Compléments sur R
Analyse exploratoire des données

Master M1 Informatique - Université d'Orléans

Christel Vrain
Christel.Vrain@univ-orleans.fr

LIFO (Laboratoire d'Informatique Fondamentale d'Orléans)
Département Informatique - Faculté Sciences
Université d'Orléans

Références

- An Introduction to R. W. N. Venables, D. M. Smith and the R Development Core Team
- Enseignements de Statistique en Biologie. A.B. Dufour D. Chessel J.R. Lobry Contributeurs S. Mousset S. Dray. Maintenance système S. Penel
  Site web : http://pbil.univ-lyon1.fr/R/enseignement.html
- Premiers pas en statistique. Y. Dodge. Springer 2001.

Un peu de terminologie

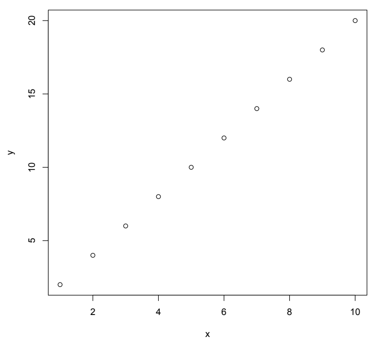
- Population : ensemble d'éléments sur lesquels porte l'analyse
- Individu (ou unité statistique) : élément de la population
- Variable (ou caractéristique ou caractère ou attribut) : sert à décrire la population
- Modalité d'une variable : valeur prise par cette variable
- Variable qualitative : un nombre fini de modalités, appelées catégories
  - échelle nominale : pas d'ordre sur les catégories
  - échelle ordinale : un ordre sur les catégories
- Variable quantitative : les modalités ont des valeurs numériques
  - variable discrète : ensemble des valeurs possibles dénombrable
  - variable continue : ensemble des valeurs possibles non dénombrable

La fonction plot

3 niveaux : création de graphique, ajouter d'informations, interaction

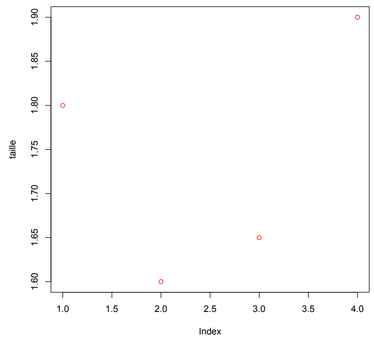
- *plot*(x, y) (x, y vecteurs) ou *plot*(u) (u liste de 2 vecteurs ou matrice à 2 colonnes)

```
> x<-1:10
> y<-seq(2,20,length=10)
> y
 [1]  2  4  6  8 10 12 14 16 18 20
>plot(x,y)
```



Graphique

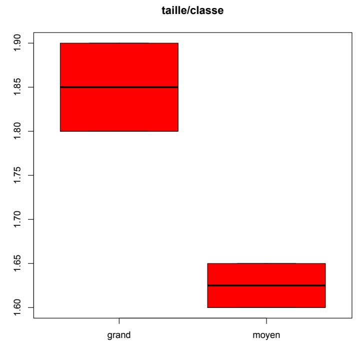
```
> taille
[1] 1.80 1.60 1.65 1.90
> plot(taille,col="red")
```



Graphique

- *plot*(f, y) si f est un facteur.

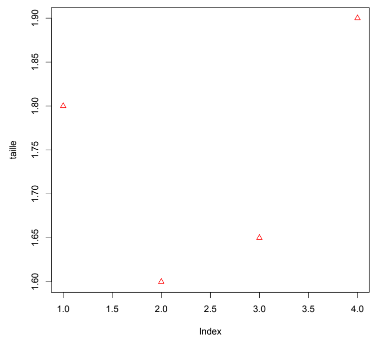
```
> taille
[1] 1.80 1.60 1.65 1.90
> et<-c("grand","moyen","moyen","grand")
> etf<-factor(et)
> etf
[1] grand moyen moyen grand
Levels: grand moyen
> plot(etf,taille,col="red",
main="taille/classe",
xlabel="taille en cm",ylabel="classe")
```



