Traitement des Données Biologiques

Master 1 Biologie Végétale (parcours BPMA) et Master 1 Bioinformatique (parcours BBS)

Maxime Bonhomme

Laboratoire de Recherche en Sciences Végétales

Traitement des Données Biologiques

Traitement des Données Biologiques : démarche et définitions

Démarche générale du biologiste

- je me pose une(des) question(s) sur mon système biologique
- **je réalise un expérience et/ou je collecte des données**
- **je décris les données récoltées**
- **j'utilise une méthode pour répondre à la question de départ**
- j'interprète

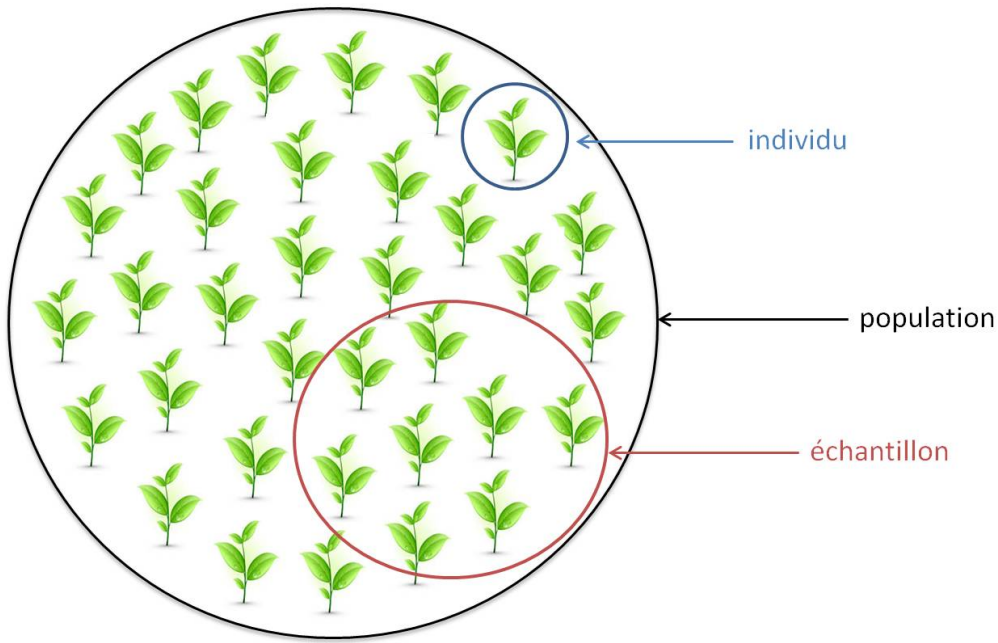
Démarche statistique parallèle

- **collecter des données : plan d'expérience et échantillonnage**
- **décrire les données : graphiques et petits calculs**
- **tester des hypothèses = tests statistiques**

Exemples de problèmes

- comparer des lignées pour un caractère (ex : analyse de mutants), effet d'un traitement (ex : un stress),..
- comparer l'expression de gènes dans différentes conditions, rechercher des corrélations entre phénotype et génotype ...

Quelques définitions



- **échantillon** : sous-ensemble de la population sur lequel sont effectuées les observations.
- **effectif** : nombre total d'individus d'une population ou de l'échantillon
- **variable ou caractère** : propriété étudiée sur les individus
  - **qualitatif : couleur, forme, sexe, direction**
  - **quantitatif discret (dénombrable) : nombre de racines latérales,...**
  - **quantitatif continu : taille, poids, concentration, temps, expression,...**

Quelques définitions

- **plan d'expérience** : dispositif expérimental permettant la collecte des données en vue de répondre à une question donnée. Il est associé à la méthode statistique utilisée pour analyser les données
  - dispositif cas-contrôle simple (un groupe traité, un non traité)
  - plans factoriels croisés : croiser systématiquement toutes les modalités de tous les facteurs expérimentaux

	traitement B (n=200)	contrôle B (n=200)
traitement A (n=200)	traitement A traitement B (n=100)	traitement A contrôle B (n=100)
contrôle A (n=200)	contrôle A traitement B (n=100)	contrôle A contrôle B (n=100)

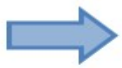
  - plans expérimental en blocs aléatoires complets -PEBAC- : effet de différents traitements entre unités expérimentales, en champs (très utilisé en agronomie). Le but est de réduire l'erreur expérimentale en éliminant la contribution de sources connues de variation entre les unités expérimentales

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

Quelques définitions

- **organisation d'un tableau de données** :

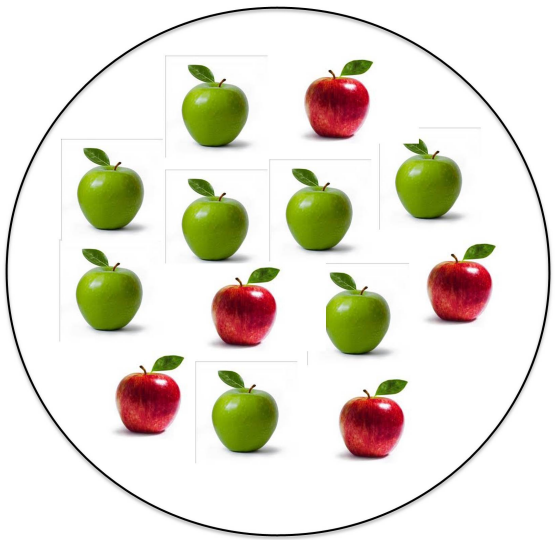
	Variable 1	Variable 2	Variable 3	Variable4
individu 1				
individu 2				
individu 3				
individu 4				
individu 5				
individu 6				
individu 7				
individu 8				
individu 9				
individu 10				
individu 11				
individu 12				
individu 13				
individu 14				
individu 15				
individu 16				
individu 17				
individu 18				
individu 19				
individu 20				
...				



	hauteur	temps_floraison	écotype
individu 1	22.6	20.3	Col-0 (USA)
individu 2	11.3	23.8	Col-0 (USA)
individu 3	18.6	41.2	Col-0 (USA)
individu 4	18.0	18.4	Col-0 (USA)
individu 5	21.7	40.9	Col-0 (USA)
individu 6	15.1	20.7	Ws-0 (Russia)
individu 7	17.3	23.1	Ws-0 (Russia)
individu 8	16.4	46.7	Ws-0 (Russia)
individu 9	23.6	19.7	Ws-0 (Russia)
individu 10	22.3	44.1	Ws-0 (Russia)
individu 11	16.2	31.3	Can-0 (Canary Isles)
individu 12	19.9	26.0	Can-0 (Canary Isles)
individu 13	22.1	35.4	Can-0 (Canary Isles)
individu 14	19.9	32.4	Can-0 (Canary Isles)
individu 15	19.3	37.1	Can-0 (Canary Isles)
individu 16	20.7	33.6	Edi-0 (Scotland)
individu 17	23.1	28.7	Edi-0 (Scotland)
individu 18	23.2	41.6	Edi-0 (Scotland)
individu 19	19.4	56.5	Edi-0 (Scotland)
individu 20	18.7	26.8	Edi-0 (Scotland)
...	...	...	...

Comment décrire les données : STATISTIQUES DESCRIPTIVES

Variable qualitative

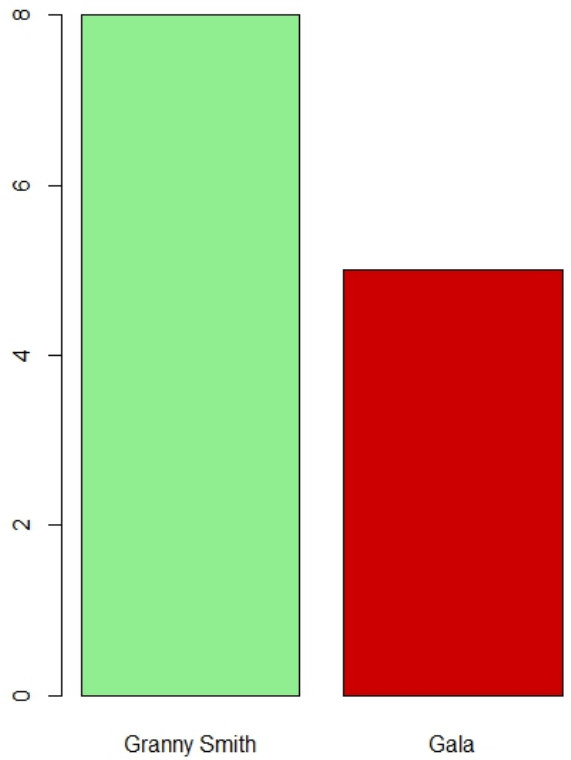


- fréquence (effectif) absolue : nombre d'observations par catégorie (n_i)
- fréquences relatives : proportion d'observations de la catégorie par rapport à l'ensemble p de catégories

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

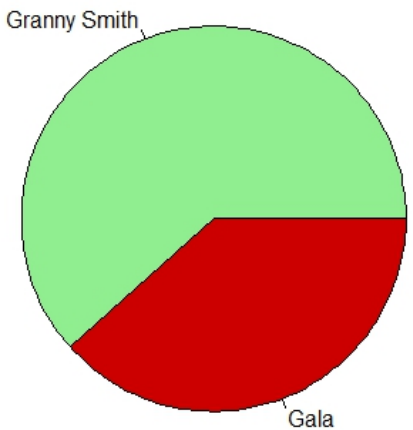
Variable qualitative

représentation : diagramme en barres

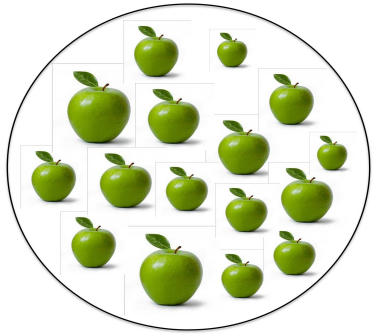


Variable qualitative

représentation : camembert



Variable quantitative



- **répartition en classes**
- fréquence (effectif) absolue : nombre d'observations par classe (n_i)
- fréquences relatives :

$$f_i = \frac{n_i}{\sum_{i=1}^{p} n_i} \tag{2}$$

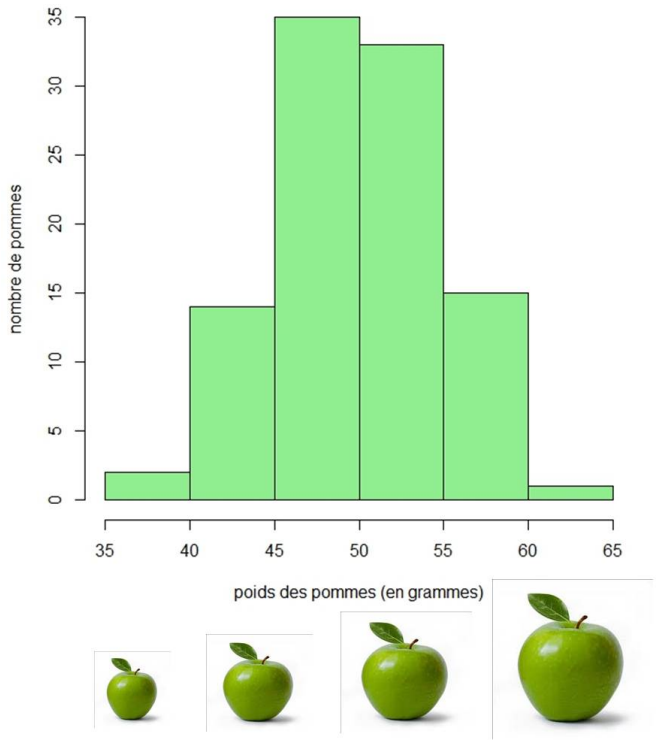
- fréquences cumulées, pour k de catégories (la somme sur toutes les catégories = 1) :

$$F_k = \sum_{i=1}^{k} f_i \tag{3}$$

Variable quantitative

représentation : histogramme

- graphique représentant la répartition des valeurs d'une variable (effectifs par classes de valeurs)



Variable quantitative

paramètres de tendance d'une distribution de valeurs (x_i, i = 1,...,n)

- **moyenne** :

$$m = \frac{1}{n} \sum_{i=1}^{n} x_i \tag{4}$$

$$m = \sum_{i=1}^{n} x_k p_k \tag{5}$$

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 de valeurs) la plus fréquente

Variable quantitative

paramètres de tendance : autres moyennes

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

$$m = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i} \tag{6}$$

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

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

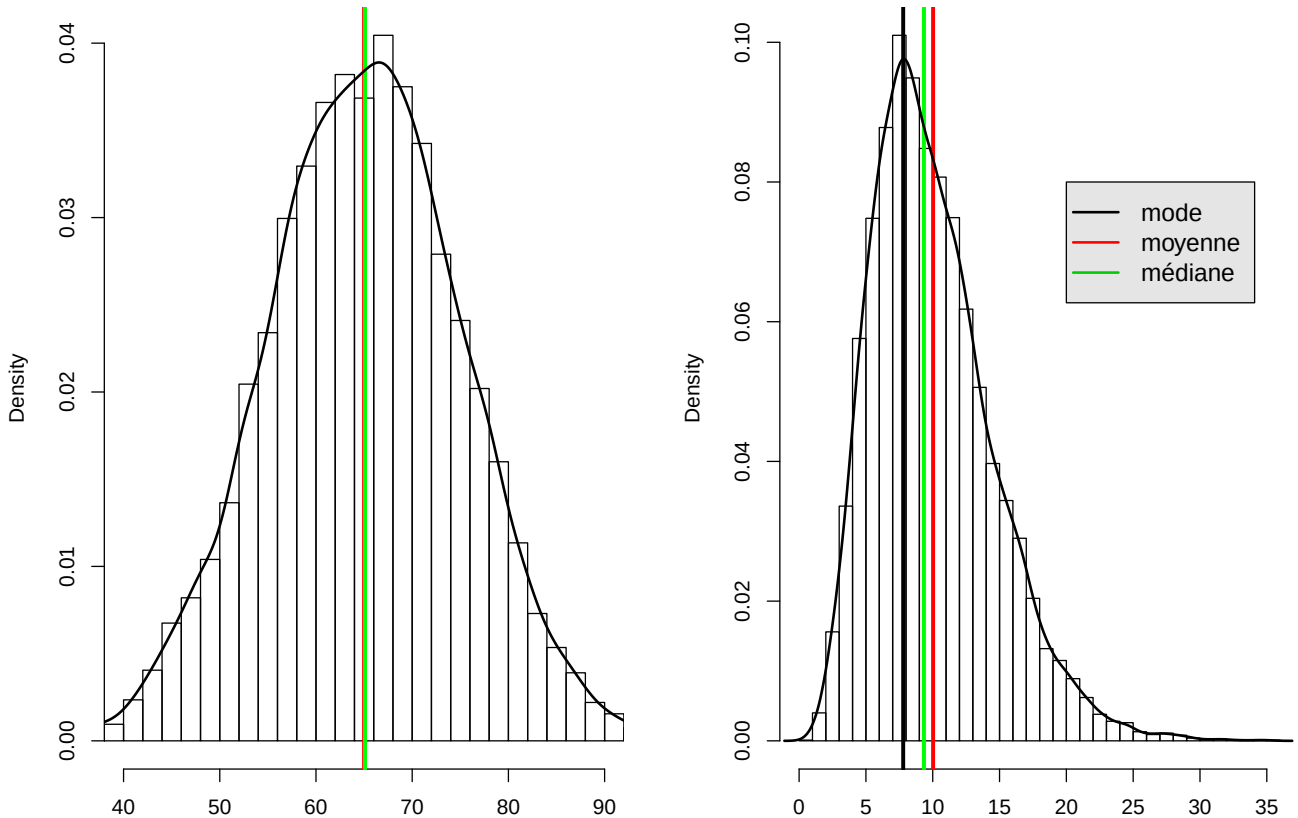
- **moyenne géometrique**, si multiplicatif ou cumulatif (ex : carré et rectangle de même surface) :

$$m = \sqrt[n]{\prod_{i=1}^{n} x_i} \tag{8}$$

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

Variable quantitative

paramètres de tendance : visualisation



Variable quantitative

paramètre de dispersion d'une distribution de valeurs (x_i, i = 1,...,n)

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

$$s_n^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2 \tag{9}$$

(10)

- valable que si on connaît 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) :

