Traitement des Données Biologiques : bases statistiques

M1 - MABS

Maxime Bonhomme

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

12 septembre 2011

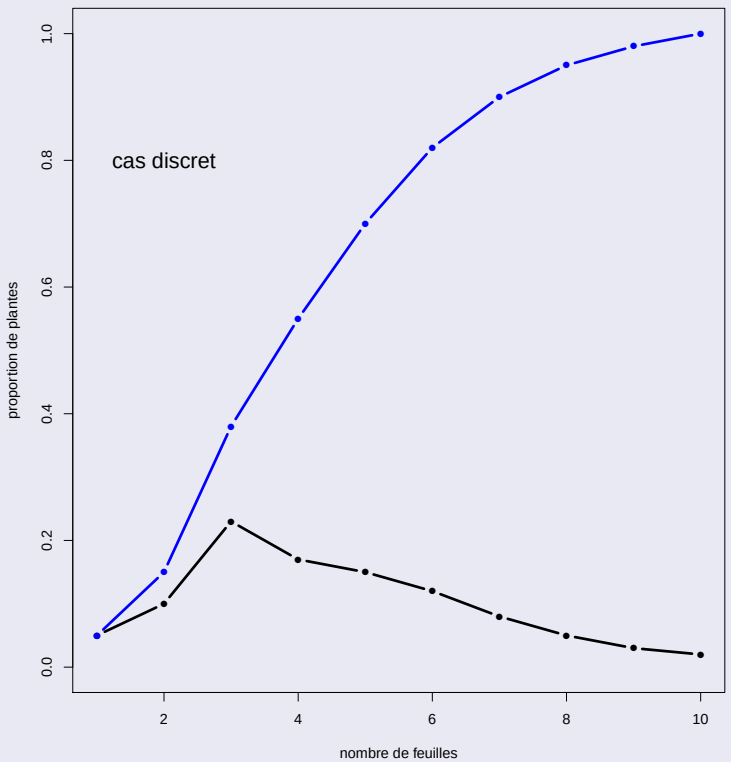
Bases statistiques pour le TDB

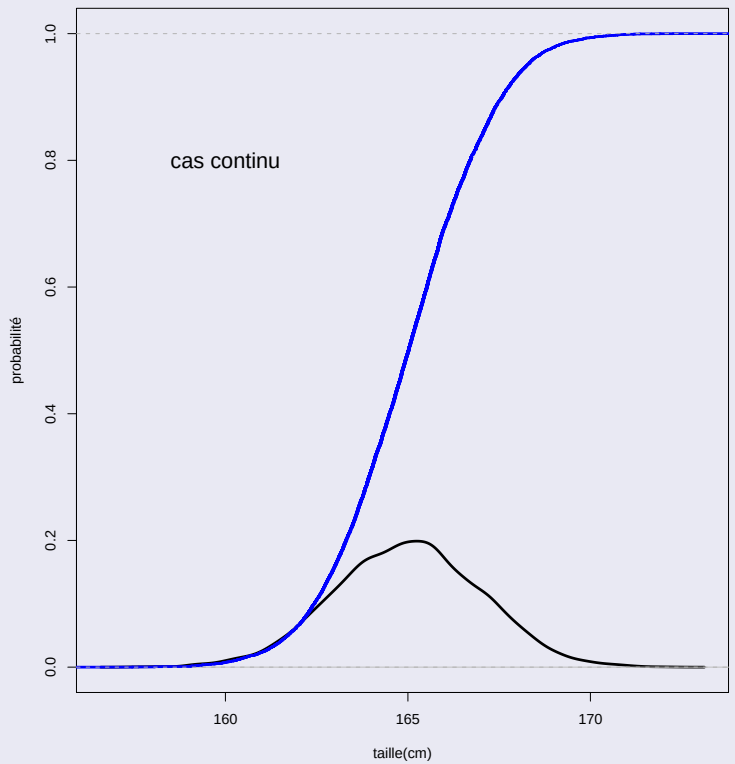
1. statistique inférentielle
   - lois de probabilités
   - tests d'hypothèse

STATISTIQUE INFERENTIELLE

Distribution (loi) de probabilité

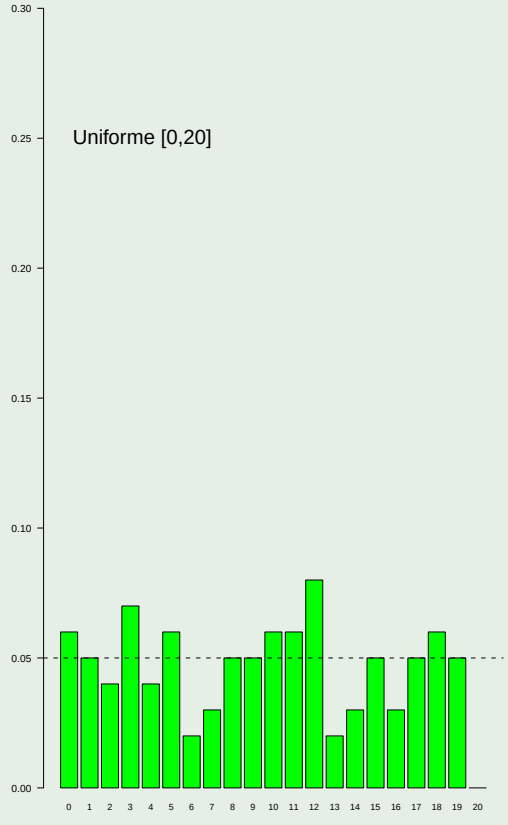
- fonction donnant une probabilité à chaque valeur ou intervalle de valeurs d'une variable aléatoire quantitative X (avec $\mathbb{P}(\Omega) = \sum_{i=1}^{n} p_i = 1$)
  - $f(k) = \mathbb{P}[X = k]$ (densité de probabilité)
  - $F(k) = \mathbb{P}[X \leq k]$ (fonction de répartition, cumulative, $\mathbb{P}[x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

