Traitement des Données Biologiques : bases statistiques

M1 - MABS

Maxime Bonhomme

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

11 septembre 2013

Bases statistiques pour le TDB

1. statistique inférentielle
   - les tests statistiques usuels
   - tests multiples
   - liaison entre plusieurs variables quantitatives
2. analyse multivariées
   - introduction aux analyses multivariées
3. références

STATISTIQUE INFERENTIELLE

Tests d'ajustement

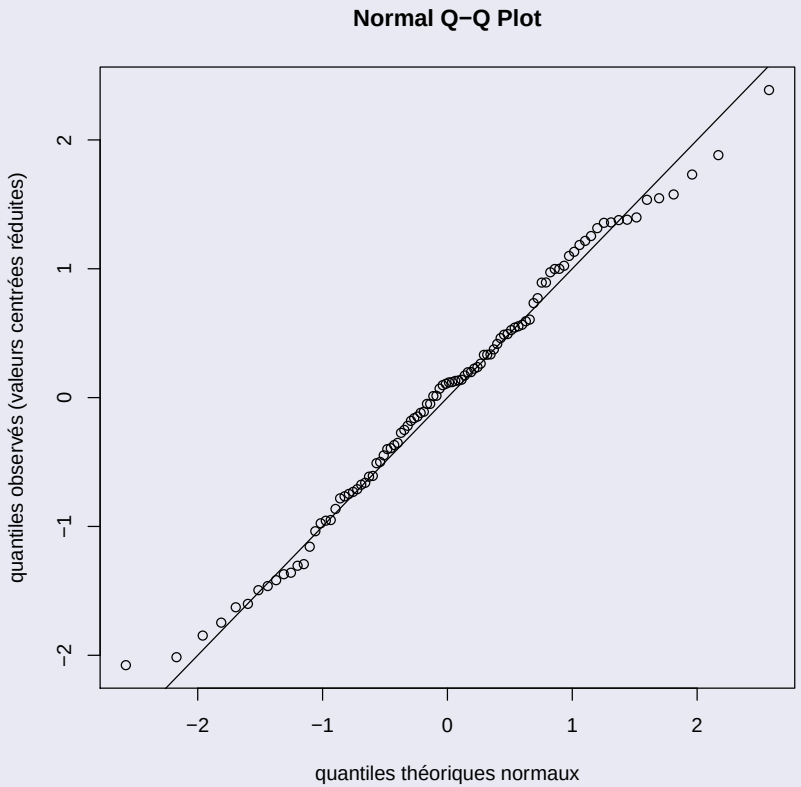
Test χ^2 d'ajustement de Pearson

- **on dispose d'une distribution observée (n observations réparties en p classes) que l'on veut comparer à une distribution théorique**
- $\chi^2_{obs} = \sum_{i=1}^{p} \frac{(n_i - nP_i)^2}{nP_i}$ suit une loi du χ^2 à $(p-1)$ ddl (si $H0$ vraie)
- on rejète l'hypothèse d'adéquation si $\mathbb{P}(\chi^2 > \chi^2_{obs}) < \alpha$
  - le test est toujours unilatéral
  - condition d'utilisation : $nP_i > 5$
- nb : si la distribution théorique n'est pas complètement définie (cas le plus fréquent) alors les valeurs des P_i doivent être déterminées à partir des observations. Dans ce cas le nombre ddl doit être diminué du nombre de paramètres estimés

Tests d'ajustement : normalité

Diagramme de probabilité (quantile-quantile plot)

- soit une série statistique (X_i), $i = 1, 2, ...n$



- les ordonnées sont déterminées de sorte que la fonction de répartition forme une droite dans le cas d'une distribution normale

Tests d'ajustement : normalité

- **test du χ^2** : distribution de fréquences groupées en classes (cf : test χ^2 d'ajustement)
- **test de Shapiro-Wilk** :
  - test puissant (pour n petit), calcul complexe
  - $W = \frac{[\sum_{i=1}^{[\frac{n}{2}]} a_i (x_{(n-i+1)} - x_{(i)})]^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$
  - a_i sont des constantes générées à partir de la moyenne et de la matrice de variance covariance des quantiles d'un échantillon de taille n suivant la loi normale. Ces constantes sont fournies dans des tables spécifiques
  - $0 < W < 1$, normalité si $W \to 1$
- **test de Kolmogorov-Smirnov** :
  - peu puissant, calcul simple
  - compare les fonctions de répartition observée et théorique, avec une distance
- si données groupées en i classes utiliser le test du χ^2 à condition que $n_i > 5$, sinon test de Kolmogorov
- si une série de n données utiliser le test de Shapiro-Wilk pour $n < 200$, sinon plutôt test de Kolmogorov

Transformations de variables

- réduire l'hétérogénéité des variances entre groupes, rendre "normale" ou "symétrique" une distribution
- **la transformation logarithmique** : $Y = log(X)$
  - loi log-normale (ex : microarray, qPCR,...)
  - quand écart-types des groupes $\propto$ moyennes des groupes
- **la transformation racine carrée** : $Y = \sqrt{X}$
  - quand variances des groupes $\propto$ moyennes des groupes
- **la transformation Box et Cox** :
  -
$$Y = \begin{cases} \frac{(X^{\lambda}-1)}{\lambda} & \text{si } \lambda \neq 0 \\ log(X) & \text{si } \lambda = 0 \end{cases}$$
  - stabilisation des variances
- **la transformation angulaire** : $Y = 2\arcsin(\sqrt{\frac{X}{n}})$
  - valeurs binomiales, pourcentages

Tests sur les fréquences

- soit une population (supposée infinie) dans laquelle la proportion d'individus présentant un caractère donné est p
- dans un échantillon de n individus, le nombre k d'individus présentant ce caractère suit une loi binomiale $B(n, p)$
  - si n est petit les tests seront basés sur la loi binomiale
  - si n est grands ($n > 50$ et $np > 10$), on appliquera l'approximation normale avec ou sans transformation angulaire

Test de conformité : $H0 : p = p_0, H1 : p \neq p_0$

- pour $n > 40$: $u_{obs} = \frac{|f - p_0|}{\sqrt{\frac{p_0(1-p_0)}{n}}}$, où f est la fréquence observée, on rejette $H0$ si $\mathbb{P}(|U| > u_{obs}) \leq \alpha$
- nombre d'observations nécessaires pour effectuer le test avec une puissance donnée : $n = \frac{p(1-p)}{d^2}(u_{1-\frac{\alpha}{2}} + u_{1-\beta})^2$, où d est la différence attendue
- exemple : pour $\alpha = 0.05$, $\beta = 0.5$, $p = 0.5$ et $d = 0.01$, alors $n \approx 10000$

Tests sur les fréquences

Comparaison de deux proportions : $H0 : p_1 = p_2, H1 : p_1 \neq p_2$ (échantillons indépendants)

- pour n faible : test exact de Fisher (loi hypergéométrique)
- pour $n > 40$: approximation normale
  - $u_{obs} = \frac{|f_1 - f_2|}{\sqrt{f(1-f)(\frac{1}{n_1} + \frac{1}{n_2})}}$, où $f_1 = \frac{x_1}{n_1}$, $f_2 = \frac{x_2}{n_2}$ et $f = \frac{x_1 + x_2}{n_1 + n_2}$ sont les fréquences observées
  - on rejette $H0$ si $\mathbb{P}(|U| > u_{obs}) \leq \alpha$

Test du χ^2 d'indépendance (tableaux de contingence)

objectif : rechercher s'il y a indépendance (ou association) entre les classes de deux variables qualitatives, par l'analyse de la répartition des effectifs au sein de ces classes

observation des couleurs des fleurs pour différents croisements de plantes

- 4 couleurs, 8 familles, 12 à 30 plantes observées/famille
- tableaux de contingence :

famille	blanc	violet clair	violet moyen	violet foncé	effectifs
1	0	6	8	0	14
2	0	7	9	1	17
...	...	...	...	...	...
effectifs	7	64	62	18	151

- question : la répartition des couleurs change-t-elle d'une famille à l'autre ? autrement dit y-a-t-il indépendance entre couleur et appartenance à une famille ?

Test du χ^2 d'indépendance (tableaux de contingence)

- statistique du χ^2 :

$$\chi^2_{obs} = \sum_{i=1}^{p}\sum_{j=1}^{q} \frac{(n_{ij} - n\hat{P}_{ij})^2}{n\hat{P}_{ij}} \quad (1)$$

- notation :
  - p=nombre de classes i
  - q=nombre de classes j
  - n_{ij}=effectif de la classe ij
  - $n_{i.}$=effectif total de la classe i
  - $n_{.j}$=effectif total de la classe j
  - n=effectif total
  - $n\hat{P}_{ij} = \frac{n_{i.}n_{.j}}{n}$
- nombre de degrés de liberté de la loi χ^2 utilisée pour $H0$: $(p-1)(q-1)$
- on rejette l'hypothèse $H0$ d'indépendance si $\mathbb{P}(\chi^2 > \chi^2_{obs}) \leq \alpha$
- condition d'utilisation du test : $n_{ij} < 5$
- cas particulier où $p = q = 2$: $\chi^2_{obs} = \frac{n(n_{11}n_{22} - n_{12}n_{21})^2}{n_{1.}n_{.1}n_{2.}n_{.2}}$
- ce test peut être utilisé pour comparer deux distributions (cf test du χ^2 d'ajustement)

Comparaison des variances

2 populations (échantillons indépendants) : test de Fisher-Snedecor

- soient 2 échantillons extraits de populations suivant des lois $\mathcal{N}(\mu_1, \sigma_1)$ et $\mathcal{N}(\mu_2, \sigma_2)$ (condition : populations normales)
- $H0 : \sigma_1^2 = \sigma_2^2$
- sous $H0$ la statistique $F_{obs} = \frac{\hat{\sigma}_1^2}{\hat{\sigma}_2^2}$ suit une loi $F_{(n_1-1, n_2-1)}$
- rejet de $H0$ si $F_{obs} \notin [F_{(n_1-1,n_2-1)}(\frac{\alpha}{2}), F_{(n_1-1,n_2-1)}(1-\frac{\alpha}{2})]$ $(H1 : \sigma_1^2 \neq \sigma_2^2)$
- en pratique, on met la plus forte variance au numérateur $(F_{obs} > 1)$ et on rejette $H0$ si $F_{obs} > F_{(n_1-1,n_2-1)}(1-\frac{\alpha}{2})$ $(H1 : \sigma_1^2 > \sigma_2^2)$

plusieurs populations normales

- soient p échantillons extraits de populations suivant des lois $\mathcal{N}(\mu_i, \sigma_i)$
- $H0 : \sigma_1^2 = \sigma_2^2 = ... = \sigma_p^2$
  - **test de Bartlett** : peu robuste à la non normalité
  - **test de Cochran** : $C = \frac{\max(s_i^2)}{\sum(s_i^2)}$, effectifs égaux
  - **test de Hartley** : $F_{max} = \frac{\max(s_i^2)}{\min(s_i^2)}$, effectifs égaux
  - **test de Levene** : ANOVA sur $Z_{ij} = |x_{ij} - m_i|$, robuste à la non normalité

