BoostR : Travaux pratiques, session 1

Vincent Guillemot

11/09/2014

Manipulation des vecteurs, facteurs et matrices

Exercice 1

Soit k un entier entre 1 et 100. Déterminez les valeurs de k pour lesquelles $\sin(k) > 0$.

Indices : `which`, `:`.

Exercice 2

Définissez un facteur `fac <- factor(c("a","b","b","b","a","b","a","a"))`. Calculez le nombre de `"a"` et de `"b"` dans `fac` utilisant la fonction `length` et des opérateurs binaires.

Que permet de faire la fonction `table` ? Appliquez la à `fac`.

Indices : `length`, `which`, `==`

Exercice 3

1. Créez un vecteur `a` contenant tous les entiers de 1 à 100.
2. Créez un vecteur `b` contenant tous les entiers pairs de 2 à 100.
3. Uniquement en utilisant `a` et `b`, créez un vecteur contenant les entiers **impairs** entre 1 et 100.

Indice: `seq`.

Exercice 4

1. Exécuter la commande `a <- rep(0:1, 50)`. Qu'a-t-on fait ?
2. Utilisez `a` pour construire une matrice `A` à 10 lignes et 10 colonnes.
3. Utilisez la fonction `t` sur cette matrice pour créer une matrice `B`. Que s'est-il passé ?
4. Que se passe-t-il après l'opération `M <- A+B` ?
5. Les commandes `A[1:5, ]` et `B[, 1:5]` permettent de récuperer respectivement les 5 premières lignes de `A` et les 5 dernières colonnes de `B`. Inspirez-vous de ces commandes pour récupérez les lignes de 1 de `A` et les colonnes de 0 de `B`.
6. Extrayez les 2 de la matrice `M`.

Listes et tableaux de données

Exercice 5

1. Créez une liste `x` contenant une variable aléatoire gaussienne de taille 10 appelée `a` et un vecteur contenant uniquement des 1 de taille 10 également. On peut accéder aux deux éléments de cette liste avec les commandes `x[[i]]` ou `x$nom_de_la_variable`. *Indice : `rnorm`.*

2. Créez un objet `y` qui est la transformation de cette liste en `data.frame`. On peut maintenant parcourir les éléments de chaque objet comme pour une matrice avec la commande `y[i,j]` !
3. Créez deux objets `z1` et `z2` contenant respectivement les 3 premières et les 3 dernières lignes de `y`. Quelle est la classe de ces deux objets ?
4. Rajoutez à la liste `x` un vecteur `alphabet` contenant les lettres de l'alphabet.
5. Essayez de transformer de nouveau `x` en `data.frame`. Que se passe-t-il ?

Fonctions

Exercice 6 : création de fonction

1. Exécutez les commandes `data(iris)` puis `print(iris)`. Nous venons de charger en mémoire l'un des nombreux jeux de données distribués avec R ! Profitez de l'aide sur ce jeu de données pour en apprendre un peu plus sur les fleurs (`?iris`) !
2. Afin de représenter l'histogramme de la première colonne de `iris`, exécutez la commande `hist(iris[,1])`. Trouvez l'argument à changer de sorte que la couleur des "barres" de l'histogramme soit égale à `"steelblue"` et la couleur des bords égale à `"white"`.
3. Nous voudrions faire les histogrammes des 4 première colonnes de `iris`. Pour cela, créez une fonction `f` en suivant la syntaxe suivante

```
f <- function(i) hist(iris[,i])
```

4. Utilisez cette fonction 4 fois pour observer les histogrammes des 4 variables numériques du jeu de données `iris`.

Remarque : pour exécuter plusieurs commandes au sein d'une même fonction, il faut utiliser des accolades `{...}`. Par exemple

```
f <- function(i) {
  letitre <- paste("Histogramme de la variable", i)
  hist(iris[,i], main=letitre, xlab="Valeur")
}
```

Exercice 7 : la fonction `ifelse`

La fonction `ifelse` permet de transformer des vecteurs booléens (c'est à dire des vecteurs contenant uniquement des valeurs `TRUE` ou `FALSE`) en tout autre valeur : une valeur sera attribuée à `TRUE` et une autre valeur à `FALSE`.

