
LECTURE 4. OUTILS POUR DECRIRE ET RESUMER LES DONNEES UNIVARIEES ET BIVARIEES

Prof. Kizungu Vumilia Roger

**UNIKIN (FACAGRO-BIOLOGIE), UNILU (FACAGRO), UEA (FACAGRO), UCB
(FACAGRO), ISS, ISTA (ENVIRONNEMENT), UPN (FACAGRO-MEDVET)**

---GII-GIII-DEA---

Release: 26 décembre 2009

Sommaire

Introduction	2
Description univariée, unidimensionnelle ou description variable par variable	4
Description bivariée, bidimensionnelle ou description simultanée de deux variables qualitatives	29
Description simultanée de deux variables qualitatives : Analyse Factorielle des Correspondances (AFC).....	35
Description bivariée, bidimensionnelle ou description simultanée de deux variables quantitatives	36
Une variable quantitative et une variable qualitative	40

Introduction

Tu sais combien l'emballage d'un produit influence son succès sur le marché des ventes. Tu sais aussi comment un joli logo, un petit slogan qui accompagne une marque concourent plus à son succès que la qualité même du produit. En Biométrie, c'est pareil : le contenu des données est une chose mais ce qu'on peut dire *dans un premier temps* sur ces données est une tout autre. Comment tu les présentes ? Sous forme d'un petit tableau des paramètres qui les résument ? Sous forme d'un graphique qui parle de lui-même ? Ceci est l'objet de ce chapitre. Il constitue ce que l'on appelle **les statistiques descriptives** ou tout simplement **analyse de données**.

Les statistiques descriptives sont un ensemble des procédés pour « presser » un tableau de données afin d'en extraire le maximum d'informations. Ces informations tournent autour de **la distribution des fréquences** des individus dans des classes ou des modalités des variables.

Les statistiques descriptives concernent les données telles qu'elles sont sans chercher à inférer dessus. Elles sont utilisées dans des problèmes des populations totales. Prenons le cas d'un sélectionneur qui a 150 accessions (individus) dans sa collection. Il a mesuré certains caractères (variables) sur ces accessions. Il dispose d'une population totale dont il veut commenter les données. Prenons aussi le cas d'un vétérinaire qui étudie la population des vaches mortes pendant une période dans une région donnée. Il dispose des fiches de chacune des vaches. C'est aussi un problème d'une population totale. Souvent les éleveurs sont tenus de déclarer les effectifs de leurs cheptels aux services spécialisés. C'est un problème d'une population totale. C'est le cas aussi de l'étude d'une population des vaux d'une ferme. Les propriétaires tiennent souvent des fiches bien suivies. Il y a lieu de décrire la population des petits nés cette année ou l'année passée. On peut compter les cas des maladies ou des morts. Les morts du cheptel peuvent même être utilisés comme un moyen d'évaluer le niveau de soin de l'élevage. Le taux de mortalité peut ainsi être lié aux divers soins. Le propriétaire du cheptel peut bien vouloir savoir le nombre de décès par sexe, par race etc. Ceci est aussi un exemple d'une population totale.

Une étude qui porte sur une population totale ne se soucie pas des autres individus en dehors de la population en étude. Il n'est pas question de généraliser le résultat aux autres individus qui ne concernent pas l'étude.

La collecte de données dans une étude de population totale débouche sur un tableau de données individus x variables. Une étude peut concerner une seule variable ou simultanément deux ou plusieurs variables.

Une étude univariée consiste à étudier seulement une variable. Par exemple, sur 10 athlètes d'une université, l'intérêt de l'étude peut porter seulement sur leur sexe. Il en résulte un ensemble de 10 données qualitatives nominales.

Sex = F, M, M, F, M, F, M, M, F, F.

Si l'intérêt porte seulement sur l'exercice physique athlètes, il en vient une variable qualitative ordinaire Exer :

Exer = Quelque fois, Non, Quelque fois, Quelque fois, Fréquemment, Fréquemment, Quelque fois, Quelque fois, Fréquemment, Quelque fois.

Si en revanche l'intérêt porte simultanément entre la taille et l'âge, alors les données consistent en 10 couples de données et l'analyse est dite bivariée. Une paire de données est par exemple (173, 18.25) et on lit que l'athlète a 173 cm de taille et a l'âge de 18.25 ans.

Si l'enquêteur prend simultanément sur les 10 athlètes (individus) les informations sur le Sexe (Sex), la longueur de la main active (Lmact), la longueur de la main inactive (Lminact), la main active (Mactive), le bras qu'il pose sur l'autre quand il les croise (Fold), le nombre de battement du cœur par minute (Poul), la main qui a déjà connu une entorse (Entor), le niveau des exercices (Exer), le niveau de fumer (Smoke), la taille (Taille) et l'âge (Age) alors il effectue une étude multivariée avec 11 caractères.

Sex	Lmact	Lminact	Mactive	Fold	Poul	Entor	Exer	Smoke	Taille	Age
F	18.5	18.0	Droite	DsurG	92	Gauche	Quelq	Jamais	173.00	18.250
M	19.5	20.5	Gauche	DsurG	104	Gauche	Non	Regul	177.80	17.583
M	20.0	20.0	Droite	Rien	35	Droite	Quelq	Jamais	165.00	23.667
F	18.0	17.7	Droite	GsurD	64	Droite	Quelq	Jamais	172.72	21.000
M	17.7	17.7	Droite	GsurD	83	Droite	Freq	Jamais	182.88	18.833
F	17.0	17.3	Droite	DsurG	74	Droite	Freq	Jamais	157.00	35.833
M	20.0	19.5	Droite	DsurG	72	Droite	Quelq	Jamais	175.00	19.000
M	18.5	18.5	Droite	DsurG	90	Droite	Quelq	Jamais	167.00	22.333
F	17.0	17.2	Droite	GsurD	80	Droite	Freq	Jamais	156.20	28.500
F	19.5	20.2	Droite	GsurD	66	Rien	Quelq	Jamais	155.00	17.500

Sans une quelconque organisation, il est difficile de ressortir les informations exploitables biologiquement de ce tableau. Dans une analyse multivariée, l'étude doit être méthodique et doit impérativement commencer par une étude univariée puis bivariée et enfin multivariée. Ce chapitre en montre progressivement les avantages.

Après l'étude de ce chapitre tu devras savoir décrire et résumer un tableau de données. Tu auras tous les éléments pour écrire le chapitre Matériels et Méthodes de ta Monographie, de ton Mémoire de fin d'études, de ton Mémoire de DEA et de ta Thèse. En d'autres termes tu devras :

- expliquer comment organiser un tableau de données sous forme d'une matrice et comment construire et interpréter les distributions de fréquences ;
- représenter les données sous forme d'un histogramme, d'un polygone de fréquence, d'un diagramme tige et feuilles, d'un diagramme en moustache etc. ;
- présenter les mesures qui résument et décrivent les propriétés de base d'un tableau de données ;
- calculer les mesures de tendance centrale comme la moyenne arithmétique, la médiane, le mode etc.
- calculer la mesure de dispersion comme l'étendue, la déviation absolue moyenne, l'écart-type etc.
- utiliser le pourcentage pour résumer les données qualitatives.

Commençons alors par l'étude univariée des variables.

Description univariée, unidimensionnelle ou description variable par variable

Les variables, comme on les a classées, sont de deux types : les variables quantitatives et les variables qualitatives. Pour les variables quantitatives, on a distingué aussi des variables quantitatives continues et les variables quantitatives discrètes. Pour les variables qualitatives, nous allons considérer les qualitatives nominales et les qualitatives ordinales. Commençons par les caractères qualitatifs.

Cas d'une variable qualitative

L'analyse d'un caractère qualitatif consiste en la connaissance de la distribution des fréquences des individus dans les différentes modalités du caractère. L'auteur d'un article peut décider de représenter la distribution des fréquences sous forme d'un tableau ou sous forme de graphique.

Le tableau des fréquences place tout simplement la modalité de la variable au regard du nombre d'individus qui lui correspond. C'est une présentation compacte et sans beaucoup de perte d'information.

Considérons la question « Quel est votre sexe ? ». Le sexe n'a que deux modalités (M et F). Combien de personnes ont noté M et combien ont noté F ? Le résultat du comptage donne le tableau de distribution des fréquences (tableau 4.1).

Tableau 4.1. Distribution de fréquences selon le critère Sexe		
Sexe	F	M
Fréquence	### = 5	### = 5
Fréquence relative	0.5	0.5

Le même résultat peut être représenté graphiquement (Fig. 4.1.)

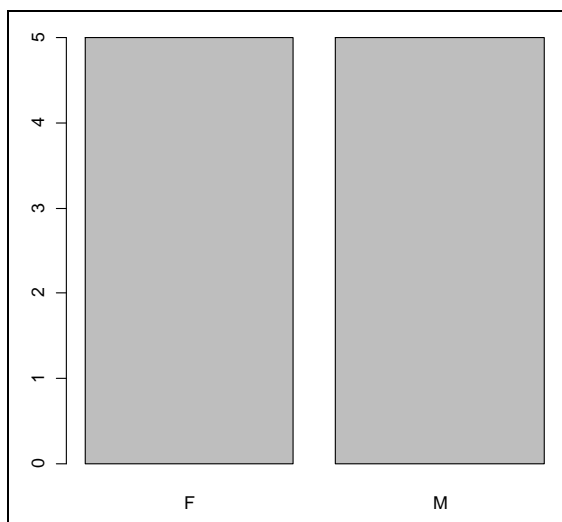


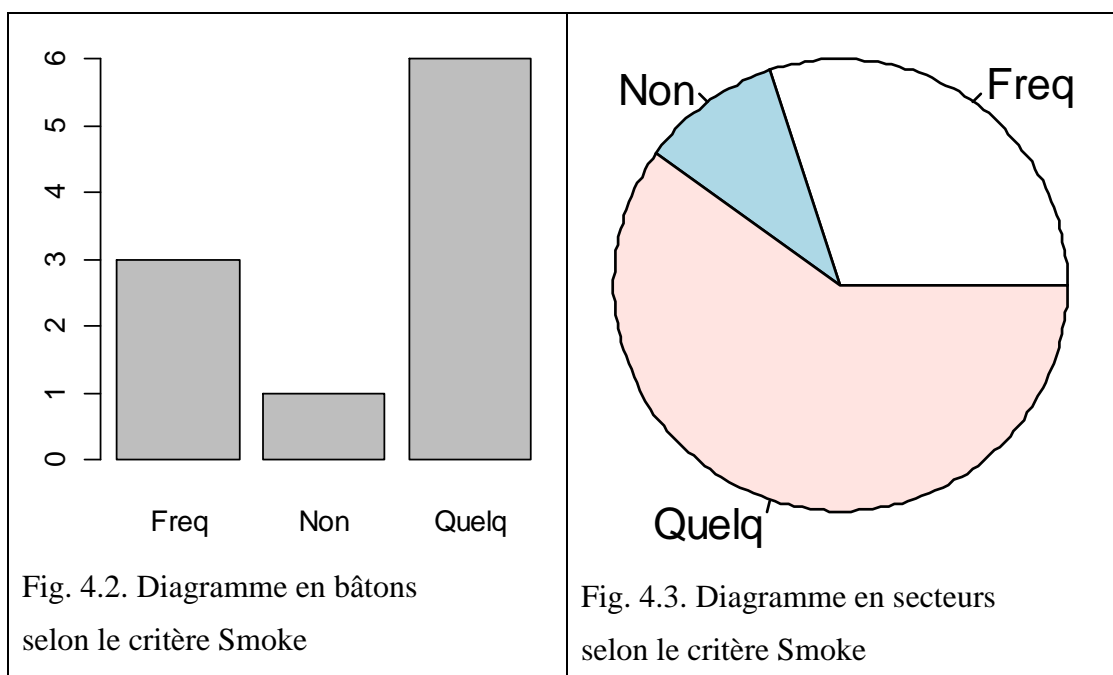
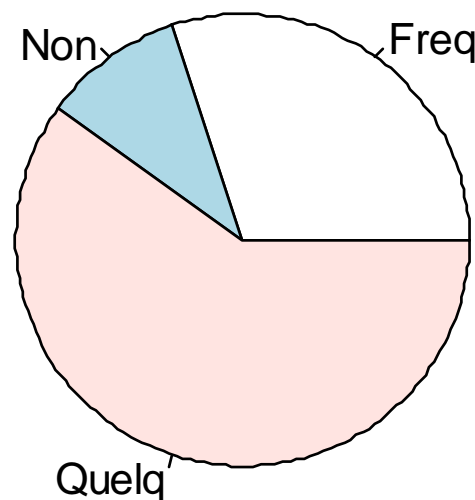
Fig. 4.1. Diagramme en bâtons

Le graphique consiste à tracer des bâtons dont la longueur est proportionnelle à la fréquence des observations. Notons en passant que ce graphique n'est pas présentable dans le cas de l'analyse de données. Tout simplement puisqu'il ne ressort aucune information qui frappe l'œil.

En revanche, si l'on considère la question « faites vous des exercices ? », le tableau de la distribution des fréquences est donnée (tableau 4.2.). Trois personnes ont répondu fréquemment, une personne a répondu Non et six personnes ont répondu Quelque fois.

Tableau 4.2. Distribution de fréquences selon le critère Smoke.			
Tu fumes ?	Fréquemment (Freq)	Non	Quelque fois (Quelq)
Fréquence	/// = 3	/ = 1	### / = 6
Fréquence relative	0.3	0.1	0.6

Une autre présentation de ces données est le diagramme en bâtons (Fig. 4.2.). Ce graphique a l'avantage de frapper l'œil sur le fait qu'il s'agit d'un petit groupe d'athlètes qui fument quelques fois ou fréquemment. Ceci peut suggérer une deuxième façon de représenter les données. Quand on veut ressortir les proportions, alors on dessine le diagramme en secteurs (Fig. 4.2.).

