Données

Les baes de donnnées seront chargée de la manièer suivante :

```
BO = read.table("http://graal.ens-lyon.fr/~fchuffart/files/data/bosson.csv", header=TRUE, sep=";")
head(BO)
FE = read.table("http://graal.ens-lyon.fr/~fchuffart/files/data/ferretti.csv", header=TRUE, sep=";")
head(FE)
HE = read.table("http://graal.ens-lyon.fr/~fchuffart/files/data/her.csv", header=TRUE, sep=" ")
head(HE)
TI = read.table("http://graal.ens-lyon.fr/~fchuffart/files/data/titanic.csv", header=TRUE, sep=";")
head(TI)
```

Description d'une variable qualitative.

Exercice 1

On travaille dans cet exercice sur la base de données `titanic.csv`. Cette base contient les informations de 1046 passagers du Titanic :

- `pclass` la classe dans laquelle ils ont voyagé (1ere, 2eme ou 3eme classe)
- `survived` : `yes` ou `no` selon s'ils ont survit au naufrage ou non
- `gender` : sexe (`F` ou `M`)
- `age` : l'age en années

1. Charger la base de données `titanic.csv`. A quoi servent les paramètres "`header`" et "`sep`" ?
2. Afficher les 6 premieres lignes de la base : `head(TI)` Ceci permet de vérifier que la base a bien été importée (nom des colonnes et format des variables) et d'en donner un premier apercu.
3. Afficher le nom des colonnes de la base en utilisant la fonction appropriée que vous chercherez.
4. Affecter la colonne `pclass` à la variable `P`, la colonne `survived` à la variable `S`, la colonne `gender` à la variable `G` et la colonne `age` à la variable `A`. Cela permettra dans la suite d'utiliser des noms raccourcis. `P <- TI[,1]` On peut aussi utiliser `P <- TI[,"pclass"]` ou `P <- TI$pclass` Quel est le type de ces 4 variables ?
5. Afficher les lignes 250 à 257. `TI[250:257, ]` Afficher les lignes 3, 45, 73. Afficher les colonnes 1 et 3. Afficher les colonnes 2 et 4 des lignes 67, 83, 101.
6. Afficher les données des passagers de première classe. `TI[P==1,] # donnees des passagers de premiere classe` ou `TI[which(P==1),] # donnees des passagers de premiere classe` Afficher les données des passagers de troisième classe. Afficher les données des passagers de deuxième et troisième classe. `TI[(P==2 | P==3),]` ou `TI[-(P==3),]`
7. Afficher les données des femmes de première classe. `TI[which(P==1 & G=="F"),]` Il faut toujours mettre des guillemets pour les valeurs de variables "chaines de caractères". Afficher les données des hommes de deuxieme classe.
8. Afficher les données des bébés (moins de 1 an). `TI[A<1,] # données des bébés` Afficher les données des enfants (moins de 18 ans), des adolescents (12-18 ans) des adultes (plus de 18 ans).
9. Calculer les effectifs et proportions des 3 classes de passagers. Quelle est la proportion des passagers en première classe ? Calculer les effectifs et proportions des survivants et des décédés. Quelle est la proportion de survivants ? Calculer les effectifs et les proportions d'hommes et femmes. Quelle est la proportion de femmes ? Représenter graphiquement ces 3 variables.
10. Tracer la répartition des 3 classes de passagers selon la survie et la survie selon la classe de voyage. Combien de passagers de première classe ont survécu ? Quelle proportion parmi tous les passagers

étaient en première classe et ont survécu ? Quelle proportion, parmi les passagers de première classe, ont survécu ? Quelle proportion de survivants était des passagers de première classe ? Répéter pour la 2ème et 3ème classe.

11. Tracer la répartition des 3 classes de passagers selon le genre et le genre selon la classe de voyage. Combien de passagers de première classe sont des femmes ? Quelle proportion parmi tous les passagers étaient en première classe et sont des femmes ? Quelle proportion, parmi les passagers de première classe, sont des femmes ? Quelle proportion de femmes était des passagers de première classe ? Répéter pour la 2ème et 3ème classe.