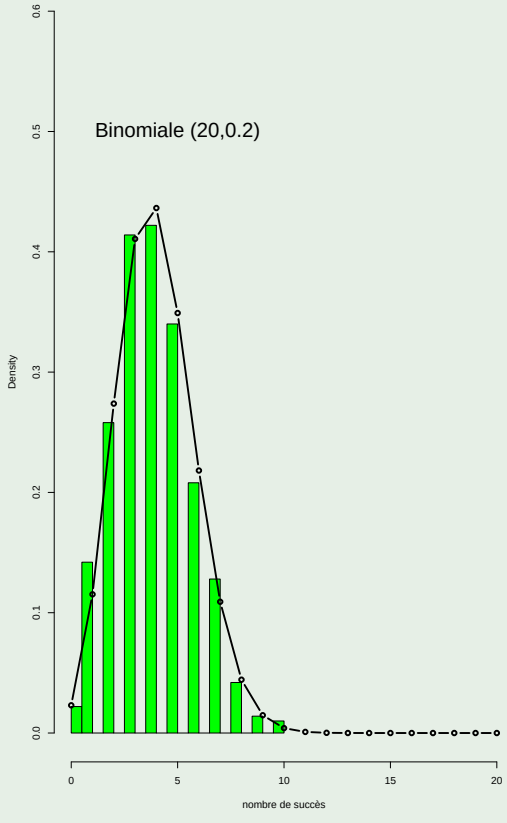


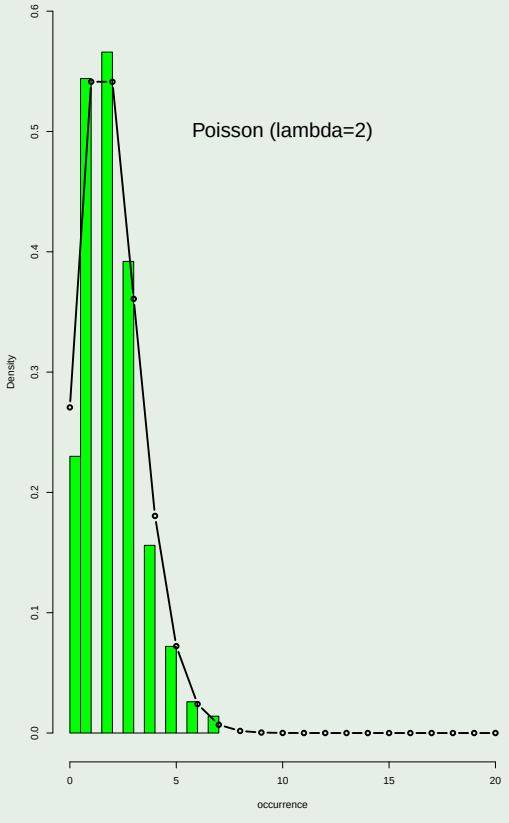


Exemple de lois de probabilité

- discrètes

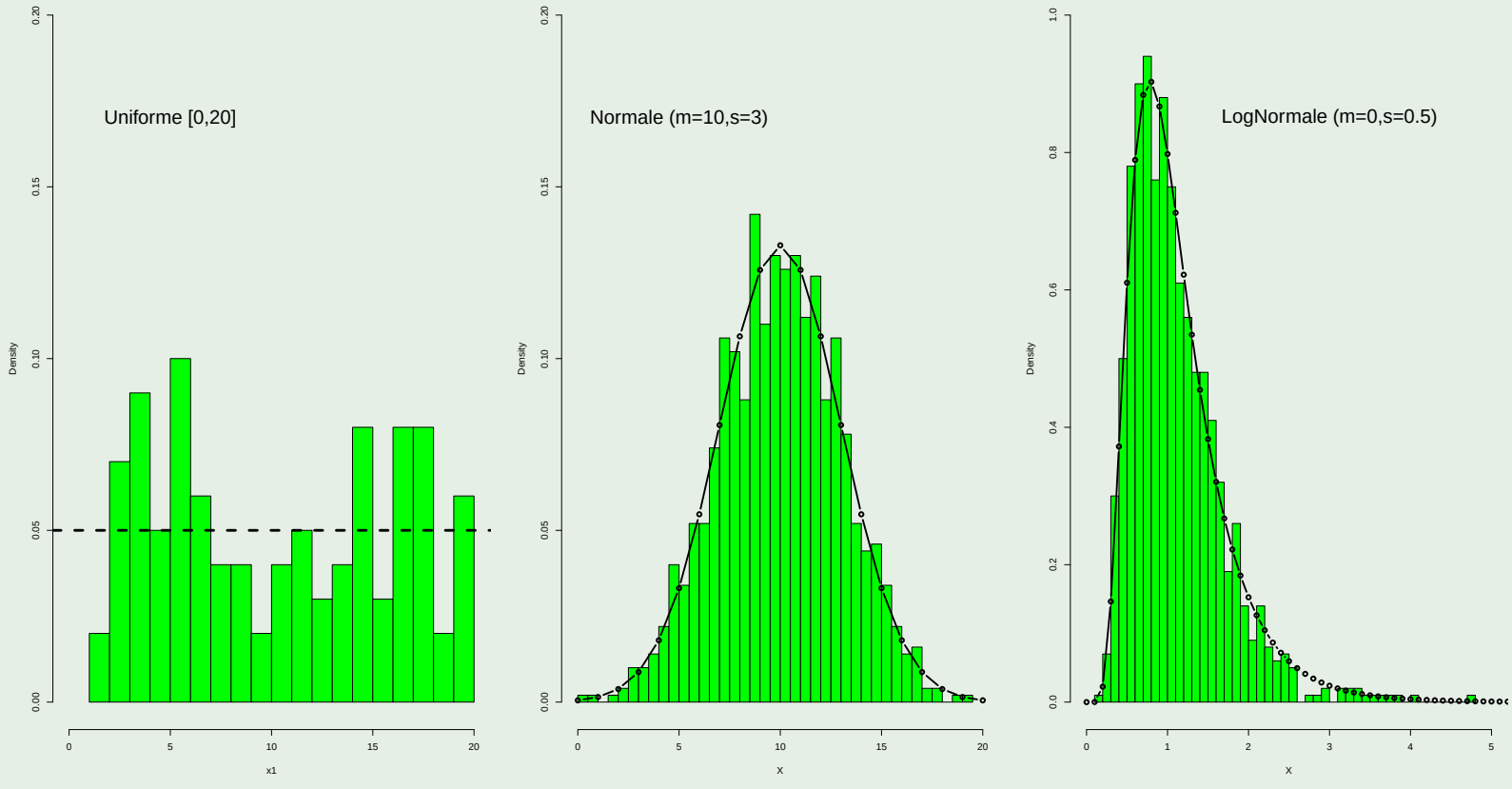






Exemple de lois de probabilité

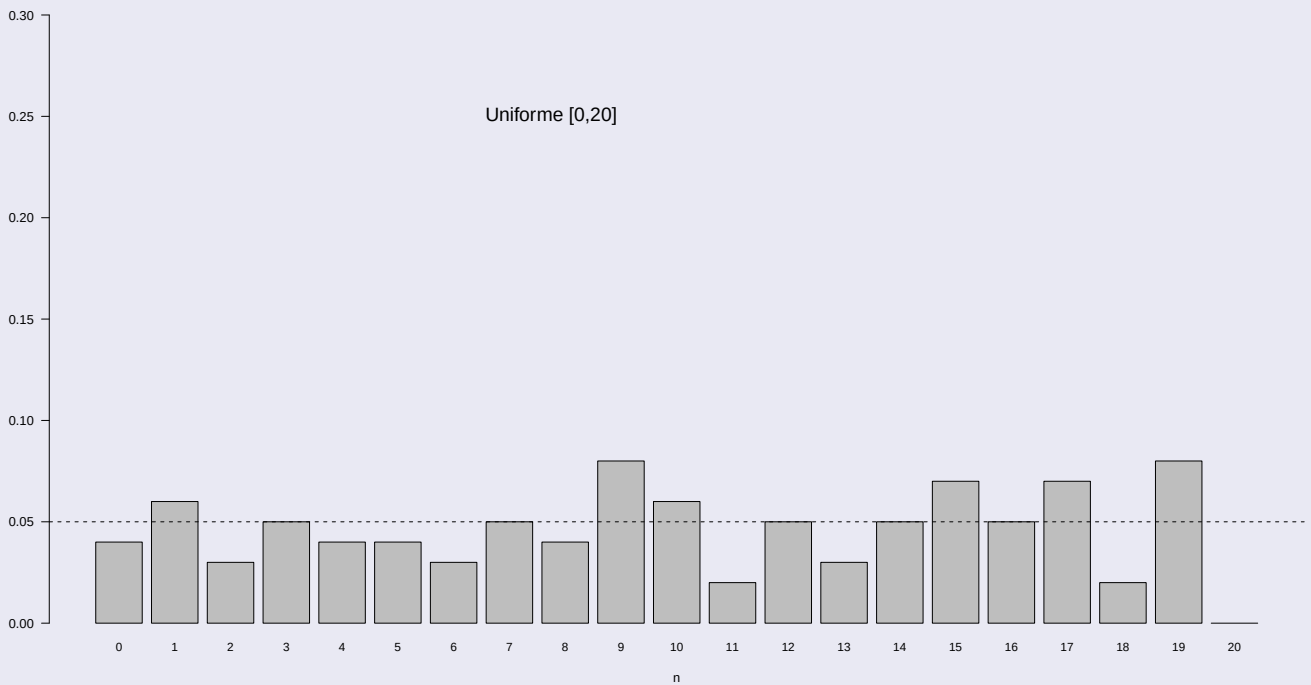
- continues



Lois discrètes : uniforme

loi uniforme : équiprobabilité entre chaque valeurs d'un ensemble fini

- $X = 1, 2, 3, ...n$, $\mathbb{P}(X = 1) = \mathbb{P}(X = 2) = ...\mathbb{P}(X = n) = \frac{1}{n}$
- $\mathbb{E}(X) = \frac{n+1}{2}$; $\mathrm{Var}(X) = \frac{n^2-1}{12}$
- exemples : jeu de dés, roulette...



Lois discrètes : Bernouilli

Bernouilli : expérience à deux issues (succès - échec)

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

X	$\mathbb{P}(X_i)$	$X_i\mathbb{P}(X_i)$
0	$1-p$	0
1	p	p

$\mathbb{E}(X) = 0 + p = p$

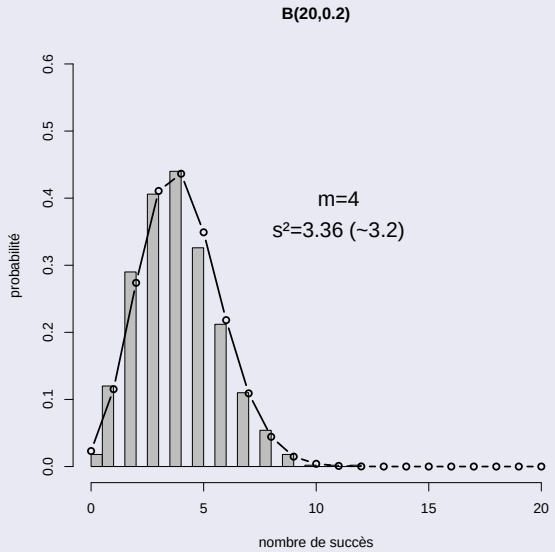
X	$X_i - \mu$	$(X_i-\mu)^2$	$\mathbb{P}(X_i)$	$\mathbb{P}(X_i)(X_i-\mu)^2$
0	$-p$	p^2	$1-p$	$p^2(1-p)$
1	$1-p$	$(1-p)^2$	p	$(1-p)^2 p$

$\mathrm{Var}(X) = p^2(1-p) + (1-p)^2 p = p(p - p^2 + 1 - 2p + p^2) = p(1-p) = pq$

Lois discrètes : Binomiale

loi Binomiale $B(n,p)$, loi d'une somme de variables de Bernouilli

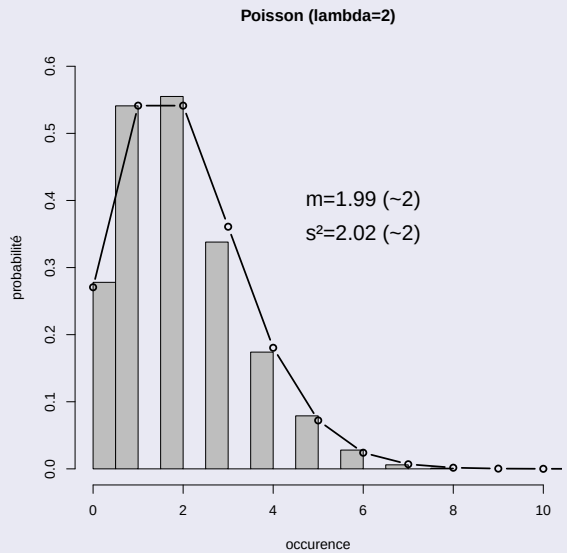
- on s'intéresse au nombre de succès k dans une expérience de n essais (épreuves de Bernouilli) avec probabilité de succès p
- $X \in \{0,...,n\}$; $\mathbb{P}(X=k) = C_n^k p^k (1-p)^{n-k}$, avec $C_n^k = \binom{n}{k} = \frac{n!}{k!(n-k)!}$
  (nombre d'échantillons possibles de k succès dans n tirages)
- $\mathbb{E}(X) = \mu = np$; $\mathrm{Var}(X) = \sigma^2 = np(1-p) = npq$
- exemples
  - si la probabilité de germination $= p$, le nombre de graines germées dans un échantillon de taille n suit une loi binomiale $B(n,p)$
  - évolution de la fréquence de deux allèles par dérive génétique dans une population (tirage avec remise avec probabilité de tirage de l'allèle $1 = p$, changeant à chaque tirage)



Lois discrètes : loi de Poisson

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

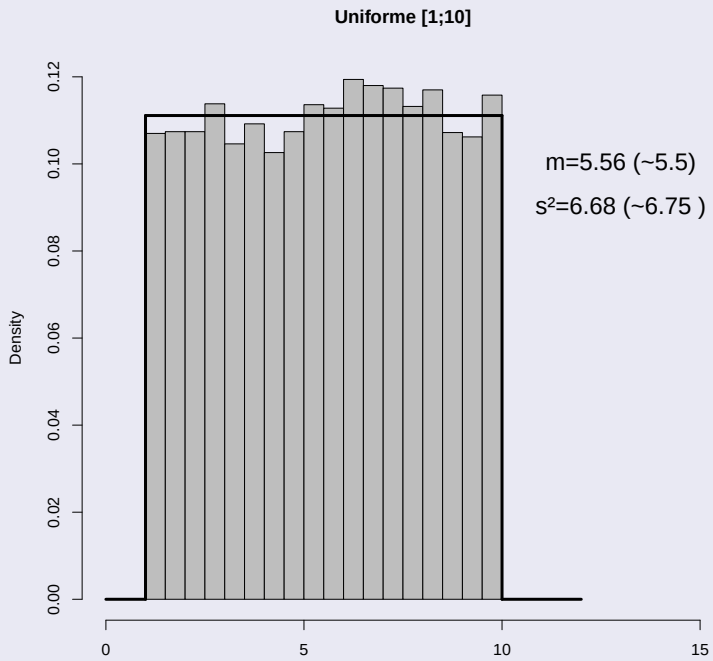
- loi de probabilité d'un événement rare, d'occurence moyenne $\lambda = np$ sur un intervalle (temps, espace) donné
- $X \in \mathbb{N}$; $\lambda \in \mathbb{R}_{\geq 0}$; $\mathbb{P}(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$
- $\mathbb{E}(X) = \lambda$; $\mathrm{Var}(X) = \sigma^2 = \lambda$
- exemple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



