Statistiques descriptives avec R

1 Statistiques descriptives : représentation de variables

1.1 Représenter une variable qualitative

Nous allons rentrer des données « à la main » pour une variable qualitative. Cette variable représente l'appartenance à un groupe (parmi 3 groupes) et prend 3 modalités `g1`, `g2` et `g3`. Les 2 premiers individus sont dans le groupe 1, les 3 suivants dans le groupe 2 et le dernier dans le groupe 3 :

```
ybrut <- c("g1","g1","g2","g2","g2","g3")
print(ybrut)
summary(ybrut)
```

Que fait le dernier ordre ci-dessus ? Nous devons transformer ce vecteur (de caractères) en variable qualitative (nommée `factor` sous R) :

```
y <- factor(ybrut)
```

Que font les ordres suivants ?

```
levels(y)
nlevels(y)
table(y)
sum(table(y))
table(y)/sum(table(y))*100
```

Tracer les effectifs de chaque modalité dans un diagramme en barre :

```
barplot(table(y))
```

Tracer les pourcentages de chaque modalité dans un diagramme en barre :

```
barplot(table(y)/sum(table(y))*100,ylab="pourcentages",xlab="groupes")
```

Que font les options `xlab` et `ylab` ?
Copier le dernier graphique dans un document word ou openoffice.
Que fait le résumé numérique d'une variable qualitative ?

```
summary(y)
```

1.2 Représenter une variable quantitative continue

Représentons une variable quantitative continue. Ouvrir le fichier `varquant.r` et exécuter son contenu (couper coller son contenu ou utiliser l'icône de tinn-R).
Que fait le résumé numérique d'une variable quantitative (continue) ?

```
summary(y)
```

Trouver sur les deux graphiques ci-dessous la différence et expliquez la.

```
hist(y,freq=TRUE)
hist(y,freq=FALSE)
```

Que font toutes les options pour ce graphique

```
hist(y,freq=FALSE,breaks=10,xlab="huile",main="Histogramme")
```

Que fait cette option

```
hist(y,freq=FALSE,breaks=c(15,18,25,30,36))
```

Expliquer tous les ordres ci-dessous

```
boxplot(y,xlab="",ylab="teneur en huile")
mean(y)
abline(h=mean(y))
quantile(y)
median(y)
abline(h=median(y),col=2)
```

Conclusion : l'histogramme est tracé grae à `hist` avec l'option `freq=FALSE`.
Un autre estimateur de la densité (estimateur à noyau) est disponible afin d'estimer la densité par une fonction continue

```
density(x, ...)
```

Retourne les coordonnées x et y d'un *estimateur* de la densité du vecteur de données `x`. L'argument `bw` indique la largeur de fenetre (plus elle est grande plus la courbe est lisse)

```
> normal=rnorm(100)
> ndens=density(normal, width=1.2)
> hist(normal, probability=T)
> lines(ndens)
```

1.3 Représenter une variable quantitative discrète

Représentons une variable quantitative discrète : le nombre d'enfant par famille. Nous allons rentrer des données « à la main » les valeurs de cette variable. La première famille possède 5 enfants, la second n'en a pas, la troisième en possède 2 enfants, la quatrième 2 et la cinquième n'en a pas.

```
y <- c(5,0,2,2,0)
```

Que font les commandes suivantes

```
unique(y)
sort(unique(y))
table(y)
```

Le diagramme en barre des effectifs est le diagramme suivant

```
plot(sort(unique(y)),table(y),type="h",ylim=c(0,max(table(y))))
```

En général, dès que les valeurs possibles sont assez nombreuses (par exemple 7 ou 10 ou plus) la variable quantitative discrète est assimilée à une variable quantitative continue. La distinction quantitatif discret ou continue n'existe pas sous R, les deux sont des variables numériques (`numeric`).

1.4 Données des tournesols

1. Importer le tableau `tournesol.csv` qui contient les variables décrites dans le tableau 1. Ce tableau sera affecté dans un objet appelé `tpropre`.