Exercice 2

On travaille dans cet exercice sur la base de données `bosson.csv` (fournies par le Professeur Jean-Luc Bosson). Cette base contient les informations de 209 patients venant de France ou du Vietnam :

— `country` : `Vietnam` ou `France`
— `gender` : `F` ou `M`
— `aneurysm` : taille de l'anévrisme en mm
— `bmi` : indice de masse corporelle
— `risk` : nombre de facteurs de risque entre 0 et 5

1. Charger la base de données `bosson.csv`.
2. Afficher les 6 premieres lignes de la base. Afficher le nom des colonnes de la base.
3. Affecter la colonne `country` à la variable `C`, la colonne `gender` à la variable `G`, la colonne `aneurysm` à la variable `1`, et la colonne `risk` à la variable `R`. Quel est le type de ces variables ?
4. Afficher les lignes 120 à 123. Afficher les colonnes 1 et 3 des lignes 67, 83, 101. Afficher les données des patients vietnamiens. Afficher le nombre de facteurs de risque des hommes.
5. Calculer les effectifs et proportions des 2 pays. Quelle est la proportion de vietnamiens ? Calculer les effectifs et les proportions d'hommes et femmes.
6. Tracer la répartition des pays selon le genre et le genre selon le pays. Combien de francais sont des hommes ? Quelle proportion parmi tous les patients sont des hommes francais ? Quelle proportion, parmi les patients vietnamiens, sont des femmes ? Quelle proportion de femmes sont vietnamiennes ?
7. Tracer la répartition des pays selon le nombre de facteurs de risque et le nombre de facteurs de risque selon le pays. Combien de vietnamiens ont 0 facteur de risque ? Quelle proportion parmi tous les patients sont des vietnamiens sans facteur de risque ? Quelle proportion, parmi les patients vietnamiens, ont 0 facteur de risque ? Quelle proportion de patients sans facteur de risque sont vietnamiens ?
8. Tracer la répartition des genres selon le nombre de facteurs de risque et le nombre de facteurs de risque selon le genre. Combien de femmes ont 0 facteur de risque ? Quelle proportion parmi tous les patients sont des femmes sans facteur de risque ? Quelle proportion, parmi les patients femmes, ont 0 facteur de risque ? Quelle proportion de patients sans facteur de risque sont des femmes ?

Description d'une variable quantitative.

Exercice 3

On étudie l'âge de mères non-fumeuses au moment de leur accouchement.

age	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35
effectifs	7	8	9	10	12	3	2	5	4	5	2	4	2	1	1

1. Est ce une variable qualitative ou quantitative ?

2. Créer un vecteur avec l'ensemble de ces données : `ages <- c(rep(21,7), rep(22,8),rep(23,9),rep(24,10),rep(25,1`
`rep(26,3),rep(27,2),`
`rep(28,5),rep(29,4),rep(30,5),rep(31,2), rep(32,4),rep(33,2), 34, 35)` Afficher le vecteur `ages`
3. Combien d'individus y a-t-il dans la base ?
4. Calculer les effectifs de chaque niveau : `EffAge<-table(ages) EffAge`
5. Calculer les fréquences de chaque age. Quelle est la proportion de mères de 25 ans ? de 30 ans ? de 35 ans ?
6. Combien de femmes ont moins de 30 ans ? plus de 28 ans ? Quelle proportion de femmes ont moins de 30 ans ? plus de 28 ans ?
7. Trouver la limite d'âge telle que 5% des femmes soient en dessous ? au dessus ?
8. Quel est l'âge moyen des mères non-fumeuses au moment de leur accouchement ? `mean(ages)`
9. Calculer la variance empirique (corrigée). `var(ages)`
10. Calculer l'ecart-type (corrigé) de l'échantillon. `sd(ages)`
11. Calculer la médiane et les quartiles de l'échantillon : `summary(ages) # quelques indicateurs statistiques median(ages) # mediane quantile(ages,probs=c(0.25,0.5,0.75)) # quartiles`
12. Tracer les effectifs empiriques avec un histogramme : `hist(ages)` Superposer une ligne verticale rouge représentant la moyenne, une bleue représentant la médiane, deux vertes représentant le premier et troisième quartiles. `abline(v=mean(ages),col="red") # ligne verticale à la moyenne abline(v=median(ages),col="nlue") # ligne verticale à la médiane abline(v=quantile(ages,0.25)` `# ligne verticale au premier quartile abline(v=quantile(ages,0.75),col="green") # ligne verticale au 3eme quartile`
13. Tracer la fonction de répartition empirique avec `plot(ecdf(ages))`. Superposer une ligne verticale bleue représentant la médiane, deux vertes représentant le premier et troisième quartiles. `abline(h=0.5,col="red") # ligne horizontale à 0.5 abline(v=median(ages),col="red")` `# ligne verticale à la médiane abline(h=0.25,col="blue") # ligne horizontale à 0.25` `abline(v=quantile(ages,0.25),col="blue") # ligne verticale au premier quartile abline(h=0.75,col="gre` `# ligne horizontale à 0.75 abline(v=quantile(ages,0.75),col="green") # ligne verticale au 3eme quartile`
14. Tracer la boite à moustaches de l'échantillon. `boxplot(ages)` Que représente chaque ligne ? Superposer une ligne horizontale rouge représentant la moyenne, une bleue représentant la médiane, deux vertes représentant le premier et troisième quartiles. `boxplot(ages)`
15. Sur quel graphique peut-on lire la fréquence de l'intervalle $[22; 25]$?