Lois continues : loi uniforme

loi uniforme : équiprobabilité entre chaque valeurs d'un ensemble fini

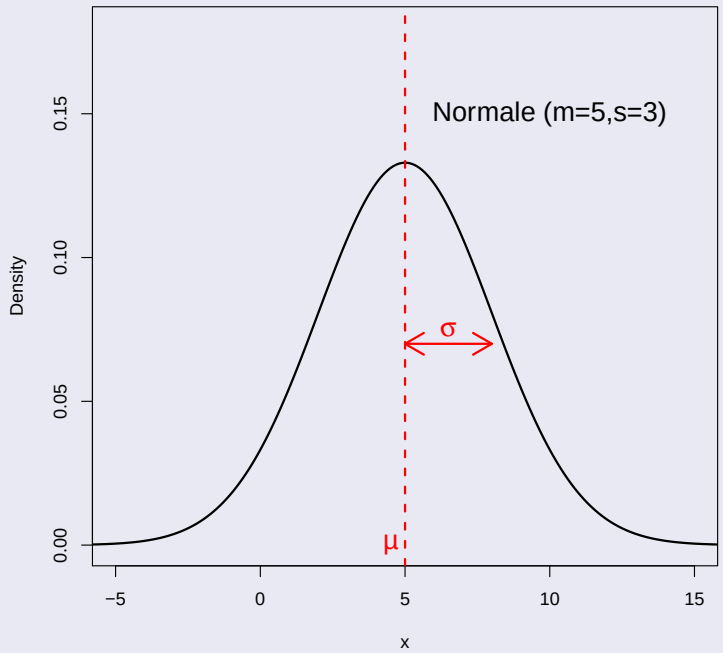
- sur un intervalle borné $[a; b]$ (généralisation de la fonction rectangle)
- $$f(x) = \begin{cases} \frac{1}{b-a} & \text{si } x \in [a; b] \\ 0 & \text{sinon} \end{cases}$$
- $\mathbb{E}(X) = \frac{a+b}{2}$; $\mathrm{Var}(X) = \frac{(b-a)^2}{12}$



Lois continues : loi Normale

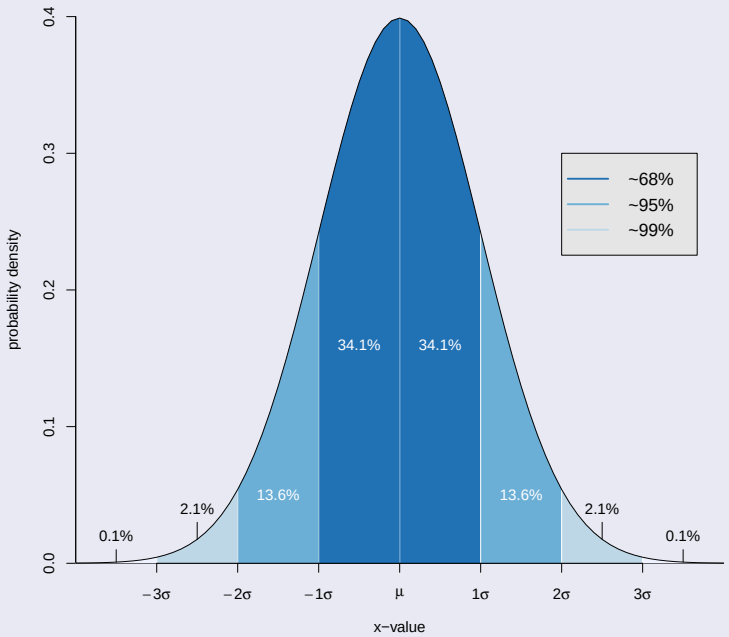
loi Normale ou de Laplace-Gauss $\mathcal{N}(\mu, \sigma)$ ("Gaussienne")

- densité : $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$
- $\mathbb{E}(X) = \mu$; $\mathrm{Var}(X) = \sigma^2$
- approximation de certaines lois (Binomiale, Poisson), dans des conditions particulières. Par exemple, la loi binomiale converge vers une loi normale quand le nombre n d'épreuves augmente. Pour n grand et p, $1-p$ de même ordre de grandeur, approximation d'une $B(n,p)$ par une $\mathcal{N}(np, \sqrt{np(1-p)})$

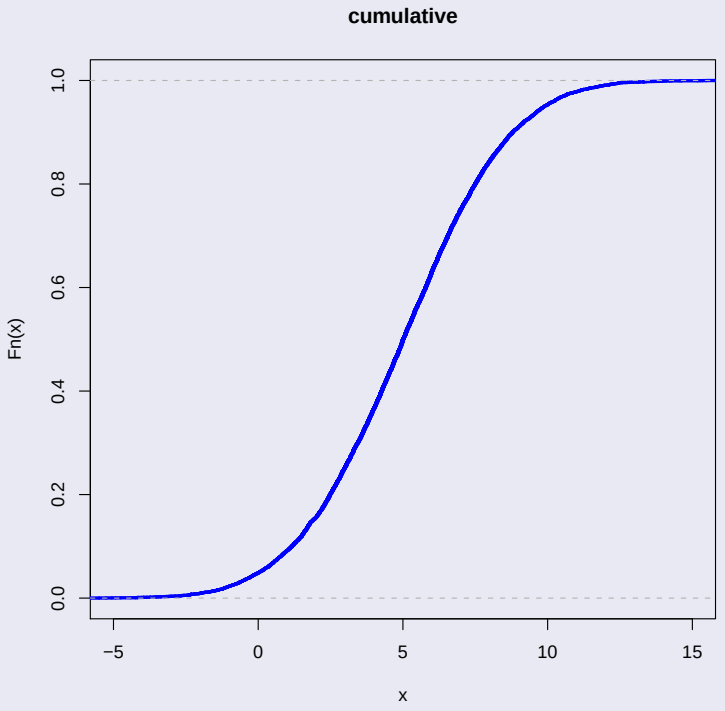


Lois continues : loi Normale

- **valeurs remarquables**



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



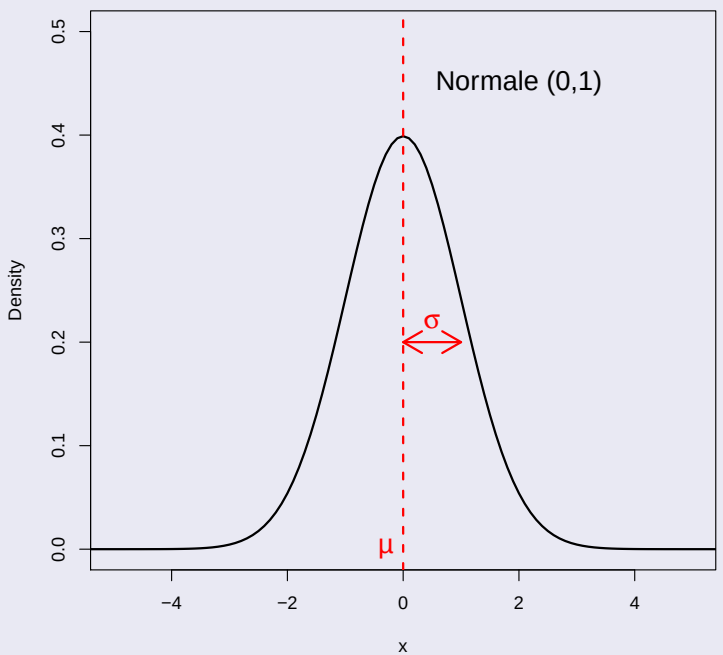
- propriétés :
  - si $X \sim \mathcal{N}(\mu, \sigma)$ alors $Y = a + bX \sim \mathcal{N}(a + b\mu, b\sigma)$
  - si $X_1 \sim \mathcal{N}(\mu_1, \sigma_1)$ et $X_2 \sim \mathcal{N}(\mu_2, \sigma_2)$ alors

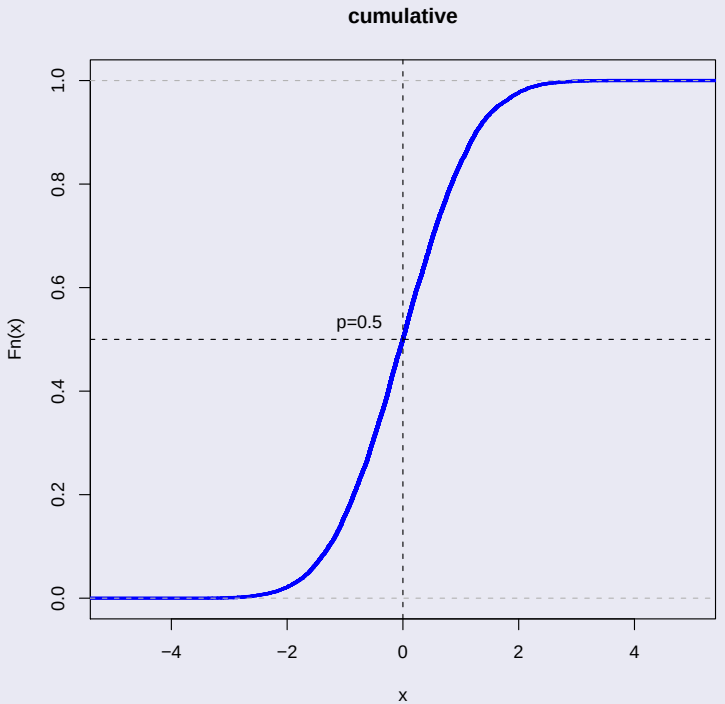
    $X_1 + X_2 \sim \mathcal{N}(\mu_1 + \mu_2, \sqrt{(\sigma_1^2 + \sigma_2^2)})$

