III. Mesures de probabilité

Une **tribu/σ-algèbre sur un ensemble** est un ensemble de parties qui est stable par complémentaire, stable par union/intersection dénombrable, et qui contient l'ensemble/le vide.

Un **espace mesurable** est un ensemble muni d'une tribu sur cet ensemble. Une **partie mesurable d'un espace mesurable** est un élément de la tribu de l'espace mesurable.

Une **mesure** sur un espace mesurable est une fonction de la tribu vers $[0, +\infty]$, tel que l'image du vide est 0 **(normalisation)** et telle que **(σ-additivité)** la mesure d'une union dénombrable de mesurables 2 à 2 disjoints est égale à la somme des mesures de chaque dénombrable.

Un **espace mesuré** est un espace mesurable muni d'une mesure sur cet espace.

Une partie d'un espace mesuré est **négligeable** ssi elle est incluse dans un mesurable de mesure nulle (ssi elle est de mesure nulle pour Lebesgue).

Une propriété $P(x \in X)$ définie sur un espace mesuré est **vrai presque partout pour la mesure μ (μ pp)** si elle est fausse sur un ensemble négligeable.

Une **mesure/loi de probabilité** est une mesure telle que la mesure de l'univers vaut 1.

Un **espace probabilisé** correspond à un espace mesurable muni d'une mesure de probabilité.

Dans ce contexte on appelle **univers** l'espace et on le note généralement Ω, on appelle **évènement**, une partie mesurable de l'univers, on appelle **évènement élémentaire** un singleton de l'univers qui est un évènement (ce n'est pas nécessairement le cas).

On dit qu'un évènement est **presque sûr** relativement à une mesure de proba P ssi il est vrai P-pp ssi son complémentaire est négligeable ssi $P(A) = 1$.

Une loi de probabilité vérifie les propriétés classiques :

Une probabilité est à valeurs dans $[0,1]$. $0 \leq P(A) \leq B$

La probabilité de l'univers est 1 et la probabilité du vide est 0. $P(\Omega) = 1,\ P(\emptyset) = 0$

σ-Additivité. La probabilité d'une union dénombrable d'évènements disjoints 2 à 2 égale la somme des probabilités de ces évènements. $\forall i \neq j\ A_i \cap A_j = \emptyset \Rightarrow\ P(\bigcup_{k\in N} A_k) = \sum_{k=0}^{\infty} P(A_k)$

La probabilité d'un évènement majore celle de ses sous évènements. $A \subseteq B \Rightarrow P(A) \leq P(B)$

$\forall A \in M\ \ P(C_\Omega A) = 1 - P(A)$

$\forall A, B \in M\ P(A \setminus B) = P(A) - P(A \cap B)$

$\forall A, B \in M\ P(A \cup B) = P(A) + P(B)$

$\forall (A_k)_{1\leq k\leq n} \in M^n\ P(\bigcup_{k=1}^{n} A_k) = \sum_{k=1}^{n} P(A_k) - \sum_{k=2}^{n} (-1)^k \sum_{1\leq i_1<\cdots<i_k\leq n} P\left(\bigcap_{j=1}^{k} A_{i_j}\right) =$

$\sum_{k=1}^{n} (-1)^{k+1} \sum_{1\leq i_1<\cdots<i_k\leq n} P\left(A_{i_1} \cap \ldots \cap A_{i_k}\right) = \sum_{\emptyset\neq J\subseteq\{1,\ldots,n\}} (-1)^{|J|+1} P\left(\bigcap_{j\in J} A_j\right)$

$\forall (A_k)_{1\leq k\leq n} \in M^n\ P(\bigcup_{k=1}^{n} A_k) \leq \sum_{k=1}^{n} P(A_k)$

$\forall (A_k)_{k\in N} \in M^N\ P(\bigcup_{k\in N} A_k) \leq \sum_{k=0}^{\infty} P(A_k) \leq \infty$

Une suite croissante d'évènements donne une suite de probabilités croissante majorée donc convergente et on a $P(\bigcup_{k\in N} A_k) = \lim_{n\to\infty} P(A_n) = \sup_{n\in N} P(A_n) \leq 1$

Une suite décroissante d'évènements donne une suite de probabilités décroissante minorée donc convergente et on a $P(\bigcap_{k\in N} A_k) = \lim_{n\to\infty} P(A_n) = \inf_{n\in N} P(A_n) \geq 0$

La mesure normalisée d'une mesure dont la mesure totale est $0 < \mu(\Omega) < \infty$, correspond à la même mesure mais divisée par son poids total : $P = \frac{\mu}{\mu(\Omega)}$. C'est donc toujours une mesure de probabilité.

Dans un espace mesurable, n'importe quel point de l'univers permet de définir la **mesure de Dirac** en ce point : $\boldsymbol{\delta_x(A)} = 1_A(x)$. Cette mesure est toujours une mesure de probabilité.

Etant donnée une famille de points $(x_k)_{1\leq k\leq N}$ dans un espace mesurable, **la loi d'équiprobabilité relativement aux points** $(\boldsymbol{x_k})_{\boldsymbol{k}}$ est définie par $P(\{x_k\}) = \frac{1}{N}$ autrement dit, $P = \frac{1}{N}\sum_{1\leq k\leq n} \delta_{x_k}$
Etant donnée une suite de points sur un espace mesurable, **la mesure de comptage relativement aux points** $(\boldsymbol{x_k})_{\boldsymbol{k}}$ est définie par $\mu = \sum_{k\in N} \delta_{x_k}$. Compte le nombre de points qui sont dans un mesurable.
La loi d'équiprobabilité n'est autre que la mesure de comptage normalisée correspondante. Ainsi la loi d'équiprobabilité est toujours une loi de probabilité, mais ce n'est pas le cas de la mesure de comptage.
Etant donnée une suite de points $(x_k)_k$ dans un espace mesurable, et une suite correspondante (même indexation) de réels positifs $(p_k)_k$ sommable de somme 1, **la mesure de probabilité discrète relativement aux points** $(\boldsymbol{x_k})_{\boldsymbol{k}}$ **de poids** $(\boldsymbol{p_k})_{\boldsymbol{k}}$ est definie par $\forall k\ P(x_k) = p_k$ c'est-à-dire par $P = \sum_{k\in N} p_k \delta_{x_k}$. C'est une mesure de probabilité absolument continue par rapport à la mesure de comptage relativement aux mêmes points.
Sur un espace mesurable dénombrable, lorsqu'on ne précise pas, la **mesure de comptage/ d'équiprobabilité/ discrète** est relative à tout l'espace.
Sur un espace mesurable dénombrable, toute mesure de probabilité est une mesure discrète.
Une mesure de probabilité est **continue** ssi tout singleton de l'univers est un évènement et est de probabilité nulle.
Une **variable aléatoire (v.a.)** correspond à une application mesurable d'un espace probabilisé, vers un espace mesurable quelconque.
Une **variable aléatoire réelle (resp. vectorielle réelle) (v.a.r.)** correspond à une application mesurable d'un espace probabilisé (Ω, M, P) vers $\big(R, B(R)\big)$ (resp. $\Big(R^d, B\big(R^d\big)\Big)$)
Une **variable aléatoire complexe (resp. vectorielle complexe) (v.a.c.)** correspond à une application mesurable d'un espace probabilisé (Ω, M, P) vers $\big(C, B(C)\big)$ (resp. $\Big(C^d, B\big(C^d\big)\Big)$)
La loi de probabilité d'une variable aléatoire $X : (\Omega, M, P) \to (\Omega', M')$ est la mesure image de P par X, ainsi $\boldsymbol{P_X = L(X)} : M' \to [0,1] : B \mapsto P_X(B) = \boldsymbol{P(X \in B)} = P\big(X^{-1}(B)\big) = P(\{\omega \in \Omega \mid X(\omega) \in B\})$ est une mesure de probabilité sur (Ω', M').
En pratique on explicite pas souvent (Ω, M, P) ni même X, les hypothèses et les calculs sont généralement formulées et menés sur les lois de probabilités des variables aléatoires en jeu.

III.2. Fonctions de répartition

La fonction de répartition (c.d.f.) d'une variable aléatoire réelle $\boldsymbol{X}$ est la fonction définie par
$\boldsymbol{F_X} : R \to R : x \mapsto F_X(x) = P_X\big((-\infty, x]\big) = P(X \leq x)$
i) Une fonction de répartition est croissante
ii) Une fonction de répartition est continue à droite et admet une limite à gauche en tout point (càdlàg)
iii) Une fonction de répartition est à valeurs dans $[0,1]$, tend vers 0 en $-\infty$ et tend vers 1 en $+\infty$.
Une fonction de $R \to R$ vérifiant i)+ii)+iii) est une fonction de répartition.
Une fonction de répartition admet au plus un nombre dénombrable de points de discontinuités car croissante.
$\forall x \in R\ F_X(x^+) = F_X(x) = P(X \leq x),\ \ F_X(x^-) = P(X < x),\ \ \ P(X = x) = F_X(x) - F_X(x^-)$
Pour une variable aléatoire réelle, la fonction de répartition caractérise la loi de probabilité : $F_X = F_Y \Leftrightarrow P_X = P_Y$. Une propriété à propos de la loi est donc par abus de langage parfois énoncée directement sur la fonction de répartition.

Par exemple, une fonction de répartition est dite **discrète** ssi P_X l'est.
La fonction de répartition d'une loi discrète finie est en escalier, et les discontinuités sont situées sur les éléments du support de la loi.
Une fonction de répartition est **continue** ssi P_X continue ssi $\forall x \in R\ P(X = x) = 0 = F_X(x) - F_X(x^-)$ ssi la fonction de répartition est bien une fonction continue ce qui justifie la confusion de vocabulaire.
Une fonction de répartition est **absolument continue/admet une densité** ssi P_X est absolument continue/admet une densité (par rapport à la mesure de Borel) ssi $\exists f_X : R \to R_+$ intégrable telle que $\forall B \in B(R)\ \ P(X \in B) = \int_B \boldsymbol{f_X} d\lambda$. Dans ce cas $\forall x \in R\ \ F_X(x) = P(X \leq x) = \int_{-\infty}^{x} f(t)dt$
Dans ce cas, une telle fonction $\boldsymbol{f_X}$ est une **densité (p.d.f.)** d'une fonction de répartition/loi de proba.
Une fonction de $R \to R$ est une densité ssi elle est positive, mesurable, λ-integrable sur R d'integrale 1, et définit donc dans ce cas une fonction de répartition, et une loi de probabilité.
Pour une variable aléatoire réelle, la fonction de répartition, caractérise la loi : $F_X = F_Y \Leftrightarrow P_X = P_Y$.
Pour une variable aléatoire réelle de loi absolument continue, la fonction de répartition, et la densité caractérisent la loi de probabilité : $f_X = f_Y \Leftrightarrow F_X = F_Y \Leftrightarrow P_X = P_Y$. On peut donc énoncer les propriétés suivant l'un des 3 points de vue indistinctement.
Une fonction de répartition peut toujours se décomposer sous la forme d'une combinaison linéaire réelle de 3 fonctions de répartitions, une discrète, une absolument continue, et une singulière (continue et non absolument continue).
Un quantile d'ordre $\boldsymbol{\alpha} \in]\mathbf{0},\mathbf{1}[$ d'une v.a.r. $\boldsymbol{X}$ est un reel x tel que $P(X \leq x) \geq \alpha$ et $P(X \geq x) \geq 1 - \alpha$ autrement dit tel que $F_X(x^-) = P(X < x) \leq \alpha \leq P(X \leq x) = F_X(x)$
En un point de continuité d'une v.a.r. X, un quantile d'ordre $\alpha \in]0,1[$ est un réel d'image α par la fonction de répartition.
Pour $n \geq 2$, **un $\boldsymbol{n}$-ile d'une v.a.r. $\boldsymbol{X}$** est un quantile de X d'ordre $\frac{k}{n}$ avec $1 \leq k \leq n-1$
Une **médiane** est un 2-ile et il y en a une seule, un **quartile** est un 4-ile et il y en a 3, un **décile** est un 10-ile et il y en a 9.
La **fonction quantile d'une v.a.r. $\boldsymbol{X}$** est la fonction definie par $\boldsymbol{F}^{\leftarrow} :]0,1[\to R : u \mapsto \inf\{x \in R \mid F(x) > u\}$
Pour une fonction de répartition/loi de proba donnée, la fonction quantile correspondante appliquée à une loi uniforme standard, donne une v.a.r. $F^{\leftarrow}(U)$ de même fonction de répartition/loi de proba.
Une fonction quantile est càdlàg sur $]0,1[$. Autres propriétés TODO
Pour $1 \leq i_1 < \cdots < i_k \leq d$, **le $(\boldsymbol{i_1}, \ldots, \boldsymbol{i_k})$-ieme vecteur aléatoire marginal d'un vecteur aléatoire $\boldsymbol{X} = (X_1, \ldots, X_d)$** sur R^d, est le vecteur $(X_{i_1}, \ldots, X_{i_k})$. Pour une notion quelconque dépendant d'un vecteur aléatoire, on définit par analogie la $(i_1, \ldots, i_k)$-ieme notion marginale comme étant celle associée au $(i_1, \ldots, i_k)$-ieme vecteur aléatoire marginal.
La fonction de répartition d'un vecteur aléatoire réel $X = (X_1, \ldots, X_d)$ est définie par $\boldsymbol{F_X}(\boldsymbol{x} = (\boldsymbol{x_1}, \ldots, \boldsymbol{x_d})) = P(X_1 \leq x_1, \ldots, X_d \leq x_d) = P(\bigcap_{k=1}^{n} X_k \leq x_k)$
La $(i_1, \ldots, i_k)$-ieme fonction de répartition marginale d'un vecteur aléatoire réel s'obtient en faisant tendre vers l'infini tous les x_i d'indices pas dans les i_j, et donc coïncide avec la fonction de répartition du vecteur extrait
$F_{(X_{i_1}, \ldots, X_{i_k})}(x_{i_1}, \ldots, x_{i_k}) = P(X_{i_1} \leq x_{i_1}, \ldots, X_{i_k} \leq x_{i_k}) = \lim_{1 \leq j \leq d, j \notin \{x_{i_1}, \ldots, x_{i_k}\}, x_j \to \infty} F_X(x_1, \ldots, x_d)$
Propriétés de la fonction de répartition vectorielle : TODO

