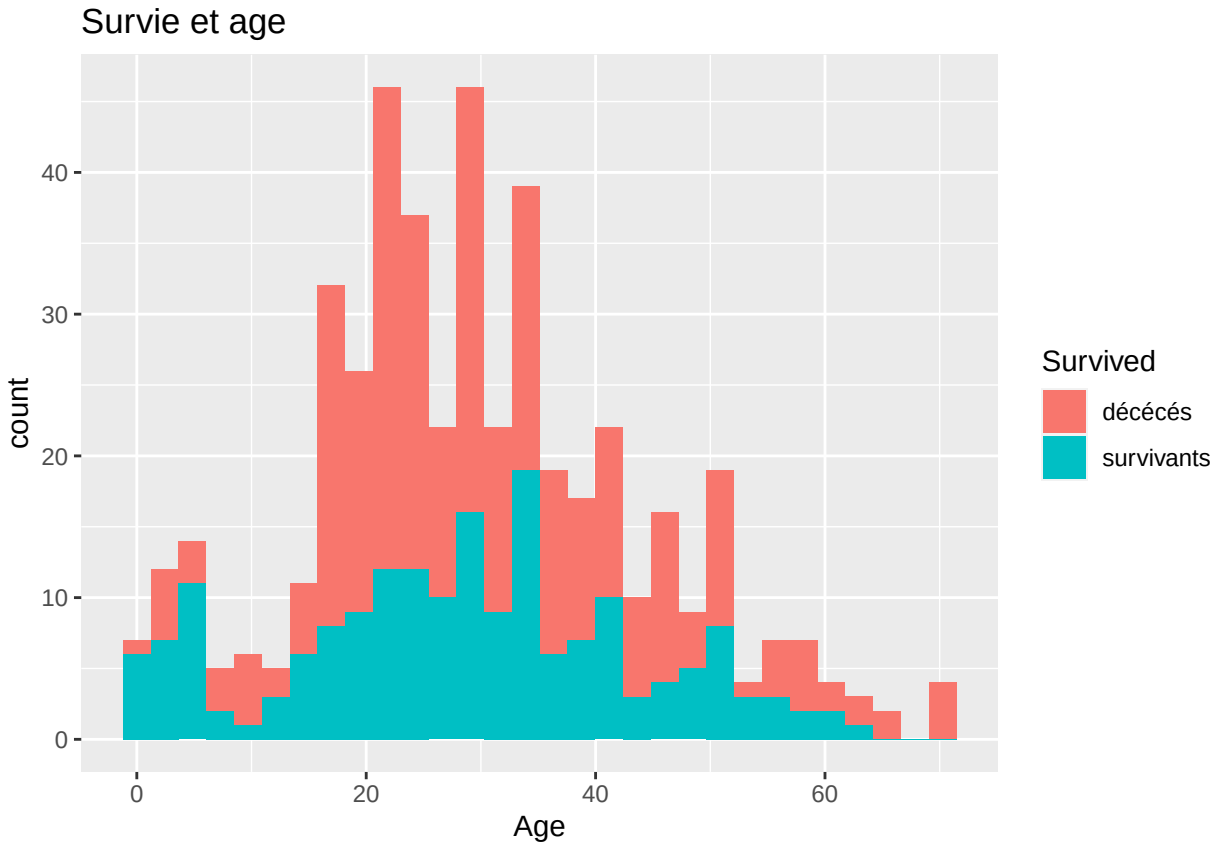
- 1 devient "Survivant" et 0 "Décédé"

```r
df_q5 <-  train%>%
  mutate(Survived= factor(Survived,
```

```
                          levels = c(0, 1, "NA"),
                          labels = c("décécés", "survivants", "blank")))
df_q5%>%
  ggplot(aes(x = Age, fill = Survived)) +
  geom_histogram() +
  labs(title = "Survie et age ")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Removed 121 rows containing non-finite values (stat_bin).
```



Grâce à cet histogramme, nous pouvons voir que les enfants et adolescents (moins de 20 ans) présentaient les plus grandes chance de survie. A contrario, les personnes agées (+60 ans ,derniere tranche de catégorie d'âge) avaient presque 0% de chance de survie. En remplaçant la variable Age avec cAge nous pouvons davantage illustrer cette situation

