Probabilités et Statistiques: Quelques petits rappels

Christine Decaestecker & Marco Saerens

ULB & UCL

Probabilités et Statistiques: Introduction

- **Démarche statistique:**
  Etude d'un **ensemble d'objets** (cas, individus, ...) sur lesquels on observe des **caractéristiques** appelées «**variables**.

  population = ensemble, collection d'objets équivalents sensés partager des propriétés communes.

  Statistique = étude des **propriétés générales des populations** plutôt que des propriétés particulières des individus.

- **Ex: contrôle de qualité de pièces usinées:**
  Pour un certain type (une population) de pièces (fabriquées dans des conditions déterminées): proportion de pièces défectueuses?

- **Echantillonage:**
  Etude exhaustive d'une population trop vaste (peut être infinie), difficile et/ou coûteuse;

  ⇒ étude d'une partie (représentative) ou **échantillon**

- **Types de variable:** (caractéristique, descripteur, attribut)
  - **qualitative**: valeurs = labels d'appartenance à une catégorie
    - **nominale** (sans notion d'ordre, ex: sexe, type de traitement, ...);
    - **ordinale** (ordre sur les labels, ex : faible, moyen, fort);
  - **quantitative**: à valeurs numériques
    - **entière** (nombre fini ou dénombrable de valeurs, ex: nombre d'accidents de la route par an);
    - **continue** (toutes les valeurs d'un intervalle réel sont observables, ex: température, pression sanguine, ...).

Variables quantitatives: les **opérations arithmétiques** (somme, moyenne, ...) ont un sens;

Variables qualitatives: seules les **opérations de comptage** ont un sens!

- **Analyse statistique**
  = étude de la **variabilité** des caractéristiques des individus.
  Des individus apparemment semblables ont généralement des valeurs différentes pour les différentes variables qui les caractérisent.

- **Utilités de l'analyse statistique:**
  - **prévoir** (de façon probabiliste) le comportement d'individus non encore observés;
  - réduire ou **expliquer** la variabilité à l'aide d'autres variables;
  - utiliser la variabilité pour **distinguer** au mieux les individus (ou groupes d'individus) entre eux.

- **Tableaux de données**:
  échantillon de n individus sur lesquels on observe p variables:

Individus	Variables						
		X_1	X_2		X_j		X_p
	1						
	2						
	i				x_{ij}		
	n						

notations vectorielles:

$i^{\text{ème}}$ observation = vecteur $\mathbf{x}_i = [x_{i1}, x_{i2}, ..., x_{ip}]^T$

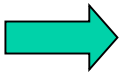
ou $\mathbf{x}(i) = [x_1(i), x_2(i),..., x_p(i)]^T$: $i^{\text{ème}}$ réalisation du vecteur aléatoire $\mathbf{x}$

Le tableau de données = une matrice (x_{ij}) avec $i = 1, ..., n$ et $j = 1, ..., p$

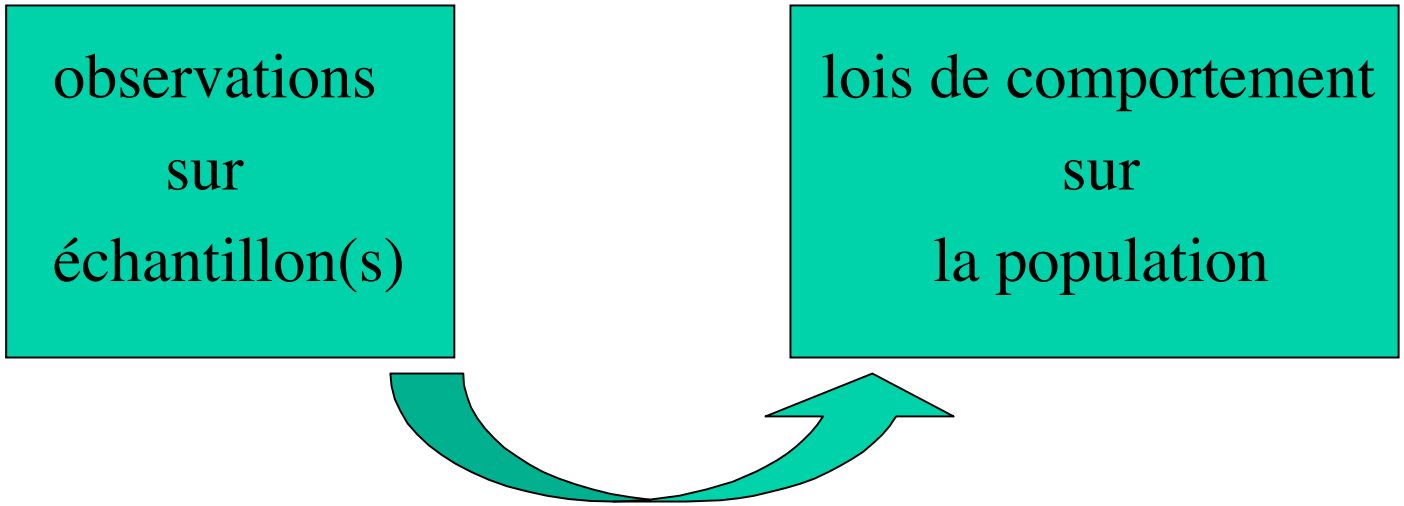
(utilisation d'outils du calcul matriciel)

- Théorie des probabilités:
  - théorie mathématique axiomatique (déconnectée de la réalité physique);
  - permet de **modéliser** des phénomènes où **le hasard** (**l'aléatoire**) intervient et d'en faire l'étude théorique (analytique).