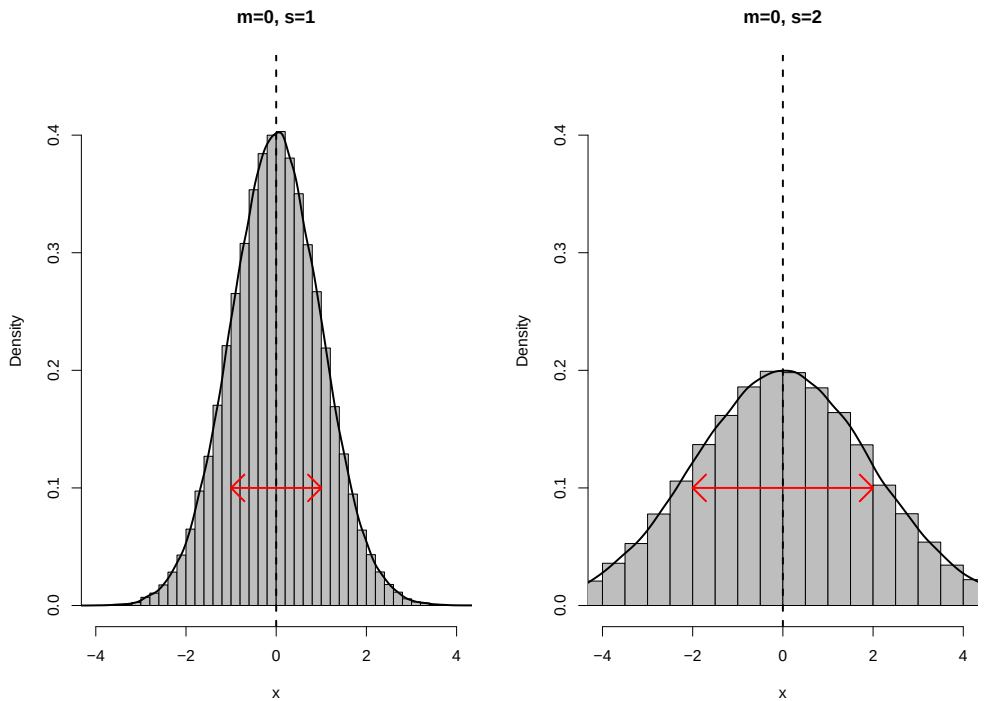
$$s_{n-1}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - m)^2 \tag{11}$$

notations

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

Variable quantitative

paramètres de dispersion : visualisation

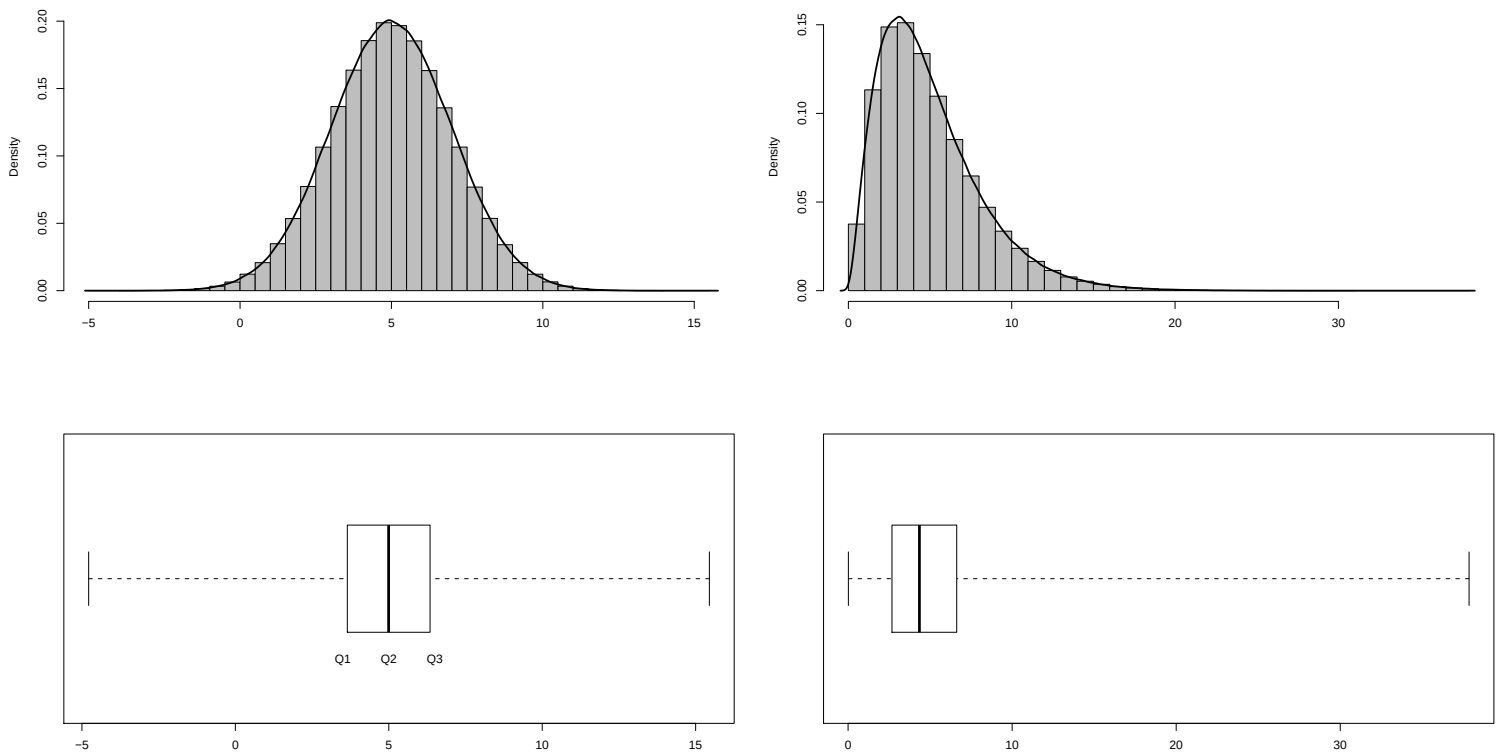


propriétés de la variance

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

Variable quantitative

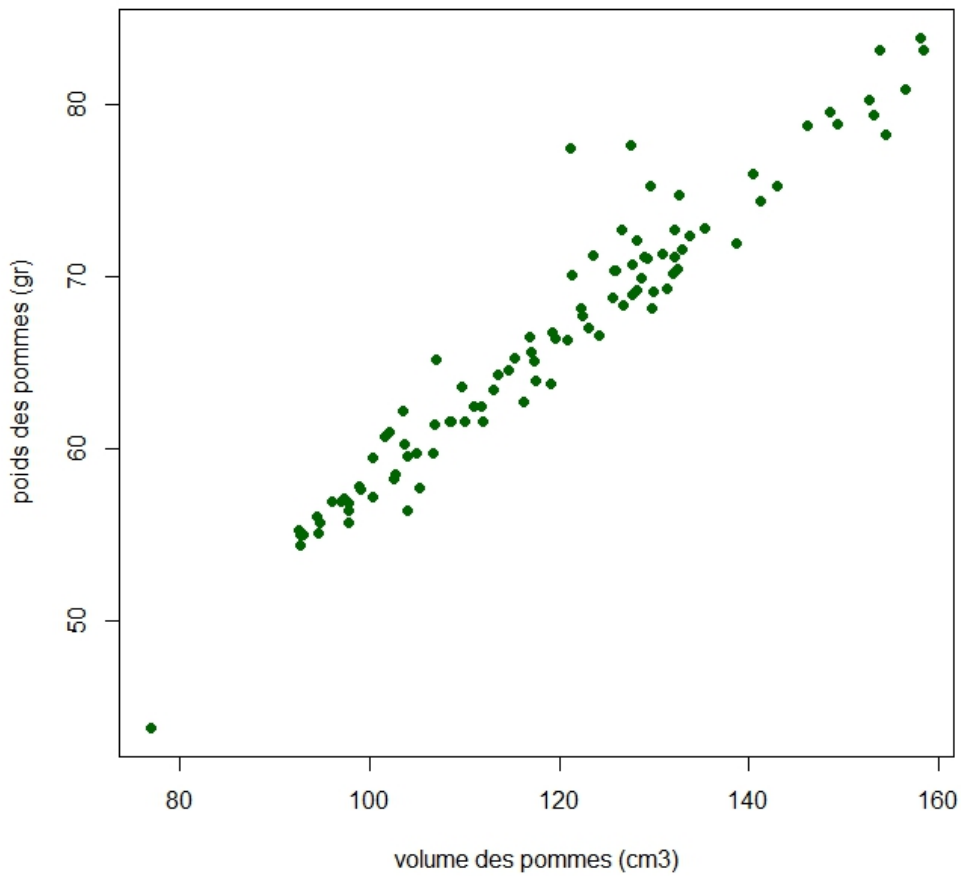
autre représentation : boîte à moustache (boxplot)



$Q1$ = 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)

2 Variables quantitatives

représentation : nuage de points



2 Variables quantitatives

liaison entre deux variables X et Y

- **covariance**

$$\mathrm{Cov}(X, Y) = \sigma_{XY} = s_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) \tag{12}$$

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

$$r = \frac{s_{xy}}{s_x s_y} \tag{13}$$

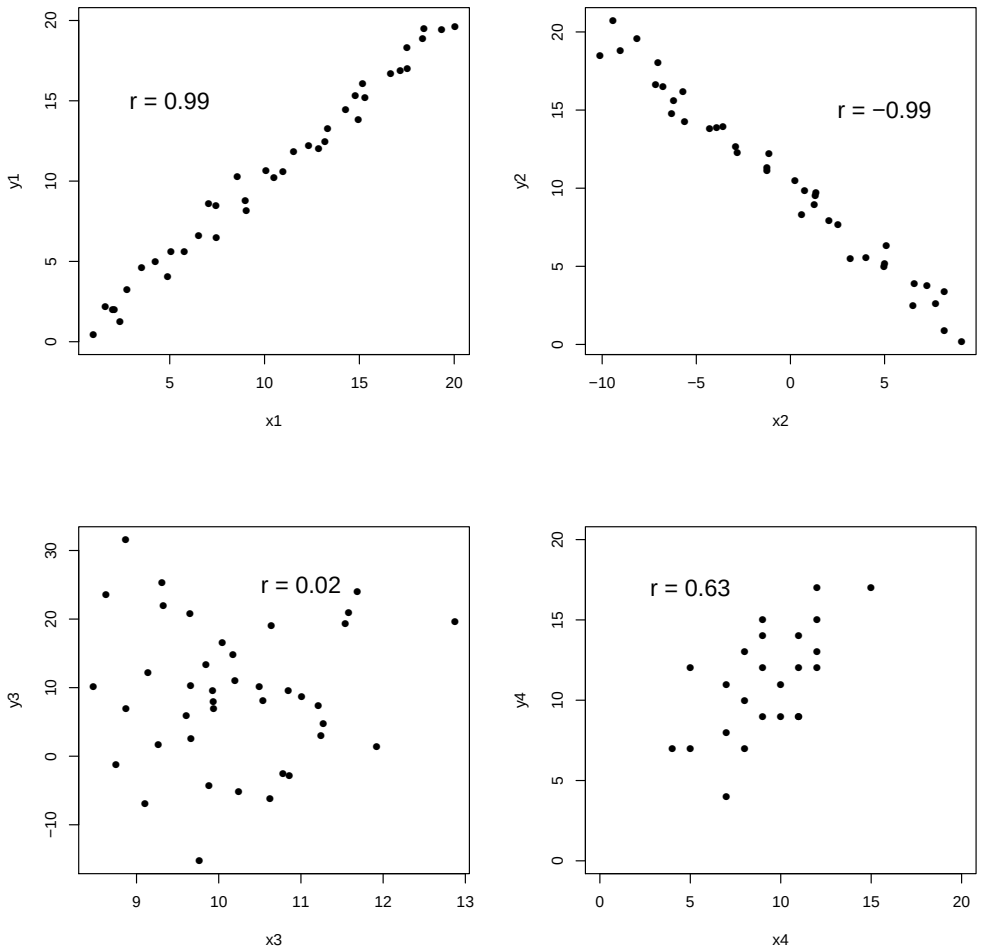
avec

- $\bar{x}$ 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
- $-\infty < s_{xy} < +\infty$, $s_{xy} = 0$ indépendance de X et Y
- $-1 < r < 1$, $r < 0$ = corrélation négative, $r > 0$ = corrélation positive, $r = 0$ pas de corrélation entre X et Y

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

2 Variables quantitatives

corrélations et nuages de points



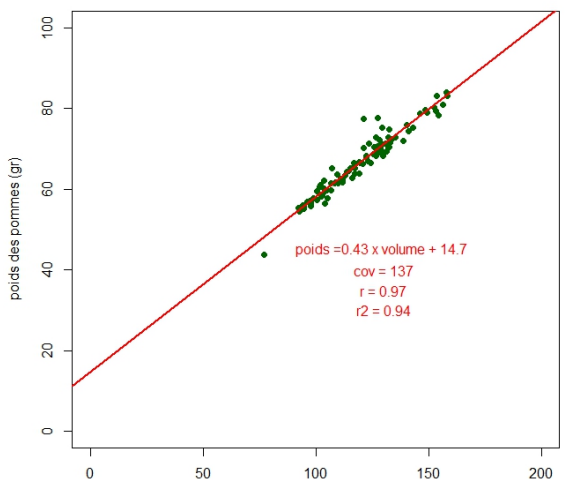
2 Variables quantitatives

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)

$$S = min(\sum_{i=1}^{n}(y_i - \hat{y}_i)^2) \tag{14}$$

- droite de régression $y = ax + b$ avec $a = \frac{s_{xy}}{s_x^2}$



Aller plus loin... : STATISTIQUES INFERENTIELLES

Je veux calculer des probabilités

- probabilités sur les données
- probabilités associées à une loi
  - loi uniforme
  - loi binomiale
  - loi de Poisson
  - loi Normale

Je souhaite généraliser à la population mon résultat sur un échantillon

- échantillon et population
- distributions d'échantillonnage et estimation par intervalle de confiance

Je veux tester des hypothèses sur mes données : quelques tests statistiques

- test du χ^2 : conformité, indépendance
- test d'adéquation (ajustement) : exemple de la normalité
- tests d'homogénéité : variances et moyennes
- analyse de la variance (ANOVA)
- notion d'erreur statistique et de tests multiples

Pfff.. mon tableau excel est gigantesque : analyses multivariées

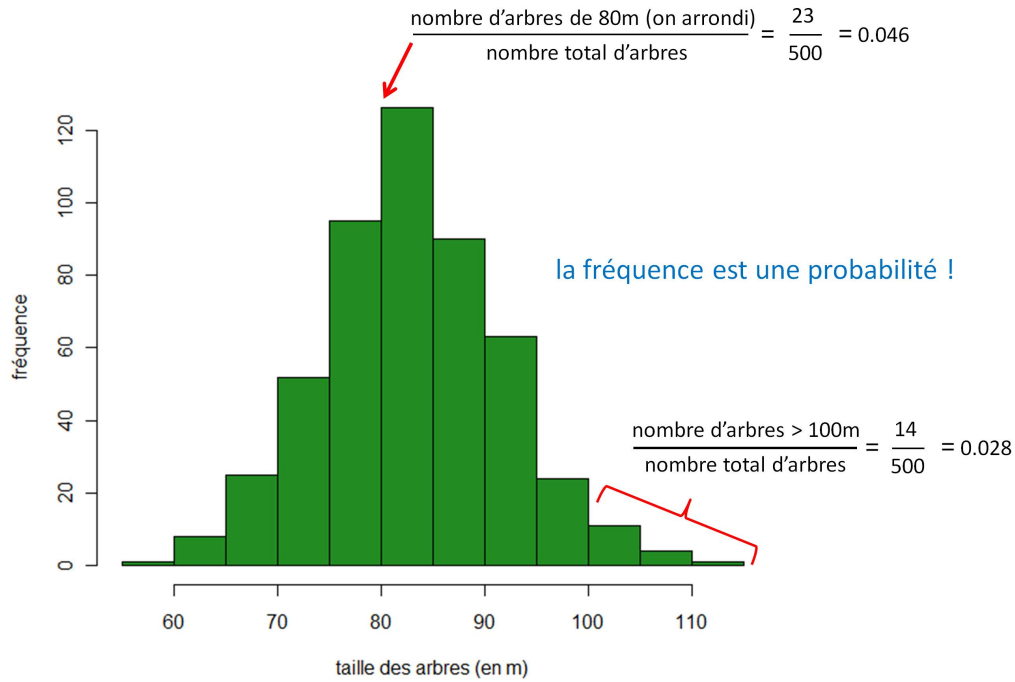
- Analyse en Composantes Principales (ACP)
- Analyse Factorielle des Correspondances (AFC)

Calcul de probabilité sur les données

un exemple

Dans le Parc national de Séquoia (USA), la hauteur de 500 Séquoias âgés a été mesurée au laser par des forestiers. Ci-dessous la distribution (l'histogramme) de la taille de cette population de séquoia.