Comparaison d'une moyenne à un standard (test de conformité)

- $H0 : \mu = \mu_0$
- **cas de grands échantillons ($n > 30$) :**
  - rappel : $\bar{x} \rightarrow \mathcal{N}(\mu, \frac{\hat{\sigma}}{\sqrt{n}})$
  - si $H0$ vraie alors $U = \frac{\bar{x} - \mu_0}{\frac{\hat{\sigma}}{\sqrt{n}}}$ suit une loi $\mathcal{N}(0,1)$
  - si $H1 : \mu \neq \mu_0$ alors on rejette $H0$ si $|U| > U_{1-\frac{\alpha}{2}}$
  - si $H1 : \mu > \mu_0$ alors on rejette $H0$ si $U > U_{1-\alpha}$
- **cas de petits échantillons ($n < 30$) :**
  - condition d'utilisation : X suit une loi normale
  - si σ est connu alors $U = \frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$ suit une loi $\mathcal{N}(0,1)$
  - si σ est inconnu alors $t = \frac{\bar{x} - \mu_0}{\frac{\hat{\sigma}}{\sqrt{n}}}$ suit une loi $\mathcal{T}(n-1)$
  - si $H1 : \mu \neq \mu_0$ alors on rejette $H0$ si $|t| > t_{1-\frac{\alpha}{2}}$
  - si $H1 : \mu > \mu_0$ alors on rejette $H0$ si $t > t_{1-\alpha}$

Comparaison des moyennes de deux populations (variances connues)

- $X_1 \to \mathcal{N}(\mu_1, \sigma_1), X_2 \to \mathcal{N}(\mu_2, \sigma_2)$, σ_1 **et** σ_2 **connues**
- $H0 : \mu_1 = \mu_2, H1 : \mu_1 \neq \mu_2$
  - $\bar{x}_1 - \bar{x}_2$ suit une loi normale de variance $\sigma^2 = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$
  - si $H0$ vraie alors $U = \frac{\bar{x}_1 - \bar{x}_2}{\sigma}$ suit une loi $\mathcal{N}(0,1)$
- on peut donc :
  - soit calculer (U_{obs}) et le comparer à α
  - soit calculer $U_{1-\frac{\alpha}{2}}$ et le comparer à U_{obs}

cas de grands échantillons (n_1 et $n_2 > 30$), et indépendants

- si $H0$ vraie alors $U = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$ suit une loi $\mathcal{N}(0,1)$
- si $H1 : \mu_1 \neq \mu_2$ alors on rejette $H0$ si $|U_{obs}| > U_{1-\frac{\alpha}{2}}$
- si $H1 : \mu_1 > \mu_2$ alors on rejette $H0$ si $U_{obs} > U_{1-\alpha}$

Comparaison des moyennes de deux populations (variances estimées)

cas de petits échantillons (n_1 et $n_2 < 30$), et indépendants : test de Student

- condition d'utilisation :
  - X_1 et X_2 suivent des lois normales
  - X_1 et X_2 **ont la même variance (homoscédasticité)**
- si $H0$ vraie alors $t = \frac{\bar{x}_1 - \bar{x}_2}{\hat{\sigma}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$ suit une loi de Student $\mathcal{T}(n_1 + n_2 - 2)$ où $\hat{\sigma}^2 = \frac{(n_1-1)\hat{\sigma}_1^2+(n_2-1)\hat{\sigma}_2^2}{n_1+n_2-2}$
- si $H1 : \mu_1 \neq \mu_2$ alors on rejette $H0$ si $|t| > U_{1-\frac{\alpha}{2}}$
- si $H1 : \mu_1 > \mu_2$ alors on rejette $H0$ si $t > U_{1-\alpha}$

cas des variances inégales

- quand les variances sont inégales une variante du test de Student peut être utilisée : **le test de Welch**
  - le principe reste le même, mais l'estimateur de la variance commune et le nombre de ddl sont différents
  - dans le doute, utiliser de préférence le test de Welch

Comparaison des moyennes de deux populations (variances estimées)

échantillons appariés

- X_1 et X_2 sont la même variable mesurée à des temps différents
- soit $X_D = X_1 - X_2$, alors l'hypothèse $H0 : \bar{X}_1 = \bar{X}_2$ peut être remplacée par l'hypothèse $H0 : \bar{X}_D = 0$
- $t = \frac{\bar{X}_D - 0}{\frac{\hat{\sigma}}{\sqrt{n}}}$ suit une loi de Student $\mathcal{T}(n-1)$
- 0 peut être remplacé par une constante μ_0 si l'on veut tester si la moyenne de la différence est significativement différente d'une valeur μ_0

Quelques tests non paramétriques

comparaison de deux échantillons indépendants : test des rangs de Mann-Whitney (ou Wilcoxon)

- classer l'ensemble des observations par ordre croissant
- déterminer leur rang
- calculer la somme des rangs (X_1) relative à l'échantillon 1
- calculer $U_1 = X_1 - \frac{n_1(n_1+1)}{2}$
- comparer la plus petite des valeurs U_1 ou $U_2 = (n_1 n_2 - U_1)$ aux valeurs critiques de la table de Mann-Whitney
- utilisation si $n_1 + n_2 > 40$
- remarque : $U_{obs} = \frac{U_1 - \frac{n_1 n_2}{2}}{\sqrt{\frac{n_1 n_2 (n_1+n_2+1)}{12}}}$ suit une loi $\mathcal{N}(0,1)$
- équivalent non paramétrique du test t de Student (plus robuste aux valeurs extrêmes car il compare les rangs, et robuste à la non normalité)

Quelques tests non paramétriques

comparaison de deux échantillons appariés : test des rangs-signés de Wilcoxon

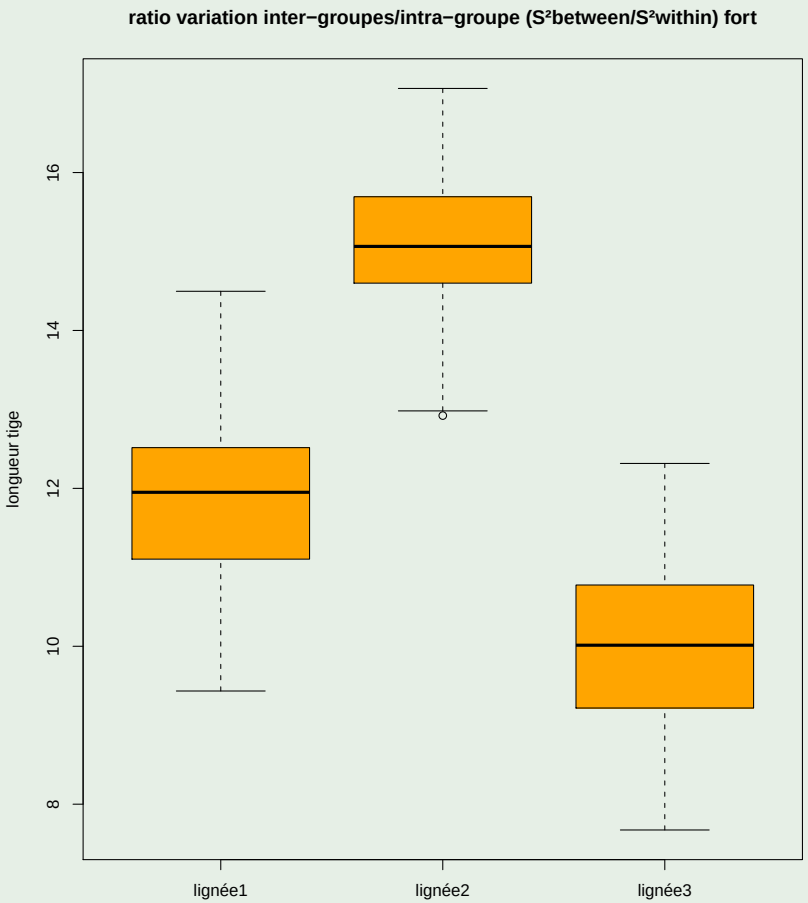
- calculer les différences $Z_i = X_i - Y_i$
- calculer les rangs des $|Z_i|$
- calculer la somme des rangs $(W+)$ des valeurs $Z_i > 0$ et la somme des rangs $(W-)$ des valeurs $Z_i < 0$
- calculer $S = \min(W+, W-)$
- comparer S à la table des valeurs critiques
- utilisation si $n > 25$
- remarque : $U_{obs} = \frac{S - \frac{n(n+1)}{4}}{\sqrt{\frac{n(n+1)(2n+1)}{24}}}$ suit une loi $\mathcal{N}(0,1)$
- équivalent non paramétrique du test t de Student pour données appariées (si données non normales, ordinales)

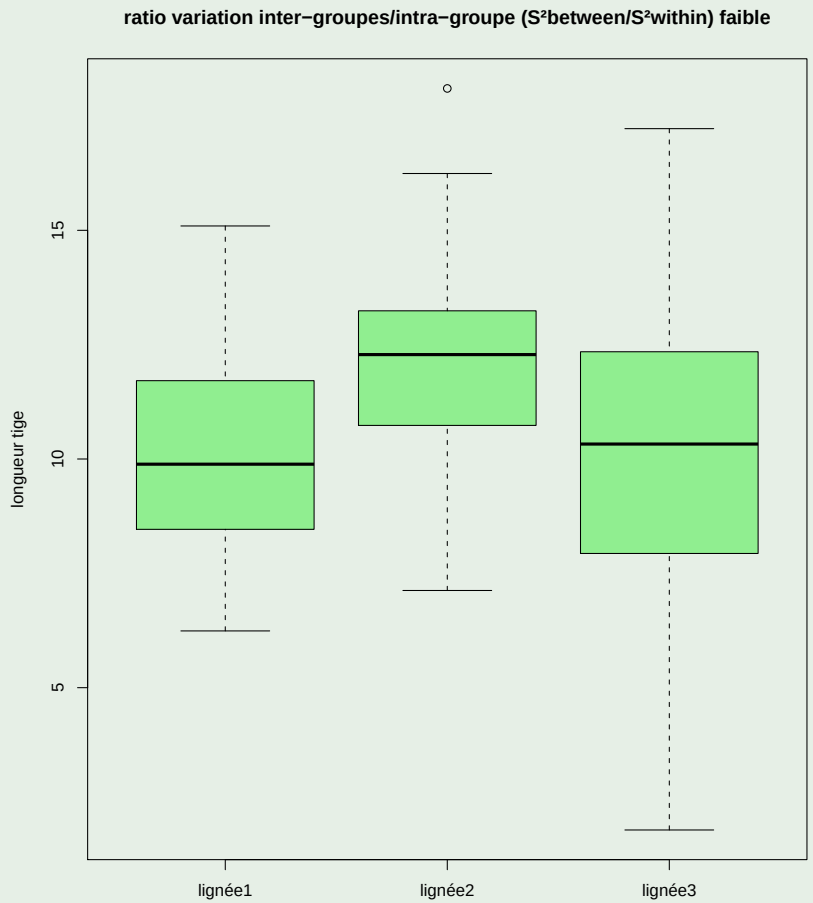
Analyse de la variance (ANOVA)

