
Fiche TP2 : Statistique descriptive

1 Les données

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

Glycine blanche	Glycine violette	Bignone	Laurier rose
-----------------	------------------	---------	--------------

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

```
haricots<-read.csv("haricots.csv",header=TRUE,sep=";",dec=",")
```

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 à l'aide de la fonction `names`. Pour visualiser les modalités d'une variable qualitative (ici la variable "`espece`", on utilise la fonction `levels`.

```
names(haricots)
levels(haricots$espece)
```

Afin de raccourcir les commandes, employez la commande `attach(haricots)` qui permet de taper simplement le nom des variables sans avoir à indiquer le nom du jeu de données la contenant (par exemple on tapera `graines` au lieu de `haricots$graines`, ou `espece` au lieu de `haricots$espece`).

2 Variable qualitative

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

```
x<-table(espece)
p<-prop.table(x)
```

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

```
plot(x,xlab="Espèce",ylab="Effectif")
```

```
plot(p,xlab="Espèce",ylab="Fréquence")
barplot(x,xlab="Espèce",ylab="Effectif")
barplot(p,xlab="Espèce",ylab="Fréquence")
```

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

```
pie(x)
```

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(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 (qui prend pour ordonnée la *densité de fréquence*) :

```
plot(freq)
```


On peut tracer la fonction de répartition empirique de cette distribution :

```
plot(ecdf(graines))
```

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

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	<code>mean(graines,na.rm=T)</code>
Mode	<code>sort(eff,decreasing=T)[1]</code>
Médiane	<code>median(graines,na.rm=T)</code>
Variance corrigée	<code>var(graines,na.rm=T)</code>
Écart-type corrigé	<code>sd(graines,na.rm=T)</code>
Quantile d'ordre a	<code>quantile(graines,a,na.rm=T,type=1)</code>
Résumé	<code>summary(graines,na.rm=T)</code>

Attention : le logiciel  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

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

plutôt que la variance non corrigée

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

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	<code>h<-hist(masse,freq=F,right=F)</code>
Classes	<code>h\$breaks</code>
Effectifs	<code>h\$counts</code>
Fréquences	
Eff. Cum. Croiss.	
Fréq. Cum. Croiss.	
Eff. Cum. Décroiss.	
Fréq. Cum. Décroiss.	
Densité de fréq.	<code>h\$density</code>

Un objet créé avec la fonction **hist** est en fait un `data.frame` contenant plusieurs informations ; pour les visualiser , taper simplement le nom de l'objet en question (ici **h**). Il est possible de choisir soi-même les classes et leur nombre en utilisant les options **breaks** ou **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	<code>mean(masse)</code>
Mode	<code>h\$mids[h\$density==max(h\$density)]</code>
Médiane	<code>median(masse)</code>
Variance corrigée	<code>var(masse)</code>
Écart-type corrigé	<code>sd(masse)</code>
Quantile d'ordre a	<code>quantile(masse,a,type=6)</code>
Résumé	<code>summary(masse)</code>

Le logiciel **R** comprend différentes méthodes de calculs pour les quantiles d'une variable quantitative continue. La plus simple est celle de **type 6** 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 de notre série statistique **masse** qui est de longueur 252 (faire `length(masse)`), on effectue $(252 + 1) \times 0.64 = 161.92$ 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 :

$$q_{0.64} = (1 - 0.92) \times 11.7 + 0.92 \times 12 = 11.976$$

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 polygone des fréquences cumulées croissantes sur lequel on s'appuiera pour déterminer les quantiles (voir l'exercice 2).

4 Exercices

Exercice 1

La répartition de 100 exploitations agricoles selon leurs superficies en hectare se présente comme suit :

Superficie (ha)	[0 ; 5[[5 ; 10[[10 ; 20[[20 ; 50[[50 ; 100[
Nombre d'exploitations	5	24	38	26	7

- Construire l'histogramme associé à cette répartition en respectant bien les classes indiquées.
- Déterminer les effectifs, fréquences, e.c.c., e.c.d., f.c.c., f.c.d. correspondants.
- Déterminer le mode de cette distribution.
- À l'aide des données brutes, calculer la superficie moyenne ainsi que sa variance corrigée.
- Déterminer la médiane des données brutes.

Exercice 2

Des biologistes marins s'intéressent à une famille de vers présents dans les sables des côtes de la Manche. Afin de mieux connaître les habitudes de cette espèce, ils décident de prélever une même quantité de sable en différents endroits et comptent le nombre de vers dans chaque prélèvement. La répartition des prélèvements selon le nombre de vers observés est donnée dans le tableau ci-dessous :

Vers	0	1	2	3	4	5	6	7
Prélèvements	13	27	28	19	8	3	1	1

- Préciser la nature de la variable et faire une représentation graphique appropriée.

- b) Établir le tableau des fréquences (avec f.c.c. et f.c.d.) de cette distribution.
- c) Calculer la moyenne et la variance de la variable observée.
- d) Déterminer le mode et le premier quartile cette série statistique.

Exercice 3

Un dénombrement de globules rouges, effectué grâce aux 500 cases d'un hématimètre, a donné le résultat suivant, où, pour chaque $i = 0, 1, \dots, 10$, 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 corrigés.

Exercice 4 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 des tailles. Déterminer l'écart interdécile.
- b) Regrouper les données en 10 classes de tailles égales allant de 118 à 178. Représenter graphiquement les données obtenues cas à l'aide d'un histogramme. Calculer la moyenne et la variance de ce regroupement.
- c) Reprendre la question précédente en regroupant les données selon les classes $[118; 132[$, $[132; 144[$, $[144; 152[$, $[152; 164[$ et $[164; 178[$.