

# TD de statistique : graphiques avec R

Jean-Baptiste Lamy

11 octobre 2007

## 1 Rappels de vocabulaire

**Données univariées** lorsqu'il n'y a qu'une seule variable

**Données bivariées** lorsqu'il y a deux variables

**Données multivariées** lorsqu'il y a plusieurs variables

**Données quantitatives, données numériques** lorsque ces variables sont des nombres sur lesquels les opérations arithmétiques (addition, soustraction, moyenne,...) ont un sens : par exemple, des températures, des concentrations, des poids, des volumes,... mais pas des numéros de département.

**Données qualitatives** dans le cas contraire : valeurs booléennes (vrai ou faux, oui ou non), ou définies par un ensemble de valeurs possible (par exemple une variable "fréquence" pouvant prendre les valeurs "rare", "fréquent", "très fréquent"; ou bien une variable "souche d'Aspergillus" pouvant prendre les valeurs "Fumigatus", "Niger", "Lentullus"; ou des numéros de département).

**Données ordonnées** lorsqu'il est possible de classer les valeurs de la variable selon un ordre. Les données quantitatives sont toujours ordonnées, et certaines données qualitatives le sont aussi (par exemple il est possible d'ordonner des fréquences mais pas des souches d'Aspergillus).

## 2 Introduction aux graphiques avec R

Faire un graphique pour représenter ses données permet de compléter l'information de la statistique descriptive. La visualisation des données est une information en soi et ouvre des pistes de travail dans l'analyse de ses données. Il existe des représentations graphiques de données très nombreuses et R offre une variété de graphiques remarquables. Pour avoir une petite idée des possibilités offertes, il est possible de taper la commande :

```
> demo(graphics)
```

Il n'est pas possible ici de détailler tous ces graphiques et les options dans les fonctions qui les utilisent. Nous nous limiterons aux principales fonctions et aux plus intéressantes.

## 3 Graphique à 1 variable

### 3.1 Préparation

Avant de tracer des graphiques, il faut d'abord charger dans R les données que l'on veut représenter, et les "attacher" (voir TP n°1), par exemple de la manière suivante :

```
> tableau = read.table("mon_fichier.csv", sep = ",", header = TRUE)
> attach(tableau)
```

### 3.2 Diagramme de dispersion (stripchart) : 1 variable numérique

Un premier graphique consiste tout simplement à mettre sur un axe des données quantitatives ordonnées : c'est le diagramme de dispersion (*stripchart* en anglais) :

```
> stripchart(variable_numérique)
```

### 3.3 Représentation point par point (dotchart) : 1 variable numérique

Le *dotchart* est similaire au diagramme de dispersion, mais chaque donnée est représentée sur une ligne différente.

```
> dotchart(variable_numérique)
```

Il peut être utile de trier les données pour faciliter la lecture du graphique :

```
> dotchart(sort(variable_numérique))
```

### 3.4 Boîtes à moustaches simple : 1 variable numérique

Pour afficher une boîte à moustache simple :

```
> boxplot(variable_numérique)
```

Lorsque l'on a plusieurs variables INDÉPENDANTES, il est possible de placer plusieurs boîtes à moustaches les unes à côté des autres, en séparant les variables par des virgules :

```
> boxplot(variable1_numérique, variable2_numérique, variable3_numérique,...)
```

### 3.5 Histogramme : 1 variable numérique

C'est une représentation plus classique et très utile ; on l'obtient avec :

```
> hist(variable_numérique)
```

Par défaut R applique la loi de Sturges (cf cours) pour déterminer le nombre de barres ; il est possible de préciser le nombre de barres manuellement (ici, 3) :

```
> hist(variable_numérique, breaks = 3)
```

ou bien les valeurs auxquelles auront lieu les coupures :

```
> hist(variable_numérique, breaks = c(0.0, 0.2, 0.5, 0.6, 1.0))
```

### 3.6 Camembert : 1 variable qualitative

Un camembert se fait avec la commande pie :

```
> pie(variable_numérique)
```

Pour faire un camembert à partir d'une donnée qualitative, il faut d'abord stratifier celle-ci avec la commande summary :

```
> pie(summary(variable_qualitative))
```

## 4 Graphique à 2 variables

En biologie, il est très fréquent d'étudier une variable en fonction d'une (ou plusieurs) autres variables : par exemple la réponse biologique d'un organisme à une substance en fonction de la dose de substance,... Nous allons maintenant voir comment étudier 2 variables simultanément, puis 3 dans la section suivante.