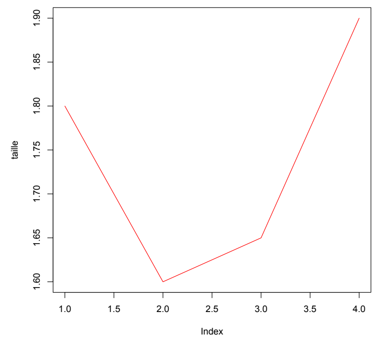
Fonction *plot* : options

- Ajout d'un titre avec le paramètre *main* : $plot(x, y, main = "calcul")$
- Choix des couleurs avec le paramètre *col* : $plot(x, y, col = "red")$
- Choix de la taille des points avec *cex* : $plot(x, y, cex = 2)$
- Choix de la forme de points avec *pch* : $plot(x, y, pch = 1)$ (1 : point, 2 : triangle, 3 : plus, 4 : croix, . . .)
- Possibilité de relier les points par des lignes avec type (Exemple : plot(x,y,type="l"))
- Limite des axes (valeur minimale et maximale) avec *xlim* et *ylim* : $plot(x, y, xlim = c(0, 2))$

Fonction *plot* : options



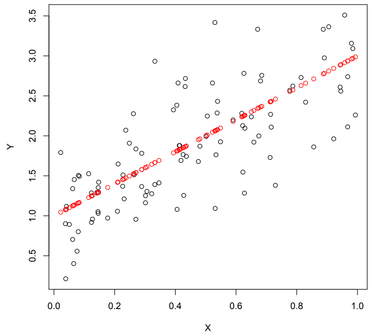

```
> plot(taille,pch=3,col="red")
```




```
> plot(taille,type="l",col="red")
```

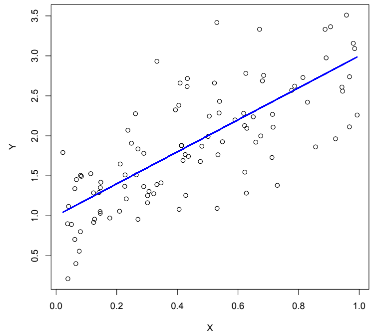
Ajout de points

```
> plot(X)
> bruit=rnorm(100,0,0.5)
# loi normale de moyenne 0
# et d'écart type 0.5
> hist(bruit,col="blue")
> Y<-1+2*X+bruit
> points(X,1+2*X,col="red")
> #points de coordonnes(x, 1+2x)
```



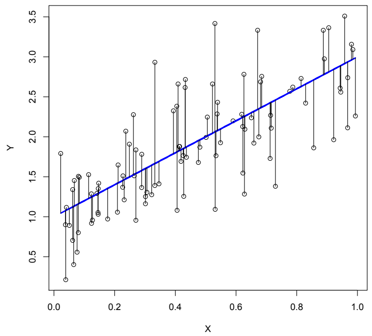
Ajout de lignes

```
> plot(X,Y)
> lines(X,1+2*X,lwd=2,col="blue")
```



Ajout de segments

```
> plot(X,Y)
> lines(X,1+2*X,lwd=2,col="blue")
> segments(X,Y,X,1+2*X)
> #segments reliant les
points (x,y) à (x,1+2x)
```



Variables qualitatives

- Variable dont le domaine a un nombre fini de valeurs (appelées *modalités*).
- Fréquence absolue pour k : n_k nombre d'occurrences de la modalité k
- Fréquence relative pour k : $\frac{n_k}{n}$ où n est le nombre d'individus.

```
> da<-read.table("weather.nominal.txt",
header=TRUE)
> da
    outlook temp   hum windy play
1     sunny  hot  high FALSE   no
2     sunny  hot  high  TRUE   no
3  overcast  hot  high FALSE  yes
4     rainy mild  high FALSE  yes
...
```