Lois continues : loi Normale

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

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





- Gaussienne dont on connait parfaitement les probabilités associées à chaque valeur de X
- 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, il est d'usage courant d'utiliser la table qui donne la fonction de répartition de la variable normale centrée réduite notée T

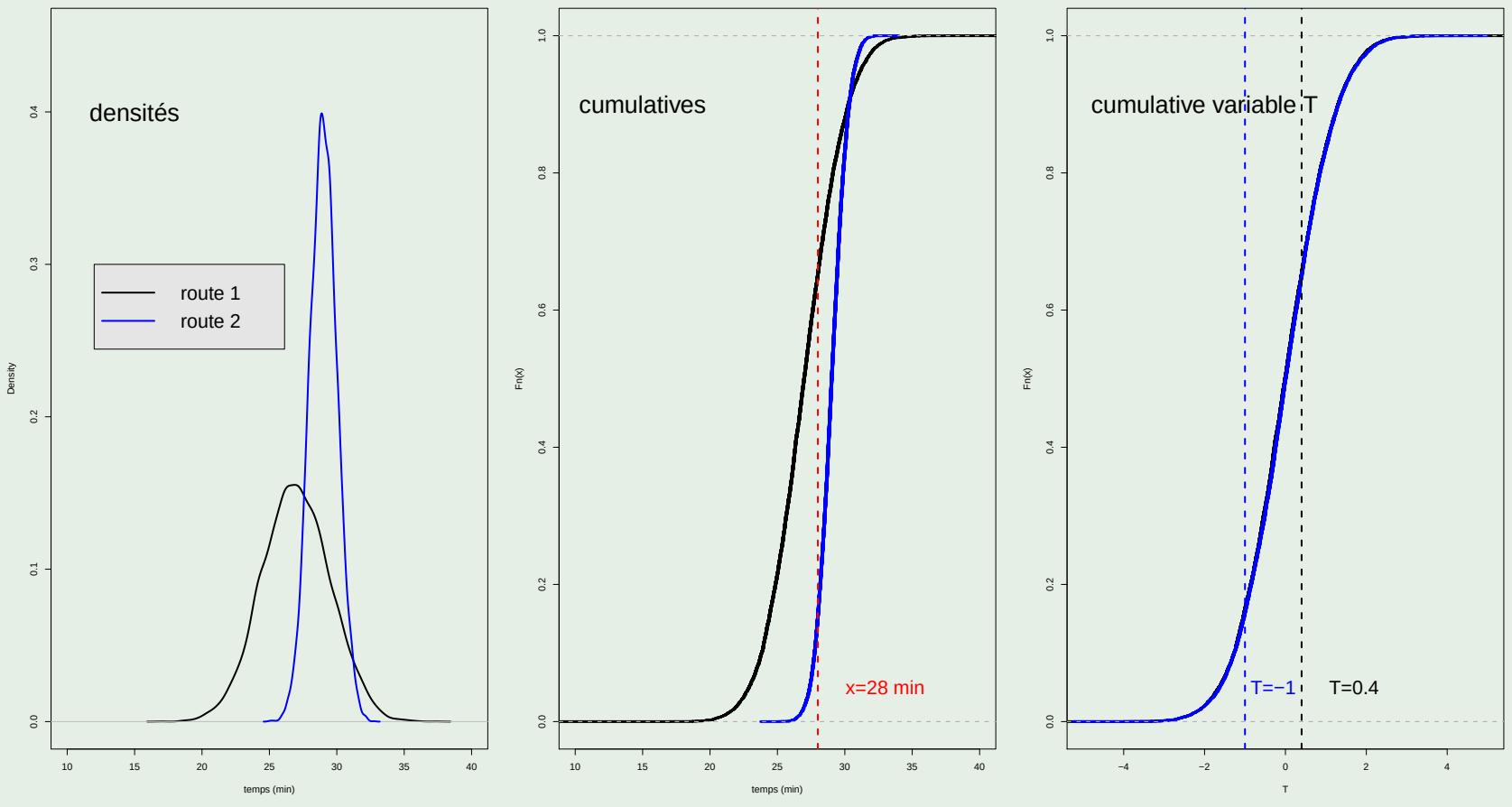
Lois continues : loi Normale

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

- pour ce rendre à son travail, un employé peut prendre 2 routes :
  - route 1 : il met 27 min en moyenne avec un s de 2,5 (2min30) - loi normale
  - route 2 : il met 29 min en moyenne avec un s de 1 (1min) - loi normale
- l'employé veut une probabilité d'arriver à l'heure la plus forte. Quelle est la meilleure route pour se rendre au travail s'il dispose de moins de 28 minutes ?
  - $\frac{D_1-27}{2.5} = \mathcal{N}(0,1)$, $\frac{D_2-29}{1} = \mathcal{N}(0,1)$
  - $\mathbb{P}(D_1 < 28) = \mathbb{P}(\frac{D_1-27}{2.5} < \frac{28-27}{2.5}) = \mathbb{P}(T < 0.4) = 0.665$
  - $\mathbb{P}(D_2 < 28) = \mathbb{P}(\frac{D_2-29}{1} < \frac{28-29}{1}) = \mathbb{P}(T < -1) = 0.159$
- CCL : par la route 1, il a 66% de chances d'arriver à l'heure et seulement 16% par la route 2 : il choisit la route 1

Lois continues : loi Normale

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



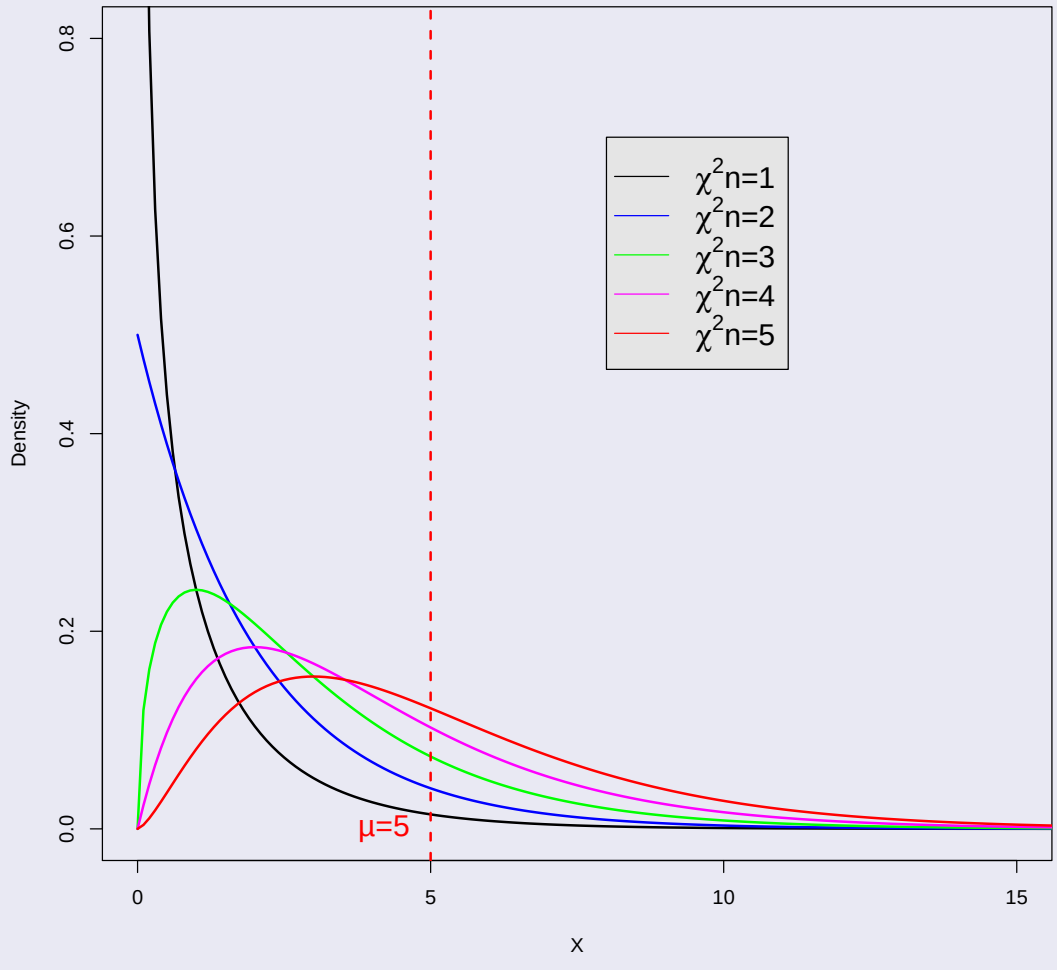
Lois continues : loi du χ^2

loi du χ^2

- soient n v.a. indépendantes $X_1, X_2, X_3, ..., X_n$ de loi $\mathcal{N}(0,1)$. La v.a. $X = \sum_i^n X_i^2$ suit une loi du χ^2 à n degrés de liberté
- $f_n(x) = \frac{1}{2^{\frac{n}{2}}\Gamma(\frac{n}{2})} x^{\frac{n}{2}-1} e^{-\frac{x}{2}}$, où Γ est la fonction gamma qui est considérée comme le prolongement de la factorielle à l'ensemble des nombres complexes (sauf entiers <0 ou $=0$)
- $\mathbb{E}(X) = n$ (=nombre de v.a.) ; $\mathrm{Var}(X) = 2n$
- l'estimateur de la variance de la population à partir d'échantillons (qui est une v.a.) suit une loi du χ^2 : $(n-1)\frac{s_{n-1}^2}{\sigma^2} \sim \chi_{n-1}^2$
- cette loi permet d'effectuer des tests (écart des observations à une hypothèse) :
  - d'adéquation (une série de données statistiques et une loi de probabilité - voir plus loin)
  - d'homogénéité (comparaison d'échantillons issus de populations differentes)
  - d'indépendance (sur variables explicatives qualitatives)

