

Travail Pratique 3

Les statistiques descriptives avec R

STT1039
BIOLOGIE QUANTITATIVE

Réalisé par
Marc Pépino



Département des sciences de l'environnement

1. Objectifs du TP3

Le travail pratique 3 (TP 3) a pour objectif général de familiariser les étudiants à l'extraction des statistiques descriptives à l'aide du logiciel R. À la fin du TP 3, l'étudiant doit savoir : obtenir les principales statistiques descriptives à l'aide des fonctions de bases, utiliser les fonctions graphiques vues au TP 2 pour extraire rapidement certaines statistiques descriptives, écrire ses propres fonctions pour calculer des statistiques spécifiques. L'accent est mis sur la documentation disponible en ligne en vue d'orienter de plus en plus les étudiants vers l'autonomie.

Note : Tout ce qui est mis en rouge et en gras dans les exemples (encadrés) sont des noms donnés par l'utilisateur. Ces noms sont à modifier en conséquence si les exemples sont utilisés avec d'autres bases de données.

Les statistiques présentées dans les exemples seront pour la plupart réalisés à partir de la base de données *TP3_Exemples.csv* disponible sur le portail du cours :

```
> #Importation de la base de données
> tab <- read.csv("D:/bureau/TP3_Exemples.csv",sep=";",dec=".",h=T)
>
> #Visualisation de la structure du tableau de données
> str(tab)
'data.frame': 100 obs. of 4 variables:
 $ x : num 185 202 203 191 198 ...
 $ y : num 530 644 682 551 616 ...
 $ f : Factor w/ 2 levels "A","B": 1 1 1 1 1 1 1 1 1 1 ...
 $ f2: Factor w/ 2 levels "J","N": 1 2 2 1 2 1 1 1 1 1 ...
```

OPTIONNEL :

```
#Une base de données similaire peut être obtenue à partir du code suivant:
#Construction des variables
x <- c(rnorm(60,200,10),rnorm(40,210,7))
f <- c(rep("A",60),rep("B",40))
f2 <- c(rep("N",50),rep("J",50))[sample(100)]
y <- rep(NA,100)
y[1:60] <- 55 + 3*x[1:60] + rnorm(60,0,50)
y[61:100] <- 60 + 3*x[61:100] + rnorm(40,0,60)

#Construction du tableau de données
tab <- as.data.frame(x)
tab$y <- y
tab$f <- as.factor(f)
tab$f2 <- as.factor(f2)
```

Pour plus de détails sur les statistiques descriptives avec R, vous pouvez consulter, entre autres, **Crawley (2015)**, **Millot (2018)**, ou **Bertrand et Maumy-Bertrand (2010)**.

2. Les fonctions de base pour obtenir les statistiques descriptives

2.1. La fonction `summary()`

La fonction `summary()` a des applications multiples (résumé d'un tableau de données, d'analyses...) comme nous le verrons dans les prochains TP. Cette fonction peut être utilisée ici pour obtenir rapidement certaines statistiques descriptives pour chacune des variables contenues dans un tableau de données:

```
> #Exemple 1
> #Obtenir les statistiques descriptives d'un tableau de données
> summary(tab)
      x          y          f          f2
Min. :173.9 Min. :529.1 A:60 J:50
1st Qu.:195.6 1st Qu.:613.2 B:40 N:50
Median:205.6 Median:655.5
Mean :203.1 Mean :662.0
3rd Qu.:210.2 3rd Qu.:709.3
Max. :225.1 Max. :880.6
```

La fonction `summary()` donne six statistiques descriptives (minimum, premier quartile, médiane, moyenne, troisième quartile, maximum) pour des variables quantitatives et un tableau de fréquence absolue pour des variables qualitatives (exemple 1).