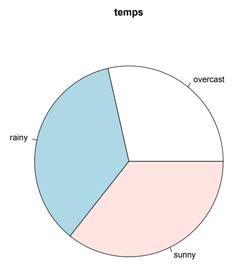
```
> summary(da)
    outlook     temp       hum
 overcast:4   cool:4   high  :7
 rainy   :5   hot :4   normal:7
 sunny   :5   mild:6

   windy          play
 Mode :logical   no :5
 FALSE:8         yes:9
 TRUE :6
 NA's :0
```

Variables qualitatives

- Représentation en secteurs - Diagramme circulaire (Pie chart)

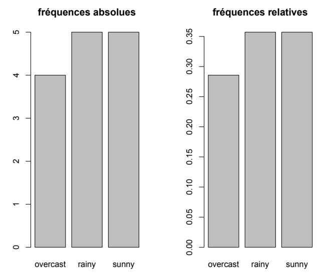
```
> look<-da$outlook
> is.factor(look)
[1] TRUE
> summary(look)
overcast     rainy     sunny
       4         5         5
> pie(summary(look))
```



Variables qualitatives

- Diagramme en bâtons, Barplot

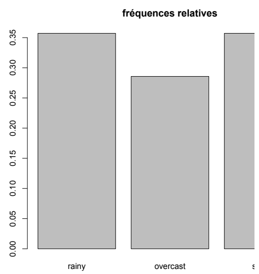
```
> freRel<-summary(look)/length(look)
> freRel
 overcast      rainy      sunny
0.2857143 0.3571429 0.3571429
> par(mfrow=c(1,2))
> barplot(summary(look),
main="fréquences absolues")
> barplot(summary(look)/length(look)
main="fréquences relatives")
```



Variables qualitatives

- Si la variable est ordonnée, faire plutôt un diagramme en bâtons en respectant l'ordre des modalités.

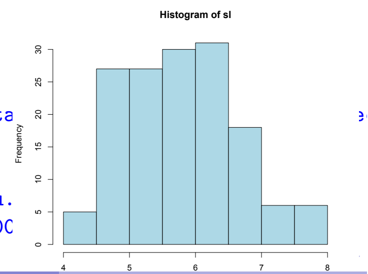
```
> look
 [1] sunny    sunny    overcast rainy
          rainy    rainy    overcast sunny
 [9] ...
Levels: overcast rainy sunny
> summary(look)
overcast     rainy     sunny
       4         5         5
> look2<-factor(look,
      levels=c("rainy", "overcast", "sunny"))
> look2
 [1] sunny    sunny    overcast rainy
       rainy    rainy    overcast sunny
 [9] ...
Levels: rainy overcast sunny
```



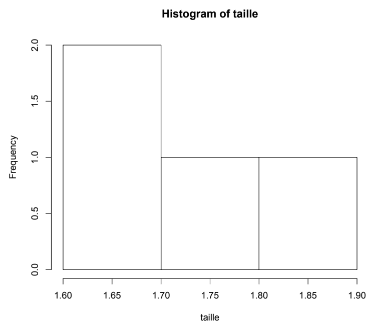
Variables quantitatives : histogramme

- Variable découpée en intervalles d'amplitude constante
- Axe horizontal : intervalles
- Axe vertical : nombre d'individus dans chaque intervalle
- Option *breaks* : valeurs du découpage
- Option *right* = *TRUE* : intervale ouvert à gauche et fermé à droite

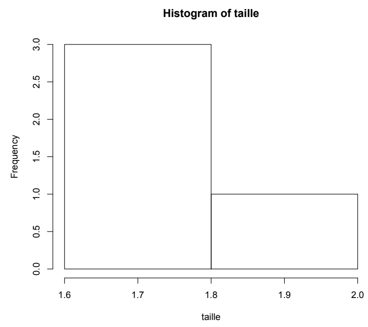
```
> names(iris)
[1] "Sepal.Length" "Sepal.Width"  "Peta
> sl<-iris$Sepal.Length
> summary(sl)
   Min. 1st Qu.  Median    Mean 3rd Qu.
  4.300   5.100   5.800   5.843   6.40
```