Lois continues : loi du χ^2

loi du χ^2



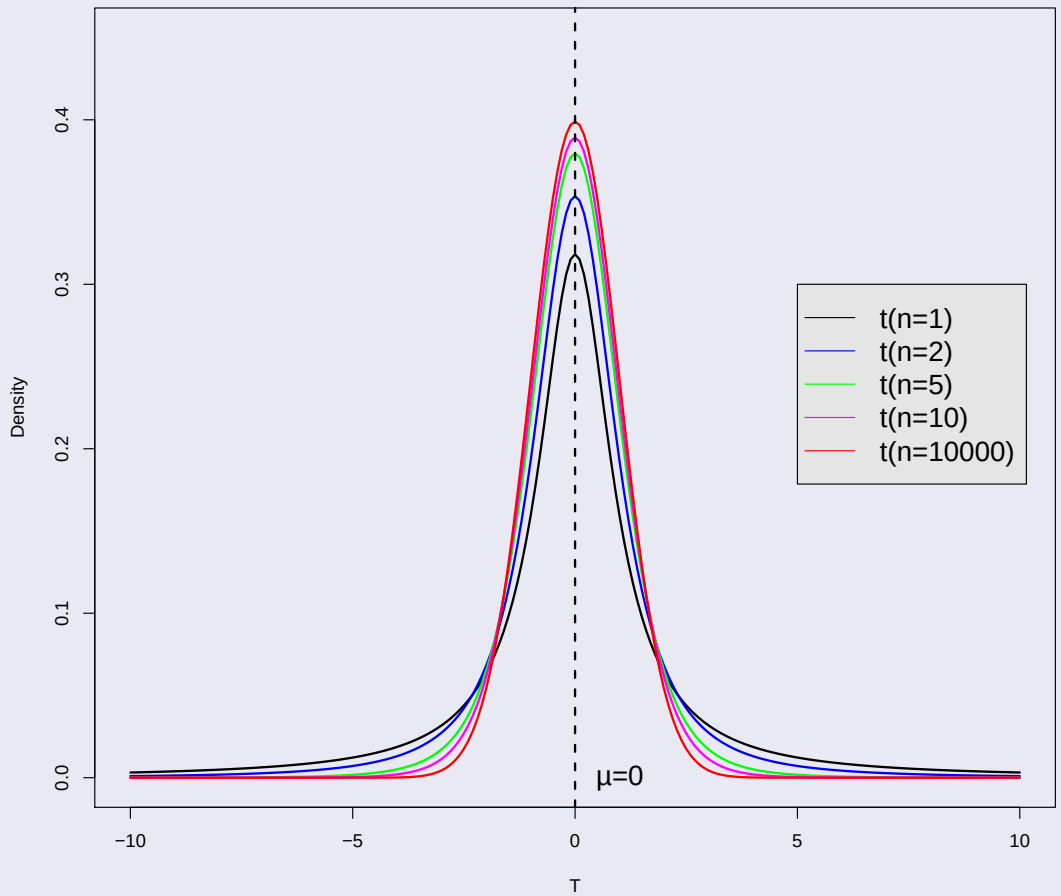
Lois continues : loi de Student

loi T de Student à n degrés de liberté

- si $X \sim \mathcal{N}(0,1)$ et $Y \sim \chi^2_{(n)}$ sont deux variables indépendantes
- alors $T = \frac{X}{\sqrt{\frac{Y}{n}}} \sim \mathcal{T}(n)$
- $f_n(x) = \frac{\Gamma(\frac{n+1}{2})}{\sqrt{n\pi}\Gamma(\frac{n}{2})(1+\frac{x^2}{n})^{\frac{n+1}{2}}}$, pour $n \geq 1$
- $\mathbb{E}(T) = 0$ pour $n > 1$; $\mathrm{Var}(T) = \frac{n}{n-2}$ pour $n > 2$
- si $n \to +\infty$, alors $\mathcal{T}(n) \to \mathcal{N}(0,1)$
- cette loi permet de tester la significativité de la différence entre deux moyennes d'échantillons indépendants (si $n \leq 30$), de construire un intervalle de confiance de cette différence, de tester la significativité de chaque paramètres d'une régression linéaire

Lois continues : loi de Student

loi T de Student à n degrés de liberté



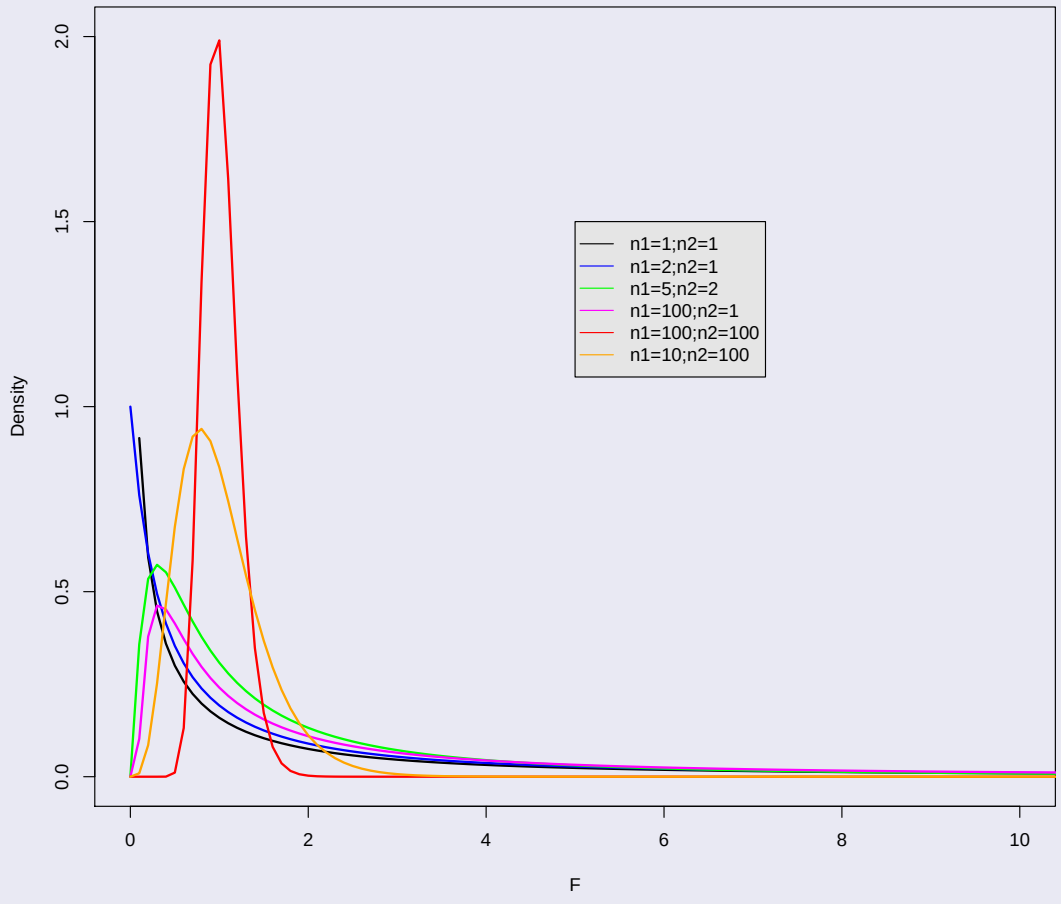
Lois continues : loi de Fisher-Snedecor

loi de Fisher-Snedecor à n_1 et n_2 degrés de liberté

- si $X_1 \sim \chi^2_{(n1)}$ et $X_2 \sim \chi^2_{(n2)}$ sont deux variables indépendantes
- alors $F = \frac{X_1/n_1}{X_2/n_2} \sim F_{(n_1,n_2)}$
- $f_{(n_1;n_2)}(x) = n_1^{\frac{n_1}{2}} n_2^{\frac{n_2}{2}} \frac{\Gamma(\frac{n_1+n_2}{2})}{\Gamma(\frac{n_1}{2})\Gamma(\frac{n_2}{2})} \frac{x^{\frac{n_1}{2}-1}}{(n_1 x+n_2)^{\frac{n_1+n_2}{2}}}$
- $\mathbb{E}(F) = \frac{n_2}{n_2-2}$ pour $n_2 \geq 3$; $\mathrm{Var}(F) = \frac{n_1+2}{n_1} \frac{n_2^2}{(n_2-2)(n_2-4)} - \frac{n_2^2}{(n_2-2)^2}$ pour $n_2 \geq 5$
- cette loi est très fréquente en tant que distribution de l'hypothèse nulle dans des tests statistiques comme les tests du ratio de vraisemblance ou encore dans l'analyse de la variance (F-test).

Lois continues : loi de Fisher-Snedecor

loi de Fisher-Snedecor à n_1 et n_2 degrés de liberté



Deux théorèmes pour continuer...

loi des grands nombres

- soit une v.a. X quelconque (non obligatoirement distribuée normalement). Réalisons un échantillon de taille n (n tirages) de X. Si $n \to +\infty$:
  - moyenne empirique m de l'échantillon $\to \mu$ (vraie moyenne théorique)
  - autrement dit m est un estimateur fortement convergent de l'espérance μ

théorème central limite (TCL)

- le TCL établit que toute somme de variables aléatoires indépendantes et identiquement distribuées tend -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 $\mathrm{Var}(S_n) = \sigma^2 n$ ($\mathbb{E}(\frac{S_n}{n}) = \mu$ et $\mathrm{Var}(\frac{S_n}{n}) = \frac{\sigma^2}{n}$)
  - si $n \to +\infty$ alors $S_n \to \mathcal{N}(n\mu, \sigma\sqrt{n})$ (convergence en loi)
  - et $Y_n = \frac{S_n - n\mu}{\sigma\sqrt{n}} \to \mathcal{N}(0,1)$
- donc la moyenne d'un échantillon $m = \frac{S_n}{n}$ tend vers une variable gaussienne lorsque $n \to +\infty$: $m \to \mathcal{N}(\mu, \frac{\sigma}{\sqrt{n}})$ (distribution de m sur tous les échantillons, avec $\frac{\sigma}{\sqrt{n}}$ = erreur standard de la mesure -standard error of the mean- SEM)

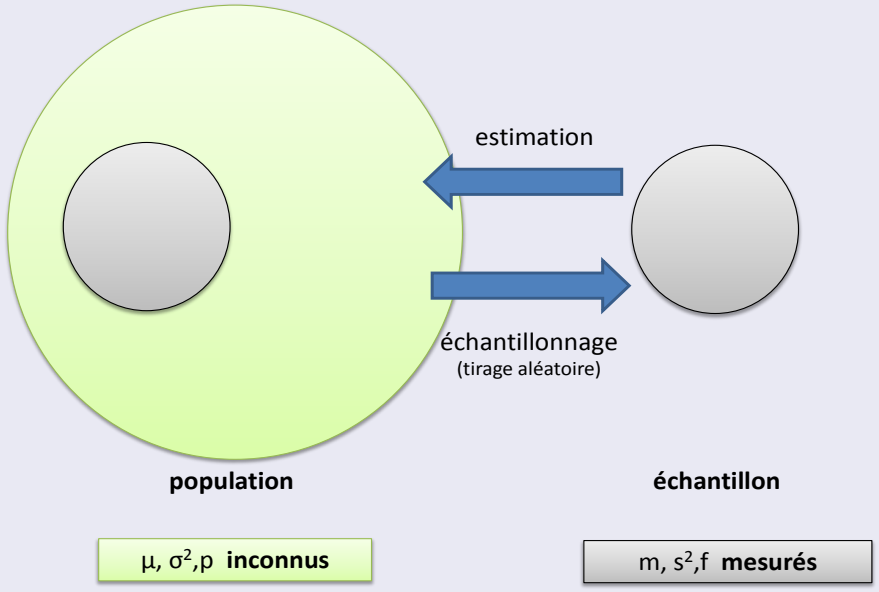
Deux théorèmes pour continuer...