- **généralisation de la comparaison de moyennes à $K > 2$ groupes**
- permet de déterminer (avec un risque α) si un facteur (**variable qualitative nominale**) ou une combinaison de facteurs a un effet sur la **variable quantitative** étudiée
- teste $H0 : \mu_1 = \mu_2 = ... = \mu_p$, H1 : au moins 2 moyennes sont significativement différentes (mais on ne sait pas lesquelles)
- méthode basée sur la **décomposition de la variance totale** (comparaison d'estimateurs de la variance avec le test de Fisher-Snedecor)
  - **analyse de variance à un facteur** : comparer plusieurs échantillons selon un facteur de variation (groupes correspondants au différents niveaux de facteur)
  - **analyse de variance à plusieurs facteurs** : comparer plusieurs échantillons selon plusieurs facteur de variation (ex : sexe, lignée)
  - **analyse de variance de modèles à effets aléatoires** : par exemple, si un nombre élevé de groupes, on pose un effet groupe aléatoire

Analyse de la variance (ANOVA)

analyse de variance à un facteur : variable quantitative = longueur de la tige, facteur = lignée (3 niveaux)





Analyse de la variance (ANOVA)

ANOVA à un facteur

- on dispose de p échantillons (avec x_{ik} : échantillon i, individu k), d'effectifs n_i et de moyennes $\bar{x}_{i(i=1,\dots,p)}$
- exemples :
  - comparaison des hauteurs d'arbres de trois forêts
  - étude de l'homogénéité des rendements fourragers de cinq prairies
- conditions d'applications
  - **échantillons aléatoires, simples et indépendants**
  - **populations normales et de même variance (homoscédasticité)**
- soit n l'effectif total et $\bar{x}$ la moyenne générale, on peut écrire :

$$\sum_{i=1}^{p}\sum_{k=1}^{n_i}(x_{ik}-\bar{x})^2 = \sum_{i=1}^{p} n_i(\bar{x}_i-\bar{x})^2 + \sum_{i=1}^{p}\sum_{k=1}^{n_i}(x_{ik}-\bar{x}_i)^2 \qquad (2)$$

  - $SCE_t = SCE_b + SCR_w$
  - avec : SCE_t la variation totale, SCE_b la variation factorielle (inter-groupes), SCE_w la variation résiduelle (intra-groupes)
- l'analyse de variance consiste à **comparer la variation factorielle à la variation résiduelle**

Analyse de la variance (ANOVA)

le tableau de l'ANOVA à un facteur

source de variation	degré de liberté	somme des carrés des écarts (SCE)	variances (carrés moyens)	F_{obs}	p-value
inter-groupes (facteur)	p-1	SCE_b	s_b^2	$\frac{s_b^2}{s_w^2}$	$\mathbb{P}_{H0}(F > F_{obs})$
intra-groupes (résidus)	n-p	SCE_w	s_w^2		
total	n-1	SCE_t	s_t^2		

- $s_b^2 = \frac{SCE_b}{p-1} = \frac{1}{p-1}\sum_{i=1}^{p} n_i(\bar{x}_i - \bar{x})^2$, p= nombre de niveaux de facteurs (i.e. nb de groupes)
- $s_w^2 = \frac{SCE_w}{n-p} = \frac{1}{n-p}\sum_{i=1}^{p}\sum_{k=1}^{n_i}(x_{ik} - \bar{x}_i)^2$
- si $H0$ est vraie, F_{obs} suit une loi $F_{(p-1,n-p)}$ ddl
- on rejète $H0$ si $\mathbb{P}_{H0}(F > F_{obs}) < \alpha$
- ATTENTION : $SCE_t = SCE_b + SCR_w$ mais $s_t^2 \neq s_b^2 + s_w^2$

Analyse de la variance (ANOVA)

ANOVA à un facteur : modèle théorique

l'ANOVA est un modèle linéaire de régression sur une variable catégorielle

- **le modèle fixe :**
  - **modèle théorique :** $X_{ik} = \mu_i + \epsilon_{ik} = \mu + \alpha_i + \epsilon_{ik}$
  - **où** α_i **: non aléatoires,** $\epsilon_{ik} \sim \mathcal{N}(0, \sigma)$
  - $H0 : \alpha_1 = \alpha_2 = ... = \alpha_p = 0$
- le modèle aléatoire :
  - modèle théorique : $X_{ik} = \mu + A_i + \epsilon_{ik}$
  - où $A_i \sim \mathcal{N}(0, \sigma_a)$, $\epsilon_{ik} \sim \mathcal{N}(0, \sigma)$
  - $H0 : \sigma_a^2 = 0$

le modèle fixe

- modèle théorique : $X_{ik} = \bar{x} + a_i + \epsilon_{ik}$
  - où a_i : non aléatoires, $\epsilon_{ik} \sim \mathcal{N}(0, \sigma)$
  - $\bar{x} = \frac{1}{n} \sum_{i=1}^{n} n_i m_i$, $\sum_{i=1}^{p} n_i a_i = 0$
- $H0 : a_1 = a_2 = ... = a_p = 0$
  - $\mathbb{E}(s_b^2) = \sigma^2 + \frac{1}{p-1} \sum_{i=1}^{p} n_i a_i^2$
  - $\mathbb{E}(s_w^2) = \sigma^2$
- si $H0$ est vraie, alors $\mathbb{E}(s_b^2) = \mathbb{E}(s_w^2) = s_t^2$

Comparaisons multiples de moyennes

- **localiser les inégalités entre moyennes suite à une analyse de variance** : quels sont les groupes qui possèdent des moyennes significativement différentes ?
- **conditions d'utilisation** :
  - populations normales, de même variance
  - échantillons aléatoires, simples et indépendants

- comparaison avec un témoin :
  - test de Dunnet : $LSD_{i1} = t_{dunnett(1-\frac{\alpha}{2})}\sqrt{s_w^2(\frac{1}{n_i}+\frac{1}{n_1})}$
  - si la différence de moyenne entre groupe i et 1 est supérieur à LSD_{i1}, on rejète $H0$ (égalité de moyennes entre i et 1)
- test LSD (Least Significant Difference) de Fisher :
  - $LSD_{ij} = t_{student(\nu;1-\frac{\alpha}{2})}\sqrt{s_w^2(\frac{1}{n_i}+\frac{1}{n_j})}$, avec $\nu = n - p$
  - si la différence de moyenne entre groupe i et j est supérieur à LSD_{ij}, on rejète $H0$ (égalité de moyennes entre i et j)
  - = test de student (t-test), en utilisant s_w^2 comme estimateur de σ^2
  - ne contrôle pas le risque global d'erreur α (voir plus loin)

Comparaisons multiples de moyennes

tests contrôlant le risque d'erreur α

ANOVA est un test global auquel est associé un risque global d'erreur α. Une fois que l'ANOVA a rejeté $H0$ on compare les moyennes 2 à 2. mais plus on compare de moyennes, plus on a de chances de rejeter $H0$ alors que $H0$ est vrai, simplement parce que la moyenne d'un échantillon sera, par pur hasard, très éloignée de la moyenne μ de l'unique distribution sous $H0$: la proportion de faux positifs (α) augmente, il faut donc effectuer une correction : $LSD_{ij} = M\sqrt{s_w^2(\frac{1}{n_i} + \frac{1}{n_j})}$, où M dépend de la méthode sélectionnée

- Bonferroni : $M = t_{student(1-\frac{\alpha'}{2};n-p)}$, avec $\alpha' = \frac{\alpha}{p(p-1)}$
  - risque de 1^{re} espèce $= \alpha$
  - très conservateur (risque β élevé, on rejette rarement $H0$)
- Tukey HSD (Honest Significant Difference) : $M = q_{\alpha,p,n-p}$
  - risque de 1^{re} espèce $\leq \alpha$
  - très conservateur (risque β élevé)
  - à préférer pour des comparaisons deux à deux
- Scheffé : $M = (p-1)F_{\alpha,p-1,n-p}$
  - risque de 1^{re} espèce $= \alpha$
  - très conservateur (risque β élevé)
  - à préférer pour des contrastes plus généraux

Comparaisons multiples de moyennes

tests contrôlant le risque d'erreur α

- méthode de Newman Keuls (multiple range test) :
  $d_{ij} = t_{NK(1-\frac{\alpha}{2})}\sqrt{s_w^2(\frac{1}{n_i}+\frac{1}{n_j})}$
  - effectifs égaux
  - calcule la plus petite amplitude significative
  - compare l'amplitude de sous-groupes à la plus petite amplitude significative
- méthode de Duncan (new multiple range test)
  - même principe que *NK*
  - risque de 1^{re} espèce plus élevé
- méthode de Ward
  - basée sur une méthode de classification numérique
  - maximise la somme des carrés des écarts entre groupes

Comparaison de plusieurs populations

test non paramétrique : test de Kruskall-Wallis

- généralisation du test de Mann-Whitney
- distributions non normales mais de même "forme"
- mêmes variances
- S_i = somme des rangs des observations du groupe i
- $H = \frac{12}{n(n+1)} \sum_{i=1}^{p} n_1 (\frac{S_i}{n_i} - \frac{n+1}{2}) \sim \chi^2_{p-1}$

Analyse de la variance (ANOVA)