Une fonction de répartition vectorielle est **absolument continue/admet une densité** ssi P_X est absolument continue/admet une densité (par rapport à la mesure de Borel) ssi $\exists f_X : R^n \to R_+$ intégrable telle que $\forall B \in B(R^d)\ \ P(X \in B) = \int_B f d\lambda$. Dans ce cas $\forall x \in R\ \ F_X(x) = P(X_1 \leq x_1, \dots, X_d \leq x_d) = \int_{-\infty}^{x_1} \dots \int_{-\infty}^{x_d} \boldsymbol{f_X}(\boldsymbol{u_1}, \dots, \boldsymbol{u_d}) du_1 \dots du_d$

Dans ce cas, une telle fonction $\boldsymbol{f_X}$ est une **densité de probabilité** de la loi P_X

Une fonction de $R^n \to R$ est une densité de probabilité ssi elle est positive, mesurable, intégrable de $(R^n, B(R^n)) \to (R, B(R))$ d'intégrale 1. Elle définit donc dans ce cas une fonction de répartition, et une loi de probabilité.

La $(i_1, \dots, i_k)$-ieme fonction de densité marginale d'un vecteur aléatoire réel s'obtient en intégrant sur R, tous les x_i d'indices pas dans les i_j, et coïncide avec la fonction de densité du vecteur extrait : Pour $I = \{i_1, \dots, i_k\}, J = \{1, \dots, d\} \setminus I$, $f_{(X_{i_1}, \dots, X_{i_k})}(x_{i_1}, \dots, x_{i_k}) = \int_{R^{n-d}} f_X(x) du_{j_1} \dots du_{j_{d-k}}$

Propriétés de la densité de probabilité vectorielle : TODO

Pour un vecteur aléatoire réelle, la fonction de répartition, caractérise la loi : $F_X = F_Y \Leftrightarrow P_X = P_Y$.

Pour un vecteur aléatoire réelle de loi absolument continue, la fonction de répartition, et la densité caractérisent la loi de probabilité : $f_X = f_Y \Leftrightarrow F_X = F_Y \Leftrightarrow P_X = P_Y$. On peut donc énoncer les propriétés suivant l'un des 3 points de vue indistinctement.

Pour une loi/cdf/pdf vectorielle absolument continue, $\frac{\partial^d F_X}{\partial x_1 \dots \partial x_d}$ existe et $\frac{\partial^d F_X}{\partial x_1 \dots \partial x_d} = f_X$ presque partout.

Pour une loi/cdf/pdf absolument continue d'une var, F_X dérivable et $F_X' = f_X$ presque partout sur R.

Changement de variables.

Soit U, V ouverts de R^n, $\varphi : U \to V$ un C^1 diffeomorphisme, X, Y deux vecteurs aléatoires réels tels que $Y = \varphi(X)\ /\ X = \varphi^{-1}(Y)$, on écrit $y = y(x) = \varphi(x),\ x = x(y) = \varphi^{-1}(y)$. Alors X est absolument continue ssi Y l'est et dans ce cas : soit $f_Y : R^n \to R$ densité de Y, $f_X : R^n \to R$ densité de X. On note $\left|\frac{dy}{dx}\right| = \left|\det J_\varphi\right|_x$

$\int_V f_Y(y) dy = \int_U f_X(x)\, dx = \int_V f_X(x) \left|\frac{dx}{dy}\right| dy$

$\forall y \in R^n\ \ f_Y(y) = f_X\left(\varphi^{-1}(y)\right) \left|\frac{dx}{dy}\right|_{x=\varphi^{-1}(y)} 1_V(y)$

Esperance.

Une v.a.r. **admet une espérance finie** ssi elle est intégrable par rapport à la loi de probabilité de son espace de départ $\int_\Omega |X|\, dP < \infty$. Autrement dit une v.a.r. admet une espérance finie ssi elle est L^1 par rapport à son espace probabilisé de départ. Dans ce cas **l'espérance d'une v.a.r.** est l'intégrale de cette v.a.r. par rapport à la loi de probabilité de son espace de départ. $\boldsymbol{E(X)} = \int_\Omega X dP$ et est toujours finie.

L'espérance d'une v.a.r. positive est l'intégrale de cette var par rapport à la loi de probabilité de son espace de départ. $\boldsymbol{E(X)} = \int_\Omega X dP$, elle est toujours définie, mais peut valoir ∞.

L'indicatrice d'une partie A d'un espace probabilisé est toujours une v.a.r. L^1 et $E(1_A) = P(A)$

Pour X, Y v.a.r. L^1 ou ≥ 0, $E(X + Y) = E(X) + E(Y)$. L'espérance est linéaire

Pour X v.a.r. L^1 ou ≥ 0, $\forall a, b \in R\ \ E(aX + b) = aE(X) + b$

Pour X, Y v.a.r. L^1 ou ≥ 0, $X \leq Y \Rightarrow E(X) \leq E(Y)$. L'espérance est croissante.

Lemme de transport. Pour X v.a.r. L^1 ou ≥ 0, $E(X) = \int_\Omega X dP = \int_R x dP_X$. Si de plus X est absolument

continue par rapport à la mesure de Borel avec densité f_X, on a $E(X) = \int_R x\, f_X(x) d\lambda$, resp. par rapport à une mesure discrète on a $E(X) = \sum_{k\in I} x_k\, P(X = x_k)$

Pour X v.a.r. L^1 ou ≥ 0 et $h : \big(R, B(R)\big) \to \big(R, B(R)\big)$ mesurable telle que $h(X)$ v.a.r. L^1 ou ≥ 0, alors $E\big(h(X)\big) = \int_\Omega h(X) dP = \int_R h(x)\, dP_X$. Si de plus X est absolument continue par rapport à la mesure de Borel, on a $E\big(h(X)\big) = \int_{-\infty}^{\infty} h(x)\, f_X(x) dx$ resp. par rapport à une mesure discrète on a $E\big(h(X)\big) = \sum_{k\in I} h(x_k)\, P(X = x_k)$

Pour une v.a.r. L^p on peut écrire $E(|X|^p) < \infty$. $\|X\|_{L^p} = \big(E(|X|^p)\big)^{\frac{1}{p}}$ est une norme sur L^p

Pour une v.a.r. L^∞ $\|X\|_{L^\infty} = \inf\{M \mid |X| \leq M \;\; P.\text{presque surement}\}$

Pour $1 \leq p < q$ on a $\|X\|_{L^p} \leq \|X\|_{L^q}$ et donc $L^q \subset L^p$

Jensen. Pour une fonction φ convexe de R dans R, et X une v.a.r. L^1 et telle que $\varphi(X)$ est L^1, on a $\varphi\big(E(X)\big) \leq E\big(\varphi(X)\big)$

Hölder. Pour $p, q \in [1, \infty]$ conjugués, on a $\|XY\|_{L^1} \leq \|X\|_{L^p} \|Y\|_{L^q}$ càd $E(|XY|) \leq E(|X|^p)^{\frac{1}{p}} E(|X|^q)^{\frac{1}{q}}$

Inégalité générale de Chebyshev. Pour une fonction quelconque φ de R dans R, A un evenement, et X une v.a.r. L^1 et telle que $\varphi(X)$ est L^1 on a $(\inf_A \varphi)\, P(X \in A) \leq E(\varphi(X) 1_A) \leq E\big(\varphi(X)\big)$

Pour $X \in L^1, \varphi(x) = |x|, A = \{|x| \geq a\}$, on obtient $aP(|X| \geq a) \leq E(|X|)$ Markov

Pour $X \in L^2, \varphi(x) = x^2, A = \{|x| \geq a\}$, on obtient $a^2 P(|X| \geq a) \leq E(X^2)$

Pour $X \in L^p, \varphi(x) = |x|^p, A = \{|x| \geq a \geq 0\}$, on obtient $a^p P(|X| \geq a) \leq E(|X|^p)$

On obtient $a^2 P(|X - E(X)| \geq a) \leq Var(X)$ Chebyshev.

Convergence monotone. Toute suite <u>croissante</u> de v.a.r. positives dans $[0, +\infty]$ converge simplement vers une v.a.r. dans $[0, +\infty]$ (la fonction supremum), et définit une suite d'espérances croissantes qui admet une limite dans [0,+∞], cette limite est égale à l'espérance de la fonction et est donc indépendante de la suite choisie. $\forall n, 0 \leq X_n \leq X_{n+1} \Rightarrow E(X_n) \to E(\sup_n X_n) = \sup_n E(X_n) = E(\lim_n X_n) = \lim_n E(X_n)$. Plus brièvement $0 \leq X_n \uparrow X \Rightarrow EX_n \uparrow EX$

Lemme de Fatou. Pour une suite de v.a.r. positives dans $[0, +\infty]$, $E(\liminf_n X_n) \leq \liminf_n E(X_n)$

Théorème de convergence dominée. Si X_n converge simplement vers X presque partout, $\forall n\ |X_n| \leq Y$ presque partout avec $E(Y) < \infty$ alors $X \in L^1$ et $E(X_n) \to_{n\to\infty} E(X)$. $E(\lim_n X_n) = \lim_n E(X_n)$.