Par exemple, `ifelse(c(TRUE, FALSE, TRUE, TRUE, FALSE), "C'est vrai !", "C'est faux !")`.

A l'aide la fonction `ifelse`, générez un facteur qui contient 100 valeurs réparties en deux "niveaux". Les deux niveaux sont `"positif"` et `"négatif"` : `"positif"` quand $sin[k] > 0$ et `"négatif"` quand $sin[k] < 0$ ($k = 1, ..., 100$).

Exercice 8 : les boucles `for`

1. Lisez l'aide sur la procédure permettant de réaliser des boucles indicées `for` (`help("for")`). *Remarque :* demander de l'aide sur cette procédure avec la syntaxe `?for` ne fonctionnera pas ! Pourquoi ?
2. Pour exemple, exécutez la boucle suivante `for (i in 1:10) print(i)`.

3. A l'aide d'une boucle, calculez la somme des entiers pairs compris entre 1 et 100.

Indices : `for`, `ifelse`, `:`, `%%`

Estimation

Exercice 9 : papillons de lumièRe

Les longueurs d'ailes de 24 papillons ont été mesurées (en cm) [Zar]:

```
butterflies <- c( 3.3, 3.5, 3.6, 3.6, 3.7, 3.8,
                  3.8, 3.8, 3.9, 3.9, 3.9, 4.0,
                  4.0, 4.0, 4.0, 4.1, 4.1, 4.1,
                  4.2, 4.2, 4.3, 4.3, 4.4, 4.5 )
```

1. Calculez des statistiques simples sur cet échantillons et l'intervalle de confiance à 95 % sur la moyenne. Quelle(s) hypothèse(s) doit-on faire pour calculer ce dernier ? Représentez un histogramme.
2. *Question subsidiaire :* calculez l'intervalle de confiance sur la variance de cet échantillon.

$$\left[\frac{(n-1)s^2}{\chi^2_{\frac{1-\alpha}{2},n-1}} ; \frac{(n-1)s^2}{\chi^2_{\frac{\alpha}{2},n-1}} \right]$$

   Indice : `qchisq`.

Exercice 10 : manipulation de données aléatoires

1. Créez un vecteur `x` de taille 100 qui est un échantillon $\mathcal{N}(0,1)$ (de loi Gaussienne, moyenne nulle et écart-type unitaire). *Indice : `rnorm`.*
2. Trouvez toutes les valeurs de `x` qui sont supérieures à son 3ème quartile (quantile à 75 %), noté `q75`. *Indice : `?quantile`.*
3. Comptez-les en suivant les étapes suivantes :
   - calculez `sum(c(TRUE, FALSE))` ;
   - que remarquez-vous ?
   - utilisez `x > q75` et la fonction `sum`.
4. Répétez les étapes 1 à 3 pour un échantillon de loi uniforme de bornes -30 et 30.

Exercice 11 : génération de données aléatoires et représentation en histogramme

Pour chaque type de variable aléatoire suivante, tracer un histogramme d'un échantillon aléatoire de taille n.

- variable uniforme, `runif`,
- variable du Khi-deux, `rchisq`,
- variable de Fisher, `rf`,
- variable de Student, `rt`.

Un exemple pour une variable aléatoire normale vous est proposé : `hist(rnorm(100))`.

Exercice 12 : Etude de la moyenne empirique sur des données simulées [Husson & Pagès]

Générez, avec le logiciel R, 1000 échantillons de taille 47 et stockez les dans une matrice (matrix) à 1000 lignes et 47 colonnes, échantillons qui seront gaussiens de moyenne nulle et de variance 1.