Fig. 4.2. Diagramme en bâtons
selon le critère SmokeFig. 4.3. Diagramme en secteurs
selon le critère Smoke

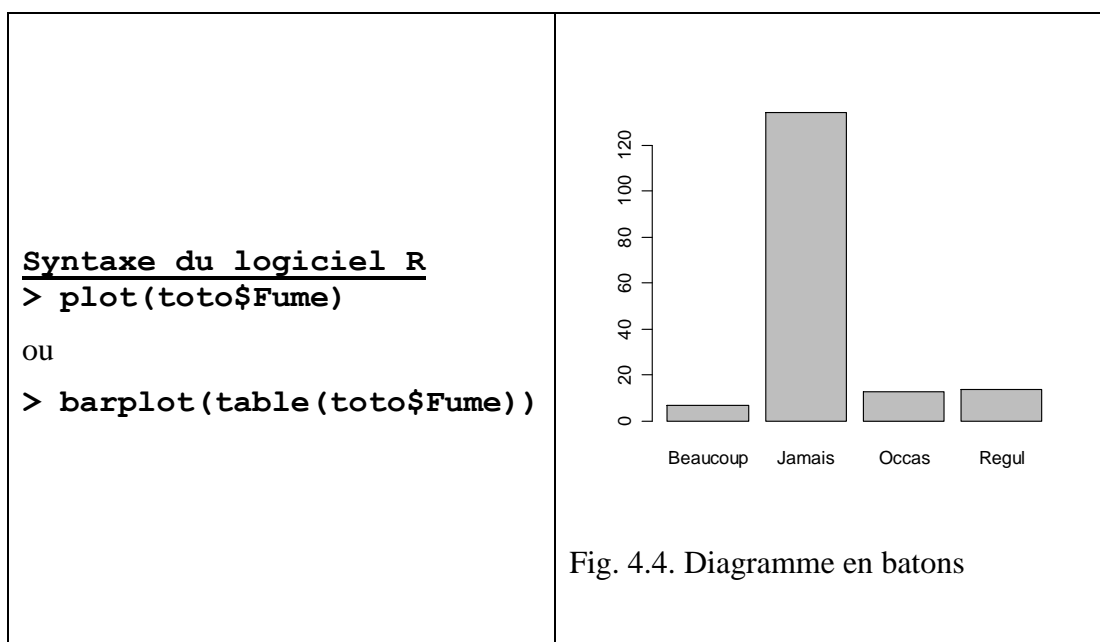
L'exemple sur dix fiches n'exige pas beaucoup d'efforts quant au comptage par niveau de modalité. En considérant les 168 fiches qui ont constitué l'étude, il y a lieu d'utiliser un ordinateur pour effectuer ces comptages sans risque d'erreur.

Le tableau de données renseigne Considérons la question « Est-ce que tu fumes ? » dont les modalités sont Jamais, Occasionnellement, Régulièrement et Beaucoup. La distribution des fréquences est donnée par le tableau 4.3.

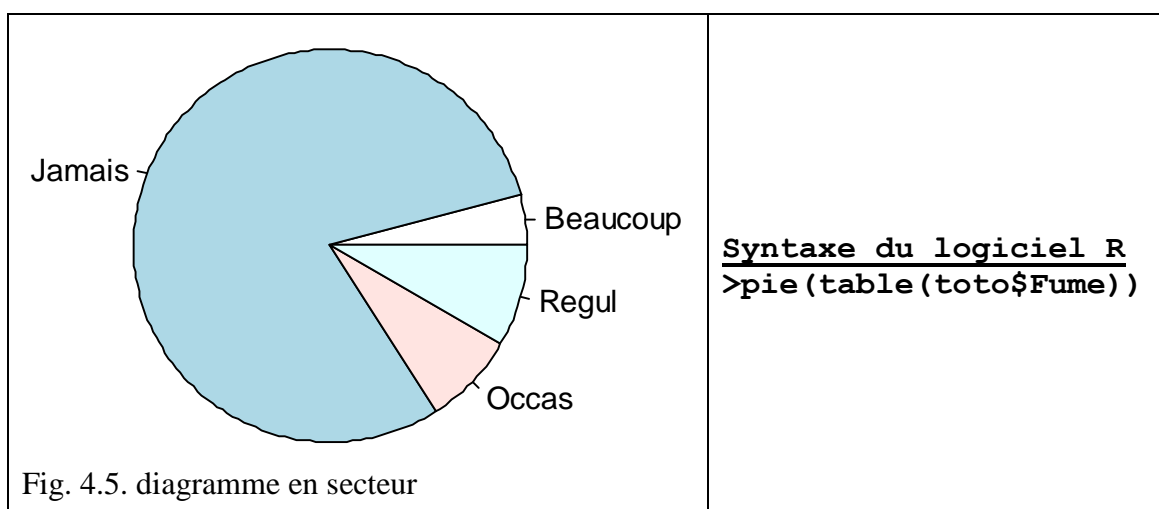
Tableau 4.3. Distribution de fréquence des étudiants selon le critère « Fume ? »			
Fume ?			
Beaucoup	Jamais	Occas	Regul
7	134	13	14

Syntaxe du logiciel R
> table(Fume)

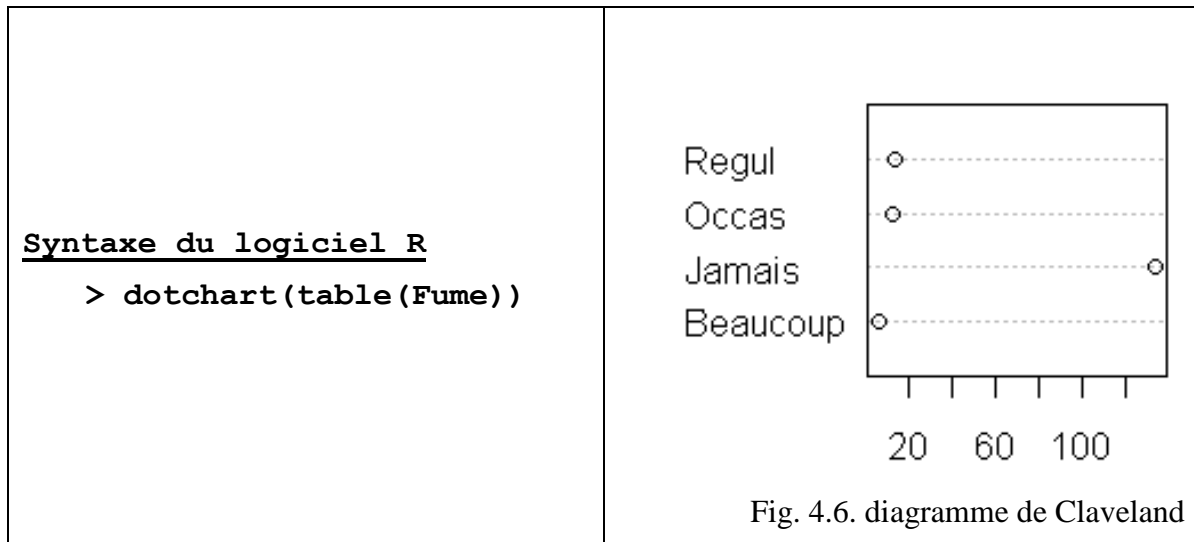
Le diagramme en batons correspondant est donné par la figure 4.4.



Il peut s'avérer important de vouloir visualiser autrement le même résultat dans l'objectif de mettre en relief des proportions. Dans ce cas, le diagramme en secteur se prête bien.



Si les proportions dans un diagramme en secteurs ne sont pas très lisibles, si les proportions ne sont pas distinguables alors on peut utiliser le diagramme de Claveland qui lui n'est pas très populaire.



Pour résumer les données issues d'une variable qualitative, trois représentations graphiques classiques s'offrent à l'auteur comme alternative au tableau de distribution des fréquences. Il s'agit de :

- diagramme en baton,
- diagramme en secteurs et de plus en plus du,
- diagramme de Claveland.

Notons ici en guise de résumé que le choix du mode de représentation des données

Cas d'une variable quantitative

L'analyse d'une variable quantitative commence toujours par la recherche de la *distribution des fréquences* des individus dans les classes créées à partir de la variable. Le premier réflexe doit être de trier les données en ordre croissant. Sur cette série ainsi triée, on peut déterminer :

- l'observation telle que 0% lui soient inférieures. C'est le **minimum**.
- l'observation telle que 25% lui soient inférieures ou égale. C'est le **premier quantile**.
- l'observation telle que 50% lui soient inférieures ou égale. C'est le **deuxième quantile ou la médiane**.
- l'observation telle que 75% lui soient inférieures ou égale. C'est le **troisième quantile**.
- l'observation telle que 100% lui soient inférieures ou égale. C'est le **maximum**.

Sur le plan pratique, la connaissance du maximum et du minimum peut révéler des erreurs de frappe dans le tableau de données. En effet, si la variable en étude est la taille

de l'homme en centimètres et qu'on a comme maximum 1722 cm alors on se rendra compte qu'il y a une virgule omise et que la bonne valeur est 172.2 cm.

Considérons la variable Age de l'étude sur les athlètes de l'université. La commande `summary()` permet rapidement aux valeurs qui renseignent sur la distribution des fréquences.

>summary(Age)					
Min.	1stQu.	Median	Mean	3rdQu.	Max.
16.92	17.67	18.58	20.43	20.17	70.42

Ce tableau permet de constater que la population enquêtée est jeune car 75 % ont un âge inférieur à 20.17 ans. Il permet ensuite de constater qu'il y a un athlète de 70.42 ans.

Le résumé des données permet souvent de détecter si une valeur dans le tableau est aberrante ou pas. Ici on est en droit de se poser la question si réellement il existait un athlète de 70 ans ou s'il s'agit d'une erreur de transcription de données.

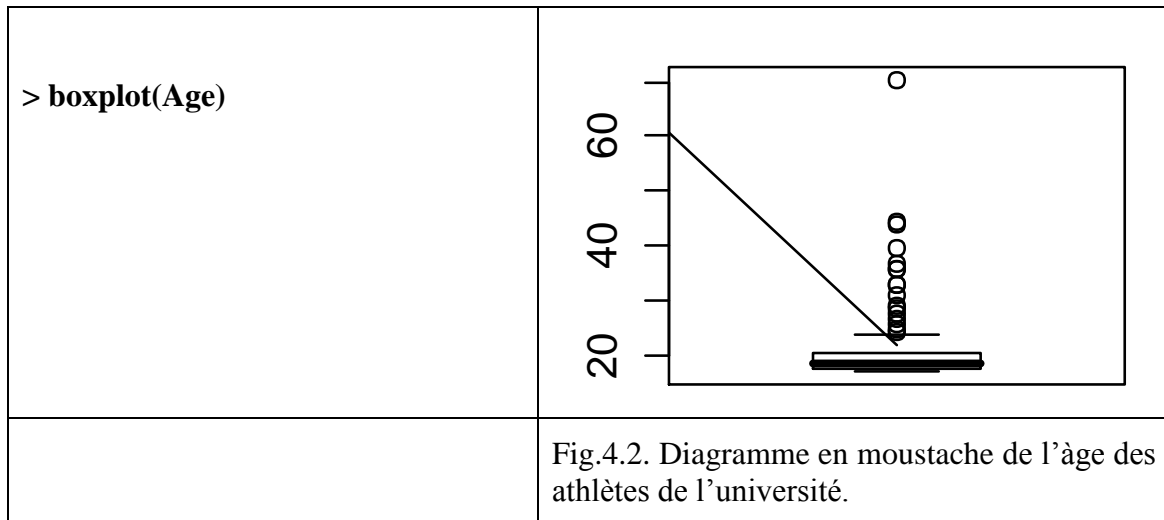
Ces statistiques peuvent être obtenues séparément, successivement par les commandes suivantes :

- `min(Age)`,
- `quantile(Age, probs=0.25)`,
- `median(Age)`,
- `quantile(Age, probs=0.75)` et
- `max(Age)`.

L'alternative de cette présentation de distribution est obtenue soit avec une boîte à moustache, soit avec un histogramme.

Le diagramme à moustache (boxplot) restitue la même information à savoir les différents quantiles. Sa représentation est souvent bénéfique pour le cas des données issues des mesures. Dans la pratique, le boxplot nous sert à avoir les premiers indices de normalité d'une variable. Il suffit de voir que la médiane est du même ordre de grandeur que la moyenne pour dire que probablement la distribution est normale. Il suffit aussi de voir si les données sont symétriques pour aller dans le même sens ou pas dans les présomptions de normalité. Nous y reviendrons après avoir défini la distribution normale.

Considérons le cas de l'âge. Le diagramme en moustache est donné par la figure 4.2.



Ce graphique ne nous offre pas l'opportunité de commenter le résultat. Ceci puisque il y a une grande dispersion causée par la présence de l'athlète très âgé.

L'alternative pour rencontrer l'interprétation du résumé est l'histogramme. Pour construire l'histogramme :

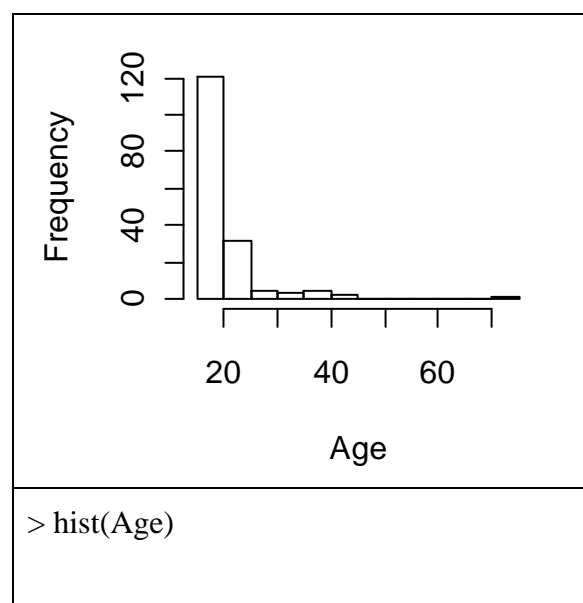
- la première étape est celle de considérer l'étendue de la série c'est-à-dire la valeur maximum moins la valeur minimum.
- la deuxième étape est celle de déterminer des intervalles ou des classes. Nous allons voir dans le paragraphe qui suit, comment créer des classes d'une variable quantitative.
- la troisième étape est celle de compter les individus correspondant à chaque intervalle. Cela se passe de la même façon que pour le diagramme en bâton sauf qu'ici, la largeur des bâtons contient une information. Et donc, la surface d'un bâton est proportionnelle à la fréquence.