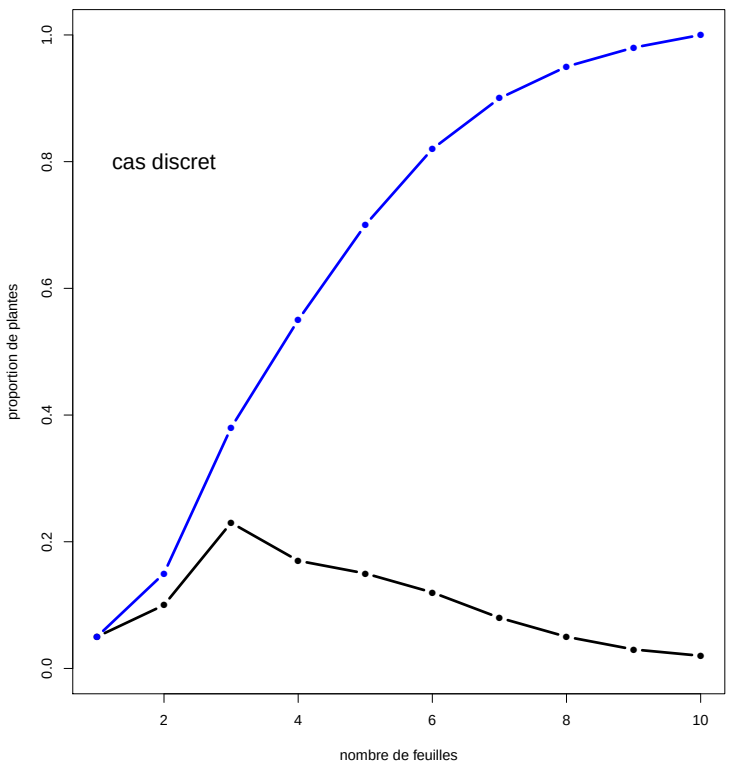
- Quelle est la probabilité qu'un arbre mesure 80m ? Qu'il mesure plus de 100 ?

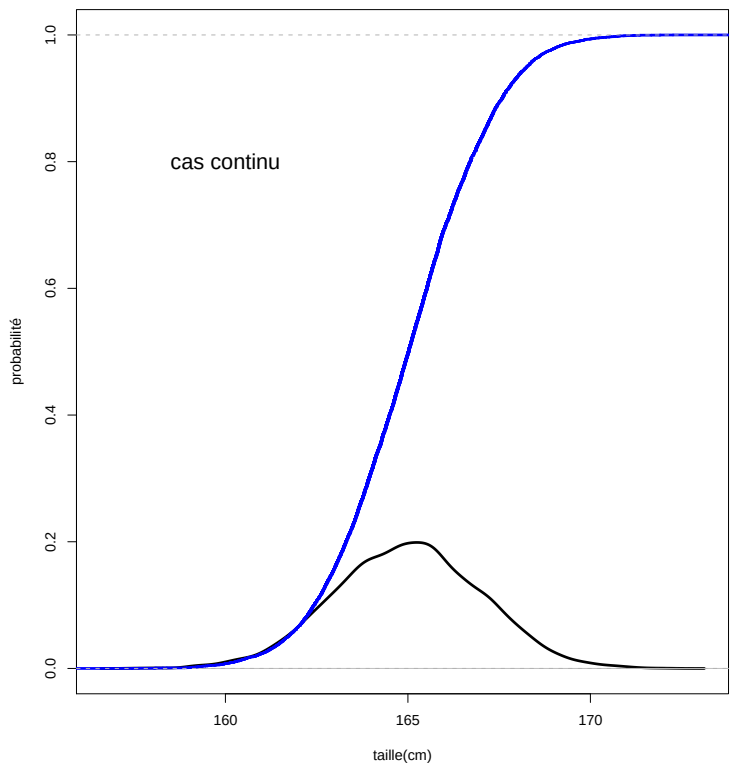


- Parfois les données ont des distributions qui sont bien connues....

Distribution (Loi) de probabilité

- fonction qui donne une probabilité à chaque valeur ou intervalle de valeurs d'une variable aléatoire quantitative X (avec $=\sum_{i=1}^{n} p_i = 1$)
  - $f(k) = \mathbb{P}[X = k]$ **(densité de probabilité)**
  - $F(k) = \mathbb{P}[X \leq k]$ **(fonction de répartition, cumulative,** $\mathbb{P}[X \leq k]$ **= intégrale sur** $min(X) \leq X \leq k$

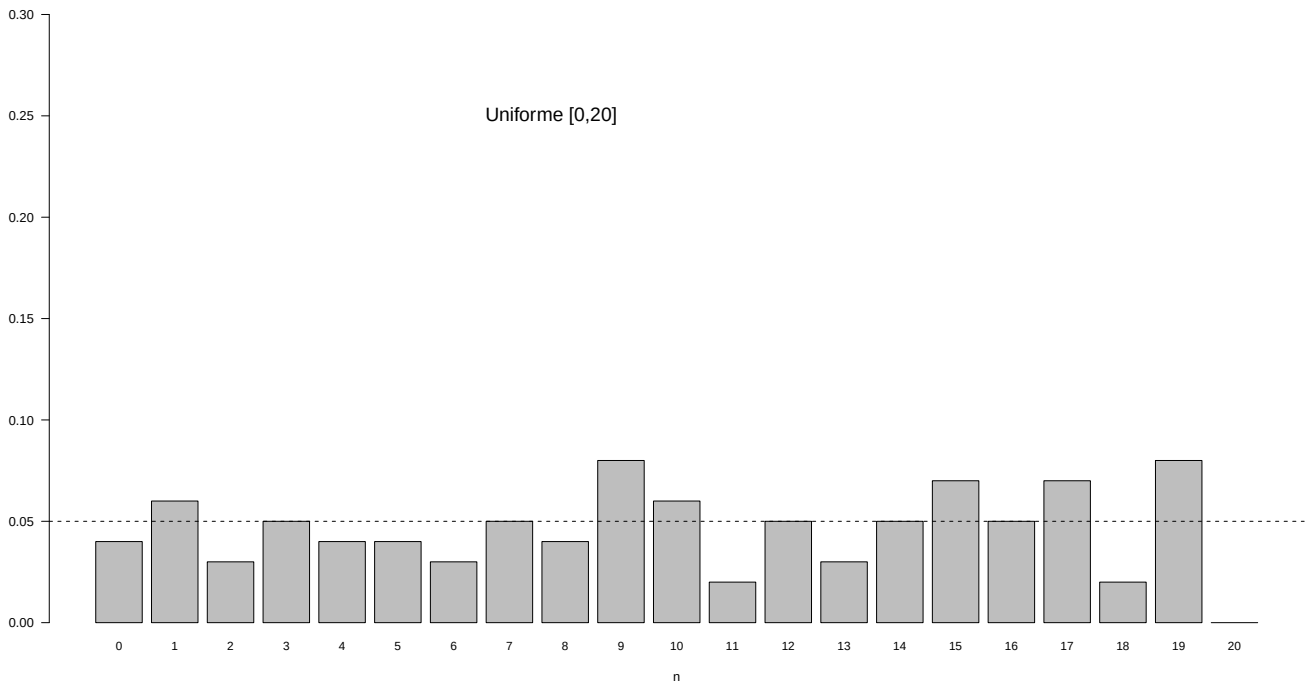




Distribution (Loi) de probabilité

Loi (discrète) uniforme : équiprobabilité entre chaque valeurs d'un ensemble fini

- $X = 1, 2, 3, ...n$, $\mathbb{P}(X = 1) = \mathbb{P}(X = 2) = ...\mathbb{P}(X = n) = \frac{1}{n}$



exemples : jeu de dés, roulette...

Distribution (Loi) de probabilité

Loi (discrète) de Bernoulli : expérience à deux issues (succès - échec)

- codage : succès = 1, échec = 0, en général non équiprobables
- $X \in \{0, 1\}$; $\mathbb{P}(X = 1) = p$; $\mathbb{P}(X = 0) = 1 - p$
- $\mathbb{E}(X) = p$; $\mathrm{Var}(X) = p(1 - p) = pq$

exemples : pile ou face, **germination**,...

- si on met une graine à germer et que la probabilité de germination est p, alors elle va germer avec cette probabilité là ! ... oui et alors ?
- ... cela devient intéressant quand on a plusieurs graines ! (on répète l'expérience de Bernoulli...)

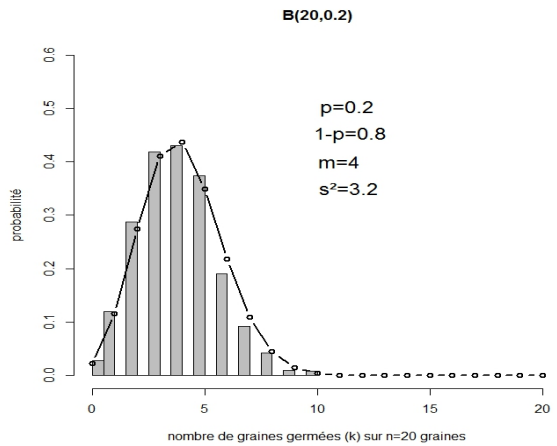
Distribution (Loi) de probabilité

Lois (discrète) Binomiale $B(n,p)$ (loi d'une somme de variables de Bernoulli)

- on s'intéresse au nombre de succès k dans une expérience de n essais (épreuves de Bernoulli) avec probabilité de succès p
- $X \in \{0,...,n\}$; $\mathbb{P}(X=k) = C_n^k p^k (1-p)^{n-k}$, avec $C_n^k = \binom{n}{k} = \frac{n!}{k!(n-k)!}$
  (nombre d'échantillons possibles de k succès dans n tirages)
- $\mathbb{E}(X) = \mu = np$; $\mathrm{Var}(X) = \sigma^2 = np(1-p) = npq$

notre exemple de la germination

- si proba germination $= p$, nombre de graines germées k dans un échantillon de taille n suit $B(n,p)$



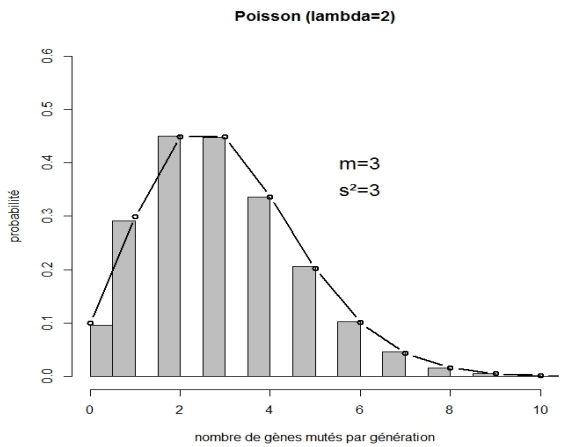
Distribution (Loi) de probabilité

Lois (discrète) de Poisson $P(\lambda)$: limite de la loi binomiale quand $p \to 0$ et $n \to \infty$

- loi de probabilité d'un événement rare, d'occurence moyenne $\lambda = np$ sur un intervalle (temps, espace) donné
- $X \in \mathbb{N}$; $\lambda \in \mathbb{R}_{\geq 0}$; $\mathbb{P}(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$
- $\mathbb{E}(X) = \lambda$; $\mathrm{Var}(X) = \sigma^2 = \lambda$

nombre de gènes d'un génome mutés à chaque génération (homme, Arabidopsis)

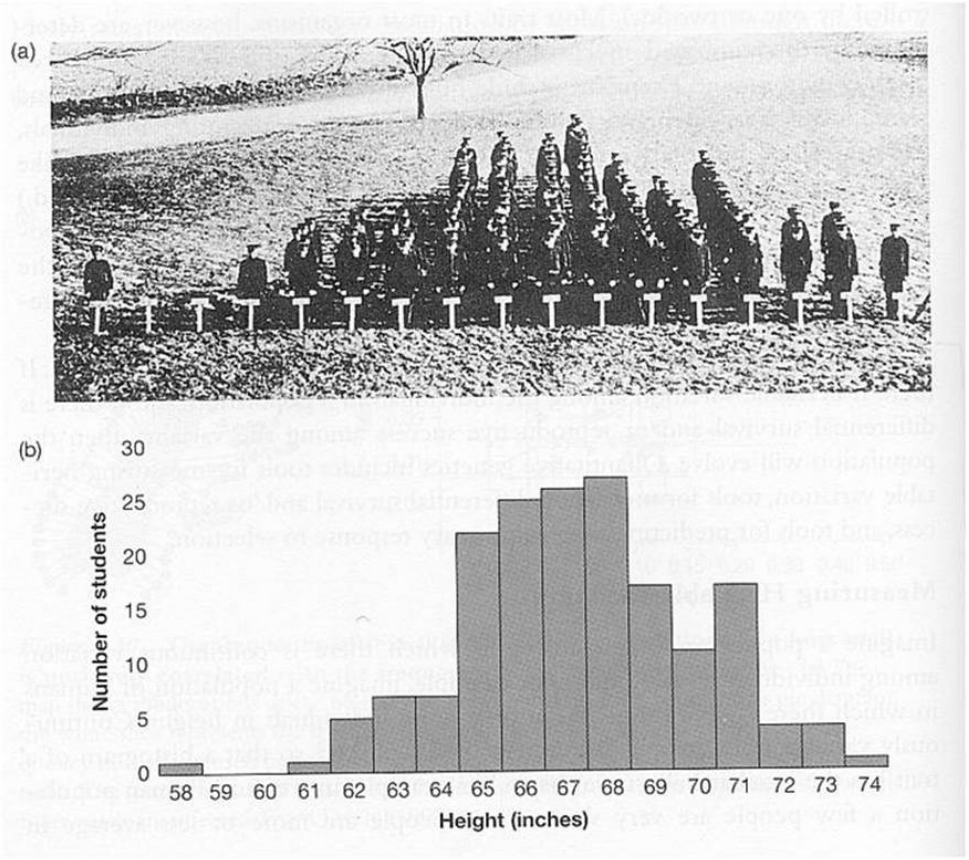
- proba qu'un gène mute $p = 10^{-4}$, nombre de gènes $n = 30000$, $\to \lambda = np = 3$



Distribution (Loi) de probabilité

loi (continue) Normale ou de Laplace-Gauss $\mathcal{N}(\mu, \sigma)$ ("Gaussienne")

- Loi inhérente à un grand nombre de phénomènes naturels



élèves d'un lycée agricole (Connecticut 1914)

Distribution (Loi) de probabilité

loi (continue) Normale ou de Laplace-Gauss $\mathcal{N}(\mu, \sigma)$ ("Gaussienne")

- ... une version plus récente !

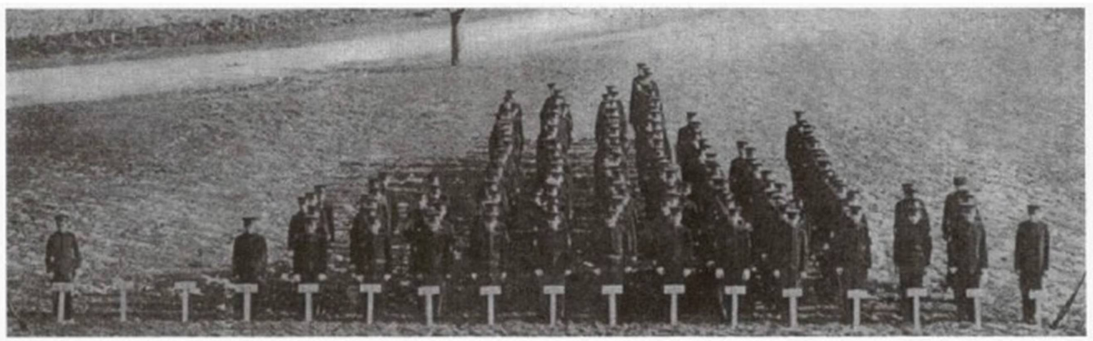


FIGURE 2.—A living histogram from Connecticut State Agricultural College (BLAKESLEE 1914). The number of men is 175, the mean height is 67.3 in., and the standard deviation is 2.7 in.

