

Big Data

TD/TP 5 : Théorie et Pratique

BUT 3

Guillaume Metzler et Antoine Rolland
Institut de Communication (ICOM)
Université de Lyon, Université Lumière Lyon 2
Laboratoire ERIC UR 3083, Lyon, France

guillaume.metzler@univ-lyon2.fr; antoine.rolland@univ-lyon2.fr


Exercice 1 : Volumétrie et tests statistiques

Dans cet exercice, on se propose d'étudier, de façon empirique, la souplesse des tests statistiques dans un contexte où nous disposons d'une grande volumétrie de données. Dans la cas présent, d'un grand nombre d'exemples.

Plus précisément, nous cherchons à étudier un test la significativité d'un test de comparaison de moyennes sur deux échantillons :

- un premier échantillon de taille n qui suit une loi normale centrée et réduite.
- un deuxième échantillon suivant une loi normale de moyenne $\mu = 0.005$ et de variance égale à 1.

On va maintenant étudier ces deux échantillons.

1. Quel test statistique doit-on effectuer pour comparer les moyennes de ces deux échantillons ? Pourquoi ? Quelle est la distribution de probabilités associée à ce test ?
2. On cherche maintenant à étudier le nombre de fois où le test de comparaison de ces deux moyennes est significatif, en fonction de la taille n de l'échantillon. Pour cela, vous devrez :
 - (a) simuler un échantillon de taille n dont les données $x_i \underset{i.i.d.}{\sim} \mathcal{N}(0, 1)$
 - (b) simuler un échantillon de taille n dont les données $x_i \underset{i.i.d.}{\sim} \mathcal{N}(0.005, 1)$
 - (c) Pour une valeur donnée de n , effectuer le test de comparaisons de moyennes sous  et déterminer si le test est significatif ou non avec un seuil $\alpha = 0.05$.
3. Que représente la valeur α dans la question précédente ?
4. Répéter les questions 2 cent fois pour une valeur de n fixée et regarder, en pourcentage, le nombre de fois où le test est significatif.
5. Répéter la question 3 et regarder comment évolue ce pourcentage en fonction de la valeur de n .

Exercice 2 : Imputations valeurs manquantes

On se propose d'étudier l'impact sur la moyenne de différentes méthodes de complétions des valeurs manquantes à l'aide de quatre méthodes distinctes :

- imputation par la moyenne
- imputation par la valeur 0
-
- imputation par l'algorithme k -NN (les k plus proches voisins)

Pour cela, on va considérer le jeu de données suivant :

```
# On fixe la graine
set.seed(1)

# On génère un jeu de données à plusieurs dimensions,
# mais avec un nombre d'exemples limité.

p = 2
n = 500

X = matrix(rnorm(n*p,3,2),ncol = 2)
Y_full = -1 + X[,1] - 2*X[,2]

# On va supprimer les 100 premières valeurs de ce vecteur

Y_miss= Y_full
Y_miss[1:100] = NA

# On peut étudier l'histogramme des valeurs de Y

hist(Y_miss,probability = TRUE, col=rgb(1,0,1,1/4),
xlab="", main = "Histogrammes")
```

On représentera et commentera l'histogramme des valeurs du vecteur Y complété par les méthodes suivantes :

1. en faisant une imputation par la valeur 0,
2. en effectuant une imputation par la moyenne des valeurs de Y observées,
3. en complétant avec une valeur au hasard tirée suivant une loi normale de moyenne la moyenne empirique observée et d'écart-type l'écart-type observé),
4. en complétant les valeurs manquantes à l'aide d'un algorithme du plus proche voisin, *i.e.* en prenant la valeur $k = 1$ (on pourrait aussi tester des valeurs de k différentes si on le souhaite).