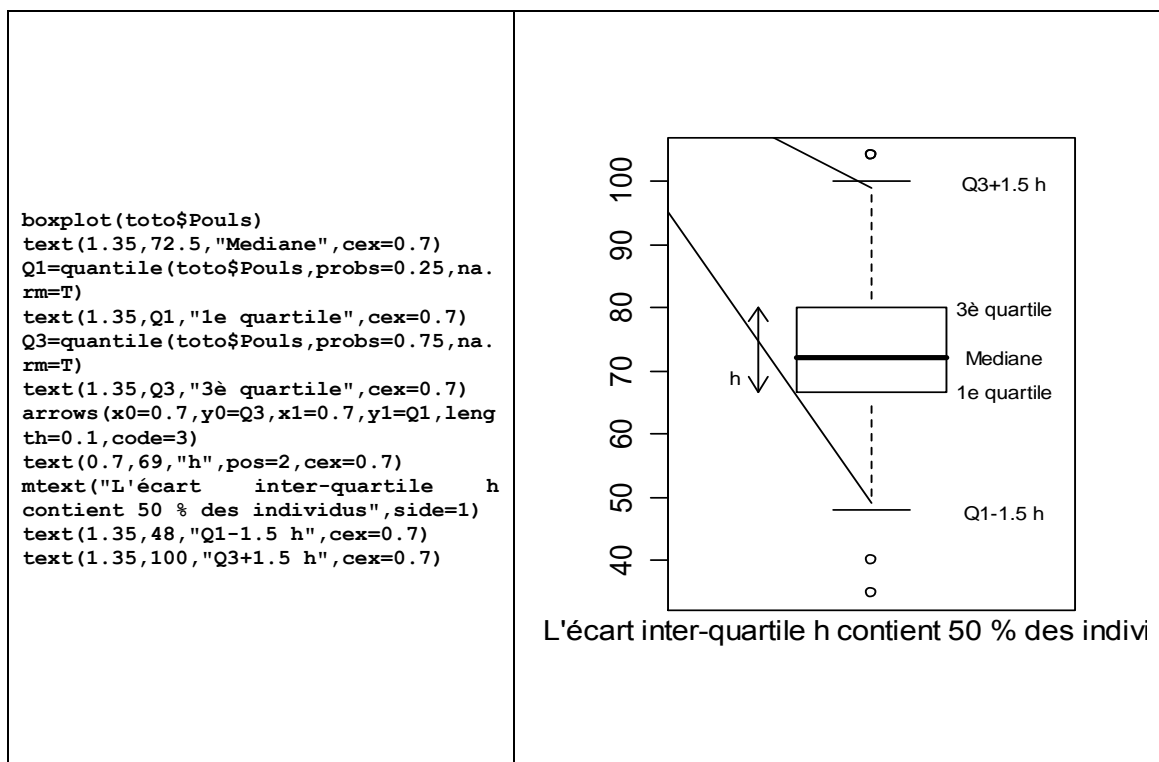
Considérons l'âge.

Contrairement au diagramme en moustache, l'histogramme restitue mieux l'information sur les quantiles et fait ressortir l'information sur l'athlète âgé.

Considérons la variable Pouls, dessinons le diagramme en moustache et commentons-le.

La figure commentée est représentée par la [figure xx](#).





Avant de passer à une deuxième illustration de la distribution d'une variable quantitative, voyons comment on la rend discrète.

Les classes ou conversion des données quantitatives en données qualitatives

Si la variable observée est continue, et si l'instrument de mesure est très précis, les tableaux de données sont caractérisés d'une part un grand nombre de lignes et d'autre part, de nombreux effectifs de faible amplitude. Ceci peut masquer les caractéristiques essentielles de la distribution. Le tableau est donc peu lisible. C'est le cas de la variable `toto$Lmact`

Les statisticiens conviennent qu'une bonne image et une information suffisante issues de ce tableau peuvent être obtenues par un groupement des données en classes dont on compterait les effectifs.

La construction d'une distribution de fréquence se fait en trois étapes :

- choisir les classes dans lesquelles les données seront groupées.
- trier les données dans les classes appropriées ;
- compter le nombre des individus qui se trouvent dans chaque classe.

Le gros problème est celui de choisir les classes. Quel doit être le nombre des classes ? Quelle doit être la largeur ou l'amplitude de la classe ? Ce choix est arbitraire. Il tient compte tout de même de l'objectif de l'étude et de la nature des données. Si l'on choisit J classes, chacune est définie par un entier j compris entre 1 et J . Chaque classe j appartenant à J est caractérisé par sa limite inférieure x_j^- , sa limite supérieure x_j^+ et son effectif n_j . La limite inférieure d'une classe est confondue avec la limite

supérieure de la classe précédente. Elle est située à mi-chemin entre la valeur la plus petite de la classe considérée et la valeur la plus grande de la classe précédente.

Le choix des classes est subjectif. Le choix du nombre des classes est guidé par le bon sens et par la pratique. Il dépend des objectifs du groupement.

Certains auteurs (Devore *et al.*, 1999) proposent de calculer le nombre des classes par la racine carrée du nombre d'observations. Quand on a 168 individus, alors le nombre des classes est 13.

D'autres (Dagnelie, xxx) par la règle de Sturges qui stipule que le nombre des classes dépend du nombre d'observations et doit être proche de la valeur k donnée par (7).

$$k = 1 + \frac{10}{3} \log_{10} n \quad (7)$$

Pour 168 observations, $k=8$. Donc il est raisonnable de prendre entre 8 et 13 classes.

Il existe aussi la méthode de Scott qui elle se fonde sur l'estimation de l'écart-type. Il existe aussi celle de Freedman-Diaconis qui elle utilise l'écart inter-quartile.

Dans le logiciel R, ces trois façons de calculer le nombre de classe sont dans la bibliothèque MASS. Les trois syntaxes correspondantes sont, pour un vecteur donné :

```
> nclass.Sturges(toto$Lmact)
[1] 9
> nclass.FD(toto$Lmact)
[1] 12
> nclass.scott(toto$Lmact)
[1] 9
```

Méthode	Nombre de classes
Sturges	9
Freedman-Diaconis	12
Scott	10

Si la variable est quantitative bricolons une formule en attendant mieux:

```
attach(toto)
x=hist(Lmact)
for (i in (1:(length(x$breaks)-1)))
{
print(c(c(x$breaks[i]),c("et < à"),c(x$breaks[i+1]),c("effectif="),c(x$counts[i])))
}
[1] "13"      "et < à"   "14"      "effectif=" "2"
[1] "14"      "et < à"   "15"      "effectif=" "1"
[1] "15"      "et < à"   "16"      "effectif=" "7"
[1] "16"      "et < à"   "17"      "effectif=" "19"
[1] "17"      "et < à"   "18"      "effectif=" "41"
[1] "18"      "et < à"   "19"      "effectif=" "35"
[1] "19"      "et < à"   "20"      "effectif=" "23"
[1] "20"      "et < à"   "21"      "effectif=" "18"
[1] "21"      "et < à"   "22"      "effectif=" "11"
[1] "22"      "et < à"   "23"      "effectif=" "8"
[1] "23"      "et < à"   "24"      "effectif=" "3"
```

Note : La répartition en classes de données quantitatives génère une variable qualitative ordoninale.

Si au lieu de considérer les classes, on considère la valeur du milieu de la classe, alors la variable générée est toujours quantitative.

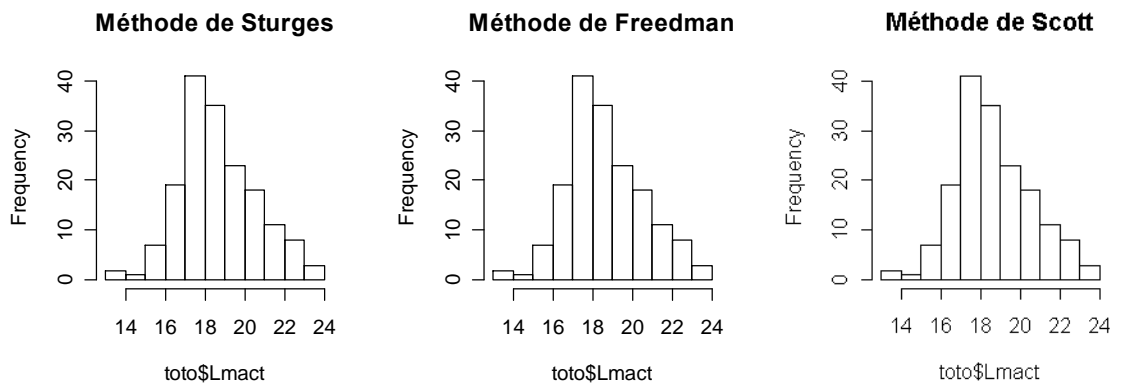
Une fois les données sont groupées de façon compacte, la distribution de fréquence peut être utilisée pour l'analyse, l'interprétation et le besoin de la communication. L'œil humain étant plus sensible aux graphiques, il est souhaitable à ce niveau de ne pas présenter seulement le tableau des fréquences mais plutôt des images de ces tableaux.

Histogramme pour une variable quantitative

L'histogramme est l'équivalent du diagramme en bâton pour les données quantitatives. Il diffère de celui-ci d'abord, par le fait que sur l'axe des abscisses on a des intervalles plutôt que des modalités. Ensuite, les barres sont proportionnelles aux effectifs.

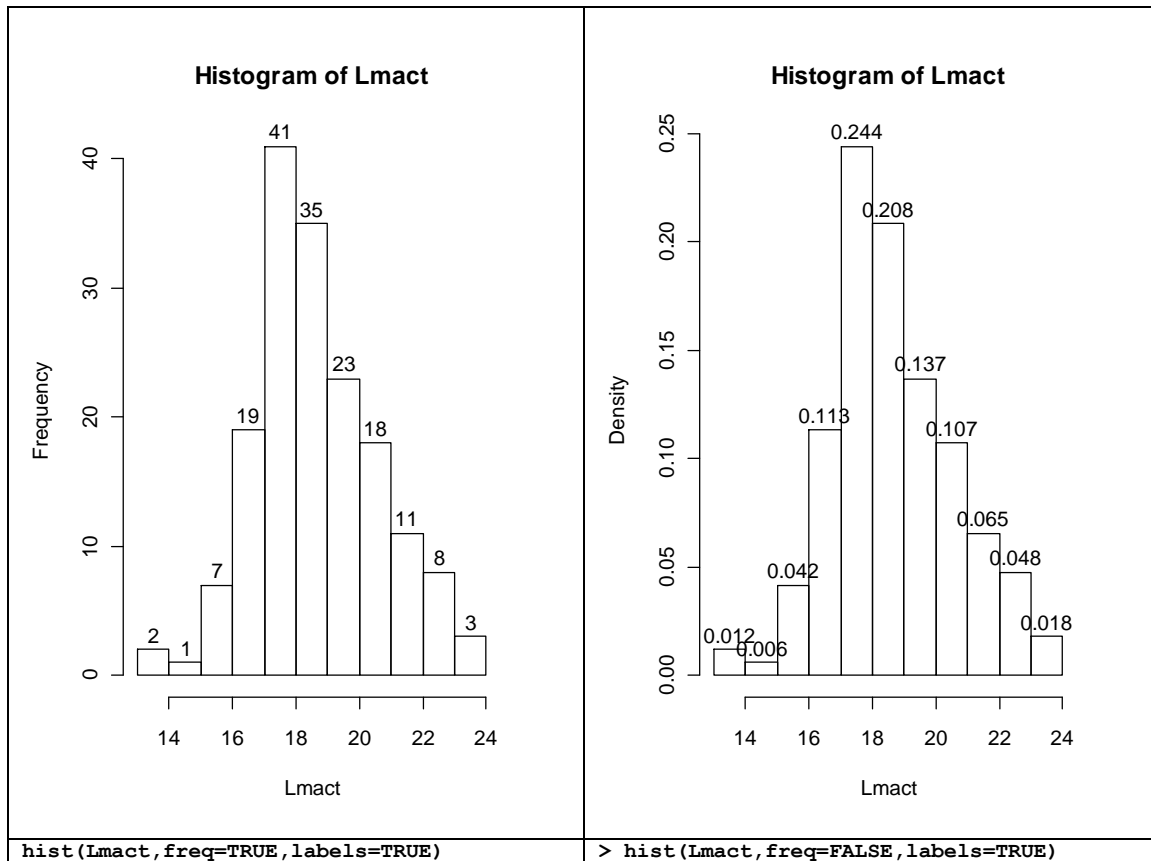
Si tu considères les données sur la longueur de main active des athlètes, tu obtiens les histogrammes suivants selon le mode de calcul des intervalles décrits plus haut.

Par défaut, la fonction hist de R utilise le nombre des classes donné par Sturges.



```
> hist(toto$Lmact,breaks = "sturges",main="Méthode de Sturges")
> hist(toto$Lmact,breaks = "fd",main="Méthode de Freedman")
> hist(toto$Lmact,breaks = "scott",main="Méthode de Scott")
```

L'histogramme peut se présenter de deux façons. La première est qu'en ordonnées on donne les fréquences et la deuxième est que l'on donne les fréquences relatives ou la densité.



