TP Introduction à R - Bases du langage
CED - MaiMoSiNE

Février 2017

1 Manipulation de données

1.1 Utilisation de la fonction apply

1. Calculer les statistiques de base (moyenne,min,max,...) des trois variables du jeu de données **ethanol** disponible dans le package **lattice**.

2. Calculer les **quantiles** de chacune des trois variables. Pour cela vous pourrez utiliser la fonction **apply** avec la fonction **quantile**.

3. Toujours avec la fonction **apply**, calculer toutes les déciles de chacune des trois variables en utilisant l'argument **probs** de la fonction **quantile**.

1.2 Fichiers de données (data.frame)

Lire le fichier texte "poidsTaille.txt" stocké dans le répertoire TP, ce fichier est un fichier texte contenant les données d'un étude du CHU de Toulouse, concernant le poids, la taille, le sexe d'enfants de 4 à 7ans. Ce fichier contient sur sa première ligne le nom des variables observées.

- Quel est la classe de la variable ainsi crée? Quelle est la taille de l'échantillon? Quel le mode et la classe des variables contenu dans cet échantillon.

- Calculer les statistiques de base à l'aide de la fonction *summary*

- Calculer séparément la moyenne, la médiane, l'étendue et les quantiles pour les variables *Poids* et *Taille* à l'aide de commandes appropriées.

- Calculer la variance (fonction *var*) des variables *Poids* et *Taille*, la variance calculée et la variance avec biais. Écrire une fonction qui calcule la variance sans biais et l'écart-type.

- Tracer un histogramme (fonction hist)en n classes de ces deux variables, on fera varier n.

- Extraire :

  - la deuxième ligne
  - la troisième colonne
  - les lignes 1, 2 et 4 avec une seule commande c()
  - les lignes 3 à 6 avec la commande :
  - tout sauf les colonnes 1 et 2.
  - toutes les lignes ayant une AGE supérieure à 70mois

1.3 Sélection et tri dans un data-frame

1. A partir du jeu de données **iris** disponible sous R, visualiser les 5 premières lignes, créer un sous jeu de données comportant uniquement les données de la modalité **versicolor** de la variable **species** , appeler ce nouveau jeu de données **iris2**.

2. Trier par ordre décroissant les données de **iris2** en fonction de la variable **Sepal.length** (vous pourrez utiliser la fonction **order**).

1.4 Tableau croisé - tableau de données

Soient les deux variables qualitatives **laine** et **tension** mesurées sur 10 individus.La variable **laine** correspond à trois types de laine: Angora, Merinos,Texel . La variable **tension** indique indique les valeurs **Faible** et **Forte à la résistance en traction**.

```
tension <- factor(c(rep("Faible",5),rep("Forte",5)))
laine <- factor(c(rep("Mer",3),rep("Ang",3),rep("Tex",4)))
```

1. Créer le tableau de contingence croisant les variables **laine** et **tension**. Observer la classe et les attributs de cet objet "tableau de contingence".

2. A partir de ce tableau, créer une matrice de caractères **tabmat** qui contient trois colonnes et autant de lignes que de croisement de modalités. Cette matrice sera remplie à chaque ligne par la tension (ligne du tableau précédent), le type de laine (colonne du tableau précédent) et l'effectif pour le croisement des modalités. Pour cela, on utilisera les fonctions **matrix** et **rep**.

3. Transformer la matrice de caractères de la question précédente en **data-frame** et contrôler le type des variables. Affecter à **n** le nombre total d'individus et à **nbefac**, le nombre de variables qualitatives.

4. Créer un compteur **iter** et une matrice **tabcomplet** de caractères (”” par exemple), de la taille du jeu de données final.

5. Faire une boucle sur le nombre de lignes de **tabmat**. Chaque ligne **i** correspond à un croisement de modalités. Si le nombre d'individus qui prennent ce croisement de modalités n'est pas nul, répéter le, sur autant de lignes de **tabcomplet** qu'il faut. Le résultat **tabcomplet** sera identique au tableau de données initial.

1.5 Ventilation

On considère la variable qualitative **Xqual** :

```
Xqual <- factor(c(rep("A",60),rep("B",20),rep("C",17),rep("D",3)))
```

1. Calculer la fréquence de chaque modalité.

2. Afficher à l'écran l'intitulé des modalités dont l'effectif est inférieur à 5% de l'effectif total.

3. Calculer les fréquences de chaque modalité sans la(les) modalité(s) de la question précédente. Le résultat sera mis dans un vecteur proba.

4. Sélectionner les individus prenant la(les) modalité(s) de la question 2.Leur donner une valeur parmi, les modalités restantes, selon un tirage dont les probabilités sont calculées en question 3(utiliser la fonction sample). Ce procédé est appelé ventilation.

2 Programmer avec R

2.1 Les listes et les fonctions

Écrire une fonction qui prend deux vecteurs de même taille en entrées et fournit la somme et le produit terme à terme de ces deux vecteurs. On utilisera un objet de type *list* pour les sorties.

Testez sur deux vecteurs et observez les résultats.

2.2 Reperer les individus manquants

Les données manquantes sont représentées sous R par NA (Not Available). Pour les retrouver, on utilise la fondtion **is.na** qui renvoie **TRUE** si la valeur vaut NA et **False** sinon.

Construire une variable mesurée sur 15 individus en effectuant un tirage selon une loi normale $\mathcal{N}(0,1)$, on affecte des données manquantes pour les individus 7,8,14, il s'agit de:

- repérer les individus qui ont des données manquantes,
- éliminer les individus qui ont des données manquantes.

2.3 Fonction de ventilation

Reprendre l'exercice de ventilation.

1. Programmer une fonction de ventilation. Cette fonction aura pour arguments une variable et le seuil à partir duquel une modalité doit être ventilée. Mettre 5% comme seuil par défaut.
2. Ecrire une fonction qui permet de ventiler toutes les variables qualitatives d'un tableau.