La **variable centrée** d'une v.a.r. X L^1 est la variable $\boldsymbol{X_c} = X - EX$

Le **moment d'ordre $\boldsymbol{p}$** d'une v.a.r. $X \in L^p$ est $E(X^p)$

Le **moment absolu d'ordre $\boldsymbol{p}$** d'une v.a.r. $X \in L^p$ est $E(|X|^p) = \|X\|_{L^p}^p$

Le **moment centré d'ordre $\boldsymbol{p}$** d'une v.a.r. $X \in L^p$ est $E(|X_c|^p)$

La **covariance entre deux v.a.r. $\boldsymbol{X, Y \in L^1}$ de produit dans $\boldsymbol{L^1}$** est $\boldsymbol{cov(X, Y)} = E(X_c Y_c) = E(XY) - E(X)E(Y)$

Inégalité Cauchy Schwarz. Hölder (p=2) $|E(XY)| \leq E(|XY|) \leq \sqrt{E(X^2)E(Y^2)}$

$(X, Y) \mapsto E(X, Y)$ est un produit scalaire sur L^2 de norme associee la norme L^2, et L^2 muni de ce produit scalaire est un espace de Hilbert.

La **covariance entre deux v.a.r. $\boldsymbol{X, Y \in L^2}$** est toujours bien définie $\boldsymbol{cov(X, Y)} = E(X_c Y_c)$

La covariance est une forme bilinéaire symétrique positive sur les $X, Y \in L^1 \times L^1$ de produit L^1 , et sur $L^2 \times L^2$ mais ce n'est pas un produit scalaire.

La **variance d'une v.a.r. $\boldsymbol{X \in L^2}$** est $\boldsymbol{Var(X)} = E(X_c^2) = E(X^2) - (EX)^2 = \inf_{a\in R} E((X - a)^2)$

Une v.a.r. L^2 est constante ssi sa variance est nulle.
Pour une v.a.r. $X \in L^2$, et $a, b \in R$, on a $Var(aX + b) = a^2 Var(X)$
L'**écart type d'une v.a.r.** $\boldsymbol{X} \in \boldsymbol{L^2}$ est $\boldsymbol{\sigma(X)} = \sqrt{Var(X)}$
Pour une famille de v.a.r. dans L^2 on a $Var\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} Var(X_i) + 2\sum_{1 \le i < j \le n} cov(X_i, X_j)$
Le **coefficient de corrélation linéaire entre deux v.a.r.** $\boldsymbol{X, Y} \in \boldsymbol{L^2}$ est défini par $\boldsymbol{\rho_{X,Y}} = \frac{cov(X,Y)}{\sqrt{var(X)var(Y)}}$. Il est à valeurs dans $[-1,1]$.
Un vecteur aléatoire réel est L^p ssi toutes ses composantes le sont.
L'espérance d'un vecteur aléatoire réel $\vec{X} \in L^1$ est definie par $\boldsymbol{E}(\vec{\boldsymbol{X}}) = \left(E(X_1), \dots, E(X_d)\right) \in R^d$
L'espérance d'une matrice aléatoire réelle $M \in L^1 \cap M_{d,n}$ est la matrice $\boldsymbol{E}(\boldsymbol{M})$ des esperances des composantes.
Pour $M, N \in L^1 \cap M_{d,n}$ $E(M + N) = E(M) + E(N)$
Pour $A \in M_{m,n}, M \in L^1 \cap M_{d,n}, B \in M_{n,k}$ $E(AMB) = AE(M)B$
La **matrice de covariance croisée** de 2 $\vec{v}$.a.r. $\vec{X}, \vec{Y} \in L^2$ est définie par

$$\boldsymbol{cov}(\vec{\boldsymbol{X}}, \vec{\boldsymbol{Y}}) = \left[cov(X_i, Y_j)\right]_{1 \le i,j \le d} = E\left(\left(\vec{X} - E(\vec{X})\right)\left(\vec{Y} - E(\vec{Y})\right)^T\right) = E\left(\overrightarrow{X_c}\overrightarrow{Y_c}^T\right) = E(\vec{X}\vec{Y}^T) - E(\vec{X})E(\vec{Y})^T$$

La **matrice de covariance** d'un $\vec{v}$.a.r. $\vec{X} \in L^2$ est définie par

$$\boldsymbol{var}(\vec{\boldsymbol{X}}) = cov(\vec{X}, \vec{X}) = \left[cov(X_i, X_j)\right]_{1 \le i,j \le d} = E\left(\left(\vec{X} - E(\vec{X})\right)\left(\vec{X} - E(\vec{X})\right)^T\right) = E\left(\overrightarrow{X_c}\overrightarrow{X_c}^T\right) = E(\vec{X}\vec{X}^T) - E(\vec{X})E(\vec{X})^T$$

La **matrice de corrélation** d'un $\vec{v}$.a.r. $\vec{X} \in L^2$ est définie par

$$\boldsymbol{corr}(\vec{\boldsymbol{X}}) = \left[\frac{cov(X_i,X_j)}{\sigma(X_i)\sigma(X_j)}\right]_{1 \le i,j \le d} = diag\left(var(\vec{X})\right)^{-\frac{1}{2}} var(\vec{X}) diag\left(var(\vec{X})\right)^{-\frac{1}{2}}$$

Pour $\vec{X}, \vec{Y} \in L^2$ $cov(\vec{X}, \vec{Y}) = E(\vec{X}\vec{Y}^T) - E(\vec{X})E(\vec{Y})^T$
Pour $\vec{X}, \vec{Y} \in L^2$ $cov(\vec{X}, \vec{Y}) = cov(\vec{Y}, \vec{X})^T$
Pour $\overrightarrow{X_1}, \overrightarrow{X_2}, \vec{Y} \in L^2$ $cov(\overrightarrow{X_1} + \overrightarrow{X_2}, \vec{Y}) = cov(\overrightarrow{X_1}, \vec{Y}) + cov(\overrightarrow{X_2}, \vec{Y})$
Pour $A \in M_{n,d}, B \in M_{d,m}, \vec{a}, \vec{b} \in R^d, \vec{X}, \vec{Y} \in L^2$, $cov(A\vec{X} + \vec{a}, B^T\vec{Y} + \vec{b}) = Acov(\vec{X}, \vec{Y})B$
Pour $\vec{X}, \vec{Y} \in L^2$ indépendants (ou plus faiblement si $\forall i \forall j\ cov(X_i, X_j) = 0$) alors $cov(\vec{X}, \vec{Y}) = [0]$
La matrice de covariance est symetrique positive
Pour une matrice A, un vecteur $\vec{a}$ et $\vec{X} \in L^2$, on a $var(A\vec{X} + \vec{a}) = Avar(\vec{X})A^T$
La matrice de covariance est un cas particulier de la matrice de covariance croisée et vérifie donc les même propriétés. La matrice de corrélation est aussi très similaire et vérifie des propriétés analogues. Il y a beaucoup d'autres propriétés (voir wiki).
La fonction caractéristique d'un $\vec{\boldsymbol{v}}$**.a.r.** $\vec{\boldsymbol{X}}$ est la fonction $\varphi_X : R^d \to C$ definie par $\boldsymbol{\varphi_{\vec{X}}}(\vec{x}) = E\left(e^{i(\vec{x}|\vec{X})}\right)$.
Pour un vecteur aléatoire réel, la fonction caractéristique, caractérise la loi : $\varphi_{\vec{X}} = \varphi_{\vec{Y}} \Leftrightarrow P_{\vec{X}} = P_{\vec{Y}}$.
La fonction caractéristique d'un $\vec{v}$.a.r. $\vec{X}$ absolument continu est en fait la transformée de Fourier de sa densité $\boldsymbol{\varphi_{\vec{X}}}(\vec{x}) = \int_{R^d} e^{i(\vec{x}|\vec{u})} f(\vec{u}) d\vec{u} = (Ff)(\vec{x})$

un $\vec{v}$.a.r. $\vec{X}$ dont la fonction caractéristique est $L^1(R^d, \lambda)$ est absolument continu de densité s'obtenant par la formule d'inversion de Fourier $f_{\vec{X}}(\vec{u}) = K \int_{R^d} e^{-i(\vec{x}|\vec{u})} \varphi_{\vec{X}}(\vec{x}) d\vec{x}$

Pour un $\vec{v}$.a.r. $\vec{X}$ $\|\varphi_{\vec{X}}\|_u = \varphi_{\vec{X}}(0) = 1$

Pour un $\vec{v}$.a.r. $\vec{X}, j \in \{1, \dots, d\}$ et $x_j \in R$ on a $\varphi_{X_j}(x_j) = \varphi_{\vec{X}}(0, \dots, 0, x_j, 0, \dots, 0)$

Pour un $\vec{v}$.a.r. $\vec{X}$ $\forall \vec{x} \in R^d$ $\overline{\varphi_{\vec{X}}(\vec{x})} = \varphi_{\vec{X}}(-\vec{x})$

Pour un $\vec{v}$.a.r. $\vec{X}$, une matrice $A \in M_{d,n}$, $\vec{b} \in R^d$, $\forall \vec{x} \in R^d$ $\varphi_{A\vec{X}+\vec{b}}(\vec{x}) = e^{i(\vec{x},\vec{b})} \varphi_{\vec{X}}(A^T \vec{x})$

Deux v.a.r. X, Y indépendantes, vérifient $\varphi_{X+Y} = \varphi_X \varphi_Y$.

La fonction caractéristique d'un vecteur aléatoire réel est uniformément continue sur R^d

Pour une v.a.r. X, $\varphi_X(t) \to_{t \to \pm\infty} 0$ (Riemann Lebesgue).

Pour une v.a.r. $X \in L^p$ avec $p \in N^*$, φ_X est p-fois dérivable et $\forall k \in \{1, \dots p\}$ $\varphi_X^{(k)}(0) = i^k E(X^k)$, et $\forall k \in \{1, \dots p\}$ $\varphi_X^{(k)}(x) = i^k E(X^k e^{ixX})$

Réciproquement si φ_X est p-fois dérivable avec p un entier pair ≥ 2 alors $X \in L^p$ autrement dit X admet tout moment d'ordre $\leq p$.

La loi d'une v.a.r. n'est en général pas caractérisée par ces moments. Toutefois elle l'est si la fonction caractéristique associée est analytique. Une condition simple d'analycité est $\forall \alpha \in R_+^*$ $E(e^{\alpha|X|}) < \infty$.

Théorème des moments. Deux v.a.r <u>à valeurs dans un intervalle borné</u> qui admettent des moments finis et égaux à tout ordre, ont même loi de probabilité.

La transformée de Laplace/fonction génératrice des moment d'un $\vec{v}$.a.r. $\vec{X}$ est la fonction $\boldsymbol{L_{\vec{X}}}: R^d \to C$ definie par $\boldsymbol{L_{\vec{X}}}(\vec{s}) = E\left(e^{(\vec{s}|\vec{X})}\right)$ <u>uniquement en les valeurs de $\vec{s}$ pour lesquelles $e^{(\vec{s}|\vec{X})}$ est L^1</u>

Une v.a.r. X telle que $e^{tX} \in L^1$ pour tout t dans un intervalle ouvert contenant 0, admet une transformée de Laplace bien définie et analytique sur un intervalle ouvert contenant 0 ou l'on peut écrire $L_X(t) = \sum_{n=0}^{\infty} \frac{t^n}{n!} E(X^n)$, en particulier pour tout $n \in N$, $L_X^{(n)}(0) = E(X^n)$

Deux $\vec{v}$.a.r. dont la transformée de Laplace est définie et finie dans un intervalle ouvert contenant 0 caractérise la loi de probabilité

Fonctions génératrices.

La fonction génératrice d'une v.a.r. $\boldsymbol{X}$ discrète à valeurs dans $\mathbb{N}$ est $\boldsymbol{G_X}: D \subseteq \mathbb{R} \to [0, \infty]: t \mapsto E(t^X) = \sum_{n=0}^{\infty} P(X = k) t^k$. Le rayon de cette série entière est ≥ 1.

2 v.a.r. discrètes vers $\mathbb{N}$, dont les fonctions génératrices coincident au voisinage de 0, ont même loi par unicité du d.s.e.

Pour une v.a.r. discrète vers $\mathbb{N}$ admettant une espérance, $G_X(1) < \infty$ et G_X dérivable en 1 et $G_X'(1) = E(X)$

Pour une v.a.r. discrète vers $\mathbb{N}$ admettant une variance (et donc une espérance), alors G_X, G_X', G_X'' sont définies en 1 et $Var(X) = G_X''(1) + G_X'(1) - G_X'(1)^2$

Deux v.a.r. discrètes vers $\mathbb{N}$ admettent la même fonction génératrice ssi elles ont même loi de proba.