Autres représentations




```
> hist(taille)
```

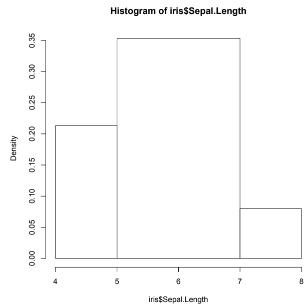



```
> hist(taille,nclass=2)
```


Autres représentations

```
> summary(iris$Sepal.Length)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  4.300   5.100   5.800   5.843   6.400   7.900
```


```
>  hist(iris$Sepal.Length)
```




```
> hist(iris$Sepal.Length,
         breaks=c(4,5,7,8))
```

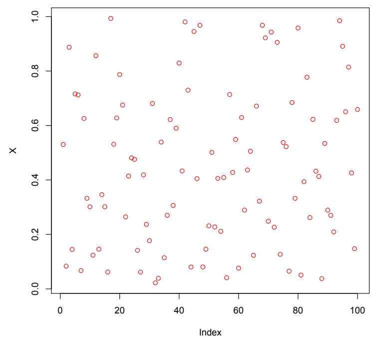

Difficulté de choix des découpages

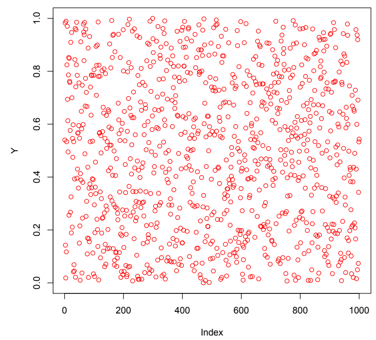
Attention, le choix des découpages influe beaucoup et peut donner une fausse perception.
Utilisation d'estimateurs locaux de la densité des points

Autre exemple

```
> X=runif(100)
> plot(X,col="red")
```

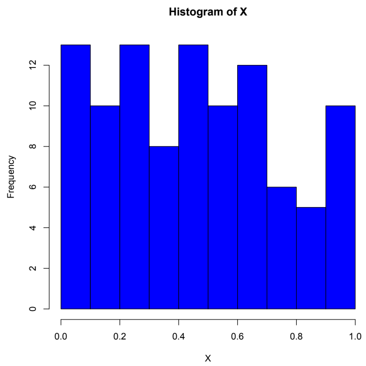
```
> Y=runif(1000)
> plot(Y,col="red")
```



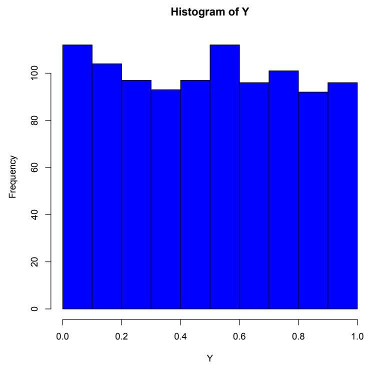


Autre exemple

```
> hist(X,col="blue")
```




```
> hist(Y,col="blue")
```



Etude des données

Regarder les données avec un oeil critique

- Mesures de tendances centrales ; moyenne, médiane, mode
- Valeurs extrémales : normales ou signes d'erreur
- Mesures de dispersion

Mesures de tendance centrale : moyenne

On dispose de k observations de la variable X : $x_1, \ldots, x_k$

- Moyenne

$$\overline{x} = \frac{x_1 + \ldots + x_k}{k} = \frac{\sum_{i=1}^{k} x_i}{k} = \left(\frac{1}{k}\right).x_1 + \ldots + \left(\frac{1}{k}\right).x_k$$

