# Fiche 4 : Statistique descriptive avec R

## 1. Les données

Pendant un été, un jardinier a ramassé des haricots de quatre espèces différentes sur son terrain :

- ⋆ "glycine blanche" (code 1) ;
- ⋆ "glycine violette" (code 2) ;
- ⋆ "bignone" (code 3) ;
- ⋆ "laurier rose" (code 4).

Pour chacun des haricots, il a relevé la masse, la taille et l'espèce de celui-ci. Quelques mois après, le jardinier a complété ses données avec deux nouvelles variables :

- ⋆ la masse sèche, relevée sur chaque haricot ;
- ⋆ le nombre de graines contenues dans les gousses des glycines blanches et violettes.

Après avoir transcrit sous `R`, dans un `data.frame` nommé `haricots`, les données initialement enregistrées dans le fichier `haricots.csv` à l'aide de la commande `read.csv`, on peut visionner tout ou partie des données à l'aide des commandes :

```
haricots
head(haricots,12)
tail(haricots,10)
```

Les nombres indiqués en option des fonctions `head` (resp. `tail`) sont les longueurs des parties de début (resp. de fin) du tableau des données. Les noms des colonnes de ce tableau sont fournies par la commande `names(haricots)`. En utilisant la fonction `class`, remarquez que `R` voit la variable `espece` comme qualitative ("`factor`") et toutes les autres comme quantitatives ("`numeric`"). Or la variable `code` n'est que le codage de la variable `espece`. Pour que `R` traite la variable `code` comme une variable qualitative, utilisez la fonction `as.factor` ; et pour vérifier les modalités d'une variable qualitative, employez la fonction `levels`.

```
class(haricots$masse)
haricots$code<-as.factor(haricots$code)
levels(haricots$code)
```

## 2. Variable qualitative

Nous allons ainsi étudier la variable qualitative `espece` (ou `code`). Pour obtenir les effectifs et les fréquences, tapez les commandes :

```
table(haricots$espece)
prop.table(table(haricots$espece))
```

Pour tracer le diagramme en bâtons ou en barres correspondant à cette distribution, on peut utiliser la fonction `plot` ou la fonction `barplot` :

```
plot(table(haricots$espece),type="h",lwd=4,xlab="Espèce",ylab="Effectif")
plot(prop.table(table(haricots$espece)),lwd=7,xlab="Espèce",ylab="Fréquence")
barplot(table(haricots$espece),xlab="Espèce",ylab="Effectif")
barplot(prop.table(table(haricots$espece)),xlab="Espèce",ylab="Effectif")
```

Enfin, le "camembert" est obtenu par la fonction `pie` :

```
pie(table(haricots$espece))
```

## 3. Variable quantitative

### 3.1. **Variable quantitative discrète.**

On s'intéresse désormais à la variable quantitative discrète `graines`. Pour calculer les effectifs, les fréquences, les effectifs cumulés croissants et décroissants, les fréquences cumulées croissantes et décroissantes, on utilise en plus des fonctions `table` et `prop.table`, les fonctions `cumsum` et `rev` :

```
eff<-table(haricots$graines)
ecc<-cumsum(eff)
ecd<-rev(cumsum(rev(eff)))
freq<-prop.table(eff)
fcc<-cumsum(freq)
fcd<-rev(cumsum(rev(freq)))
```

Le diagramme en bâtons se construit à l'aide de la fonction `plot` comme pour une variable qualitative (préférez `plot` à `barplot` pour obtenir des bâtons et non des barres qui peuvent prêter à confusion avec un histogramme). On peut tracer la fonction de répartition empirique de cette distribution :

```
plot(ecdf(haricots$graines))
```

où `ecdf` désigne la fonction de répartition empirique sous `R`.

Par contre, on ne peut pas calculer les différents indicateurs numériques de cette série statistique à cause des éléments non renseignés `NA` (Not Available). Il est cependant possible de le faire à l'aide des commandes suivantes et en ajoutant une option :

| Indicateur | Commande |
|---|---|
| Moyenne | `mean(haricots$graines,na.rm=T)` |
| Mode | `sort(table(haricots$graines),decreasing=T)[1]` |
| Médiane | `median(haricots$graines,na.rm=T)` |
| Variance corrigée | `var(haricots$graines,na.rm=T)` |
| Écart-type corrigé | `sd(haricots$graines,na.rm=T)` |
| Quantile d'ordre $p$ | `quantile(haricots$graines,p,na.rm=T,type=1)` |
| Étendue | `diff(range(haricots$graines,na.rm))` |
| Étendue interquartile | `IQR(haricots$graines,na.rm)` |
| Résumé | `summary(haricots$graines,na.rm)` |

**Remarque 1. Attention** : le logiciel `R` calcule **les variances corrigées et les écarts-types corrigés** ! En effet, pour une série statistique $(x_i)_{i=1,\dots,n}$, il calcule la variance corrigée

$$\sigma_c^2(x) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2,$$

plutôt que la variance non corrigée

$$\sigma^2(x) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2.$$

Il fournit aussi l'écart-type corrigé qui est la racine carrée de la variance corrigée.

