Traitement des Données Biologiques : bases statistiques

M1 - MABS

Maxime Bonhomme

UMR CNRS-UPS 5546, Laboratoire de Recherche en Sciences Végétales, Castanet-Tolosan

12 septembre 2011

Bases statistiques pour le TDB

GENERALITES

définition

- science formelle, méthode et technique (ensemble de méthodes)
- **science de collecter, organiser, analyser et interpréter des données** (analyser les variations entre observations)
- le but est de disposer d'un outil d'aide à la décision

démarche générale

- collecte des données : plan d'expérience, échantillonnage
- traitement des données : description, estimation de paramètres, tests d'hypothèses
- interprétation et conclusion

exemples de problèmes abordés

- effet d'un traitement, comparaison phénotypique de lignées (ex : analyse de mutants)
- analyse d'expression (microarrays), association génotype phénotype...

- **population** : ensemble fini ou infini des objets d'une étude statistique
- **individu** (unité de base ou unité statistique) : entité sur laquelle porte l'étude, et élément de la population
  - étudiant : notes, âge, catégorie,...
  - plante : taille, taux de croissance, teneur,...
  - lignée : taux de germination, précocité,...
  - gène : nombre de transcrits, activation,...
- **échantillon** : sous-ensemble de la population sur lequel sont effectuées les observations. Petit Robert : "petite quantité d'une marchandise que l'on montre pour donner une idée de l'ensemble"
- **effectif** : nombre total d'individus d'une population ou de l'échantillon
- **variable ou caractère** : propriété étudiée sur les individus (variable aléatoire)
  - **qualitative :**
    - nominale non ordonnée : couleur, forme,...
    - nominale binaire : sexe, gauche/droite,...
    - ordinale : petit/moyen/grand, score, expression...
  - **quantitative :**
    - discrète (dénombrable) : nombre de racines latérales,...
    - continue : taille, poids, concentration, temps, ratio d'expression,...

  distinction intervalle / rapport : le zéro est arbitraire dans une variable d'intervalle (ex : température), et signifie l'absence de caractère dans une variable de rapport (ex : poids, vitesse)

- **série statistique** : suite d'observations réalisées sur un échantillon ou une population
- **variable aléatoire** : fonction définie sur l'ensemble des éventualités, c'est-à-dire l'ensemble des résultats possibles d'une expérience aléatoire. En particulier, si on change d'échantillon les résultats ou valeurs changent
- **statistique descriptive** :
  - organisation et description d'un ensemble de données
  - extraction d'information
- **statistique inférentielle** :
  - généralisation de l'échantillon à la population (tests d'hypothèses)
  - estimation de paramètres

plan d'expérience

- dispositif expérimental permettant la collecte des données en vue de répondre à une question donnée
- associé à la méthode statistique utilisée pour analyser les données
  - plans factoriels (exemple : deux traitements sur le même lot de personnes, sans interaction entre traitements)

	traitement B (n=200)	placebo de B (n=200)
traitement A (n=200)	traitement A traitement B (n=100)	traitement A placebo B (n=100)
placebo A (n=200)	placebo A traitement B (n=100)	placebo A placebo B (n=100)

  - plans expérimental en blocs aléatoires complets -PEBAC- (exemple : effet de différents traitements entre unités expérimentales, en champs). Le but est de réduire l'erreur expérimentale en éliminant la contribution de sources connues de variation entre les unités expérimentales

3	2	4	2	1	4
1	5	6	5	6	3
5	3	4	5	2	4
6	1	2	3	6	1

PEBAC relatif à la comparaison de six éléments : exemple de six fumures différentes, numérotées de 1 à 6 au sein de quatre blocs

STATISTIQUE DESCRIPTIVE

Variable qualitative

variable qualitative

- fréquence (effectif) absolue : nombre d'observations par catégorie (n_i)
- fréquences relatives : proportion d'observations de la catégorie par rapport à l'ensemble p de catégories