Nous allons calculer l'age des survivants et des personnes décédées et les stocker dans des vecteurs

```
survived_age= train$Age[train$Survived ==1]
deceased_age =train$Age[train$Survived ==0]
summary(survived_age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    0.75   19.00   29.00   28.37   38.00   62.00      36
```

```
summary(deceased_age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    1.00   21.00   28.00   30.36   38.25   71.00      85
```

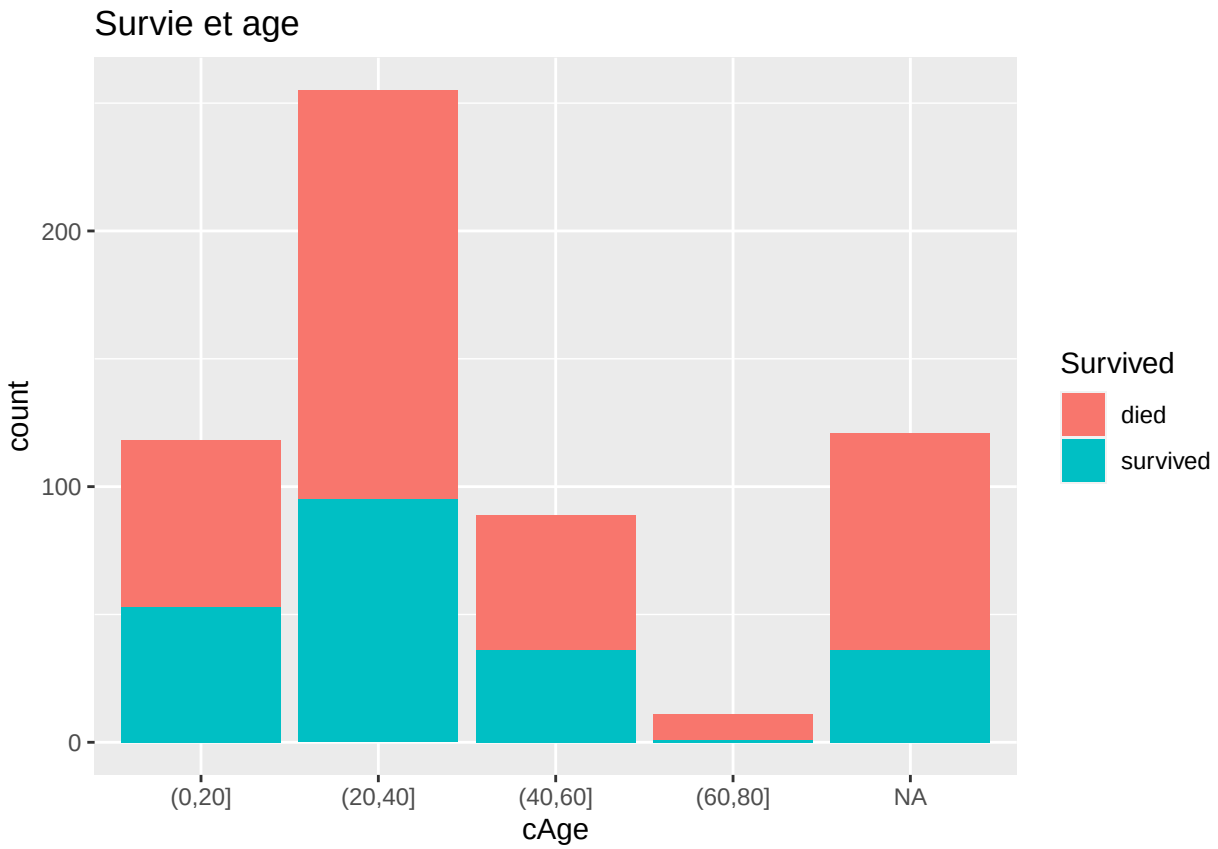
En moyenne les personnes ayant survécu au naufrage avaient 28.37 ans , tandis que les personnes décédées avaient 30.36 en moyenne.

