

A2DI: Proba/Stat

John Klein

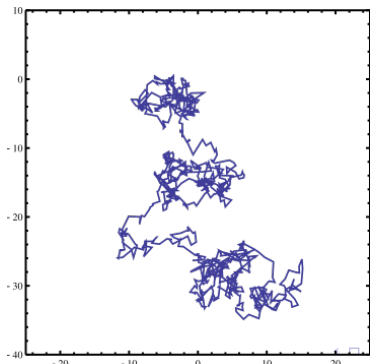
Lille1 Université - CRIStAL UMR CNRS 9189



Pourquoi des probas en ML ?

Raison n°1 :

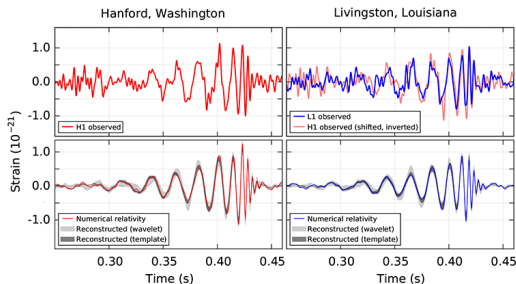
- Si on fait du ML, c'est parce que la **solution** exacte du problème est **inconnue**.
- Elle est inconnue parce que souvent l'univers est trop **complexe** pour trouver un modèle analytique.
- Les probabilités sont justement un moyen très pertinent pour obtenir une **solution approximative mais simple**.



Pourquoi des probas en ML ?

Raison n°2 :

- En ML, on extrait un modèle des **données**.
- La plupart du temps, les données sont **polluées** par des **incertitudes** (erreur de mesure, bruit, arrondi, etc.).
- Les probabilités forment le **modèle d'incertain** le plus couramment utilisé.



Plan du chapitre

1 Probabilités

2 Statistiques

- **Probabilités** : modèle mathématique pour représenter l'**incertain**.
- **Statistiques** : proba + **data**
 - On cherche un modèle probabiliste cohérent avec les données (**fit**),
 - On extrait de ce modèle des attributs ou des **estimés** (classe, prédiction, ..)

Intuitivement, une **proba** c'est quoi ?

- Une **fréquence** d'occurrence dans une expérience **aléatoire**.
- Cas d'école : le lancé de dé.
- $\mathbb{P}(2) = \lim_{n \rightarrow +\infty} \frac{\#\{\text{lancé} = 2\}}{n}$ où n est le nombre de lancés.
- L'ensemble des possibles, ou **univers**, est $\Omega = \{1; 2; 3; 4; 5; 6\}$.
- Pour atteindre la proba, il faut une **infinité de réalisations** de l'expérience aléatoire.

Intuitivement, une **proba** c'est quoi ?

- Une **fréquence** d'occurrence dans une expérience **aléatoire**.
- Parfois, certaines probas peuvent d'obtenir par simple **dénombrement**.
- Cas d'école : le poker.
- $\mathbb{P}(\{\text{paire d'as}\}) =$

Intuitivement, une **proba** c'est quoi ?

- La représentation d'une **connaissance partielle**.
- Ex : Quelle est l'**âge** du président ?
- **info** n°1 : Il a 3 enfants.
- **info** n°2 : Il est plus jeune que son prédécesseur âgé de 70 ans.
- Ces deux informations sont des éléments de preuve (**evidence**) qui induisent des valeurs probables pour l'**âge** du président.
- Il n'y a **aucun aléa**, on ne peut pas prendre une infinité de présidents et calculer une fréquence !
- On parle alors de probabilités **subjectives** par opposition aux autres appelées **objectives** ou **fréquentistes**.

Mathématiquement, une **proba** c'est quoi ?

- Une **mesure** normalisée à 1.

Définition

Soit Ω un ensemble et 2^Ω l'ensemble des sous-parties de Ω .

Soit μ une application de 2^Ω dans $[0; +\infty[$. On dit que μ est une **mesure** ssi :

- $\mu(\emptyset) = 0$,
- pour tout A et B sous-ensembles de Ω tels que $A \cap B = \emptyset$, on a $\mu(A \cup B) = \mu(A) + \mu(B)$.

pour aller plus loin : une mesure est en fait définie sur une tribu (ou σ -algèbre) associée à Ω .

Mathématiquement, une **proba** c'est quoi ?

- Une **mesure** normalisée à 1.

Définition

Soit μ une mesure sur 2^Ω . On dit que μ est une **mesure de probabilité** ssi :

- $\mu(\Omega) = 1$.

Notion de Variable aléatoire

- Imaginons un jeu :
 - La partie coûte 20€.
 - On lance un dé à 6 faces équiprobables.
 - Le gain est égal au carré de la face obtenue.
- Comment exprimer simplement le retour sur investissement en fonction de l'issue du jeu ?

Variable aléatoire discrète

- On dit d'une variable aléatoire X qu'elle est **discrète** si l'ensemble des valeurs qu'elle prend est typiquement \mathbb{Z} ou \mathbb{N} ou un ensemble fini comme $\{1; \dots; \ell\}$.
- Exemple :
 - Pile ou Face, $X \in \{F; P\}$,
 - Lancé de dé, $X \in \{1; 2; 3; 4; 5; 6\}$,
 - Classe d'un exemple $X \in \{c_1, \dots, c_\ell\}$,
 - Nombre de personne dans la file du R.U. $X \in \mathbb{N}$.