ANOVA à deux facteurs : **le modèle fixe (fixed effect model)**

- modèle théorique : $X_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ijk}$
- où α_i, β_j et γ_{ij} : non aléatoires, $\epsilon_{ijk} \sim \mathcal{N}(0, \sigma)$
- $\bar{x}_{ij} = \frac{1}{n_{ij}} \sum X_{ij}$
- $\bar{x}_{i.} = \frac{1}{q} \sum_{j=1}^{q} \bar{x}_{ij}$
- $\bar{x}_{.j} = \frac{1}{p} \sum_{i=1}^{p} \bar{x}_{ij}$
- $\bar{x} = \frac{1}{pq} \sum_{i=1}^{p} \sum_{j=1}^{q} \bar{x}_{ij} (= \frac{1}{p} \sum_{i=1}^{p} \bar{x}_{i.} = \frac{1}{q} \sum_{j=1}^{q} \bar{x}_{.j})$
- hypothèses nulles :
  - effets principaux : $H0 : \alpha_1 = \alpha_2 = ... = \alpha_p = 0$, $H0' : \beta_1 = \beta_2 = ... = \beta_q = 0$
  - interaction : $H0 : \gamma_{11} = \gamma_{22} = ... = \gamma_{pq} = 0$
- si les hypothèses nulles sont vraies alors les ratios des sommes de carrés d'écarts suivent une loi F de Fisher-Snédecor

Analyse de la variance (ANOVA)

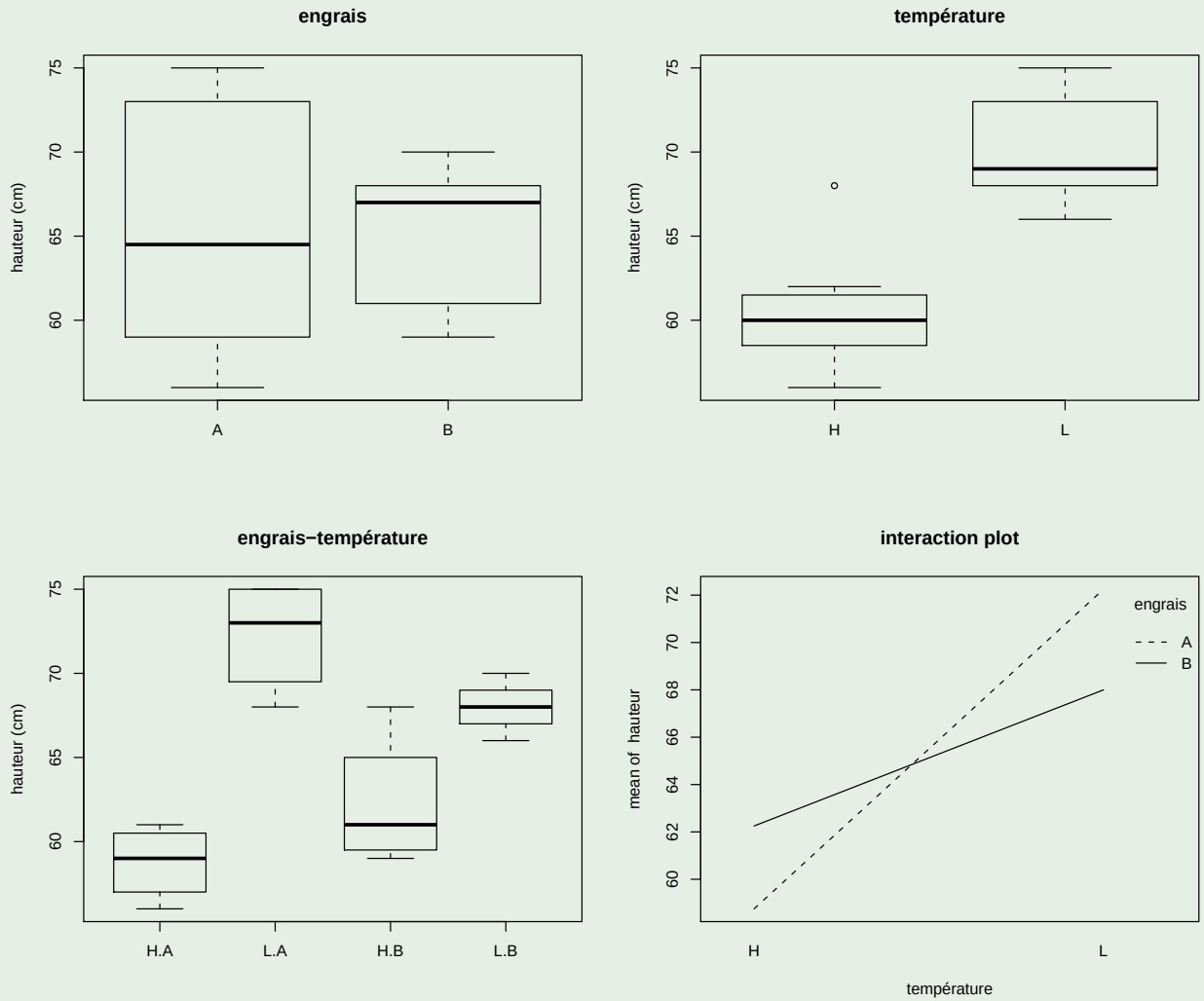
le tableau de l'ANOVA à deux facteurs (effets fixes)

source de variation	degré de liberté	somme des carrés des écarts (SCE)	variances (carrés moyens)	F_{obs}
facteur A	p-1	SCE_a	s_a^2	$\frac{s_a^2}{s_w^2}$
facteur B	q-1	SCE_b	s_b^2	$\frac{s_b^2}{s_w^2}$
interaction AxB	(p-1)(q-1)	SCE_{ab}	s_{ab}^2	$\frac{s_{ab}^2}{s_w^2}$
variation résiduelle	n-pq	SCE_w	s_w^2	
total	n-1	SCE_t	s_t^2	

- tester l'interaction (F_{AB})
- en l'absence d'interaction ($\mathbb{P}(F > F_{AB}) < \alpha$) on peut tester les effets principaux (F_A et F_B)

Analyse de la variance (ANOVA)

ANOVA à 2 facteurs : effet du type d'engrais (A et B) et des températures (L=low et H=high) sur la hauteur de plantes



Analyse de la variance (ANOVA)

ANOVA à 2 facteurs : effet du type d'engrais (A et B) et des températures (L=low et H=high) sur la hauteur de plantes

source de variation	degré de liberté	somme des carrés des écarts (SCE)	variances (carrés moyens)	F_{obs}	p-value
température	1	370.56	370.56	41.85	$3.07 x 10^{-5}$
engrais	1	0.56	0.56	0.06	0.80
interaction	1	60.06	60.06	6.78	0.023
intra-groupes (résidus)	12	106.25	8.85		

- p-valeurs tests de Student comparaison 2 à 2 : $0.00056, 0.03833, 0.1790, 0.06529$ pour "AL vs AH","BL vs BH","BH vs AH","BL vs AL"
- p-valeurs corrigées pour comparaisons multiples de moyenne : $0.0002, 0.0545, 0.1326, 0.1326$
- il n'y a pas d'effet global du facteur "engrais", il y a un effet "température" mais essentiellement en présence de l'engrais "A" (interaction)

Analyse de la variance (ANOVA)

ANOVA à deux facteurs : le modèle aléatoire (random effect model)

- 2 facteurs aléatoires A et B (ex : différents appareils de mesure et différents opérateurs)
- modèle théorique : $X_{ijk} = \mu + A_i + B_j + C_{ij} + \epsilon_{ijk}$
- où $A_i \sim \mathcal{N}(0, \sigma_a)$, $B_j \sim \mathcal{N}(0, \sigma_b)$, $C_{ij} \sim \mathcal{N}(0, \sigma_{ab})$, et $\epsilon_{ijk} \sim \mathcal{N}(0, \sigma)$

source de variation	degré de liberté	somme des carrés des écarts (SCE)	variances (carrés moyens)	F_{obs}
facteur A	p-1	SCE_a	s_a^2	$\frac{s_a^2}{s_{ab}^2}$
facteur B	q-1	SCE_b	s_b^2	$\frac{s_b^2}{s_{ab}^2}$
interaction AxB	(p-1)(q-1)	SCE_{ab}	s_{ab}^2	$\frac{s_{ab}^2}{s_w^2}$
variation résiduelle	n-pq	SCE_w	s_w^2	
total	n-1	SCE_t	s_t^2	

- tester l'interaction : $F_{AB} = \frac{s_{ab}^2}{s_w^2}$
- en l'absence d'interaction on peut tester les effets principaux $F_A = \frac{s_a^2}{s_{ab}^2}$ et $F_B = \frac{s_b^2}{s_{ab}^2}$
- $\mathbb{E}(s_a^2) = \sigma^2 + nq\sigma_a^2 + n\sigma_{ab}^2$
- $\mathbb{E}(s_b^2) = \sigma^2 + np\sigma_b^2 + n\sigma_{ab}^2$
- $\mathbb{E}(s_{ab}^2) = \sigma^2 + n\sigma_{ab}^2$, $\mathbb{E}(s_w^2) = \sigma^2$

Analyse de la variance (ANOVA)

ANOVA à deux facteurs : **le modèle mixte (mixed effect model)**

- un facteur fixe A, un facteur aléatoire B (ex : 4 traitements - effet fixe "traitement", 3 répétitions biologiques - effet aléatoire "individu", 2 analyses par échantillon)
- modèle théorique : $X_{ijk} = \mu + \alpha_i + B_j + C_{ij} + \epsilon_{ijk}$
- où α_i non aléatoire, $B_j \sim \mathcal{N}(0, \sigma_b)$, $C_{ij} \sim \mathcal{N}(0, \sigma_{ab})$, et $\epsilon_{ijk} \sim \mathcal{N}(0, \sigma)$

source de variation	degré de liberté	somme des carrés des écarts (SCE)	variances (carrés moyens)	F_{obs}
facteur A (fixe)	p-1	SCE_a	s^2_a	$\frac{s^2_a}{s^2_{ab}}$
facteur B (aléatoire)	q-1	SCE_b	s^2_b	$\frac{s^2_b}{s^2_w}$
interaction AxB	(p-1)(q-1)	SCE_{ab}	s^2_{ab}	$\frac{s^2_{ab}}{s^2_w}$
variation résiduelle	n-pq	SCE_w	s^2_w	
total	n-1	SCE_t	s^2_t	

- tester l'interaction : $F_{AB} = \frac{s^2_{ab}}{s^2_w}$
- en l'absence d'interaction on peut tester les effets principaux $F_A = \frac{s^2_a}{s^2_{ab}}$ (effet fixe vs interaction)

  et $F_B = \frac{s^2_b}{s^2_w}$ (effet aléatoire vs erreur résiduelle)

Analyse de la variance (ANOVA)