$$f_i = \frac{n_i}{\sum_{i=1}^{p} n_k} \tag{1}$$

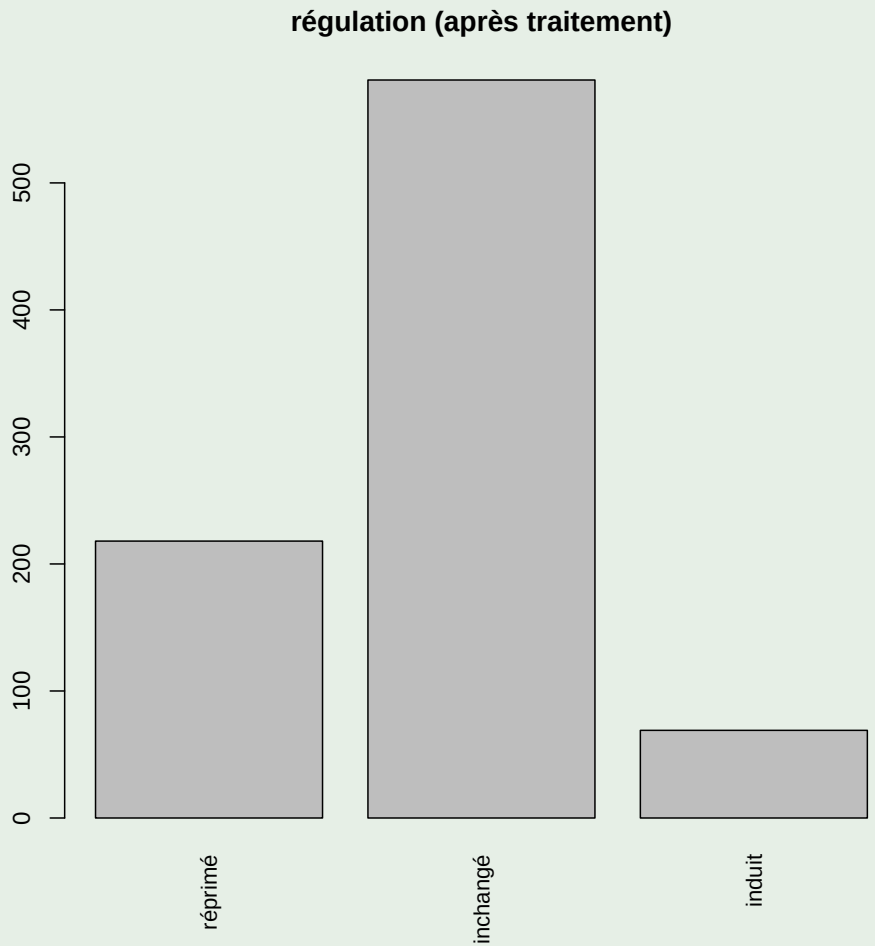
- fréquences cumulées (si variable ordonnée) :

$$N_i = \sum_{i=1}^{i} n_k \tag{2}$$

$$F_i = \sum_{i=1}^{i} f_k \tag{3}$$

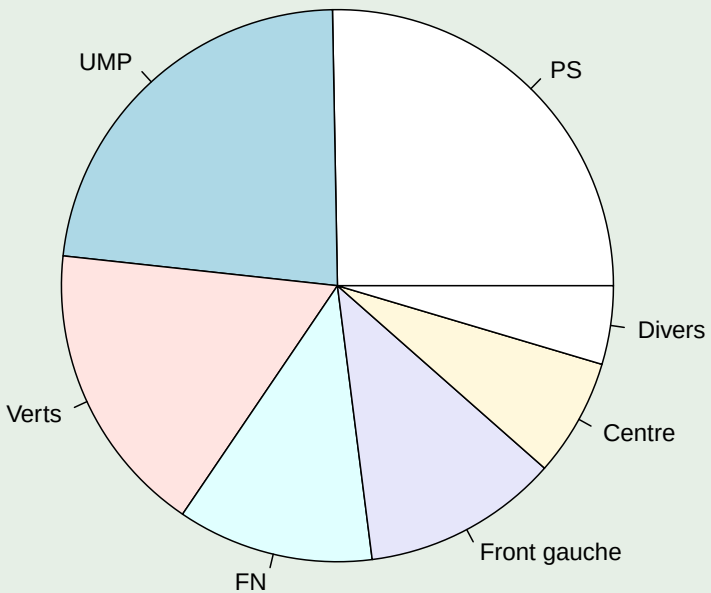
Variable qualitative

représentation : diagramme en barres



Variable qualitative

représentation : camembert



Variable quantitative

variable quantitative

- **répartition en classes**
- fréquence (effectif) absolue : nombre d'observations par classe (n_i)
- fréquences relatives : proportion d'observations de la classe par rapport à l'ensemble p des classes

$$f_i = \frac{n_i}{\sum_{i=1}^{p} n_k} \tag{4}$$

- fréquences cumulées (si variable ordonnée) :