### 4.1 Boîtes à moustaches : 1 variable numérique + 1 variables qualitatives

Pour afficher une boîte à moustache simple :

```
> boxplot(variable_numérique)
```

Pour afficher une boîte à moustache de la variable 1 pour chaque valeur possible de la variable 2 (la variable 1 est numérique et en général la variable 2 est qualitative) :

```
> boxplot(variable1_numérique ~ variable2_qualitative)
```

NB : On comprend mieux cette appellation de boîte à moustache si on la représente horizontalement :

```
> boxplot(len, range = 0, horizontal = TRUE)
```

### 4.2 Graphique à deux dimensions (X, Y) : 2 variables numériques

La première variable est placée en X, la seconde en Y. Il est possible d'utiliser des variables numériques ou qualitatives, cependant ce type de graphique est surtout utile avec des variables qualitatives.

```
> plot(variable_numerique1, variable_numerique2)
```

### 4.3 Barre cumulée : 2 variables qualitatives

Un diagramme en "barre cumulée" permet d'étudier 2 variables qualitatives :

```
> plot(variable_qualitative1, variable_qualitative2)
```

## 5 Graphique à 3 variables et plus

Réponse :

## 5.1 Boîtes à moustaches : 1 variable numérique + 2 variables qualitatives

Pour afficher une boîte à moustache sur trois variables (ou plus) : ici on affiche une boîte à moustache pour la variable 1 pour chaque combinaison des variables 2 et 3 :

```
> boxplot(variable1_numerique ~ (variable2_qualitative + variable3_qualitative))
```

**Astuce** Le nombre de boîtes peut vite devenir important ; l'ensemble est souvent plus lisible à l'horizontal, et en mettant les étiquettes des axes à l'horizontal (`las = 2`) :

```
> boxplot(variable1_numerique ~ (variable2_qualitative + variable3_qualitative), horizon-  
tal = TRUE, las = 2)
```

## 5.2 Graphique de niveau : 1 variable numérique + 2 variables qualitatives

Avant d'utiliser ce type de graphique, il faut importer le module d'extention "lattice" (il suffit de le faire une seule fois) :

```
> require(lattice)
```

Les graphiques de niveau permettent d'étudier une variable numérique en fonction de deux variables qualitatives :

```
> levelplot(variable_numerique1 ~ variable_qualitative2 * variable_qualitative3)
```

Chaque "case" du graphique obtenue correspond à une valeur de la variable 2 et une valeur de la variable 3 ; la couleur de la case indique la valeur moyenne de la variable 1.

## 5.3 Nuage en 3D : 3 variables numériques

Avant d'utiliser ce type de graphique, il faut importer le module d'extention "lattice" (il suffit de le faire une seule fois) :

```
> require(lattice)
```

Les nuages de point en 3D permettent d'étudier le comportement de 3 variables numériques simultanément :

```
> cloud(variable_numerique1 ~ variable_numerique2 * variable_numerique3)
```

## 6 Mise en forme des graphiques

Il existe de nombreuses options pour mettre en forme les graphiques. En voici quelques-unes :

**main** titre du graphique

**sub** sous-titre du graphique

**xlab** titre de l'axe X

**ylab** titre de l'axe Y

D'autres options sont disponibles ; pour les obtenir, consultez l'aide, par exemple pour les histogrammes :

```
> help(hist)
```

## 7 Exercice 1

Nous allons reprendre l'étude du TP 1. Pour rappel, cette étude portait sur l'efficacité de trois herbicides sur trois plantes : blé, chiendent et liseron. Pour cela, des cultures de ces plantes ont été mises en présence de l'un des trois herbicides, ou d'aucun d'entre eux. Le nombre de plants vivants dans la culture a été compté avant l'expérience, et 10 jours après. Chaque combinaison plante - herbicide a fait l'objet de 20 expérimentations, plus un témoin sans herbicide (soit 240 expérimentation en tout).

Le tableau de donnée est disponible dans le fichier `herbicide2.csv` (et inclus la variable "survivants" calculer lors du TP 1, qui représente le taux de plantes survivantes dans chaque expérimentation).