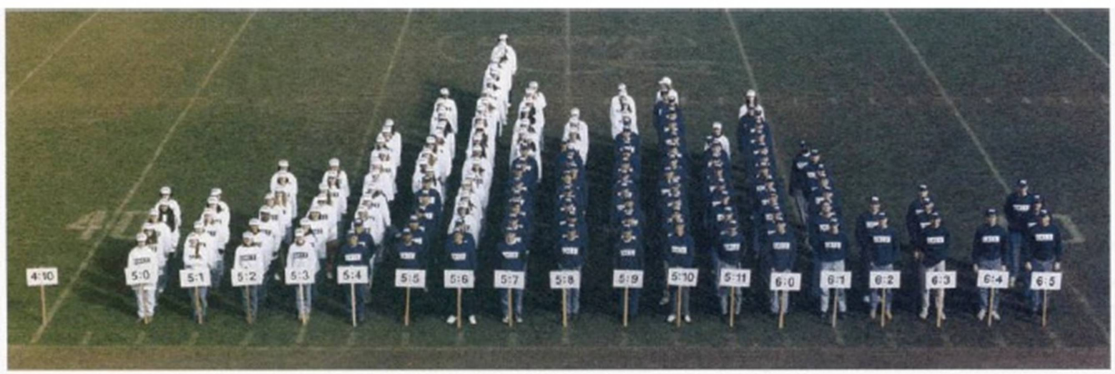
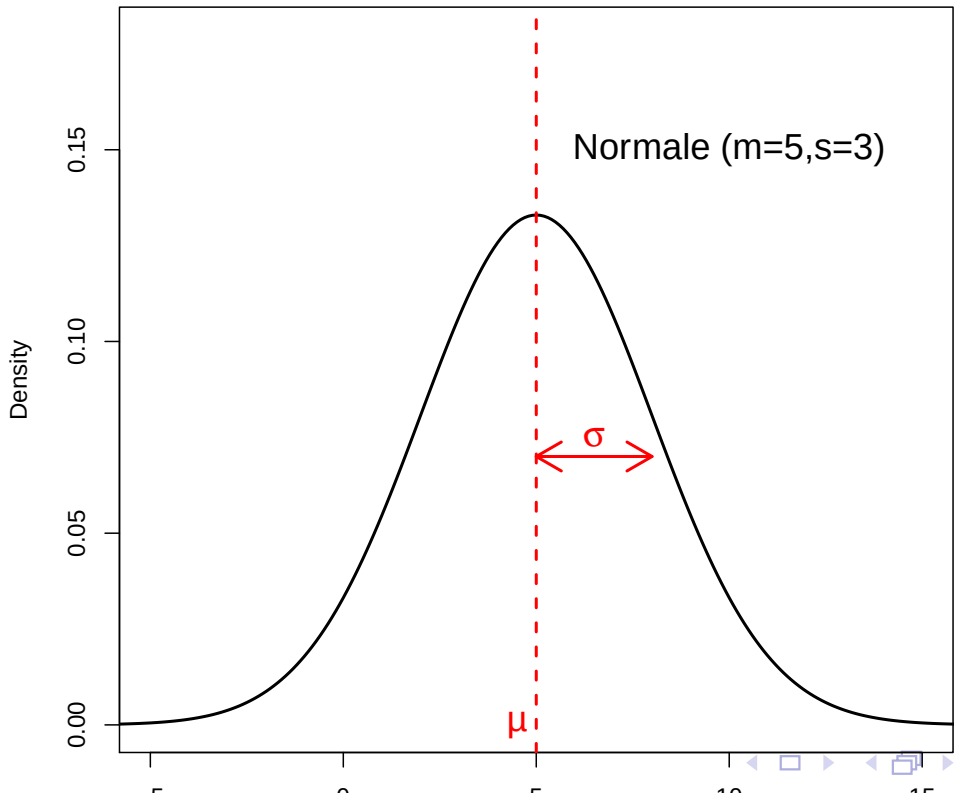


FIGURE 3.—A modern version of Figure 2, from Connecticut State University in 1996. The means and standard deviations in inches are as follows: males, 70.1 ± 3.0; females, 64.8 ± 2.7; combined, 67.6 ± 4.0. Photo from LINDA STRAUSBAUGH.

Distribution (Loi) de probabilité

loi (continue) Normale ou de Laplace-Gauss $\mathcal{N}(\mu, \sigma)$ ("Gaussienne")

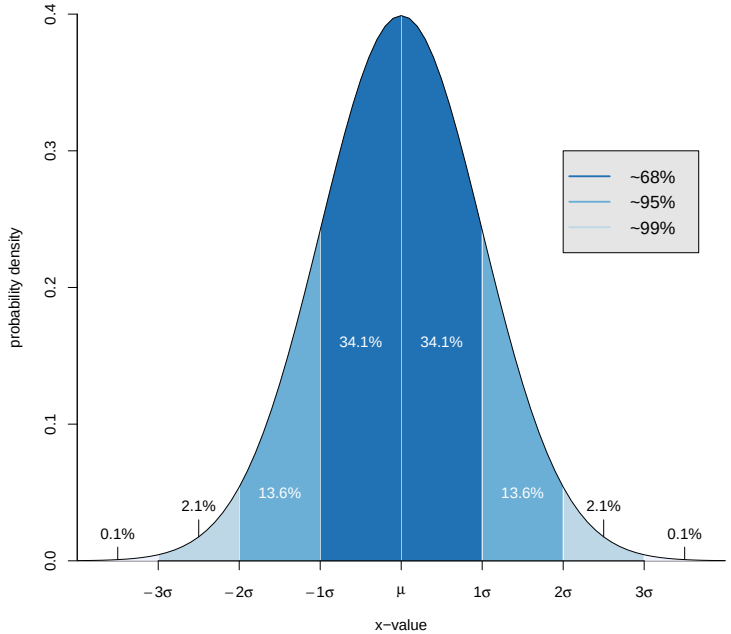
- densité : $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$
- $\mathbb{E}(X) = \mu$; $\mathrm{Var}(X) = \sigma^2$

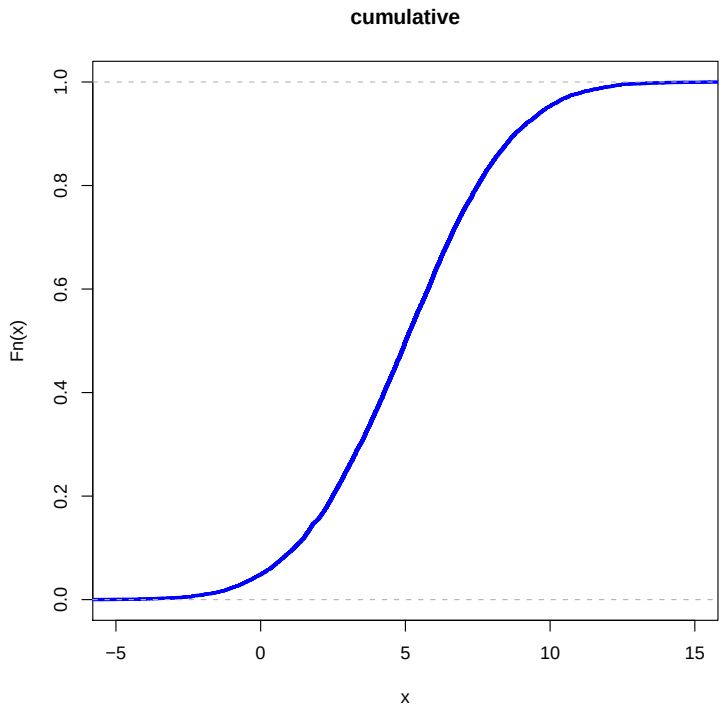


Distribution (Loi) de probabilité

loi (continue) Normale ou de Laplace-Gauss $\mathcal{N}(\mu, \sigma)$ ("Gaussienne")

- **valeurs remarquables** **fonction de répartition ($F(x)$)**



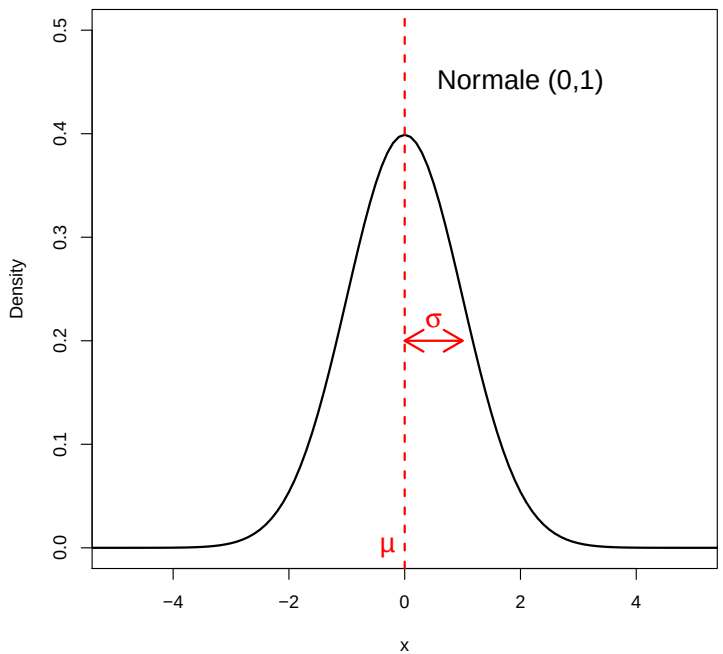


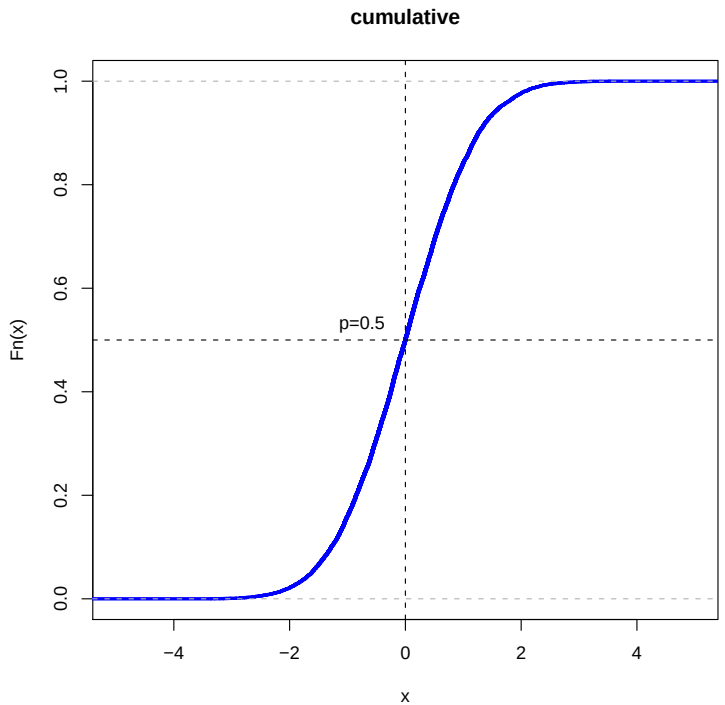
- une propriété parmis d'autres : si $X_1 \sim \mathcal{N}(\mu_1, \sigma_1)$ et $X_2 \sim \mathcal{N}(\mu_2, \sigma_2)$ alors $X_1 + X_2 \sim \mathcal{N}(\mu_1 + \mu_2, \sqrt{(\sigma_1^2 + \sigma_2^2)})$

Distribution (Loi) de probabilité

loi (continue) Normale centrée réduite $\mathcal{N}(0,1)$

- densité : $f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$





- toute $X \sim \mathcal{N}(\mu, \sigma)$ peut être transformée en $T \sim \mathcal{N}(0,1)$ par la transformation $T = \frac{X-\mu}{\sigma}$
- pour un calcul de probabilité sur X, on utilise la table qui donne la fonction de répartition de la variable T

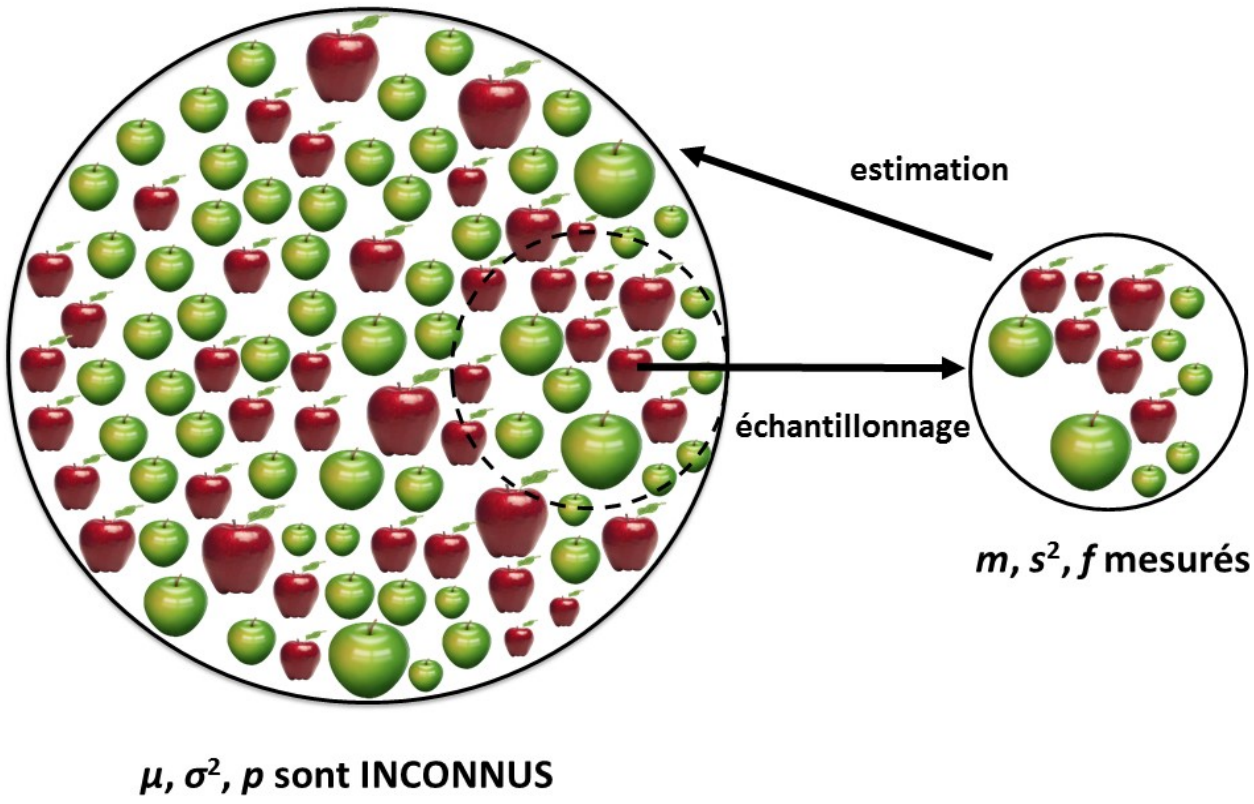
Distribution (Loi) de probabilité

$\mathcal{N}(0,1)$: application à des calculs de probabilité

- Le Parc touristique A affirme qu'il y a plus de chances d'observer des Séquoia de plus de 100m de hauteur chez lui que dans le Parc B, car ses arbres font en moyenne 90m, alors qu'ils font 85m en moyenne dans le Parc B. Est-ce vrai ? :
  - Parc A : la hauteur H_A des arbres suit une $\mathcal{N}(90,5)$
  - Parc B : la hauteur H_B des arbres suit une $\mathcal{N}(85,10)$
  - $\frac{H_A-90}{5} = \mathcal{N}(0,1)$, $\frac{H_B-85}{10} = \mathcal{N}(0,1)$
  - $1-\mathbb{P}(H_A<100) = 1-\mathbb{P}(\frac{H_A-90}{5} < \frac{100-90}{5}) = 1-\mathbb{P}(T<2) = 0.0227$
  - $1-\mathbb{P}(H_B<100) = 1-\mathbb{P}(\frac{H_B-85}{10} < \frac{100-85}{10}) = 1-\mathbb{P}(T<1.5) = 0.067$
- conclusion : Dans le Parc B, il y a 6.7% de chances d'observer un arbre de plus de 100m, et seulement 2.2% de chances dans le Parc A

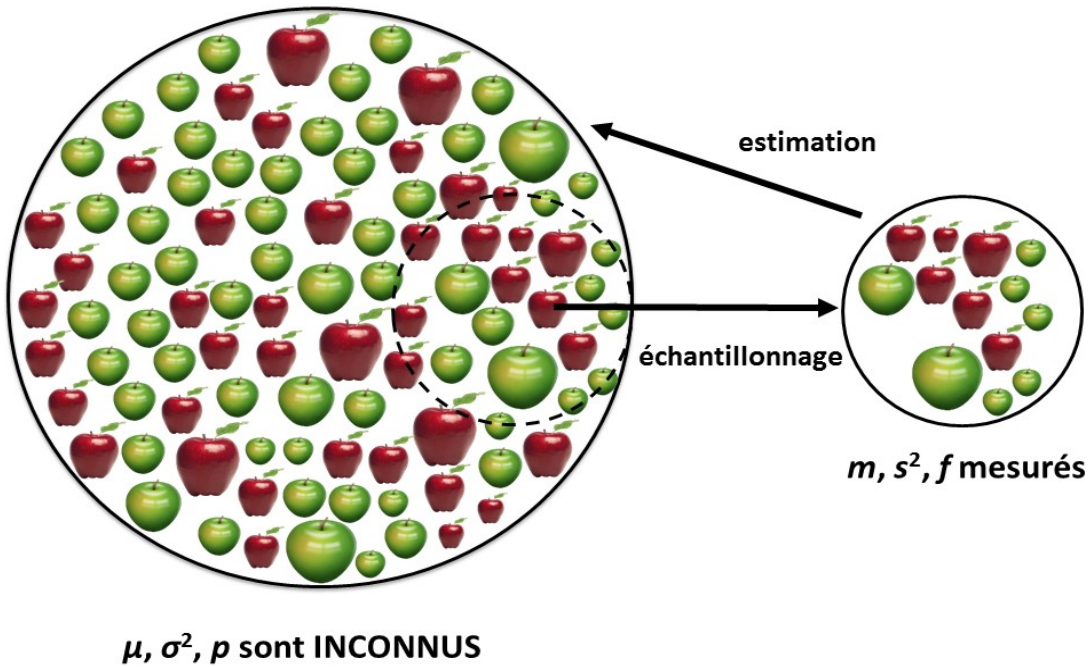
Echantillonnage - Estimation

Comment passer de l'échantillon à la population ?