Le logiciel R offre l'opportunité d'avoir le détail sur l'histogramme. Il suffit de nommer l'objet hist par un nom, ici x et d'appeler l'objet. Ceci donne les limites des classes (`$break`), les fréquences par classe (`$count`), les intensités ou les fréquences relatives (`$intensities`), les milieux des classes (`$mids`).

```
> x=hist(Lmact, freq=TRUE, labels=TRUE)
> x
$breaks
 [1] 13 14 15 16 17 18 19 20 21 22 23 24

$count
 [1] 2 1 7 19 41 35 23 18 11 8 3

$intensities
 [1] 0.011904760 0.005952381 0.041666667 0.113095238 0.244047619 0.208333333
 [7] 0.136904762 0.107142857 0.065476190 0.047619048 0.017857143

$density
 [1] 0.011904760 0.005952381 0.041666667 0.113095238 0.244047619 0.208333333
 [7] 0.136904762 0.107142857 0.065476190 0.047619048 0.017857143

$mids
 [1] 13.5 14.5 15.5 16.5 17.5 18.5 19.5 20.5 21.5 22.5 23.5

$xname
 [1] "Lmact"

$equidist
 [1] TRUE

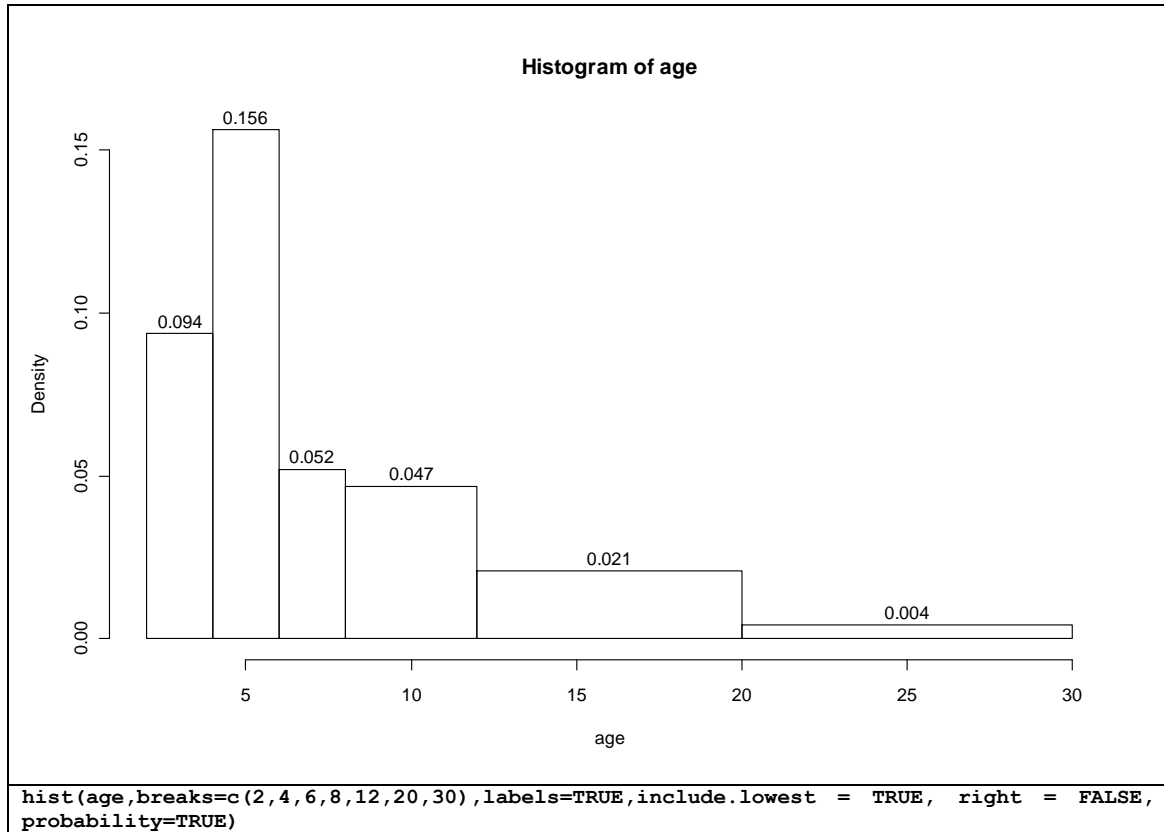
attr(,"class")
 [1] "histogram"
```

Exercice

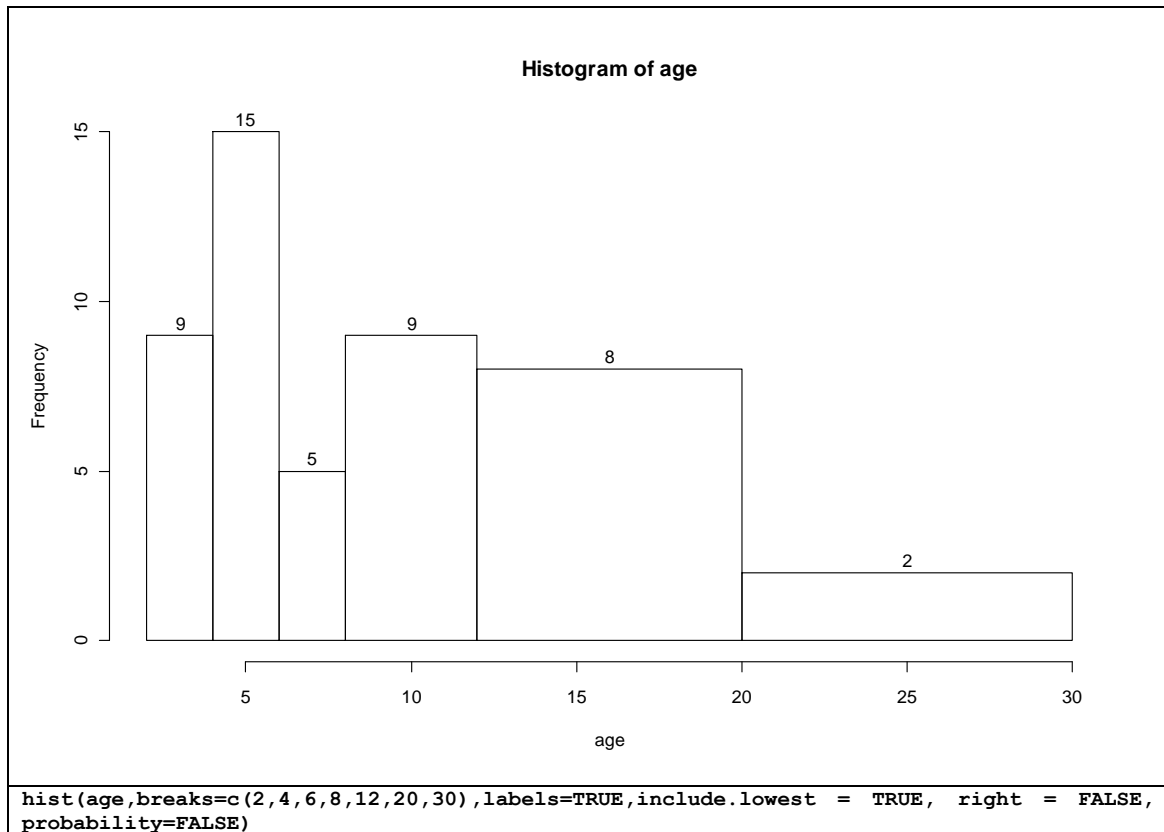
Soit une population d'élèves dont les âges sont données par le vecteur `age = c(11.5, 12.1, 9.9, 9.3, 7.8, 6.2, 6.6, 7.0, 13.4, 17.1, 9.3, 5.6, 5.7, 5.4, 5.2, 5.1, 4.9, 10.7, 15.2,`

8.5, 4.2, 4.0, 3.9, 3.8, 3.6, 3.4, 20.6, 25.5, 13.8, 12.6, 13.1, 8.9, 8.2, 10.7, 14.2, 7.6, 5.2, 5.5, 5.1, 5.0, 5.2, 4.8, 4.1, 3.8, 3.7, 3.6, 3.6, 3.6).

Dessiner l'histogramme avec les classes suivantes : $2 <= x < 4$, $4 <= x < 6$, $6 <= x < 8$, $8 <= x < 12$, $12 <= x < 20$, $20 <= x < 30$. En d'autres termes, la limite inférieure comprise et la limite supérieure non comprise



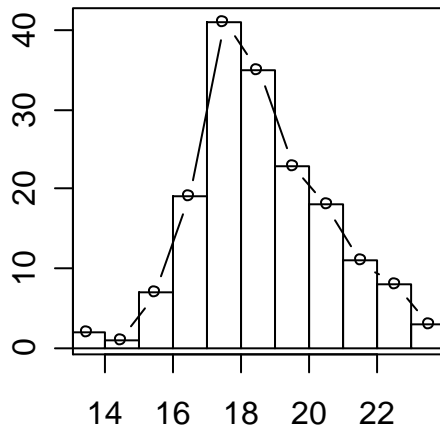
La fréquence relative est déduite par largeur de la classe multipliée par la densité. C'est la l'aire de la surface correspondant à la classe.



Polygone de fréquence

Un des graphiques populaires est aussi le polygone de fréquence. Il est souvent utilisé comme alternative d'un histogramme. Il faut pour cela utiliser la distribution des fréquences sur le milieu des classes.

Le polygone de fréquence utilise donc les mêmes données que l'histogramme. Ils sont dessinés à la même échelle. Sur la figure ci-dessous, l'histogramme est dessiné en arrière plan. Remarquer que le polygone de fréquence passe par les points milieux de chaque classe. Ces points sont tout simplement reliés par des segments



```
> x=hist(toto$Lmact,add=T)
> plot(x$mids,x$count,type="b")
> hist(toto$Lmact,add=T)
```

Diagramme tige et feuille (stem and leaf)

```
> stem(toto$Lmact)
The decimal point is at the |
13 | 0
14 | 0
15 | 04569
16 | 00023455579
17 | 00000000000123555555555555555566677899
18 | 0000000000000002223355555555555555556668888999
19 | 000001224445555555555678
20 | 00000011125555555558
21 | 00000345558
22 | 00000255558
23 | 00122
```

Le principe de ce diagramme est de produire une forme d'histogramme.

La tige est le nombre qui est à gauche d'une barre verticale. C'est le digit commun pour les nombres dont le reste des nombres est à droite de la barre verticale. La feuille est un nombre unique qui représente la donnée considérée. Pour cet exemple, il y a 168 feuilles qui constituent l'ensemble des données observées.

Pour les obtenir, on ordonne les nombres, on les divise par 1000 pour isoler la tige et on arrondi à un chiffre après la virgule ;

```
> tiges=round(sort(toto$Lmact/1),2)
> str(tiges)
Num [1:168] 13 14 15 15.4 15.5 15.6 15.9 16 16 16 ...
```

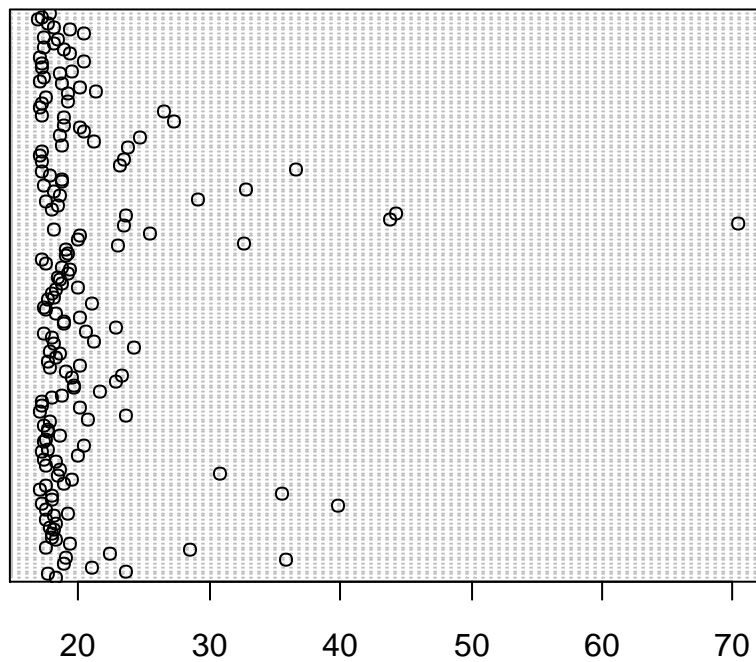
la feuille est le premier chiffre à gauche de la virgule et les feuilles sont les nombres qui suivent.

Diagramme de dispersion des points

(dotplot actuellement renommé dotchart dans R et stripchart)

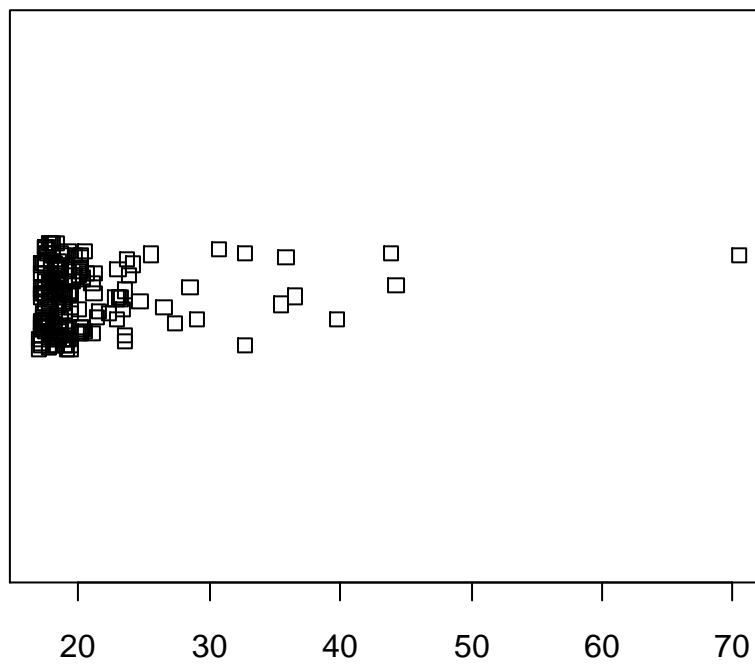
Le dot chart est une des analyses préliminaires que l'on effectue sur les données. Il permet de montrer comment les données sont concentrées autour de telle ou telle valeur. Ici rapidement on peut dire que les étudiants interrogés étaient en majorité jeunes de moins de 20 ans.

```
>dotchart (toto$Age)
```



C'est pareil avec stripchart. L'analyse vise à savoir où les données sont concentrées chez les jeunes. Cette première analyse peut s'avérer importante quand on compare deux séries de données. Une peut se concentrer autour d'une moyenne pendant que l'autre l'est autour d'une autre.