conséquence pratique

- comme on sait :
  - estimer μ et $\frac{\sigma^2}{n}$
  - que toute variable aléatoire peut-être ramenée à une $\mathcal{N}(0,1)$ par la transformation $T = \frac{X-\mu}{\sigma}$
- il s'en suit qu'au delà d'un certain effectif, on sait ramener la distribution de la moyenne de toutes les variables aléatoires à UNE distribution unique : la $\mathcal{N}(0,1)$
- cette distribution étant connue, cette propriété fondamentale nous permet donc de disposer d'un moyen simple pour :
  - estimer par intervalle
  - faire des tests (ex : test de comparaison de moyennes -t de Student- voir plus loin)

Echantillonnage

théorie de l'échantillonnage



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

Echantillonnage

distribution d'échantillonnage (on suppose la population connue)

- distribution d'échantillonnage d'une proportion -ou fréquence- p (caractère qualitatif) :
  - $B(n,p)$: $\mathbb{E}(X) = \mu = np$; $\mathrm{Var}(X) = \sigma^2 = np(1-p) = npq$
  - $\mathbb{E}(\frac{X}{n}) = \mu = p$; $\mathrm{Var}(\frac{X}{n}) = \sigma^2 = \frac{pq}{n}$
  - d'où $\mu_p = p$ et $\sigma_p = \sqrt{\frac{pq}{n}}$
  - approximation normale pour n grand : $\mathcal{N}(p, \sqrt{\frac{pq}{n}})$
- distribution d'échantillonnage de la moyenne μ (caractère quantitatif) :
  - si la population est infinie ou que l'échantillonnage est avec remise (équivalent) : $\mu_m = \mu$ et $\sigma_m = \frac{\sigma}{\sqrt{n}}$
  - si la population est de taille finie N et $N > n$ et que l'échantillonnage est sans remise : $\mu_m = \mu$ et $\sigma_m = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$
  - approximation normale pour n grand : $\mathcal{N}(\mu, \frac{\sigma}{\sqrt{n}})$
- distribution d'échantillonnage de la variance σ^2 (caractère quantitatif) :
  - pour tout échantillon de taille n prélevé avec remise : $\mu_{s^2} = \frac{n-1}{n}\sigma$

Estimation

estimation ponctuelle

- moyenne : la moyenne m d'un échantillon prélevé au hasard dans une population est un bon estimateur de la moyenne inconnue de la population, μ
- fréquence (proportion) : la fréquence f des éléments possédant une certaine propriété dans un échantillon prélevé au hasard dans une population est un bon estimateur de la proportion inconnue p des éléments de cette population ayant cette propriété.
- variance (écart-type) : on montre que l'estimateur
  - $s_n^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - m)^2$ est biaisé car $\mathbb{E}(s_n^2) = \frac{n-1}{n}\sigma^2$
  - $s_{n-1}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - m)^2$ est sans biais car $\mathbb{E}(s_{n-1}^2) = \sigma^2$
  - d'où estimateur s_{n-1}^2 (ou $\frac{n}{n-1}s_n^2$)
- construction d'estimateurs :
  - méthodes des moments (égalité entre moments théoriques et empiriques) = les estimateurs ci-dessus.
  - maximum de vraisemblance

Estimation

estimation par maximum de vraisemblance (ML) : principe

- estimateur ML d'un paramètre = valeur estimée du paramètre telle que les probabilités des observations soient maximum. Dans la fonction de vraisemblance $L(x_1, x_2, x_3, ..., x_n, \theta)$, les données sont des constantes et les paramètres à estimer sont des variables
- le maximum de la fonction, si elle est dérivable, correspond à $L'_\theta = \frac{\partial L(x_1, x_2, x_3, ..., x_n, \theta)}{\partial \theta} = 0$
- on travaille souvent avec la log-vraisemblance

exemples

- estimation de λ de la loi de Poisson par ML
  - $\mathbb{P}(X = x_i) = e^{-\lambda} \frac{\lambda^{x_i}}{x_i!}$
  - $L(x_1, x_2, x_3, ..., x_n, \lambda) = e^{-n\lambda} \frac{\lambda^{\sum_{i=1}^{n} x_i}}{\prod x_i!}$
  - $lnL = -n\lambda + (\sum x_i) ln\lambda - ln \prod (x_i!)$
  - $(lnL)'_\lambda = -n + \frac{\sum x_i}{\lambda} = 0$
  - $\hat{\lambda} = \frac{\sum x_i}{n} = m$ (moyenne empirique des observations)
- estimateurs des paramètres d'une $\mathcal{N}(\mu, \sigma^2)$
  - estimateur ML de μ : $m = \frac{1}{n} \sum_{i=1}^{n} x_i$
  - estimateur ML de σ^2 : $s^2_{n-1} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - m)^2$

Estimation

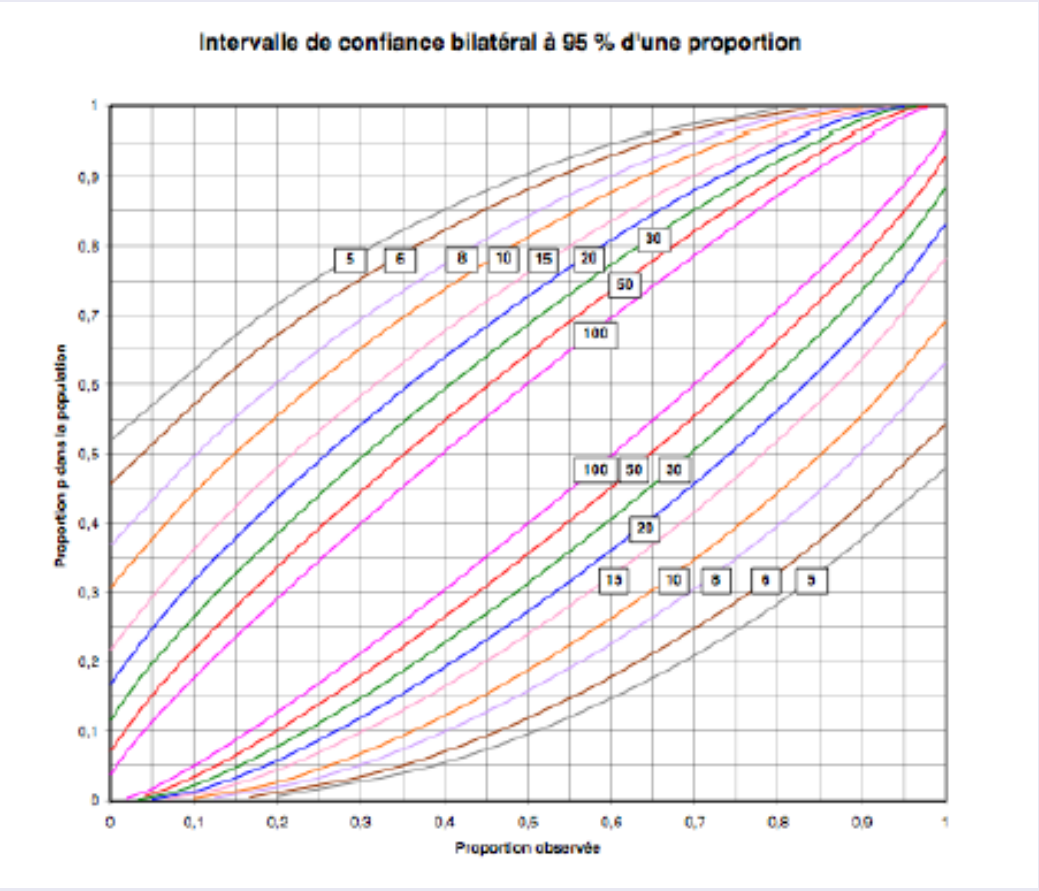
estimation par intervalle de confiance

- intervalle de confiance (IC) : intervalle ayant une probabilité donnée de contenir la vrai valeur du paramètre (ex : $p, \mu, \sigma...$)
- α = probabilité que l'IC ne contienne pas la vraie valeur
- on cherche un intervalle $[a, b]$ centré sur la valeur estimée du paramètre inconnu θ, contenant la vraie valeur de ce paramètre avec une probabilité $1 - \alpha$ fixée *apriori*, soit $\mathbb{P}[a < \theta < b] = 1 - \alpha$
- 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 $\mathcal{N}(0, 1)$
  - $IC_{(p)}(95\%) = [p - 1.96\sqrt{\frac{p(1-p)}{n}}, p + 1.96\sqrt{\frac{p(1-p)}{n}}]$, avec $u_{(0.975)} = 1.96$ $(\alpha = 0.05)$
  - exemple : on sème 50 graines dont 35 germent : $p = \frac{35}{50} = 0.7$, $IC_{(p)}(95\%) = [0.57, 0.83]$ **(0.7 ± (1.96 * √(0.7 * 0.3/50)))**

Estimation

estimation par intervalle de confiance

- IC d'une proportion :
  - abaques pour IC(95%) d'une proportion



Estimation

estimation par intervalle de confiance

- IC d'une moyenne (cas d'une $\mathcal{N}(\mu, \sigma)$ où σ^2 est **connu)** :
  - $IC_{(\mu)} = [m - u\frac{\sigma}{\sqrt{n}}, m + u\frac{\sigma}{\sqrt{n}}]$
  - 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 $u_{(0.995)} = 2.6$ $(\alpha = 0.01)$
- IC d'une moyenne (cas d'une $\mathcal{N}(\mu, \sigma)$ où σ^2 est **inconnu)** :
  - $IC_{(\mu)} = [m - t\frac{s_n}{\sqrt{n}}, m + t\frac{s_n}{\sqrt{n}}]$
  - où t est le quantile d'ordre $1 - \frac{\alpha}{2}$ de la loi $\mathcal{T}(n - 1)$
  - quand $n \to +\infty$, en pratique $n > 100$, on approxime $\mathcal{T}(n - 1)$ par $\mathcal{N}(0, 1)$
- 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.