Echantillonnage - Estimation

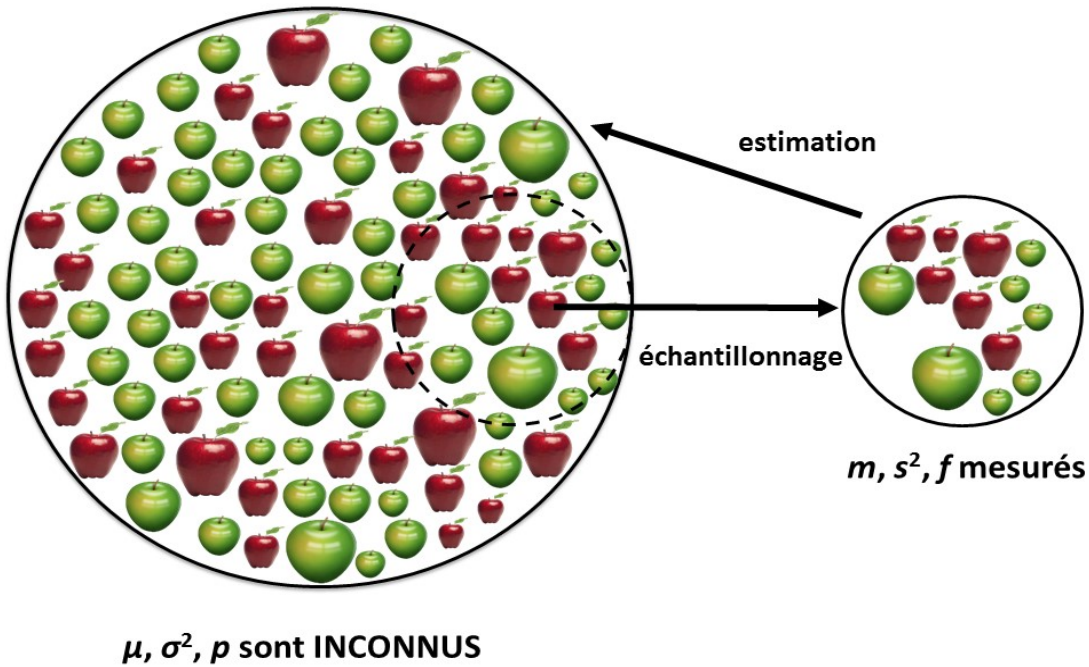
Loi des grands nombres



Si l'échantillon est très grand, alors la moyenne m, la variance s^2 ou la fréquence f dans l'échantillon sera très proche des valeurs de la population (μ, σ^2, p).

Echantillonnage - Estimation

Théorie de l'échantillonnage (si n est petit)

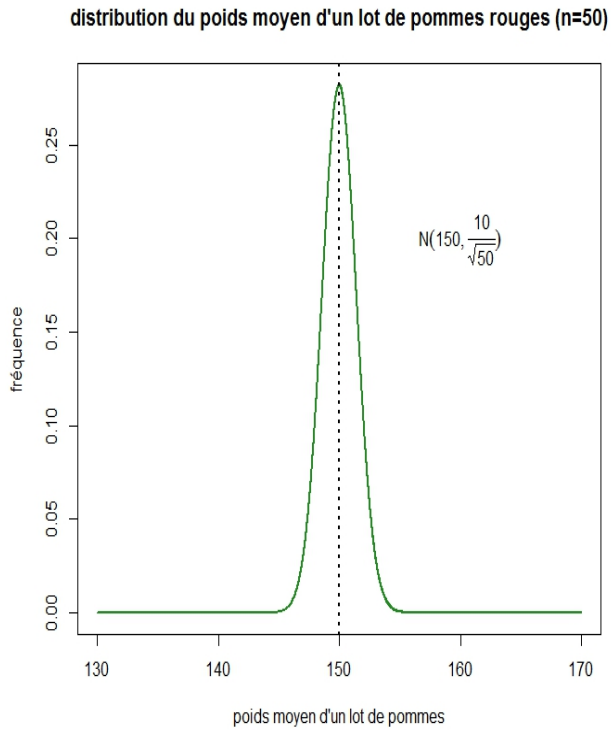


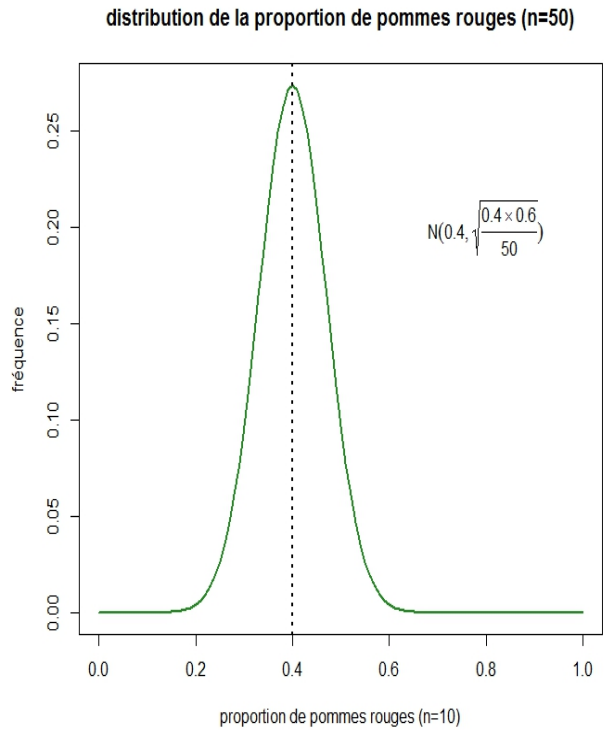
tout comme une variable aléatoire X suit une certaine loi de distribution de moyenne μ et de variance σ^2 (de la population), m, s^2 et f varient d'un échantillon à un autre de la même population, et sont donc aussi des variables aléatoires. Chaque paramètre possède alors une distribution d'échantillonnage au même titre que X.

Echantillonnage - Estimation

Connaître les distributions d'échantillonnage...

$$m \rightarrow \mathcal{N}(\mu, \frac{\sigma}{\sqrt{n}}) \qquad f \rightarrow \mathcal{N}(p, \sqrt{\frac{pq}{n}})$$





NB : grâce au théorème central limite (voir démonstration en annexe 1)

Echantillonnage - Estimation

... permet d'estimer par **intervalle de confiance**

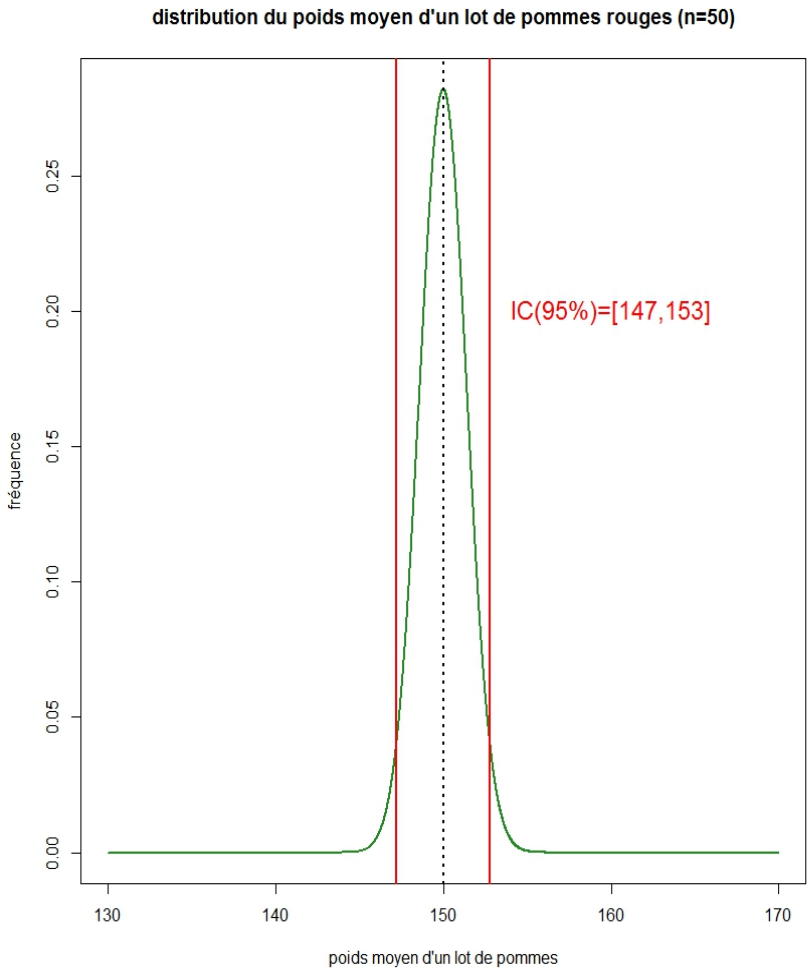
- **intervalle de confiance (IC) : intervalle ayant une probabilité $1-\alpha$ donnée de contenir la vrai valeur du paramètre (ex : p, μ, σ...)**
- **IC d'une proportion** :
  - pour n grand on utilise l'approximation normale :
    $IC_{(p)} = [p - u\sqrt{\frac{p(1-p)}{n}}, p + u\sqrt{\frac{p(1-p)}{n}}]$
  - où u est le quantile d'ordre $1 - \frac{\alpha}{2}$ de la loi $\mathcal{N}(0,1)$
  - $IC_{(p)}(95\%) = [p - 1.96\sqrt{\frac{p(1-p)}{n}}, p + 1.96\sqrt{\frac{p(1-p)}{n}}]$
  - exemple : 20 pommes rouges sur 50 pommes rouges + vertes :
    $p = \frac{20}{50} = 0.4$, $IC_{(p)}(95\%) = [0.26, 0.54]$ **(0.4 ± (1.96 * √(0.4 * 0.6/50)))**

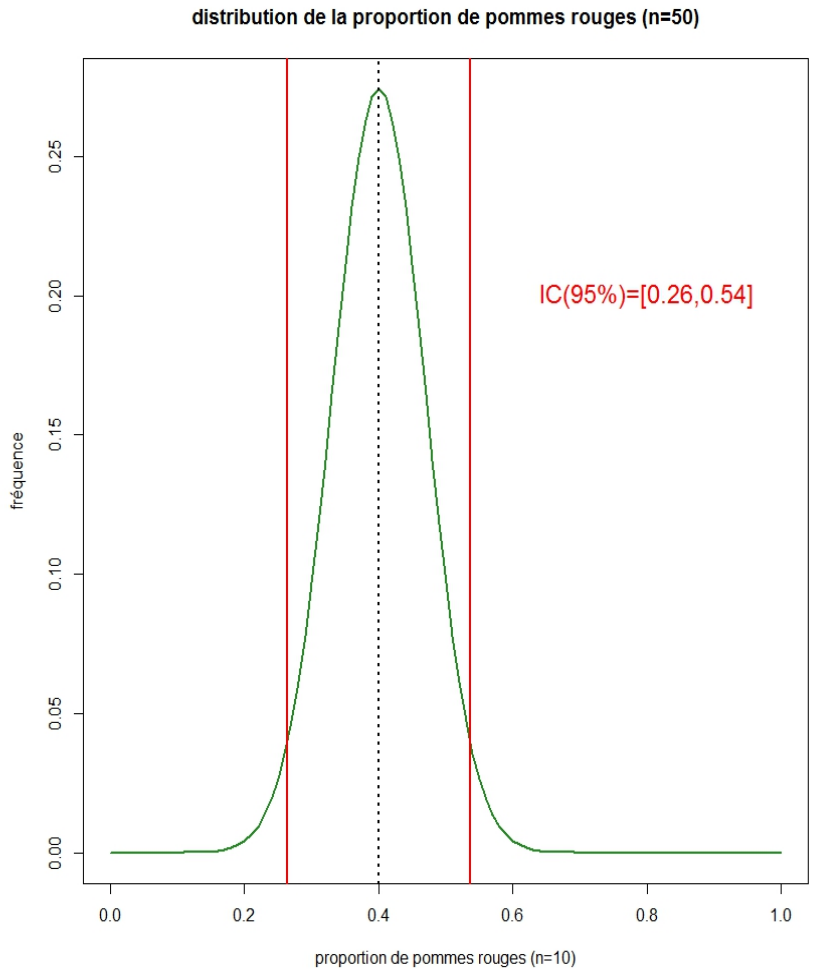
Echantillonnage - Estimation

... permet d'estimer par **intervalle de confiance**

- **IC d'une moyenne** :
  - $IC_{(\mu)} = [m - u\frac{\sigma}{\sqrt{n}}, m + u\frac{\sigma}{\sqrt{n}}]$
  - où u est le quantile d'ordre $1 - \frac{\alpha}{2}$ de la loi $\mathcal{N}(0,1)$
  - $IC_{(\mu)}(95\%) = [m - 1.96\frac{\sigma}{\sqrt{n}}, m + 1.96\frac{\sigma}{\sqrt{n}}]$, avec $u_{(0.975)} = 1.96$ $(\alpha = 0.05)$
  - $IC_{(\mu)}(99\%) = [m - 2.6\frac{\sigma}{\sqrt{n}}, m + 2.6\frac{\sigma}{\sqrt{n}}]$, avec $u_{(0.995)} = 2.6$ $(\alpha = 0.01)$
  - exemple : poids moyen d'un lot de 50 pommes rouge $= 150g$, avec $\sigma = 10$: $IC_{(\mu)}(95\%) = [147, 153]$ $(150 \pm (1.96\frac{10}{\sqrt{50}})$
- **pour augmenter la confiance, il faut élargir l'intervalle**
- **pour obtenir un intervalle plus fin avec même degré de confiance, il faut augmenter la taille n de l'échantillon.**

Echantillonnage - Estimation





Test statistique : comment ça marche

Principe :

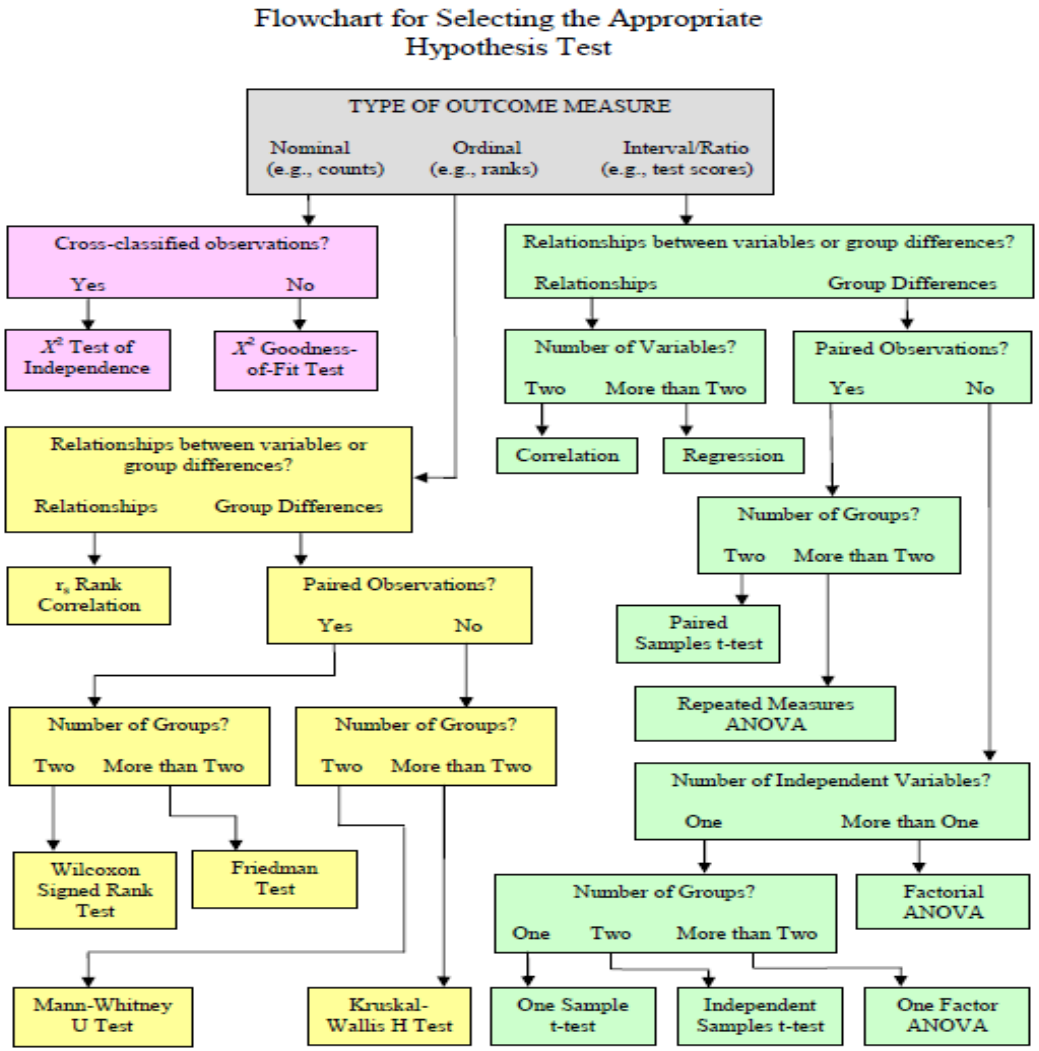
Démarche consistant à rejeter ou ne pas rejeter une hypothèse statistique, appelée hypothèse nulle ($H0$), en fonction d'un jeu de données (échantillon). **Le rejet de l'hypothèse $H0$ est nécessairement associé à un risque d'erreur (car l'hypothèse $H0$ est toujours possible, même si elle est parfois peu probable)**, le but étant que ce risque soit le plus faible possible (souvent $< 5\%$).