La fonction génératrice caractérise la loi d'une v.a.r. discrète vers $\mathbb{N}$

Deux v.a.r. discrètes vers $\mathbb{N}$ indépendantes vérifient $G_{X+Y} = G_X G_Y$ réciproque fausse.

Composition de fonctions génératrices. Pour $(X_n)_n$ suite de v.a.r. discrètes vers $\mathbb{N}$ de même loi X, et N v.a.r. discrète vers $\mathbb{N}$, et $(N, X_1, X_2, \dots)$ indépendantes alors $G_{\sum_{n=1}^{N} X_n} = G_N \circ G_X$

IV. Indépendance. Rappel sur les tribus produits.

Un **rectangle élémentaire mesurable** sur une famille/produit fini ou dénombrables d'espaces mesurables est un produit cartésien (de même indexation que la famille) de parties mesurables de ces espaces.

La **tribu produit d'une famille finie/dénombrable d'espaces mesurables** $\otimes_{i=1}^{n} \boldsymbol{T_i}$ est la tribu engendrée par les rectangles élémentaires mesurables sur cette famille d'espace mesurables.

La tribu produit d'une famille finie d'espaces mesurables est la plus petite tribu sur le produit des espaces qui rende chaque projection mesurable.

Une fonction d'un espace mesurable vers un produit fini/dénombrable d'espaces mesurable muni de la tribu produit, est mesurable ssi chacune de ses composante l'est vis-à-vis de son propre espace.

La tribu produit fini des tribus boréliennes d'espaces topologiques est inclus dans la tribu borélienne de la topologie produit fini. Il y a égalité si les espaces topologiques sont engendres par des bases dénombrables.

IV.1. Indépendance

Dans un espace probabilisé, **deux évènements sont indépendants** et on note $\boldsymbol{A} \perp\!\!\!\perp \boldsymbol{B}$ ssi $P(A \cap B) = P(A)P(B)$

Pour deux évènements d'un espace probabilisé, $A \perp\!\!\!\perp B \Leftrightarrow \overline{A} \perp\!\!\!\perp B \Leftrightarrow A \perp\!\!\!\perp \overline{B} \Leftrightarrow \overline{A} \perp\!\!\!\perp \overline{B}$

Dans un espace probabilisé, une **famille quelconque d'évènements est mutuellement indépendante** et on note $\perp\!\!\!\perp_{\boldsymbol{i \in I}} \boldsymbol{A_i}$ ssi pour toute sous-famille <u>finie</u>, la probabilité de l'intersection est le produit des probabilités. $\forall J$ fini $\subseteq I, P\left(\bigcap_{j \in J} A_j\right) = \prod_{j \in J} P\left(A_j\right)$

Une famille d'évènements mutuellement indépendante est une famille d'évènements indépendants 2 a 2, mais la réciproque est fausse.

Dans un espace probabilisé, une **famille quelconque de sous-tribus (resp. algèbres) est mutuellement indépendante** et on note $\perp\!\!\!\perp_{\boldsymbol{i \in I}} \boldsymbol{M_i}$ ssi toute famille d'évènements sur cette famille de sous-tribus (resp. algèbres) est mutuellement indépendante : $\forall (A_i \in M_i)_{i \in I} \ \perp\!\!\!\perp_{i \in I} A_i$

Les tribus engendrées respectives de deux algèbres indépendantes sont indépendantes.

Une famille quelconque de variables aléatoires $(X_i)_{i \in I}$ partant d'un même espace probabilisé forment une **famille (mutuellement) indépendante de variables aléatoires** et on note $\perp\!\!\!\perp_{\boldsymbol{i \in I}} \boldsymbol{X_i}$ ssi la famille des sous-tribus engendrées par les variables est mutuellement indépendante : $\perp\!\!\!\perp_{i \in I} \sigma(X_i)$ ssi $\forall J$ fini $\subseteq I$, $\forall \left(B_j \in M'\right)_{j \in J} P\left(\bigcap_{j \in J} \{X_j \in B_j\}\right) = \prod_{j \in J} P\left(X_j \in B_j\right)$ ssi $\forall (B_i \in M')_{i \in I} \perp\!\!\!\perp_{i \in I} \{X_i \in B_i\}$

Une famille quelconque d'évènements est mutuellement indépendante $\perp\!\!\!\perp_{i \in I} A_i$ ssi $\perp\!\!\!\perp_{i \in I} 1_{A_i}$ ssi $\perp\!\!\!\perp_{i \in I} \sigma(A_i)$

Une famille quelconque de sous-tribus est mutuellement indépendante $\perp\!\!\!\perp_{i \in I} M_i$ ssi $\forall (A_i \in M_i)_{i \in I} \ \perp\!\!\!\perp_{i \in I} A_i$ ssi $\forall (A_i \in M_i)_{i \in I} \ \perp\!\!\!\perp_{i \in I} 1_{A_i}$ ssi $\forall \left(X_i \text{ v. a. r. definie sur } (\Omega, M_i, P)\right)_{i \in I} \ \perp\!\!\!\perp_{i \in I} X_i$

Une famille de v.a. mutuellement indépendante est une famille de v.a. indépendantes 2 a 2, mais la réciproque est fausse. Exemple : $Z = XY$ avec X, Y v.a. de Bernoulli. $\{X, Y, Z\}$

Pour une famille quelconque de sous-tribus mutuellement indépendante $\perp\!\!\!\perp_{i \in I} M_i$, on peut partitionner la famille $I = \cup_{j \in J} J_j, J_j \cap J_k = \emptyset$ et pour chaque groupe de sous-tribus former la tribu engendrée par l'union, les nouvelles sous-tribus obtenues sont encore mutuellement indépendante : $\perp\!\!\!\perp_{j \in J} \sigma\left(\cup_{i \in J_j} M_i\right)$

Pour $(\varphi_i : (E_i, B_i) \to (E_i', B_i') \text{ mesurable})_{1 \leq i \leq n}$ et $(X_i : (\Omega, M, P) \to (E_i, B_i) \ v.a.)_{1 \leq i \leq n}$ alors $\perp\!\!\!\perp_{i=1}^{d} X_i \Rightarrow$

$\perp\!\!\!\perp_{i=1}^{d} \varphi_i(X_i)$

Une famille quelconque de variables aléatoires $(X_i)_{i\in I}$ partant d'un même espace probabilisé est mutuellement indépendante ssi $\forall J$ fini $\subseteq I$, $\forall\left(\phi_j \text{ fonction borelienne} \mid \phi_j(X_j) \in L^1\right)_{j\in J}$

$E\left(\prod_{j\in J}\phi_j(X_j)\right) = \prod_{j\in J} E\left(\phi_j(X_j)\right)$

Dans ce cas on a en particulier pour des v.a.r. indépendantes : $\perp\!\!\!\perp_{i=1}^{d} X_i \Rightarrow E(X_1 \dots X_d) = E(X_1) \dots E(X_d)$

Un produit fini d'espaces probabilisés est encore un espace mesurable que l'on peut munir de la mesure de probabilité produit $P = \otimes_{i=1}^{n} P_i$ grâce au théorème de prolongement des prémesures.

La mesure de probabilité produit est unique pour un produit fini d'espaces probabilisés σ-finis.

La loi de probabilité d'un $\vec{v}$.a. est égale au produit des probabilités marginales : $P_{\vec{X}} = P_{X_1} \otimes \dots \otimes P_{X_d}$ ssi ses v.a. composantes sont mutuellement indépendantes : $\perp\!\!\!\perp_{i\in I} X_i$

La fonction caractéristique d'un $\vec{v}$.a.r. est égale au produit des fonctions caractéristiques marginales : $\forall(t_1, \dots, t_d) \in R^d\ \varphi_{\vec{X}}(\vec{t}) = \varphi_{X_1}(t_1) \dots \varphi_{X_d}(t_d)$ ssi ses v.a.r. composantes sont mutuellement indépendantes : $\perp\!\!\!\perp_{i\in I} X_i$

Deux v.a.r. admettant une covariance sont **non corrélés** ssi $cov(X,Y) = 0$ ssi $E(XY) = E(X)E(Y)$

Deux $\vec{v}$.a.r. L^2 sont **non corrélés** ssi ils donnent une covariance nulle ssi $cov(\vec{X},\vec{Y}) = 0$ ssi $E(\vec{X}\vec{Y}^T) = E(\vec{X})E(\vec{Y})^T$. Deux $\vec{v}$.a.r. indépendants sont toujours non corrélés, mais la réciproque est fausse : $X \sim N(0,1)$, et $Y = X^2$

Des v.a.r. 2 a 2 non corrélées vérifient $var\left(\sum_{i=1}^{d} X_i\right) = \sum_{i=1}^{d} var(X_i)$

Exemple : Toute combinaison linéaire de d points fixes de R^d a coefficients dans $[0,1]$ peut etre approximée a $\frac{\sqrt{d}}{2}$ près par une combinaison linéaire a coefficient dans $\{0,1\}$.

Exemple : Pour une suite $(a_n)_n \in R^N$ qui tend vers l'infini, alors $\frac{1}{n} card\{k \le n \mid |v(k) - \ln\ln n| > a_n\sqrt{\ln\ln n}\} \to_{n\to\infty} 0$, avec $v(k)$ le nombre de diviseurs premiers de n.

IV.2. Sommes de variables aléatoires indépendantes

On dit qu'une famille de v.a. est **iid** ssi elle est mutuellement indépendante et identiquement distribuée.

Une famille finie de d v.a. L^2 iid, vérifie $\forall t \in R_+^*\ P\left(\left|\sum_{i=1}^{d}(X_i - E(X_i))\right| \ge t\sqrt{d}\right) \le \frac{Var(X_1)}{t^2}$, ainsi l'ordre de grandeur de la somme est au plus $\sqrt{d}$ et elle ressemble donc a un terme déterministe $nE(X_1)$ de l'ordre de d plus un terme aleatoire d'ordre au plus $\sqrt{d}$

La somme de deux v.a.r. indépendantes sur un même espace probabilisé est une v.a.r sur ce même espace de loi de probabilité donnée par le **produit de convolution des lois des deux var $\boldsymbol{P_X \star P_Y}$** défini par $\forall\phi$ fonction borelienne bornee $\int_R \phi\ d(P_X \star P_Y) = \int_R \int_R \phi(x+y) dP_Y(y)\, dP_X(x) = \int_R \int_R \phi(x+y)\, dP_X(x)\, dP_Y(y)$

Le produit de convolution de lois vérifie la commutativité, l'associativité, la distributivité, et admet le Dirac de centre 0 comme élément neutre.