    = outil pour la statistique:

Statistique		**Probabilités**
Observation de phénomènes concrets		***Modélisation de phénomènes aléatoires***
• Observation de données: possibles imprécisions, erreurs		• Modélisation des déviations entre vraies valeurs et valeurs observées
• Distribution des valeurs d'une variable sur une population		• Modélisation par des lois de probabilité théoriques (gaussienne, poisson, …)
• Echantillonage des individus observés par tirage au hasard (assurer la représentativité)		• Propriétés observées = variables aléatoires ➔ étude des répartitions, des liens éventuels

Probabilités = outils essentiels pour extrapoler:



STATISTIQUE INFERENTIELLE
(essentiellement uni- et bi-variée)

1ère partie du cours
(revisions)

autre aspect de la statistique:

STATISTIQUE EXPLORATOIRE
(essentiellement multivariée)

2ème partie du cours

- **Statistique exploratoire:**
  (statistique descriptive, analyse de données, "data mining")

But:

synthétiser, structurer (extraire) l'information contenue dans les données

Démarche

représentations des données:
tableaux, graphiques (visualisations), indicateurs numériques

↓

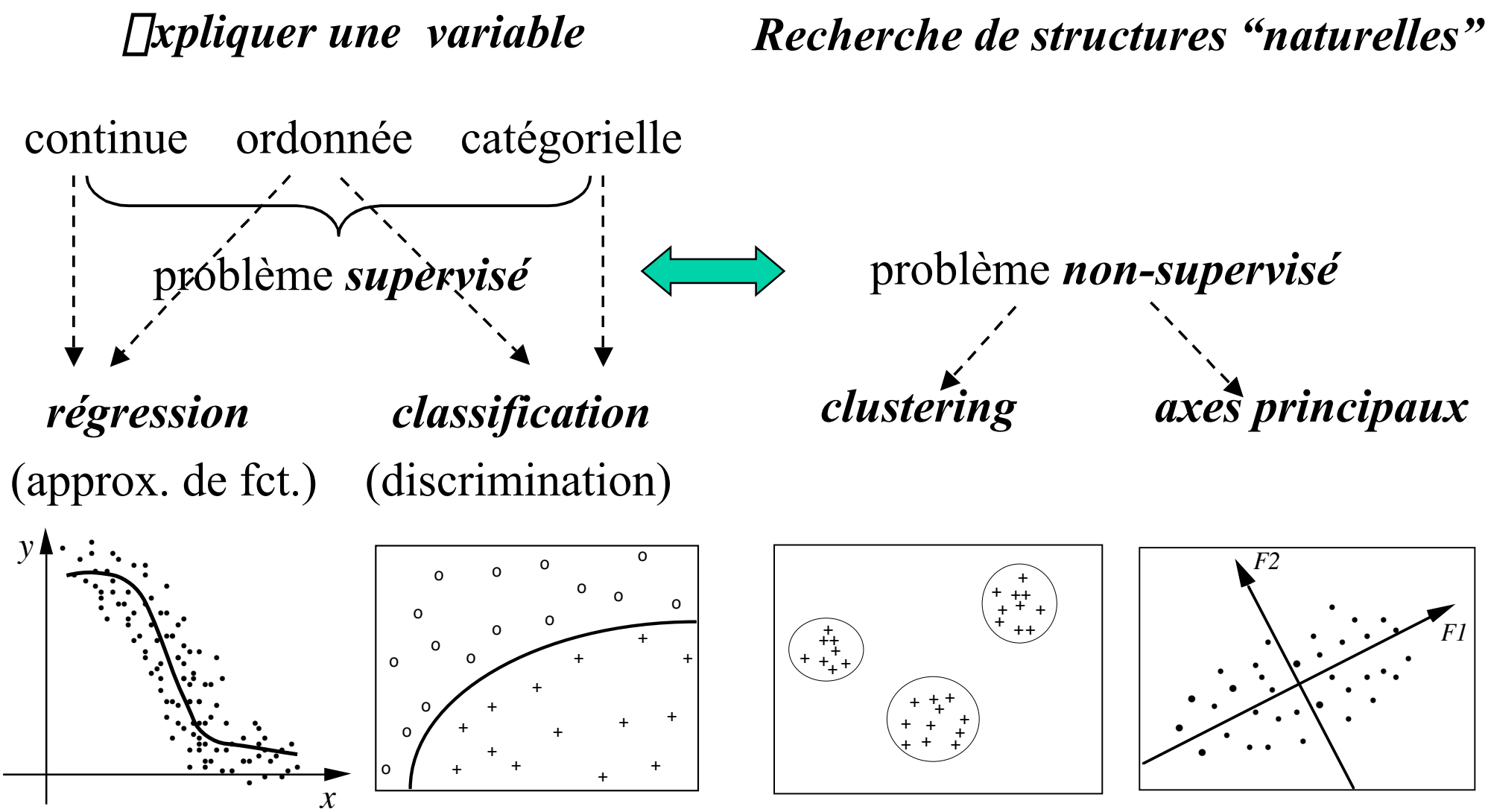
mise en évidence de propriétés de l'échantillon

↓

suggérer des hypothèses sur la population, le phénomène,
le problème étudié

outils de base: algèbre linéaire, calcul matriciel, géométrie euclidienne, informatique.

- **Exemples de tâches exploratoires:**



- **Démarche scientifique:**
  1) phase exploratoire (observations),
  2) formulation d'hypothèses, de modèles, de règles,
  3) validation.