Classification selon $H0$:

- **test de conformité, d'adéquation** : confronter un paramètre (ou une distribution) calculé(s) sur l'échantillon, à une valeur (ou une distribution) pré-établie.
- **test d'homogénéité (ou de comparaison)** : vérifier que K ($K \geq 2$) échantillons (groupes) proviennent de la même population (i.e. la distribution de la variable d'intérêt est la même dans les K échantillons).
- **test d'indépendance (ou d'association)** : rechercher une liaison entre 2 variables (qualitatives, quantitatives).

Test statistique : comment ça marche

Quel test utiliser ?



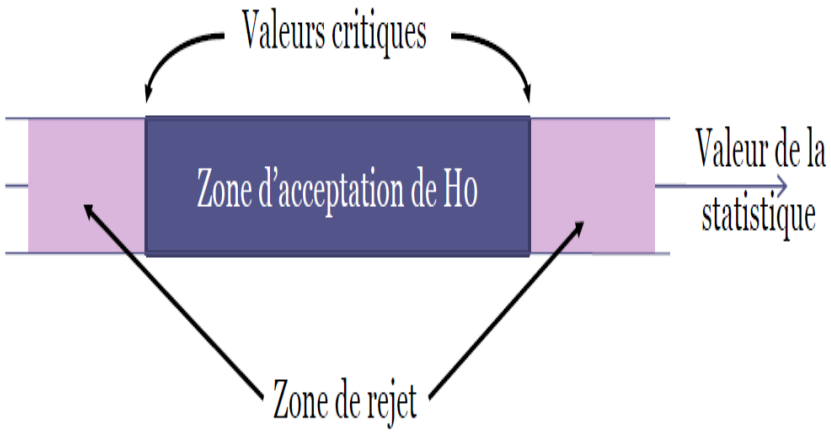
Test statistique : comment ça marche

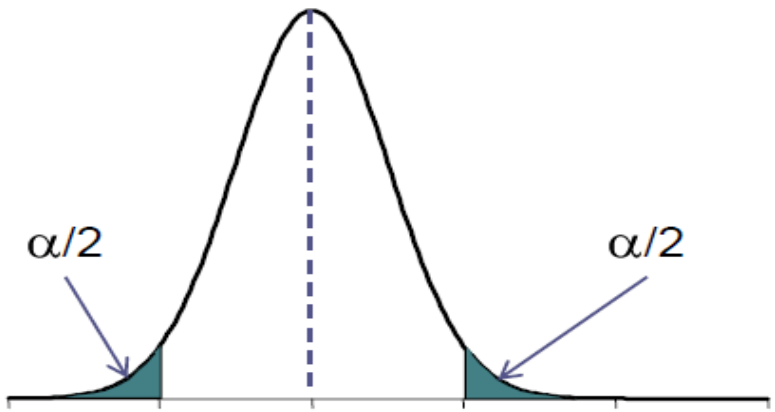
Une fois que j'ai choisi mon test : les étapes

- **Expérience et/ou récolte des données** : tableau de données (compatibles avec les conditions d'application du test ?)
- **Réalisation du test**
  - Quelle est l'hypothèse nulle $H0$? L'hypothèse alternative $H1$? (dépend si uni ou bi-latéral)
  - Qu'est-ce que je calcule sur mes données ? ($\pm$ important pour le biologiste)
  - Quelle est la distribution des valeurs possibles de la statistique sous $H0$? ($\pm$ important pour le biologiste)
  - Pour quel(s) intervalle(s) de valeurs $H0$ est peu probable ($\leq \alpha$) ? (et donc que je vais me permettre de la rejeter avec un risque α)
  - J'accepte ou je rejette $H0$? J'en conclus quoi ?

Test statistique : comment ça marche

Distribution *H*0 et risque d'erreur lors du rejet de *H*0





erreur de 1re espèce (type I) : rejeter une hypothèse alors qu'elle est vraie (α = risque d'erreur de type I)

Test statistique : comment ça marche

EN PRATIQUE

- connaître $H0$ et $H1$
- fixer le risque d'erreur α autorisé (ex : $5\% = 0.05$, $1\% = 0.01$,...)
- calculer la statistique du test sur les données
- avec la distribution sous $H0$: calculer la probabilité d'obtenir sous $H0$ une valeur supérieure ou égale à la statistique calculée sur les données (c'est la $p - value$)
  - si $p - value < \alpha$: on rejette $H0$ au risque d'erreur α (on accepte donc $H1$)
  - si $p - value > \alpha$: on accepte $H0$
- **remarque : pour ce qui est en bleu c'est le logiciel qui fait le boulot pour moi !**

explorons cela avec des exemples...

Test sur les effectifs (variable qualitative)

Test de χ^2 de conformité ($H0$)

- **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 (voir annexe 2) à $(p-1)$ ddl (si $H0$ vraie)
- on rejette l'hypothèse d'adéquation si $\mathbb{P}(\chi^2 > \chi^2_{obs}) \leq \alpha$ (le test est toujours unilatéral)

Exemple en génétique Mendelienne

- fréquences génotypiques attendues après autoféconda ton d'une F1 hybride, si 2 allèles ($A1$,$A2$) à un gène donné : $\frac{1}{4}$ de $A1A1$, $\frac{1}{2}$ de $A1A2$, $\frac{1}{4}$ de $A2A2$
- par croisement, sur $n = 100$ descendants, on obtient 20 $A1A1$, 54 $A1A2$, 26 $A2A2$
- $\chi^2_{obs} = \frac{(20-25)^2}{25} + \frac{(54-50)^2}{50} + \frac{(26-25)^2}{25} = 1.36$
- $\mathbb{P}(\chi^2 > 1.36) = 0.5066 > \alpha$ (pour une distribution χ^2 avec $ddl = 2$ et $\alpha = 5\%$)
- conclusion : on accepte $H0$ = conformité = le croisement suit bien les proportion Mendeliennes attendues

Test sur les effectifs (variable qualitative)

Test du χ^2 d'indépendance ($H0$)

- **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**

Résistance qualitative des plantes à un pathogène

- Question : la résistance ou sensibilité à un pathogène est elle indépendante de l'écotype d'*Arabidopsis thaliana* ? Autrement dit y-a-t-il indépendance entre **phénotype** et **écotype** ?
- tableau de contingence :

	Col-0	Ws-0	Can-0	**effectifs**
Résistant	15	5	3	23
Sensible	0	19	16	35
effectifs	15	24	19	58

Test sur les effectifs (variable qualitative)

Test du χ^2 d'indépendance ($H0$)

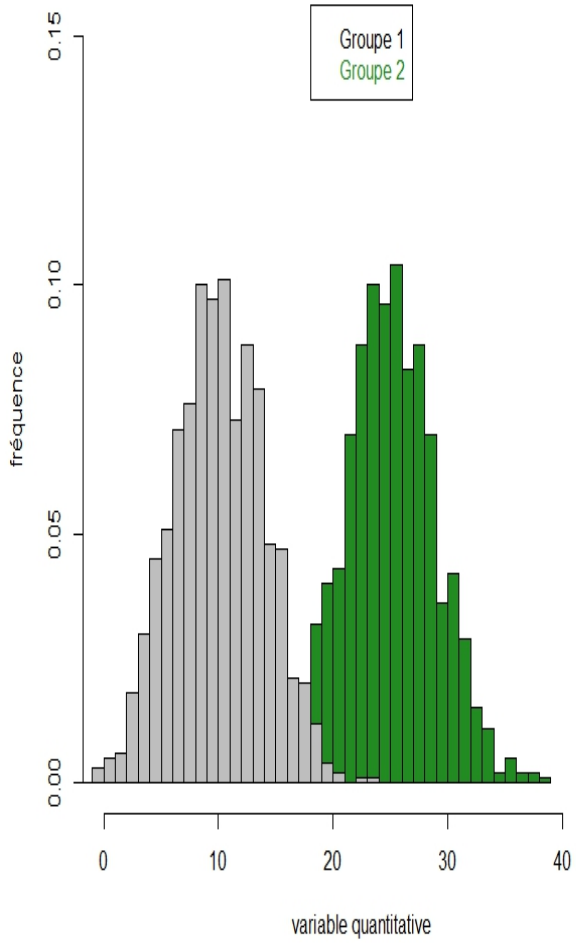
- $\chi^2_{obs} = \sum_{i=1}^{p} \sum_{j=1}^{q} \frac{(n_{ij} - n\hat{P}_{ij})^2}{n\hat{P}_{ij}}$ suit une loi du χ^2 à $(p-1)(q-1)$ ddl (si $H0$ vraie)
- notation :
  - p=nombre de classes i
  - q=nombre de classes j
  - n_{ij}=effectif de la classe ij
  - n=effectif total
  - $n\hat{P}_{ij}$ = effectif attendu de la classe ij (=produit des marges, divisé par n)
- on rejette $H0$ d'indépendance si $\mathbb{P}(\chi^2 > \chi^2_{obs}) \leq \alpha$

Résistance qualitative des plantes à un pathogène

- $\chi^2_{obs} = \frac{(15-6)^2}{6} + \frac{(5-10)^2}{10} + \frac{(3-8)^2}{8} + \frac{(0-9)^2}{9} + \frac{(19-14)^2}{14} + \frac{(16-11)^2}{11} = 30.9$
- $\mathbb{P}(\chi^2 > 30.9) = 2x10^{-7} < \alpha$ (pour une distribution χ^2 avec $ddl = 2$ et $\alpha = 5\%$)
- conclusion : on rejette $H0$ = pas indépendance = la résistance / sensibilité au pathogène dépend de l'écotype (les plantes Col-0 sont significativement plus résistantes)

Test d'homogénéité : comparaison (variable quantitative) de 2 groupes

Marche à suivre :



Comparer ces 2 groupes = **comparer les valeurs moyennes**
(en tenant compte de la variabilité!)

au préléable:

Vérifier que les 2 distributions sont normales
(test d'ajustement à la loi normale (H0): **Shapiro**, χ^2, Kolmogorov-Smirnov, quantile-quantile plot)

OUI — NON

Vérifier que les variances sont homogènes (H0: $\sigma^2_1 = \sigma^2_2$)
(test de Fisher sur les variances)

Comparer les 2 groupes avec un **test non paramétrique**
(test Mann - Whitney – Wilcoxon, voir annexe 6)

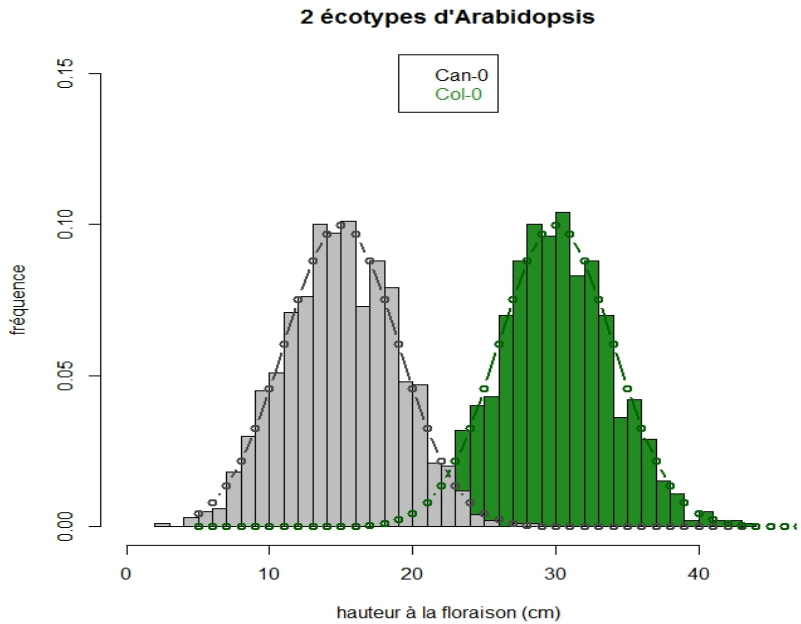
OUI — NON

Comparer les moyennes
(H0: $\mu_1 = \mu_2$)
(test de Student)

Comparer les moyennes
(H0: $\mu_1 = \mu_2$)
(test de Welch)

Test d'homogénéité : comparaison (variable quantitative) de 2 groupes

Différence de croissance entre 2 écotypes d'*Arabidopsis thaliana* ?

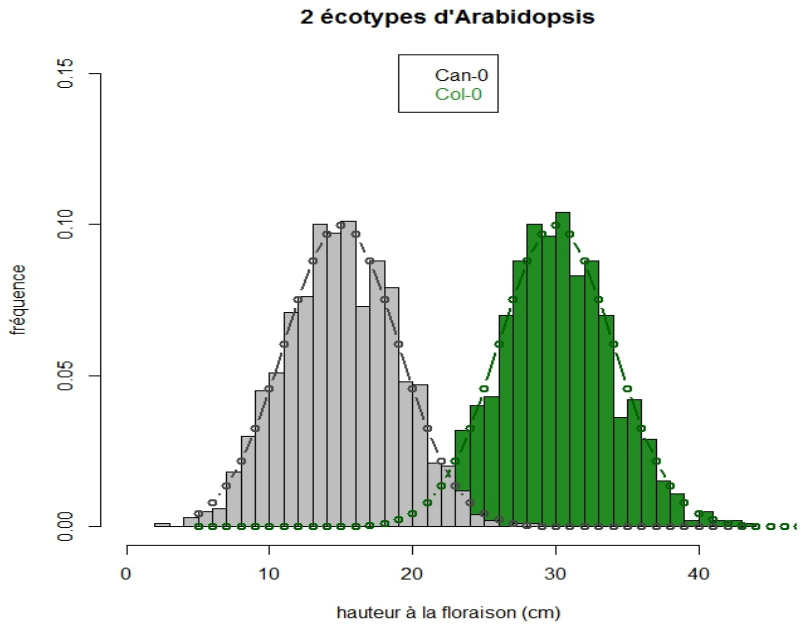


Normalité des données : test de Shapiro (*H*0 : normalité)

- Groupe "Can-0" : $W = 0.9988$, $p - value = 0.77 > \alpha$, on accepte $H0$
- Groupe "Col-0" : $W = 0.9989$, $p - value = 0.80 > \alpha$, on accepte $H0$
- détails de la statistique de Shapiro ; voir annexe 3

Test d'homogénéité : comparaison (variable quantitative) de 2 groupes

Différence de croissance entre 2 écotypes d'*Arabidopsis thaliana* ?