### 3.2. **Variable quantitative continue.**

3.2.1. *Données brutes.*
Prenons par exemple le cas de la variable quantitative continue `masse` dont les valeurs observées sont fournies sans regroupement en classes. Pour construire le tableau des fréquences et l'histogramme correspondant, il faut normalement tout d'abord définir les classes et dénombrer leurs effectifs. La fonction `hist` permet d'obtenir à la fois l'histogramme et les effectifs et fréquences :

| Objet | Commande |
|---|---|
| Histogramme | `histo<-hist(haricots$masse,freq=F)` |
| Classes | `histo$breaks` |
| Effectifs | `histo$counts` |
| Fréquences | `histo$counts/sum(histo$counts)` |
| Eff. Cum. Croiss. | `cumsum(histo$counts)` |
| Fréq. Cum. Croiss. | `cumsum(histo$counts)/sum(histo$counts)` |
| Eff. Cum. Décroiss. | `rev(cumsum(rev(histo$counts)))` |
| Fréq. Cum. Décroiss. | `rev(cumsum(rev(histo$counts))/sum(histo$counts))` |

Il est possible de choisir soi-même les classes et leur nombre en utilisant les options `breaks` et `nclass` (consultez l'aide de la fonction `hist`).

Les indicateurs numériques s'obtiennent à l'aide des mêmes fonctions que pour une variable quantitative discrète :

| Indicateur | Commande |
|---|---|
| Moyenne | `mean(haricots$masse)` |
| Mode | `sort(hist(haricots$masse)$density,decreasing=T)[1]` |
| Médiane | `quantile(haricots$masse,0.5,type=4)` |
| Variance corrigée | `var(haricots$masse)` |
| Écart-type corrigé | `sd(haricots$masse)` |
| Quantile d'ordre $p$ | `quantile(haricots$masse,p,type=4)` |
| Étendue | `diff(range(haricots$masse))` |
| Étendue interquartile | `IQR(haricots$masse,type=4)` |

Le logiciel `R` comprend différentes méthodes de calculs pour les quantiles d'une variable quantitative continue. La plus simple est celle de `type 4` qui consiste en une interpolation linéaire de la fonction de répartition empirique des données. Le principe est le suivant : pour calculer le quantile d'ordre 0.64 par exemple, on effectue $252 \times 0.64 = 161.28$ et on regarde ainsi quelles sont les 161 et 162 èmes valeurs de la série statistique **rangée dans l'ordre croissant** (à l'aide de la fonction `sort`). Les valeurs correspondantes sont 11.7 et 12, donc la valeur du quantile d'ordre 0.64 est :

$$q_{0.64} = (1 - 0.28) \times 11.7 + 0.28 \times 12 = 11.784$$

3.2.2. *Données regroupées en classes.* Lorsqu'on est en présence de données déjà regroupées en classe, sans avoir à disposition les données brutes correspondantes, on fera les calculs *à la main* car le logiciel `R` ne propose pas d'outils adaptés à cette situation. On utilisera notamment les formules du cours pour calculer les indicateurs numériques usuels, on construira histogramme et polygône des fréquences cumulées croissantes sur lequel on s'appuiera pour déterminer les quantiles.

## 4. Exercices

**Exercice 1.** Un dénombrement de globules rouges, effectué grâce aux 500 cases d'un hématimètre, a donné le résultat suivant, où, pour chaque $i = 0, 1, \ldots, 10$, $n_i$ est le nombre de cases de l'hématimètre qui contiennent $i$ globules rouges.

| $i$ | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | total |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $n_i$ | 12 | 42 | 91 | 111 | 100 | 66 | 46 | 21 | 8 | 2 | 1 | 500 |

a) Préciser la nature de la variable et faire une représentation graphique appropriée.
b) Établir le tableau des fréquences complet de cette distribution statistique.
c) Calculer la moyenne, le mode et la médiane de la variable observée.
d) Déterminer la variance et l'écart-type (non corrigés).

**Exercice 2.** On a mesuré la taille (en cm) de 40 élèves d'une classe et on a obtenu les résultats suivants :

| | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 138 | 164 | 150 | 132 | 144 | 125 | 149 | 157 | 146 | 158 | 140 | 147 | 136 | 148 | 152 | 144 | 168 | 126 | 138 | 176 |
| 163 | 119 | 154 | 165 | 146 | 173 | 142 | 147 | 135 | 153 | 140 | 135 | 161 | 145 | 135 | 142 | 150 | 156 | 145 | 128 |

a) Calculer la moyenne et la variance (non corrigée) des tailles.
b) Déterminer l'écart interdécile.
c) Regrouper les données en 10 classes, puis en 5 classes. Représenter graphiquement les données obtenues dans les deux cas à l'aide d'un histogramme. Calculer la moyenne dans les deux cas. Commenter les résultats obtenus.

**Exercice 3.** Un biologiste a obtenu les données suivantes sur le nombre de vertèbres des poissons d'une certaine espèce, capturés dans un fjord danois.

| nombre de vertèbres | 104 | 105 | 106 | 107 | 108 | 109 | 110 | 111 | 112 | 113 | 114 | 115 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| nombre de poissons | 1 | 2 | 10 | 12 | 25 | 40 | 35 | 32 | 11 | 5 | 2 | 2 |

a) Réaliser un tableau de fréquences et de fréquences cumulées croissantes.
b) Proposer une représentation graphique des données.
c) Calculer la moyenne, le mode et la médiane du nombre de vertèbres.
d) Calculer l'étendue, l'étendue interquartile, la variance et l'écart type (non corrigés) de ce nombre.

**Exercice 4.** La répartition des salaires en France obtenue pour un échantillon effectué en 1990 est la suivante :

| salaire en $10^3$ francs | $[3\,;\,4[$ | $[4\,;\,6[$ | $[6\,;\,8[$ | $[8\,;\,9[$ | $[9\,;\,10[$ | $[10\,;\,12[$ |
|---|---|---|---|---|---|---|
| effectif | 12 | 30 | 120 | 210 | 90 | 25 |

a) Calculer le salaire moyen, le salaire médian et le premier quartile. Interpréter ces résultats.
b) Pourquoi avoir choisi des classes de longueurs inégales pour grouper les valeurs ? Comment pensez-vous que cela a influencé les résultats obtenus à la question précédente ?