La somme de deux v.a.r. indépendantes sur un même espace probabilisé admet pour fonction caractéristique $\varphi_{X+Y} = \varphi_X \varphi_Y$. Ici c'est bien un produit car dimension 1 et $\varphi_{X+Y} : R \to C$

IV.3. Applications de l'indépendance

Soit une suite dénombrable d'espaces probabilisés $(\Omega_i, M_i, P_i)_{i\in N}$. On cherche à construire un espace probabilisé (Ω, M, P) et une suite dénombrable de v.a. $\left(X_i : (\Omega, M, P) \to (\Omega_i, M_i)\right)_{i\in N}$ mutuellement

indépendantes telles que $\forall i \in N\ P_{X_i} = P_i$.
On pose $\Omega = \prod_{i\in N} \Omega_i$, et X_i la projection sur la i-eme cordonnée.
On pose $M = \otimes_{i\in N} M_i = \sigma((X_i)_{i\in N}) = \sigma(C)$ avec $C = \{C_n \times \prod_{i\geq n+1} \Omega_i : n \in N, C_n \in \otimes_{0\leq i\leq n} M_i\}$
C est une algèbre appelée algèbre des cylindres.
On pose $Q : C \to [0,1] : C_n \times \prod_{i\geq n+1} \Omega_i \mapsto (\otimes_{0\leq i\leq n} P_i)(C_n)$
Théorème de Kolmogorov : La fonction d'ensemble Q se prolonge en une unique probabilité P sur (Ω, M). De plus dans (Ω, M, P) les X_i sont mutuellement indépendantes.
En conséquence de ce théorème, on peut parler plus librement d'une suite $(X_n)_{n\in N}$ de v.a.r. ou même de $\vec{v}$.a.r. indépendants sur un espace probabilisé (Ω, M, P).
Tribu terminale. Dans de nombreux problèmes, on est intéressé par le comportement limite d'une suite de variables aléatoires. Un exemple élémentaire est la suite des proportions de piles dans un tirage successif à pile ou face. Dans de telles situations, les événements dans une tribu engendrée par un nombre fini de variables ont peu d'intérêt, et on ne s'intéresse qu'aux événements définis ultimement.
Soit une suite de tribus $(M_n)_{n\in N}$ indépendantes (par exemple $M_n = \sigma(X_n)$ avec $\perp\!\!\!\perp_{n\in N} X_n$) et soit $F_n = \sigma(\bigcup_{k\geq n} M_k)$, (donc $F_{n+1} \subseteq F_n$). La **tribu terminale/des évènements terminaux adaptée aux tribus M_n** est $F_\infty = \bigcap_{n\in N} F_n$
Loi du 0-1. Tout évènement d'une tribu terminale est de probabilité soit 0 soit 1.
Pour une suite d'évènements $(A_n)_{n\in N}$ d'un espace probabilisé, on modélise l'évènement $A =$« $(A_n)_{n\in N}$ **se produit une infinité de fois/infiniment souvent/i.s.** » par $A = \bigcap_{n\in N} \bigcup_{k\geq n} A_k$. Cet évènement est un évènement terminal pour la tribu terminale adaptée aux tribus $\sigma(A_n)$ et donc $P(A) \in \{0,1\}$. On peut écrire : $\omega \in A \Leftrightarrow \forall n \in N\ \exists k_\omega \geq n\ \omega \in A_k$ attention car k depend de ω.
Pour une suite de v.a. indépendantes $\perp\!\!\!\perp_{n\in N} X_n$ et une suite $(a_n)_n$ tendant vers ∞, pour tout k_0 l'évènement $\left\{\frac{1}{a_n}\sum_{k=k_0}^{n} X_k \text{ converge quand } n \to \infty\right\}$ est un évènement final adapté aux $\sigma(X_n)$.
Lemmes de Borel-Cantelli. Une suite d'évènements dont la somme des probabilités est finie, est une suite se produisant presque surement un nombre fini de fois. $\sum_{n\in N} P(A_n) < \infty \Rightarrow P(A_n\ i.s.) = 0$
Une suite d'évènements <u>indépendants</u> dont la somme des probabilités est infinie, est une suite se produisant presque surement infiniment souvent. $\perp\!\!\!\perp_{n\in N} A_n$ et $\sum_{n\in N} P(A_n) = \infty \Rightarrow P(A_n\ i.s.) = 1$
En fait l'indépendance 2 à 2 des évènements suffit.
Soit $(X_n)_n$ suite iid de $Ber\left(\frac{1}{2}\right)$. Par Borel-Cantelli, il y a p.s. une infinité de 0, et p.s. une infinité de 1.
IV. 4. Liste de distributions et propriétés : TODO Wikipédia devrait suffire.

IV.4. Vecteurs aléatoires gaussiens et loi gaussiennes
Exemple fondamental.
Pour $X \sim N(\mu, \sigma^2)$ $\forall t \in R\ \varphi_X(t) = e^{i\mu t - \frac{1}{2}\sigma^2 t^2}$
Soit $\vec{X} = (X_1, \dots, X_d) \mid \perp\!\!\!\perp_{i=1}^{d} X_i \mid \forall i\ X_i \sim N(\mu_i, \sigma_i^2)$, $\vec{\mu} = (\mu_i)_i$.
$$E(\vec{X}) = \vec{\mu},\ var(\vec{X}) = \begin{pmatrix} \sigma_1^2 & & 0 \\ & \ddots & \\ 0 & & \sigma_d^2 \end{pmatrix},\ var(\vec{X})^{-1} = \begin{pmatrix} \sigma_1^{-2} & & 0 \\ & \ddots & \\ 0 & & \sigma_d^{-2} \end{pmatrix}$$
$\forall \vec{t} \in R^d\ \varphi_{\vec{X}}(\vec{t}) = \prod_{i=1}^{d} \varphi_{X_i}(t_i) = e^{i(\vec{\mu}|\vec{t}) - \frac{1}{2}(\vec{t}^T var(\vec{X})\vec{t})}$

Pour $\alpha \in R^d,\ \forall t \in R\ \ \varphi_{(\vec{\alpha}|\vec{X})}(t) = e^{i(\vec{\mu}|\vec{\alpha})t - \frac{1}{2}(\vec{\alpha}^T var(\vec{X})\vec{\alpha})t^2}$ donc $(\vec{\alpha}|\vec{X}) \sim N\left((\vec{\mu}|\vec{\alpha}), \vec{\alpha}^T var(\vec{X})\vec{\alpha}\right)$

Lorsque $var(\vec{X})$ est définie positive càd $\exists i\ \sigma_i \neq 0$ cad $\det\left(var(\vec{X})\right) > 0$, on a

$\forall \vec{x} \in R^d\ f_{\vec{X}}(\vec{x}) = \prod_{i=1}^{d} f_{X_i}(x_i) = \frac{1}{\sqrt{2\pi}^d} \cdot \frac{1}{\sqrt{\det var(\vec{X})}} \cdot e^{-\frac{1}{2}(\vec{x}-\vec{\mu})^T var(\vec{X})^{-1}(\vec{x}-\vec{\mu})} =$

$\frac{1}{\prod_{i=1}^{d} \sigma_i \sqrt{(2\pi)^d}} e^{-\frac{1}{2}\sum_{i=1}^{d}\left(\frac{x_i-\mu_i}{\sigma_i}\right)^2}$

On appelle **gaussienne multidimensionnelle standard centrée** et on note $\boldsymbol{N_d(\overrightarrow{0_d}, I_d)}$, la loi de probabilité caractérisée par la fonction caractéristique $\varphi(\vec{t}) = e^{-\frac{1}{2}(\vec{t}^T\vec{t})}$, autrement dit c'est la loi de probabilité d'un vecteur aléatoire $\vec{X} = (X_1, \dots, X_d)\ \big|\!\perp\!\!\!\perp_{i=1}^{d} X_i\ \big|\forall i\ X_i \sim N(0,1)$.

On appelle **gaussienne multidimensionnelle d'espérance $\boldsymbol{\vec{\mu} \in R^d}$ et matrice de covariance $\boldsymbol{\Gamma}$ <u>symétrique semi-définie positive</u>** ($\Gamma = AA^T$) et on note $\boldsymbol{N_d(\vec{\mu}, \Gamma)}$ la loi de $A\vec{X} + \vec{\mu}$ avec $\vec{X} \sim N_d(\overrightarrow{0_d}, I_d)$

Autrement dit c'est la loi de probabilité caractérisée par la fonction caractéristique : $\varphi(\vec{t}) = e^{i(\vec{\mu}|\vec{t}) - \frac{1}{2}(\vec{t}^T\Gamma\vec{t})}$

Un **vecteur gaussien** est un vecteur aléatoire réel dont toute combinaison linéaire des composantes donne une v.a.r. de loi gaussienne.

La loi d'un vecteur gaussien est caractérisée par son vecteur espérance et sa matrice de covariance.

En fait tout $\vec{v}$.a.r. $\vec{X} \sim N_d(\vec{\mu}, \Gamma)$ est un vecteur gaussien tel que $E(\vec{X}) = \vec{\mu}, var(\vec{X}) = \Gamma$, et réciproquement tout vecteur gaussien vérifie $\vec{X} \sim N_d\left(E(\vec{X}), \mathrm{var}(\vec{\mathrm{X}})\right)$.

Ainsi un vecteur gaussien correspond à un vecteur aléatoire réel de distribution une gaussienne multidimensionnelle, on peut identifier les deux concepts.

Tout sous-vecteur d'un vecteur gaussien est un vecteur gaussien.

Un vecteur gaussien/une gaussienne multidimensionnelle est **non dégénéré** ssi la matrice de covariance Γ est inversible ssi son déterminant est non nul.

Un vecteur gaussien non dégénéré est absolument continu de densité $f_{\vec{X}}(\vec{x}) = \frac{1}{\sqrt{(2\pi)^d}} \cdot \frac{1}{\sqrt{\det\Gamma}} \cdot e^{-\frac{1}{2}(\vec{x}-\vec{\mu})^T\Gamma^{-1}(\vec{x}-\vec{\mu})}$

$AN_d(\vec{\mu}, \Gamma) + \vec{a} \sim N_d(A\vec{\mu} + \vec{a}, A\Gamma A^T)$

Pour un vecteur gaussien $\vec{X} \sim N_d(\vec{\mu}, \Gamma)$ on a $\perp\!\!\!\perp_{i=1}^{d} X_i$ ssi Γ est diagonale.

Deux v.a.r. formant un vecteur gaussien, sont indépendantes ssi elles sont non corrélées.

Un vecteur gaussien, à ses composantes mutuellement indépendantes ssi elles sont 2 à 2 non corrélées.

V. Convergence de suites de variables aléatoires

V.1. Convergence presque sûre.

Pour une suite de v.a.r. $(X_n)_{n\in N}$ et une v.a.r. X sur un même espace probabilisé $\{X_n \to_{n\to\infty} X\} = \bigcap_{p\geq 1} \bigcup_{m\in N} \bigcap_{n\geq m} \left\{|X_n - X| < \frac{1}{p}\right\}$ est bien un évènement.

Une suite de v.a.r. $(X_n)_{n\in N}$ sur un espace probabilisé **converge presque surement (p.s.) vers une v.a.r. $\boldsymbol{X}$** sur ce même espace probabilisé et on écrit $\boldsymbol{X_n \to_{n\to\infty} X}$ **p.s.** ssi la suite de fonction X_n converge simplement vers la fonction X p.s. ssi $P(X_n \to_{n\to\infty} X) = 1$

ssi $\forall \varepsilon > 0\ P(\exists m \in N\ \forall n \geq m\ |X_n - X| < \varepsilon) = 1$

ssi $\forall \varepsilon > 0\ P(|X_n - X| \geq \varepsilon \text{ i.s.}) = 0$
ssi $\forall \varepsilon > 0\ P(\sup_{n \geq m} |X_n - X| \geq \varepsilon) \to_{m \to \infty} 0$ càd $\sup_{n \geq m} |X_n - X| \to^P_{m \to \infty} 0$
ssi $\forall \varepsilon > 0\ P(\exists m \in N\ \forall n \geq m\ |X_n - X_m| < \varepsilon) = 1$ càd $P((X_n)_{n \in N} \text{ de Cauchy}) = 1$
Si $\forall \varepsilon > 0\ \sum_{n=0}^{\infty} P(|X_n - X| \geq \varepsilon) < \infty$ alors $X_n \to_{n \to \infty} X$ p.s. Réciproque vraie si les $(Y_n = X_n - X)_{n \in N}$ sont une famille indépendante mutuellement (ou 2 a 2 suffit).
Donc si $\perp\!\!\!\perp_{n \in N} X_n$ alors $X_n \to_{n \to \infty} 0$ p.s. $\Leftrightarrow\ \forall \varepsilon > 0\ \sum_{n=0}^{\infty} P(|X_n| \geq \varepsilon) < \infty$
On utilise souvent le lemme de Borel Cantelli pour montrer une convergence presque sûre.

Convergence en probabilité.