ANOVA à deux facteurs : **modèle à facteurs imbriqués - hiérarchisé (nested model)**

- un facteur fixe ou aléatoire A
- un facteur subordonné, généralement aléatoire, $B(A)$
- exemple : un facteur fixe "type de prairie", un facteur subordonné "la prairie"
- modèle théorique : $X_{ijk} = \mu + \alpha_i + B_{j(i)} + \epsilon_{ijk}$
- test des effets de chaque facteur :
  - $F_A = \frac{s_a^2}{s_{b(a)}^2}$
  - $F_B = \frac{s_{b(a)}^2}{s_w^2}$

Test du rapport de vraisemblance (likelihood ratio test)

- au lieu du test F on peut utiliser le LRT pour tester $H0 : \alpha_1 = \alpha_1 = ... = \alpha_p = 0$
- ce test est utilisé pour comparer 2 modèles "emboîtés" (pas des facteurs emboîtés !)
- modèle1 : $y_i = \alpha_0 + e_i$
- modèle1 : $y_i = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + ... + \alpha_p x_p + e_i$
- la log-vraisemblance est définie par $l(\mu_i, y_i) = log[f_Y(y_i, \theta)]$
- la déviance entre les modèles 1 et 2 est définie par $D(y, \mu) = -2[l(\mu, y) - l(y, y)]$ (i.e. la différence entre le modèle estimé et le modèle saturé - autant de paramètres que d'observations)
- si un modèle M_1 est imbriqué dans un autre modèle M_2, alors la différence des déviances $D_1 - D_2$ suit un $\chi^2_{p-2-p_1}$
- pour tester un seul coefficient du modèle, on utilise $D = -2log \frac{vraisemblancedumodelsanslavariable}{vraisemblancedumodelaveclavariable}$ qui suit un χ^2_1
- pour comparer des modèles non emboîtés :
  - AIC (Akaike Information Criterion) : $AIC = 2k - 2ln(L)$, avec k le nombre de paramètres du modèle théorique, et L la valeur maximum de la vraisemblance pour le modèle estimé ("fit" du modèle estimé par rapport au modèle théorique)
  - BIC (Bayesian Information Criterion) : $BIC = -2ln(L) + kln(n)$

Extension de l'analyse de la variance

Analyse de la covariance (ANCOVA)

- a pour but d'étudier **l'effet d'une variable qualitative (facteur) et d'une variable continue (covariable) sur une variable réponse** (comparaisons de moyennes en tenant compte des effets d'une variable auxiliaire -covariable- quantitative)
- modèle théorique : $Y_{ik} = \mu + \alpha_i + \beta(X_{ik} - \bar{x}) + \epsilon_{ijk}$
- où Y = variable étudiée et X la covariable

Modèle linéaire généralisé (GLM)

- les modèles linéaires généralisés permettent d'étudier la liaison entre une variable réponse Y et un ensemble de variables explicatives ou prédicteurs $X_1, X_2 \ldots X_K$
- ils englobent :
  - le modèle linaire général (régression multiple, ANOVA, ANCOVA)
  - le modèle log-linéaire
  - la régression logistique
  - la régression de Poisson
- formule générale : $y = X\beta + \epsilon$, où :
  - y = vecteur des observations
  - X = matrice des valeurs des variables explicatives (régression linéaire) ou indicatrices (analyse de variance)
  - β = vecteur des paramètres du modèle
  - ϵ = vecteur des termes d'erreurs

Notion de correction pour les tests multiples

comparer simultanément l'expression de milliers de gènes entre 2 conditions

- si chaque test est réalisé à un risque α
- si l'on réalise G test indépendants (pour les G gènes)
- ALORS nombre moyen de faux positifs $= G\alpha$
- $G = 10000$, $\alpha = 0.05$, alors 500 gènes sont déclarés différentiellement exprimés à tort
- d'où l'idée de contrôler les risques du test multiple dont $H0$="tous les gènes n'ont pas de différence d'expression entre les traitements", et $H1$="il existe au moins un gène différentiellement exprimé"
- après un test multiple : P gènes sont déclarés différentiellement exprimés (dont des faux positifs) et N gènes sont déclarés non différentiellement exprimés (dont des faux négatifs)

Notion de correction pour les tests multiples

	déclaré non diff. exp.	déclaré diff. exp.
m_0 gènes non diff. exp.	vrais négatifs	faux positifs
m_1 gènes diff. exp.	faux négatifs	vrais positifs
$G = m_0 + m_1$	N	P

- test simple d'hypothèse : on cherche à contrôler le risque α
- test multiple : on cherche à contrôler une fonction du nombre de faux positifs
- Contrôle du Family-wise error rate (FWER) :
  - FWER = probabilité d'avoir au moins un faux positif (sur l'ensemble des gènes testés)
  - procédure de Bonferroni la plus connue. Si $G = 10000$ et $FWER \leq 0.05$, alors chaque test est réalisé avec un risque $\alpha' = 5 * 10^{-6}$ $(\frac{\alpha}{G})$
  - mais plus il y a de tests, moins on rejète $H0$
  - et problème de non indépendance des tests (corégulation des gènes) : procédure de Westfall and Young
  - très conservatifs (peu de gènes sont déclarés différentiellement exprimés), plutôt recherche de candidats pour des analyses fonctionnelles
- False Discovery Rate (FDR) : proportion moyenne de faux positifs
  - l'idée est plutôt d'avoir une idée de l'espérance du nombre de faux-positifs, parmi les gènes déclarés différentiellement exprimés
  - si $\alpha = 0.05$, 5% des gènes significatifs après correction seront des faux positifs
  - moins conservatif (plus de faux-positifs), utile quand l'objectif de l'expérience transcriptomique est exploratoire

Régression linéaire simple

test de corrélation de Spearman

- coefficient de corrélation de Spearman : $\rho = 1 - \frac{6 \sum_{i=1}^{n} d_i^2}{n(n^2-1)}$, où $d_i = x_i - y_i$
- question : à partir de quelles valeurs on va considérer qu'il y a dépendance ou indépendance entre x et y ?
- le test est non paramétrique, mais pour $n > 10$ on a $t = \rho\sqrt{\frac{n-2}{1-\rho}}$ qui suit une loi de Student $\mathcal{T}(n-2)$
- $H0$: il n'y a pas de corrélation des rangs
- $H1$: il existe une corrélation "monotone"

Régression linéaire simple

tests sur les paramètres

- rappel : la régression linéaire permet d'estimer les paramètres $\hat{a}$ et $\hat{b}$ de la droite de régression de manière à ce que $y_i = \hat{a}x_i + \hat{b} + \epsilon_i$
- on veut valider le modèle $\hat{y}_i = ax_i + b$, donc on veut tester si $\hat{a}$ et $\hat{b}$ sont significativement différents de 0 (hypothèse $H0$)
- on dispose des variances de chaque estimateur :
  - $s^2_{\hat{a}} = \frac{\sigma^2}{\sum_{x=1}^{n}(x_i - \bar{x})^2}$, où $\sigma^2 = \frac{\sum_{i=0}^{n}(y_i - \hat{y}_i)^2}{n-2}$
  - $s^2_{\hat{b}} = \frac{\sigma^2 \sum_{i=0}^{n} x_i^2}{n^2 s_x^2}$
- test de Student bilatéraux :
  - pour $\hat{a}$: $t = \frac{\hat{a}}{\frac{s_{\hat{a}}}{\sqrt{n}}}$
  - pour $\hat{b}$: $t = \frac{\hat{b}}{\frac{s_{\hat{b}}}{\sqrt{n}}}$

Régression linéaire multiple

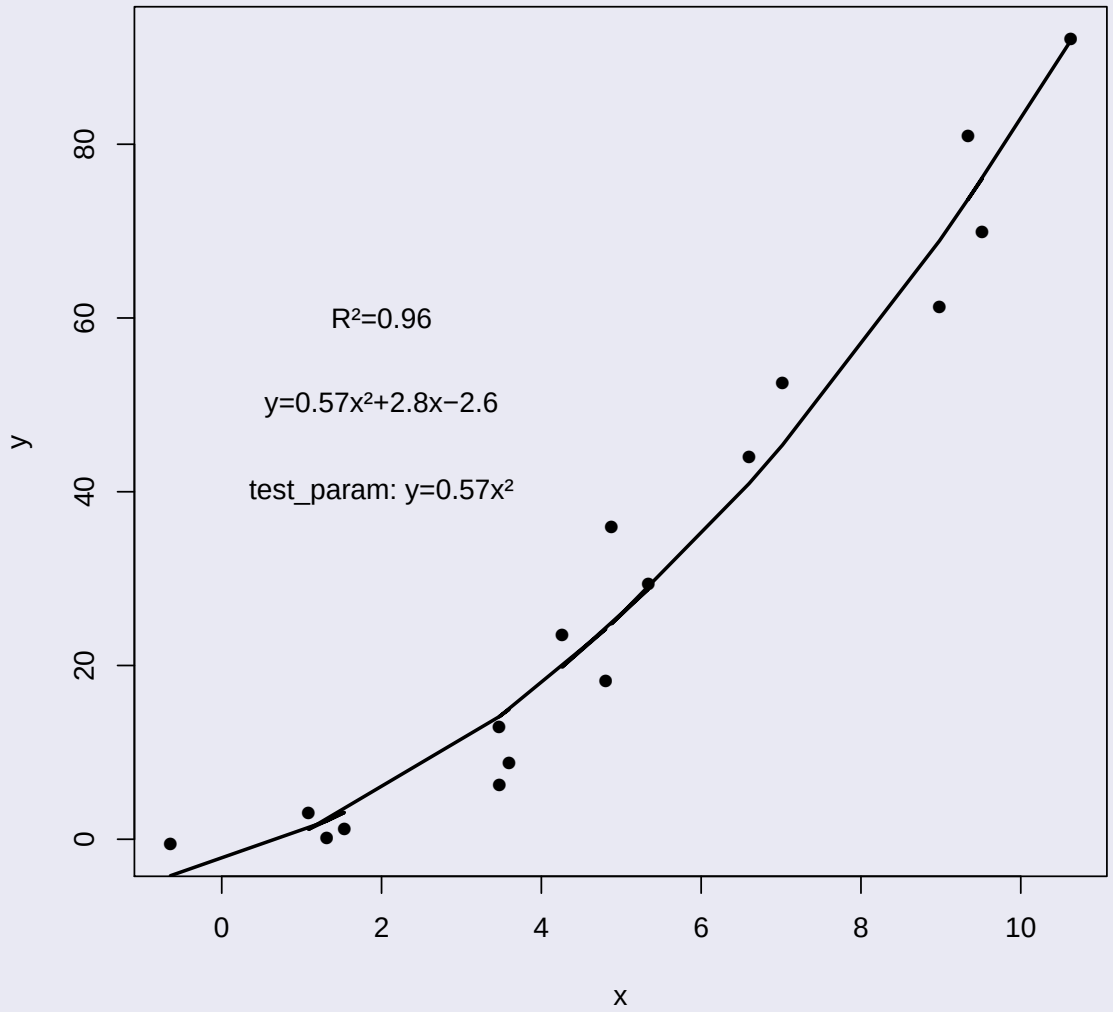
- modèle théorique : $Y_i = \beta_0 + \sum_{j=1}^{p} \beta_j X_{ij} + \epsilon_i$ où ϵ_i indépendante $\sim \mathcal{N}(0, \sigma)$
- exemple : relation entre le poids des rejets de chicorée (endives) et le poids des feuilles et le poids des racines : $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$
- les valeurs des paramètres β sont estimées par la méthode des moindres carrés : $min \sum_{i=1}^{n} (y_i - b_0 - b_1 x_{i1} - ... - b_p x_{ip})^2$
- Les coefficients b_i sont appelés coefficients de régression partielle
- distributions d'échantillonnage :
  - variance résiduelle : $\chi^2_{(n-p-1)}$
  - coefficients de régression : loi de Student $\mathcal{T}(n-p-1)$
  - coefficients de détermination (part de la variance expliquée par la régression) et de corrélation multiple (corrélation entre Y observées et Y estimées) : $F(p, n-p-1)$

Régression logistique sur données binaires (modèle linéaire !)