Code variable	Descriptif variable
ecotype	code plante
plt	numéro du plant d'un écotype donné
etat	état d'origine de la plante (aux USA)
longitude	longitude du lieu de collecte (aux USA)
latitude	latitude du lieu de collecte (aux USA)
haut	hauteur des plants
semflo	jour de floraison (écart en jour par rapport au premier mai)
rambas	note de ramification basale (entre 0 aucune et 4 maximum)
longfeu	longueur du cumulée du limbe et du pétiole (cm ?)
grlon	longueur maxi de la graine (mm, moyenne sur 15 graines minimum)
huile	pourcentage d'huile

TABLE 1 – Variables mesurées sur les tournesols (dans la station d'essai aux environs de Montpellier).

2. Donner pour chaque variable son type (variable qualitative, quantitative discrète, quantitative continue).
3. Effectuer un résumé numérique du tableau de données `tpropre` :

   ```
   summary(tpropre)
   ```

4. Quelles sont les variables qui sont reconnues comme variables quantitatives et comme variables qualitatives ?
5. Donner à chaque variable le type voulu grâce à `factor` ou `as.numeric`

1.5 Deux variables quantitatives continues

Par defaut R trace des points (`type="p"`) aux coordonnées fournies (ci-dessous l'ordonnée est la variable `huile` et l'abscisse la variable `grlon`). Détailler le rôle des options

```
plot(huile~grlon,data=tpropre)
plot(huile~grlon,data=tpropre,pch="+")
plot(huile~grlon,data=tpropre,col=2,pch="+")
```

Traçons des lignes

```
plot(huile~grlon,data=tpropre,type="l")
```

Qu'a t-on fait ?

1.6 Deux variables qualitatives : tableau de contingence

Utilisez l'ordre suivant

```
table(tpropre[,"ecotype"],tpropre[,"etat"])
```

Que renvoit il ?

1.7 Données des tournesols (suite)

1. Calculer la moyenne empirique des variables `huile, grlon` et `longfeu`.
2. Pour ces mêmes variables donner leurs quartiles empiriques.

3. Pour ces mêmes variables les représenter par un boxplot.
4. Pour ces mêmes variables calculer leur variance empirique.
5. Représenter graphiquement chacune des variables et exporter ces représentations graphiques dans un document word ou openoffice.

2 Manipuler des données

2.1 Importation (exercice)

Importer les tableaux `test1.csv`, `test2.csv` et `test3.csv` dans les variables `don1`, `don2`, `don3`.

2.2 Fusionner des tableaux

Exécuter et commenter :

```
> toto.1 <- cbind(don1,don3)
> toto.1
> montab = rbind(don1,don2)
> montab
> rbind(don1,don3)
> objects()
> rm(toto.1,montab)
> objects()
```

2.3 Fusionner des tableaux

Il est possible de fusionner deux tableaux selon une clef (cf. fusion de 2 tables dans les bases de données), grâce à l'ordre classique `merge`