Estimation

estimation par intervalle de confiance

- IC d'une moyenne : application

Intervalle de confiance d'une moyenne (population normale)

alpha= 0.05 moyenne 157.4 Ecart-type 8.5

	Ecart-type connu		Ecart-type estimé	
n	Inf	Sup	Inf	Sup
5	149.9596	164.8604	146.8559	167.9641
10	152.1417	162.6783	151.3295	163.4905
15	153.1085	161.7115	152.7029	162.1171
20	153.6848	161.1352	153.4319	161.3881
25	154.0781	160.7419	153.9014	160.9186
30	154.3684	160.4516	154.2360	160.5840
40	154.7759	160.0441	154.6916	160.1284
50	155.0540	159.7660	154.9943	159.8257
100	155.7440	159.0760	155.7234	159.0966

Estimation

estimation par intervalle de confiance

- IC d'une variance (cas d'une $\mathcal{N}(\mu, \sigma)$ où μ est **connu**) :
  - on a $n\frac{s_n^2}{\sigma^2} \sim \chi^2_n$
  - $IC_{(\sigma^2)} = [n\frac{s_n^2}{\chi^2_{(1-\frac{\alpha_2}{2})}}, n\frac{s_n^2}{\chi^2_{(\frac{\alpha_1}{2})}}]$
  - où $\chi^2_{\frac{\alpha_1}{2}}$ est le quantile d'ordre α_1 de la loi $\chi^2_{(n)}$
  - où $\chi^2_{1-\frac{\alpha_2}{2}}$ est le quantile d'ordre $1-\alpha_2$ de la loi $\chi^2_{(n)}$
  - intervalle non centré car la loi du χ^2 n'est pas symétrique
- IC d'une variance (cas d'une $\mathcal{N}(\mu, \sigma)$ où μ est **inconnu**) :
  - on a $(n-1)\frac{s_{n-1}^2}{\sigma^2} \sim \chi^2_{n-1}$
  - $IC_{(\sigma^2)} = [n-1\frac{s_{n-1}^2}{\chi^2_{(1-\frac{\alpha_2}{2})}}, n-1\frac{s_{n-1}^2}{\chi^2_{(\frac{\alpha_1}{2})}}]$

Test : principe et classification

- **test d'hypothèse** : démarche consistant à rejeter ou à ne pas rejeter (rarement accepter) une hypothèse statistique, appelée hypothèse nulle (H0), en fonction d'un jeu de données (échantillon). Il s'agit de statistique inférentielle : à partir de calculs réalisés sur des données observées, nous émettons des conclusions sur la population, en leur rattachant des risques de se tromper.
- les tests selon leur finalité :
  - **test de conformité** : confronter un paramètre calculé sur l'échantillon à une valeur pré-établie (ex : tests portant sur la moyenne ou sur les proportions).
  - **test d'adéquation** : vérifier la comptabilité des données avec une distribution choisie a priori (ex : test d'adéquation à la loi normale)
  - **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'association (ou d'indépendance)** : rechercher une liaison entre 2 variables. Les techniques utilisées diffèrent selon que les variables sont qualitatives nominales, ordinales ou quantitatives.

Test : principe et classification

- tests paramétriques / non paramétriques :
  - **tests paramétriques** : on stipule que les données sont issues d'une distribution paramétrée. Dans ce cas, les caractéristiques des données peuvent être résumées à l'aide de paramètres estimés sur l'échantillon, la procédure de test subséquente ne porte alors que sur ces paramètres. L'hypothèse de normalité sous jacente des données est le plus souvent utilisée, la moyenne et la variance suffisent pour caractériser la distribution. Concernant les tests d'homogénéité par exemple, pour éprouver l'égalité des distributions, il suffira de comparer les moyennes et/ou les variances.
  - **tests non paramétriques** : on ne fait aucune hypothèse sur la distribution sous-jacente des données. Pas besoin d'estimer les paramètres des distributions avant de procéder au test. Pour les données quantitatives, les tests non paramétriques transforment les valeurs en rangs (appellation "tests de rangs"). Lorsque les données sont qualitatives, seuls les tests non paramétriques sont utilisables.

Quelques exemples de questions biologiques

- effet d'un traitement sur le taux de germination de graines
- effet d'un traitement sur la croissance de plantules d'*Arabidopsis*
- comparer les taux de lignine dans différentes variétés d'Eucalyptus
- comparer les effets de plusieurs doses d'engrais sur le rendement de plusieurs variétés
- existence dune corrélation entre génotype et phénotype, entre deux variables biochimiques

Quelques types de méthodes statistiques

- **ajustement de distributions (adéquation)**
  - test de normalité (χ^2, Shapiro-Wilk, Kolmogorov-Smirnov,...)
  - comparaison de distributions (χ^2, Kolmogorov)
- **test de conformité**
  - comparaison de la valeur d'un paramètre à une valeur connue (χ^2, t de Student,...)
  - intervalle de confiance
- **comparaison de proportions**
- **comparaison de populations**
  - tests non parametriques (Mann-Whitney - Wilcoxon, Kruskal-Wallis...)
  - tests paramétriques (t de Student, Levene,...)
  - analyse de variance (ANOVA) dans le cas de plusieurs populations (test de Fisher)
- **méthodes relatives a la régression (2 variables ou plus)**

Facteurs de choix d'une méthode statistique

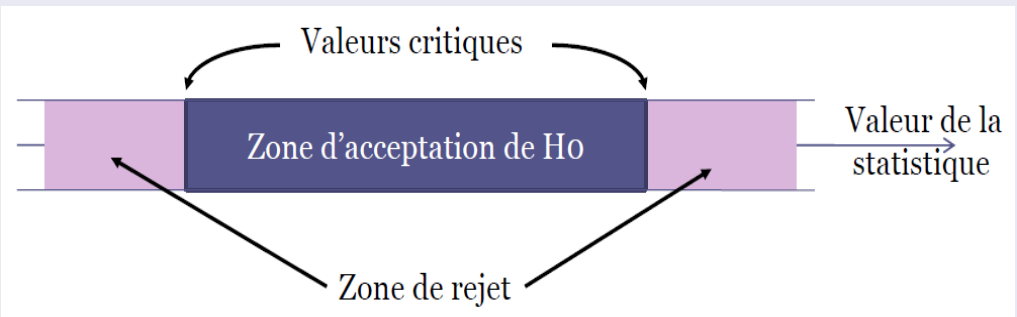
- **structure des données**
  - nombre de variables
  - plan d'expérience : nombre et nature des facteurs
  - échantillons indépendants ou correlés
- **nature des données**
  - qualitatives, quantitatives
- **objectifs poursuivis**
  - estimation (limites de confiance), tests d'hypothèse
  - comparaisons avec un témoin, interactions entre facteurs
- **propriétés des méthodes statistiques**
  - normalité, robustesse, puissance

Les étapes d'un test statistique

- **choix d'un test statistique** (paramétrique, non paramétrique), souvent guidé par les contraintes de taille de l'échantillons
- **réalisation de l'étude** : tableau de données (compatibles avec les conditions d'application du test ?)
- **réalisation du test**
  - **hypothèses** : le test consiste a trancher entre 2 hypotheses : H0, l'hypothèse nulle, et H1, l'hypothèse alternative
  - **statistique de test** : la fonction statistique liée au test qui servira a choisir quelle hypothèse retenir
  - **loi de la statistique de test sous** $H0$: Il s'agit de la loi de la statistique choisie précedemment, soumise aux conditions de H0
  - **région critique** : c'est l'intervalle de valeur de notre statistique de test ou l'on rejette H0 avec un probabilite α (variable suivant test unilateral ou bilateral)
  - **résolution** : on calcule la statistique sur la base des données
  - **conclusion** : on peut finalement conclure, si les données observées sont en région critique qu'on rejette H0 , sinon qu'on rejette H1

Hypothèses $H0$ et $H1$

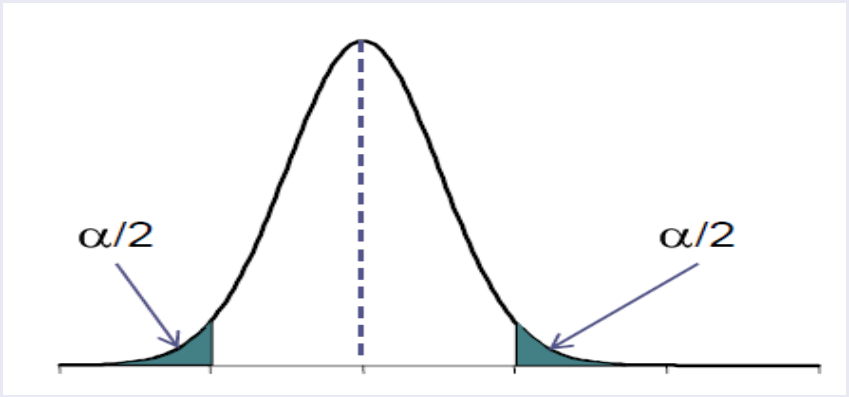
- **hypothèse nulle ($H0$)** : hypothèse que l'on veut tester
  - distribution (normalité), indépendance
  - égalité de moyennes, de variances
  - comparaison de distributions
- **hypothèse alternative ($H1$)** : test bilatéral ou unilatéral (ex : moyenne de A est différente ou supérieure à moyenne de B)
- **règle de décision**
  - si, lorsque $H0$ est vraie, la valeur calculée de la statistique de test a une «trop» faible probabilité d'être observée alors on rejette l'hypothèse $H0$



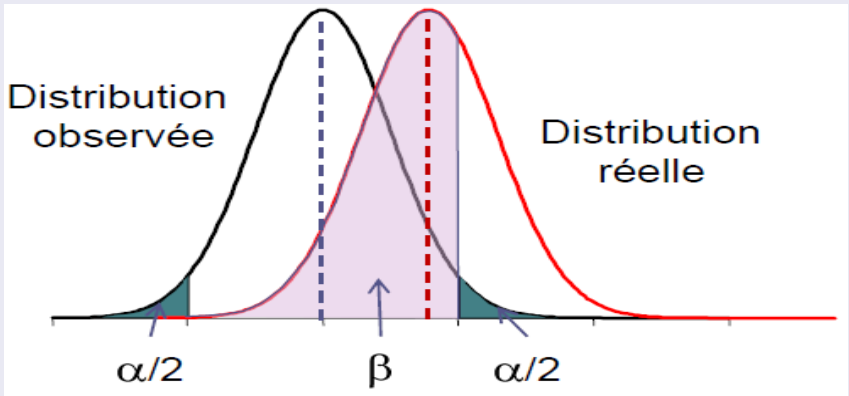
  - la zone de rejet depend de $H1$:
    * si le test est bilatéral la zone de rejet est répartie de chaque coté de la zone d'acceptation
    * si le test est unilatéral la zone de rejet se situe d'un seul coté de la zone

