Travaux Pratiques Analyse Statistique et R 01

Analyses descriptives sur R

Alison PATOU
patou.alison@gmail.com
30/01/2020

Introduction

Dans ce TP nous allons approfondir la manipulation des objets sous R, faire des analyses descriptives et nous verrons également les tests d'hypothèse.

Exercice 1

Info: Nous travaillons sur le dataset **House Pricing** qui contient 1460 lignes et 81 colonnes.

Question 1

- Charger le jeu de données House Princing (fichier house pricing.csv)
  Donner quelques indicateurs sur la colonne *LotArea* (minimum, maximum, range, moyenne, médiane, ...) sans utiliser la fonction summary().

Question 2

- Calculer la variance sur la colonne *YrSold* grâce à la fonction associée.

Question 3

- Calculer la variance non-corrigée variance sur la colonne *YrSold* sans utiliser la fonction mais en se ramenant à la formule :

$$V(x) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}$$

Question 4

- Calculons maintenant l'écart-type toujours sur sur la colonne *YrSold* grâce à la fonction sd().
  Calculer l'écart-type en prenant la racine carrée de la variance de la question 3.

Question 5

- Quels sont les effectifs associés à chacune des dates contenues dans la colonne *YrSold* ?
  Visualiser cela avec une courbe.

Question 6

• Découpez votre dataset pour ne garder que les 15 dernières colonnes, à savoir :

```
 [1] "WoodDeckSF"   "OpenPorchSF"   "EnclosedPorch" "3SsnPorch"     "ScreenPorch"   "PoolArea"
 [7] "PoolQC"       "Fence"         "MiscFeature"   "MiscVal"       "MoSold"        "YrSold"
[13] "SaleType"     "SaleCondition" "SalePrice"
```

NB : c'est la fonction *names()* qui permet d'obtenir ce résultat (=affichage des noms de colonnes)

Question 7

• Créer une matrice de corrélation grâce au nouveau dataset créé à la question 6 (attention, la correlation ne se fait que sur des variables quantitatives !).

Question 8

•Visualiser les résultats dans une matrice de corrélation (rappel : vous devez importer la bibliothèque corrplot).

Question 9

•Quel est le top 3 des variables que l'on devrait garder pour prédire notre prix ?

Exercice 2

On a relevé les pourcentages de satisfaction des cours reçus par des étudiants d'une école d'ingénieur. Les résultats sont les suivants :

74, 85, 95, 84, 68, 93, 84, 87, 78, 72, 81, 91, 80, 65, 76, 81, 97, 69, 70, 98.

Question 1

• Créer la séquence satisfaction avec la fonction c()

Question 2

• Afficher le tableau des effectifs et vérifier que chacune des modalités est égale à 1 avec la fonction table()

Question 3

• Combien de résultats ont été récoltés ? Utilisez une fonction R pour déterminer ce résultat

Question 4

•Afficher la visualisation boite à moustache de cette séquence. Retrouver les éléments clés grâce à la fonction summary()
Visualiser l'histogramme de ces pourcentages de satisfcation avec 4 classes.
Que pouvez-vous en dire ?

Exercice 3

Info: On traite 3 échantillons de de patients ayant reçu un traitement différent T1, T2, T3. On obtient les résultats suivants :

	Traitement T1	Traitement T2	Traitement T3
Ont guéri	50	72	64
N'ont pas guéri	10	15	11

Question 1

- Créer le jeu de données grâce aux lignes de commande R suivantes :

```
guerison <- matrix(c(50,72,64,10,15,11),nrow=2,byrow=T)
rownames(guerison)<-c("gueri","Pas gueri")
colnames(guerison)<-c("Traitement 1","Traitement 2","Traitement 3")
guerison <- as.table(guerison)
```

Question 2

- On souhaite savoir si le type de traitement à une influence sur la guérison. Pour cela, faites un test du khi2

Question 3

- Afficher la p-value. Que pouvez-vous en dire?

Exercice 4

Info: Nous travaillons sur le fichier Notes_Info_Data.csv qui regroupent les notes des élèves à différents QCM selon leur filière (Info ou Data).

Question 1

- Importer le fichier puis calculer la moyenne des notes.

Question 2

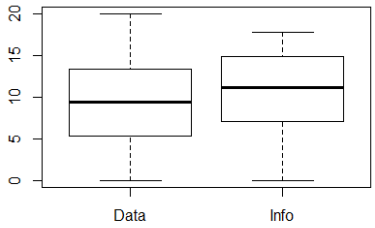
- Maintenant calculer la moyenne des notes selon la filière (Info vs Data). Pour cela vous allez devoir utiliser une condition de sélection.
  exemple : dataset$Note[dataset$Filiere=="Info"]

Question 3

- Calculer l'écart-type de la filière Info puis celui de la filière Data.

Question 4

- Représentez à l'aide d'un boxplot, les notes des Info et ceux des Data, côté à côte.



- NB : en R pour comparer des champs, on utilise souvent le '~'
  exemple : data$age~data$sexe

Question 5

- Calculer le test de Student entre les filière Info et les Data

Question 6

- Quel est la p-value associée ? Que pouvez-vous en conclure : la filière info est-elle meilleure que la filière data ?