- un modèle permettant de prédire les valeurs prises par une variable catégorielle, le plus souvent binaire, à partir d'une série de variables explicatives continues et/ou binaires.
- exemple : trouver les facteurs qui caractérisent un groupe de plantes malades par rapport à plantes saines.
- Y prends les valeurs 0 ou 1
- $p(X|1)$ (resp. $p(X|0)$) est la distribution conditionnelle des X sachant la valeur prise par Y
- $p(1|X)$ est la probabilité a posteriori d'obtenir la modalité 1 de Y sachant la valeur prise par X
- hypothèse fondamentale : $ln\frac{p(X|1)}{p(X|0)} = a_0 + a_1x_1 + ... + a_jx_j$
- modèle LOGIT de $p(1|X)$: $ln\frac{p(1|X)}{1-p(1|X)} = b_0 + b_1x_1 + ... + b_jx_j$
- régression logistique car loi de probabilité logistique : $p(1|X) = \frac{e^{b_0+b_1x_1+...+b_jx_j}}{1+e^{b_0+b_1x_1+...+b_jx_j}}$
- estimation des b_j par maximum de vraisemblance

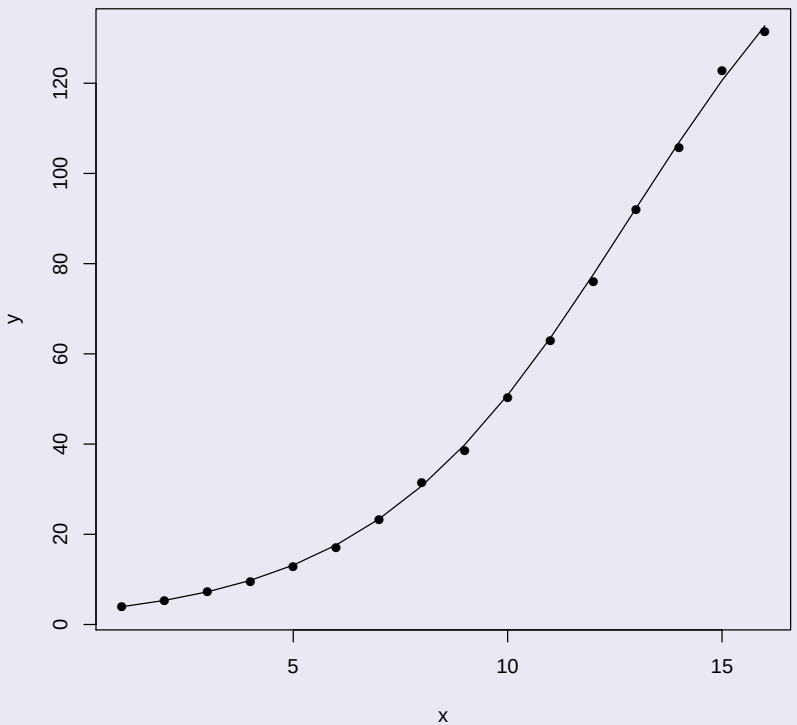
Régression curvilinéaire -polynomiale- (modèle linéaire !)

- **régression polynomiale** : $y_i = \beta_0 + \beta_1 x + \beta_2 x^2 + ... + \beta_n x^n + \epsilon_i$



Régression non-linéaire

exemple du modèle logistique généralisé



- $Y = (a + bc^X)^d$

Analyse en Composantes Principales - ACP

- décrire et analyser des données multivariées : tableau individus (n) x caractères (p)
- représenter graphiquement
  - ressemblances entre individus (notion de distance)
  - corrélations entre variables
- représentations planes d'un ensemble de points (individus) contenus dans R^p ou d'un ensemble de vecteurs (variables) dans R^n qui maximisent l'information (variance)

- transformer des variables "corrélées" en nouvelles variables indépendantes les unes des autres
- les "composantes principales", ou axes, réduisent l'information en un nombre plus petit de variables ("composantes")

- changement de repère dans l'espace des individus : recherche des axes principaux (ou axes factoriels) permettant d'obtenir les "meilleures" représentations planes des distances entre individus
- plan factoriel : défini par deux axes factoriel
- composante principale : combinaison linéaire des variables associée à un axe factoriel
- inertie d'un axe factoriel = variance de la composante principale associée (représente l'information contenue dans la composante)
- ACP normée : variables normées avant l'analyse

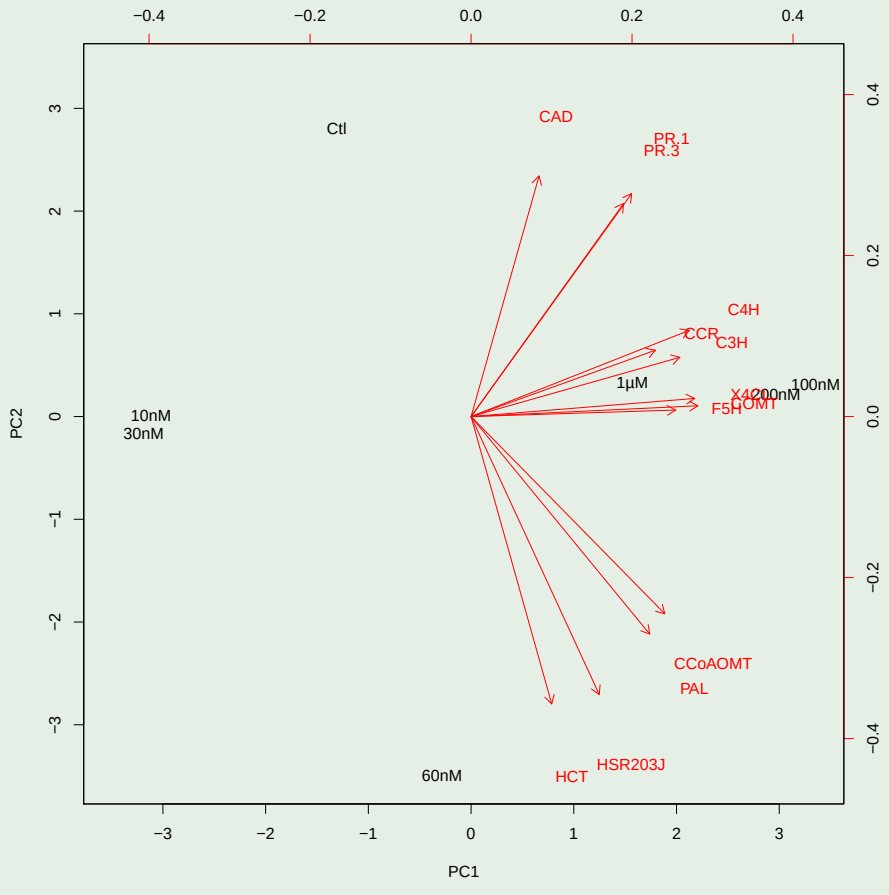
Analyse en Composantes Principales - ACP

exemple : expression de gènes dans différentes conditions expérimentales

	4CL	C3H	C4H	CAD	CCoAOMT	CCR	COMT	F5H	HCT	HSR203J	PAL	PR.1	PR.3
Ctl	2.131	1.386	1.648	1.171	10.059	1.372	1.429	2.127	3.094	0.475	2.371	1.000	1.000
10nM	1.755	0.692	1.000	0.078	11.328	0.058	0.263	9.102	1.468	15.767	1.687	0.018	0.034
30nM	1.040	0.267	1.000	0.018	16.531	0.135	0.516	5.152	2.486	16.361	2.687	0.006	0.011
60nM	3.532	1.066	0.978	0.181	204.626	0.687	2.538	28.282	50.953	121.505	31.562	0.105	0.165
100nM	14.992	6.349	4.159	0.436	185.948	6.264	7.908	134.845	17.109	55.242	25.630	0.970	0.615
200nM	10.207	3.683	4.582	0.602	172.678	2.619	9.849	216.955	16.464	80.688	21.202	0.653	0.957
1µM	8.448	5.837	2.849	0.307	128.631	7.271	5.157	66.978	12.825	54.408	18.717	0.672	0.624