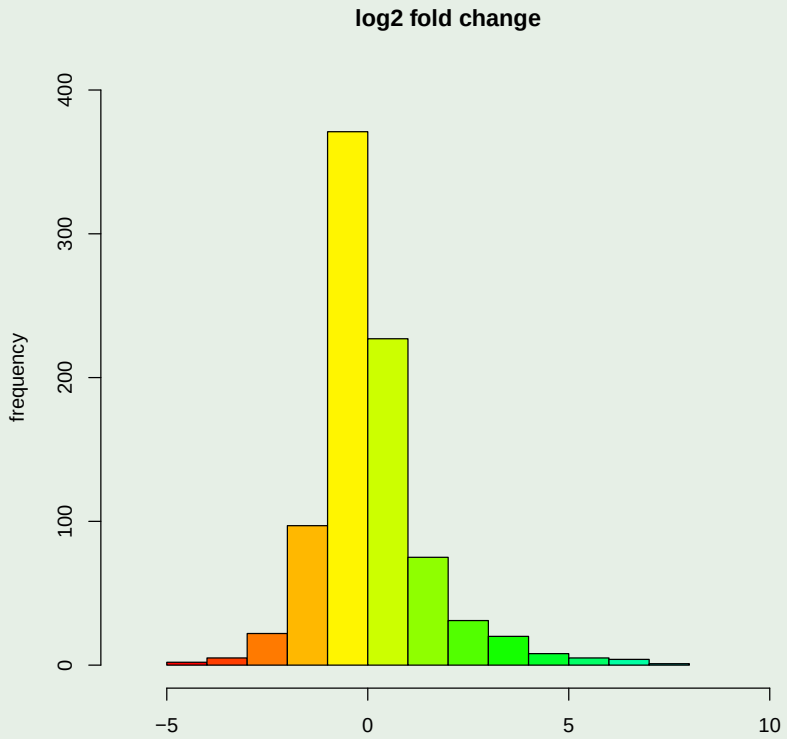
$$N_i = \sum_{i=1}^{i} n_k \tag{5}$$

$$F_i = \sum_{i=1}^{i} f_k \tag{6}$$

Variable quantitative

représentation : histogramme

- graphique représentant une distribution statistique par des rectangles verticaux de surface proportionnelle aux effectifs



règle de Sturges : Nb classes $\sim log2(n) + 1$

Tendance centrale

paramètres d'une distribution (x_i, i=1,...,n) : tendance centrale

- **moyenne** :

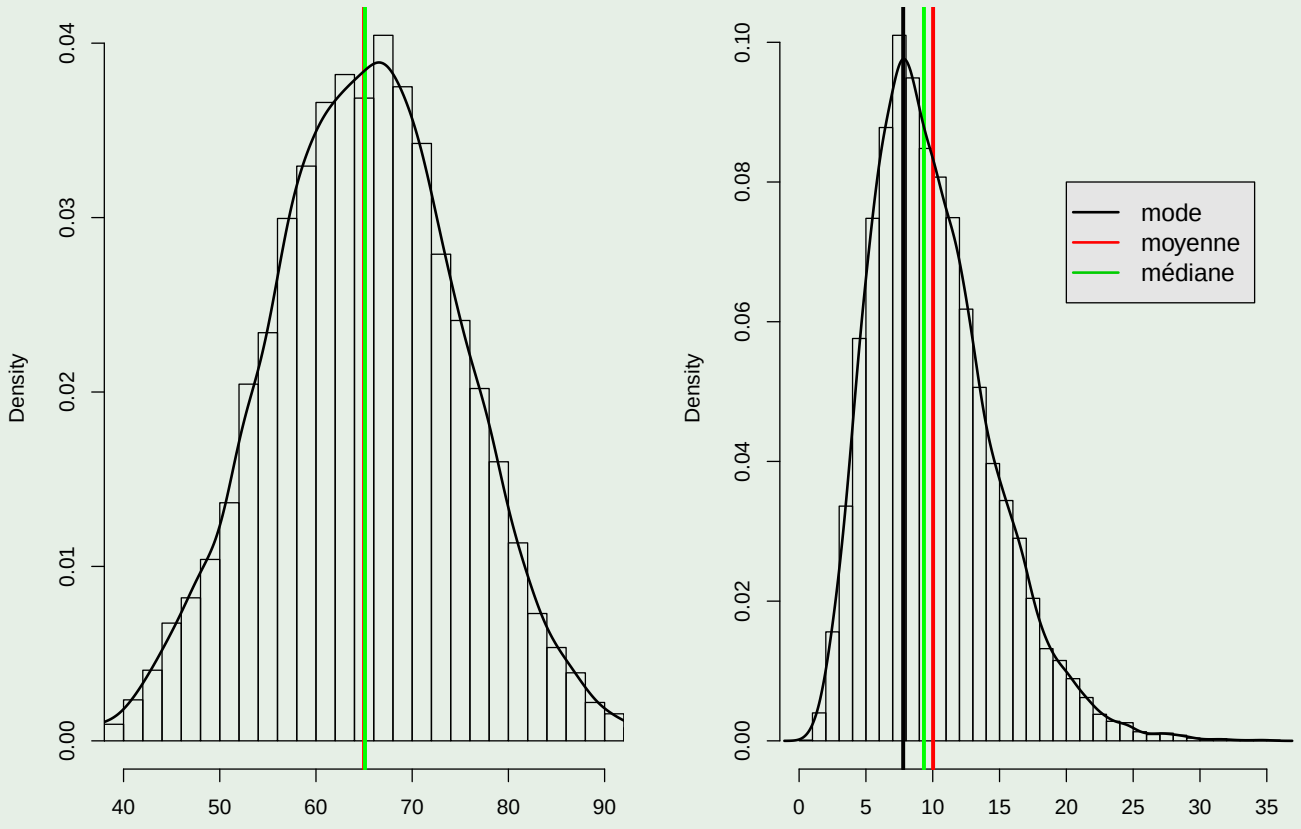
$$m = \frac{1}{n} \sum_{i=1}^{n} x_i \tag{7}$$