Une suite de v.a.r. $(X_n)_{n \in N}$ sur un espace probabilisé **converge en probabilité/mesure/L_0 vers une v.a.r. X** sur ce même espace probabilisé et on écrit $\boldsymbol{X_n \to^P_{n \to \infty} X}$
ssi $\forall \varepsilon > 0\ \ P(|X_n - X| \geq \varepsilon) \to_{n \to \infty} 0$
ssi $\forall \varepsilon > 0\ \ P(|X_n - X| < \varepsilon) \to_{n \to \infty} 1$
ssi $|X_n - X| \to^P_{n \to \infty} 0$
Sur un espace probabilisé (Ω, M, P), $d(X,Y) = E(\min(1, |X - Y|))$ et $d(X,Y) = E\left(\frac{|X-Y|}{1+|X-Y|}\right)$ sont des choix possible d'écarts qui rendent l'espace probabilisé semi-métrique. La topologie de cet espace semi-métrique est celle de la convergence en probabilité. $X_n \to^P_{n \to \infty} X$ ssi $d(X_n, X) \to_{n \to \infty} 0$
$X_n \to^P_{n \to \infty} X$ ssi toute suite extraite de X_n admet elle-même une suite extraite convergeant presque surement vers X. La convergence p.s. n'est pas métrisable car si elle l'était, elle coïnciderait avec la convergence en probabilité.
L'espace $(L^0(\Omega, M, P), d)$ est complet pour la distance métrisant la convergence en probabilité.
où $L^0(\Omega, M, P) = \{X \text{ v. a.}\}$
Une suite de v.a.r. $(X_n)_{n \in N}$ sur un espace probabilisé converge en probabilité ssi elle vérifie le **critère de Cauchy en probabilité** : $\forall \varepsilon > 0\ \ \exists N \in N\ \forall n \geq m \geq N\ P(|X_n - X_m| \geq \varepsilon) \leq \varepsilon$
Et même ssi elle vérifie le **critère <u>faible</u> de Cauchy en probabilité** : $\forall \varepsilon > 0\ \ \exists N \in N\ \forall n \geq N\ P(|X_n - X_N| \geq \varepsilon) \leq \varepsilon$ (A vérifier)

Convergence dans Lp.

Pour un espace probabilisé, L^p ($p \in (0, \infty)$) correspond à l'ensemble des v.a.r. telles que $E(|X|^p) < \infty$ quotiente par la relation d'équivalence « être égal presque partout » et $\|X\|_{L^p} = E(|X|^p)^{\frac{1}{p}}$ est une norme sur L^p. Remarque $\|X\|_{L^p}$ est toujours défini (peut valoir $+\infty$)
Une suite de v.a.r. $(X_n)_{n \in N}$ sur un espace probabilisé **converge en norme L^p vers une v.a.r. X** sur ce même espace probabilisé et on écrit $\boldsymbol{X_n \to^{L^p}_{n \to \infty} X}$ ssi $\|X_n - X\|_{L^p} \to_{n \to \infty} 0$
Si $X_n \to^{L^p}_{n \to \infty} X$ alors $E(X_n^p) \to_{n \to \infty} E(X^p)$ mais la réciproque est fausse.
Si $X_n \to^p_{n \to \infty} X$ et $\sup_n E(|X_n|^r) < \infty$ alors $X_n \to^{L^p}_{n \to \infty} X$ pour $p < r$.
L^p est complet.

Convergence in law/distribution

A sequence of real-valued random variables $(X_n)_{n \in N}$ is said to **converge in distribution**, or **converge weakly**, or **converge in law** to a random variable X iff $\mathrm{F}_{X_n}(x) \to_{n \to \infty} F_X(x)$ for every number $x \in \mathbb{R}$ at which F_X is continuous.
The requirement that only the continuity points of F_X should be considered is essential. For example, if X_n are distributed uniformly on intervals $\left(0, \frac{1}{n}\right)$, then this sequence converges in distribution to a

degenerate random variable $X = 0$. Indeed, $F_{X_n}(x) = 0$ for all *n* when $x \leq 0$, and $F_{X_n}(x) = 1$ for all $x \geq \frac{1}{n}$ when $n > 0$. However, for this limiting random variable $F_X(0) = 1$ even though $F_{X_n}(0) = 0$ for all n. Thus the convergence of cdfs fails at the point $x = 0$ where F_X is discontinuous.

Convergence in distribution may be denoted as $\boldsymbol{X_n \to^d X, X_n \to^D X, X_n \to^L X, X_n \Rightarrow X, X_n \to^d \mathcal{L}_X}$ where $\mathcal{L}_X$ is the law (probability distribution) of X. For example, if X is standard normal we can write $X_n \to^d \mathcal{N}(0, 1)$.

For d-random vectors $\left(\overrightarrow{X_n}\right)_{n \in N}$ the convergence in distribution is defined similarly. We say that this sequence **converges in distribution** to a random d-vector $\vec{X}$ iff $\mathrm{P}\left(\overrightarrow{X_n} \in A\right) \to_{n \to \infty} \mathrm{P}(\vec{X} \in A)$ for every A which is a continuity set of $\vec{X}$.

Convergence in law/distribution properties:

Since $F_X(a) = P(X \leq a)$ the convergence in distribution means that the probability for X_n to be in a given range is approximately equal to the probability that the value of X is in that range, provided n is sufficiently large.

In general, convergence in distribution does not imply that the sequence of corresponding probability density functions will also converge. As an example one may consider random variables with densities $f_n(x) = (1 - \cos(2\pi n x))1_{(0,1)}$. These random variables converge in distribution to a uniform $U(0,1)$, whereas their densities do not converge at all.

However, according to Scheffé's theorem, convergence of the probability density functions implies convergence in distribution.

The **portmanteau lemma** provides several equivalent definitions of convergence in distribution. Although these definitions are less intuitive, they are used to prove a number of statistical theorems. The lemma states that $X_n \to^d X$ iff any of the following statements are true:

$\mathrm{P}(X_n \leq x) \to \mathrm{P}(X \leq x)$ for all continuity points of F_X: $x \mapsto \mathrm{P}(X \leq x)$

$\mathrm{E}f(X_n) \to \mathrm{E}f(X)$ for all bounded, continuous functions f (utile pour preuves)

$\mathrm{E}f(X_n) \to \mathrm{E}f(X)$ for all bounded, Lipschitz functions f

$\mathrm{E}f(X_n) \to \mathrm{E}f(X)$ for all C_c^∞ functions f (utile pour preuve réciproque Levy)

$\liminf \mathrm{E}f(X_n) \geq \mathrm{E}f(X)$ for all nonnegative, continuous functions f

$\liminf \mathrm{P}(X_n \in G) \geq \mathrm{P}(X \in G)$ for every open set G

$\limsup \mathrm{P}(X_n \in F) \leq \mathrm{P}(X \in F)$ for every closed set F

$\mathrm{P}(X_n \in B) \to \mathrm{P}(X \in B)$ for all continuity sets B of random variable X

$\limsup \mathrm{E}f(X_n) \leq \mathrm{E}f(X)$ for every upper semi-continuous function f bounded above

$\liminf \mathrm{E}f(X_n) \geq \mathrm{E}f(X)$ for every lower semi-continuous function f bounded below.

The continuous mapping theorem applies to convergence in distribution

Note however that convergence in distribution of X_n to X and Y_n to Y does in general *not* imply convergence in distribution of $X_n + Y_n$ to $X + Y$ or of $X_n Y_n$ to XY.

Lévy's continuity theorem: the sequence X_n converges in distribution to X iff the sequence of corresponding characteristic functions converges pointwise to the characteristic function φ of X.

$X_n \to^d_{n \to \infty} X$ ssi $\forall t \in \mathbb{R}\ \varphi_{X_n}(t) \to_{n \to \infty} \varphi(t)$

Critère de Lévy. Si $\forall t \in \mathbb{R}\ \varphi_{X_n}(t) \to_{n \to \infty} \varphi(t)$ alors φ est la fonction caractéristique d'une v.a.r X et donc il y a convergence en loi vers cette variable par Lévy.

Convergence in distribution is metrizable by the Lévy–Prokhorov metric.

Random variable convergence properties.

Continuous mapping theorems.

Pour toute fonction $\phi \in C^0(R,R)$ et $X_n \to^P_{n\to\infty} X$ alors $\phi(X_n) \to^P_{n\to\infty} \phi(X)$

Pour toute fonction $\phi \in C^0(R,R)$ et $X_n \to^d_{n\to\infty} X$ alors $\phi(X_n) \to^d_{n\to\infty} \phi(X)$

Pour toute fonction $\phi \in C^0(R,R)$ et $X_n \to_{n\to\infty} X$ p.s. alors $\phi(X_n) \to_{n\to\infty} \phi(X)$ p.s.

Provided the probability space is complete:

If $X_n \to^p X$ and $X_n \to^p Y$, then $X = Y$ almost surely.

If $X_n \to^{\text{a.s.}} X$ and $X_n \to^{\text{a.s.}} Y$, then $X = Y$ almost surely.

If $X_n \to^{L^r} X$ and $X_n \to^{L^r} Y$, then $X = Y$ almost surely.

If $X_n \to^p X$ and $Y_n \to^p Y$, then $aX_n + bY_n \to^p aX + bY$ (for any real numbers a and b) and $X_n Y_n \to^p XY$.

If $X_n \to^{\text{a.s.}} X$ and $Y_n \to^{\text{a.s.}} Y$, then $aX_n + bY_n \to^{\text{a.s.}} aX + bY$ (for any real numbers a and b) and $X_n Y_n \to^{\text{a.s.}} XY$.

If $X_n \to^{L^r} X$ and $Y_n \to^{L^r} Y$, then $aX_n + bY_n \to^{L^r} aX + bY$ (for any real numbers a and b).

None of the above statements are true for convergence in distribution.

The chain of implications between the various notions of convergence are noted in their respective sections. They are, using the arrow notation:

$$\begin{array}{ccccccc} \overset{L^q}{\to} & \Rightarrow_{q\geq p\geq 1} & \overset{L^p}{\to} & \Rightarrow & \overset{L^1}{\to} & & \\ & & & & \Downarrow & & \\ \overset{\text{a.c.}}{\to} & \Rightarrow & \overset{\text{a.s.}}{\to} & \Rightarrow & \overset{\text{P}}{\to} & \Rightarrow & \overset{\text{d}}{\to} \end{array}$$

Convergence in probability implies there exists a sub-sequence (k_n) which almost surely converges:

$X_n \to^p X \quad \Rightarrow \quad X_{k_n} \to^{\text{a.s.}} X$

$X_n \to^d c \quad \Rightarrow \quad X_n \to^p c$, provided *c* is a constant.

$X_n \to^d X,\ |X_n - Y_n| \to^p 0 \quad \Rightarrow \quad Y_n \to^d X$

Contre exemples. $\left(X_n \sim Ber(p_n)\right)_n$ v.a. indépendantes avec $p_n \to 0$ et $\sum_{n=1}^{\infty} p_n = \infty$, alors $(X_n)_n$ converge en proba vers 0, mais ne converge pas p.s. vers 0.

X_n v.a. qui vaut n avec proba $\frac{1}{n}$ et 0 avec proba $1 - \frac{1}{n}$ converge en proba vers 0 mais pas vers 0 dans L^1

Slutsky. If X_n converges in distribution to *X* and Y_n converges in distribution to a constant *c*, then the joint vector converges in distribution: $X_n \to^d X,\ Y_n \to^d c \quad \Rightarrow \quad (X_n, Y_n) \to^d (X, c)$ provided *c* is a constant.

Note that the condition that Y_n converges to a constant is important, if it were to converge to a random variable *Y* then we cannot conclude that $(X_n, Y_n) \to^d (X, Y)$.

If X_n converges in probability to *X* and Y_n converges in probability to *Y*, then the joint vector converges in probability to : $X_n \to^p X,\ Y_n \to^p Y \quad \Rightarrow \quad (X_n, Y_n) \to^p (X, Y)$

If X_n converges in probability to *X*, and if 1}} for all *n* and some *b*, then X_n converges in *r*th mean to *X* for all . In other words, if X_n converges in probability to *X* and all random variables X_n are almost surely bounded above and below, then X_n converges to *X* also in any *r*th mean.