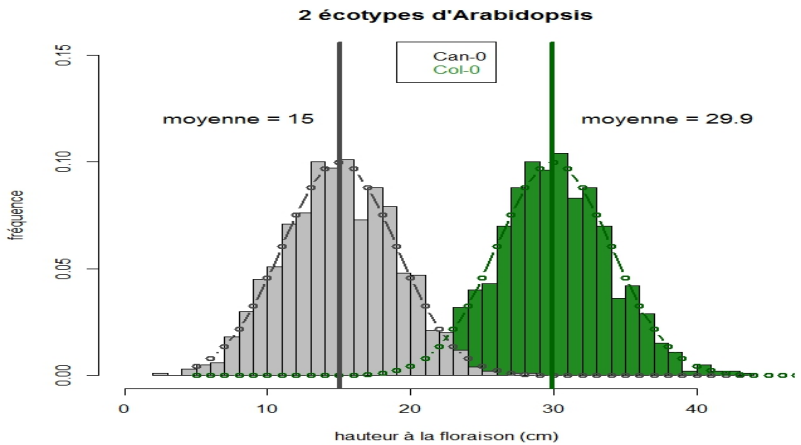


Homogénéité des variances : test F de Fisher ($H0 : \sigma_1^2 = \sigma_2^2$)

- sous $H0$ la statistique $F_{obs} = \frac{\hat{\sigma}_1^2}{\hat{\sigma}_2^2}$ suit une loi $F_{(n_1-1, n_2-1)}$
- $F = 0.959$, $p - value = 0.51 > \alpha$, on accepte $H0$
- détails de la statistique de Fischer ; voir annexe 4. Remarque : sous $H0$ on attend $F_{obs} \sim 1$

Test d'homogénéité : comparaison (variable quantitative) de 2 groupes

Différence de croissance entre 2 écotypes d'*Arabidopsis thaliana* ?



Comparaison des moyennes : test de Student ($H0 : \mu_1 = \mu_2$)

- sous $H0$ (avec n_1 et $n_2 > 30$), $t = \dfrac{\bar{x}_1 - \bar{x}_2}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}}$ suit une loi $\mathcal{N}(0, 1)$
- sous $H0$ (avec n_1 et $n_2 < 30$), $t = \dfrac{\bar{x}_1 - \bar{x}_2}{\hat{\sigma}\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}}$ suit une loi de Student $\mathcal{T}(n_1 + n_2 - 2)$
- $t = -84.04$, $p - value < 2.2x10^{-16} < \alpha$, on rejette $H0$, les moyennes sont significativement différentes. "Col-0" a une croissance significativement plus rapide que "Can-0".
- détails du test de Student ; voir annexe 5. Remarque : sous $H0$ on attend $t \sim 0$

Test d'homogénéité : comparaison (variable quantitative) de > 2 groupes

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** de variation (**variable qualitative nominale**) ou une combinaison de facteurs a un effet sur la **variable quantitative** étudiée
- **les groupes sont des modalités du (des) facteur(s)**. Exemple : facteur "écotype", avec 3 modalités "Col-0","Can-0" et"Ws-0"
- on 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)
  - **ANOVA à 1 facteur** : comparer plusieurs échantillons selon 1 facteur de variation (ex : on compare la croissance pour les 3 "écotypes" différents (= le facteur : "écotype"))
  - **ANOVA à 2 facteurs** : comparer plusieurs échantillons selon 2 facteurs de variation (ex : on compare la croissance pour les 3 écotypes différents (= premier facteur : "écotype"), en présence/absence d'un pathogène (= deuxième facteur : "pathogène"))

Test d'homogénéité : comparaison (variable quantitative) de > 2 groupes

ANOVA à 1 facteur

- on dispose de p échantillons (avec x_{ik} : échantillon i, individu k), d'effectifs n_i et de moyennes $\bar{x}_{i(i=1,\ldots,p)}$
- conditions d'applications
  - **échantillons aléatoires et indépendants**
  - **populations normales et de même variance (homoscédasticité)** (sinon test non paramétrique de Kruskall-Wallis ou transformation de variables ; voir annexe 7 et 8)
- soit n l'effectif total et $\bar{x}$ la moyenne générale, on peut écrire :

$$\sum_{i=1}^{p}\sum_{k=1}^{n_i}(x_{ik}-\bar{x})^2 = \sum_{i=1}^{p} n_i(\bar{x}_i-\bar{x})^2 + \sum_{i=1}^{p}\sum_{k=1}^{n_i}(x_{ik}-\bar{x}_i)^2 \qquad (15)$$

  - $SCE_t = SCE_b + SCE_w$
  - avec : SCE_t la variation totale, SCE_b la variation factorielle (inter-groupes), SCE_w la variation résiduelle (intra-groupes)
- l'analyse de la variance consiste à **comparer la variation factorielle à la variation résiduelle** (avec test de Fisher)
- l'analyse de la variance s'écrit sous la forme d'un modèle linéaire ! (voir annexe 9)

Test d'homogénéité : comparaison (variable quantitative) de > 2 groupes

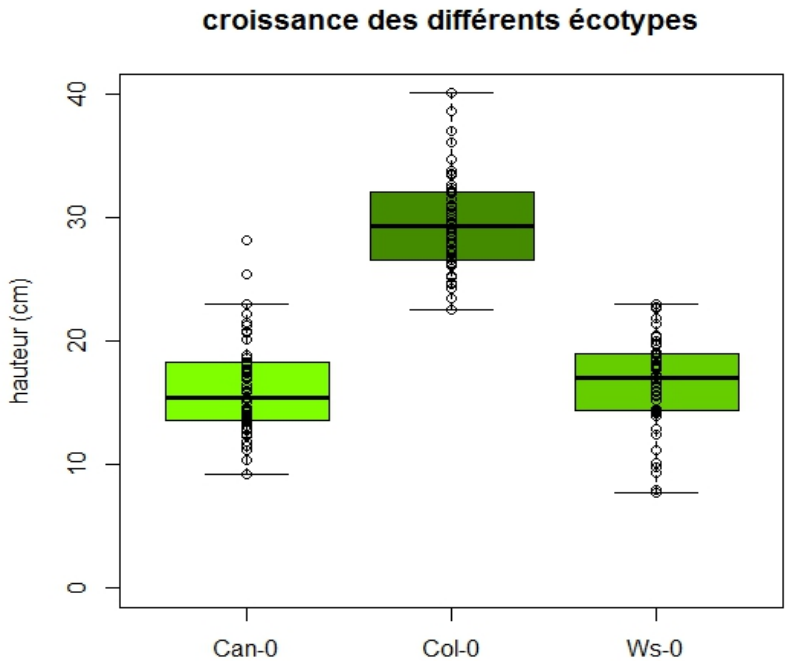
ANOVA à 1 facteur : le tableau !

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	$s_b^2 = \frac{SCE_b}{p-1}$	$\frac{s_b^2}{s_w^2}$	$\mathbb{P}_{H0}(F > F_{obs})$
intra-groupes (résidus)	n-p	SCE_w	$s_w^2 = \frac{SCE_w}{n-p}$		
total	n-1	SCE_t	s_t^2		

- si $H0$ est vraie, F_{obs} suit une loi $F_{(p-1,n-p)}$ ddl
- on rejète $H0$ si $\mathbb{P}(F > F_{obs}) \leq \alpha$

Test d'homogénéité : comparaison (variable quantitative) de > 2 groupes

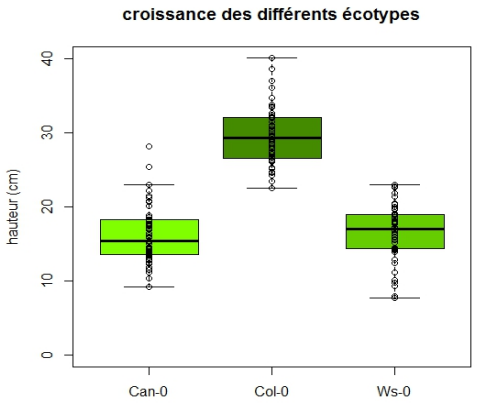
Différence de croissance entre 3 écotypes d'*Arabidopsis thaliana* ?



- normalité (test de Shapiro sur les résidus de l'ANOVA) : $W = 0.9904$, $p - value = 0.3983$, donc on accepte $H0$ = normalité des données de chaque groupe
- homogénéité des variances (test de Bartlett) : $K - squared = 0.1293$, $p - value = 0.9374$, donc on accepte $H0$ = homogénéité
- ==> on peut donc faire l'ANOVA

Test d'homogénéité : comparaison (variable quantitative) de > 2 groupes

Différence de croissance entre 3 écotypes d'*Arabidopsis thaliana* ?



ANOVA à 1 facteur : le tableau

	df	sum of squares	variances	F_{obs}	p-value
écotype	2	5860	2930.1	196.2	$2.2x10^{-16}$***
résidus	147	2196	14.9		

- $p - value < 2.2x10^{-16} < \alpha$, donc on rejette $H0$, il y a un effet du facteur "écotype" sur la croissance des plantes = il y a au moins une moyenne qui diffère significativement des autres
- tests de Student des groupes 2 à 2 : $p - value = 0.67 > \alpha$ uniquement pour "Can-0" *vs* "Ws-0", donc c'est "Col-0" qui diffère significativement (meilleure croissance ici)

Test d'homogénéité : comparaison (variable quantitative) de > 2 groupes

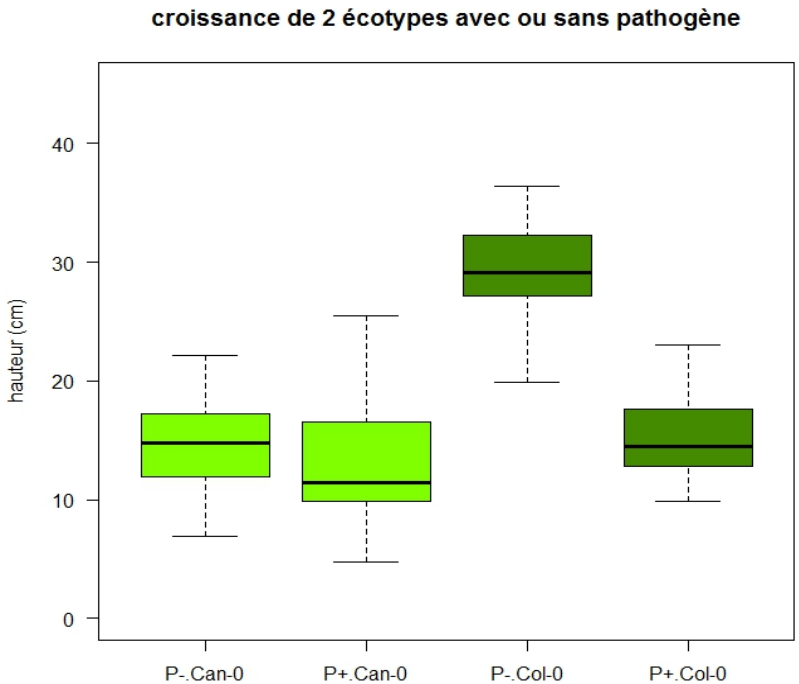
ANOVA à 2 facteurs : le tableau !

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})** (il y a un effet d'un facteur mais pas sur toutes les modalités de l'autre facteur)
- en l'absence d'interaction ($\mathbb{P}(F > F_{AB}) > \alpha$) on peut tester les effets principaux (F_A et F_B)

Test d'homogénéité : comparaison (variable quantitative) de > 2 groupes

Différence de croissance en fonction de l'écotype et de la présence/absence du pathogène ?



- normalité des groupes : $W = 0.9891, p - value = 0.5911$, on accepte $H0$
- homogénéité des variances : $K - squared = 2.5411, p - value = 0.4679$, on accepte $H0$
- ==> on peut donc faire l'ANOVA

Test d'homogénéité : comparaison (variable quantitative) de > 2 groupes

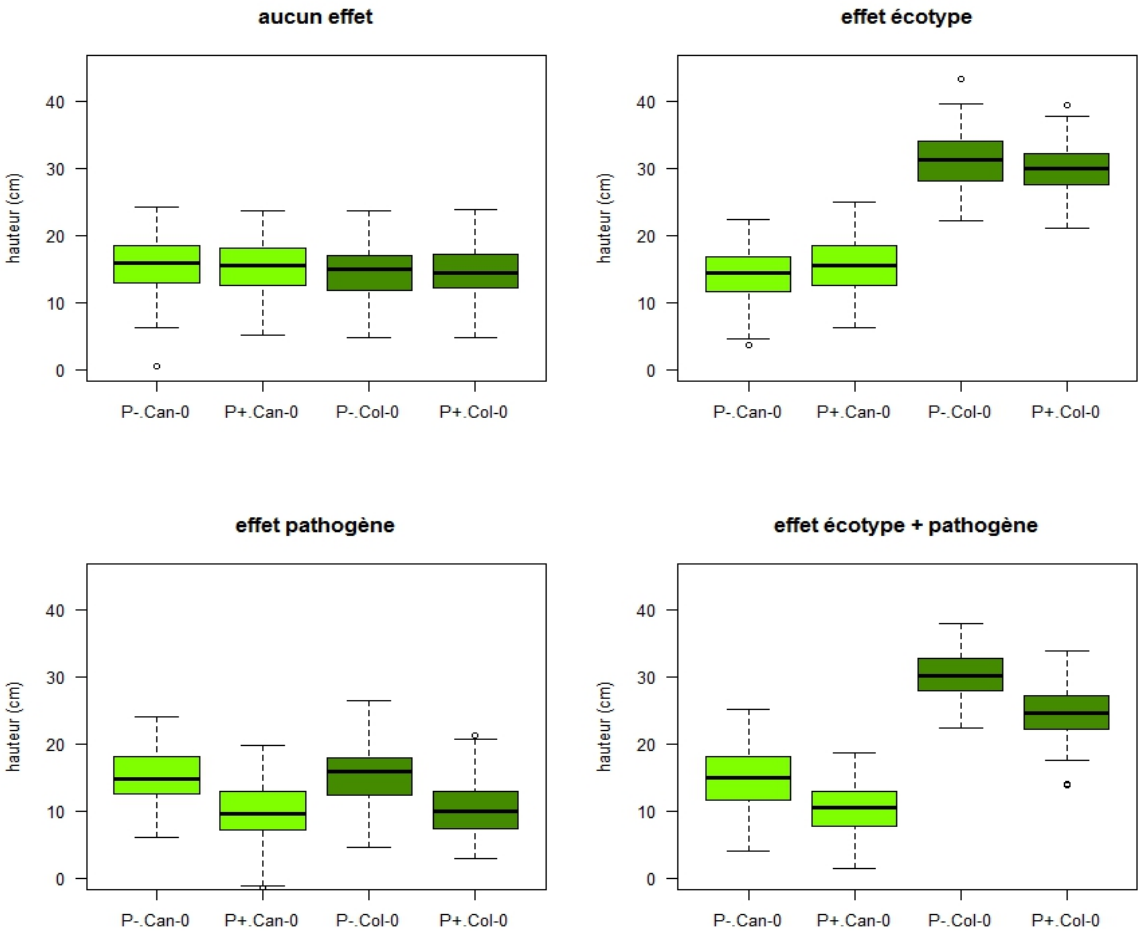
ANOVA à 2 facteur : le tableau

	df	sum of squares	variances	F_{obs}	p-value
écotype	1	1763	1763	93.2	$8.2x10^{-16}$***
pathogène	1	1522	1522	80.48	$2.4x10^{-14}$***
écotype : pathogène	1	935	935	78.4	$3x10^{-10}$***
résidus	144	2668	19		

- interaction écotype x pathogène significative ($p - value < 3x10^{-10} < \alpha$)
- (==> l'effet du pathogène est écotype-dépendant)
- tests 2 à 2 (ex : Tukey Honest Significant Difference) :
  - effet néfaste du pathogène pour "Col-0" ($p - value < 10^{-7}$), PAS pour "Can-0" ($p - value = 0.52 > \alpha$)
  - croissance significativement meilleure pour "Col-0" en absence du pathogène ($p - value < 10^{-7} < \alpha$)

Test d'homogénéité : comparaison (variable quantitative) de > 2 groupes

Les autres cas possibles avec une ANOVA à 2 facteurs



Notion de correction pour les tests multiples

Exemple : on compare l'expression génique entre 2 conditions expérimentales pour des milliers de gènes

(cf cours de Roland Barriot)

- pour chaque gène, je compare les 2 moyennes d'expression (test de Student)
  - chaque test est réalisé à un risque α
  - si l'on réalise G test indépendants (pour les G gènes) ALORS le nombre moyen de faux positifs $= G\alpha$
  - si $G = 10000$ et $\alpha = 0.05$, alors 500 gènes sont déclarés différentiellement exprimés à tort !
- test multiple : on cherche à contrôler le risque de faux-positifs associé au grand nombre de tests effectués
- 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- ($G = N + P$)