$$m = \sum_{i=1}^{n} x_k p_k \tag{8}$$

avec $p_k = n_k / n$

- **médiane** : valeur en dessous de laquelle sont situées 50% des observations
- **quartiles** : valeurs à 25%, 50% et 75% de l'effectif
- **centiles** : valeurs à x% de l'effectif
- **mode** : valeur (ou classe) la plus fréquente

Tendance centrale



Tendance centrale

autres moyennes

- **moyenne arithmétique pondérée** : valeurs $(X = x_1, x_2, ..., x_n)$ affectées de coefficients $(W = w_1, w_2, ..., w_n)$.

$$m = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i} \tag{9}$$

- **moyenne harmonique**, si fractions (ex : calcul de la vitesse moyenne) :

$$m = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}} \tag{10}$$

- **moyenne géometrique**, si multiplicatif ou cumulatif (ex : carré et rectangle de même surface) :

$$m = \sqrt[n]{\prod_{i=1}^{n} x_i} \tag{11}$$

ex : le carré (rectangle moyen à deux côtés égaux) qui a même surface qu'un rectangle de côtés 3 et 7 a pour côté $\sqrt[2]{3 * 7} = 4.58$

Dispersion

paramètres d'une distribution (x_i, i=1,...,n) : dispersion

- **variance (= moment centré d'ordre 2)** :

$$s_n^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - m)^2 \tag{12}$$

$$s_n^2 = \sum_{i=1}^{n}(x_k - m)^2 p_k \tag{13}$$

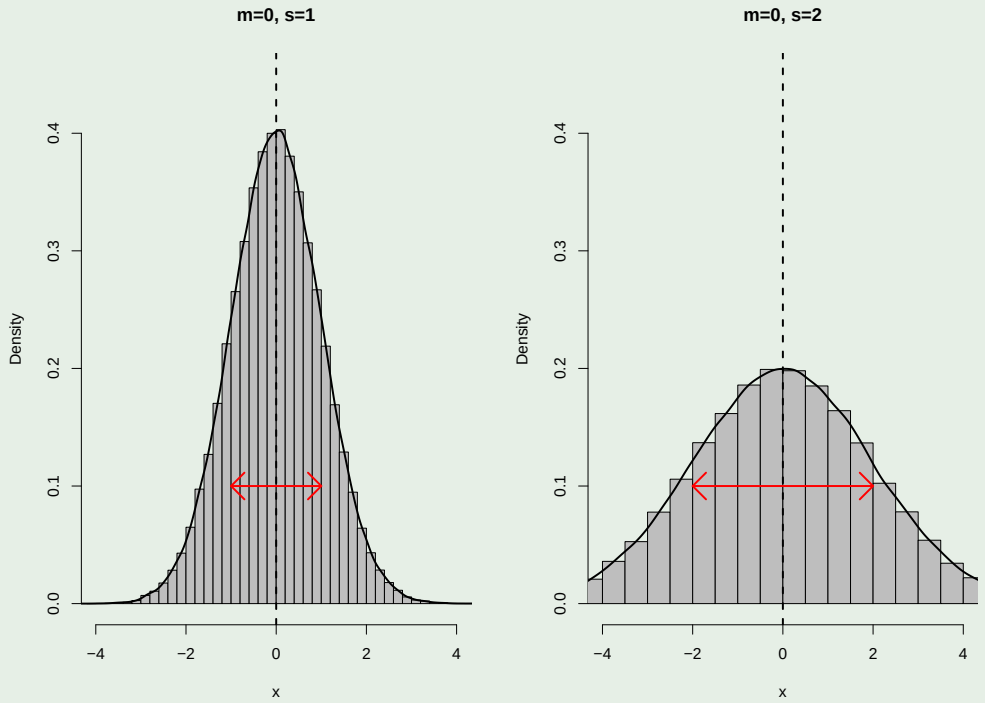
- valable que si on connait la vraie moyenne de la population. Donc 1 degré de liberté de moins correspondant au calcul de la moyenne (ddl = nb de valeurs qui sont libres de varier dans le calcul final de la statistique) :

$$s_{n-1}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - m)^2 \tag{14}$$

notations

- écart-type (standard déviation -SD) : s
- m et s^2 (s) : estimateurs de la moyenne et de la variance (écart-type) de la population à partir de l'échantillon
- μ et σ^2 (σ) : vraie moyenne et variance (écart-type) de la population
- $\mathbb{E}(X)$ et $\mathrm{Var}(X)$: espérance (moyenne) et variance de la variable aléatoire X
- coefficient de variation $cv = s/m$