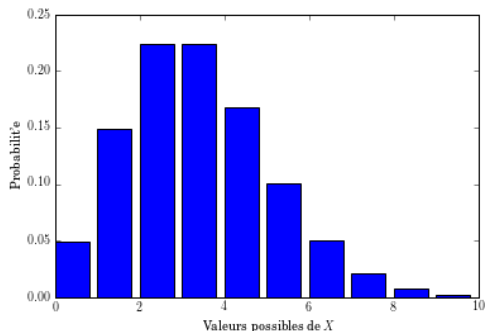
Variable aléatoire continue

- On dit d'une variable aléatoire X qu'elle est **continue** si l'ensemble des valeurs qu'elle prendre est typiquement \mathbb{R} (ou segment de \mathbb{R}).
- Exemple :
 - Sortie d'un capteur de température, $X \in [-273.15; +\infty]$,
 - Cours d'une action, $X \in [0; \infty]$,
 - Proportion de mâles dans une population $X \in [0, \dots, 1]$,
 - Solution d'un problème de régression $X \in \mathbb{R}$.

Variable aléatoire discrète : **distribution** Soit \mathbb{X} l'ensemble des valeurs possibles de X .

Définition

Les probabilités associées à chaque valeur possible d'une variable aléatoire discrète sont regroupées dans une fonction appelée **Loi** ou **Distribution** de X et qu'on notera $p_X : \mathbb{X} \rightarrow [0; 1]$ et $p_X(i) = \mathbb{P}(\{X = i\})$



Variable aléatoire continue : densité Soit \mathbb{X} l'ensemble des valeurs possibles de X .

- $\forall a \in \mathbb{X}$, on a $\mathbb{P}(X = a) = 0!$
- Cela signifie que pour ces v.a. une probabilité nulle n'implique pas qu'un événement est impossible !
- On obtient (éventuellement) des probas non nulles que pour des événements du type $X \in [a; b]$ avec $a < b$.
- On doit utiliser une autre fonction pour résumer nos croyances sur les chances d'observer une valeur a plutôt que b .

Définition

On appelle **fonction de répartition** d'une v.a. la fonction $F_X: \mathbb{X} \rightarrow [0; 1]$ telle que :

$$F_X(a) = \mathbb{P}(X \leq a). \quad (1)$$

Densité de probabilité

- La définition de la **fonction de répartition** ou **distribution cumulée** s'applique aussi aux v.a. discrètes.
- Pour les v.a., sous réserve de pouvoir dériver F_X , on introduit une autre fonction qui caractérise la concentration des chances pour une valeur particulière :

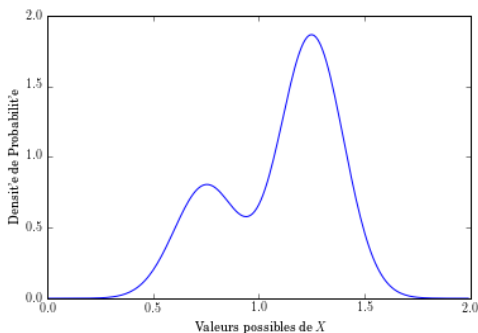
Définition

On appelle **densité de probabilité** d'une v.a. continue la fonction $p_X: \mathbb{X} \rightarrow [0; +\infty]$ telle que :

$$p_X(a) = F'_X(a). \quad (2)$$

Densité de probabilité

- On a $\int p_X(u) du = 1$.
- En revanche, il est possible d'avoir $p_X(u) > 1$!



Notation (abusive) : $p_X(\textcolor{red}{A}) = \int_{\textcolor{red}{A}} p_X(u) du = \mathbb{P}(X \in \textcolor{red}{A})$

Espérance

- Reprenons l'exemple du **jeu** :
 - La partie coûte 20€.
 - On lance un dé à 6 faces équiprobables.
 - Le gain est égal au carré de la face obtenue.
- Quel **retour sur investissement** puis-je **espérer** après un grand nombre de parties ?

partie n°1	1	2	3	4	5	6	7	8	9	10	11	12
issue	1	6	5	2	2	2	4	4	4	2	4	4
gain	-19	16	5	-16	-16	-16	-4	-4	-4	-16	-4	-4

Espérance

Définition

On appelle **espérance** d'une fonction f de la v.a. X , la quantité notée $\mathbb{E}_X[f]$ telle que :

$$\mathbb{E}_X[f] = \begin{cases} \sum_{a \in \mathbb{X}} f(a) p_X(a) & \text{si } X \text{ est discrète} \\ \int_{\mathbb{X}} f(u) p_X(u) du & \text{si } X \text{ est continue} \end{cases} . \quad (3)$$

Cas particulier : si $f = \mathbb{I}_A$ est la **fonction indicatrice** sur $A \subset \mathbb{X}$.

$$\mathbb{I}_A(u) = \begin{cases} 1 & \text{si } u \in A \\ 0 & \text{sinon} \end{cases} \quad \text{et donc } \mathbb{E}_X[\mathbb{I}_A] = p_X(A). \quad (4)$$

→ L'espérance est une notion plus générale que la distribution.