- **Outils de validation**:
  - **Inférence**: s'appuie sur des modèles probabilistes,
    => démarche analytique,
    => modèles mathématiques relativement "simples",
    => vue simplifiée parfois contestable du phénomène étudié: les données doivent "coller" suffisamment au modèle pour que la démarche soit valable.
  - **Procédure informatique**: ne s'appuie que sur les données,
    => rééchantillonnage et vérification (à recommencer x fois),
    => outils informatiques: tests intensifs pour avoir une bonne ***perception générale*** du phénomène
    => nécessite une ***quantité de données*** *importante*

- **Exemple**: estimation d'une moyenne, intervalle de confiance

  échantillon de n valeurs d'une variable X: $x_1, x_2, ..., x_n$

  moyenne échantillon: $\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$ dépend de l'échantillon analysé

  intervalle de confiance = intervalle autour de $\bar{x}$: $[\bar{x} - \delta\bar{x}, \bar{x} + \delta\bar{x}]$
  ayant une probabilité fixée (95%) de contenir la moyenne population m

  Méthode d'inférence:
  suppose une **distribution théorique** (ex: normale) pour déterminer l'intervalle.

  Méthode de rééchantillonnage: (principes de base)
  - génère un grand nombre N d'échantillons similaires à l'échantillon de départ,
  - calcule les N moyennes échantillon => **distribution empirique,**
  - détermine l'intervalle contenant 95% des valeurs.

Outils probabilistes de base (rappels)

- **Expérience aléatoire**: expérience où le hasard intervient rendant le résultat imprévisible (Ex: lancer un dé)

  ensemble de tous les résultats possibles = **univers des possibles** = Ω (Ex: Ω = {1, 2, ..., 6})

  Evénement: assertion relative au résultat d'une expérience, se réalise ou non (Ex: obtenir un nbre pair)

  = toute partie de Ω (Ex: {2, 4 , 6})

  ↪ outils ensemblistes

- Soit $\mathscr{C}$ un ensemble d'événements = ensemble de parties de Ω satisfaisant les propriétés suivantes (algèbre de Boole):

  $\forall A \in \mathscr{C}$: son contraire $\overline{A} \in \mathscr{C}$ ($\overline{A} = \Omega \setminus A$)

  $\forall A_1, A_2, \ldots, A_n \in \mathscr{C}$: $\bigcup A_i \in \mathscr{C}$

  $\Omega \in \mathscr{C}$

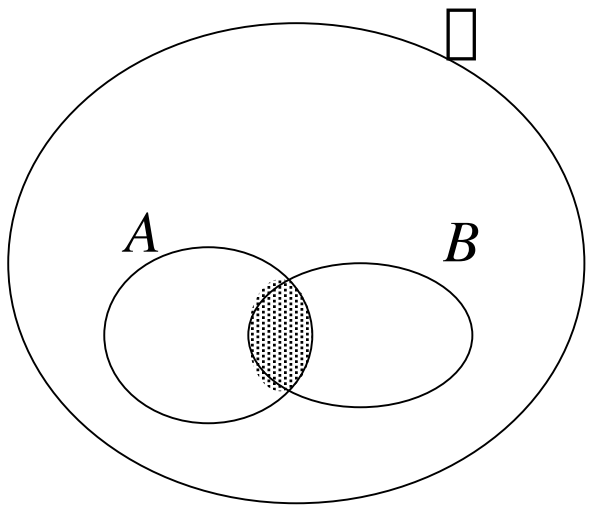
  ⇒ (Ω , $\mathscr{C}$) est un **espace probabilisable**

- **Loi de probabilité *P***: $(\Omega, \mathscr{C}) \rightarrow [0,1]$ telle que

  $P(\Omega) = 1$ et $\forall A_1, A_2, \ldots, A_n \in \mathscr{C}$ tels que $A_i \cap A_j = \varnothing$ $(\forall i \neq j)$,

  on a $P(\bigcup A_i) = \sum P(A_i)$

  $\Rightarrow$ $(\Omega, \mathscr{C}, P)$ est un **espace probabilisé**



- **Propriétés élémentaires**

  $P(\varnothing) = 0$

  $P(\neg A) = 1 - P(A)$

  $P(A) \leq P(B)$ si $A \subseteq B$

  $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

- **Probabilités conditionelles - Indépendance**

  Proba. conditionelle de *A* sachant *B* (de proba. non nulle):

  $$P(A \mid B) = P(A \cap B) / P(B)$$

  A* indépendant de *B ssi $P(A \mid B) = P(A)$

  ssi $P(A \cap B) = P(A)\,P(B)$

- **Théorème des probabilités totales**:

  $\forall$ $A_1, A_2, ..., A_n$ formant une partition de Ω $(A_i \cap A_j = \varnothing \ \forall\ i \neq j$, et $\bigcup A_i = \Omega)$

  $$\forall B \qquad P(B) = \sum_i P(B \cap A_i)$$
  $$= \sum_i P(B \mid A_i)\, P(A_i)$$

  en particulier: $P(B) = P(B \mid A)\, P(A) + P(B \mid \neg A)\, P(\neg A)$

- **Théorème de Bayes:**

  $$P(A \mid B) = P(B \mid A)\, P(A) \,/\, P(B)$$

  Généralisation: si $A_1, A_2, ..., A_n$ forment une partition de Ω

  $$P(A_i \mid B) = P(B \mid A_i)\, P(A_i) \,/\, \sum_k P(B \mid A_k)\, P(A_k)$$

  Ex d'utilisation: aide au diagnostique:
  probabilité d'une affection étant donnés les résultats d'examens médicaux.