Dispersion



propriétés de la variance

- $\sigma^2(X) = \mathrm{Var}(X) = \mathbb{E}(X^2)\text{-}\mathbb{E}(X)^2$
- $\sigma^2(X+Y) = \mathrm{Var}(X+Y) = \mathrm{Var}(X)+\mathrm{Var}(Y)$ (si X et Y indépendantes)
- $\sigma^2(X-Y) = \mathrm{Var}(X-Y) = \mathrm{Var}(X)+\mathrm{Var}(Y)$ (si X et Y indépendantes)

Dissymétrie et aplatissement

paramètres d'une distribution (x_i, i=1,...,n) : dissymétrie et aplatissement

- **aplatissement (kurtosis)** :

$$[\frac{n(n+1)}{(n-1)(n-2)(n-3)}\sum_{i=1}^{n}(\frac{x_i-\bar{x}}{s})^4] - \frac{3(n-1)^2}{(n-2)(n-3)} \quad (15)$$

  - $= 0$ pour une loi normale centrée réduite
  - > 0 pour une distribution "pointue"
  - < 0 pour une distribution "aplatie"

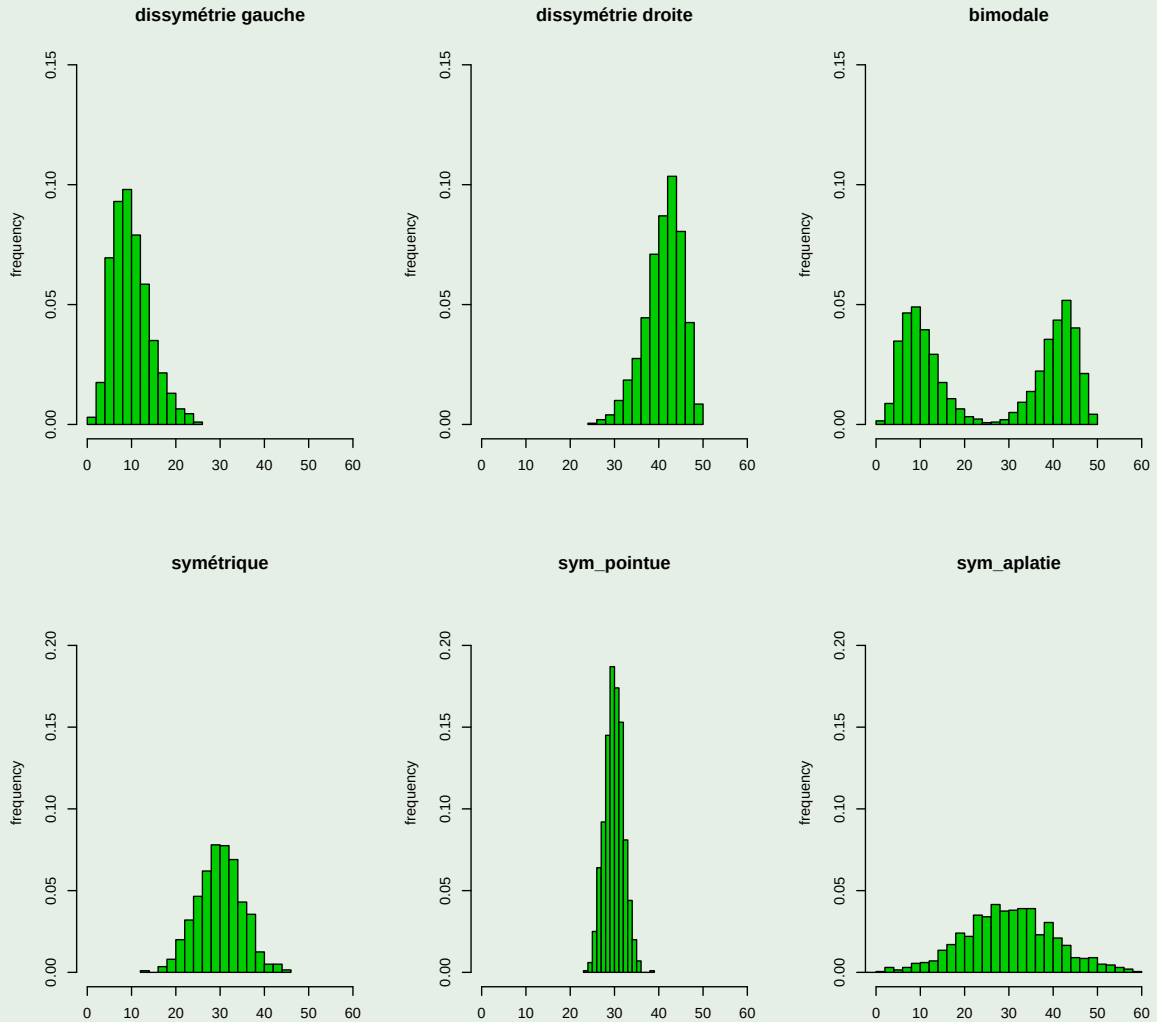
- **dissymétrie (skewness)** :