2.2. Variables quantitatives

Le tableau suivant donne les principales fonctions de bases pour extraire des statistiques descriptives de variables quantitatives :

Tableau 1. Statistiques descriptives pour des variables quantitatives

Fonction	Définition
<code>mean(x)</code>	Moyenne de x
<code>median(x)</code>	Médiane de x
<code>min(x), max(x)</code>	Minimum et maximum de x
<code>var(x)</code>	Variance de x
<code>sd(x)</code>	Écart type de x
<code>sum(x)</code>	Somme de x
<code>quantile(x)</code>	Quantile de x (l'argument <i>prob</i> permet de modifier les quantiles calculés par défaut)
<code>cov(x,y)</code>	Covariance entre x et y
<code>cor(x,y)</code>	Corrélation entre x et y

Note : l'argument « **na.rm = T** » permet d'omettre les données manquantes avant le calcul des statistiques descriptives présentées au tableau 1 :

```

> #Exemple 2
> #Créer une variable x dans laquelle la troisième valeur est manquante
> x <- tab$x
> x[3] <- NA
> #Calculer le quantile 95% de x
> quantile(x,prob=0.95)
Erreur dans quantile.default(x, prob = 0.95) : les valeurs manquantes et
les NaN ne sont pas autorisés si 'na.rm' est FALSE
> #Calculer le quantile 95% de x
> quantile(x,prob=0.95,na.rm=T)
  95%
219.4383

```

2.3. Variables qualitatives

La fonction **table()** produit un tableau de fréquence absolue si elle est appliquée à une variable qualitative et un tableau de contingence si elle est appliquée à deux variables qualitatives :

```

> #Exemple 3
> #Tableau de fréquence absolue pour une variable qualitative
> table(tab$f)
tab$f
 A  B
60 40
> #Tableau de contingence pour deux variables qualitatives
> table(tab$f,tab$f2)
      tab$f2
tab$f  J  N
  A 31 29
  B 19 21

```

2.4. Statistiques descriptives détaillées par groupe

Les fonctions **tapply()** et **aggregate()** peuvent servir à calculer rapidement des statistiques descriptives à des sous-groupes de données (ex. : moyenne des longueurs suivant le sexe). Le premier argument correspond à la variable pour laquelle on souhaite appliquer une fonction statistique, le second argument se présente sous forme de liste et contient les groupes servant à séparer la base de données, et le troisième argument contient la fonction statistique à appliquer (voir exemple 4). On peut aussi utiliser la fonction **summary()** vue à la section 2.1.

```

> #Exemple 4
> #Exemple d'utilisation de la fonction aggregate
> aggregate(tab$x,list(tab$f,tab$f2),mean)
  Group.1 Group.2      x
1      A      J 199.0391
2      B      J 208.3086
3      A      N 199.5342
4      B      N 209.1761
> #Exemple d'utilisation de la fonction tapply
> tapply(tab$x,list(tab$f,tab$f2),mean)
      J      N
A 199.0391 199.5342
B 208.3086 209.1761

```

Note : la fonction *aggregate()* retourne un objet se présentant sous la forme d'un tableau de données (data frame).

3. Utilisation de l'histogramme : mode et tableau de fréquence

3.1. Calcul du mode

La fonction *hist()*, qui génère un histogramme, peut-être utilisée pour trouver le **mode** d'une distribution (soit le centre de classe ayant la plus haute fréquence). La méthode n'est cependant pas parfaite. Elle ne permet pas, par exemple, de déterminer tous les modes de distributions bi- ou multi-modales. Pour cela, il est toujours conseillé de faire une représentation graphique des données sous forme d'histogramme et de déterminer les modes directement par lecture du graphique.

```

> #Exemple 5
> #calculer le mode d'une variable
> #Attention: valable pour des distributions unimodales
> h <- hist(tab$x,plot=F)
> h$mids[which(h$counts==max(h$counts))]
[1] 207.5

```

Note : la fonction *which()* détermine l'élément du centre de classe pour lequel la fréquence de la classe est maximale.

3.2. Calcul du tableau de fréquence absolue d'une variable quantitative

La fonction *hist()* peut aussi être utilisée pour générer un tableau de fréquence absolue d'une variable quantitative. Par défaut, les classes sont construites en se basant sur la règle de Sturges (voir l'aide en ligne pour plus de détails en tapant « ?hist »).

```

> #Exemple 6
> #calculer le tableau de fréquence absolue d'une variable quantitative
> h <- hist(tab$x,plot=F)
> freq <- cbind(h$mids, h$counts)
> freq
      [,1] [,2]
[1,] 172.5  2
[2,] 177.5  1
[3,] 182.5  2
[4,] 187.5  9
[5,] 192.5  8
[6,] 197.5 13
[7,] 202.5 13
[8,] 207.5 24
[9,] 212.5 20
[10,] 217.5  3
[11,] 222.5  4
[12,] 227.5  1

```

Note : il est possible d'exporter le tableau de fréquence en utilisant la fonction *write.table()* vue au TP 1.