- **Assignement d'une probabilité à un événement**
  - ***vision classique*** (jeux de hasard):
    Ω est un **ensemble fini** de cas possibles dont chaque singleton (***événement élémentaire***) a la même probabilité de se réaliser.
    (Ex: lancer un dé parfait => Ω est constitué de 6 éléments équiprobables)

    d'où:
    $$P(A) = \frac{\textbf{Nbre de cas favorables}}{\textbf{Nbre de cas possibles}}$$

    utilisation de l'***analyse combinatoire***

    
    ne convient pas si Ω n'est pas fini
    vision utopiste de la réalité (équiprobabilité)
  - ***vision "fréquentiste"*** (loi des grands nombres):
    répéter un grand nbre de fois N l'expérience et observer le nbre de fois que l'événement d'intérêt A se produit:

    d'où:
    $$f(A) = \frac{\textbf{Nbre d'occurrences de } A}{\textbf{Nbre d'expériences } (N)} \quad \textbf{et} \quad P(A) = \lim_{N\to\infty} f(A)$$

    
    ne convient pas si l'événement n'est pas reproductible!

– ***conception subjectiviste:***

La probabilité objective d'un événement n'existe pas et n'est pas une grandeur mesurable!

Probabilité = mesure d'incertitude variant avec les circonstances et l'observateur

=> **mesure subjective**

Contrainte: satisfaire aux axiomes du calcul des probabilités

=> permet de **probabiliser des événements non reproductibles et non aléatoires** !

(Ex: Quelle est la probabilité qu'il pleuve demain?)

=> permet de **modéliser l'incertain**

Modélisation de l'incertain et de l'imprécis:
nouveaux développements théoriques (depuis les années 1950):

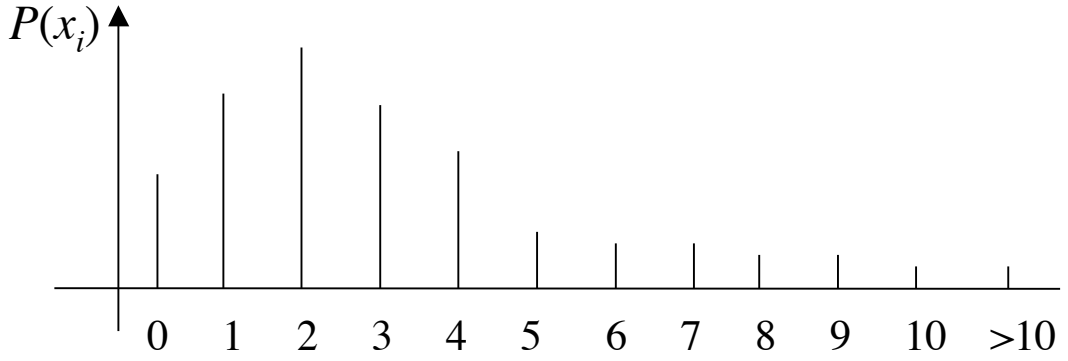
théories des possibilités, des fonctions de croyance et des ensembles flous

généralisation des axiomes des probabilités => plus grande souplesse
plus proche du raisonnement humain

Variables aléatoires et distributions

- **Variable aléatoire**: entité prenant différentes valeurs ('variable'), chacune avec une certaine probabilité ('aléatoire')
  - nbre fini ou dénombrables de valeurs: variable **discrète**
  - toute valeur dans un certain intervalle de ¬ : variable **continue**
- **Loi de probabilité** d'une variable aléatoire X:
  assignation des probabilités sur les différentes valeurs de X (discrète)
  ou sur des intervalles de valeurs de X (continue)
  Pour une variable discrète: ***masses ponctuelles*** $P(X = x_i)$
  Pour une variable continue: ***densité de probabilité*** $P(a < X < b)$

Ex: X = nbre d'enfants
des familles européennes

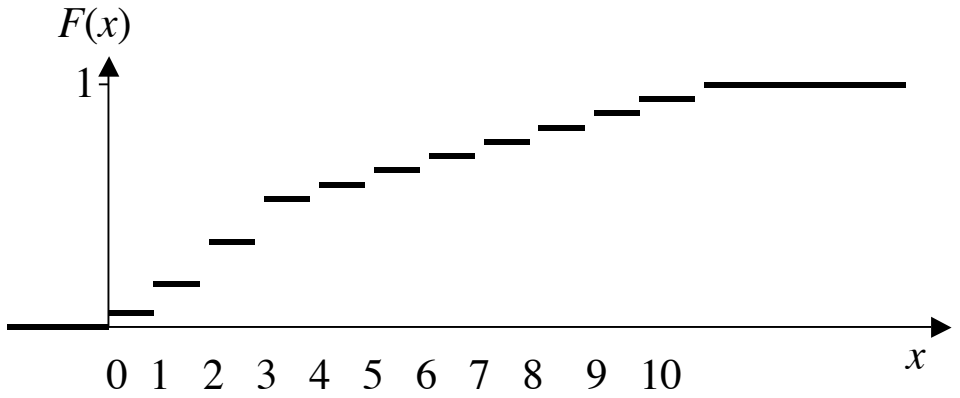


- **Fonction de répartition** d'une variable aléatoire X:

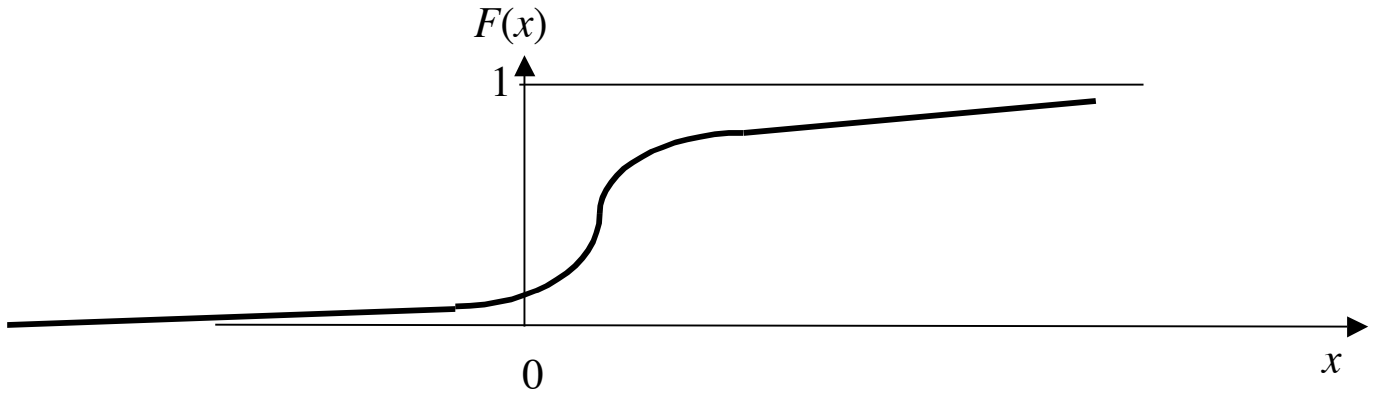
  $F(x) = P(X < x)$ (fct monotone croissante)

  d'où: $P(a \leq X < b) = F(b) - F(a)$

  Ex: fct de répartition d'une variable discrète :



  Ex: fct de répartition d'une variable continue :



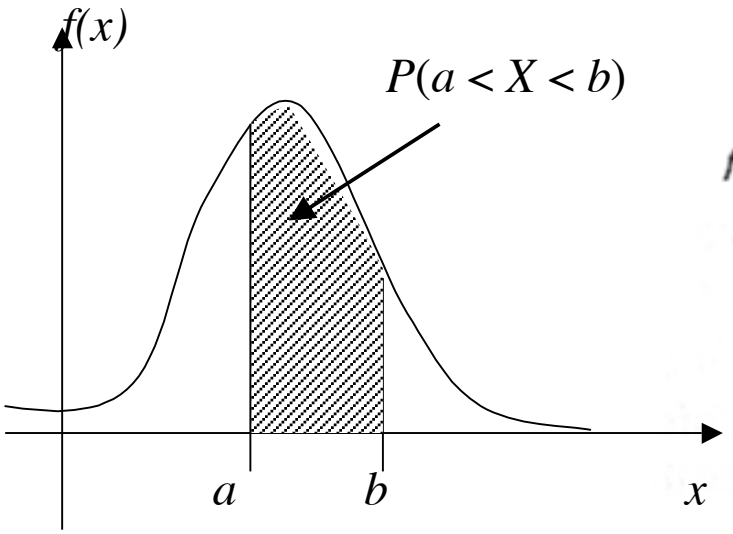
- **Densité de probabilité** (variable **continue**):
  $f(x)$ est la fct de densité pour une variable X si pour tout intervalle $[a\ ,\ b]$ de ¬ on a:

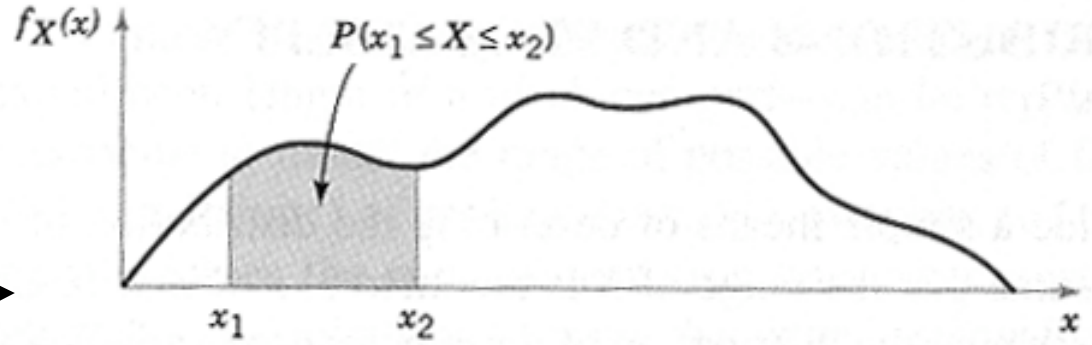
$$P(a < X < b) = \int_a^b f(x)\mathrm{d}x$$

= ***aire sous la courbe*** $f(x)$ au dessus de $[a, b]$

d'où $\int_{-\infty}^{+\infty} f(x)\mathrm{d}x = 1$

et $$F(a) = P(X < a) = \int_{-\infty}^{a} f(x)\mathrm{d}x$$





- **Moments d'une variable aléatoire *X*:**
  valeurs typiques:
  - centrales: **moyenne**
  - de dispersion: **variance, écart-type (déviation standard)**
  - de forme de distribution: coefficient **d'asymétrie** ('*skewness*'), **d'aplatissement** ('*kurtosis*').

  notion d'***espérance mathématique***: $E(X)$ = moyenne (= centre de masse):

  - var. discrète: $\mu = E(X) = \sum_i x_i P(X = x_i)$
  - var. continue de densité $f(x)$: $\mu = E(X) = \int x f(x)\, dx$ (n'existe pas tjrs!)
  - propriétés élémentaires: $E(a) = a$

    $E(aX) = a\, E(X)$

    $E(X + Y) = E(X) + E(Y)$

variance: $V(X) = \sigma^2 = E\big((X - E(X))^2\big)$ (moment centré d'ordre 2)
$= E(X)^2 - \mu^2$