$$\frac{n}{(n-1)(n-2)}\sum_{i=1}^{n}(\frac{x_i-\bar{x}}{s})^3 \quad (16)$$

  - $= 0$ pour une distribution symétrique
  - > 0 pour une distribution étalée à droite
  - < 0 pour une distribution étalée à gauche

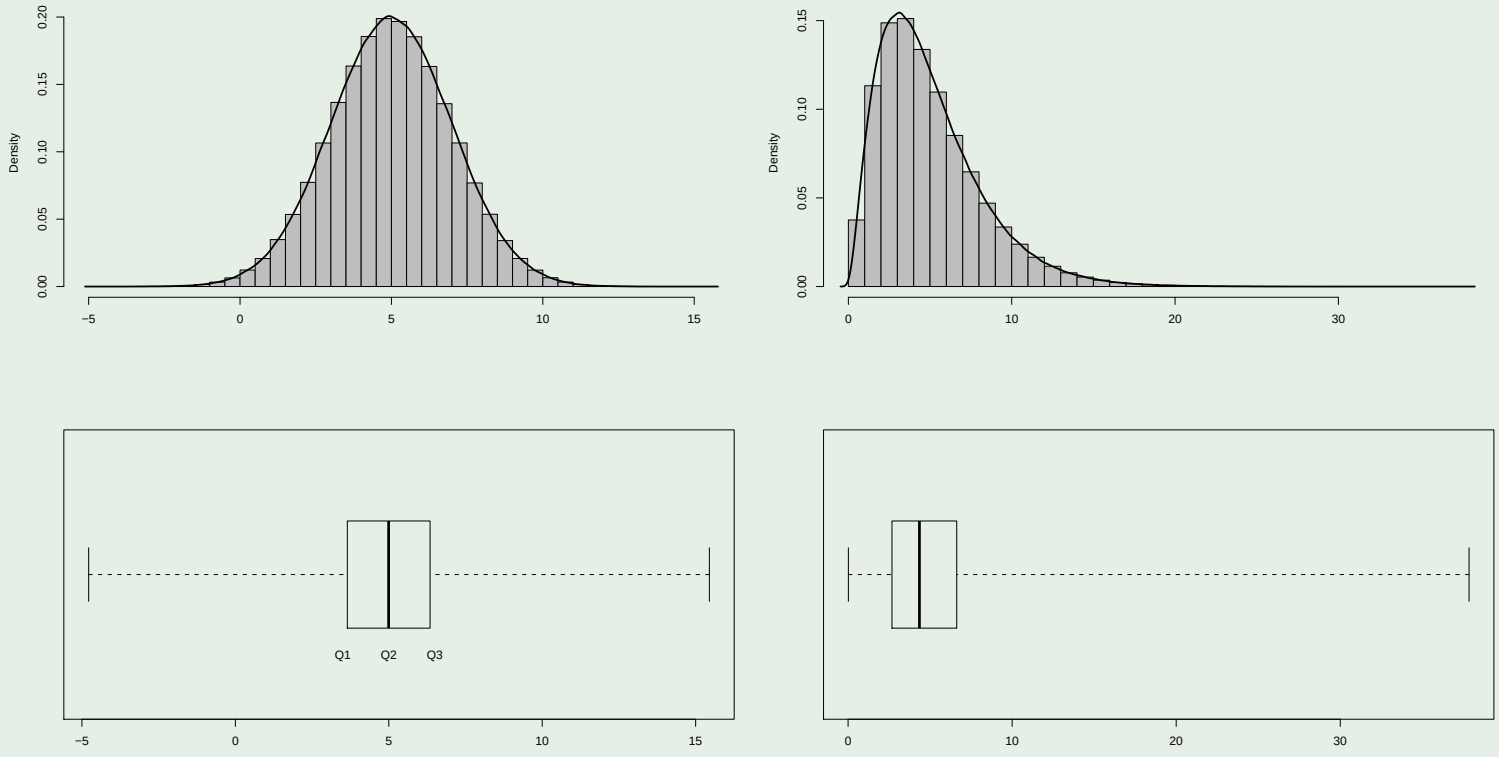
Représentation d'une série statistique

exemples de distributions de fréquences



Représentation d'une série statistique

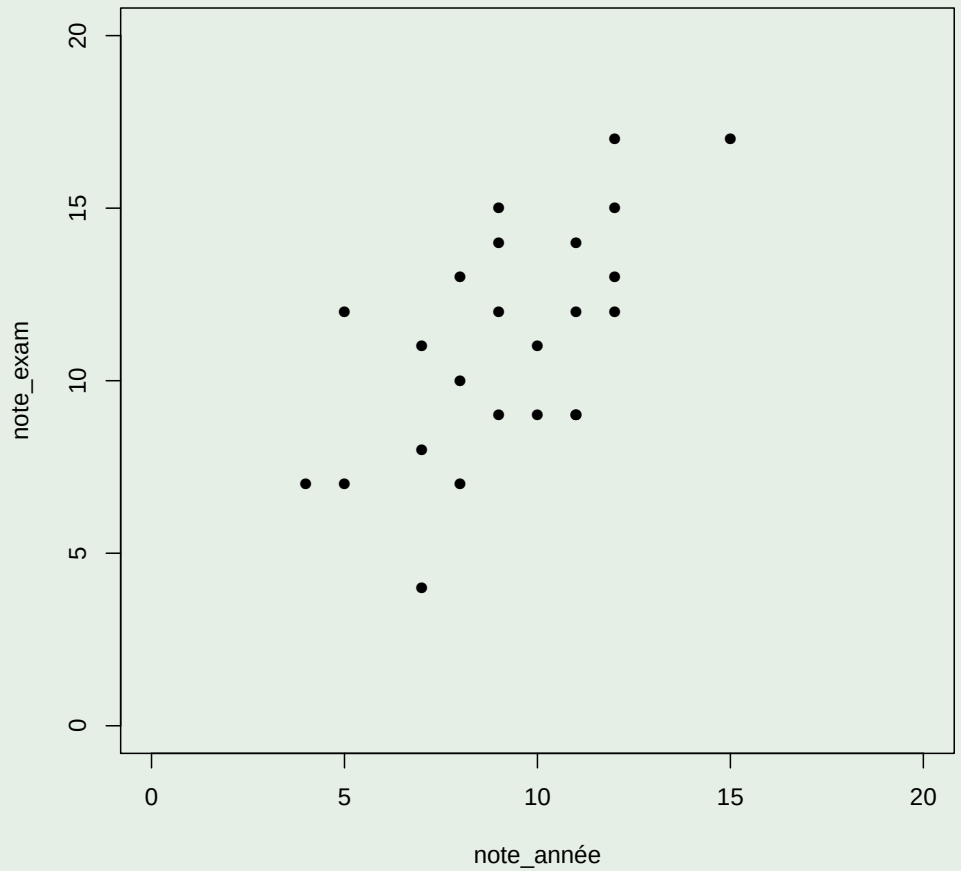
boîte à moustache (boxplot)



$Q1$ = quartile 1 (1er quart des données), $Q2$ = médiane, $Q3$ = quartile 1, 3ème quart des données ;
(nb : dans le cas d'une loi Normale, environ 95% des valeurs sont comprises entre les deux extêmes)

Représentation

nuage de points



Liaison entre deux variables quantitatives X et Y

- **covariance**

$$\mathrm{Cov}(X, Y) = \sigma_{XY} = s_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) \tag{17}$$