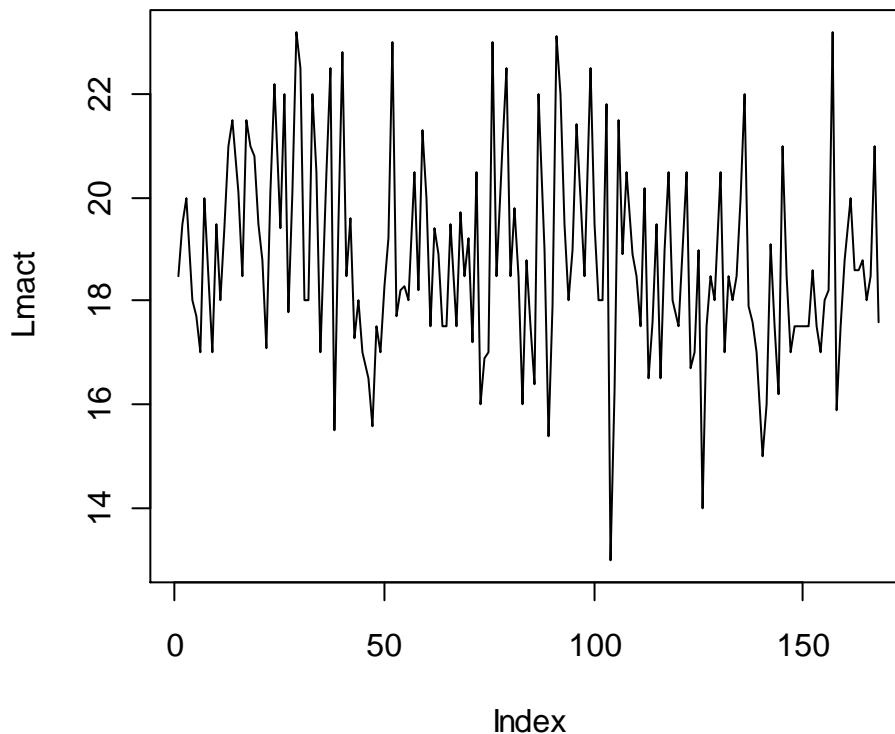
```
>stripchart (toto$Age, method="jitter")
```



Graphique de tendance

Un simple graphique en lignes reliant les différentes données représentées dans l'ordre de leur prise peut révéler des surprises. Elle peut dans le cas des séries temporelles révéler des tendances.

```
>plot(toto$Lmact, type="l")
```

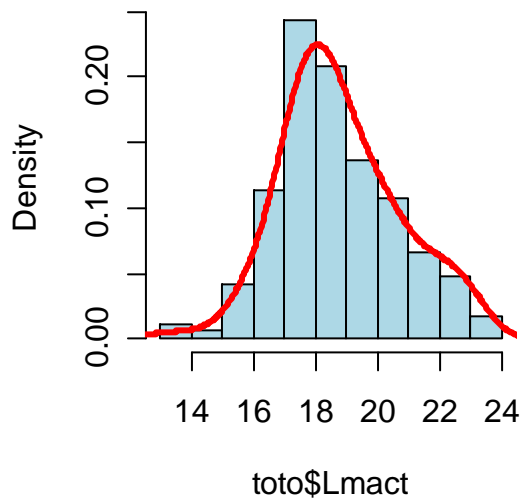


Histogramme et densité

On peut remplacer l'histogramme par le graphe d'une fonction. Si on voit les données comme une somme de masses de Dirac, il suffit de faire la convolution avec un "noyau" bien choisi, par exemple, la densité de la loi gaussienne. Pour affiner la courbe de densité, on joue avec la valeur de bw.

```
hist(toto$Lmact, probability=TRUE,
breaks="Sturges", col="light blue",main="Histogramme et densité")
points(density(toto$Lmact,bw=.7), type='l', col='red', lwd=3)
```

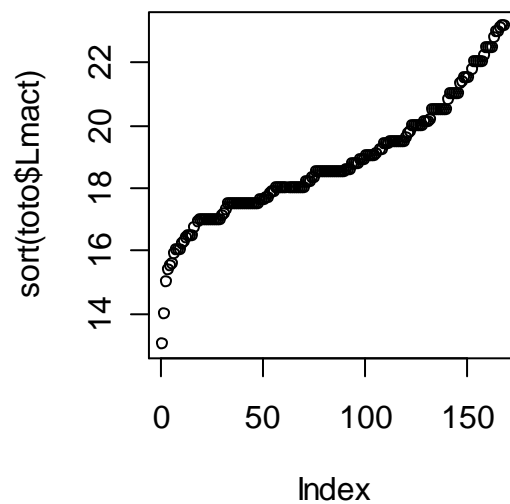
Histogramme et densité



Les courbes de densité présentent aussi le premier problème des histogrammes : un choix de noyau différent donne une courbe différente -- mais le second problème disparaît : celui du nombre des classes.

Plot sur données triées

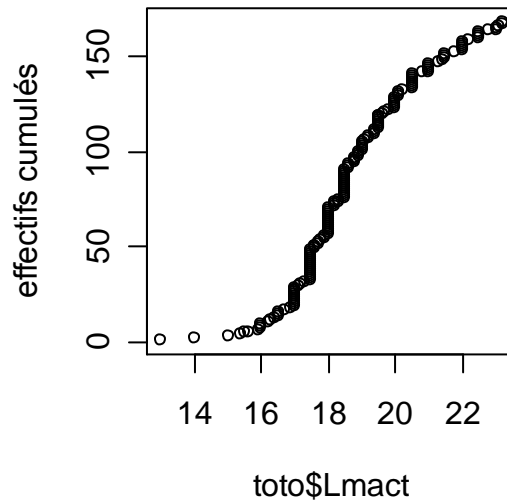
```
plot(sort(toto$Lmact))
```



Un troisième graphique consiste en la même représentation mais après les avoir triées, Les paliers horizontaux correspondent aux pics de l'histogramme.

Effectifs cumulés

```
effectifs.cumules <- function (x) {
  x.name <- deparse(substitute(x))
  n <- length(x)
  plot( 1:n ~ sort(x), xlab=x.name,
        ylab='effectifs cumulés')
}
effectifs.cumules(toto$Lmact)
```



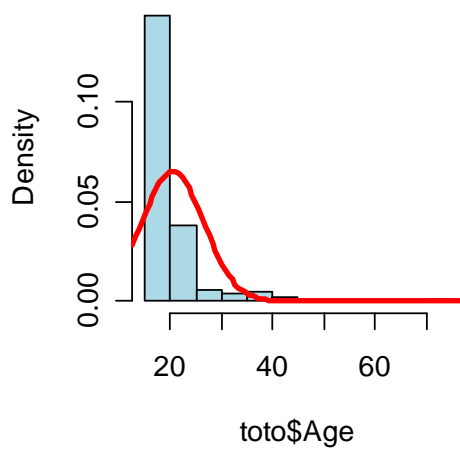
On peut aussi considérer le graphe des effectifs cumulés (c'est juste le symétrique du précédent).

Dans certains cas, les individus portent des noms : on peut les faire figurer sur le graphique (c'est le même type de graphique que précédemment, mais tourné de 90 degrés).

Graphique quantile-quantile

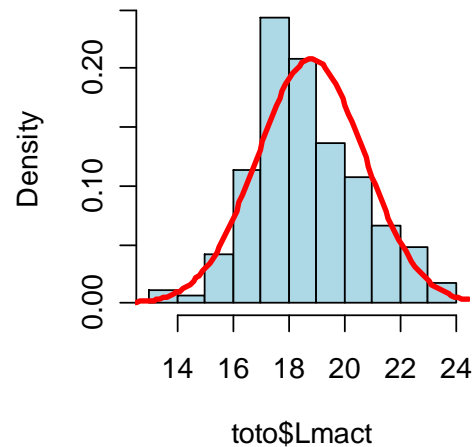
Nous venons de présenter une manière (graphique) de voir si une distribution est symétrique. Une autre chose que l'on aime bien, dans les données qu'on est amené à étudier, ce sont les variables normales.

Histogram of toto\$Age



```
hist(toto$Age, probability=TRUE, col="light
blue")
f <- function(t) {
dnorm(t, mean=mean(toto$Age),
sd=sd(toto$Age) )
}
curve(f, add=T, col="red",
lwd=3)
```

Histogram of toto\$Lmact

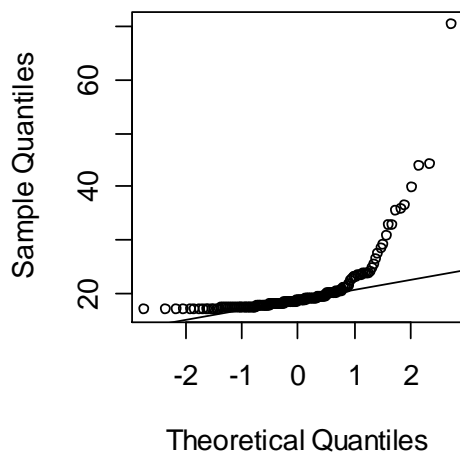


```
hist(toto$Lmact, probability=TRUE,
col="light blue")
f <- function(t) {
dnorm(t, mean=mean(toto$Lmact),
sd=sd(toto$Lmact) )
}
curve(f, add=T, col="red",
lwd=3)
```

Dans certains cas, il est flagrant que la variable n'est pas normale : c'est le cas de l'âge. Dans d'autres cas, c'est moins flagrant, par exemple pour la longueur de la main. Une première manière de s'en apercevoir consiste à comparer la densité de nos données avec une densité normale. On peut aussi voir graphiquement si une variable est normale : il suffit de représenter les quantiles d'une distribution normale en fonction des quantiles de l'échantillon.

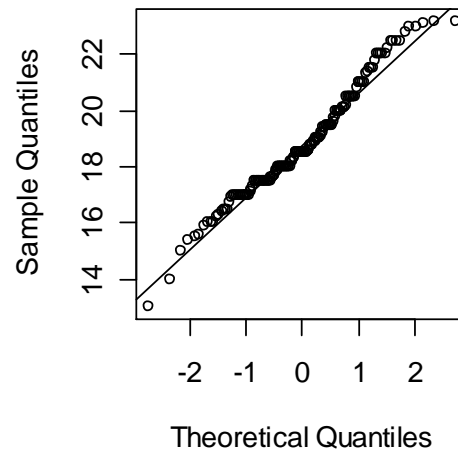
Il y a déjà une fonction qui fait cela. (La fonction `qqline` trace une droite passant par les premier et troisième quartiles.) Dans cet exemple, les données sur `Lmact` sont à peu près normales. En revanche le cas de l'âge est un exemple typique d'une donnée non normale.

Normal Q-Q Plot



```
qqnorm(toto$Age)
qqline(toto$Age)
```

Normal Q-Q Plot

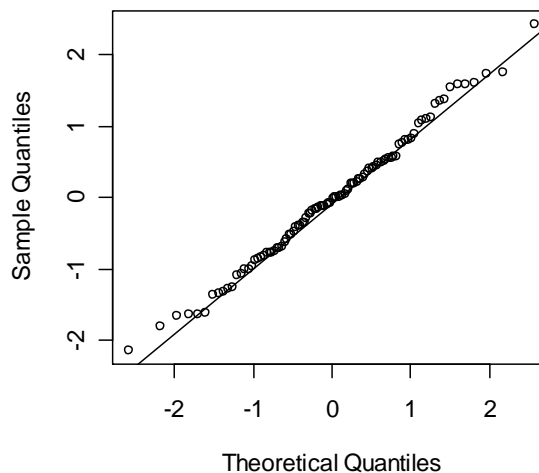


```
qqnorm(toto$Lmact)
qqline(toto$Lmact)
```

Si on modélise une donnée y normale alors on a le cas suivant:

```
y <- rnorm(100)
qqnorm(y)
qqline(y)
```

Normal Q-Q Plot



Exercices

Générer les lois diverses et tracer le qq plot

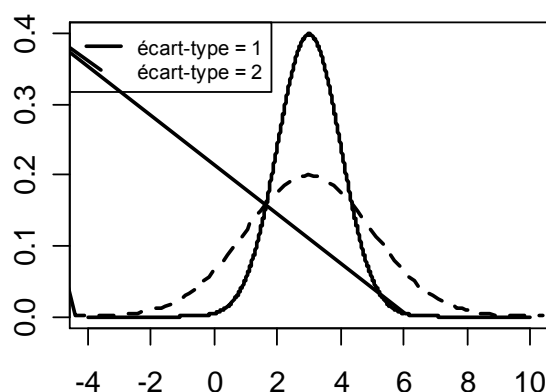
Mesure des tendances centrales d'une variable quantitative

Les livres classiques de la statistique donnent les paramètres qui mesurent les tendances centrales. On vient de voir comment calculer la moyenne arithmétique, les quartiles ou quantiles et la médiane par la commande summary. Il y en a d'autres comme le mode dit aussi valeur dominante, la moyenne amputée (trimmed mean), la moyenne

géométrique et la médiale. Pour en savoir plus il faut consulter les classiques de statistique.

Mesure de la dispersion et de la position relative d'une variable qualitative

Que signifie que deux ensembles de données ont même moyenne ? Est-ce dire que ces deux ensembles sont identiques ? Peut-être oui, peut-être non ! La **figure xx** montre que deux ensembles de données ont même moyenne mais ne sont pas identiques. Ce qui les différencie c'est la dispersion autour de la moyenne telle que visible sur l'axe des abscisses.



La mesure de la dispersion est une des mesures qui évaluent la variabilité dans une population ou la variabilité intra-groupe. On l'exprime souvent dans les mêmes unités que les données de départ.

La mesure de la dispersion est importante pour deux raisons :

- elle nous aide à porter un jugement sur la précision de la mesure. Si la dispersion est grande, alors la moyenne n'est pas représentative de l'ensemble des données.
- elle aide à décider à un certain moment de prendre une action pour contrôler la dispersion.

Trois mesures de dispersions sont souvent calculées. Je m'intéresse à celles qui sont appliquées aux données non groupées. Car les données groupées ont la fâcheuse faiblesse de perte d'informations et aussi elles ont été imaginées pour faciliter les calculs. Les mesures sont :

- l'écart-type et le coefficient de variation
- la déviation absolue moyenne
- L'écart moyen à la médiane
- L'écart interquartile
- L'amplitude ou Etendue

L'amplitude ou Etendue

L'étendue la plus simple mesure de la dispersion. C'est tout simplement la différence entre le maximum et le minimum de la série.

```
> etendue= range(toto$Iact)[2] - range(toto$Iact)[1]
> etendue
[1] 10.2
```

L'écart moyen absolu (MAD pour mean absolute deviation)

L'écart-moyen absolu ou écart moyen (MAD) est la moyenne arithmétique des valeurs absolues des écarts par rapport à la moyenne. Certains auteurs le calculent aussi par rapport à la médiane.

```
> mad(toto$Iact)
[1] 1.4826
```

L'Ecart-type et le Coefficient de variation

L'écart-type est la plus populaire des grandeurs qui évaluent la dispersion des données. C'est la racine carrée de la moyenne arithmétique des carrés des écarts par rapport à la moyenne (dite aussi variance).