Notion de correction pour les tests multiples

2 **procédures possibles**

- Contrôle du Family-wise error rate (FWER) :
  - **FWER = probabilité d'avoir au moins un faux positif (sur l'ensemble des gènes testés)**
  - **correction la plus connue : Bonferroni**. Si $G = 10000$ et $FWER \leq 0.05$, alors chaque test est réalisé avec un risque $\alpha' = \frac{\alpha}{G} = 5x10^{-6}$
  - mais plus il y a de tests, moins on rejette $H0$ = très conservatifs (peu de gènes sont déclarés différentiellement exprimés)
  - à utiliser plutôt pour rechercher des gènes candidats sérieux pour des analyses fonctionnelles
- False Discovery Rate (FDR) :
  - **l'idée est plutôt de contrôler le nombre moyen de faux-positifs parmi les gènes déclarés différentiellement exprimés**
  - contrôler le FDR à un niveau $q = 5\%$ signifie que 5% des gènes déclarés positifs sont des faux positifs
  - moins conservatif (plus de faux-positifs), mais aussi plus puissant (plus de vrai positifs)
  - à utiliser quand l'objectif de l'expérience transcriptomique est exploratoire (détecter des groupes de gènes)

Comment décrire de grands tableaux de données : ANALYSE MULTIVARIEES

Analyse en Composantes Principales -ACP- (variables quantitatives)

- on souhaite décrire un tableau (n) individus x (p) variables quantitatives
- l'ACP permet de représenter graphiquement :
  - les corrélations entre variables
  - les ressemblances entre individus (notion de distance)
- l'ACP transforme un grand nombre des variables "corrélées" en un plus petit nombre de variables indépendantes les unes des autres, les "composantes principales", ou "axes" (=réduction de dimensionnalité)
- on visualise les variables initiales et les individus dans le plan factoriel (les 2 premiers axes) :
  - dans le cercle des corrélations, cosinus = coefficient de corrélation
  - dans le plan des individus, les individus ressemblants sont proches
- chaque axe explique une proportion de la variabilité du tableau initial
- remarque : analyses impliquant des calculs matriciels complexes (calculs de valeurs propres et de vecteurs propres)

Analyse en Composantes Principales -ACP- (variables quantitatives)

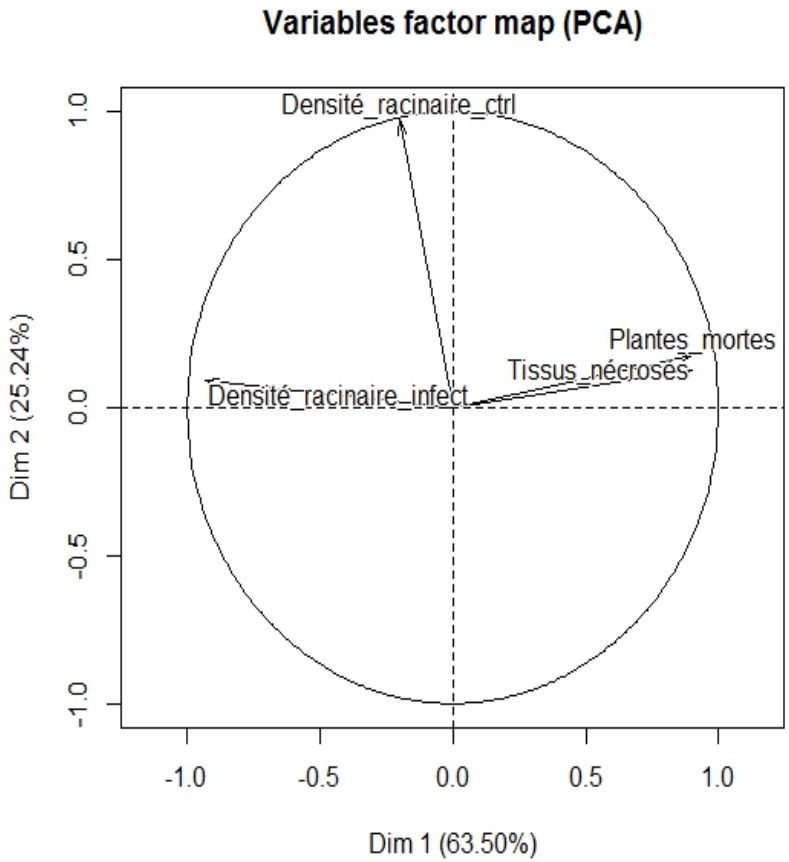
Mesures de symptômes et de développement lors de la réponse à un pathogène racinaire

	Variable 1	Variable 2	Variable 3	Variable 4....
individu 1				
individu 2				
individu 3				
individu 4				
individu 5				
individu 6				
individu 7				
individu 8				
individu 9				
individu 10				
individu 11				
individu 12				
individu 13				
individu 14				
individu 15				
individu 16				
individu 17				
individu 18				
individu 19				
individu 20				
...				

	%_tissus_nécrosés	%_plantes_mortes	Densité_racinaire (infection)	Densité_racinaire (contrôle)
1	100	0.72	10.6	14.8
2	68	0	12	5.9
3	100	0.17	4.6	7
4	50	0	10.6	7.1
5	71	0.04	9.5	5.2
6	100	0.45	4.1	11.4
7	63	0	10.6	9.2
8	62	0	13	10.2
9	100	0.5	1.9	3
10	100	0.79	0.4	9.9
11	100	0.42	7.8	5.1
12	100	0.85	2	7.4
13	57	0	7.7	6.4
14	100	0.38	2.6	7.8
15	100	1	1.2	9.8
16	100	0.75	1	6.6
19	100	0.71	0.2	7.9
20	100	0.43	0.7	7.6
21	100	0.67	4	8.9
23	69	0	7.8	8.7
...	...	...	...	...

Analyse en Composantes Principales -ACP- (variables quantitatives)

Mesures de symptômes et de développement lors de la réponse à un pathogène racinaire



Individuals factor map (PCA)

Analyse Factorielle des Correspondances -AFC- (table de contingence)

- l'AFC s'utilise avec 2 variables qualitatives qui possèdent 2 ou plus de 2 modalités
- l'objectif est d'étudier les "correspondances" entre les modalités de chaque variables
- l'AFC peut être vue comme une ACP avec une distance particulière : la métrique du χ^2
- la visualisation simultanée (en 2 dimensions) des modalités de chaque variable permet de représenter les écarts relatifs à l'indépendence entre les deux variables
- remarque :
  - on peut faire une AFC avec les données prévues pour une ACP (on traite les valeurs numériques comme des catégories)
  - on ne peut pas faire une ACP avec des données prévues pour une AFC

Analyse Factorielle des Correspondances -AFC- (table de contingence)

Fréquence des codons chez différentes espèces de Bactéries et d'Archaea

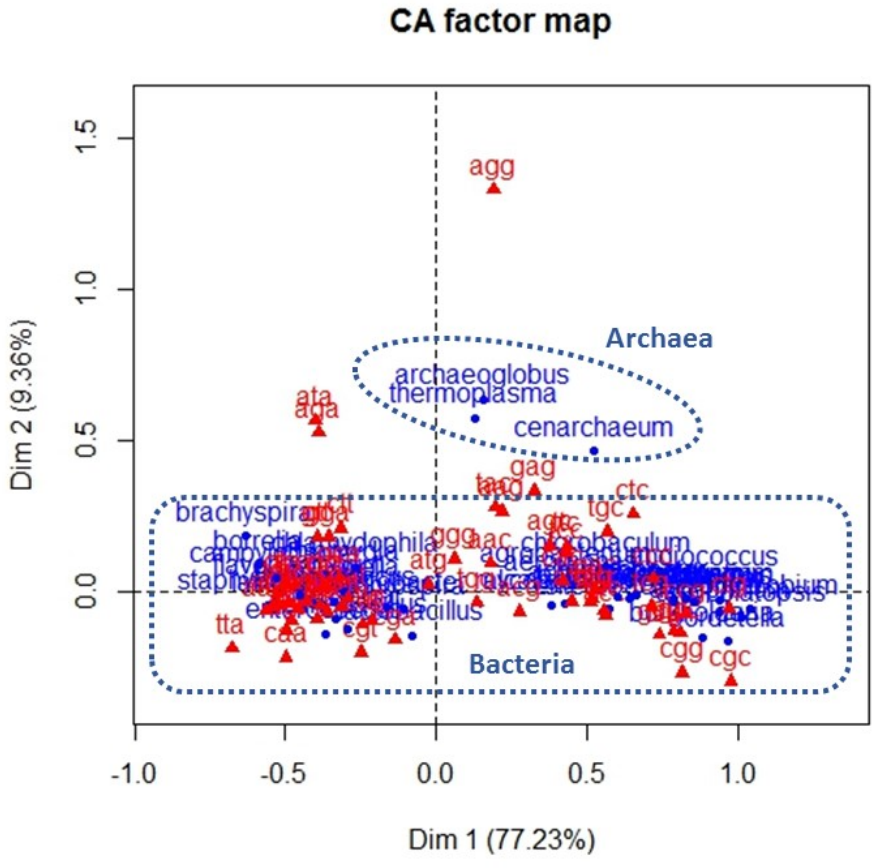
Variable Y	Variable X	Modalité 1	Modalité 2	Modalité 3	Modalité 4...
	Modalité 1				
	Modalité 2				
	Modalité 3				
	Modalité 4				
	Modalité 5				
	Modalité 6				
	Modalité 7				
	Modalité 8				
	Modalité 9				
	Modalité 10				
	Modalité 11				
	Modalité 12				
	Modalité 13				
	Modalité 14				
	Modalité 15				
	Modalité 16				
	Modalité 17				
	Modalité 18				
	Modalité 19				
	Modalité 20				
	...				

fréquence du codon dans l'espèce

Espèce	Codon	aaa	aac	aag	aat	aca	acc	acg	act	aga
	achromobacter	216	417	691	149	134	428	429	92	48
	acinetobacter	3442	1233	1429	2590	1271	1220	908	1266	462
	actinobacillus	7536	2049	1645	5197	1597	1813	1483	1824	936
	aeromicrobium	0	215	197	6	22	598	529	15	8
	aeromonas	2879	4170	4937	2578	1005	5352	1255	1264	754
	agrobacterium	1295	1809	2415	1253	801	1938	1714	567	495
	alcaligenes	702	741	876	589	341	910	598	288	124
	allochromatium	92	578	802	137	35	784	263	44	13
	amycolatopsis	26	273	290	15	24	592	426	20	12
	anabaena	1664	688	506	1756	994	415	171	1100	314
	angiococcus	66	562	921	128	115	962	765	81	71
	archaeoglobus	30604	28630	56856	12805	12839	15420	15264	10378	28331
	arthrobacter	487	935	893	270	336	1688	707	238	119
	arthrospira	1161	485	460	889	291	576	221	527	213
	azoarcus	80	650	1066	120	21	859	637	24	13
	azorhizobium	95	398	679	245	60	769	448	56	61
	azospirillum	218	2395	2998	268	171	3367	1901	144	64
	azotobacter	472	2319	2529	416	106	3077	565	238	71
	bacillus	240429	56603	77692	143413	119725	11428	57655	54100	40562
	Bartonella	653	124	307	537	377	100	143	235	74
	...	...	...	...	...	...	...	...	...	...

Analyse Factorielle des Correspondances -AFC- (table de contingence)

Fréquence des codons chez différentes espèces de Bactéries et d'Archaea



ANNEXES

ANNEXE 1 : Echantillonnage - Estimation

Théorème central limite (TCL) : démonstration

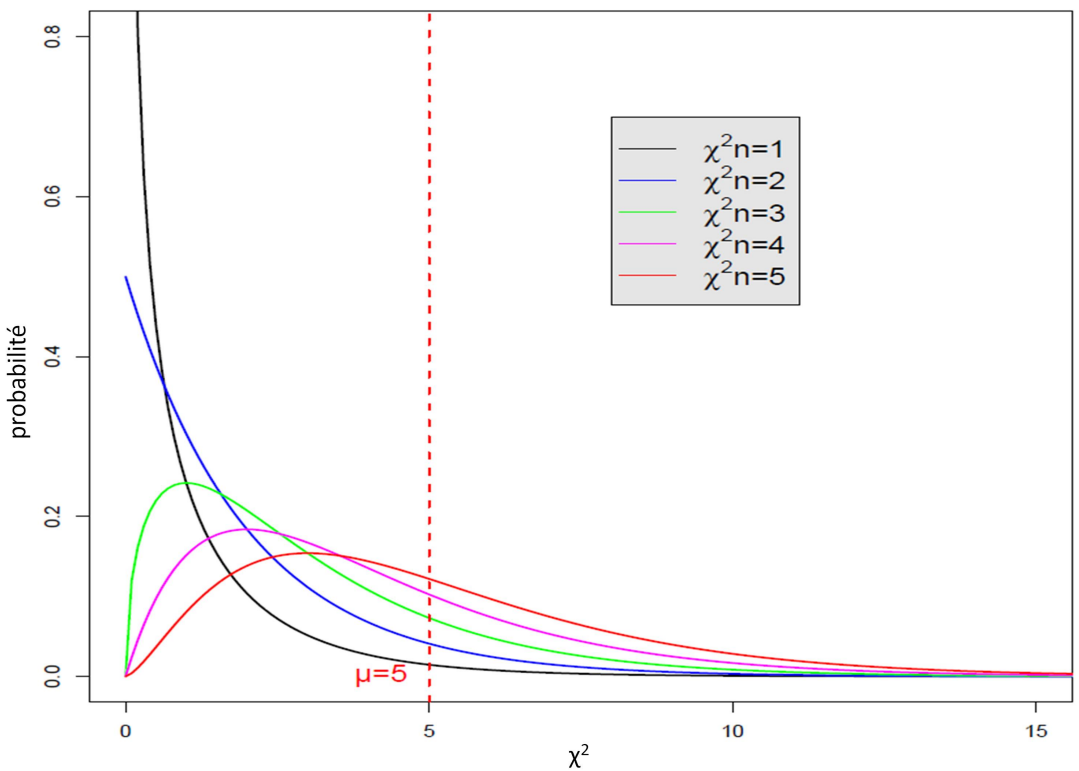
- le TCL établit que toute somme de variables aléatoires indépendantes et identiquement distribuées **(et par extension toute moyenne ou variance)** converge vers une variable aléatoire gaussienne
  - si X suit une loi d'espérance μ et d'écart-type σ, et si $S_n = X_1 + X_2 + ...X_n$
  - alors $\mathbb{E}(S_n) = n\mu$ et $\mathrm{Var}(S_n) = \sigma^2 n$
  - et $\mathbb{E}(\frac{S_n}{n}) = \mu$ et $\mathrm{Var}(\frac{S_n}{n}) = \frac{\sigma^2}{n}$
  - si $n \to +\infty$ alors $m \to \mathcal{N}(\mu, \frac{\sigma}{\sqrt{n}})$ (convergence en loi)
- $m \to \mathcal{N}(\mu, \frac{\sigma}{\sqrt{n}})$ (distribution de m sur tous les échantillons, avec $\frac{\sigma}{\sqrt{n}}$ = erreur standard de la mesure -standard error of the mean- SEM)

ANNEXE 2 : Test sur les effectifs (variable qualitative)

Loi de χ^2

soient n v.a. indépendantes $X_1, X_2, X_3, ..., X_n$ de loi $\mathcal{N}(0,1)$. La v.a. $\chi^2 = \sum_i^n X_i^2$ suit une loi du χ^2 à n degrés de liberté

$\mathbb{E}(\chi^2) = n$ (=nombre de v.a.) ; $\mathrm{Var}(\chi^2) = 2n$



ANNEXE 3 : 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

ANNEXE 4 : Test d'homogénéité : 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

ANNEXE 5 : Comparaison moyennes de 2 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 $U_{obs} > U_{1-\alpha}$

ANNEXE 5 : Comparaison moyennes de 2 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

ANNEXE 6 : 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)

ANNEXE 7 : tests non paramétriques

test non paramétrique : test de Kruskall-Wallis

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

ANNEXE 8 : 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

ANNEXE 9 : Analyse de la variance (ANOVA)

ANOVA à un facteur : modèle linéaire théorique

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

- **le modèle théorique à effet(s) fixe(s) :**
  - $X_{ik} = \mu_i + \epsilon_{ik}$
  - $X_{ik} = \mu + \alpha_i + \epsilon_{ik}$
  - où α_i : non aléatoires, $\epsilon_{ik} \sim \mathcal{N}(0, \sigma)$
  - $H0 : \alpha_1 = \alpha_2 = ... = \alpha_p = 0$
- **le modèle théorique à effet(s) aléatoire(s) :**
  - $X_{ik} = \mu + A_i + \epsilon_{ik}$
  - où $A_i \sim \mathcal{N}(0, \sigma_a)$, $\epsilon_{ik} \sim \mathcal{N}(0, \sigma)$
  - $H0 : \sigma_a^2 = 0$
- **modèle à effet fixe ou aléatoire ?**
  - effet fixe : si je suis intéressé par l'effet de chaque niveau du facteur, et si il y a un faible nombre de niveaux de facteurs
  - effet aléatoire : s'il y a potentiellement une "population" de niveaux de facteurs, et si je ne suis pas directement intéressé par l'effet de chaque niveau mais plutôt par la distribution des effets (variance)