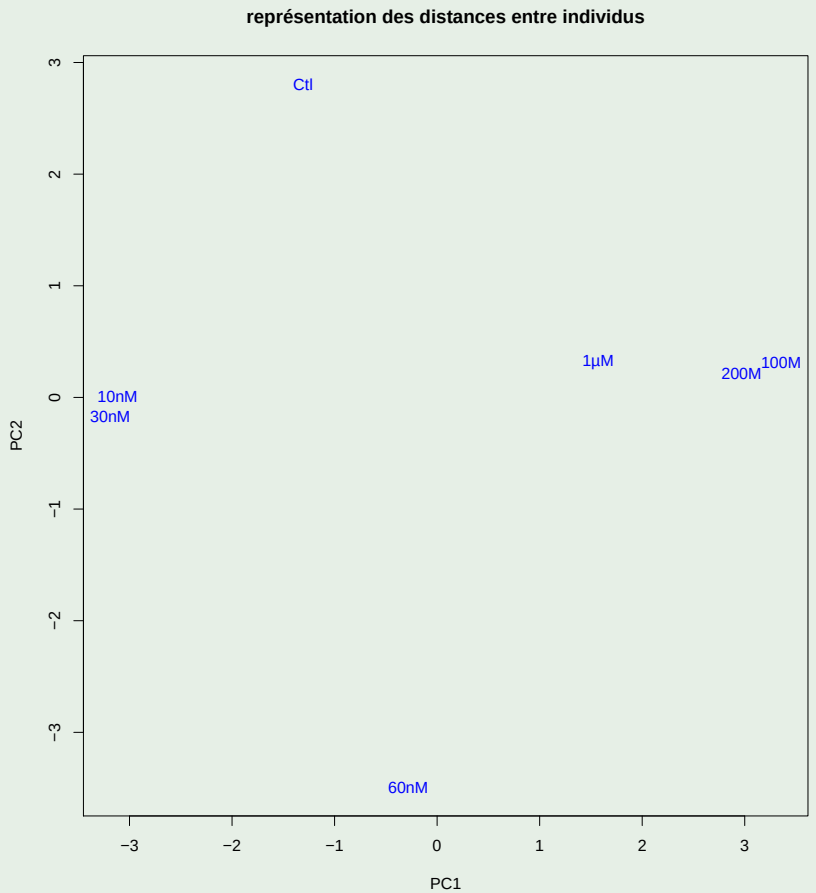
Analyse en Composantes Principales - ACP

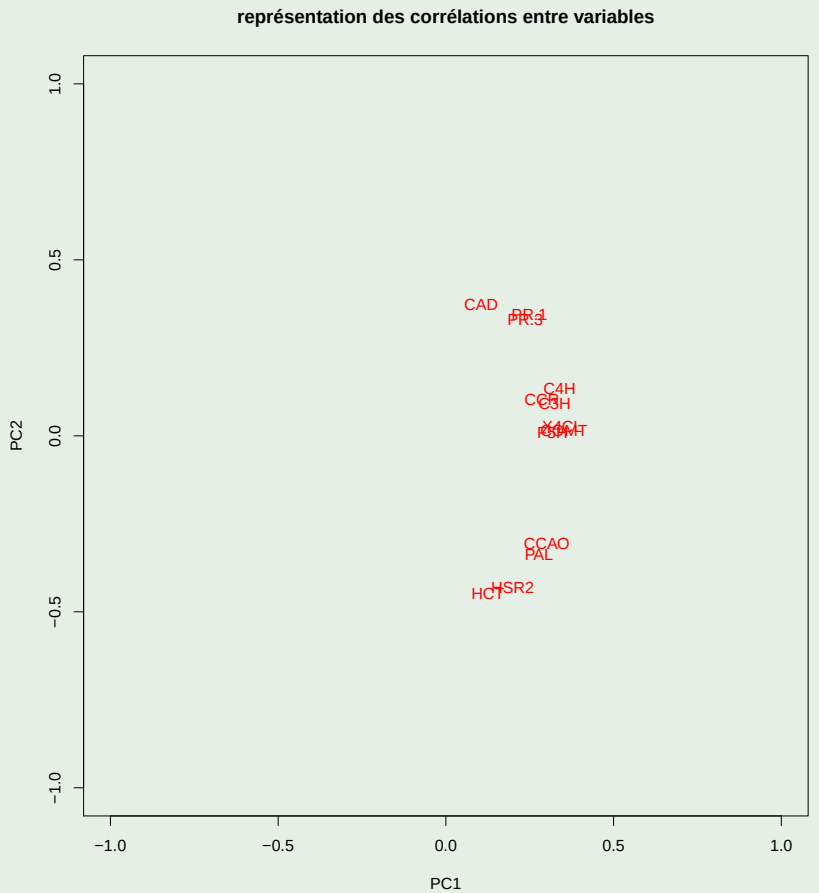
exemple : expression de gènes dans différentes conditions expérimentales