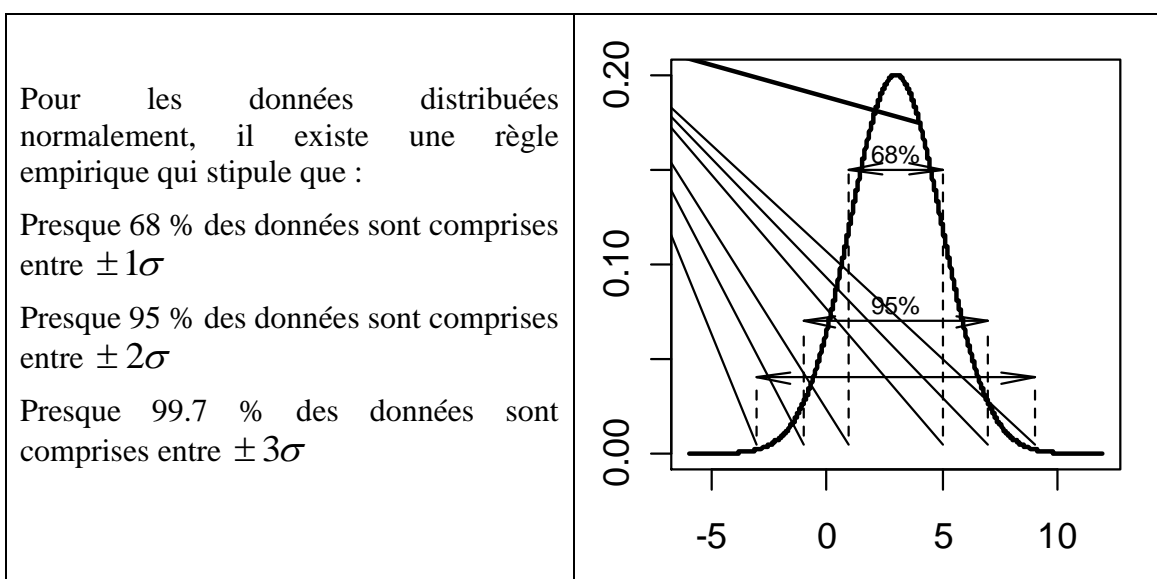
Avant de calculer l'écart-type, il faut toujours déterminer si les données sont celles issues de la population entière ou de l'échantillon de cette population.

Théorème de Chebyshev : la proportion des données d'une série qui se trouvent entre k écart-types par rapport à la moyenne ($k \geq 1$) est au moins $1 - \frac{1}{k^2}$.

Ce théorème permet d'interpréter l'écart-type comme ceci :

Pour $k=2$, $1 - \frac{1}{k^2} = 3/4$. Ce qui veut dire que pour n'importe quelle

distribution, au moins 75 % des données sont comprises entre deux fois l'écart-type.



En agronomie, on préfère utiliser la notion du coefficient de variation ou coefficient de variabilité V . C'est le rapport entre l'écart-type et la moyenne. Cette grandeur dépouillée de l'unité, permet de voir la variabilité relative entre deux parcelles différentes.

Exercice :

Soit une variable $x = 3 \ 4 \ 7 \ 12 \ 15 \ 18 \ 19$.

Calculer semi-manuellement avec le logiciel R, la variance

```
> x
[1] 3 4 7 12 15 18 19
> moy=mean(x)
> ecart=x-moy
> SCE=sum(ecart^2)
> SCE
[1] 258.8571
> length(x)
[1] 7
> var(x)
[1] 43.14286
> data.frame(x,moy,ecart,ecart^2)
```

x	Moy	Ecart	ecart.2
3	11.14286	-8.1428571	66.3061224
4	11.14286	-7.1428571	51.0204082
7	11.14286	-4.1428571	17.1632653
12	11.14286	0.8571429	0.7346939
15	11.14286	3.8571429	14.8775510
18	11.14286	6.8571429	47.0204082
19	11.14286	7.8571429	61.7346939
			SCE= 258.8571
		variance=SCE/(length(x)-1)	43.14286

```
> variance=SCE/(length(x)-1)
> variance
[1] 43.14286
>
```

Exercice :

correction de Sheppard dans le cas des distributions en cloche (Dagnelie).

```
> sd(x)
[1] 6.568322
> sqrt(variance)
[1] 6.568322
>
```

Calculer semi-manuellement le coefficient de variation

```
> sd(x)
[1] 6.568322
> cv=sd(x)/moy
> cv
[1] 0.5894648
```

Propriétés :

- Si les valeurs observées sont identiques alors la variance, l'écart-type et le coefficient de variation sont nuls.
- La variance et l'écart-type sont influencés par un changement d'unités. Ils sont indépendants du changement d'origine.

L'écart moyen à la médiane

L'écart médian ou l'écart probable ou écart équiprobable est la médiane des valeurs absolues des écarts calculées par rapport à la moyenne ou à la médiane.

```
> mad(toto$Age, center = median(toto$Age), constant = 1.4826, na.rm = FALSE, low = FALSE, high = FALSE)
[1] 1.605656
```

Note :

low: si 'TRUE', on calcule la médiane inférieure pour l'échantillon d'effectif impair. Il ne prend pas la moyenne mais la plus petite.

high: si 'TRUE', on calcule la médiane supérieure

La valeur constante par défaut est 1.4826. C'est égal approximativement égal à $1/\Phi^{(-1)(3/4)} = '1/qnorm(3/4)'$ et qui garantit la consistance c'est-à-dire l'espérance $E[\text{mad}(X_1, \dots, X_n)] = \sigma$ pour les X_i distribués $N(\mu, \sigma^2)$ et pour n grand

Ecart interquartile

C'est la différence entre le quartile d'ordre 3 et le quartile d'ordre 1. Cet intervalle englobe la moitié des observations qui se situent au centre des observations.

Dans R, il se calcule avec la commande IQR

```
> IQR(Lmact, na.rm = FALSE)
[1] 2.5
```

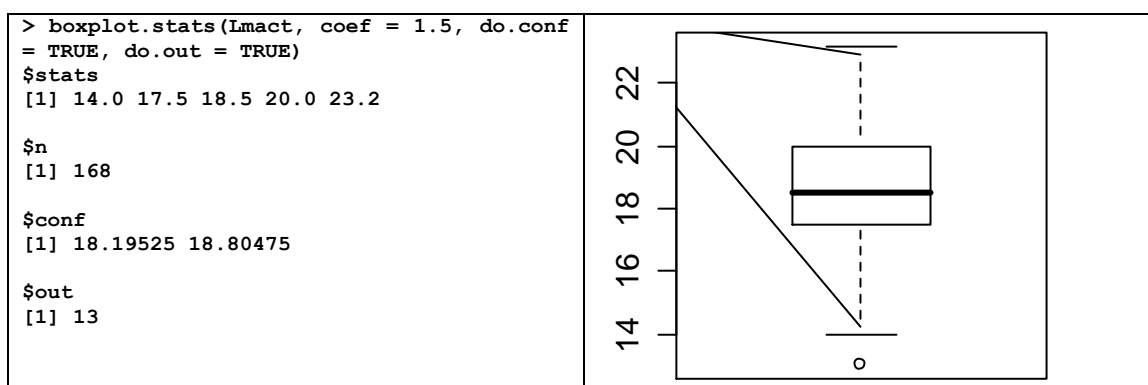
Cette fonction calcule les quartiles en utilisant la formule suivante recommandée par Tukey:

```
> IQR= quantile(Lmact,3/4) - quantile(Lmact,1/4)
> IQR
75%
2.5
```

Le semi-interquartile égal à la moitié de l'écart interquartile. On montre qu'il est égal à l'écart médian.

Note : IQR est simple à calculer mais moins robuste.

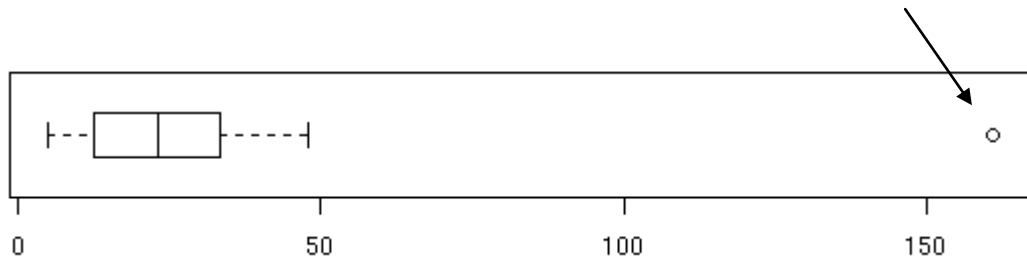
Exercice : Résumé-boxplot.stats



Cette commande permet de d'avoir le minimum, le maximum, le premier quartile, le deuxième et le troisième.

Si on ne précise pas l'option `<< range=0 >>`, le graphe va mettre en évidence les points atypiques, i.e., très éloignés ("outliers", en anglais), qui sont au delà de 1.5 fois l'intervalle interquartile.

Si on tombe sur un graphique qui se présente comme suit, on suspectera une donnée aberrante.



Dans un tel cas, une décision doit être prise. Si l'on est conscient qu'il s'agit d'une erreur alors on la corrige.

Pour nous résumer, nous venons d'apprendre :

- Une petite organisation des données permet de comprendre la biologie qui se cache derrière. Celle-ci passe, pour les variables qualitatives aussi bien pour les variables quantitatives, par la distribution des fréquences.
- Dans un article scientifique, la distribution de fréquences peut se donner sous forme de tableau ou sous forme de graphique. Nous avons vu le diagramme en barre ou le diagramme en secteur pour les variables qualitatives. Nous avons vu l'histogramme ou le graphique tige-feuille pour les variables quantitatives.
- L'histogramme et le diagramme tige-feuille permettent :
 - D'identifier la valeur typique ou représentative de la variable
 - La hauteur de la dispersion de la variable,
 - La présence des gaps entre les données,
 - La symétrie de la variable,
 - Le nombre et la localisation des pics ou des modes,
 - La présence des données suspectes.
- Pour enrichir l'interprétation des données, nous avons calculé les paramètres de position et les paramètres de dispersion.
- Dans la pratique, les premiers indices de la distribution des données peuvent être révélés par les paramètres calculés jusqu'ici. Nous avons vu qu'une moyenne proche de la médiane est un premier indice. La symétrie est un deuxième. Nous avons anticipé avec les graphiques qqplot et qqline pour voir le comportement d'une distribution normale.

Exercices

Exercice 2.1. Soit une série de données sur la température d'un corps en degré Fahrenheit.

TempDegreF=c(40,83,67,45,66,70,69,80,58,72,73,70,57,63,70,78,52,67,61,70,81,76,79,75,76,58,31).

Que dire a priori de ces données ?

```
> length(TempDegreF)
27
```

Il y a 27 températures enregistrées. En termes statistique on dit que la série a un effectif de 27.

Sans une quelconque organisation de ces données, il est difficile d'en sortir quoi que ce soit. Il suffit de les trier en ordre croissant :

```
> sort(TempDegreF)
31 40 45 52 57 58 58 61 63 66 67 67 69 70 70 70 70 72 73 75 76
76 78 79 80 81 83
```

On sait que la plus petite température enregistrée est 31 degrés. La plus grande est 83 degrés. Si on divise la série en deux parties égales, on constate que le point milieu est à la $27/2 = 23.5$ ème position. Ce point correspond à la température 70 degrés. De part et d'autres il y a 23 observations. On dira que 50 % des observations sont inférieures à 70. C'est la médiane. Si on divise la série en 4 parties, on constate que le premier quart correspond à la position $27/4 = 6.75$.

```
> summary(TempDegreF)
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 31.00  59.50  70.00  66.19  75.50  83.00
Tracer le dotplot
```

Exercice 2.2. On a enregistré les rendements de manioc sur certains sites. Les données sont :

RdtManioc=c(5.9,7.2,7.3,6.3,8.1,6.8,7,7.6,6.8,6.5,7,6.3,7.9,9,8.2,8.7,7.8,9.7,7.4,7.7,9.7,7.8,7.7,11.6,11.3,11.8,10.7)

Les données sont issues de combien de sites ? Que dire de ces données ?

