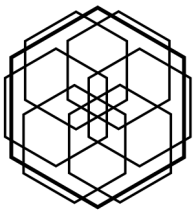


PrivateTeacher

Cours Privés de Science

Traitement des données avec R
Exercice d'entrainement

Julien RUPPEN

20 March, 2024

Abstract

Ce document est un exercice pratique pour te permettre de pratiquer le traitement des données avec R. A l'aide des conditions sur les lignes et les colonnes, exerce-toi à selectionner des données dans un dataframe.

Enoncé

Des mesures ont été faite au sein d'une classe d'étudiants-es de l'université de Heidelberg en Allemagne. En raison de contraintes techniques et par manque de temps, seul une centaine d'individus ont été selectionné parmi cette classe de 300 personnes. L'échantillon ainsi formé est cependant jugé suffisamment grand pour être considéré représentatif de la population.

Observations

Chaque étudiants-es est interrogé au sujet de 5 caractéristiques de sa personne. Deux d'entre elles prennent des valeurs numériques et sont donc des observations quantitatives. On les désigne par les en-tête Obs_01 et Obs_02. Trois autres caractéristique sont de nature qualitatives. Il s'agit de: la couleur de leurs yeux, leur genre et enfin leur appartenance à l'un des trois groupes A, B ou C.

Fichier de donnée

Toutes les observations réalisée sur cet échantillons d'étudiants-es ont été réunie dans un même fichier de donnée. Il s'agit du fichier:

```
## Data-6650-Traitement-Donnees-PrivateTeacher.csv
```

Structure du fichier

Chaque ligne correspond à un individu et dispose donc de 5 colonnes, une colonne par observation. Le fichier de donnée se présente sous la forme d'une liste d'observation dont la structure est la suivante:

```
##    Obs_01 Obs_02 Groupe Genre Yeux
## 1    8.20  16.46      B femme vert
## 2    8.45  10.27      B femme bleu
## 3    2.90   3.83      C homme brun
## 4   16.94  23.15      C femme vert
## 5    8.08  13.08      C femme bleu
## 6    5.17   3.89      C femme brun
## 7    5.41  24.75      A homme bleu
## 8   11.35  19.04      C homme vert
## 9    9.14  20.33      B femme vert
## 10   5.55   8.95      B femme bleu
```

Se concentrer sur la démarche

Le mystère le plus absolu est gardé sur ce que sont les observations 1 et 2. Il peut s'agir de n'importe quelle valeur numérique telle que le poids, la taille, l'age ou le nombre de cheveux

de la personne.. Ce mystère est gardé à dessein, pour te permettre de faire abstraction des détails et de te concentrer sur la démarche.

Cette démarche est toujours la même et consiste en les étapes suivantes:

1) Sélectionne les observations 1 de chaque sous-groupe dans le fichier de donnée.
2) Place ces valeurs dans une variable de ton choix
3) Calcule la moyenne et la variance à l'aide des fonction mean() et var()
4) Trace l'histogramme à l'aide de la commande hist()

Il s'agit là d'une séquence d'étape que nous allons adopter systématiquement !

Questions Pratiques

Une relation de proportionnalité est cachée au sein de notre échantillon. On ne sait pas cependant au sein de quel groupe se trouve cette relation. Cela peut être le groupe des femmes, cela peut-être le groupe des personnes aux yeux bleu. Personne ne le sait. Votre mission si vous l'acceptez est donc de trouver au sein de quels groupes se cache la relation de proportionnalité.