- Moyenne pondérée : x_i apparaît n_i fois ($n = n_1 + \ldots + n_k$)

$$\overline{x} = \frac{n_1.x_1 + \ldots + n_k.x_k}{n_1 + \ldots + n_k} = \frac{\sum_{i=1}^{k} n_i.x_i}{\sum_{i=1}^{k} n_i} = \left(\frac{n_1}{n}\right).x_1 + \ldots + \left(\frac{n_k}{n}\right).x_k$$

- Moyenne d'une distribution de fréquence

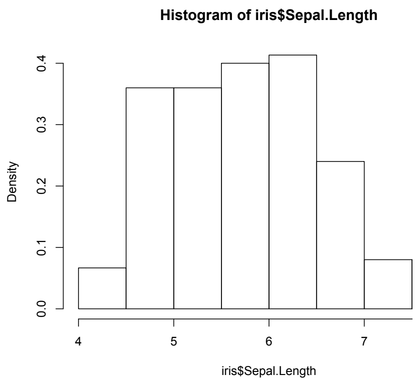
$$\overline{x} = \sum_{i=1}^{k} f_i.x_i$$

où f_i est la fréquence de x_i ($f_i = \frac{n_i}{n_1 + \ldots + n_k}$)

Mesures de tendance centrale : moyenne

```
> hist(iris$Sepal.Length)
> summary(iris$Sepal.Length)
  Min. 1st Qu.  Median
 4.300   5.100   5.800

 Mean 3rd Qu.    Max.
 5.843   6.400   7.900
```



Propriétés de la moyenne

- Toutes les observations ont le même poids : les valeurs extrêmes influencent autant que les autres
- La somme des écarts à la moyenne est nulle :

$$\sum_{i=1}^{k}(x_i - \overline{x}) = 0$$

- La somme des carrés des distances à la moyenne est plus faible que la somme des carrés des distances à toute autre valeur.
  $\rightarrow$ La moyenne est la mesure qui minimise la somme des carrés des écarts à elle-même.

Autres moyennes

- Moyenne géométrique

$$G = \sqrt[k]{x_1 \times \ldots \times x_k} \quad ou \quad log(G) = \frac{1}{k}(log(x_1) + \ldots + log(x_k))$$

- Moyenne géométrique pondérée (avec $n = n_1 + \ldots + n_k$)

$$G = \sqrt[n]{x_1^{n_1} \times \ldots \times x_k^{n_k}} \quad ou \quad log(G) = \frac{1}{n}\sum_{i=1}^{k}(n_i.log(x_i))$$

- Moyenne harmonique

$$H = \frac{k}{\frac{1}{x_1} + \ldots + \frac{1}{x_k}}$$

- Moyenne quadratique

$$Q = \sqrt[2]{\frac{1}{k}(x_1^2 + \ldots + x_k^2)}$$

Autres moyennes

```
> summary(iris$Sepal.Length)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.

  4.300   5.100   5.800   5.843   6.400   7.900
```

- Moyenne géométrique

```
> prod(SL)^(1/length(SL))
[1] 5.78572
```

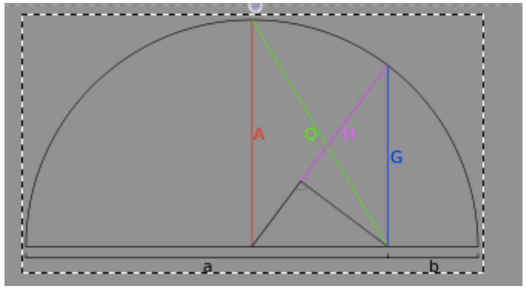
- Moyenne harmonique

```
> length(SL)/sum(1/SL)
[1] 5.728905
```

