UP1 – Majeure Science des Données 2025-2026

Probabilités Avancées

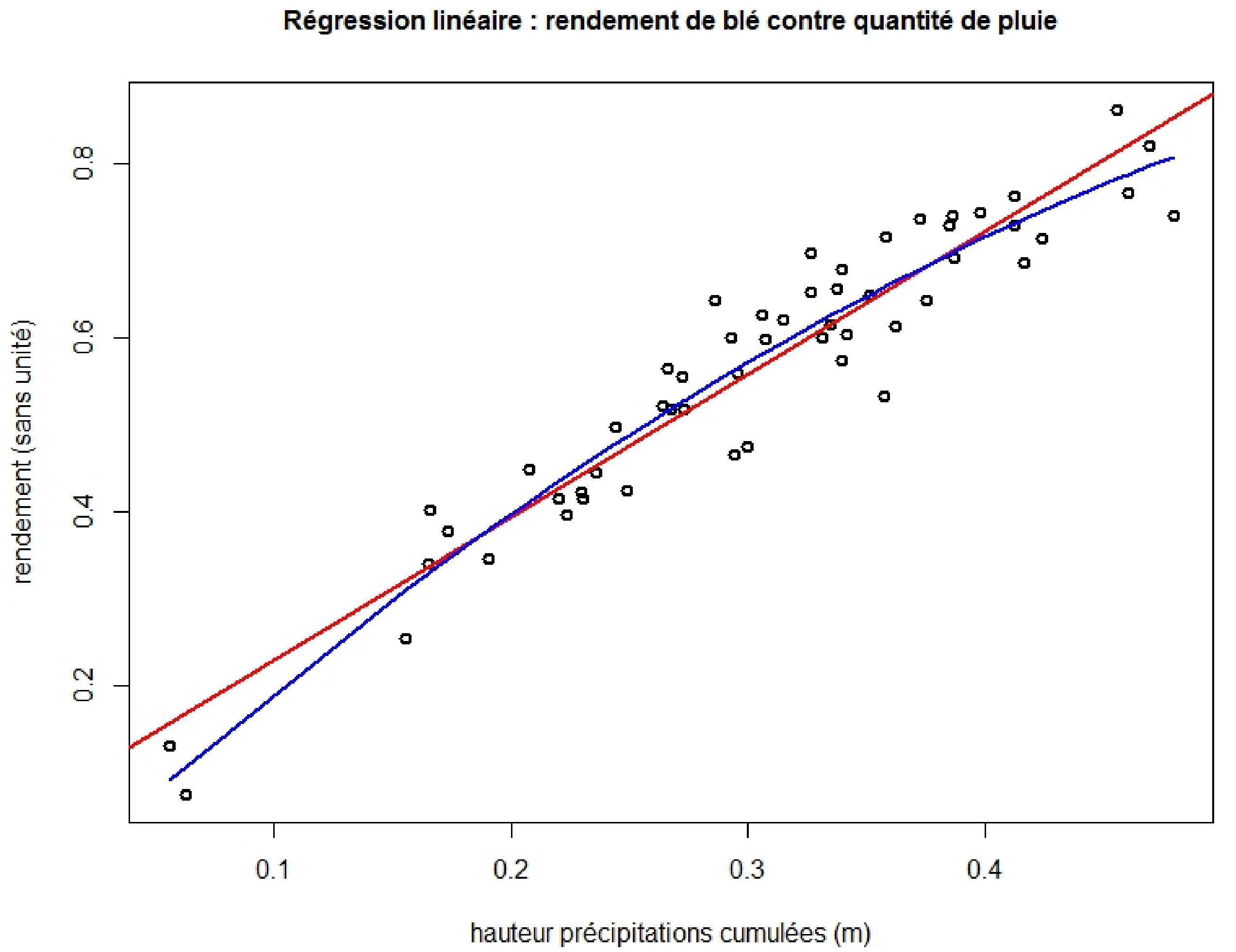
Chapitre I : Vecteurs Gaussiens (VGs)

Chapitre II : Lois conditionnelles

☞ chapitre 1 au cœur de la statistique classique (Régression Linéaire, Séries Temporelles, …)

☞ chapitre 2 indispensable pour bien comprendre les problématiques de prédiction (régression, classification supervisée, …) ou de prévision

Motivation pour l'étude des VGs : Régression Linéaire Multiple (RLM)



☞ **objectif** : prédire le rendement d'une parcelle de blé juste avant récolte…

On devine une relation de la forme (droite en rouge)

$$\textbf{rendement} = \boldsymbol{\beta_0} + \boldsymbol{\beta_1}\times\textbf{pluie} + \textbf{Erreur}$$

☛ modèle de **régression linéaire simple (RLS)** de la forme $\boldsymbol{y = \beta_0 + \beta_1 x + \varepsilon}$ avec p = 1 prédicteur : x = pluie

ou d'une forme « plus complexe »

$$\textbf{rendement} = \boldsymbol{\beta_0} + \boldsymbol{\beta_1}\times\textbf{pluie} + \boldsymbol{\beta_2}\times\textbf{pluie}^2 + \textbf{Erreur}$$

☛ modèle de **régression linéaire multiple** de la forme $\boldsymbol{y = \beta_0 + \beta_1 x^{(1)} + \beta_2 x^{(2)} + \varepsilon}$ avec p = 2 prédicteurs : $x^{(1)}$ = pluie et $x^{(2)}$ = pluie^2

☛ la variable x est **aléatoire** ou **non contrôlée**. Même dans le cas où ce prédicteur serait **contrôlé**, la réponse y = « rendement » serait aléatoire compte tenu du terme d'**Erreur,** composante résiduelle qui intègre tous les autres facteurs (aléatoires ou non) influençant le rendement…

Modèle général de régression linéaire multiple (décliné sur la population des n individus) **:**

$$\mathbf{1 \leq i \leq n, \quad y_i = \beta_0 + \beta_1 x_i^{(1)} + \ldots + \beta_p x_i^{(p)} + \varepsilon_i \quad \text{où } \varepsilon_i \text{ résidu (théorique)}}$$

Sous forme matricielle : $\mathbf{y = X\beta + \varepsilon}$

- $y = (y_1 \ldots y_n)^T$ vecteur colonne des réponses de taille n
- X matrice de taille $n \times (p+1)$
- $\beta = (\beta_0 \ldots \beta_p)^T$ de taille (p+1) : paramètres pour la **réponse espérée**
- $\varepsilon = (\varepsilon_1 \ldots \varepsilon_n)^T$ vecteur colonne des **résidus théoriques** ou **erreurs** de régression (**bruit**)

Hypothèses par défaut des logiciels de Statistique ou de « data science » :

$\varepsilon_1, \ldots, \varepsilon_n$ réalisations de **variables aléatoires indépendantes de même loi** supposée normale $N(0, \sigma^2)$ de densité :

$$x \to \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{x^2}{2\sigma^2}\right)$$

☹ **Une difficulté : $\varepsilon_1, \ldots, \varepsilon_n$ ne sont pas observés directement !**

Toute l'analyse et compréhension du modèle repose sur le fait que le vecteur

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \in \mathbb{R}^n$$

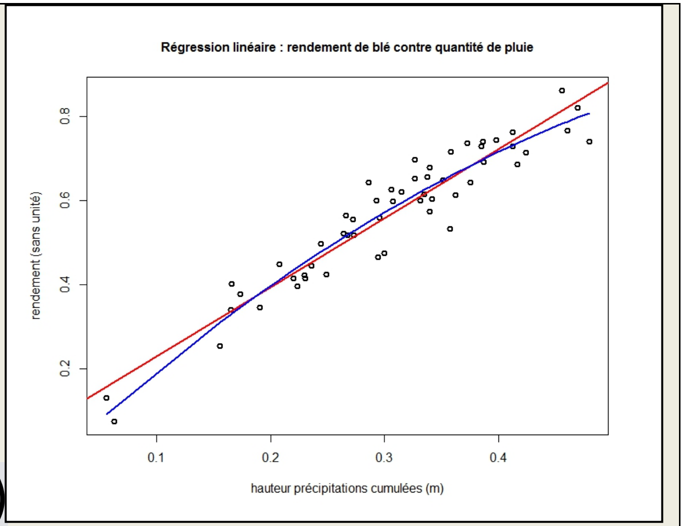
est un **vecteur aléatoire de loi gaussienne n-dimensionnelle**