Almost sure representation. Usually, convergence in distribution does not imply convergence almost surely. However, for a given sequence X_n which converges in distribution to X_0 it is always possible to find a new probability space (Ω, *F*, P) and random variables {Y_n, *n* = 0, 1, ...} defined on it such that Y_n is

equal in distribution to X_n for each , and Y_n converges to Y_0 almost surely.
We say that X_n **converges almost completely**, or **almost in probability** towards *X* iff for all $\varepsilon > 0$, $\sum_n \mathbb{P}(|X_n - X| > \varepsilon) < \infty$. When X_n converges almost completely towards *X* then it also converges almost surely to *X*. In other words, if X_n converges in probability to *X* sufficiently quickly (i.e. the above sequence of tail probabilities is summable for all $\varepsilon > 0$), then X_n also converges almost surely to *X*.
If S_n is a sum of *n* real independent random variables: $S_n = X_1 + \cdots + X_n$ then S_n converges almost surely if and only if S_n converges in probability.
The dominated convergence th. gives sufficient conditions for a.s. convergence to imply L^1-convergence:
$$\begin{cases} X_n \to^{a.s} X \\ |X_n| < Y \\ E(Y) < \infty \end{cases} \Rightarrow X_n \to^{L^1} X$$
A necessary and sufficient condition for L^1 convergence is $X_n \to^P X$ and the sequence $(X_n)_n$ is uniformly integrable.

V.5. Les lois faible et forte des grands nombres, le théorème limite central

Pour une suite de $\vec{v}$.a.r. $\left(\overrightarrow{X_n}\right)_{n\geq 1}$ on note $\overrightarrow{\overline{X_n}} = \frac{1}{n}\sum_{i=1}^{n} \overrightarrow{X_k}$
Soit $c \in R$ et X_n suite de v.a.r., $X_n \to^{L^2} c$ ssi $E(X_n) \to c$ et $var(X_n) \to 0$
Loi faible des grands nombres. Pour une suite $(X_n)_{n\in N}$ de v.a.r. iid et L^1 $(E(|X_1|) < \infty)$ alors $\overline{X_n} \to^P_{n\to\infty} E(X_1) \in R$
Loi forte des grands nombres. Pour une suite $(X_n)_{n\in N}$ de v.a.r. iid, $X_1 \in L^1$ ssi $E|X_1| < \infty$ ssi $\overline{X_n} \to^{p.s.} E(X_1)$
Donc pour une suite $(X_n)_{n\geq 1}$ de v.a.r. iid L^1, on a toujours $\overline{X_n} \to^{p.s.} E(X_1)$.
Loi forte des grands nombres dans L^2. Pour une suite $(X_n)_{n\geq 1}$ de v.a.r. iid L^2, $\overline{X_n} \to^{L^2} E(X_1)$
Théorème de la limite centrale. Pour une suite $\left(\overrightarrow{X_n}\right)_{n\geq 1}$ de $\vec{v}$.a.r. iid L^2 alors
$$\sqrt{n}\left(\overrightarrow{\overline{X_n}} - E\left(\overrightarrow{X_0}\right)\right) \to^d N\left(0, var\left(\overrightarrow{X_0}\right)\right)$$
Dimension 1 : Pour une suite $(X_n)_{n\geq 1}$ de v.a.r. iid L^2 avec $\sigma(X_0) > 0$ alors $\sqrt{n}\frac{\left(\overline{X_n} - E(X_0)\right)}{\sigma(X_0)} \to^d N(0,1)$
Etant donnée une telle suite X_n on peut poser $Z_n = \frac{X_n - E(X_0)}{\sigma(X_0)}$ et alors $\sqrt{n}\overline{Z_n} \to^d N(0,1)$
Réciproquement, une suite $(Z_n)_{n\geq 1}$ de v.a.r. iid telles que $\sqrt{n}\overline{Z_n}$ converge en loi, converge automatiquement vers $N(0,1)$ et vérifie $E(Z_1) = 0, Z_1 \in L^2$.
Théorème limite centrale poissonien. Une suite de v.a.r. $(S_n)_{n\geq 1}$ de loi binomiale $B(n, p_n)$ avec $\forall n \in N\ p_n > 0$ telle que $np_n \to \lambda > 0$ alors $S_n \to^d P(\lambda)$

VI. Probabilités et espérances conditionnelles

Disclaimer : Cette section adopte des conventions personnelles, peu orthodoxes. J'avais envie de réfléchir autrement, mais comme je peux me tromper, je ne garantis rien au lecteur.
La **tribu induite/tribu trace** sur une partie fixée d'un espace mesurable par une tribu, est la tribu (c'en est bien une) obtenu en intersectant tous les éléments de la tribu initiale avec la partie fixée.
L'**espace mesuré induit** sur une partie mesurable d'un espace mesure est l'espace mesure formé par la partie, la tribu trace induite sur cette partie, et la restriction de la mesure a cette tribu induite.
L'**espace probabilisé induit** par un évènement B d'un espace probabilisé (Ω, M, P) de probabilite non nulle $P(B) \neq 0$ est l'espace probabilisé formé par l'évènement, la tribu trace induite sur cet évènement, et la mesure de probabilité restreinte à cet évènement, normalisée par le poids total de cet évènement de sorte à ce qu'elle reste une mesure de probabilité sur ce nouvel espace : On note $\left(B, M|B, P_{|B} = P(.)/P(B)\right)$
Pour un évènement $A \in M$, l'évènement induit $\boldsymbol{A|B}$ est l'évènement $A|B = A \cap B \in M|B$.

La probabilité conditionnelle d'un évènement est la probabilité induite de l'évènement induit :
$\boldsymbol{P}(\boldsymbol{A}|\boldsymbol{B}) = P_{|B}(A|B) = P_{|B}(A \cap B) = P(A \cap B)/P(B)$
Le point de vue orthodoxe est de considérer $P(.|B)$ définie sur tout l'espace mesurable initial, je préfere l'intuition de se placer dans l'espace induit et de considérer A comme induit dans cet espace en l'intersectant avec B.
Pour une fonction X: $(\Omega, \mathrm{M}, \mathrm{P}) \to (\Omega', M')$ mesurable, la fonction induite $\boldsymbol{X}|\boldsymbol{B}$: $\left(B, M\middle|B, P_{|B}\right) \to (\Omega', M')$ coïncidant avec X sur B est mesurable. $\forall A' \in M'\ \{\omega \in B \mid (X|B) \in C\} = \{\omega \in \Omega\ |X \in C\} \cap B \in M|B$
De façon générale $\boldsymbol{P}((\boldsymbol{X} \in \boldsymbol{C})|\boldsymbol{B}) = \boldsymbol{P}\big((\boldsymbol{X}|\boldsymbol{B}) \in \boldsymbol{C}\big) = P_{|B}\big((X|B) \in C\big) = P\big(B \cap (X \in C)\big)/P(B)$
En employant la notation $|B$ quelque part je suppose implicitement m'être place dans l'espace induit, donc la probabilité que j'utilise doit obligatoirement être celle de l'espace induit. Toutes les variables aléatoires, évènements (pas ceux de l'espace d'arrivée) et probabilité considérés doivent être induits, c'est l'intuition de « conditionner » suivant un évènement, c'est-à-dire supposer qu'il s'est produit.
L'avantage théorique de ce point de vue est qu'il n'y a pas besoin de définitions spécifiques pour les concepts d'espérance/probabilité/indépendance conditionnelle a un evenement fixe, puisque le concept de variable aléatoire conditionnelle a un évènement fixé est défini.

Fomules

En commettant l'abus d'écriture $\boldsymbol{p}(\boldsymbol{X} = \boldsymbol{x}) = P(X \in [x, x + dx]) = f_X(x)dx$, les formules continu/discret sont analogues. (il suffit de remplacer $\int$ par $\sum$)
$P(X \in A) = \int_{x \in A} p(X = x)$, $P(Y \in B) = \int_{y \in B} p(Y = y)$
$\boldsymbol{p}(\boldsymbol{Y} = \boldsymbol{y}|\boldsymbol{X} = \boldsymbol{x}) = \frac{p\big((X,Y)=(x,y)\big)}{p(X=x)} = \frac{p(X=x,Y=y)}{p(X=x)}$ (écriture abusive néanmoins utile)
$p(Y = y|X = x)p(X = x) = p\big((X, Y) = (x, y)\big) = p(X = x, Y = y)$
$p(Y = y) = \int_{x \in R} p(Y = y|X = x)p(X = x) = \int_{x \in R} p(X = x, Y = y)$ (Bayes)
Plus généralement
$\boldsymbol{P}(\boldsymbol{Y} \in \boldsymbol{B}\ |\boldsymbol{X} \in \boldsymbol{A}) = P(X \in A, Y \in B)/P(X \in A)$
$P(Y \in B \mid X \in A)P(X \in A) = P(X \in A, Y \in B) = P((X, Y) \in A \times B)$
Ces notations me paraissent plus intuitives à utiliser et à mémoriser.

Interprétation (divagation?) épistémique des probabilités : Ces intuitions apparaissent souvent dans les exo d'espérance conditionnelle sans être bien expliquées. Ce qui suit représente ma vision personnelle en l'état. J'ai du mal à trouver des informations précises sur ce sujet.

Intuition de la tribu :

Un espace probabilisé (Ω, M, P) représente la simulation d'un phénomène aléatoire.
La tribu considérée M représente l'ensemble des informations disponibles a posteriori, c'est-à-dire après la simulation du phénomène.
Dire qu'un évènement $A \subseteq \Omega$ appartient à la tribu M c'est-à-dire $A \in M$ signifie « on pourra répondre à la question binaire : A s'est-il produit (oui ou non) ?» autrement dit « on saura si A s'est produit » après avoir « simulé » le phénomène aléatoire.
A posteriori, Si, on sait si A s'est produit, alors, on sait si $\overline{A}$ s'est produit, donc il est clair qu'une tribu doit être stable par complémentaire d'après cette intuition.
De plus si on sait si A s'est produit, et on sait si B s'est produit, alors on sait si $A \cup B$,$A \cap B$ s'est produit, donc une tribu doit être stable par réunion/intersection.
On suppose cette stabilité s'étend au cas dénombrable infini, pour généraliser notre intuition car rien ne l'empêche, on ne peut pas généraliser au-delà du cas dénombrable car problèmes liés à la définition d'une mesure.
On sait toujours que Ω se produit et que $\emptyset$ ne se produit pas donc ces deux la doivent au moins être dans toute tribu.
Une tribu est donc un ensemble d'informations, chaque partie élément de la tribu représente une de ces