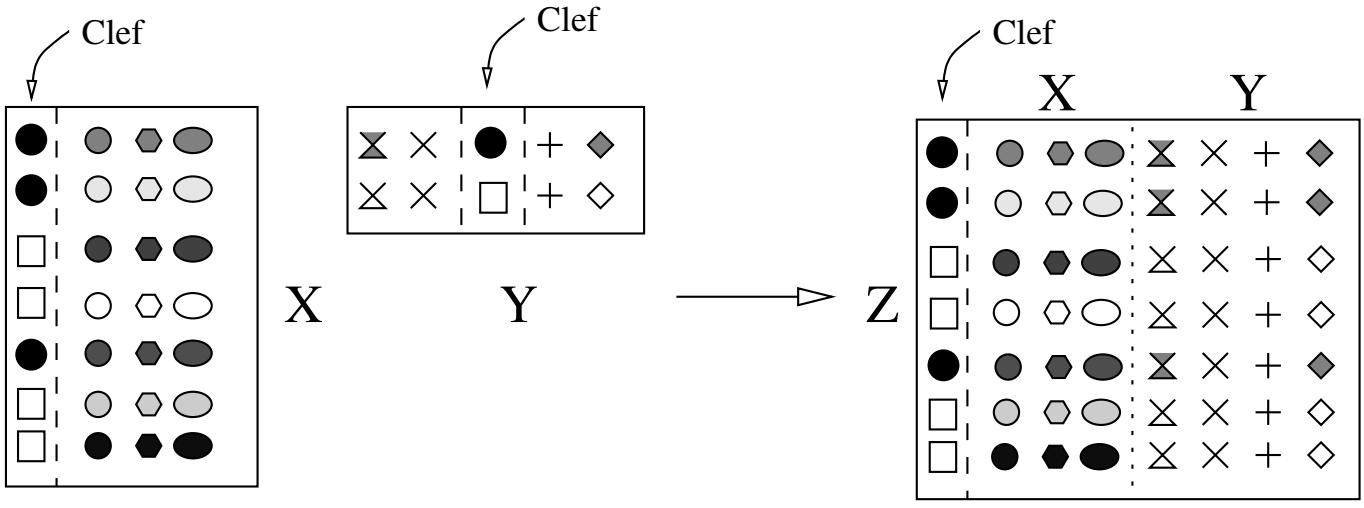

FIGURE 1 – Fusion par clef : `merge(X,Y,by="clef")` où `clef` est le nom d'une variable commune à `X` et `Y`.

Importer les données du fichier `tournesol_propre.csv` (dans que l'on nommera `tpropre`) ainsi que celle du fichier `meteo_tournesol.csv` (dans un tableau nommé `meteo`). Faire fusionner les tableaux `tpropre` et `meteo` grâce à la clef `ecotype`.

3 Facteurs et fonctions

1. Réimporter éventuellement l'objet `mtcars` grâce à `data(mtcars)`.
2. Donner un résumé numérique de chaque variable de la matrice `mtcars` (voir `summary`).
3. Créer un facteur (de nom `conso`) en découpant en classe la variable `mpg` (de la matrice `mtcars`). Le découpage sera fait de 5 en 5 en partant de 10 (voir `cut, seq`).
4. Afficher les niveaux (ou modalités) de `conso` (`levels`).
5. Créer `conso2` égal à `conso`.
6. Fusionner la premiere et la seconde modalité de `conso2` en une seule de nom `"fuse"` (`levels`).
7. Transformer `conso` en facteur ordonné (`ordered`).
8. Voir les effectif de chaque modalité (`table`)
9. Voir les pourcentages de chaque modalité (`table, length`)
10. Faire un résumé numérique de `conso` et voir la différence avec une variable numérique (question 2).
11. Faire une moyenne de chaque variable de la matrice `mtcars` (`apply` ET `colMeans`).
12. Faire une somme de chaque variable de la matrice `mtcars` (`apply` et ET `colSums` ET par une boucle)).
13. Calculer la médiane somme de chaque variable de la matrice `mtcars` (`apply`)).
14. Créer une fonction pour calculer $n! = 1 \times 2 \times \ldots (n-1) \times n$
    - par une boucle `for`,
    - par une boucle `while`,
    - sans boucle (`prod`),
    - par une fonction mathématique intégrée.
15. Créer une fonction pour les écarts absolus à la moyenne d'un vecteur `x` ($MAD = \frac{1}{n}\sum_{i=1}^{n} |x_i - \bar{x}|$ où $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$.
16. utiliser cette fonction pour calculer l'écart à la moyenne (MAD) pour chaque colonne de la matrice `mtcars` (`apply`).
17. Calculer la moyenne, pour chaque niveau de `conso`, de toutes les colonnes de la matrice `mtcars` (`apply`)) sauf la première ie.
18. Calculer la moyenne par niveau de `conso` de toutes les colonnes de la matrice `mtcars` sauf la première (ipour les 10 colonnes de mtcars il faut 5 moyennes, une par niveau de `conso`) ; voir `aggregate`.