☛ **Y est appelé vecteur gaussien (VG)**

```
> summary(lm(rend ~ pluie + pluie2))

Residuals:
      Min        1Q    Median        3Q       Max
-0.125759 -0.031894 -0.000287  0.035384  0.093880

Coefficients:
              Estimate    Std. Error    t value   Pr(>|t|)
(Intercept)   -0.04246    0.04512       -0.941    0.35109
pluie         2.50171     0.31868       7.850 2. 49e-10 ***
pluie2        -1.52278    0.54701       -2.784    0.00752 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 0.04893 on 51 degrees of freedom
Multiple R-squared:  0.914,     Adjusted R-squared:  0.9106
F-statistic:   271 on 2 and 51 DF,  p-value: < 2.2e-16
```



☞ le calcul des probabilités et la Statistique sont au cœur de la science des données
☞ tous ces résultats résultent des propriétés fondamentales des vecteurs gaussiens et du TCL multi-dimensionnel !

X v.a.r. est dite de loi *normale* (ou *gaussienne*) de *moyenne* $\mu \in \mathbb{R}$ et de *variance* $\sigma^2 > 0$ (*écart-type* $\sigma > 0$) si sa *loi de probabilité* admet la *densité* suivante :

$$\mathbf{f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(- \frac{(x-\mu)^2}{2\sigma^2} \right)}$$

Ainsi,

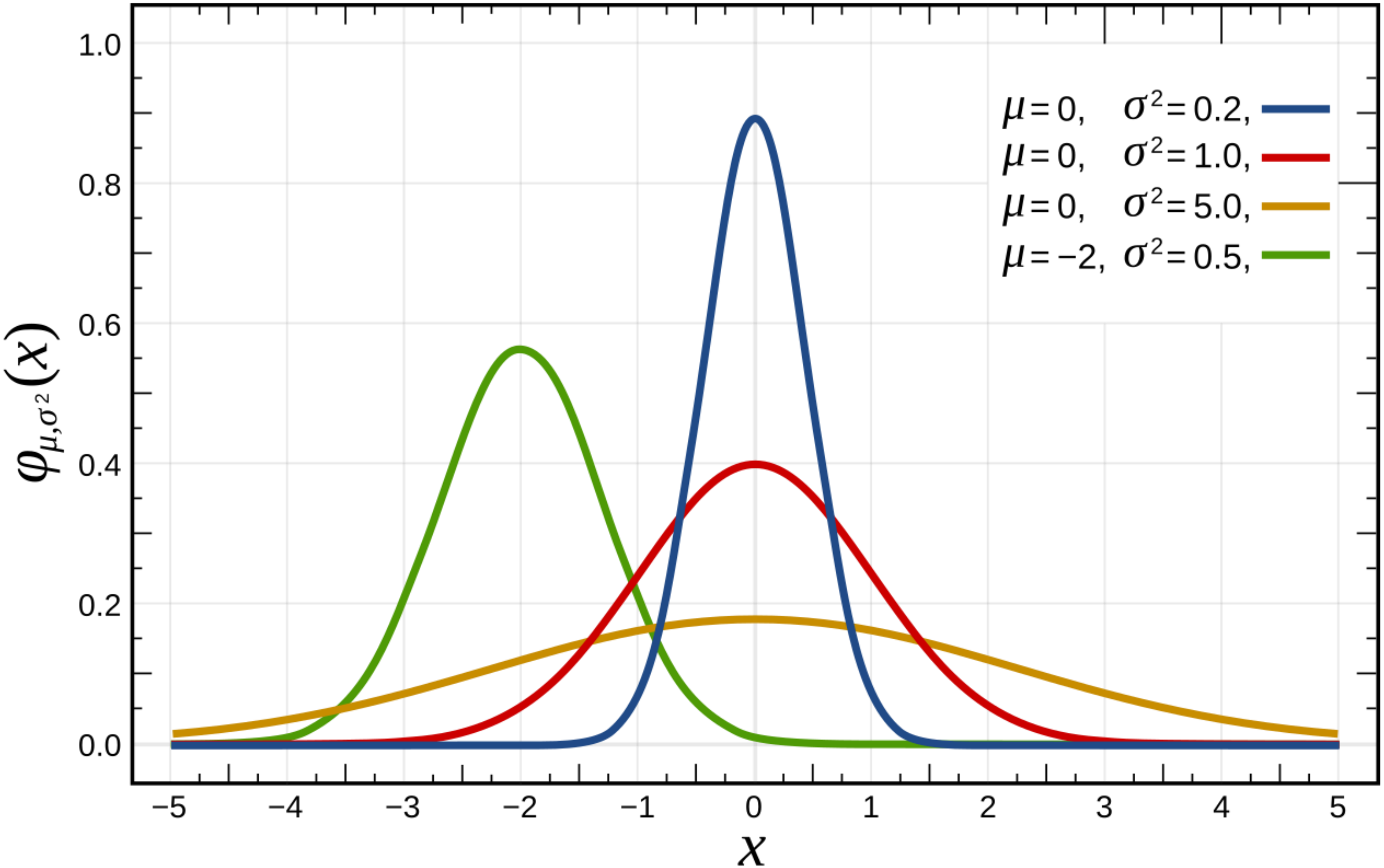
$$P(X \in [a, b]) = \int_a^b f_X(x)\,dx$$

On notera $X \sim N(\mu, \sigma^2)$.

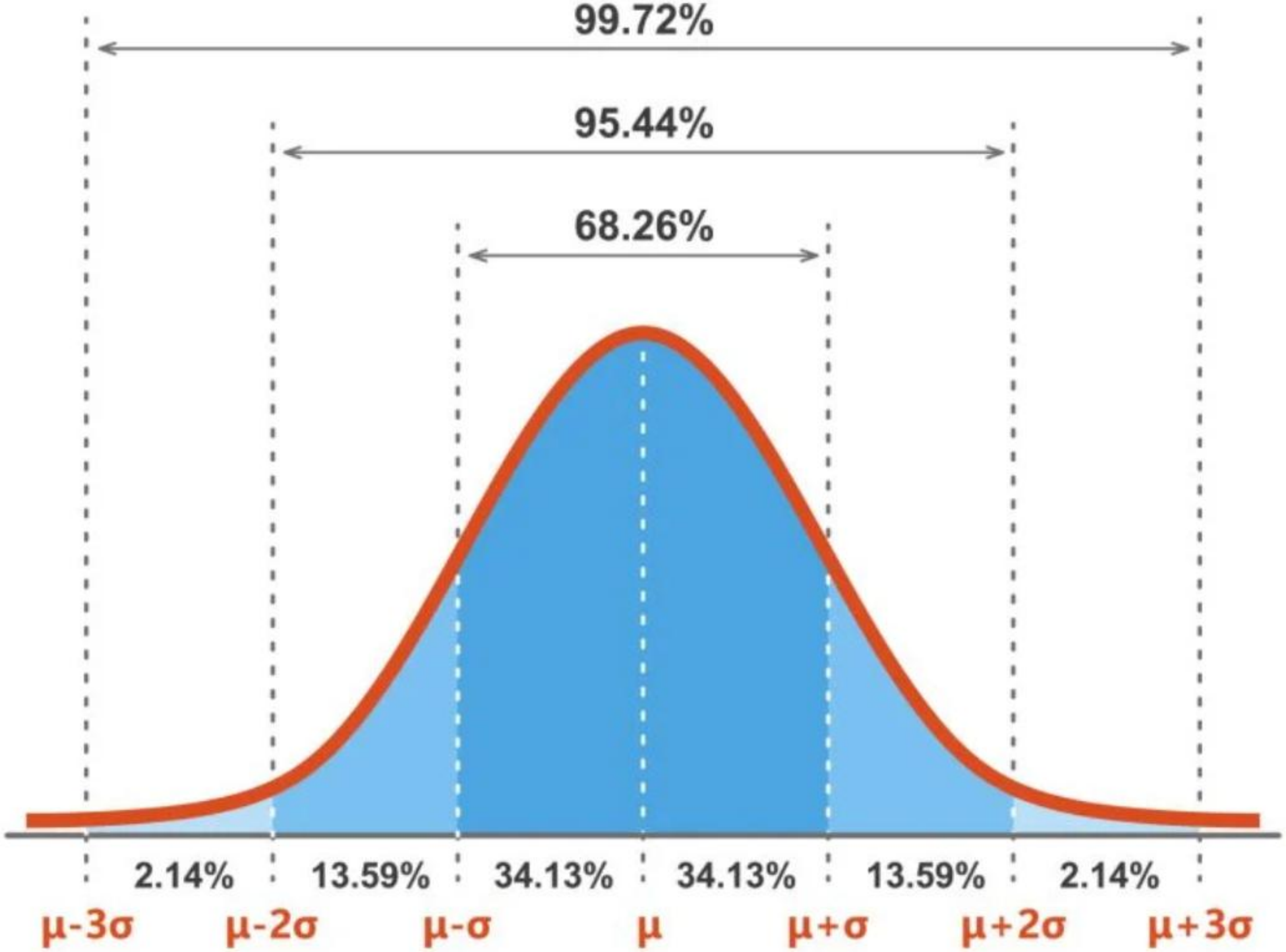
Convention. $X = \mu$ est gaussienne dégénérée, i.e. $X \sim N(\mu,0)$.