Remarque : les fréquences cumulées s'obtiennent à l'aide de la fonction *cumsum()*. Dans cet exemple, on obtiendrait les fréquences cumulées en tapant « *cumsum(h\$counts)* ». Le pourcentage de fréquences cumulées s'obtient en divisant les fréquences cumulées par l'effectif de l'échantillon : « *cumsum(h\$counts) / sum(h\$counts)* »

4. Écrire ses propres fonctions

Pour obtenir des statistiques particulières qui ne sont pas calculées par défaut, il est possible d'écrire (connaissant l'équation mathématique) sa propre fonction à l'aide de la fonction *function()*. Cette fonction contient en argument la ou les variables sur lesquelles seront faites le calcul. À l'extérieur de la parenthèse, on écrit en langage R la formule permettant de calculer la statistique. Une fois créée, on peut utiliser cette fonction comme n'importe quelle autre fonction. Il est possible, par exemple, de se construire une fonction qu'on appellera *moyenne()* et qui calculera la moyenne d'une variable. On peut vérifier par la suite que les fonctions *moyenne()* et *mean()* donnent le même résultat !

```

> #Exemple 7
> #Créer un fonction pour le calcul de la moyenne
> moyenne <- function(x) sum(x)/length(x)
> moyenne(tab$x)
[1] 203.0727
> mean(tab$x)
[1] 203.0727

```

4.1. Calcul du coefficient d'asymétrie

La fonction suivante, **skew()**, est écrite pour calculer le coefficient d'asymétrie :

```
> #Exemple 8
> #Coefficient d'asymétrie ; extrait de Crawley 2005 (p. 69)
> skew<-function(x){
+ m3<-sum((x-mean(x))^3)/length(x)
+ s3 <-sqrt(var(x))^3
+ m3/s3}
> skew(tab$x)
[1] -0.550315
```

4.2. Calcul du coefficient d'aplatissement

La fonction suivante, **kurtosis()**, est écrite pour calculer le coefficient d'aplatissement :

```
> #Exemple 9
> #Coefficient d'aplatissement ; extrait de Crawley 2005 (p. 72)
> kurtosis<-function(x){
+ m4<-sum((x-mean(x))^4)/length(x)
+ s4 <-var(x)^2
+ m4/s4 - 3}
> kurtosis(tab$x)
[1] 0.0318387
```

Note : Le coefficient d'aplatissement est centré sur 0.

4.3. Réaliser un tableau détaillé des statistiques descriptives

Le script suivant permet de générer un tableau détaillé des statistiques descriptives de toute variable quantitative. Le script contient les fonctions **skew()** et **kurtosis()** vues aux exemples 8 et 9. La fonction **stat()** est créée pour générer le tableau récapitulatif des statistiques descriptives. Il est possible d'exporter le tableau grâce à la fonction **write.table()** vue au TP 1, ce qui permettra d'intégrer facilement le tableau récapitulatif dans un document tel un rapport.

```

> #Exemple 10
> #Faire ses propres statistiques descriptives#####
> stat<-function(x){
+ val <- c(length(x),mean(x),median(x),sd(x),var(x),min(x),max(x),diff(range(x)),
+ sum(x),skew(x),kurtosis(x))
+ nom <- c("n","moyenne","médiane","écart type",
+"variance","minimum","maximum","étendue","somme","a3","a4")
+ tab <- as.data.frame(nom)
+ tab$val <- val
+ tab}
> #####
> fin <- stat(tab$x)
> fin

```

	nom	val
1	n	100.0000000
2	moyenne	203.0726632
3	médiane	205.6319276
4	écart type	10.7179709
5	variance	114.8749009
6	minimum	173.8523638
7	maximum	225.1063600
8	étendue	51.2539962
9	somme	20307.2663154
10	a3	-0.5503150
11	a4	0.0318387

Note : a3 et a4 correspondent respectivement aux coefficients d'asymétrie et d'aplatissement.

5. Conseils pratiques et astuces à retenir

summary(). Cette fonction permet d'avoir un aperçu rapide de la distribution des variables d'un tableau de données.

Tableau 1. Il reprend les principales fonctions pour obtenir les statistiques descriptives de base.