Exercice 2.3. Longueur des feuilles

```
Longueur=c(2.97, 4.00, 5.20, 5.56, 5.94, 5.98, 6.35, 6.62, 6.72, 6.78,
6.80, 6.85, 6.94, 7.15, 7.16, 7.23, 7.39, 7.62, 7.62, 7.69,
7.73, 7.87, 7.93, 8.00, 8.26, 8.29, 8.37, 8.47, 8.56, 8.58,
8.61, 8.67, 8.69, 8.81, 9.07, 9.27, 9.37, 9.43, 9.52, 9.58,
9.60, 9.76, 9.82, 9.83, 9.83, 9.84, 9.96, 10.04, 10.21, 10.28,
10.28, 10.30, 10.35, 10.36, 10.40, 10.49, 10.50, 10.64, 10.95, 11.09,
11.12, 11.21, 11.29, 11.43, 11.62, 11.70, 11.70, 12.16, 12.19, 12.28,
12.31, 12.62, 12.69, 12.71, 12.91, 12.92, 13.11, 13.38, 13.42, 13.43,
13.47, 13.60, 13.96, 14.24, 14.35, 15.12, 15.24, 16.06, 16.90, 18.26)
```

Tracer histogramme avec

```
> x= hist(Longueur,breaks=seq(1,19,by=2))
> x
$breaks
 [1] 1 3 5 7 9 11 13 15 17 19

$counts
 [1] 1 1 11 21 25 17 9 4 1
```

```

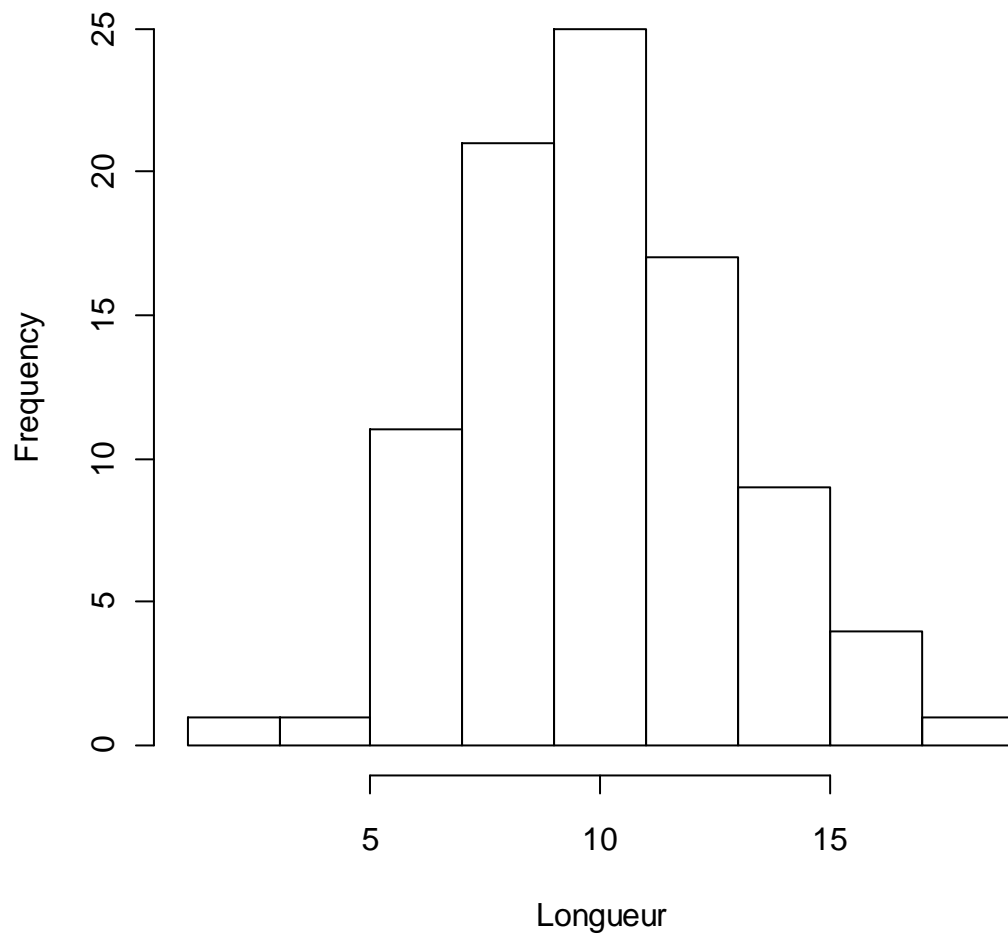
$intensities
[1] 0.005555554 0.005555556 0.061111111 0.116666667 0.138888889
[6] 0.094444444 0.050000000 0.022222222 0.005555556

$density
[1] 0.005555554 0.005555556 0.061111111 0.116666667 0.138888889
[6] 0.094444444 0.050000000 0.022222222 0.005555556

$mids
[1] 2 4 6 8 10 12 14 16 18

```

Histogram of Longueur



Description bivariée, bidimensionnelle ou description simultanée de deux variables qualitatives

Ce chapitre traite de la statistique descriptive à deux dimensions. Il met en évidence la forme, l'intensité et le sens de la relation qui existe entre deux séries d'observation considérées simultanément.

Distribution de fréquence : Tableau de Contingence

Considérons un tableau individus x caractère. Soit une étude qui porte simultanément sur deux variables notées x et y. Ces variables peuvent être toutes qualitatives ou toutes quantitatives ou même une qualitative et une autre quantitative.

Soient données deux variables qualitatives x de l modalités et y de k modalités. Les données peuvent se regrouper sous forme d'un tableau croisé où les n_{ij} représentent les fréquences des individus qui présentent à la fois les modalités x_i et y_j . Ce tableau est appelé tableau de contingence à l lignes et k colonnes (Tableau II.3.1.).

Tableau II.3.1. : tableau de contingence						
	y_1	y_2	...	y_j	...	y_k
x_1	n_{11}	n_{12}	...	n_{1j}	...	n_{1k}
x_2	n_{21}	n_{22}		n_{2j}	...	n_{2k}
.
.
.
x_i	n_{i1}	n_{i2}	...	n_{ij}	...	n_{ik}
.
.
.
x_s	n_{r1}	n_{r2}	...	n_{rj}	...	n_{rk}

Exemple :

Soient des observations portées sur l'âge et le nombre des fruits portés pendant une année par 12 arbres fruitiers (Tableau II.3.1.).

Tableau II.3.1. Nombre des fruits et âge d'un arbre fruitier

Individu	Nombre De Fruits x10	Age De l' arbre
1	8	1
2	11	2
3	5	4
4	5	3
5	9	3
6	9	3
7	9	3
8	11	3
9	10	4
10	6	5
11	8	5
12	9	6

Il y a à noter dans cet exemple que les deux variables sont continues discrètes. On peut rendre les deux variables qualitatives en considérant les classes d'âge [1,3[, [3,5[et [5,7[et les classes de nombre des fruits [4,8[, [8,10[et [10,12[.

Ces données peuvent donc se regrouper sous forme d'un tableau de contingence (Tableau II.3.2.).

Tableau II.3.2.								
		j=1	J=2	j=3				
		[1, 3[y ₁ =2	[3, 5[y ₂ =4	[5, 7[y ₃ =6	n _i .	n _i .x _i	n _i .x _i ²	$\sum_j n_{ij}x_i y_j$
i=1	[4, 8[x ₁ =6	0	2	1	3	18	108	84
i=2	[8, 10[x ₂ =9	1	3	2	6	54	486	234
i=3	[10, 12[x ₃ =11	1	2	0	3	33	363	110
	n _j	2	7	3	12	105	957	428
	n _j y _j	4	28	18	50			
	n _j y _j ²	8	112	108	228			
	$\sum_i n_{ij}x_i y_j$	40	244	144	428			

Lois Marginales

Effectifs marginaux

$$n_{i0} = \sum_j n_{ij}$$

$$n_{0j} = \sum_i n_{ij}$$

Fréquences marginales

$$f_{i0} = \frac{n_{i.}}{n}$$

$$f_{0j} = \frac{n_{.j}}{n}$$

Moyenne marginale

$$\bar{x} = \frac{1}{n_{00}} \sum_i n_{i0} x_i$$

$$\bar{x} = \frac{105}{12} = 8,75$$

$$\bar{y} = \frac{1}{n_{00}} \sum_j n_{0j} y_j$$

$$\bar{y} = \frac{50}{12} = 4,17$$

Variance marginale

$$V(x) = \frac{1}{n_{00}} \sum_i n_i x_i^2 - \bar{x}^2$$

$$V(x) = \frac{1}{12} 957 - (8,75)^2 = 3,19$$

$$V(y) = \frac{1}{n} \sum_j n_j y_j^2 - \bar{y}^2$$

$$V(y) = \frac{1}{12} 228 - (4,17)^2 = 1,61$$

Lois Conditionnelles

Effectifs conditionnels

Fréquences conditionnelles

$$f_{i|j} = \frac{n_{ij}}{n_j}$$

$$f_{j|i} = \frac{n_{ij}}{n_i}$$

Moyenne conditionnelles

La moyenne conditionnelle de x si y=yj La moyenne conditionnelle de y si x=xi

$$\bar{x}_j = \frac{1}{n_{0j}} \sum_i n_{ij} x_i$$

$$\bar{y}_i = \frac{1}{n_{i0}} \sum_j n_{ij} y_j$$

Variance conditionnelles

$$V(x)_j = \frac{1}{n_j} \sum_i n_{ij} \bar{x}_j - \bar{x}_j^2$$

$$V(y)_i = \frac{1}{n_i} \sum_j n_{ij} \bar{y}_i - \bar{y}_i^2$$

Relation entre fréquences marginale et conditionnelle

$$f_i \cdot f_{ji} = f_{ij}$$

$$f_j \cdot f_{ij} = f_{ij}$$

La démonstration est facile. Il suffit de remplacer les fréquences par leurs définitions.

Relation entre moyennes marginale et conditionnelle

$$\bar{x} = \frac{1}{n} \sum_i n_j \bar{x}_i$$

$$\bar{y} = \frac{1}{n} \sum_i n_i \bar{y}_j$$

Mesure de la contingence globale

La mesure de la contingence globale est donnée par le χ^2 de Pearson :

$$\sum_{i=1}^r \sum_{j=1}^s \left[\frac{n_i n_j}{n} - n_{ij} \right]^2 \cdot \frac{n \cdot}{n_i n_j}$$

Relation entre variances marginale et conditionnelle

$$V(x) = \frac{1}{n} \sum_j n_j V(x)_i + \frac{1}{n} \sum_j n_j (\bar{x}_j - \bar{x})^2$$

$$V(x) = \bar{V}(x)_j + V(\bar{x})_j$$

Le deuxième membre est la somme entre la moyenne des variances conditionnelles et la variance des moyennes conditionnelles

Moments et covariances

Le moment d'ordre k en x et d'ordre l en y par rapport à c pour x et d pour y est donné par :

$$m_{kl} = \frac{1}{n} \sum_i \sum_j n_{ij} (x_i - c)^k (y_j - d)^l$$

En particulier, pour les moments centrés ($c = \bar{x}, d = \bar{y}$)

$$m_{11} = \text{COV}(x, y)$$

$$m_{20} = s_x^2$$

$$m_{02} = s_y^2$$

Correlation entre deux variables qualitatives : Coefficient de Sperman

Dans une analyse sensorielle, lors d'une dégustation, on veut savoir si les juges sont en accord (homogènes) ou pas.

Soient 2 personnes A et B qui jugent 9 échantillons de pain à base du melange manioc-blé.

N° échantillon	1	2	3	4	5	6	7	8	9
A	3	5	1	9	8	7	2	4	6
B	3	9	1	5	2	8	4	7	6
d_i (différence absolue)	0	4	0	4	6	1	2	3	0

Si les juges sont homogènes, alors $d_i = 0$ pour tous les couple AB

On démontre que :

$$r = 1 - 6 \sum_i \frac{d_i^2}{n(n^2 - 1)}$$

C'est le coefficient de corrélation de Spearman.

- Si $r=1$ alors les juges sont en accord parfait
- Si $r=-1$ alors les juges sont en désaccord parfait
- Si $r=0$ alors il y a indépendance entre les traitements.

Plus r est proche de 1, plus les classements sont identiques.

Description simultannée de deux variables qualitatives : Analyse Factorielle des Correspondances (AFC)

Dans cette méthode, on part d'un tableau de nombres (les valeurs des variables ou une table de contingence), on regarde ses colonnes comme des points d'un espace de dimension n (idem pour les lignes), on cherche un sous-espace de dimension deux qui contient le maximum d'information. On projette le nuage des points sur ce sous-espace, au sens du produit scalaire canonique de R^n ou d'un autre produit scalaire).

L'analyse des correspondances s'intéresse aux tableaux de contingence, i.e., aux variables qualitatives. Plaçons-nous dans la cas de deux variables. On commence par transformer le tableau en deux tableaux : celui des profils-lignes (la somme des éléments de chaque ligne égale 1 (ou 100%)) et celui des profils-colonnes.

Si les deux variables sont indépendantes, on a

$f(i,j)=f(i,*)f(*,j)$. On peut utiliser le test du Chi2 de Pearson pour comparer les distributions de $f(i,j)$ et de $f(i,*)f(*,j)$.

D'un point de vue technique, l'analyse des correspondances ressemble beaucoup à l'analyse des composantes principales. Il s'agit de représenter graphiquement les points correspondant aux différentes lignes du tableau (resp, les colonnes), de manière à avoir un nuage de points les plus dispersé possibles (i.e., en maximisant la variance). On procède donc exactement comme pour l'ACP, à ceci près qu'on ne mesure pas les distances avec la métrique canonique, mais à l'aide de la « distance du chi2 »

Exemple

Nous allons considerer un exemple classique qui intéresse les biologistes généticiens. Il s'agit de voir la correspondance entre la couleur des yeux (Eye = Brown, Blue, Hazel et green) et la couleur des cheveux (Hair = Black, Brown, Red, Blond)

```
library(MASS)
data(HairEyeColor)
> y <- HairEyeColor
> y
, , Sex = Male
Eye
Hair   Brown Blue Hazel Green
Black  32   11  10    3
Brown  38   50  25   15
Red    10   10   7    7
Blond   3   30   5    8
, , Sex = Female
Eye
Hair   Brown Blue Hazel Green
Black  36    9   5    2
Brown  81   34  29   14
Red    16    7   7    7
Blond   4   64   5    8
> x <- HairEyeColor[,1]+HairEyeColor[,2]
> x
Eye
Hair   Brown Blue Hazel Green
Black  68   20  15    5
Brown 119   84  54   29
Red    26   17  14   14
Blond   7   94  10   16
>
> afc=corresp(x, nf = 2)
```

```

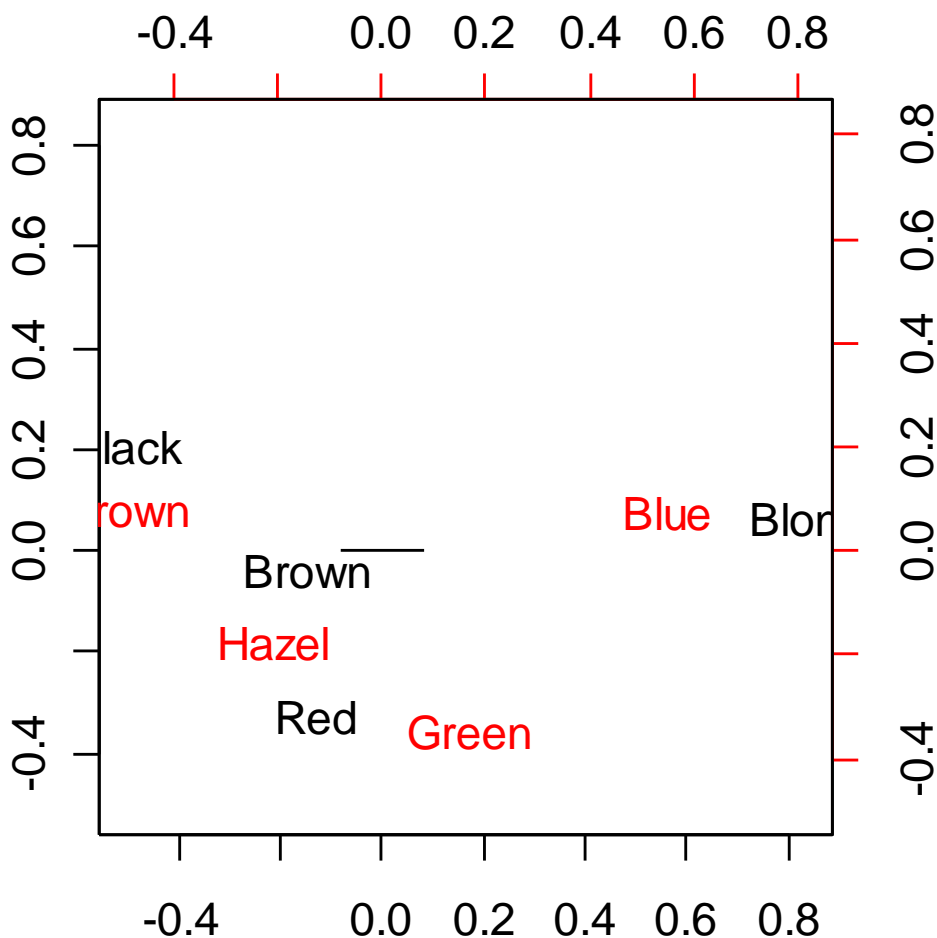
> afc
First canonical correlation(s) : 0.4569165 0.1490859
Hair scores:
[,1]      [,2]
Black -1.1042772  1.4409170
Brown -0.3244635 -0.2191109
Red    -0.2834725 -2.1440145
Blond  1.8282287  0.4667063
Eye scores:
[,1]      [,2]
Brown -1.0771283  0.5924202
Blue   1.1980612  0.5564193
Hazel  -0.4652862 -1.1227826
Green  0.3540108 -2.2741218
>

```

Elle accepte les matrices, les data.frames, les facteurs.