Propriété : $X = \mu + \sigma Z$ où Z est de loi normale N(0, 1), soit la loi normale dite centrée-réduite

I.1. Loi gaussienne sur $\mathbb{R}$



I.1. Loi gaussienne sur IR



Importance de la loi normale – Théorème Central Limite (TCL):

$X_1, X_2, \ldots$ **indépendantes et identiquement distribuées** (en abrégé i.i.d.) de moyenne μ et de variance σ^2

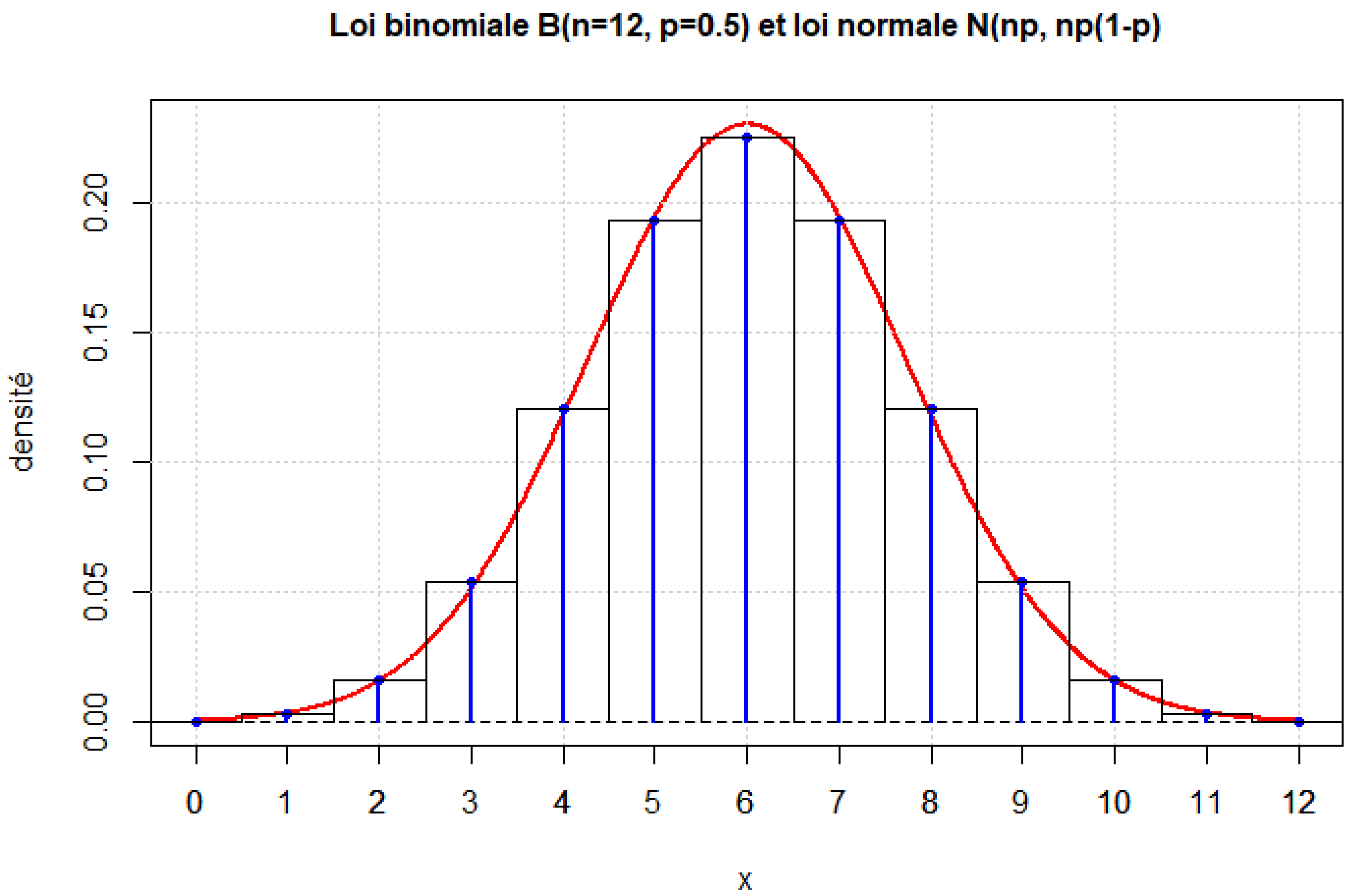
Considérons la suite des sommes partielles : $\mathbf{S_n = X_1 + \ldots + X_n}$ $(n \geq 1)$.

Alors, pour tout $a < b$:

$$\mathbf{P\left(a \leq \frac{\sqrt{n}}{\sigma} \times \left(\frac{S_n}{n} - \mu \right) \leq b \right) = P\left(a \leq \frac{S_n - n\mu}{\sigma\sqrt{n}} \leq b \right) \to_n \int_a^b \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2} \right) dx}$$

En pratique : $\frac{S_n}{n} \approx_{n \text{ grand}} N(\mu, \frac{\sigma^2}{n}) \Leftrightarrow S_n \approx_{n \text{ grand}} N(n\mu, n\sigma^2)$

Si les X_k sont $N(\mu, \sigma^2)$, alors $\frac{S_n}{n}$ est de loi $N(\mu, \frac{\sigma^2}{n})$ et S_n est de loi $N(n\mu, n\sigma^2)$.



Loi binomiale B(n=12, p=0.5) et loi normale N(np, np(1-p)
densité
0.00
0.05
0.10
0.15
0.20
0
1
2
3
4
5
6
7
8
9
10
11
12
x

Définition. $X = (X_1, \ldots, X_d) \in \mathbb{R}^d$ où $X_1, \ldots, X_d$ sont des v.a. est appelé **vecteur aléatoire réel** (V.a.r.) de dimension d (ou V.a.r. d-dimensionnel).

