

PrivateTeacher

Cours Privés de Science

Traitement des données avec R Exercice d'entraînement

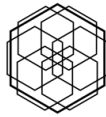
Julien RUPPEN

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 sélectionner des données dans un dataframe.





Enoncé

Des mesures ont été faites 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 sélectionnés parmi cette classe de 300 personnes. L'échantillon ainsi formé est cependant jugé suffisamment grand pour être considéré représentatif de la population.

Observations

Chaque étudiant-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-têtes `Obs_01` et `Obs_02`. Trois autres caractéristiques 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 réalisées sur cet échantillon d'étudiants-es ont été réunies dans un même fichier de donnée. Il s'agit du fichier:

```
## Data-2220-Traitement-Donnees-PrivateTeacher.csv
```

Structure du fichier

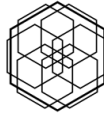
Chaque ligne correspond à un individu et dispose donc de 5 colonnes, une colonne par observation. Le fichier de donnée se présente sous la forme d'une liste d'observation dont la structure est la suivante:

##	Obs_01	Obs_02	Groupe	Genre	Yeux
## 1	4.15	10.62	B	homme	brun
## 2	4.11	18.02	A	homme	bleu
## 3	5.15	17.36	B	femme	bleu
## 4	5.53	29.88	A	femme	bleu
## 5	1.49	6.33	B	homme	brun
## 6	8.08	28.44	B	femme	vert
## 7	6.88	6.49	C	femme	bleu
## 8	12.34	10.56	C	femme	vert
## 9	1.81	8.66	B	homme	brun
## 10	19.33	24.25	C	femme	vert

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'âge ou le nombre de cheveux





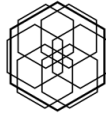
de la personne.. Ce mystère est gardé à dessein, pour te permettre de faire abstraction des détails et de te concentrer sur la démarche.

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

```
## Sélectionner les valeurs appartenant aux femmes  
## dont les yeux sont de couleur vert.
```

Question 02

a)

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

b)

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

c)

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

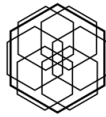
Question 03

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

Question 04

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





Corrigé Détaillé

Question 00

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

```
## data = read.csv("Data-2220-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 caractère de séparation est un espace `sep=" "`.

Question 01

Pour sélectionner les données demandée, on formule tout d’abord une condition sur les colonnes:

```
## condition_genre = data$Genre == "femme"
## condition_yeux = data$Yeux == "vert"
```

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 condition sur les ligne de notre variable “data”

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

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

```
## data_vert = data[condition_yeux,]
```

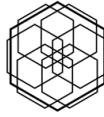
Comme on souhaite sélectionner les deux on écrira donc:

```
## data_femme_vert = data[condition_genre & condition_yeux,]
```

les 5 premières lignes de la selection sont les suivantes:

```
##   Obs_01 Obs_02 Groupe Genre Yeux
## 6    8.08  28.44      B femme vert
## 8   12.34  10.56      C femme vert
## 10  19.33  24.25      C femme vert
## 17   2.65  15.27      B femme vert
## 26   6.41  19.51      B femme vert
```





Question 02

a)

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

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

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

```
## data_femme_vert[, "Obs_01"]
```

Ou plus simplement

```
## data_femme_vert$Obs_01
```

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

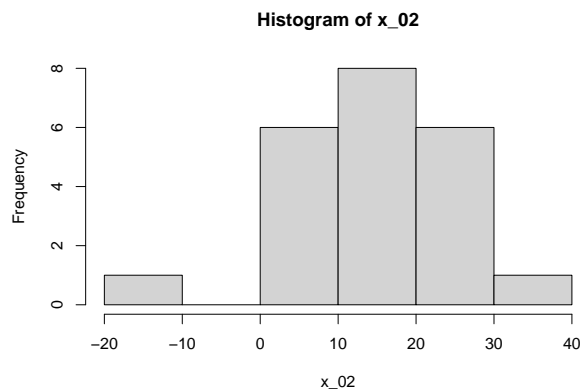
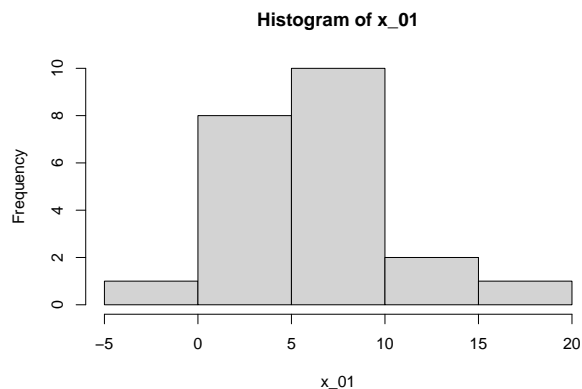
```
## x_01 = data_femme_vert$Obs_01
```

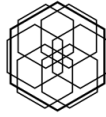
```
## x_02 = data_femme_vert$Obs_02
```

On peut afficher les histogrammes correspondants aux observation 1 et 2 des groupes sélectionné 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 respectivement, 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] 5.852273
## [1] 15.02091

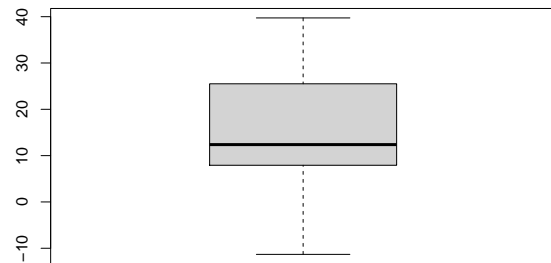
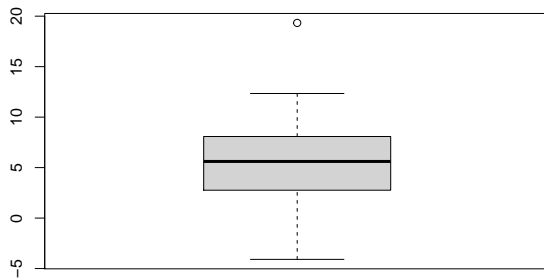
## var(x_01)
## var(x_02)

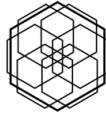
## [1] 22.63467
## [1] 125.2137
```

c)