Espérance

Souvent, on note $\mathbb{E}_X[id] = \mathbb{E}[X]$.

On les **propriétés** suivantes :

- $\mathbb{E}[cte] = cte$,
- $\mathbb{E}[aX + Y] = a\mathbb{E}[X] + \mathbb{E}[Y]$,
- $\mathbb{E}[X] \mathbb{E}[Y] \neq \mathbb{E}[XY]$.

Couple de v.a. : (X, Y)

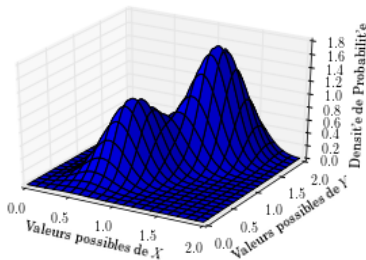
En ML, on doit souvent manipuler des ensembles de v.a. :

- Les exemples d'apprentissage sont souvent **multi-dimensionnels** et chaque dimension engendre une v.a. X_i .
- Si X représente les exemples alors $\mathbf{X} = [X_1 \dots X_d]$ sera un **vecteur aléatoire**.
- La **prédiction** est elle aussi **incertaine** et modéliser par une v.a. Y

Couple de v.a. : (X, Y)

On généralise la notion de distribution pour un couple et on parle de **loi jointe** notée $p_{X,Y}$:

- $p_{X,Y}(a, b) = \mathbb{P}(X=a \text{ et } Y=b)$ si X et Y sont **discrètes** et $a \in \mathbb{X}$, $b \in \mathbb{Y}$.
- $p_{X,Y}(A, B) = \mathbb{P}(X \in A \text{ et } Y \in B)$ si X et Y sont **continues** et $A \subset \mathbb{X}$, $B \subset \mathbb{Y}$.
- $p_{X,Y}(A, b) = \mathbb{P}(X \in A \text{ et } Y=b)$ si X est **continue** tandis que Y est **discrète** et $A \subset \mathbb{X}$, $b \in \mathbb{Y}$.



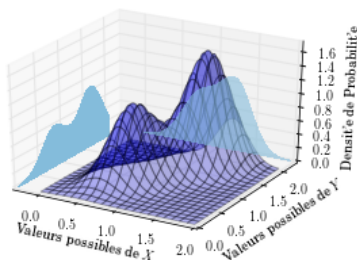
Couple de v.a. (X, Y) : marginales

Définition

On appelle **loi marginale** la loi p_X d'une v.a. X deduite d'une loi jointe $p_{X,Y}$. On a :

$$p_X(a) = \begin{cases} \sum_{b \in \mathbb{Y}} p_{X,Y}(a, b) & \text{si } Y \text{ est discrète} \\ \int_{\mathbb{Y}} p_{X,Y}(a, y) dy & \text{si } Y \text{ est continue} \end{cases} \quad (5)$$

En général, il n'est pas possible de remonter à la jointe à partir des marginales.

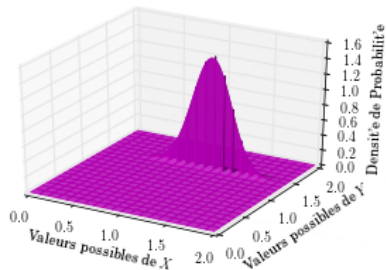
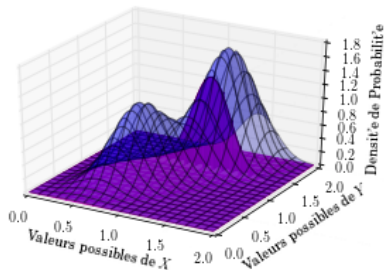


Couple de v.a. (X, Y) : loi conditionnelle

Définition

On appelle **loi conditionnelle** la loi $p_{X|Y=b}$ d'une v.a. X après avoir observé l'événement $Y = b$ de probabilité (ou de densité) non nulle. On a :

$$p_{X|Y=b}(a) = \frac{p_{X,Y}(a, b)}{p_Y(b)}. \quad (6)$$



Couple de v.a. (X, Y) : Théorème de Bayes

En ML, on souhaite pouvoir renverser un conditionnement. Le théorème de Bayes est un élément-clé d'un tel processus :

$$p_{Y|X=a}(b) = \frac{p_{X|Y=b}(a) p_Y(b)}{p_X(a)}, \quad (7)$$

$$= \frac{p_{X|Y=b}(a) p_Y(b)}{\int_{\mathbb{Y}} p_{X|Y=u}(a) p_Y(u) du}, \quad (8)$$

$$\propto p_{X|Y=b}(a) p_Y(b). \quad (9)$$

Couple de v.a. (X, Y) : Indépendance

Définition

On dit que deux v.a.s X et Y qu'elles sont **indépendante**, noté $X \perp\!\!\!\perp Y$ si la jointe est le produit des marginales :

$$p_{X,Y}(A, B) = p_X(A) \times p_Y(B). \quad (10)$$

Exemple

X est le résultat d'un lancé de dé.