Question 00

```
## Charger le fichier de donnée dans la mémoire de travail (la mémoire vive).
```

Question 01

```
## Selectionner les valeurs appartenant aux membres du groupe C
## qui sont des femmes.
```

Question 02

a)

```
## Tracer un histogramme des valeurs des observations 1 et 2
## au sein du groupe selectionné
```

b)

```
## Calculer la moyenne et la variance des observations 1 et 2
## au sein du groupe selectionné
```

c)

```
## Représenter un boxplot des valeurs des observations 1 et 2
## au sein du groupe selectionné
```

Question 03

```
## Existe-t-il une relation de proprotionnalité
## entre les observations 1 et 2 du groupe selectionné ?
```

Question 04

```
## Quelle est la significativité statistique du coefficient de correlation ?
```

Corrigé Détaillé

Question 00

Pour charger le fichier de donnée dans la mémoire vive, on utilise la commande:

```
## data = read.csv("Data-6650-Traitement-Donnees-PrivateTeacher.csv", sep=" ")
```

Par cette commande, on donne l'instruction au logiciel de lire un fichier dont le format est csv (comma separated value) et de le mettre dans une variable nommée "data". Le format de fichier csv est un format de fichier texte dans lequel chaque valeur est séparée par un caractère de séparation. Dans cette commande, on précise à R que le charactère de séparation est un espace sep=" ".

Question 01

Pour selectionner les données demandée, on formule tout d'abord une condition sur les colonnes:

```
## condition_groupe = data$Groupe == "C"
## condition_genre = data$Genre == "femme"
```

Le résultat de ces condition se trouve maintenant dans les variables "condition". Il s'agit d'un masque que l'on peut utiliser pour n'afficher que les sous-groupe qui nous intéresse. On les utilise comme une contition sur les ligne de notre variable "data"

```
## data_C = data[condition_groupe,]
```

Si on avait voulu faire seulement la deuxième selection, on aurait choisi:

```
## data_femme = data[condition_genre,]
```

Comme on souhaite selectionner les deux on écrira donc:

```
## data_C_femme = data[condition_groupe & condition_genre,]
```

les 5 premières lignes de la selection sont les suivantes:

```
##    Obs_01 Obs_02 Groupe Genre Yeux
## 4   16.94  23.15      C femme vert
## 5    8.08  13.08      C femme bleu
## 6    5.17   3.89      C femme brun
## 16  18.52  23.74      C femme bleu
## 19  15.72  18.60      C femme bleu
```

Question 02

a)

Après avoir selectionné les lignes correspondantes au groupe indiqués en posant nos conditions sur les lignes, on peut maintenant selectionner les colonnes qui contiennent les observations 1 et 2 en posant une condition sur les colonnes. On écrit:

```
## data[condition_groupe & condition_genre, "Obs_01"]
```

On peut aussi utiliser directement la variable qui contient les données selectionnées auparavant:

```
## data_C_femme[,"Obs_01"]
```

Ou plus simplement

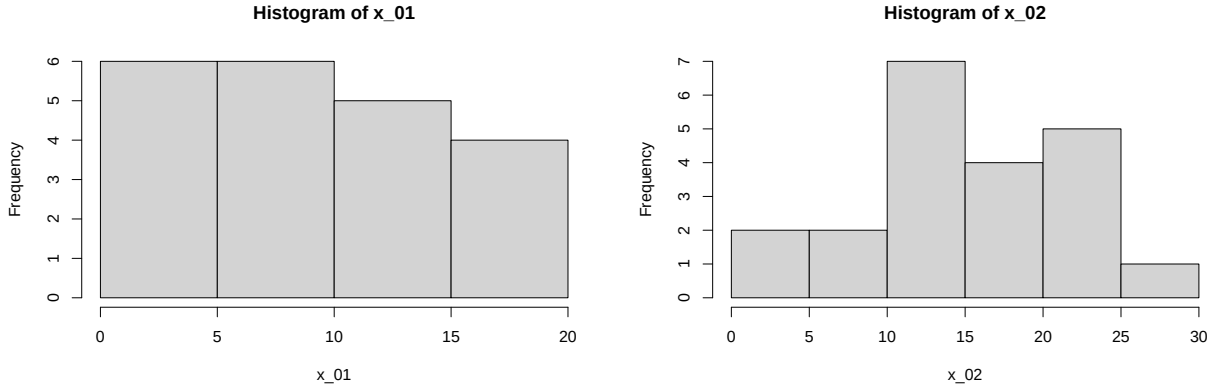
```
## data_C_femme$Obs_01
```

En placant maintenant le résultat de cette sélection dans les variables x_01 et x_02:

```
## x_01 = data_C_femme$Obs_01
## x_02 = data_C_femme$Obs_02
```