Relation entre cA et S
Nous avons rajouté une colonne cAge au DF train

```
new_train_df <-train
new_train_df$cAge <- cut( new_train_df$Age, breaks=c(0,20,40,60,80))
head(new_train_df)
```

```
##     PassengerId Survived Pclass                                          Name
## 707         707        1      2                 Kelly, Mrs. Florence "Fannie"
## 706         706        0      2 Morley, Mr. Henry Samuel ("Mr Henry Marshall")
## 566         566        0      3                           Davies, Mr. Alfred J
## 244         244        0      3                 Maenpaa, Mr. Matti Alexanteri
## 825         825        0      3                  Panula, Master. Urho Abraham
## 754         754        0      3                            Jonkoff, Mr. Lalio
##        Sex Age SibSp Parch            Ticket    Fare Cabin Embarked    cAge
## 707 female  45     0     0            223596 13.5000  <NA>        S (40,60]
## 706   male  39     0     0            250655 26.0000  <NA>        S (20,40]
## 566   male  24     2     0         A/4 48871 24.1500  <NA>        S (20,40]
## 244   male  22     0     0 STON/O 2. 3101275  7.1250  <NA>        S (20,40]
## 825   male   2     4     1           3101295 39.6875  <NA>        S  (0,20]
## 754   male  23     0     0            349204  7.8958  <NA>        S (20,40]
```

```
df_q5 <- new_train_df%>%
  mutate(Survived= factor(Survived,
                          levels = c(0, 1, "NA"),
                          labels = c("died", "survived", "blank")))
#df_q5 <-na.omit(df_q5$cAge)
df_q5%>%
  ggplot(aes(x = cAge, fill = Survived)) +
  geom_bar() +
  labs(title = "Survie et age ")
```



Ce graphique vient confirmer le propos précédent : les personnes les plus jeunes sur le bateau avaient le plus de chance de survivre.

Question 6

Commenter les résultats obtenus en formulant une première hypothèse quant à la survie des passagers selon les différentes valeurs de P, Sx, et cA.

Nous pouvons émettre plusieurs hypothèses :

- Une femme avait plus de chance de survivre qu'un homme lors du naufrage.
- Un voyageur en classe 1 avait plus de chance de survivre que celui voyageant en classe 2 ou 3.
- Une personne agée entre 0 et 20 ans avait plus de chance de survie.

Une personne qui croise toutes ces caractéristiques (ie être une femme agée entre 0 et 20 ans et voyageant en classe 1) est supposé présenter le plus fort taux de survie sur le bateau.

Question 7

A l'aide de tests statistiques:

- **vérifier si l'age moyenne des passagers est différente de 30**
- **vérifier si l'age moyenne des passagers ayant survécu est inférieure à 30**
- **vérifier si l'age moyenne des passagers n'ayant pas survécu est supérieure à 30**

- *Vérifier si l'age moyenne des passagers est différente de 30*

```r
#on teste h1 aussi avec le test de student
ans <-t.test(train$Age, mu=30, conf.level = .90)
ans
```

```
##
##  One Sample t-test
##
## data:  train$Age
## t = -0.63866, df = 472, p-value = 0.5234
## alternative hypothesis: true mean is not equal to 30
## 90 percent confidence interval:
##  28.48981 30.66663
## sample estimates:
## mean of x
##  29.57822
```

H0 = l'age moyen des passagers est différent de 30.
H1 = l'age moyen des passagers est de 30 ans.
H1 est vraie si la moyenne n'est pas égal à 30. H1 est vraie car x est dans l'intervalle de confiance. H0 est rejeté

Vérifier si l'age moyen des passagers ayant survécu est inférieure à 30

H0 = l'age moyen des passagers ayant survécu est inférieur à 30.
H1 = l'age moyen des passagers ayant survécu est supérieur à 30.
L'âge moyen est dans l'intervalle de confiance, on ne rejette pas H0

```r
# la je verifie h0
ans <-t.test(survived_age, mu=30, alternative = "greater", conf.level= .90)
ans
```

```
##
##  One Sample t-test
##
## data:  survived_age
## t = -1.4788, df = 184, p-value = 0.9295
## alternative hypothesis: true mean is greater than 30
## 90 percent confidence interval:
##  26.94778      Inf
## sample estimates:
## mean of x
##  28.36757
```