Y est le sexe du lanceur.

Le sexe du lanceur n'a aucune influence sur le résultat du lancé, d'où $X \perp\!\!\!\perp Y$.

Couple de v.a. (X, Y) : Indépendance

On peut également caractériser l'indépendance de deux v.a.s X et Y par le conditionnement :

$$X \perp\!\!\!\perp Y \iff p_{X|Y=b}(a) = p_X(a), \forall a \in \mathbb{X}, b \in \mathbb{Y}. \quad (11)$$

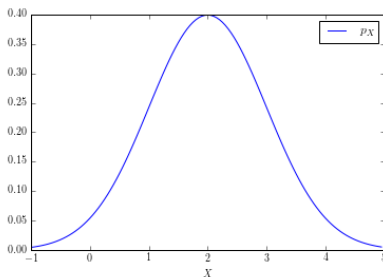
Connaître Y ne nous apporte aucune information sur X !

Couple de v.a. (X_1, X_2) : Indépendance / Le cas gaussien

On dit qu'une v.a. continue X suit une loi gaussienne (ou normale), noté $X \sim \mathcal{N}(\mu, \sigma)$, si sa densité de probabilité vaut :

$$p_X(u) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(u-\mu)^2}{2\sigma^2}}. \quad (12)$$

La famille des distributions gaussiennes est paramétrée par μ et σ .
Exemple pour $\mu = 2$ et $\sigma = 1$.



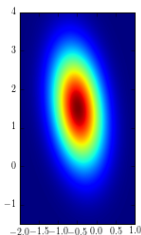
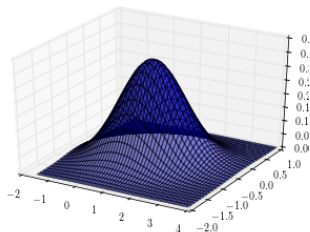
Couple de v.a. (X_1, X_2) : Indépendance / Le cas gaussien multivarié

On dit qu'un vecteur aléatoire continu \mathbf{X} suit une loi gaussienne multivariée, noté $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, si sa densité de probabilité jointe vaut :

$$p_{\mathbf{X}}(\mathbf{u}) = \frac{1}{(2\pi)^{d/2} \det(\boldsymbol{\Sigma})^{1/2}} e^{-\frac{1}{2}(\mathbf{u}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{u}-\boldsymbol{\mu})}. \quad (13)$$

La famille des distributions gaussiennes en dimension d est paramétrée par le vecteur $\boldsymbol{\mu}$ et la matrice $\boldsymbol{\Sigma}$.

Exemple pour $d=2$, $\boldsymbol{\mu} = \begin{pmatrix} 1 \\ -0.5 \end{pmatrix}$ et $\boldsymbol{\Sigma} = \begin{pmatrix} 1 & 0.1 \\ 0.1 & 0.2 \end{pmatrix}$.



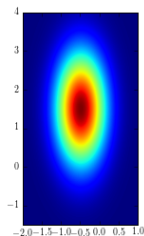
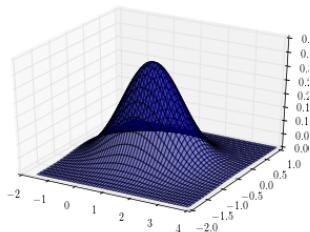
Couple de v.a. (X_1, X_2) : Indépendance / Le cas gaussien multivarié

Les composantes d'un vecteur aléatoire gaussien $\mathbf{X} = (X_1 \dots X_d)^T$ sont indépendante ssi la matrice Σ est diagonale.

$$p_{\mathbf{X}}(\mathbf{u}) = \frac{1}{(2\pi)^{d/2} \det(\Sigma)^{1/2}} e^{-\frac{1}{2}(\mathbf{u}-\mu)^T \Sigma^{-1}(\mathbf{u}-\mu)}. \quad (14)$$

Couple de v.a. (X_1, X_2) : Indépendance / Le cas gaussien multivarié

Exemple pour $d=2$, $\mu = \begin{pmatrix} 1 \\ -0.5 \end{pmatrix}$ et $\Sigma = \begin{pmatrix} 1 & 0 \\ 0 & 0.2 \end{pmatrix}$.



Couple de v.a. (X, Y) : Indépendance

ATTENTION

Ce n'est pas parce $p_X = p_Y$ qu'on a X dépend de Y !

Avoir les mêmes probas ne signifie être lié d'une quelconque manière.

Exemple

X est le résultat d'un lancé de dé.

Y est le résultat d'un lancé d'autre dé.

On a $p_X = p_Y$ mais le résultat du 1er lancé est **indépendant** du 2ème !

Triplet de v.a. (X, Y, Z) : Indépendance conditionnelle

Définition

On dit que deux v.a.s X et Y qu'elles sont **conditionnellement indépendante** sachant Z , noté $(X \perp\!\!\!\perp Y) | Z$ si la jointe sachant Z est le produit des marginales sachant Z :

$$p_{X,Y|Z=c}(A, B) = p_{X|Z=c}(A) \times p_{Y|Z=c}(B). \quad (15)$$

Exemple

X est une v.a. binaire représentant la possibilité d'être atteint de la grippe.