- Moyenne quadratique

```
> sqrt(sum(SL^2)/length(SL))
[1] 5.901328
```


Propriétés des moyennes pour 2 valeurs a et b



source : http://fr.wikipedia.org/wiki/Moyenne_harmonique

$Q^2 = (a^2 + b^2)/2 = ((a+b)/2)^2 + ((a+b)/2 - b)^2$

$G^2 = ab = ((a+b)/2)^2 - ((a-b)/2)^2$

Soit α : angle antre axe des abscisses et H

H tel que défini dans le diagramme est la moyenne harmonique, car $sin(\alpha) = G/A = H/G$ d'où $H = G^2/A$ $\quad (\frac{2}{1/a+1/b} = \frac{ab}{(a+b)/2})$

Autres moyennes

```
> da
  x nx
1 2  5
2 3  2
3 4  1
```

- Moyenne pondérée

```
> sum(x*nx)/sum(nx)
[1] 2.5
```

- Moyenne quadratique

```
> sqrt(sum(SL^2)/length(SL))
[1] 5.901328
```


Comparaison

- Toutes ces moyennes se généralisent en : le nombre M tel que

$$f(M) = \frac{1}{n}(n_1.f(x_1) + \ldots + n_k.f(x_k))$$

  où f est une fonction croissante ou décroissante de la variable x.
- Plus grande importance attribuée aux valeurs élevées pour les moyennes arithmétiques et quadratiques (encore plus la moyenne quadratique)
- Réduction de l'influence des valeurs les plus grandes pour les moyennes géométriques et harmoniques et augmentation de l'influence des valeurs les plus petites (encore plus vrai pour la moyenne harmonique)
- $$H < G < \overline{x} < Q$$

Médiane

- Médiane : point qui partage la distribution d'une série d'observations en deux parts égales
  $\Rightarrow$ partage l'histogramme (densité) en deux parties de même surface.
- Ne s'applique que pour des observations pouvant être ordonnées.
- Minimise la somme des distances absolues de toutes les observations à ce point

Estimation

1. Classement des observations en ordre croissant
2. Calcul
   - Si le nombre d'observations est impair, la médiane est la $(k+1)/2$-ème observation.
   - Si le nombre d'observations est pair, toute observation entre la $k/2$-ème observation et la $(k/2+1)$-ème observation.
   - $\rightarrow$ la moyenne de la $k/2$-ème observation et de la $(k/2+1)$-ème observation.
   - $\rightarrow$ Si les données sont groupées par classe, une valeur plus précise sous l'hypothèse d'une répartition uniforme des observations

$$med = L + \frac{n/2 - u}{v}.c$$

     - L : limite inférieure de la classe médiane
     - n :nombre total d'observations
     - u: somme des fréquences absolues des classes se situant avant la classe médiane
     - v : fréquence de la classe médiane
     - c : largeur de la classe médiane

Mode pour les variables qualitatives

Mode : valeur qui possède la fréquence la plus élevée

- N'est pas toujours une valeur centrale de la distribution
- une distribution peut avoir plusieurs modes: distribution unimodale, bimodale, plurimodale
- instable lorsque le nombre de données est faible, sensible à la taille et au nombre d'intervalles choisis pour regrouper les données

Comparaison

- moyenne : tient compte des valeurs de la distribution
- mode : indique une seule valeur de la distribution
- médiane : indique un rang
- si la distribution est unimodale et symétrique : moyenne, mode, médiane sont confondus
- si la distribution est bimodale et symétrique, moyenne et médiane sont confondues
- si la distribution est asymétrique, ils peuvent avoir des valeurs différentes
  - distribution étirée à droite : en général, mode puis médiane puis moyenne
  - distribution étirée à gauche : moyenne, médiane, mode

Variance

- Variable quantitative avec k observations, x_1, ldots, x_k

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

- Variable quantitative discrète à k valeurs x_1, ldots, x_k de fréquence n_1, ldots, n_k

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

  où $n = n_1 + \ldots + n_k$
- Ecart-type : $\sigma = \sqrt{\sigma^2}$
- Si les données sont groupées en intervalles et les observations individuelles non connues, estimation de la variance (par exemple en prenant comme valeur le point central de l'intervalle)