Définition (V.a. continu). $X = (X_1, \ldots, X_d) \in \mathbb{R}^d$ est dit **continu** de densité :

$$\boldsymbol{x = (x_1, \ldots, x_d) \in \mathbb{R}^d \to f_X(x)}$$

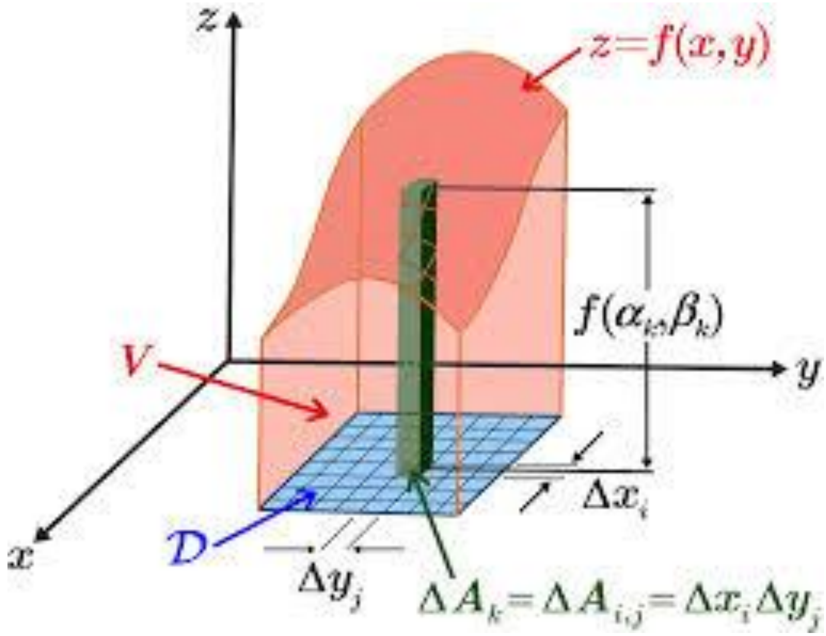
si (par définition)

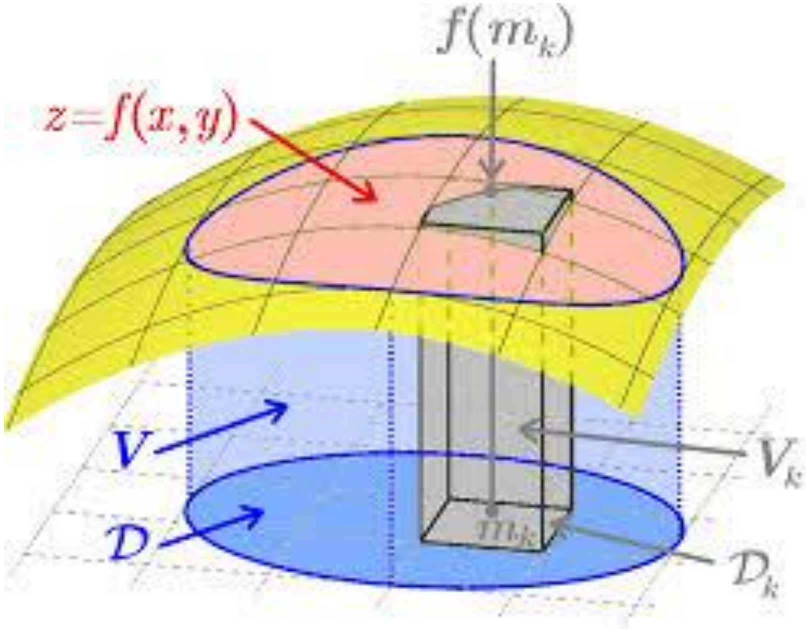
$$\mathbf{P(X \in [a_1, b_1]\times\ldots\times[a_d, b_d]) = \int_{[[a,\, b]]} f_X(x)dx}$$

où $[[a, b]]$ désigne un hyper-rectangle quelconque $[a_1, b_1]\times\ldots\times[a_d, b_d]$ de $\mathbb{R}^d$.

Remarque : $\mathbf{f_X \geq 0}$ **et** $\displaystyle \int_{\mathbb{R}^d} \mathbf{f_X(x)dx = 1}$

Intégrale multiple (comme limite de sommes de Riemann)



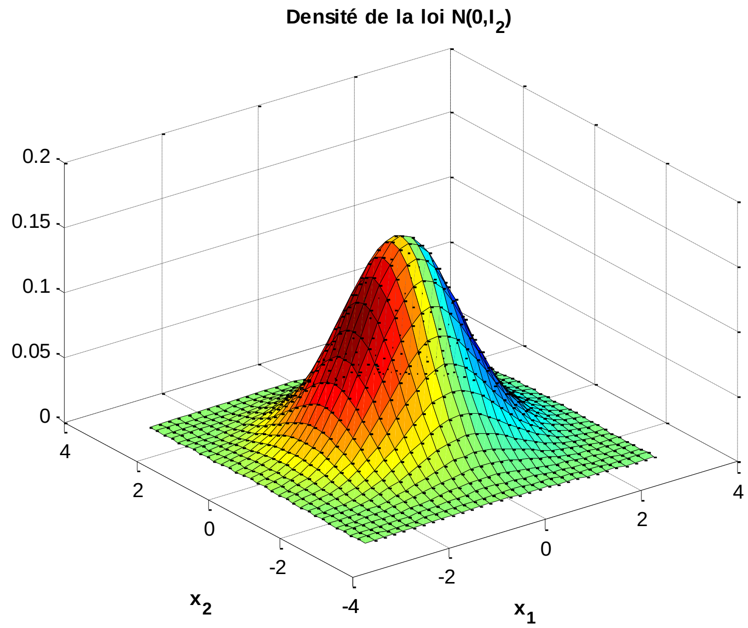


Cas particulier : si $f(x) = f_1(x_1) \times \dots \times f_d(x_d)$, alors :

$$\int_{[[a,\, b]]} \mathbf{f(x)dx} = \int_{a_1}^{b_1} f_1(x_1)dx_1 \times \dots \times \int_{a_d}^{b_b} f_d(x_d)dx_d$$

Exemple simple : $X = (X_1, X_2)$ où X_1, X_2 sont i.i.d. de loi $N(0, 1)$, alors X admet la densité de probabilité suivante :

$$\boldsymbol{x = (x_1, x_2) \rightarrow \frac{1}{2\pi} \exp\left(-\frac{\|x\|^2}{2}\right)}$$



☞ cloche de Gauss bidimensionnelle

Lois marginales d'un V.a.r. : ce sont les lois des composantes X_1, …, X_d.

Dans le **cas continu**, ces composantes admettent des densités f_{X_1}, …, f_{X_d} données par

$$\mathbf{f_{X_1}(\boldsymbol{x_1}) = \int_{\mathrm{IR}^{d-1}} f_X(\boldsymbol{x_1}, \ldots, \boldsymbol{x_d})\, d\boldsymbol{x_2} \times \ldots \times d\boldsymbol{x_d}\ ,\ \ etc.}$$

Cas indépendance : X = X = (X_1, …, X_d) V.a. **continu** alors

$$X_1, \ldots, X_d \text{ indépendantes} \Leftrightarrow f_X(x) = f_{X_1}(x_1) \times \ldots \times f_{X_d}(x_d)$$

Définition – moyenne et matrice de (variance-)covariance d'un V.a.r. X d-dimensionnel :

- $\mu = (EX_1, \ldots, EX_d) \in \mathbb{R}^d$ **moyenne de X : $\mu = E(X)$**

- $\Gamma = (\ E[(X_i - \mu_i)(X_j - \mu_j)]\ )_{i,\,j}$ matrice de taille $d \times d$ appelée **covariance de X : $\Gamma = \mathrm{Cov}(X)$** (matrice des variances-covariances du vecteur X)

Convention usuelle : tout vecteur de $\mathbb{R}^d$ est vu comme un vecteur colonne de taille d (matrice de taille $d \times 1$).

Moyenne d'une matrice aléatoire : $E(M) = (EM_{i,j})$

Ainsi, en langage matriciel, $\boldsymbol{\mu = E(X)}$ **et** $\boldsymbol{\Gamma = E[\ (X - \mu)(X - \mu)^T\ ]}$

Propriété 1. La matrice Γ de terme général $\mathbf{\Gamma_{ij} = Cov(X_i, X_j)}$ est une matrice **symétrique positive**. Sa diagonale est formée des variances individuelles des variables : $\mathbf{\Gamma_{ii} = Cov(X_i, X_i) = Var(X_i)}$.