écart-type: $\sigma = \sqrt{V(X)}$

Définition générale: moment centré d'ordre $k = \mu_k = E\big((X - E(X))^k\big)$

Coefficients d'asymétrie α_1 (*skewness*) et d'aplatissement α_2 (*kurtosis*):

$$\alpha_1 = \frac{\mu_3}{\sigma^3} \qquad \alpha_2 = \frac{\mu_4}{\sigma^4}$$

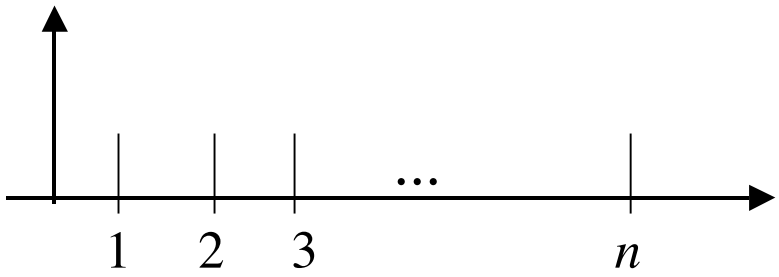
- **Médiane et percentile:**
  médiane: valeur x_{50} telle que $P(X \leq x_{50}) = 0.50$
  percentile $p\%$: valeur x_p telle que $P(X \leq x_p) = p\%$

Lois de probabilité d'usage courant

Usage: Modélisation de systèmes physiques et de leurs résultats,
permet des prédictions à propos de ces systèmes.

Lois discrètes:

- **Loi discrète uniforme:**
  $X = \{1, 2, ..., n\}$ avec $P(X = k) = 1/n$



- **Loi ou schéma de Bernouilli de paramètre *p*:**
  $X = \{0,1\}$ avec $P(X = 1) = p$ (apparition d'un événement) et
  $P(X = 0) = 1 - p$ (non-apparition d'un événement)

- **Loi binomiale $\mathcal{B}(n, p)$:**

  $X = \sum_{i=1}^{n} X_i$ somme de n variables de Bernouilli X_i indépendantes de même paramètre p

- **Loi binomiale (suite)**
  X = nbre d'apparitions d'un événement parmi n expériences aléatoires indépendantes dont le résultat est l'apparition ou non d'un événement de probabilité p. D'où: $P(X = k) = C_n^k p^k (1 - p)^{n-k} = \frac{n!}{k!\ (n-k)!} p^k (1 - p)^{n-k}$

- **Loi de Poisson $\mathcal{P}(\lambda)$:**
  X à valeur naturelle (entière, positive ou nulle) et satisfait:
  $$P(X = k) = \exp(-\lambda)\frac{\lambda^k}{k!}$$
  = probabilité d'obtenir x événements indépendants pendant un temps T, si le nbre moyen d'événements par unité de temps est c et $\lambda = cT$

  Bonne approximation d'une binomiale lorsque p est petit (événement rare):
  $$\mathcal{B}(n, p) \sim \mathcal{P}(n\,p)$$

  Ex d'application:
  loi du nbre de suicides par an dans un pays donné
  loi du nbre d'appels téléphoniques pendant un intervalle de temps T

- **Autres lois discrètes utilisant le schéma de Bernouilli:**
  - Loi géométrique: X = nbre d'essais jusqu'au 1er succès

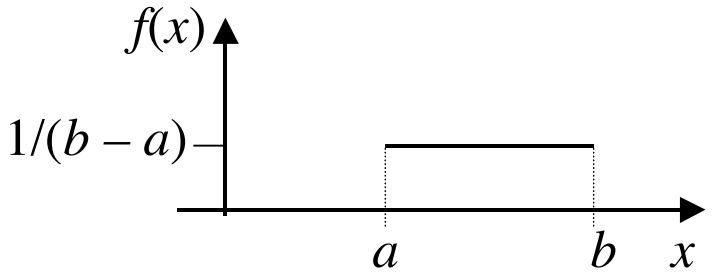
$$P(X = k) = p(1 - p)^{k-1}$$

  - Loi binomiale négative: X = nbre d'essais jusqu'au $r$ème succès

$$P(X = k) = C_{k-1}^{r-1} p^r (1 - p)^{k-r}$$

Lois continues

- **Loi uniforme** sur un intervalle [a, b]:
  même probabilité sur tout l'intervalle: $f(x) = 1/(b - a)$

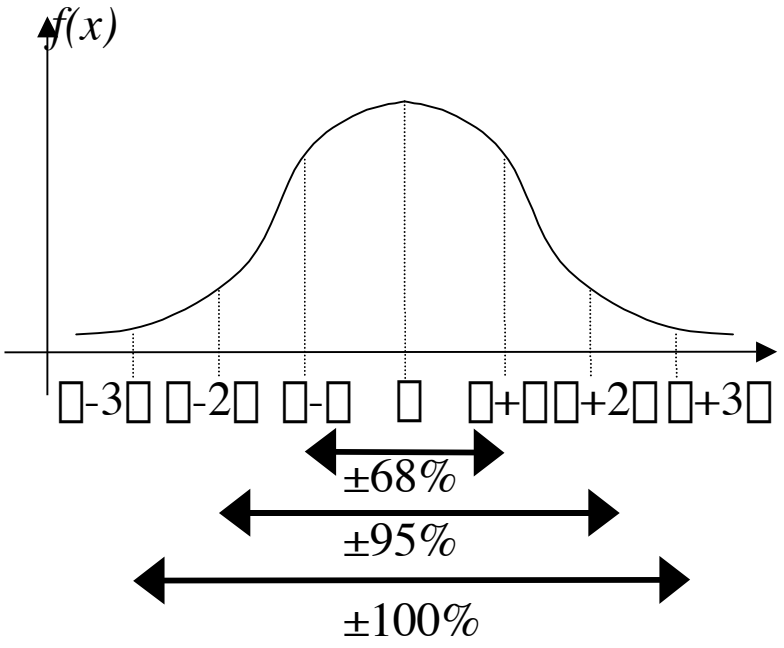


- **Loi normale (ou gaussienne): $N(\mu, \sigma)$**
  moyenne = μ et variance = σ^2

  standardisation: $Z = (X - \mu)/\sigma$

  $\Rightarrow \mu = 0$ et $\sigma = 1$

$$f(x) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]$$



- Rôle fondamental, fréquemment utilisé:
  - **loi limite** sur des échantillons de grandes tailles;
  - approximation des lois binomiale (n grand) et poisson (μ grand);
  - **théorème central-limite:**
    la somme de n variables aléatoires de même loi de moyenne μ et d'écart-type σ tend vers une normale.