- **coefficient de corrélation linéaire (Pearson)**

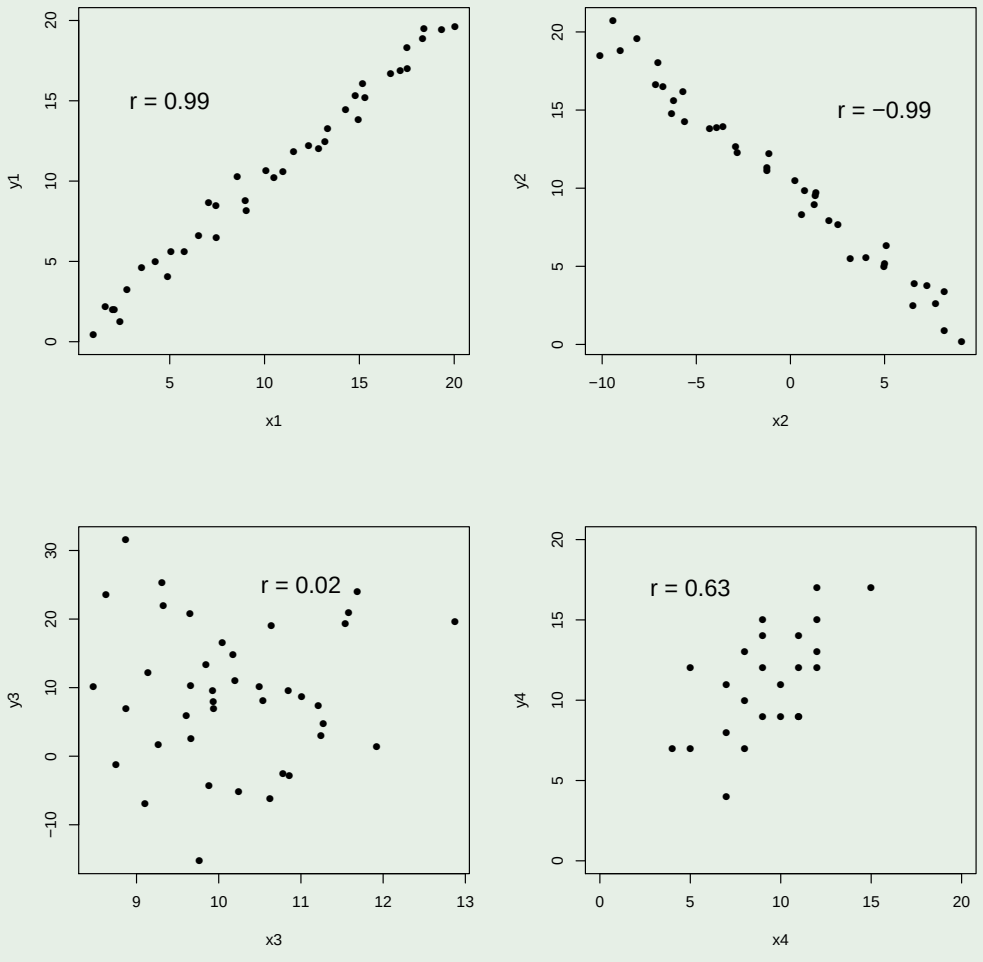
$$r = \frac{s_{xy}}{s_x s_y} \tag{18}$$

avec

- $\bar{x}$ et $\bar{y}$: estimateurs de la moyenne des variables X et Y
- s_x et s_y : estimateurs de l'écart-type des variables X et Y
- $-1 < r < 1$, $r < 0$ = corrélation négative, $r > 0$ = corrélation positive, $r = 0$ pas de corrélation entre X et Y
- $-\infty < s_{xy} < +\infty$, $s_{xy} = 0$ indépendance de X et Y

- **coefficient de détermination** $= r^2$
  - 1 = ajustement parfait
  - $0.7 < r < 1$ = ajustement justifié
  - $r < 0.7$ = ajustement non justifié

exemple de corrélation

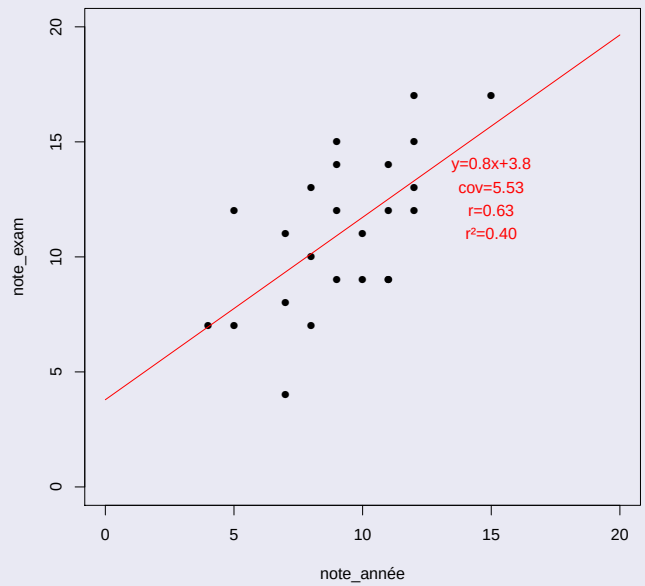


Notion de régression

- BUT : faire passer une droite qui passe au plus près des points
- droite de régression $\hat{y_i} = ax_i + b$
- les coefficients de la droite de régression sont calculés de manière à minimiser la somme des carrés des écarts entre les valeurs observées y_i et les valeurs estimées $\hat{y_i}$ (méthode des moindres carrés)

$$S = min(\sum_{i=1}^{n}(y_i - \hat{y_i})^2) \tag{19}$$

- droite de régression $y = ax + b$ avec $a = \frac{s_{xy}}{s_x}$



check-list

- individu ?
- population étudiée ?
- échantillon ou population ?
- effectif ?
- variables :
  - nombre
  - nature
  - nombre de catégories (cas de var qualitative)
- séries -variables- indépendantes ou appariées **(ex : mesure à deux temps proches, correction d'un ensemble de copies par deux examinateurs) ?**
- variable
  - fixée **(25 plantes choisies dans chacune des 4 parcelles d'une récolte : "parcelle" = fixée)**
  - aléatoire **(100 plantes choisies au hasard sur les 4 parcelles d'une récolte : "parcelle" = aléatoire, d'où accès à la distribution de la variable)**