1. On veut calculer la moyenne empirique d'échantillons de taille 2, 5, 10 et 30. Utilisez les 2 premières colonnes pour calculer 1000 moyennes sur 2 individus, les colonnes 3 à 7 pour 5 individus, les colonnes 8 à 17 pour 10 individus et 18 à 47 pour 30 individus.
2. Construisez un histogramme de ces 1000 moyennes pour chaque taille d'échantillon.
3. Reprendre les mêmes questions pour une loi uniforme sur $[-1, 1]$.

Exercice 13 : Calcul de la taille d'un échantillon pour une précision donnée [Husson & Pagès]

On a pesé 15 poulpes mâles adultes pêchés au large des côtes Mauritaniennes. On suppose que, pour cette espèce de poulpes, les poids sont répartis selon une loi normale d'espérance μ et de variance σ^2. Le tableau ci-dessous donne l'échantillon des 15 valeurs obtenues :

```
x <- c( 1150, 1500, 1700, 1800, 1800, 1850, 2200, 2700,
        2900, 3000, 3100, 3500, 3900, 4000, 5400 )
```

1. Donnez une estimation de μ et σ^2 à partir des données.
2. Construisez un intervalle de confiance pour μ au niveau $\alpha = 5\%$. Donnez l'amplitude de cet intervalle.
3. Si n désigne la taille de l'échantillon, construisez une fonction permettant de donner l'amplitude de l'intervalle de confiance pour μ au niveau $\alpha = 5\%$ en fonction de n. Les valeurs de moyenne et de variance sont supposées rester les mêmes.

Exercice 14

19 malades atteints d'un cancer du poumon ont été traités chirurgicalement (ablation du poumon atteint) dans un même service de chirurgie et suivis jusqu'a leur décès. La série des durées de survie, mesurées en semaines a partir de la date de l'intervention chirurgicale jusqu'a celle du décès, est la suivante :

```
x <- c(25,  45, 238, 194, 16, 23, 30, 16, 22, 123, 51, 412, 162,  14, 72, 35, 30, 91, 45)
```

Dans l'étude de durée de survie, on définit la fonction de survie empirique S par $S(x) = 1 - F(x)$, où F est la fonction de répartition empirique.

1. Déterminez la médiane et les quartiles de la série statistique. Calculez l'écart interquartile.
2. Calculez la moyenne $\overline{x}$ et l'écart type $\widehat{\sigma}$ de la série.
3. Construisez le diagramme de Tukey (boite a moustaches).
4. Représentez graphiquement $F(x)$ et *attention, question difficile* $S(x)$.

Exercice 15

On considère une série statistique de 60 taux d'hémoglobine dans le sang (gr/l) mesures sur des adultes presumes en bonne santé.

Valeurs mesurées chez les hommes :

```
hom <- c(141 , 144 , 146 , 148 , 149 , 150 , 150 , 151 , 153 , 153 , 153 ,
         154 , 155 , 156 , 156 , 160 , 160 , 160 , 163 , 164 , 164 , 165 ,
         166 , 168 , 168 , 170 , 172 , 172 , 176 , 179.1)
```

Valeurs mesurées chez les femmes :

```
fem <- c(105 , 110 , 112 , 112 , 118 , 119 , 120 , 120 , 125 , 126 , 127 ,
         128 , 130 , 132 , 133 , 134 , 135 , 138 , 138 , 138 , 138 , 142 ,
         145 , 148 , 148 , 150 , 151 , 154 , 154 , 158)
```

1. Calculez les moyennes et quartiles de ces échantillons.
2. Déterminez les étendues de ces distributions.
3. Calculez l'écart interquartile pour chacune des trois distributions.
4. Tracez les diagrammes de Tukey (boites a moustache) des trois distributions.
5. Calculez les variances et les écarts-types de ces distributions.
6. Utilisez la fonction `boxplot` avec l'option `notch = TRUE` : que dit l'aide de la fonction et que pouvez vous conclure ?