On peut afficher les histogrammes correspondants aux observation 1 et 2 des groupes selectionné à l'aide des deux commandes suivantes:

```
## hist(x_01)
## hist(x_02)
```



b)

Avec les valeurs des observations 1 et 2 dans les variable x_01 et x_02 respectivemetn, on calcule tout aussi simplement les valeur de la moyenne et de la variance à l'aide des commandes:

```
## mean(x_01)
## mean(x_02)
```

```
## [1] 8.977143
```

```
## [1] 14.89143
```

```
## var(x_01)
## var(x_02)
```

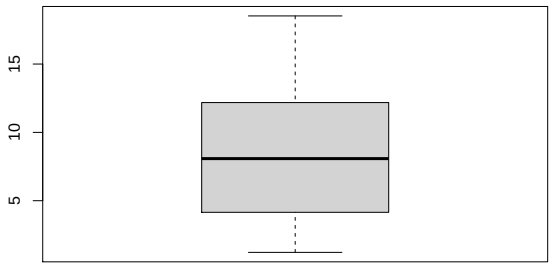
```
## [1] 28.62677
```

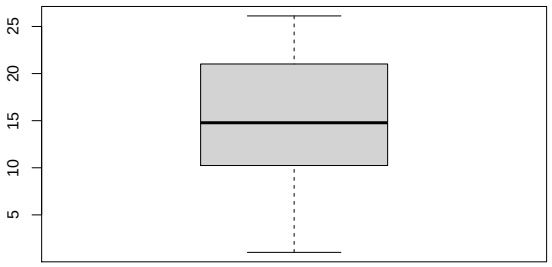
```
## [1] 48.70326
```

c)

De même pour le tracé des boxplots, la commande s'écrit

```
## boxplot(x_01)
## boxplot(x_02)
```

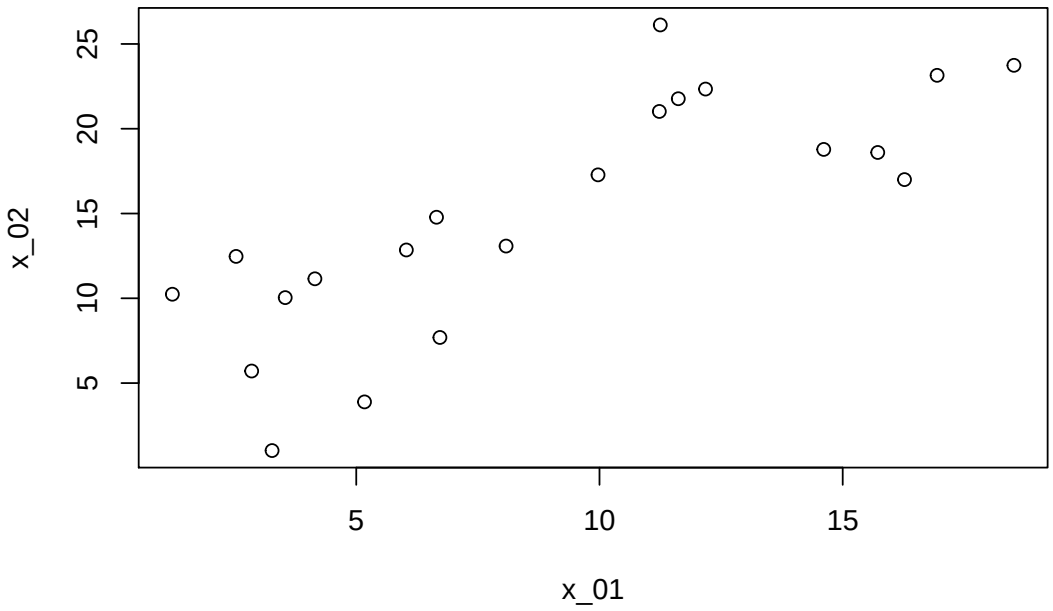




Question 03

Pour savoir s'il existe une relation de proportionnalité entre les observations 1 et 2 du sous-groupe sélectionné, on commence par tracer le nuage de point (scatterplot) entre les observations 1 et 2. On utilise pour cela la commande:

```
## plot(x_01,x_02)
```