Propriété 2. Si A est une application linéaire de $\mathbb{R}^d$ dans $\mathbb{R}^d$ (identifiée à une matrice d×d) et b un vecteur de $\mathbb{R}^d$, le vecteur aléatoire $\mathbf{Y = AX + b}$ vérifie

$$\boldsymbol{\mu_Y = A\mu_X + b \text{ et } \Gamma_Y = A\Gamma_X A^T}$$

où $\mu_Y = E(Y)$ et $\Gamma_Y = Cov(Y)$ (se généralise à A matrice rectangulaire).

Propriété 3. Si $\Gamma = Cov(X)$ est de rang r, alors $X \in Im(\Gamma)$ sous-espace vectoriel de dimension r. Dit autrement, X a seulement **r degrés de liberté**. Si donc r < d, X n'admet pas de densité définie sur $\mathbb{R}^d$.

Définition d'un Vecteur Gaussien (VG) : soit $\mu \in \mathbb{R}^d$ et Γ une matrice **symétrique définie positive** de taille d×d (le rang de Γ est d). Un vecteur aléatoire réel $X = (X_1, \ldots, X_d) \in \mathbb{R}^d$ est dit **gaussien** (ou multi-gaussien) de moyenne μ et de covariance Γ s'il admet la densité dite (multi-)gaussienne ou normale suivante :

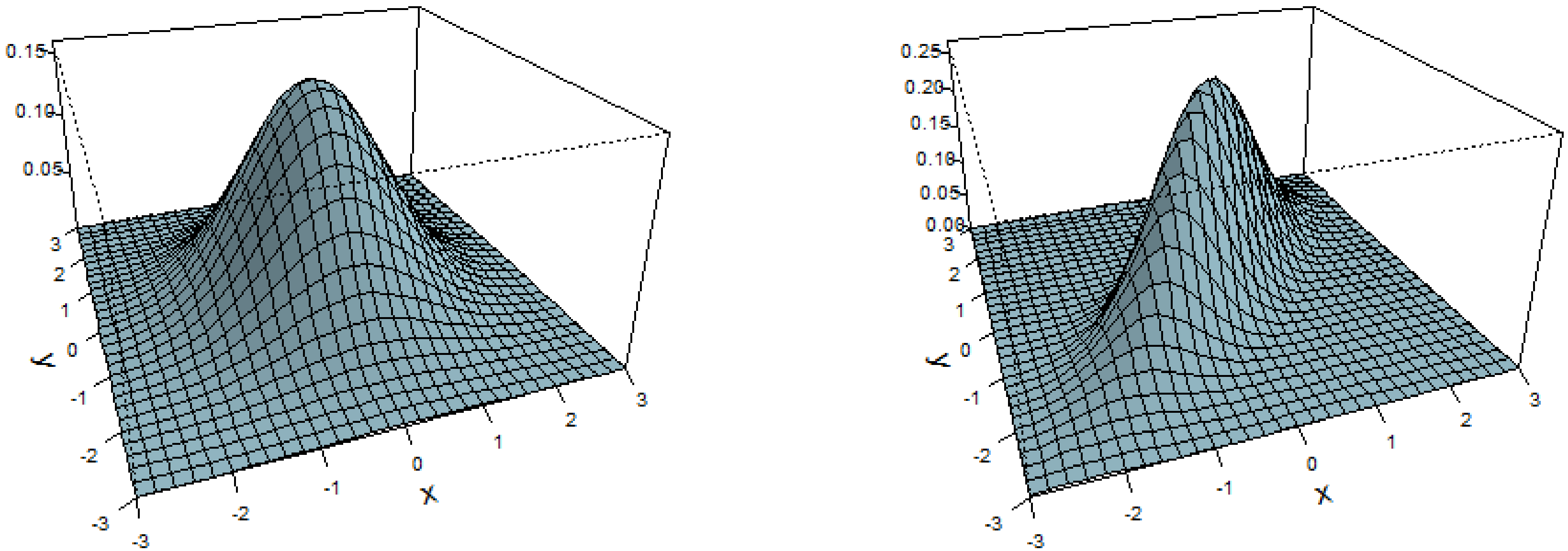
$$\boldsymbol{x \to \frac{1}{\sqrt{\det(\Gamma)}\ \sqrt{2\pi}^{\,d}} \exp\left(-\frac{1}{2} < x-\mu \mid \Gamma^{-1}(x-\mu) > \right)}$$

Cette expression définit bien une densité associée à un vecteur aléatoire de moyenne μ et de covariance Γ et on a la décomposition $\mathbf{X = \mu + U}\Sigma Z$ avec les variables $\mathbf{Z_1, \ldots, Z_d}$ **i.i.d. N(0, 1)** et où les matrices Σ et U sont obtenues par diagonalisation en base orthonormée de la matrice de covariance :

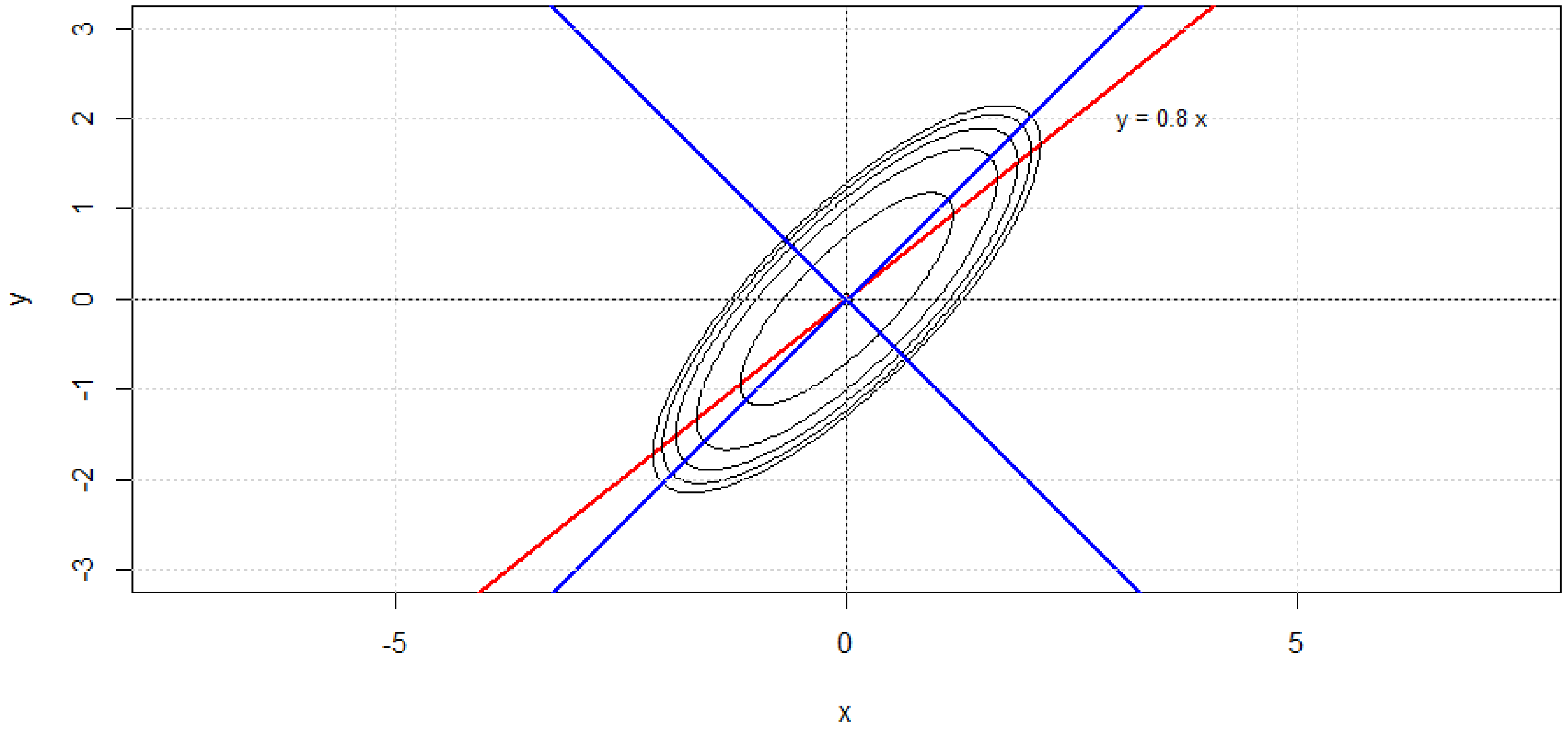
$$\mathbf{\Gamma = U\Lambda U^T = U\Sigma^2 U^T}\ \ (\Sigma = \sqrt{\Lambda})\ \text{avec}\ \mathbf{U^T U = I_d}$$

On notera **N(μ, Γ)** la loi normale multivariée de moyenne μ et de covariance Γ.