Simulation d'un échantillon

Exercice 4

1. Définir `x` le vecteur `c(0,1)`.
2. Permuter aléatoirement `x`. Répéter 10 fois. Combien de fois avez vous eu les deux permutations possibles ?
3. Pour `n=100`, créer un vecteur `E` d'un échantillon de `n` valeurs tirées aléatoires dans `x` avec remise.
4. Calculer les effectifs et les proportions de 0 et de 1 dans `E`. Tracer un diagramme en barre.
5. Calculer la moyenne de `E`. Est ce que la différence entre la moyenne et 0.5 est grande ?
6. Répéter avec `n=10000`.
7. Recommencer les questions 4 et 5 en tirant un échantillon de taille `n=1000` de valeurs de `x`, où la probabilité de 0 est 1/3 et la probabilité de 1 est 2/3. Quelle est la moyenne ?

Calcul de probabilités et de quantiles

Exercice 5

On sait par expérience qu'une certaine opération chirurgicale a 80% de chance de succès. On réalise cette opération chez 5 patients. Soit X la variable aléatoire qui modélise le nombre de succès parmi ces 5 patients.

1. Quelle loi (quel modèle) proposez vous pour X ?
2. On veut calculer la probabilité que l'opération échoue les 5 fois.
   (a) Ecrire mathématiquement cette probabilité. La calculer avec RStudio.
   (b) On peut la calculer directement avec RStudio, en ayant spécifié au préalable les valeurs de n et p : `dbinom(0,n,p)`
   (c) On peut aussi la retrouver en faisant des simulations de la variable X. Par exemple, on commence en réalisant 100 fois l'expérience (chez 500 patients au total). Simuler X 100 fois `n <- 5; p <- 0.8 # parametres N <- 100 # taille de l'échantillon X <- rbinom(N,n,p) # échantillon simulé`
   (d) Calculer les fréquences empiriques de votre échantillon. Combien de fois la modalité 1 est-elle apparue ? `table(X) table(X)/N`
   (e) Visualiser votre échantillon à l'aide d'un diagramme en barres. `barplot(table(X))`
   (f) Pour calculer numériquement la probabilité d'avoir 5 échecs, il faut augmenter la taille de l'échantillon à $N = 100\,000$. Simuler votre échantillon.
   (g) A quelle modalité que l'on appellera `mod` l'événement "5 échecs" correspond-il ? Calculer sa fréquence avec la commande : `F <- length(which(X==mod))/N`
   (h) Pour une meilleure estimation de la probabilité d'avoir 5 échecs, il faut augmenter la taille de l'échantillon à $N = 100\ 000$. Simuler votre échantillon et calculer la fréquence de l'évènement "5 échecs".
   (i) Comparer ces valeurs avec la valeur théorique.
3. On s'intéresse maintenant à la probabilité que l'opération échoue exactement 2 fois parmi les 5 patients.
   (a) Calculer exactement cette probabilité avec RStudio.
   (b) On va calculer numériquement cette probabilité avec des simulations.
   (c) A quelle modalité correspond l'évènement d'intérêt ? Calculer la fréquence empirique de cet événement.

Distribiutions

Exercice 6

Dans cette première partie, on cherche à tracer la densité d'une loi normale et à comprendre l'influence de ses paramètres.

1. Représenter sur le même graphique les densités des lois normales $\mathcal{N}(0,1)$ et $\mathcal{N}(0,10)$. On rappelle que la fonction dnorm prend en argument l'écart-type et pas la variance : `curve(dnorm(x), xlim=c(-3,3), ylim=c(0,1.5), col="red", lwd=2)`
   `curve(dnorm(x,0,sqrt(10)), add=TRUE, col="green", lwd=2)`
2. Rajouter sur le même graphique les densités des lois normales $\mathcal{N}(0,0.1)$ et $\mathcal{N}(1,1)$. Commenter.
3. Simuler un échantillon de taille 100 de la loi $\mathcal{N}(0,1)$ `X<-rnorm(100,0,1)`
4. Représenter un histogramme en fréquences de cet échantillon. Superposer sur le même graphique la densité de la loi $\mathcal{N}(0,1)$.
5. Recommencer avec un échantillon de taille 5000. Commenter.