Maintenant que l'on se fait une représentation visuelle de la relation qui existe peut-etre entre les deux observations, il est utile de la quantifier en terme de coefficient de correlation. Nous utiliserons ici les coefficient de correlation de Pearson:

```
## cor(x_01, x_02, method="pearson")
## [1] 0.7951669
## Le coefficient de correlation est plus grand que 0.6, il est donc raisonnable
## d'affirmer qu'il existe une relation de proportionnalité entre les deux séries
## d'observations faites au sein du groupe selectionné. Le graphique semble confirmer
## cette conclusion en montrant un alignement du nuage de points. Une relation
## n'existe peut-être pas, mais alors elle est très plausible.
## A ce stade, il faut conduire un test d'hypothèse pour s'en assurer.
```

Question 04

A ce stade, il convient donc de tester la significativité statistique du coefficient calculé. On exécute donc la commande suivante

```
## result = cor.test(x_01,x_02)
```

```
##
##  Pearson's product-moment correlation
##
## data:  x_01 and x_02
## t = 5.7159, df = 19, p-value = 1.649e-05
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.5534637 0.9133384
## sample estimates:
##       cor
## 0.7951669
```

Avec cette commande, nous testons l'hypothèse selon laquelle les variables ne sont pas proportionnelles. (on dit aussi: ne sont pas corrélées linéairement). L'hypothèse de départ H0 est donc que la valeur du coefficient de corrélation est égale à 0.

On retrouve ici le coefficient de correlation calculé auparavant mais cette fois-ci, la valeur-p associée nous permet de quantifier la significativité statistique de la valeur du coefficient.

Une valeur-p supérieur à 5% nous indique que le coefficient de correlation n'est pas statistiquement significatif. Il est donc raisonnable d'accepter H0. Dans le contexte de notre test, cela signifie donc que la valeur du coefficient de correlation est probablement proche de 0.

Si au contraire la valeur-p est inférieur à 5%, cela signifie que la valeur calculée du coefficient de corrélation est statistiquement significative. Cela constitue un élément en faveur de H1. Il est donc raisonnable d'affirmer que la valeur du coefficient de correlation est significativement différente de 0.

Conclusion

Et c'est ainsi que se conduit une analyse de donnée reposant sur la sépéaration entre sous groupe. Tu est à présent capable de selectionner des donnée au sein d'une série d'observation et d'en faire une analyse. Cette analyse est certe élémentaire, mais elle constitue la base de ce qui te permettra dorénavant de faire toute sorte d'analyse élémentaire, aussi bien que des analyse plus poussée.

Ce tutoriel pour vous montrer qu'une fois que les données sélectionner, toutes les autres commande sont élémentaire

Continue à pratiquer, c'est cela qui te permettra de répondre rapidement aux questions d'examen

Ces exercices sont fait pour vous entrainer.

Reprenez ces feuilles pour vous entrainer. Placez vous en situation d'examen et s'exercer à répondre aux question.

Tu trouvera ici tous les paramètre utilisé pour générer les data. Tu peux t'en servire pour vérifier les valeurs des moyenne et variance ainsi que les paramètre de la droite pour tous les sous-groupe du data set dont il est question dans cette série.

Paramètres de la simulation

```
##   Groupe N_Obs Mean Variance Slope Intercept
## 1      A    35    9        4     4         5
## 2      B    41   10        4     1         5
## 3      C    38    9        5     1         5
```

```
## [1] "Random seed is: 6650"
```

Tableau de de toute les moyenne et variance de tous les sous groupe

Les valeurs selectionnée pour formuler la question sont:

```
## Genre:femme, Yeux:bleu, Groupe:C
```