Faire des graphiques. Les graphiques servent à la fois à visualiser la distribution des données mais aussi à extraire certaines statistiques (mode, tableau de fréquence...).

stat(). En compilant les exemples 8 à 10, vous pouvez obtenir un tableau récapitulatif détaillé des statistiques descriptives en utilisant la fonction *stat()*.

Package psych. Vous pouvez installer cette bibliothèque (voir TP 1) et utiliser la fonction *describe()* pour obtenir les principales statistiques descriptives.

6. Exercices

Répondez aux exercices dans un fichier texte « Nom_Prenom_TP3.txt ». Au début du fichier, identifiez-vous par votre nom et prénom, ajouter la date et le numéro du TP, le tout en utilisant des commentaires. Utilisez des commentaires pour structurer vos réponses.

Pour les exercices, importez le tableau de données *TP3_Exercices.csv* disponible sur votre portail étudiant. **Optionnel** : vous pouvez utiliser également le code mentionné à la fin de l'exercice 6 pour refaire les exercices avec une nouvelle base de données.

Exercice 1. Calculer le minimum, le premier quartile, la médiane, la moyenne, le troisième quartile, et le maximum pour les variables x et y .

Exercice 2. Calculer l'écart type pour x et y . Calculer la corrélation entre x et y à l'aide de la fonction *cor()*. Retrouver ce résultat à l'aide de la covariance et des écarts types.

Exercice 3. Déterminer le mode et le tableau de fréquence relative de la variable y . Vérifier que la somme des fréquences relatives est égale à 1. Exporter le tableau de fréquence relative sur le bureau.

Exercice 4. Calculer les quantiles 5% et 95% de x pour chaque groupe (A et B) de la variable f .

Exercice 5. Réaliser un tableau récapitulatif des statistiques descriptives de la variable y à l'aide de la fonction *stat()* de l'exemple 10 du TP 3. Comparer ces statistiques descriptives avec celles obtenues à l'aide de la fonction *describe()* du package *psych*.

Code pour créer une base de données similaire à celle du fichier *TP3_Exercices.csv* :

```
#Construction des variables
x <- c(rnorm(60,200,10),rnorm(40,210,7))
f <- c(rep("A",60),rep("B",40))
f2 <- c(rep("N",50),rep("J",50))[sample(100)]
y <- rep(NA,100)
y[1:60] <- 55 + 3*x[1:60] + rnorm(60,0,50)
y[61:100] <- 60 + 3*x[61:100] + rnorm(40,0,60)

#Construction du tableau de données
tab <- as.data.frame(x)
tab$y <- y
tab$f <- as.factor(f)
tab$f2 <- as.factor(f2)

#Visualisation de la structure du tableau de données
str(tab)
```

Bibliographie

- Bertrand, F. et Maumy-Bertrand, M. 2010. Initiation à la statistique avec R. Dunod, Paris, France.
- Crawley, M. J. 2013. The R book (2^{ème} éd.). John Wiley and Sons, Chichester, England.
- Crawley, M. J. 2015. Statistics: an introduction using R (2^{ème} éd.). John Wiley and Sons, Chichester, England.
- Legendre, P. 2009. Introduction au langage statistique R.
http://biol09.biol.umontreal.ca/BIO6077/Introduction_R.pdf.
- Millot, G. 2018. Comprendre et réaliser les tests statistiques à l'aide de R : manuel pour les débutants (4^{ème} éd.). De Boeck, Bruxelles, Belgique.
- Murrell, P. 2011. R graphics (2^{ème} éd.). Chapman & Hall/CRC The R series, Boca Raton, Floride.
- Paradis, E. 2005. R pour les débutants.
http://cran.r-project.org/doc/contrib/Paradis-rdebuts_fr.pdf.
- Short, T. 2004. R Reference Card. <http://cran.r-project.org/doc/contrib/Short-refcard.pdf>.

Liens utiles :

- Site officiel de R.
<http://www.r-project.org>
- Ateliers R du CSBQ.
<https://qcbs.ca/wiki/r>
- Quick R.
<http://www.statmethods.net/>
- Apprentissage en ligne du logiciel R.
<https://www.datacamp.com/>
- Beginners statistics introduction.
http://influentialpoints.com/Training/basic_statistics_introduction.htm
- R Graphics Cookbook.
<http://www.cookbook-r.com/Graphs/>
- Liste des bases de données disponibles sous R.
<https://vincentarelbundock.github.io/Rdatasets/datasets.html>