informations.
Une mesure de probabilité n'est définie que sur les évènements mesurables, c'est-à-dire les éléments de la tribu.
Conditionner par rapport à une sous-tribu, autrement dit se placer dans une sous-tribu, revient à restreindre les informations disponibles a posteriori, cela signifie perdre des informations.
Ceci est à contraster avec le fait de conditionner par rapport à un évènement fixé. Quand on conditionne par rapport à un évènement on peut interpréter ça comme se placer dans la sous-tribu trace de la tribu initiale, mais on ne perd pas réellement d'information, on élague/fait un zoom sur la tribu initiale car on considère qu'un évènement est déjà connu et fixé a priori, ce qui simplifie le problème.
Intuition de la mesurabilité : les éléments et seuls ceux dont on pourra dire a posteriori s'ils se sont produits ou non peuvent être affectés d'une probabilité.
Une variable aléatoire/fonction mesurable est une fonction telle que pour toute partie d'arrivée dont on aura l'information, alors on aura l'information de « est-ce que la variable tombe dedans ?» cad $\{X \in B\}$ cad $X^{-1}(B)$.
Les variables aléatoires considérées modélisent les variables que l'on observe et ne peuvent dépendre que des informations dont on dispose a posteriori, c'est ce que traduit la condition de mesurabilité.
Cette condition est bien nécessaire pour justifier l'intuition d'information.
Tribu engendrée par un truc = information représentée par ce truc
La tribu engendrée par un évènement seul $A \subseteq \Omega$ est $\boldsymbol{\sigma(A)} = \{\emptyset, \Omega, A, \overline{A}\}$ elle représente la même chose que $A \in M$ car $A \in M \Leftrightarrow \sigma(A) \subseteq M$. C'est l'information de l'évènement A (si oui ou non il s'est produit).
La tribu engendrée par une famille quelconque de partie $F \subseteq P(\Omega)$, $\boldsymbol{\sigma(F)}$ représente « on pourra répondre à toute question logique d'occurrence, faisant intervenir les évènements de la famille (en nb fini ou dénombrable) »
Par exemple $A \cap (\overline{B} \cup C) \in \sigma(A, B, C)$ car « on saura répondre a : Est-ce que A s'est produit et (B ne s'est pas produit ou C s'est produit) ?» à partir de l'information de A, B, C.
La tribu engendrée par une variable aléatoire seule X représente « on pourra répondre à toute question logique formulée sur X (plus précisément à toute question logique d'occurrence faisant intervenir des évènements (en nb fini ou dénombrable) $X \in A$ devant être eux-mêmes des informations)».
La tribu engendrée par une famille quelconque de variables aléatoires, $\boldsymbol{\sigma(X_i : i \in I)}$ représente « on pourra répondre à toute question logique formulée sur les variables de la famille (en nb fini ou dénombrable) »
Un avantage d'une tribu engendrée par une variable aléatoire est qu'elle peut servir à modéliser l'information correspondant à des questions non binaires plus complexes.
Exemple : Soit un dé a 6 faces numérotées de 1 à 6 ou certaines faces partagent une même couleur par exemple $\{1,2\}$ de couleur rouge, $\{3,6\}$ de couleur verte, $\{5,4\}$ de couleur bleue, on suppose X est une v.a. donnant le résultat d'un lancer de dé entre 1 et 6. On peut définir la fonction couleur et l'appliquer a X pour obtenir une nouvelle v.a. donnant la couleur du résultat. σ(couleur de X) représente l'information de la couleur de X, on a σ(couleur de X)$\subseteq \sigma(X)$ car l'information de X concerne toutes les questions sur X, en particulier celles concernant sa couleur.
Conditionner par rapport à une tribu, signifie conditionner sachant un ensemble d'informations a posteriori. En général on cherche à calculer l'espérance conditionnelle sachant la tribu.
Si on ne conditionne pas, on garde la tribu initiale, on aura donc toutes les informations a posteriori $E(X|M) = X$ car « la meilleure estimation de X en étant omniscient a posteriori, est X »
Pour une sous-tribu G de la tribu initiale M,
$E(X|G)$ représente « la meilleure estimation de X sachant les informations fournies par G a posteriori »
En fait on rend la variable X, G mesurable, en la rendant plus grossière, en perdant de l'information.

Meilleur estimation signifie, que l'on projette X sur l'espace des fonctions G mesurables de sorte a minimiser une metrique, celle de rigueur etant la norme L^2 associée au produit scalaire $E(XY)$.
Exemple : $E(X|\sigma(\text{couleur de } X))$ signifie qu'a posteriori on ne dispose que de l'information de la couleur de X, on ne saura pas sur quelle valeur X est tombée. Il faut donc séparer les cas suivant les seuls informations dont on disposera a posteriori : la couleur de X. $E\left(X\middle|\sigma\left(\text{couleur(X)}\right)\right)$ est donc une v.a. telle que pour chaque $c \in \{R, V, B\}$ $E\left(X\middle|\sigma\left(\text{couleur(X)}\right)\right)(c) = E(X|\text{couleur}(X) = c)$
donc $E\left(X\middle|\sigma\left(\text{couleur(X)}\right)\right) : R \mapsto 1.5, V \mapsto 4.5, B \mapsto 4.5$ (si équiprobabilité initialement)
On a restreint la tribu initiale a une sous-tribu, et rendu X plus grossier en prenant la meilleure estimation respectivement à la norme naturelle L^2. Il n'y a pas vraiment de concept unique de variable conditionnée par une tribu car ce processus de perte d'information n'est pas nécessairement unique, l'espérance conditionnelle est le processus traditionnel, appelé ainsi car il consiste à compresser les pertes d'information locales par leur espérance. L'espérance conditionnelle donne lieu à beaucoup de propriétés remarquables grâce à sa simplicité théorique.

VI.4. Espérances conditionnelles dans les espaces gaussiens

Définition de l'espérance conditionnelle

TODO

Propriétés

All the following formulas are to be understood in an almost sure sense. The σ-algebra $\mathcal{H}$ could be replaced by a random variable Z.

Pulling out independent factors: If X is independent of $\mathcal{H}$, then $E(X \mid \mathcal{H}) = E(X)$.
Let $B \in \mathcal{H}$. Then X is independent of 1_B, so we get that

$$\int_B X \, dP = E(X1_B) = E(X)E(1_B) = E(X)P(B) = \int_B E(X) \, dP.$$

Thus the definition of conditional expectation is satisfied by the constant random variable $E(X)$, as desired.

If X is independent of $\sigma(Y, \mathcal{H})$, then $E(XY \mid \mathcal{H}) = E(X)\,E(Y \mid \mathcal{H})$. Note that this is not necessarily the case if X is only independent of $\mathcal{H}$ and of Y.

If X, Y are independent, $\mathcal{G}, \mathcal{H}$ are independent, X is independent of $\mathcal{H}$ and Y is independent of $\mathcal{G}$, then $E(E(XY \mid \mathcal{G}) \mid \mathcal{H}) = E(X)E(Y) = E(E(XY \mid \mathcal{H}) \mid \mathcal{G})$.

Stability:

If X is $\mathcal{H}$-measurable, then $E(X \mid \mathcal{H}) = X$.

If Z is a random variable, then $\mathrm{E}(f(Z) \mid Z) = f(Z)$. In its simplest form, this says $\mathrm{E}(Z \mid Z) = Z$.

Pulling out known factors:

If X is $\mathcal{H}$-measurable, then $E(XY \mid \mathcal{H}) = X\,E(Y \mid \mathcal{H})$.

If Z is a random variable, then $\mathrm{E}(f(Z)Y \mid Z) = f(Z)\mathrm{E}(Y \mid Z)$.

Law of total expectation: $E(E(X \mid \mathcal{H})) = E(X)$.

Tower property:

For sub-σ-algebras $\mathcal{H}_1 \subset \mathcal{H}_2 \subset \mathcal{F}$ we have $E(E(X \mid \mathcal{H}_2) \mid \mathcal{H}_1) = E(X \mid \mathcal{H}_1)$.

A special case is when Z is a $\mathcal{H}$-measurable random variable. Then $\sigma(Z) \subset \mathcal{H}$ and thus $E(E(X \mid \mathcal{H}) \mid Z) = E(X \mid Z)$.

Doob martingale property: the above with $Z = E(X \mid \mathcal{H})$ (which is $\mathcal{H}$-measurable), and using also $\mathrm{E}(Z \mid Z) = Z$, gives $E(X \mid E(X \mid \mathcal{H})) = E(X \mid \mathcal{H})$.

For random variables X, Y we have $E(E(X \mid Y) \mid f(Y)) = E(X \mid f(Y))$.

For random variables X, Y, Z we have $E(E(X \mid Y, Z) \mid Y) = E(X \mid Y)$.
Linearity: we have $E(X_1 + X_2 \mid \mathcal{H}) = E(X_1 \mid \mathcal{H}) + E(X_2 \mid \mathcal{H})$ and $E(aX \mid \mathcal{H}) = a\,E(X \mid \mathcal{H})$ for $a \in \mathbb{R}$.
Positivity: If $X \geq 0$ then $E(X \mid \mathcal{H}) \geq 0$.
Monotonicity: If $X_1 \leq X_2$ then $E(X_1 \mid \mathcal{H}) \leq E(X_2 \mid \mathcal{H})$.
Monotone convergence: If $0 \leq X_n \uparrow X$ then $E(X_n \mid \mathcal{H}) \uparrow E(X \mid \mathcal{H})$.
Dominated convergence: If $X_n \to X$ and $|X_n| \leq Y$ with $Y \in L^1$, then $E(X_n \mid \mathcal{H}) \to E(X \mid \mathcal{H})$.
Fatou's lemma: If $E(\inf_n X_n \mid \mathcal{H}) > -\infty$ then $E(\liminf_{n\to\infty} X_n \mid \mathcal{H}) \leq \liminf_{n\to\infty} E(X_n \mid \mathcal{H})$.
Jensen's inequality: If $f\colon \mathbb{R} \to \mathbb{R}$ is a convex function, then $f(E(X \mid \mathcal{H})) \leq E(f(X) \mid \mathcal{H})$.
Conditional variance: $\mathrm{Var}(X \mid \mathcal{H}) = \mathrm{E}((X - \mathrm{E}(X \mid \mathcal{H}))^2 \mid \mathcal{H})$
Algebraic formula for the variance: $\mathrm{Var}(X \mid \mathcal{H}) = \mathrm{E}(X^2 \mid \mathcal{H}) - (\mathrm{E}(X \mid \mathcal{H}))^2$
Law of total variance: $\mathrm{Var}(X) = \mathrm{E}(\mathrm{Var}(X \mid \mathcal{H})) + \mathrm{Var}(\mathrm{E}(X \mid \mathcal{H}))$.
Martingale convergence: For a random variable X, that has finite expectation, we have $E(X \mid \mathcal{H}_n) \to E(X \mid \mathcal{H})$, if either $\mathcal{H}_1 \subset \mathcal{H}_2 \subset \ldots$ is an increasing series of sub-σ-algebras and $\mathcal{H} = \sigma(\bigcup_{n=1}^{\infty} \mathcal{H}_n)$ or if $\mathcal{H}_1 \supset \mathcal{H}_2 \supset \ldots$ is a decreasing series of sub-σ-algebras and $\mathcal{H} = \bigcap_{n=1}^{\infty} \mathcal{H}_n$.
Conditional expectation as L^2-projection: If X, Y are in the Hilbert space of square-integrable real random variables (real random variables with finite second moment) then
for $\mathcal{H}$-measurable Y, we have $E(Y(X - E(X \mid \mathcal{H}))) = 0$, i.e. the conditional expectation $E(X \mid \mathcal{H})$ is in the sense of the $L^2(P)$ scalar product the orthogonal projection from X to the linear subspace of $\mathcal{H}$-measurable functions. (This allows to define and prove the existence of the conditional expectation based on the Hilbert projection theorem.)
Moreover the mapping $X \mapsto \mathrm{E}(X \mid \mathcal{H})$ is self-adjoint: $\mathrm{E}(X\mathrm{E}(Y \mid \mathcal{H})) = \mathrm{E}(\mathrm{E}(X \mid \mathcal{H})\mathrm{E}(Y \mid \mathcal{H})) = \mathrm{E}(\mathrm{E}(X \mid \mathcal{H})Y)$
Conditioning is a contractive projection of L^p spaces $L^p(\Omega, \mathcal{F}, P) \to L^p(\Omega, \mathcal{H}, P)$. I.e., $\mathrm{E}(|\mathrm{E}(X \mid \mathcal{H})|^p) \leq \mathrm{E}(|X|^p)$ for any $p \geq 1$
Doob's conditional independence property:[1] If X, Y are conditionally independent given Z, then $P(X \in B \mid Y, Z) = P(X \in B \mid Z)$ (equivalently, $E(1_{\{X \in B\}} \mid Y, Z) = E(1_{\{X \in B\}} \mid Z)$)

VII. Martingales (à temps discret)
VII.1 Généralités
VII.2 Théorèmes de convergence
VII.3 Application à la loi des grands nombres
VIII. Chaînes de Markov (à espace d'états dénombrable)
VIII.1. La propriété de Markov
VIII.2. Calcul des lois marginales
VIII.3. Généralisation de la propriété de Markov
VIII.4. Comportement asymptotique. Mesures invariantes
VIII.5. Récurrence et transience
VIII.6. Comportement asymptotique d'une chaîne de Markov

1