Exemples



$\mu = 0$; Var(X) = Var(Y) = 1 ; $\rho = 0$ (à gauche) et $\rho = 0.8$ (à droite)



Courbes d'iso-densité du cas ρ = 0.8

Théorème Central Limite multidimensionnel

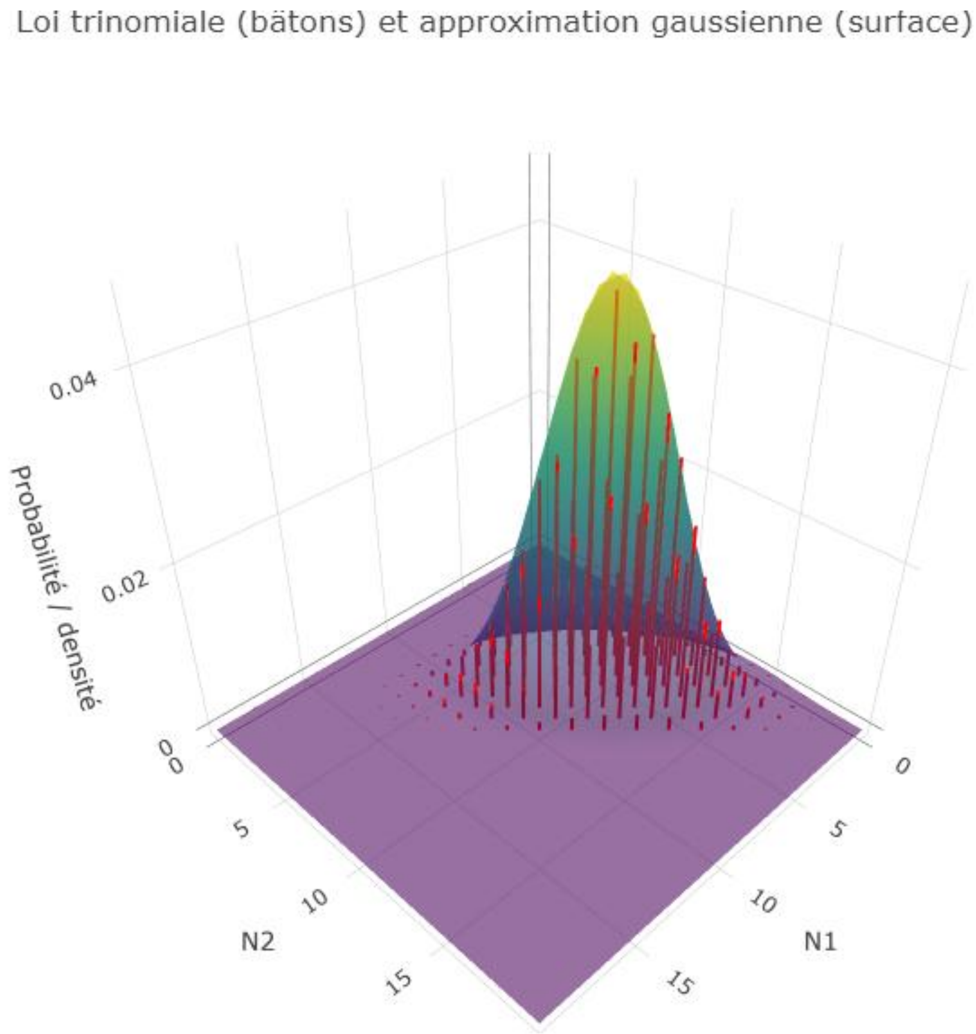
Soit $X^{(1)}$, $X^{(2)}$, … une suite de vecteurs aléatoires **indépendants et identiquement distribués** dans $\mathbb{R}^d$ de vecteur moyenne μ et de matrice de covariance Γ supposée **inversible**. Pour n ≥ 1, on note encore $\mathbf{S_n = X^{(1)} + \ldots + X^{(n)}} \in \mathbb{R}^d$.

Alors, pour tout hyper-rectangle [[a, b]] de $\mathbb{R}^d$

$$\mathbf{P}\left(\sqrt{\mathbf{n}} \times \left(\frac{\mathbf{S_n}}{\mathbf{n}} - \mu\right) = \frac{\mathbf{1}}{\sqrt{\mathbf{n}}} \times (\mathbf{S_n} - \mathbf{n}\mu) \in [[\mathbf{a}, \mathbf{b}]]\right) \to_{\mathbf{n}} \int_{[[\mathbf{a}, \mathbf{b}]]} \mathbf{f_X}(\boldsymbol{x})\mathbf{d}\boldsymbol{x}$$

où f est la densité de la loi gaussienne multivariée N(0, Γ).

En pratique : $$\frac{\mathbf{S_n}}{\mathbf{n}} \approx \mathbf{N}\left(\mu, \frac{\mathbf{1}}{\mathbf{n}}\Gamma\right)$$



Paramètres de la loi trinomiale :

n = 20
p_1 = 0.3
p_2 = 0.5

(et donc p_0 = 0.2)

Définition. X est dit *gaussien centré-réduit* si **X ~ N(0, $\mathbf{I_d}$)** où $\mathbf{I_d}$ désigne la matrice *identité* de taille d. S**es composantes X_1, …, X_d sont indépendantes et de loi N(0, 1)**. Sa densité est :

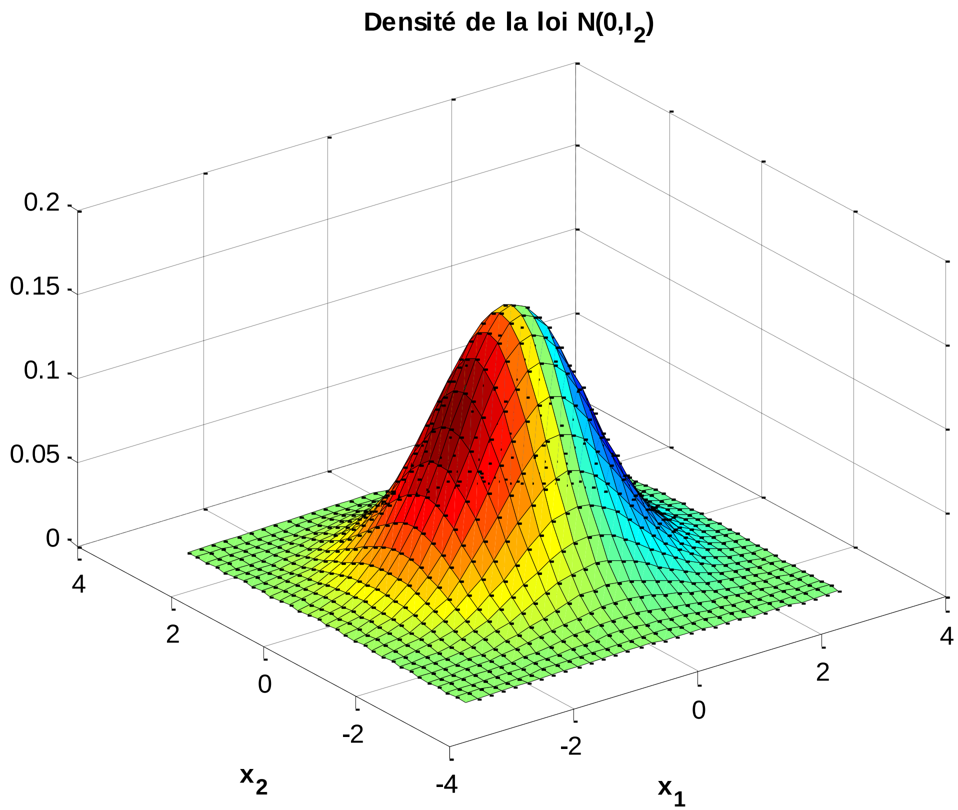
$$\boldsymbol{x} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_d)^T \to \frac{1}{\sqrt{2\pi}^{\,d}} \exp\left(-\frac{\|\boldsymbol{x}\|^2}{2}\right)$$