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

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

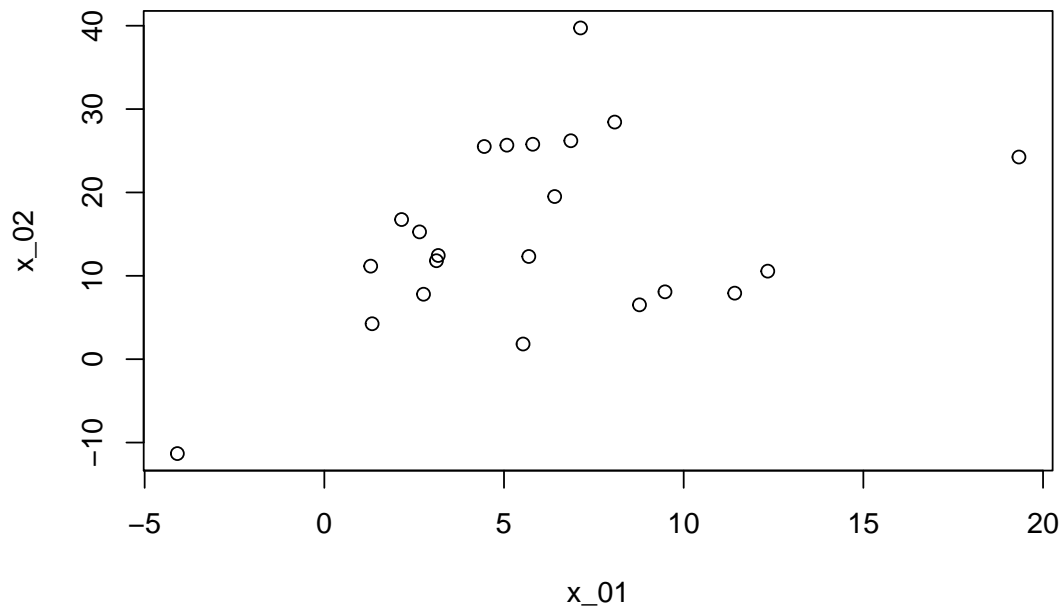




Question 03

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)
```



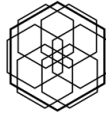
Maintenant que l'on se fait une représentation visuelle de la relation qui existe peut-être entre les deux observations, il est utile de la quantifier en terme de coefficient de corrélation. Nous utiliserons ici les coefficient de corrélation de Pearson:

```
## cor(x_01, x_02, method="pearson")
```

```
## [1] 0.3769562
```

```
## Le coefficient de corrélation est plus petit que 0.6. Il est donc raisonnable
## d'affirmer qu'il n'existe pas de relation de proportionnalité entre les deux
## séries d'observations faites au sein du sous-groupe sélectionné dans cet exercice.
## Le graphique semble confirmer cette conclusion en ne montrant pas d'alignement
## du nuage de point. Une relation existe peut-être mais alors, elle est très faible.
## 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 = 1.8201, df = 20, p-value = 0.08375
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
## -0.05308964  0.68905465
## sample estimates:
##          cor
## 0.3769562
```

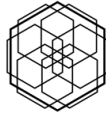
Avec cette commande, nous testons l'hypothèse selon laquelle les variables ne sont pas proportionnelles. (on dit aussi: ne sont pas corrélées linéairement). L'hypothèse de départ H_0 est donc que la valeur du coefficient de corrélation est égale à 0.

On retrouve ici le coefficient de corrélation 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 corrélation n'est pas statistiquement significatif. Il est donc raisonnable d'accepter H_0 . Dans le contexte de notre test, cela signifie donc que la valeur du coefficient de corrélation 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 H_1 . Il est donc raisonnable d'affirmer que la valeur du coefficient de corrélation est significativement différente de 0.





Conclusion

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

Ce tutoriel pour vous montrer qu'une fois que les données sont sélectionnées, toutes les autres commandes sont élémentaires.

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

Ces exercices sont faits pour vous entraîner.

Reprenez ces feuilles pour vous entraîner. Placez-vous en situation d'examen et s'exercez à répondre aux questions.

Tu trouveras ici tous les paramètres utilisés pour générer les données. Tu peux t'en servir pour vérifier les valeurs des moyennes et variances ainsi que les paramètres de la droite pour tous les sous-groupes du data set dont il est question dans cette série.

Paramètres de la simulation

```
##   Groupe N_Obs Mean Variance Slope Intercept
## 1      A    45     5         1         4         4
## 2      B    35     3         5         3         5
## 3      C    41    10         5         1         2

## [1] "Random seed is: 2220"
```

Tableau de toutes les moyennes et variances de tous les sous-groupes

Les valeurs sélectionnées pour formuler la question sont:

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