1. Charger le fichier `herbicide2.csv` et afficher les données.

**Réponse :**

```
> t = read.table("herbicide2.csv", sep=";", header=TRUE)  
> attach(t)  
> t
```

2. Quels sont les variables dans cette étude ? De quel type sont-elles : numérique, qualitative ordonnée, qualitative non-ordonnée ?

**Réponse :** plante : qualitative non-ordonnée

herbicide : qualitative non-ordonnée

survivants : numérique

(+ deux variables numériques sans intérêt car inclus dans survivants : nb\_plants, nb\_plants\_survivants).

3. Tracer un diagramme de dispersion représentant le taux de plantes survivantes.

**Réponse :**

```
> stripchart(survivants)
```

4. Tracer un histogramme représentant le taux de plantes survivantes. Quel graphique vous semble le plus lisible : le diagramme de dispersion ou l'histogramme ?

**Réponse :**

```
> hist(survivants)
```

5. Tracer un camembert représentant la proportion des différentes plantes dans l'étude.

**Réponse :**

```
> pie(summary(plante))
```

6. Tracer un graphique représentant le taux de survivants en fonction de l'espèce de plante. Dans l'ensemble, quelle espèce résiste le mieux aux trois herbicides ?

**Réponse :**

```
> boxplot(survivants ~ plante)
```

Le blé résiste mieux.

7. Tracer un graphique représentant le taux de survivants en fonction de l'herbicide. Dans l'ensemble, quel herbicide semble le plus efficace ?

**Réponse :**

```
> boxplot(survivants ~ herbicide)
```

L'herbicide 3 est le plus efficace, tout type de plante confondu.

8. Tracer un graphique représentant le taux de survivants en fonction de l'herbicide et de l'espèce de plante. Commenter l'efficacité de chaque herbicide sur chaque type de plante.

**Réponse :**

```
> require(lattice)
```

```
> levelplot(survivants ~ plante * herbicide)
```

9. En vous aidant du graphique précédent, répondre aux questions suivantes :

- (a) quel herbicide est le plus approprié pour appliquer sur une route, où l'on souhaite qu'aucune plante ne pousse ?

**Réponse :** L'herbicide 3.

- (b) quel herbicide est le plus approprié pour appliquer sur un champ de blé, où l'on souhaite que le blé pousse, mais pas le chiendent ni le liseron ?

**Réponse :** L'herbicide 1 (surtout efficace contre le chiendent) ou le 2 (surtout efficace contre le liseron).

- (c) quelle expérience peut-on envisager pour améliorer l'efficacité du traitement de ce champ de blé ?

**Réponse :** Tester l'association des herbicides 1 et 2.

## 8 Exercice 2

La formule de Cockcroft permet de calculer la clairance rénale à partir d'un dosage de la créatininémie (dans le sang). Chez la femme, la formule est la suivante :

$$clairance = \frac{(140 - age) \times poids \times 1,04}{creatininemie}$$

Nous souhaitons vérifier la validité de cette formule pour des patientes âgées. Pour cela, la clairance rénale a été mesurée dans les urines (de 24h) sur un échantillon de 80 patientes âgées, et nous allons comparer cette mesure à l'estimation fournie par la formule de Cockcroft.

1. Charger le fichier cockroft.csv.

**Réponse :**

```
> t = read.table("cockroft.csv", sep=",", header=TRUE)
> attach(t)
```

2. Représenter graphiquement l'âge des patientes, et leur poids.

**Réponse :** Attention, on ne s'intéresse pas ici à l'interaction entre ces deux variables!

```
> hist(age)
> hist(poids)
```

3. Calculer la clairance rénale pour chaque patient, et la mettre dans une nouvelle colonne du tableau de donnée, que l'on appellera "clairance\_cockroft".

**Réponse :**

```
> t["clairance_cockroft"] = (140 - age) * poids * 1.04 / creatininemie
```

4. Représenter graphiquement la clairance rénale mesurée chez les patientes en fonction de la clairance calculée.

**Réponse :**

```
> plot(clairance_cockroft, clairance)
```

5. Ajouter un titre au graphique précédent, et modifier les axes pour qu'ils indiquent "Clairance calculée par la formule de Cockroft" et "Clairance mesurée sur les urines de 24h".

**Réponse :**

```
> plot(cl, clairance, main="mesure de clairance", xlab="Clairance calculée par la for-
mule de Cockroft", ylab="Clairance mesurée sur les urines de 24h")
```