- **Exemples d'autres lois continues:**
  - loi log-normale: si $\ln X$ est distribué selon une gaussienne
  - loi exponentielle: $f(x) = \lambda \exp(-\lambda x)$ pour $x > 0$

    ...

Inégalité de Chebychev (1867):

Pour TOUTE variable aléatoire (discrète ou continue)

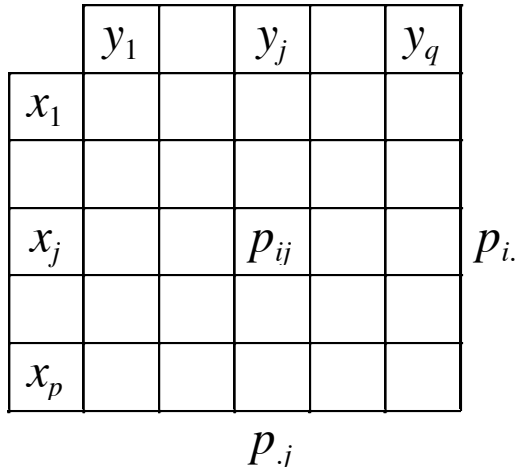
X de moyenne μ et de variance σ^2,

$$P(|X - \mu| \geq c\sigma) \leq 1/c^2$$

Ex: c	$P(\lvert X - \mu \rvert \geq c\sigma)$	pour $N(\mu, \sigma)$
1.5	≤ 44.4%	13.4%
2.0	≤ 25.0%	4.6%
3.0	≤ 11.1%	0.27%
4.0	≤ 6.3%	0.01%

Couple de variables aléatoires

- **Lois associées à un couple (X, Y) de var. discrètes:**
  - ***loi jointe*** = loi du couple (X,Y) définit par la table:
    où $p_{ij} = P(X = x_i \text{ et } Y = y_j)$



  - ***lois marginales*** = lois des variables individuelles:
    loi marginale de X: $$P(X = x_i) = \sum_{j=1}^{q} p_{ij} = p_{i.}$$
    loi marginale de Y: $$P(Y = y_j) = \sum_{i=1}^{p} p_{ij} = p_{.j}$$
  - ***lois conditionnelles:*** loi d'une variable étant donné la valeur prise par l'autre

    loi de X si $Y = y_j$: $p_{i|j} = P(X = x_i \mid Y = y_j) = p_{ij} / p_{.j}$

    loi de Y si $X = x_i$: $p_{j|i} = P(Y = y_j \mid X = x_i) = p_{ij} / p_{i.}$
  - ***Indépendance de X et Y:*** ssi $p_{ij} = p_{i.}\, p_{.j}$ ssi $p_{i|j} = p_{i.}$ ssi $p_{j|i} = p_{.j}$

- **Conditionnement d'une var. continue *Y* par une var. discrète *X***

  Fct de répartition conditionnelle: $F_{Y|x}(y) = P(Y < y \mid X = x)$

  Fct de densité conditionnelle (si elle existe): $f_{Y|x}(y)$ telle que

  $$P(a < Y < b \mid X = x) = \int_a^b f_{Y|x}(y)\,dy$$

  Densité marginale de Y: $f_Y(y) = \sum_x f_{Y|x}(y)\, P(X = x)$

- **Couple (*X*, *Y*) de variables continues (espace à 2 dimension)**

  Fct de répartition jointe de (X, Y): $F_{XY}(x, y) = P(X < x \text{ et } Y < y)$

  Fcts de répartition marginales: $F_X(x) = P(X < x) = F_{XY}(x, \;\;)$ (pas de cond. sur y)

  $$F_Y(y) = P(Y < y) = F_{XY}(\;\;, y) \quad \text{(pas de cond. sur } x\text{)}$$

  Fct de densité jointe de (X, Y) (si elle existe): $f_{XY}(x, y)$ si pour toute région A de l'espace à 2 dimensions:

  $$P((X,Y) \in A) = \iint_A f_{XY}(x, y)\,dx\,dy$$

  = ***volume sous la courbe*** $f_{XY}(x, y)$ au-dessus de la région A

  Indépendance de X et Y: $F_{XY}(x, y) = F_X(x)\, F_Y(y)$ $f_{XY}(x, y) = f_X(x)\, f_Y(y)$

Fcts de densité marginales: $f_X(x) = \int f_{XY}(x, y)\mathrm{d}y \qquad f_Y(y) = \int f_{XY}(x, y)\,\mathrm{d}x$

Loi conditionnelle : $f_{Y|x}(y) = f_{XY}(y) \big/ f_X(x) \quad \text{pour } f_X(x) > 0$

indépendance ssi $f_{Y|x}(y) = f_Y(y)$ ssi $f_{X|y}(x) = f_X(x)$

Espérance conditionnelle: $E(Y \mid x) = \int y\, f_{Y|x}(y)\,\mathrm{d}y$

(rappel: pour une variable X de densité $f(x)$: $E(X) = \int x\, f(x)\,\mathrm{d}x$)

Généralisation à un nbre $n > 2$ de variables:

- variables discrètes: $P(X_1 = x_1, X_2 = x_2, \ldots X_n = x_n)$
- variables continues: A une région dans l'espace à n dimensions:

$$P((X_1, \ldots, X_n) \in A) = \int\!\!\int\!\!\cdots\!\!\int_A f_{X_1 \ldots X_n}(x_1, \ldots, x_n)\,\mathrm{d}x_1 \ldots \mathrm{d}x_n$$

➡ ***distributions marginales, conditionnelles, conditions d'indépendance***

- **Covariance et corrélation:**

  = mesures de la variation conjointe de 2 variables:

covariance: $\sigma_{XY} = \text{cov}(X,Y) = E\big((X - E(X))\,(Y - E(Y))\big)$

$= E(X\,Y) - E(X)\,E(Y)$

avec $E(X\,Y) = \sum_i \sum_j x_i\, y_j\, p_{ij}$ pour des variables X et Y ***discrètes***

$= \int\int x\, y\, f_{XY}(x, y)\mathrm{d}x\,\mathrm{d}y$ " " ***continues***

corrélation: $\rho_{XY} = \text{cor}(X,Y) = \text{cov}(X,Y) \,/\, \sigma_X\, \sigma_Y$

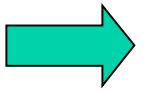
Propriétés:

1) $V(X + Y) = V(X) + V(Y) + 2\,\text{cov}(X,Y)$.

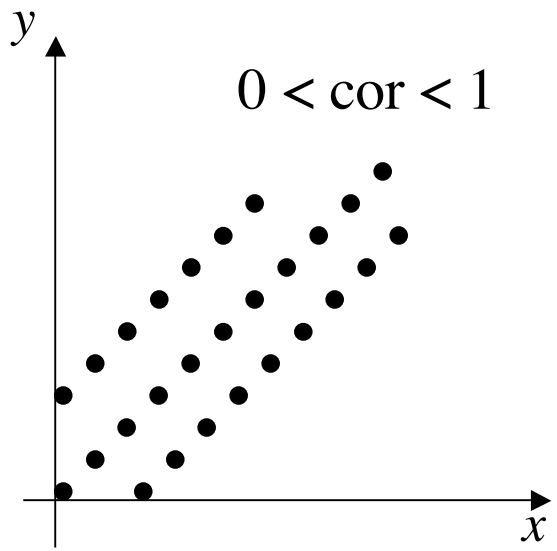
2) SI X et Y indépendantes ALORS $\text{cov}(X,Y) = 0$ (car $E(X\,Y) = E(X)\,E(Y)$)

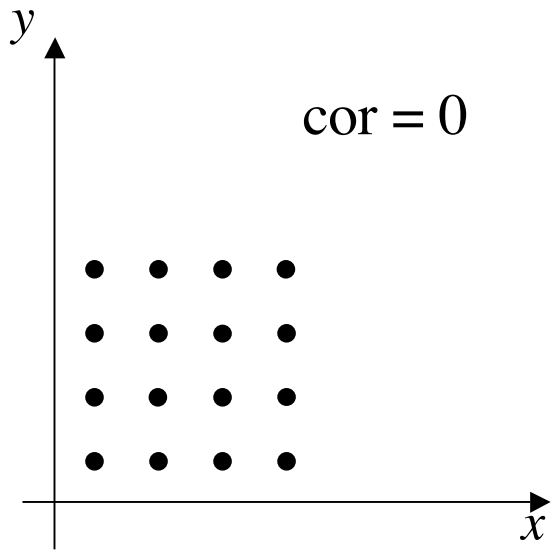
la réciproque est fausse: $\text{cov}(X,Y) = 0$ n'implique pas l'indépendance

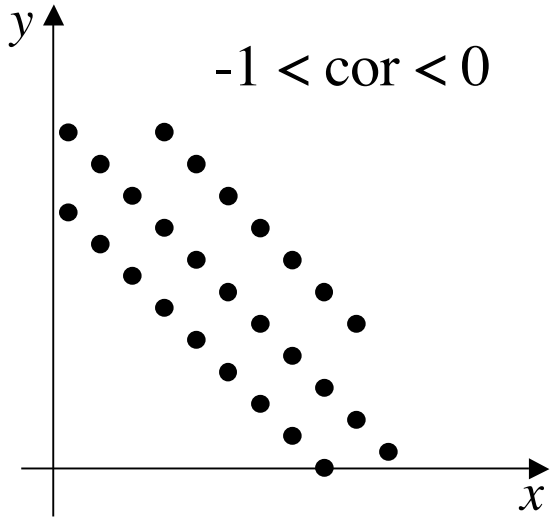
3) $-1 \le \rho_{XY} \le 1$

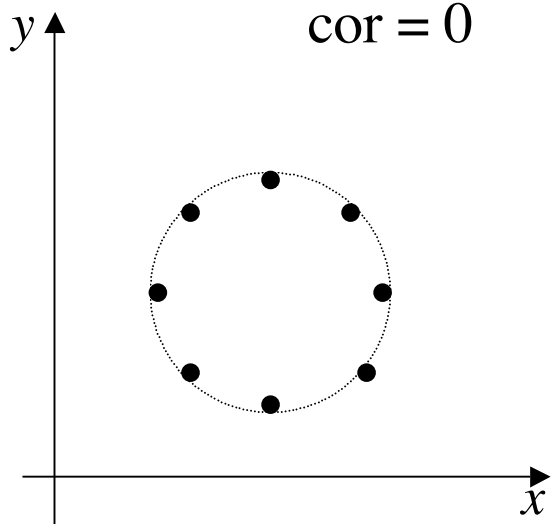
4) $\rho^2_{XY} = 1$ ssi $Y = aX + b$  ρ_{XY} **et** ρ_{XY} **= mesures de relation linéaire**

Soit distribution équiprobable des points (x, y):









- Distribution normale bivariée