Y est une v.a. binaire représentant la possibilité d'avoir de la fièvre.

Z est une v.a. binaire représentant la possibilité de souffrir de maux de tête.

X est une pathologie tandis que Y et Z sont des symptômes.

On a :

Triplet de v.a. (X, Y, Z) : Indépendance conditionnelle et Causalité

Dans l'exemple précédent :

- X est une cause,
- Y et Z sont des effets,
- Y et Z ne sont pas indépendantes, il y a de bonne chances d'avoir $Y = \text{true}$ quand $Z = \text{true}$,
- mais il n'y a pas de lien de causalité en Y et Z .

Triplet de v.a. (X, Y, Z) : **Indépendance conditionnelle**

L'indépendance conditionnelle s'exprime aussi comme suit :

$$(X \perp\!\!\!\perp Y) | Z \iff p_{X|Y=b, Z=c}(a) = p_{X|Z=c}(a), \forall a \in \mathbb{X}, b \in \mathbb{Y}. \quad (16)$$

Une fois Z connu, la connaissance de Y n'apporte rien concernant la valeur de X .

ATTENTION

$$(X \perp\!\!\!\perp Y) | Z \not\Leftarrow X \perp\!\!\!\perp Y, \quad (17)$$

$$(X \perp\!\!\!\perp Y) | Z \not\Rightarrow X \perp\!\!\!\perp Y. \quad (18)$$

Plan du chapitre

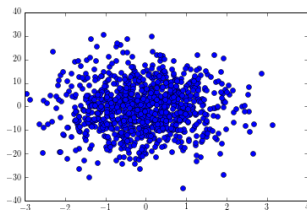
1 Probabilités

2 Statistiques

.. et les **données** furent !

Le mot **statistique** peut désigner :

- un **échantillon** recueilli $\{x_i\}_{i=1}^n$ où chaque x_i est tiré selon une même loi L , noté $x_i \sim L$. Elle est représentative de la **population** générale.



- un **calcul** opéré sur une loi L ou un échantillon suivant une loi L et permettant de décrire le comportement de L .

$$s = \sum_{j=1}^{\ell} \frac{n \left(p_X(j) - \frac{1}{n} \sum_{i=1}^n \mathbb{I}_j(x_i) \right)^2}{p_X(j)} \quad (19)$$

Statistique : Moments

Les moments sont les statistiques descriptives les plus répandus permettant de caractériser une distribution.

Définition

Soit X une v.a. et $\mu_X^{(i)}$ Son **moment** d'ordre i est donné par :

$$\mu_X^{(i)} = \mathbb{E} [X^i]. \quad (20)$$

Définition

Soit X une v.a. et $\nu_X^{(i)}$ Son **moment centré** d'ordre i est donné par :

$$\nu_X^{(i)} = \mathbb{E} [(X - \mathbb{E}[X])^i]. \quad (21)$$

Statistique : Moments

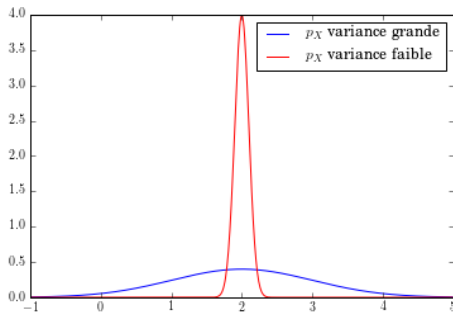
Cas particuliers de moments centrés :

Définition

Le moment centré d'**ordre 2** d'une v.a. X est appelé **variance** de X , notée $\text{Var}[X]$.

La racine de la variance est appelée **écart-type**, noté $\sigma_X = \sqrt{\text{Var}[X]}$.

La variance caractérise l'**étalement** d'une distribution.



Statistique : Moments

Définition

Soit X une v.a. et $\nu_X^{(i)}$ Son **moment centré réduit** d'ordre i est donné par :

$$m_X^{(i)} = \mathbb{E} \left[\left(\frac{X - \mathbb{E}[X]}{\sigma_X} \right)^i \right]. \quad (22)$$

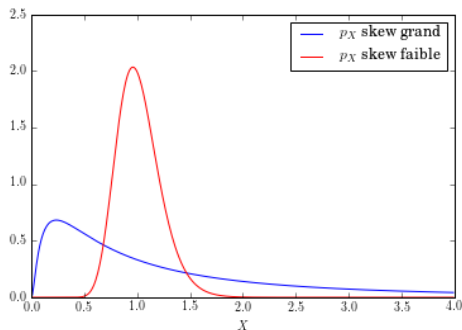
Statistique : Moments

Cas particuliers de moments centrés réduits :

Définition

Le moment centré réduit d'ordre 3 d'une v.a. X est appelé **skew** de X .

La variance caractérise l'**asymétrie** d'une distribution.