Vérifier si l'age moyenne des passagers n'ayant pas survécu est supérieure à 30

H0 = l'age moyen des passagers n'ayant survécu est supérieur à 30
H1 = l'age moyen des passagers n'ayant survécu est inférieur à 30
L'age moyen est dans l'intervalle de confiance, on ne rejette pas H0

```r
# la je verifie h0
ans <-t.test(deceased_age, mu=30, alternative = "less", conf.level= .90)
ans
```

```
##
##  One Sample t-test
##
## data:  deceased_age
## t = 0.43455, df = 287, p-value = 0.6679
## alternative hypothesis: true mean is less than 30
## 90 percent confidence interval:
##      -Inf 31.40794
## sample estimates:
## mean of x
##   30.3559
```

Question 8 :

On peut estimer la probabilité de survie conditionnellement à la valeur d'une autre variable, à l'aide de formules du type avec n1,female = nombre de survivants parmi tous les passagers femmes et nfemale = nombre total de passagers femmes. Estimer

- $\mathbf{P(S = 1|Sx = female)}$
- $\mathbf{P(S = 1|Sx = male)}$
- $\mathbf{P(S = 1|P = 1)}$
- $\mathbf{P(S = 1|P = 2)}$
- $\mathbf{P(S = 1|P = 3)}$
- $\mathbf{P(S = 1|cA = (0,20])}$
- $\mathbf{P(S = 1|cA = (20,40])}$
- $\mathbf{P(S = 1|cA = (40,60])}$
- $\mathbf{P(S = 1|cA = (60,80])}$

```
genre =train$Sex[train$Survived == 1]
c_female = count(genre=="female")
c_male = count(genre=="male")

class = train$Pclass[train$Survived== 1]
c_class1 = count(class == 1)
c_class2 = count(class == 2)
c_class3 = count(class == 3)

tranche_age = cAge[train$Survived == 1]
cA_0_20 = count(tranche_age == "(0,20]")
cA_20_40 = count(tranche_age == "(20,40]")
cA_40_60 = count(tranche_age == "(40,60]")
cA_60_80 = count(tranche_age == "(60,80]")

#selon le genre
sprintf("• P(S = 1|Sx = female) : %f" ,sum(c_female)/sum(Sx=="female") )
```

```
## [1] "• P(S = 1|Sx = female) : 0.756219"
```

```
sprintf("• P(S = 1|Sx = male): %f",sum(c_male)/sum(Sx=="male"))
```

```
## [1] "• P(S = 1|Sx = male): 0.175573"
```

```r
#selon la classe du billet
sprintf("• P(S = 1|P = 1) : %f",sum(c_class1) /sum(train$Pclass==1))
```

```
## [1] "• P(S = 1|P = 1) : 0.654676"
```

```r
sprintf("• P(S = 1|P = 2): %f",sum(c_class2) /sum(train$Pclass==2))
```

```
## [1] "• P(S = 1|P = 2): 0.451613"
```

```r
sprintf("• P(S = 1|P = 3): %f",sum(c_class3) /sum(train$Pclass==3))
```

```
## [1] "• P(S = 1|P = 3): 0.223565"
```

```r
#selon la catégorie d'age
sprintf("• P(S = 1|cA = (0,20]): %f",sum(cA_0_20, na.rm= TRUE)/
          sum(cAge=="(0,20]",na.rm = TRUE))
```

```
## [1] "• P(S = 1|cA = (0,20]): 0.449153"
```

```r
sprintf("• P(S = 1|cA = (20,40]): %f",sum(cA_20_40, na.rm= TRUE)/
          sum(cAge=="(20,40]",na.rm = TRUE))
```

```
## [1] "• P(S = 1|cA = (20,40]): 0.372549"
```

```r
sprintf("• P(S = 1|cA = (40,60]), %f",sum(cA_40_60, na.rm= TRUE)
        / sum(cAge=="(40,60]",na.rm = TRUE))
```

```
## [1] "• P(S = 1|cA = (40,60]), 0.404494"
```

```r
sprintf("• P(S = 1|cA = (60,80]): %f",sum(cA_60_80, na.rm= TRUE)
        / sum(cAge=="(60,80]",na.rm = TRUE))
```

```
## [1] "• P(S = 1|cA = (60,80]): 0.090909"
```

Les femmes avaient 76% de chance de survie sur le bateau, tandis que les hommes en avait 18% environ.