où $\|x\| = \sqrt{x_1^2 + \ldots + x_d^2}$ désigne la norme euclidienne usuelle de $\mathbb{R}^d$.

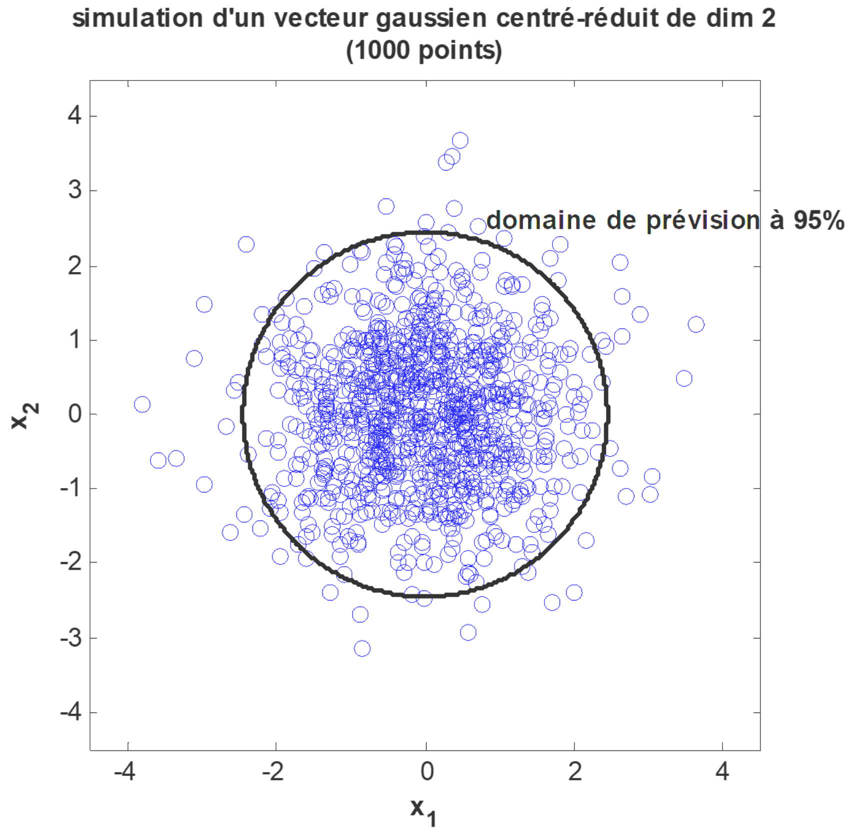
Interprétation. Les composantes du vecteur **X** forment un bruit « blanc », ce qui renvoie aussi à la terminologie de **bruit blanc gaussien** (standard).

Dans toute la suite, ε désignera un vecteur aléatoire de loi **N(0, $\mathbf{I_d}$)**.

Cas de la dimension d = 2 :



I.3. Vecteur gaussien centré-réduit



Propriété fondamentale de ε gaussien centré-réduit :

Soit U une **isométrie** (linéaire) : $U^T U = I_d$.

- $\varepsilon \sim N(0, I_d) \Rightarrow U\varepsilon \sim N(0, I_d)$

- $Z = U^T\varepsilon$ coordonnées (euclidiennes) du vecteur ε dans la base orthonormée associée aux vecteurs colonne $u_1, \ldots, u_d$ de la matrice U, alors :

$$\varepsilon = Z_1 u_1 + \ldots + Z_d u_d$$

avec les (nouvelles) variables $Z_1 + \ldots + Z_d$ i.i.d. N(0, 1)

THÉORÈME DE COCHRAN*

Soit P un projecteur orthogonal, i.e. $P^2 = P$ et $P^T = P$. Alors :

- **les vecteurs aléatoires Pε et (I – P)ε** sont **indépendants**, **de degrés de liberté p et d – p où p = dim(ImP**).

- Les variables aléatoires $||P\varepsilon||^2$ et $||(I - P)\varepsilon||^2$ sont **indépendantes** de lois respectives

$$\boldsymbol{||P\varepsilon||^2 \sim \chi_p^2 \text{ et } ||(I - P)\varepsilon||^2 \sim \chi_{d-p}^2}$$

(*William Gemmel Cochran, 1909-1980)

Définition générale d'un vecteur gaussien

Un vecteur aléatoire $X = (X_1, \ldots, X_p) \in \mathbb{R}^p$ est dit **gaussien** s'il existe $\mu \in \mathbb{R}^p$, une matrice A de taille $p \times q$ et $\varepsilon \sim N(0, I_q)$ tels que

$$X = \mu + A\varepsilon$$

X image par une **transformation affine** de $\mathbb{R}^q$ dans $\mathbb{R}^p$ d'un VG centré-réduit de dimension q.

Moyenne et matrice de covariance :

$$E(X) = \mu$$

$$Cov(X) = AA^T$$

Densité et décomposition de Prasanta Mahalanobis (1893-1972)

- $\Gamma = \mathrm{Cov}(X) = AA^T$ est **inversible**, alors X admet la densité

$$\boldsymbol{x \to \frac{1}{\sqrt{\det(\Gamma)}\ \sqrt{2\pi}^{\,p}} \exp\left(-\frac{1}{2} \langle x-\mu, \Gamma^{-1}(x-\mu)\rangle\right)}$$

et

$$\mathbf{X} = \boldsymbol{\mu} + \sqrt{\boldsymbol{\lambda_1}}\, \mathbf{Z_1 U_1} + \ldots + \sqrt{\boldsymbol{\lambda_p}}\, \mathbf{Z_p U_p}\ ;\ \mathbf{Z} \sim \mathbf{N(0, I_p)}$$

en écrivant que $\mathbf{\Gamma = U\Lambda U^T}$ où $U^TU = I_p$ et $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p \geq 0$ les valeurs propres de Γ.

- Γ est **non inversible**, X est dit « dégénéré » et « vit » en réalité dans un sous-espace affine de dimension $d < p$ (et $d \leq q$). De manière plus précise, on a

$$\mathbf{X} = \boldsymbol{\mu} + \sqrt{\boldsymbol{\lambda_1}}\, \mathbf{Z_1 U_1} + \ldots + \sqrt{\boldsymbol{\lambda_d}}\, \mathbf{Z_d U_d}\ ;\ \mathbf{Z} \sim \mathbf{N(0, I_d)}$$

Dans les deux cas, on dira que X est de loi N(μ, Γ).

I.4. Vecteurs gaussiens quelconques

CAS DE L'INDÉPENDANCE. ☺

Si $X \in \mathbb{R}^d$ est un vecteur gaussien, alors

$$\mathbf{X_1, \ldots, X_d \text{ indépendantes} \Leftrightarrow Cov(X_i, X_j) = 0 \text{ pour } i \neq j}$$
$$\mathbf{\Leftrightarrow \Gamma \text{ matrice diagonale}}$$

SIMULATION DE $X \sim N(\mu, \Gamma)$.