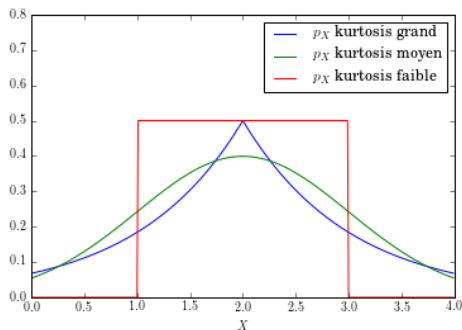
Statistique : Moments

Cas particuliers de moments centrés réduits :

Définition

Le moment centré réduit d'ordre 4 d'une v.a. X est appelé **kurtosis** de X .

La variance caractérise la **platitude** d'une distribution.



Couple de v.a. (X_1, X_2) : Covariance

En ML, on a souvent besoin de savoir si les **variations** d'une v.a. X sont proches de celle d'une autre v.a. Y :

- Quand X baisse est-ce que Y baisse aussi ?
- Quand X augmente est-ce que Y augmente aussi ?

Le calcul de la covariance permet de répondre en partie à cette question :

Définition

On note $\text{cov}(X, Y)$ la **covariance** des v.a.s X et Y définie par :

$$\text{cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])]. \quad (23)$$

On a :

- $\text{cov}(X, X) = \text{var}(X)$,
- $\text{cov}(aX + Y, Z) = a\text{cov}(X, Z) + \text{cov}(Y, Z)$.

Couple de v.a. (X_1, X_2) : Covariance

L'indépendance entraîne une covariance nulle :

$$X \perp\!\!\!\perp Y \quad \Rightarrow \quad \text{cov}(X, Y) = 0. \quad (24)$$

Covariance et causalité :

- Exemple des maladies et symptômes (c.f. indep. cond.).
- Les symptômes ont une covariance positive.
- La covariance n'est pas une preuve de causalité.

Attention à tous ces articles de presse du style « manger bio fait gagner 5ans d'espérance de vie »

Couple de v.a. (X_1, X_2) : Covariance

Pour un vecteur aléatoire $\mathbf{X} = (X_1 \dots X_d)$, on appelle **matrice de variance-covariance**, la matrice carrée $d \times d$ définie positive \mathbf{S} dont les éléments sont donnés par $S_{ij} = \text{cov}(X_i, X_j)$.

Exercice : Prouver que la matrice de variance-covariance d'un vecteur $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ est $\boldsymbol{\Sigma}$ (pour $d = 2$).

Couple de v.a. (X_1, X_2) : **Corrélation**

La corrélation est une forme **normalisée** de la covariance :

Définition

On note $\rho(X, Y)$ le coefficient de **corrélation** (de Pearson) des v.a.s X et Y défini par :

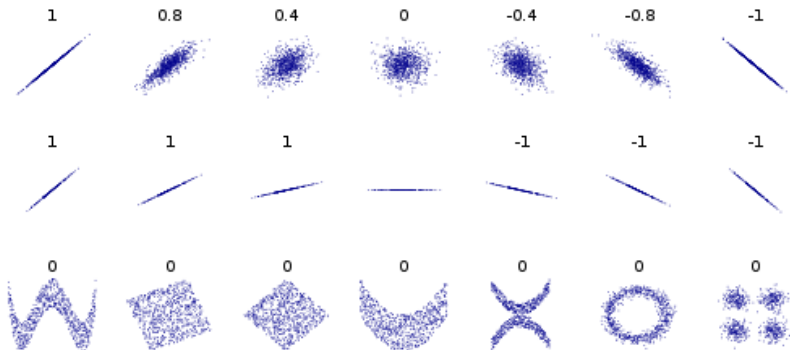
$$\rho(X, Y) = \frac{\text{cov}(X, Y)}{\sqrt{\text{var}(X) \text{var}(Y)}}. \quad (25)$$

On a $\rho(X, Y) \in [-1; 1]$:

- $\rho(X, Y) = 1$ signifie que X et Y sont liés linéairement
- $\rho(X, Y) = -1$ signifie la même chose mais, leurs variations ont un signe opposé
- $\rho(X, Y) = 0$ ne signifie pas grand chose..

Couple de v.a. (X_1, X_2) : Covariance

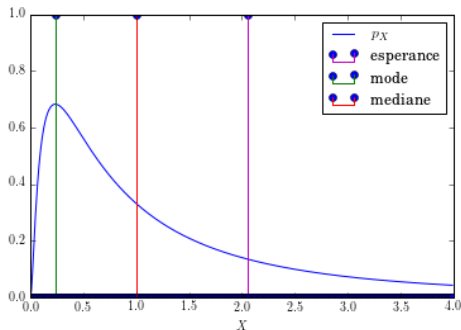
Exemples de coefficients de corrélation pour différents échantillons :



Comment résumer une distribution ?