Risque d'erreur

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

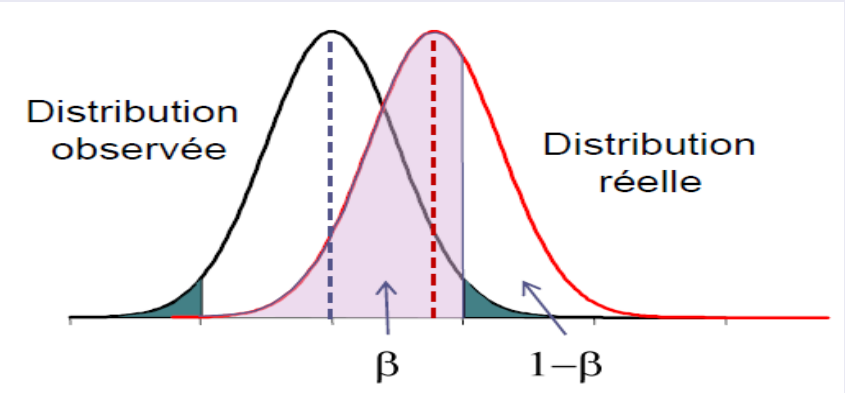


- **erreur de 2^{me} espèce (type II)** : accepter une hypothèse alors qu'elle est fausse (β = risque d'erreur de type II)



Prise de décision : risques associé au test

- **puissance du test** $= 1 - \beta$: capacité du test à rejeter une hypothèse fausse



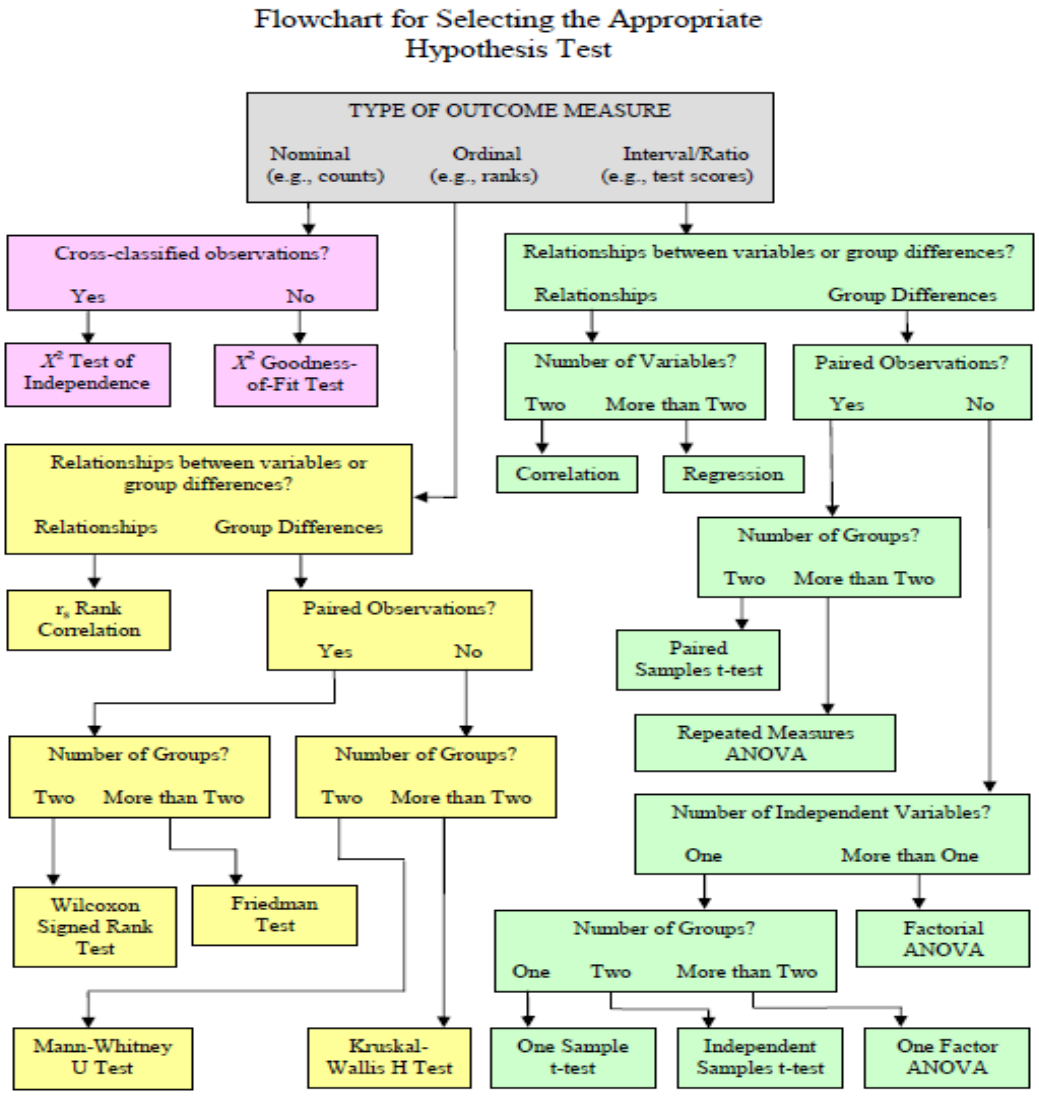
- la puissance dépend de la taille de l'échantillon, de σ^2, et de la différence que l'on veut détecter
- dans certains cas on peut calculer $1 - \beta$ *a priori* (ou bien la simuler)
- risques associé au test

réalité décision	$H0$	$H1$
$H0$	$1 - \alpha$ (décision correcte)	β
$H1$	α	$1 - \beta$ (décision correcte)

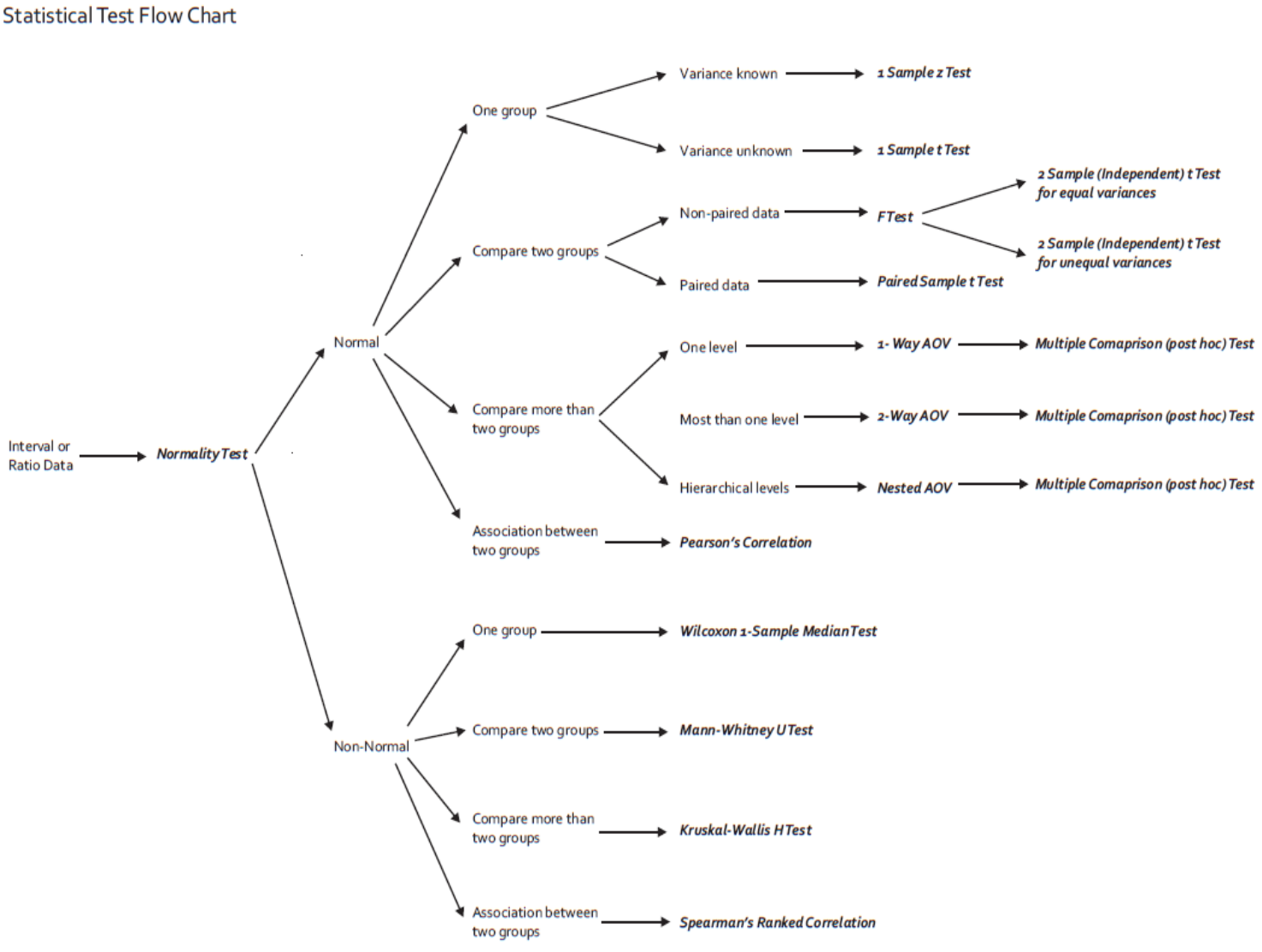
Test statistique : EN PRATIQUE

- **définir** $H0$ **et** $H1$
- **fixer** *a priori* **une valeur maximale pour** α (ex : 0.05, 0.01,...)
- **acquérir les données**
- avec le test approprié : **calculer la probabilité** p **de la statistique observée** $(P - value)$ **sous l'hypothèse** $H0$
  - si $p < \alpha$: **on rejette** $H0$ **et on accepte** $H1$ **au risque d'erreur** $< \alpha$
  - si $p > \alpha$: **on accepte** $H0$

Quel test utiliser ?



Quel test utiliser ?



Quel test utiliser ?