Les probabilités de survie conditionellement à la classe de voyage occupée par le passager sont les suivantes :

- 65% pour la classe 1
- 45% pour la classe 2
- 37% pour la classe 3

Les probabilités de conditionnement selon les tranches d'âge :

- [0:20] ans : ils avaient 45% de chance de survie , soit le plus fort taux.
- [20:40] ans : ils avaient 37% de chance de survie.
- [40:60] ans : ils avaient 40% de chance de survie .
- [60:80] ans : ils avaient 9% de chance de survie , soit le plus faible taux.

Questions bonus :Classification naive de Bayes

Question 9

```
#selection de variables

tc_class = tally(~ cAge + Pclass, data = train, margins = T)
tc_class
```

```
##          Pclass
## cAge        1   2   3 Total
##   (0,20]    9  23  86   118
##   (20,40]  61  74 120   255
##   (40,60]  42  19  28    89
##   (60,80]   6   3   2    11
##   <NA>     21   5  95   121
##   Total   139 124 331   594
```

tally() nous donne une table de contingence, mais nous voulons des fréquences , nous allons donc utliser prop.table

```
S_P <- prop.table(table(train$Pclass,train$Survived ), margin = 2)
S_cA <- prop.table(table(new_train_df$cAge,train$Survived ), margin = 2)
S1 <- prop.table(table(train$Survived ), margin = 1)
S_Sx <-  prop.table(table( train$Sex ,train$Survived), margin = 2)

#On peut donner des noms aux lignes et aux colonnes
#pour faciliter l'accès aux différents éléments de la table:
#S1 : 1 lignes 2 colonnes
names(S1)<-c('0', "1")

#S_P : 3 lignes 2 colonnes
rownames(S_P)<-c("1","2","3")
colnames(S_P)<-c("0","1")
#S_cA : 4 lignes 2 colonnes
rownames(S_cA)<-c("1","2","3","4")
colnames(S_cA)<-c("0","1")
#S_Sx  : 2 lignes 2 colonnes
rownames(S_Sx)<-c("1","2")
colnames(S_Sx)<-c("0","1")

S_Sx
```

```
##
##             0         1
##   1 0.1313673 0.6877828
##   2 0.8686327 0.3122172
```

```
S_P
```

```
##
##             0         1
##   1 0.1286863 0.4117647
##   2 0.1823056 0.2533937
##   3 0.6890080 0.3348416
```

```
S_cA
```

```
##
##               0           1
##   1 0.225694444 0.286486486
##   2 0.555555556 0.513513514
##   3 0.184027778 0.194594595
##   4 0.034722222 0.005405405
```

```
S1
```

```
## 0 1
## 1 1
```

Question 10

Remarque : la fonction ne retourne pas un pourcentage, d'où la dernière ligne en commentaire

```
prob_prediction<-function(Sex, Pclass, c_Age) {
  numerator =   S_cA[c_Age,"1"] *S1["1"] *S_Sx[Sex,"1"] * S_P[Pclass,"1"]
  denominator =  S_cA[c_Age,"1"] *S1["1"] *S_Sx[Sex,"1"] * S_P[Pclass,"1"]
  + S_cA[c_Age,"0"] *S1["0"]* S_Sx[Sex,"0"] * S_P[Pclass,"0"]

  return(numerator/denominator)
}


#prob_prediction("female",3,"(20;40]")
```

Conclusion

Dans ce devoir d'analyse il nous a été demandé d'analyser principalement quatre variables : age , sex, survie et classe de voyage.
L'analyse de ce jeu de données nous a permis de mettre en évidence les cas de figure où la probabilité de survivre était la plus élevée : être une femme , une personne jeune (entre 0 et 20 ans) et voyager en 1ere classe.
Pour aller plus loin, il serait intéressant de croiser davantage de variables afin d'expliquer le faible taux de survie des hommes. Cela peut se faire également en implementant des modèles de prédictions plus poussés.