1°/ avec un seul point

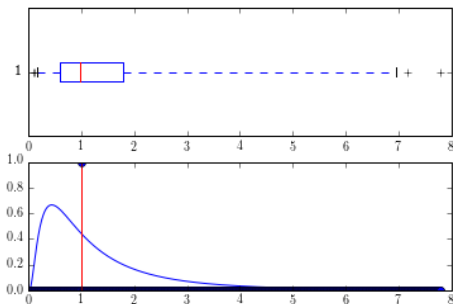
- avec l'espérance $= \mathbb{E}(X)$: c'est la valeur dont la moyenne d'un échantillon sera le plus proche,
- ou le mode $= \arg \max_{u \in \mathbb{X}} p_X(u)$: c'est la valeur la plus probable,
- ou la médiane $= F_X^{-1}(0.5)$: c'est la valeur qui sépare les autres en 2 groupes de probabilité 0.5.



Comment résumer une distribution ?

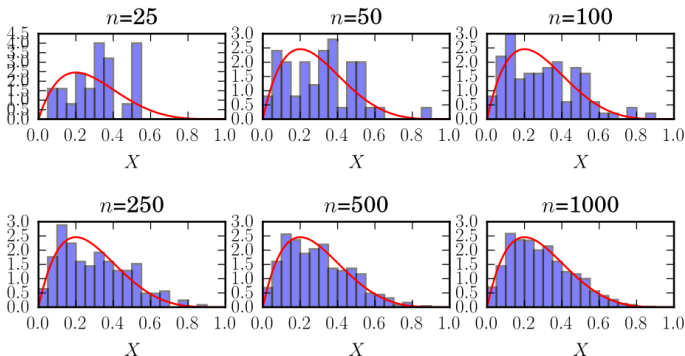
2°/ avec 5 statistiques

- la **médiane** = $F_X^{-1}\left(\frac{1}{2}\right)$,
- et le 1er **quartile** = $F_X^{-1}\left(\frac{1}{4}\right)$ et 3ème **quartile** = $F_X^{-1}\left(\frac{3}{4}\right)$,
- et le 2ème **percentile** = $F_X^{-1}\left(\frac{2}{100}\right)$ et 98ème **percentile** = $F_X^{-1}\left(\frac{98}{100}\right)$.



Comment remonter à la distribution p_X qui a généré nos données $\{x_i\}_{i=1}^n$?
Rangeons les données dans des cases (bins) !

- On découpe \mathbb{X} en r bins (en général de taille égale).
- On pose $\hat{p}_X(A_i) = \frac{1}{n} \sum_{i=1}^n \mathbb{I}_{A_i}(x_i)$ (distribution empirique)
- Quand n est grand, $\hat{p}_X(A_i) \approx p_X(A_i)$.



Pourquoi ça marche? Loi des grands nombres

- Voyons nos données comme un **vecteur** de taille n : $[x_1 \dots x_n]^T$.
- Ce vecteur est l'unique réalisation d'un vecteur aléatoire **\mathbf{X}** dont chaque entrée suit une **distribution identique** L et **indépendante** des autres entrées.
- On parle d'échantillon indépendant et identiquement distribué (**iid**).
- Notons $\mathbf{Y}_n = \frac{1}{n} \sum_{i=1}^n X_i$ la **moyenne** des v.a.s correspondant aux entrées du vecteur.
- Soit μ l'espérance de L .

Théorème

En reprenant les notations ci-dessus, on a :

$$\mathbb{P} \left(\lim_{n \rightarrow +\infty} \mathbf{Y}_n = \mu \right) = 1 \quad (26)$$

Pourquoi ça marche? Loi des grands nombres

- Ce résultat signifie que quand n est très grand, une réalisation de Y_n ne sera plus une v.a. mais une constante égale à μ !
- Appliquons ce résultat à $\mathbf{R} = \mathbb{I}_A \circ \mathbf{X}$ avec $A \subset \mathbb{X}$. On a alors :
 - $Y_n = \hat{p}_X(A)$,
 - $\mu = \mathbb{E}_X[\mathbb{I}_A] = p_X(A)$.

A quelle **vitesse** converge-t-on ? **Théorème Central Limite**

On aimerait pouvoir **garantir** en fonction de **n** un résultat du type

$$\mathbb{P}(|\hat{p}_X(A; n) - p_X(A)| > \tau) = \epsilon. \quad (27)$$

Théorème Central Limite

Soient X_1, \dots, X_n n v.a. indépendantes suivant une même loi L d'espérance finie μ et de variance finie non nulle σ^2 . Soit $Y_n = \frac{1}{n} \sum_{i=1}^n X_i$. On a

$$Y_n \xrightarrow{\text{loi}} \mathcal{N}\left(\mu; \frac{\sigma^2}{n}\right) \quad (28)$$

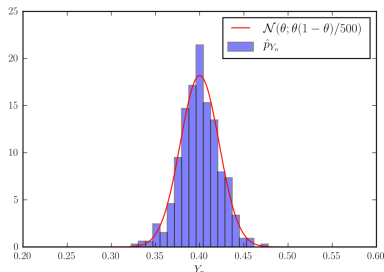
A quelle **vitesse** converge-t-on ? **Théorème Central Limite**

Exemple : soit X une variable aléatoire **binaire** : $\mathbb{X} = \{0; 1\}$. Il existe $\theta \in [0; 1]$ avec

$$p_X(0) = \theta, \quad (29)$$

$$p_X(1) = 1 - \theta. \quad (30)$$

On dit que X suit une loi de **Bernoulli**, noté $X \sim \text{Ber}(\theta)$. Prenons $n = 500$ tirage de la loi $\text{Ber}(\theta)$. Répétons $m = 400$ fois l'expérience et construisons alors l'histogramme de Y_n à comparer avec la distribution théorique $\mathcal{N}\left(\theta; \frac{\theta(1-\theta)}{500}\right)$.