Nf est le nombre des facteurs

```
> biplot(corresp(x, nf = 2))
```



Conclusion : les personnes qui ont les cheveux blonds ont souvent les yeux bleus...

Description bivariée, bidimensionnelle ou description simultanée de deux variables quantitatives

Dans ce paragraphe, je vais considérer comment deux mesures sont en relation l'une et l'autre. La seule condition requise est que chaque couple de données soit prise sur un même individu.

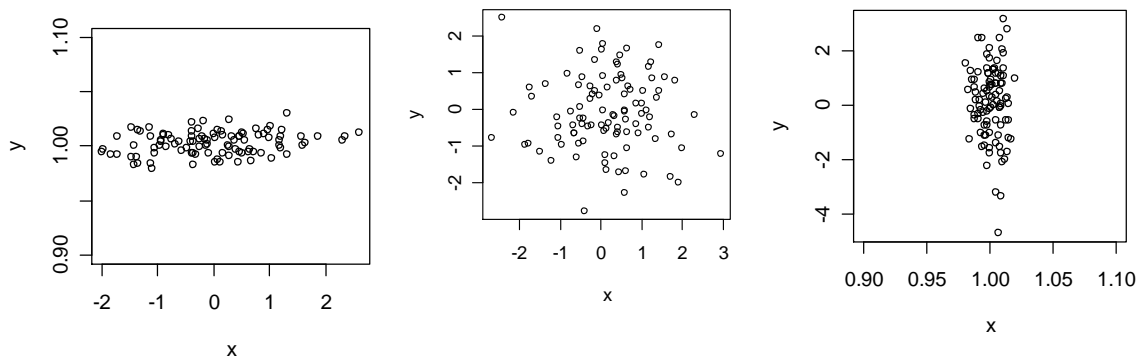
On distingue 3 types des données bivariées :

- Expérimentation stimulus/réponse : dans ce type d'expérimentation, le stimulus est sous le contrôle de l'expérimentateur.
- Expérimentation type série temporelle : Il arrive que l'on veuille étudier une grandeur dans le temps.
- Observation de la relation entre deux variables : dans une enquête qui utilise un questionnaire génère beaucoup de données sur une même personne. On peut étudier la relation entre les données de deux questions. On peut être tenté de voir la relation entre l'âge et le poids d'une personne. Est-ce que le poids diminue avec l'âge ?

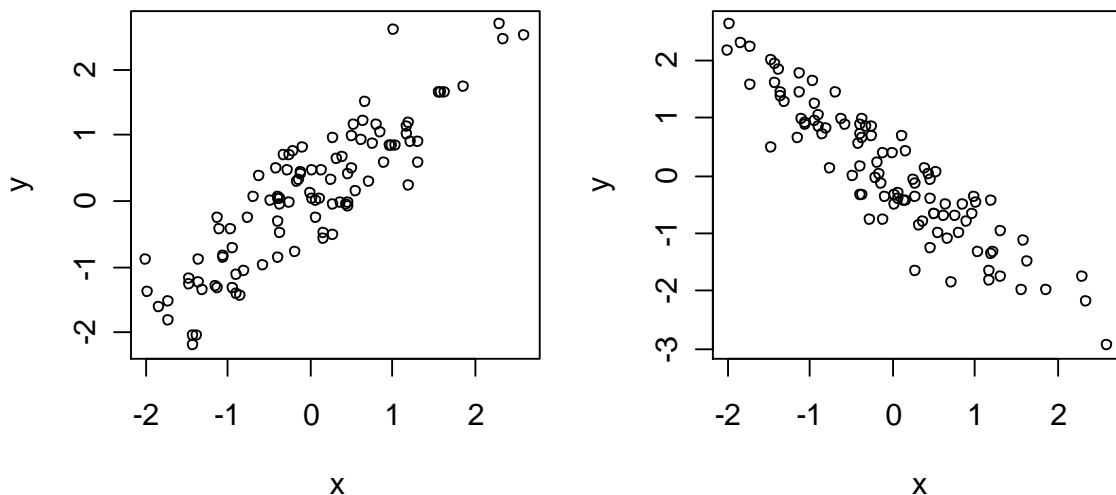
Deux variables quantitatives : Nuage des points

Le nuage des points est un graphique à deux dimensions qui peut montrer une tendance ou ne rien révéler du tout.

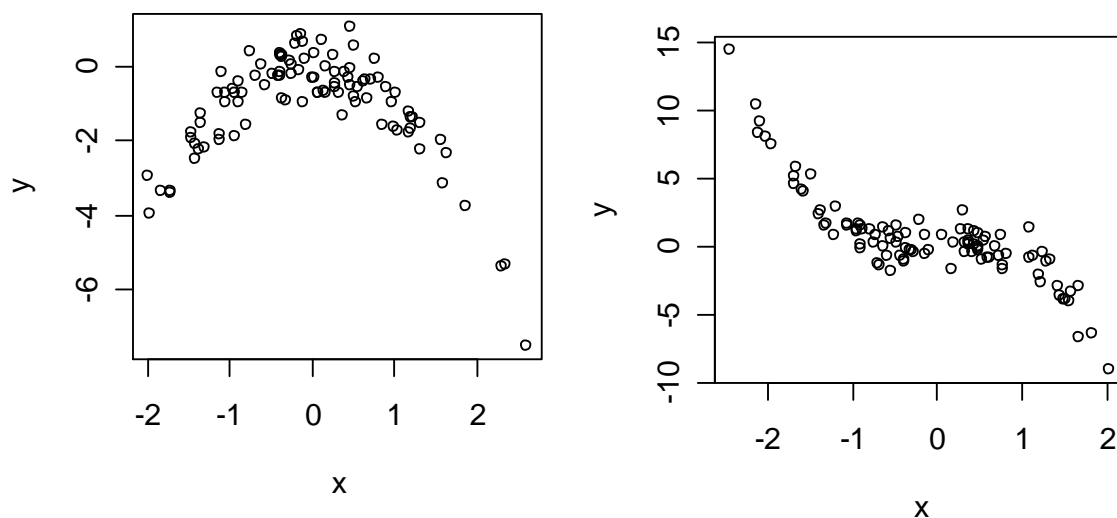
Nuage des points : pas de relation entre x et y



Nuage des points : tendance de croissance et de décroissance



Nuage des points : tendance non linéaire



Analyse de la relation linéaire

Considérons le cas où la représentation graphique accuse une relation linéaire. La corrélation va donner la mesure de combien est forte la relation entre les deux variables.

Corrélation vue comme mesure de l'association

Définition

Si la tendance est la croissance ou la décroissance, on dira que les deux variables sont corréées. La mesure standard pour mesurer cette relation est le coefficient de corrélation.

Le coefficient de corrélation r est donné par la relation suivante :

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}} = \frac{S_{xy}}{S_{xx} S_{yy}}$$

Interprétation du coefficient de corrélation

Le coefficient de corrélation est un nombre compris entre -1 et 1. Les valeurs exactes de 1 ou -1 veulent dire que les points observés tous sont sur une droite. La valeur nulle indique l'absence de relation entre les deux variables. Les valeurs comprises entre 0 et -1 indiquent une relation décroissante et les valeurs entre 0 et 1 indiquent une relation croissante.

Corrélation et causalité

La corrélation n'implique pas une relation de cause à effet. Si on dispose des données qui montrent une corrélation entre le fait de fumer et le cancer, une première possibilité

est que fumer entraîne le cancer. Il est aussi possible qu'en ayant le cancer, cela pousse à fumer.

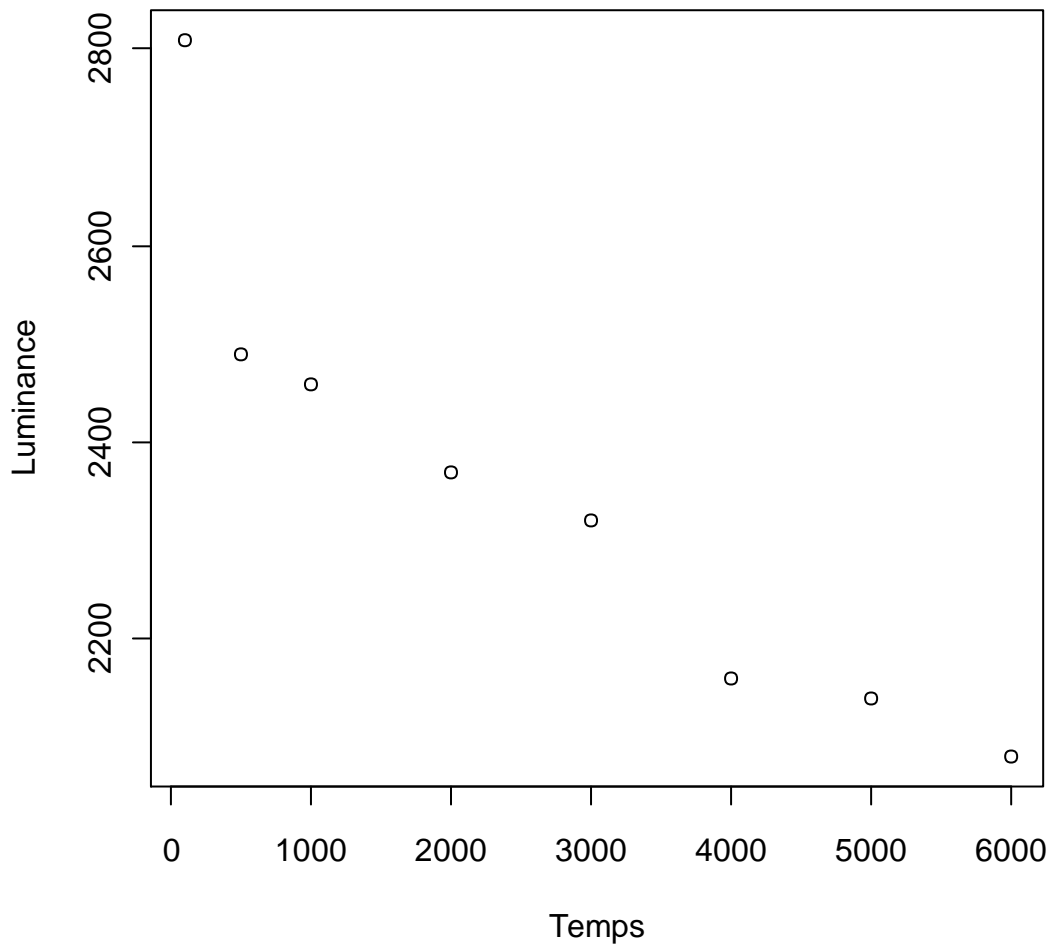
Exercice : Une lampe voit sa quantité de lumière (Luminance) diminuer avec le temps.

$Temps = c(100, 500, 1000, 2000, 3000, 4000, 5000, 6000)$

$Luminance = c(2810, 2490, 2460, 2370, 2320, 2160, 2140, 2080)$.

Etudier la relation.

```
> Temps= c(100, 500, 1000, 2000, 3000, 4000, 5000, 6000)
> Luminance= c(2810, 2490, 2460, 2370, 2320, 2160, 2140, 2080)
>
> Tempsbar=rep(mean(Temps), 8)
> Lumbar=rep(mean(Luminance), 8)
> data.frame(Temps, Luminance, Tempsbar, Lumbar)
  Temps Luminance Tempsbar Lumbar
1   100     2810     2700 2353.75
2   500     2490     2700 2353.75
3  1000     2460     2700 2353.75
4  2000     2370     2700 2353.75
5  3000     2320     2700 2353.75
6  4000     2160     2700 2353.75
7  5000     2140     2700 2353.75
8  6000     2080     2700 2353.75
>
> Syy=sqrt(sum((Luminance-Lumbar)^2))
> Sxx=sqrt(sum((Temps-Tempsbar)^2))
> r=Sxy/(Sxx*Syy)
> r
[1] -0.9215474
> cor(Temps, Luminance)
[1] -0.9215474
```



Une variable quantitative et une variable qualitative

Soit le rendement de manioc en Tonnes récoltés sur 3 sites A, B et C.

RdtManSite=c(6.1,8.4,7.6,7.5,4.4, 8.8,8.3,5.9,7.4,7.6,8,7.5,7.6.8,9.3)

Site=c(rep("A",5), rep("B",5),rep("C",5))

Résumé des données

```
> summary(toto)
  RdtManSite  Site
Min.   :4.400  A:5
1st Qu.:6.900  B:5
Median :7.500  C:5
Mean   :7.373
3rd Qu.:8.150
Max.   :9.300
```