Exercices supplémentaires de niveau avancé

Exercice 16 : Propriétés des estimateurs de la moyenne et de la variance [Husson & Pagès]

Dans une population de taille $N = 5$, une variable X peut prendre uniquement les valeurs suivantes de façon équiprobable

$$-6, -3, 0, 3, 6$$

De façon générale, pour une variable aléatoire X pouvant prendre des valeurs X_i avec une probabilité p_i, les moyenne et variance théoriques se calculent de la façon suivante

$$\mu = \sum_i p_i X_i \text{ et } \sigma^2 = \sum_i p_i (X_i - \mu)^2.$$

1. Calculez la moyenne μ et la variance σ^2 (théoriques) de X dans cette population.
2. On effectue des prélèvements d'échantillons de taille $n = 2$ sans remise dans cette population. Enumérez tous les échantillons possibles. Pour chacun d'entre eux, calculez leur moyenne empirique et leur variance empirique. *Indices : utilisez la fonction* `expand.grid` *ainsi qu'une boucle* `for`*. Attention, vous aurez très probablement besoin de convertir le résultat de* `expand.grid` *en* `matrix` *!*
3. Vérifiez que la moyenne empirique est un estimateur sans biais de la moyenne μ. (Même chose pour la variance)

Exercice 17

Il y a une grande colline entre les deux villages A et B. Le train prends 2h pour aller de A à B, et 2h30 min de B à A. La vitesse du train est de 30 km/h quand il monte et de 40 km/h quand il descend. Calculez la distance entre A et le sommet de la colline S ainsi que la distance entre B et S.

Indice : On peut résoudre le système à deux équations et deux inconnues suivant

$$\begin{cases} x + y = 1 \\ x + 2y = 2 \end{cases}$$

à l'aide de la commande `solve(matrix(c(1,1,1,2),2,2), c(1, 2))`.

Exercice 18

Créez une fonction qui accepte trois arguments numériques `a`, `b` et `c` et qui permet de calculer les solutions de l'équation quadratique

$$ax^2 + bx + c = 0.$$

Utilisez-la pour résoudre les trois équations suivantes :

$$x^2 + 2x + 1 = 0$$

$$x^2 + 3x + 2 = 0$$

$$-x^2 + x - 2 = 0$$

Exercice 19 (difficile)

La fonction exponentielle peut s'approximer à l'aide de séries entières. Mathématiquement :

$$\lim_{n \to +\infty} \sum_{k=0}^{n} \frac{x^k}{k!} = e^x,$$

où $k! = 1 \times 2 \times ... \times k$ est la fonction *factorielle*.

Nous allons montrer graphiquement la qualité de cette approximation avec les étaês suivantes.

1. Initialisez deux vecteurs `a` et `b` à la valeur 1.
2. Initialisez un réel `x` à la valeur 0.2.
3. Initialisez un entier `n` à la valeur 10.
4. A l'aide d'une boucle `for`, faites en sorte que `a` et `b` soient tels que

   $$a[1] = b[1] = 1 \text{ puis } a[k] = x^{k-1} \text{ et } b[k] = (k-1)! \text{ pour } k = 2, ..., n.$$

   (Indice : utilisez la fonction `factorial`)
5. Divisez `a` par `b` et faites en la somme ! Vous obtenez ainsi une approximation à 10 termes de $e^{0.2}$.
6. Répétez l'opération pour obtenir toutes les approximations de $e^{0.2}$ à 2, 3, ..., 8 et 9 termes.
7. Faites un graphe contenant ces valeurs et observez la vitesse de convergence !

Références

- [Zar] : Biostatistical analysis de Jerrold H. Zar.
- [Husson & Pagès] : Statistiques générales pour utilisateurs de François Husson et Jérôme Pagès.