A quelle **vitesse** converge-t-on ? **Théorème Central Limite**

Revenons à

$$\mathbb{P}(|\hat{p}_X(A; n) - p_X(A)| > \tau) = \epsilon. \quad (31)$$

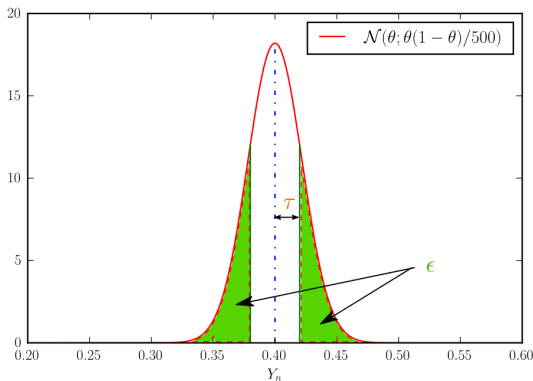
- Supposons que je tire **n** échantillons $\{x_1, \dots, x_n\}$ selon la loi de X .
- La probabilité d'avoir $x_i \in A$ est $p_X(A)$.
- La probabilité d'avoir $x_i \notin A$ est $1 - p_X(A)$.
- Je peux transformer mes échantillons x_i en échantillons **binaires** $z_i = \mathbb{I}_A(x_i)$.
- Les z_i sont tirés selon $\text{Ber}(\theta = p_X(A))$!

A quelle **vitesse** converge-t-on ? **Théorème Central Limite**

Revenons à

$$\mathbb{P}(|\hat{p}_X(A; n) - p_X(A)| > \tau) = \epsilon. \quad (32)$$

- Quand **n** est grand, ma probabilité **ε** correspond à la surface suivante :



Comment remonter à la distribution p_X qui a généré nos données $\{x_i\}_{i=1}^n$?
Supposons que p_X appartienne à une famille paramétrée $\{f_\theta\}_{\theta \in \Theta}$.

Comment remonter à la distribution p_X qui a généré nos données $\{x_i\}_{i=1}^n$?

Supposons que p_X appartienne à une famille paramétrée $\{f_\theta\}_{\theta \in \Theta}$.

→ On peut alors calculer la fonction de vraisemblance $\mathcal{L}(\theta)$:

$$\begin{aligned}\mathcal{L}(\theta) &= p(\mathcal{D}|\theta), \\ &= \prod_{i=1}^n f_\theta(\mathbf{x}_i).\end{aligned}\tag{33}$$

Comment remonter à la distribution p_X qui a généré nos données $\{x_i\}_{i=1}^n$?

Supposons que p_X appartienne à une famille paramétrée $\{f_\theta\}_{\theta \in \Theta}$.

→ On peut alors calculer la fonction de vraisemblance $\mathcal{L}(\theta)$:

$$\begin{aligned}\mathcal{L}(\theta) &= p(\mathcal{D}|\theta), \\ &= \prod_{i=1}^n f_\theta(\mathbf{x}_i).\end{aligned}\tag{33}$$

ATTENTION

$\mathcal{L}(\theta)$ n'est pas une distribution :

$$\int \mathcal{L}(\theta) d\theta \neq 1.\tag{34}$$

Comment remonter à la distribution p_X qui a généré nos données $\{x_i\}_{i=1}^n$?
 Supposons que p_X appartienne à une famille paramétrée $\{f_\theta\}_{\theta \in \Theta}$.

→ On peut alors calculer la fonction de vraisemblance $\mathcal{L}(\theta)$:

$$\begin{aligned}\mathcal{L}(\theta) &= p(\mathcal{D}|\theta), \\ &= \prod_{i=1}^n f_\theta(\mathbf{x}_i).\end{aligned}\tag{33}$$

ATTENTION

$\mathcal{L}(\theta)$ n'est pas une distribution :

$$\int \mathcal{L}(\theta) d\theta \neq 1.\tag{34}$$

En ML, on préfère souvent la *negative log-Likelihood* :

$$\text{NLL}(\theta) = -\log \mathcal{L}(\theta).$$

Comment remonter à la distribution p_X qui a généré nos données $\{x_i\}_{i=1}^n$?
 → vraisemblance $\mathcal{L}(\theta)$.

Exemple

Soient les données suivantes $\mathbf{x}_i \sim \text{Ber}(\theta) : \{0; 0; 0; 1; 0; 0; 1; 0; 1; 0\}$.
 On a $n = 10$ et

$$\begin{aligned}\mathcal{L}(\theta) &= \prod_{i=1}^{10} \theta^{1-x_i} (1-\theta)^{x_i}, \\ &= \theta^7 (1-\theta)^3.\end{aligned}$$