Propriétés

- La variance est toujours positive.
- Si elle est nulle, toutes les observations sont identiques.
- Si on ajoute la même constante à chaque observation, la variance ne change pas.
- Si on multiplie les observations par une même constante positive ou négative, on modifie la variance à un facteur multiplicatif (le carré de la constante).
  $\sigma^2(aX + b) = a^2\sigma^2(X)$ et $\sigma(aX + b) = |a|\sigma(X)$.
- Expression de la variance d'une population à partir de la variance et de la moyenne de deux sous-ensembles (σ^2 en fonction de $\overline{x}(1)$, $\sigma^2(1)$, $\overline{x}(2)$, $\sigma^2(2)$?)

Autres mesures

- Empan: différence entre la valeur la plus élevée et la moins élevée
- Ecart moyen

$$E.M. = \frac{1}{k}\sum_{i=1}^{k}|x_i - \overline{x}|$$

- Ecart médian mad

$$E.Med = \frac{1}{k}\sum_{i=1}^{k}|x_i - med|$$

- Ecart géométrique

$$log(E.geom) = \frac{1}{k}\sum_{i=1}^{k}(log(x_i) - log(G))$$

- Intervalle interquartile

Intervalle interquartile

Intervalle comprenant 50% des observations le plus au centre de la distribution.
Se définit à partir des quantiles qui sont des positions particulières (par exemple la médiane)

- quartiles : divisent l'ensemble d'observations en 4 parties égales
- déciles : idem mais en 10 parties égales
- centiles : item mais en 100 parties égales

Variables quantitatives : résumé à 5 valeurs et box plot

- résumé à 5 valeurs

  Médiane
  Premier quartile Troisième quartile
  Extrême inférieur Extrême supérieur

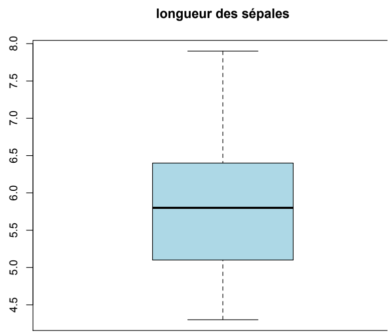
  où si n est le nombre d'observations et si les données sont rangées par ordre croissante
  - rang médiane $= (n+1)/2$
  - rang quartile $= (\lfloor rang\ médiane \rfloor + 1)/2$

  La médiane et les quartiles sont les données correspondant aux rangs calculés (la moyenne entre les valeurs les plus proches si le rang n'est pas entier)
- représentation graphique de ce résumé à 5 valeurs
  Lorsque les observations sont très dispersées, définition de 2 valeurs
  - $a_1 = 1er\ quartile - 1.5 \times I.Q$
  - $a_2 = 3eme\ quartile + 1.5 \times I.Q$

    ($I.Q.$ intervalle interquartile := 3eme quartile - 1er quartile)

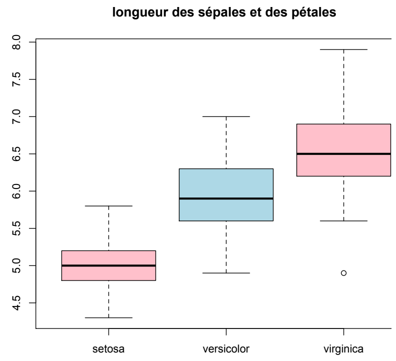
Variables quantitatives : résumé à 5 valeurs et box plot

```
> summary(iris$Sepal.Length)
   Min. 1st Qu.  Median    Mean 3rd
  4.300   5.100   5.800   5.843   6.
> quantile(SL,c(0,0.25,0.5,0.75,1))
  0\%  25\%  50\%  75\% 100\%
 4.3  5.1  5.8  6.4  7.9
> IQR(iris$Sepal.Length)
[1] 1.3
> boxplot(sl,main="longueur des
sépales",
col="lightblue")
>
```