Analyse en Composantes Principales - ACP

exemple : expression de gènes dans différentes conditions expérimentales

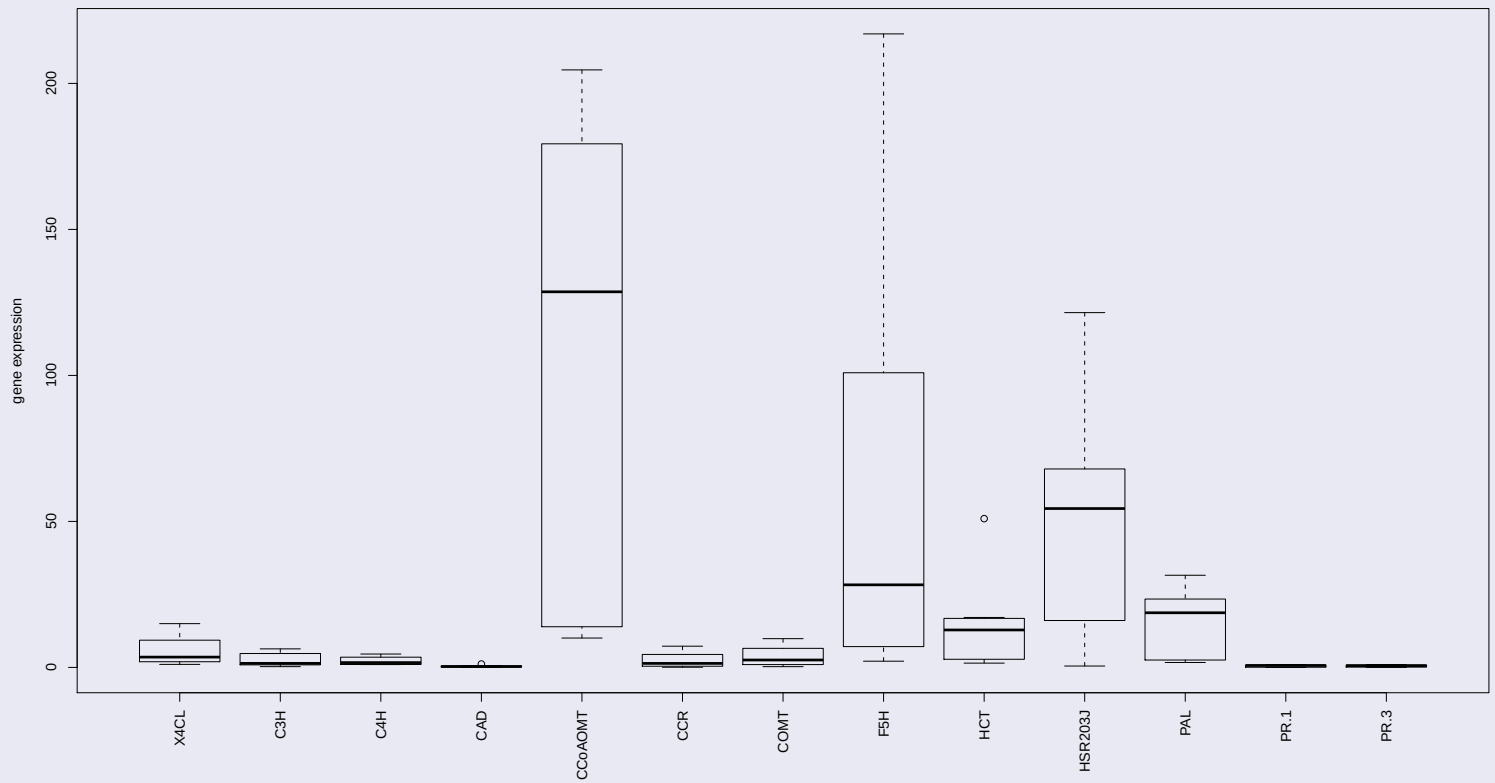




Analyse en Composantes Principales - ACP

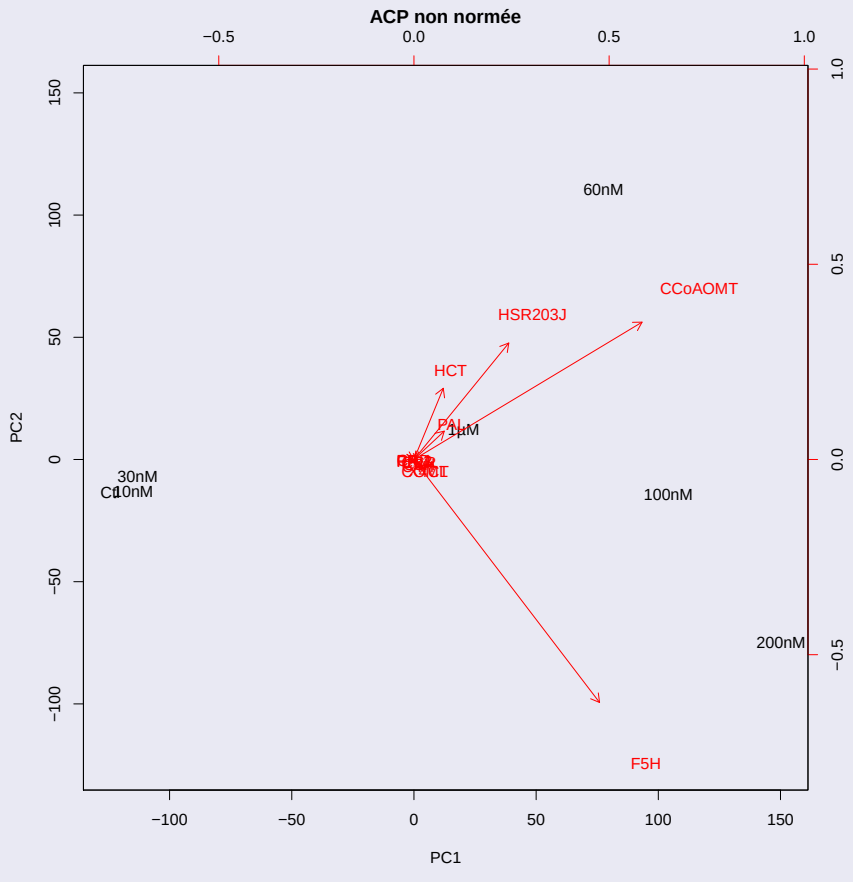
ACP normée et non normée

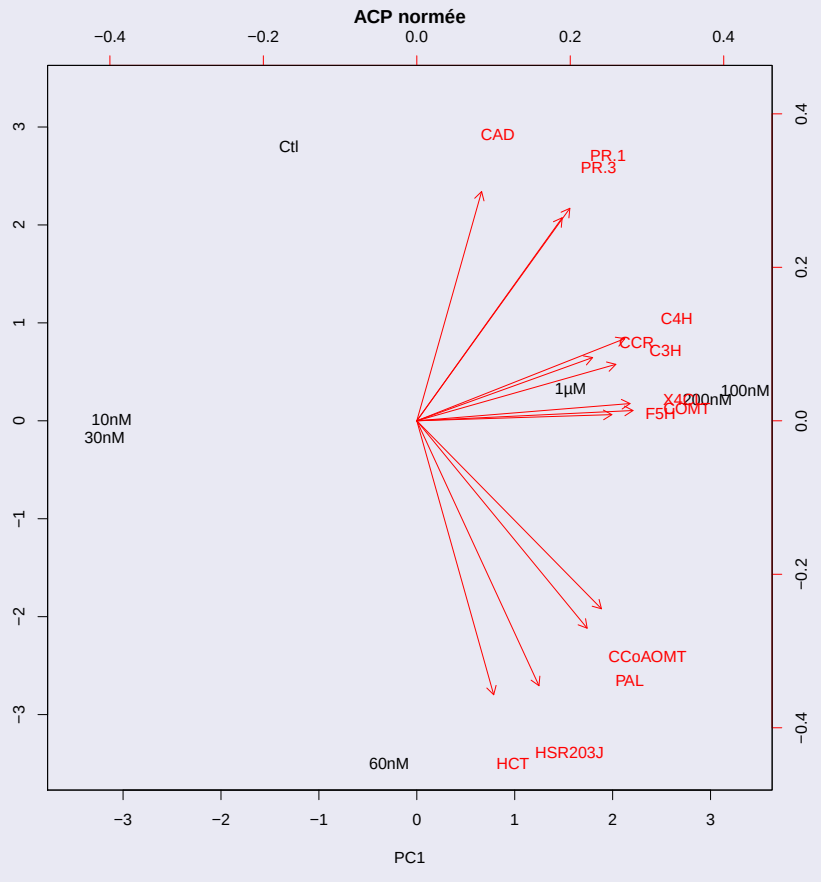
- la variance de l'expression diffère d'un gène à l'autre : faire une ACP normée



Analyse en Composantes Principales - ACP

ACP normée et non normée



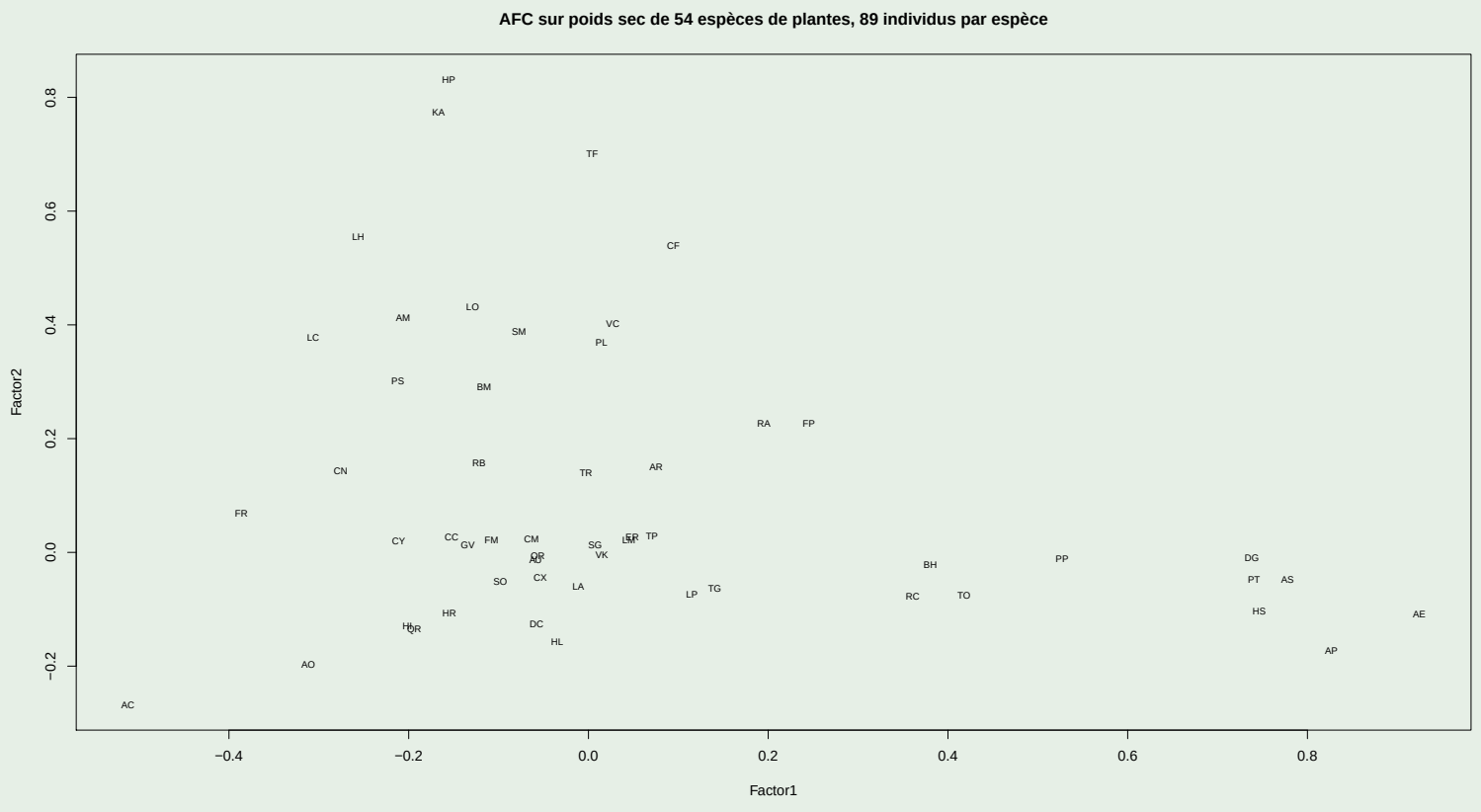


Analyse Factorielle des Correspondances - AFC

ACP normée et non normée

- pourquoi des "correspondances" ?
  - variables numériques : corrélation (ACP)
  - variables nominales : correspondance (AFC)
  - l'AFC s'utilise avec des variables qualitatives qui possèdent 2 ou plus de 2 modalités. Elle offre une visualisation en deux dimensions des tableaux de contingences.
  - il faut savoir que l'AFC peut être vue comme une ACP avec une distance particulière : la distance du χ^2
- pourquoi "factorielle" ?
  - il s'agit de décomposer le tableau original en une somme de tableaux/matrices qui sont chacun le produit de facteurs simples.
  - autrement dit, on les « met en facteurs »
- on peut faire une AFC avec des données permettant de faire une ACP (on traite les valeurs numériques comme des catégories)
- on ne peut pas faire une ACP avec des données permettant de faire une AFC

Analyse Factorielle des Correspondances - AFC

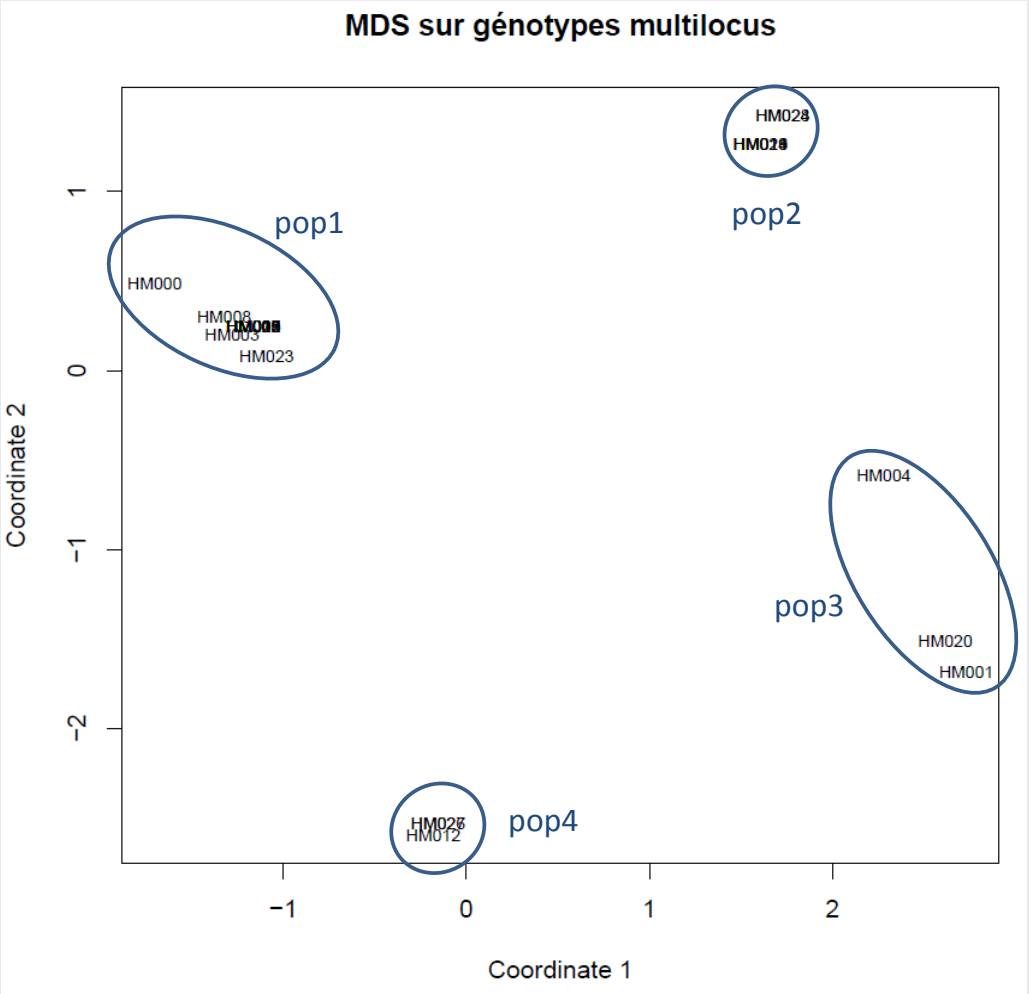


sur facteur 1 : forte corrélation positive pour AE, AP et AS, et forte corrélation négative pour AC, AO et FR

ACP sur tableau de distances : "multidimensional scaling" - MDS

MDS

- explorer la (dis)similarité entre observations (ex : génotypes multi-locus de plusieurs individus)



Références

 *Probability and statistical inference. Hogg RV, Tanis EA*

 *Probabilités - Estimation statistique. Lethielleux M, Edition Express*