Variables quantitatives : résumé à 5 valeurs et box plot

```
> boxplot(sl ~ iris[[5]],
   col=c("pink", "lightblue"),
   main="longueur des sépales
   et des pétales ")
```



Autres mesures

- différence moyenne : moyenne des valeurs absolues des différences des couples pris 2 à 2 ($k2$ couples)
- coefficients de dispersion relative
  - coefficient quartile (ou semi-interquartile relatif)
    $Q_r = E.med$ ou $Q'_r = \frac{Q3-Q1}{Md}$ ou $Q'_r = \frac{Q3-Q1}{Q_3+Q_1}$ ($= Q_r$ s i la distribution est symétrique)
  - coefficient de variation
    $V = \frac{\sigma}{\overline{x}}$

Dispersion pour les variables qualitatives

- Variable dichotomique (2 valeurs 0 ou 1) : mêmes définitions que pour les variables qualitatives
  $\overline{x} = \frac{1}{n}\sum_{i=1}^{k} x_i$: proportion des observations satisfaisant $x = 1$
  $1 - \overline{x}$: proportions des observations satisfaisant $x = 0$
  $\sigma^2 = \overline{x}(1 - \overline{x})$
  $\sigma^2 = \frac{1}{k}\sum_{i=1}^{k}(x_i - \overline{x})^2 = \frac{1}{k}\sum_{i=1}^{k}(x_i^2 - 2.x_i.\overline{x} + \overline{x}^2) = \frac{1}{k}\sum_{i=1}^{k} x_i - 2.\overline{x}^2 + \overline{x}^2 = \overline{x} - \overline{x}^2$
  (car $x_i^2 = x_i$)
- Variable multicatégorielle

$$\sigma^2(X) = p_1.\ldots.p_j$$

où p_i proportion d'observation dans la catégorie i.

Mesures de forme : asymétrie

Coefficient de Yule

$$s = \frac{(Q_3 - med) - (med - Q_1)}{(Q_3 - med) + (med - Q_1)}$$

- $s = 0$: symétrie parfaite
- $s > 0$: oblique à gauche ou étalement à droite
- $s < 0$: oblique à droite ou étalement à gauche

Mesures de forme : asymétrie

Coefficient de Pearson

- $$p = \frac{\overline{x} - mod}{s}$$
  même interprétation en remplaçant s par p (performant pour des distributions faiblement asymétriques)
- $$\beta_1 = \frac{\mu_3^2}{\mu_2^3}$$
  où $\mu_r = \frac{1}{n}\sum_{i=1}^{k} n_i(x_i - \overline{x})^r$: moment centré d'ordre r
  ($m_r = \frac{1}{n}\sum_{i=1}^{k} n_i x_i^r$: moment d'ordre r)
- coefficient de Fisher : racine carré du coefficient β_1 de Pearson

Mesures d'aplatissement

- coefficient de Pearson
  $$\beta_2 = \frac{\mu_4}{\mu_2^2}$$
  d'autant plus faible que la courbe est platicurtique
  $\beta^2 = 3$ pour une distribution normale
- coefficient de Fisher
  $$\gamma_2 = \beta_2 - 3$$
  positif pour une courbe est leptocurtique
  $\gamma_2 = 0$ pour une distribution normale

Un exemple

```
> par(mfrow=c(2,3))
> x<-rnorm(500,0,1)
> hist(x,proba=TRUE)
> lines(density(x))
> y<-rnorm(500,8,1)
> hist(y,proba=TRUE)
> lines(density(y))
> z<-x+y
> z<-c(x,y)
> hist(z,proba=TRUE)
> lines(density(z))
> t<-rnorm(500,1,1)
> hist(t,proba=TRUE)
> u<-c(x,t)
> lines(density(t))
> hist(u,proba=TRUE)
> lines(density(u))
> da<-data.frame("x"=x,"y"=y,"z"=z,"t"=t,"u"=u)
> write.table(da,"data_bimod.txt")
```


Exemple (suite)