- Faire la décomposition de Cholesky de Γ: $\Gamma = SS^T$ avec S triangulaire inférieure (et diagonale positive)
- Simuler $\varepsilon \sim N(0, I_d)$
- Calculer $X = \mu + S\varepsilon$

Quelques propriétés à retenir pour la pratique :

- $X = \mu + A\varepsilon$ avec $\varepsilon \sim N(0, I_q)$ est un VG : $X \sim N(\mu, AA^T)$

- toute transformation affine d'un VG est un VG. En particulier, les composantes d'un VG sont toutes de loi normale sur $\mathbb{R}$

- si $X = (X_1, \dots, X_p)$ est un VG de loi $N(\mu, \Gamma)$ et u un vecteur de $\mathbb{R}^p$, alors

  $u^TX = u_1X_1 + \dots + u_pX_p$ est de variance $\mathbf{u^T\Gamma u}$, donc de loi $\mathbf{N(u^T\mu, u^T\Gamma u)}$

- indépendance = non corrélation dans le cas gaussien

Décomposition en valeurs singulières (DVS ou SVD):

A étant donnée de taille p×q, il existe une matrice p×q Σ diagonale à termes diagonaux ≥ 0 telle que

$$\mathbf{A = U\Sigma V^T} \quad \text{avec} \qquad \mathbf{V^T V = I_q} \text{ et } \mathbf{U^T U = I_p}$$

Ainsi, $\quad \mathbf{\Gamma = AA^T = U\Lambda_p U^T} \qquad$ **avec** $\mathbf{\Lambda_p} = \Sigma\Sigma^{\mathbf{T}}$ **diagonale** p×p

$$\mathbf{A^T A = V\Lambda_q V^T} \qquad \textbf{avec } \mathbf{\Lambda_q} = \Sigma^{\mathbf{T}}\Sigma \textbf{ diagonale } q\times q$$

Les termes diagonaux de Σ sont appelés ***valeurs singulières*** de la matrice A et sont le plus souvent rangées par ordre décroissant.

Soient Y v.a.r. et X v.a. **discrète** $X \in \{ x_1, x_2, \ldots, x_p\}$.

On appelle *loi conditionnelle de Y sachant* $X = x_k$ la loi de probabilité notée $\mathbf{P_{Y|X=x_k}}$ définie par

$$\mathbf{P_{Y|X=x_k}([a, b]) = P(Y \in [a, b] \mid X = x_k) = \frac{P(Y \in [a, b] \text{ et } X = x_k)}{P(X = x_k)}}$$

L'ensemble des p lois de probabilité $P_{Y|X=x_k}$ $(1 \leq k \leq p)$ est appelé *système des lois conditionnelles de Y sachant X.*

Exemple 1. Loi du pile ou face deux fois : Y = nombre total de pile, X résultat du premier lancer

Cas où Y est également discrète : $Y \in \{ y_1, y_2, \ldots, y_q\}$

les p lois conditionnelles de Y sachant X s'écrivent

$$P_{Y|X=x_i} = \sum_{j=1}^{q} p_{j|i}\,\delta_{y_j} \text{ où } p_{j|i} = \frac{p_{ij}}{p_{i.}} = \frac{P(Y = y_j \text{ et } X = x_i)}{P(X = x_i)}$$

On a (!)

X, Y indépendantes $\Leftrightarrow P_{Y|X=x_i} = P_Y$ pour tout i

$$\Leftrightarrow p_{j|i} = \frac{P(Y = y_j \text{ et } X = x_i)}{P(X = x_i)} = P(Y = y_j) = p_{\bullet j}$$

II.2. Lois conditionnelles dans le cas continu

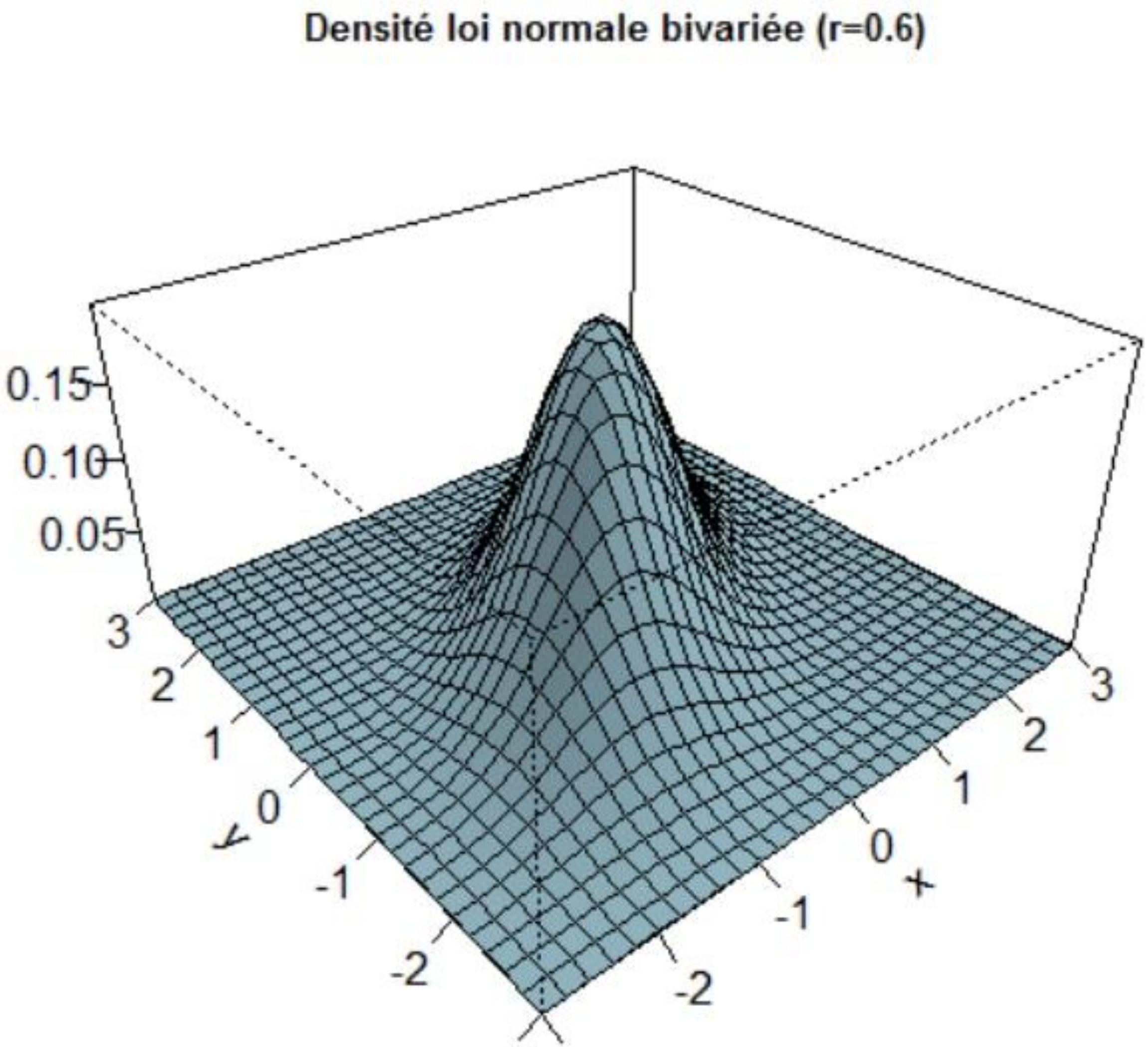
$(X, Y) \in \mathbb{R}^2$ **est un vecteur continu de densité** $(x, y) \rightarrow f_{(X, Y)}(x, y)$

La loi conditionnelle de Y sachant X = x est définie par la *densité conditionnelle* suivante

$$\boxed{f_{Y \mid X = x}(y) = \frac{f_{(X, Y)}(x, y)}{f_X(x)}}$$

pour tout x tel que $f_X(x) = \int_{-\infty}^{+\infty} f_{(X, Y)}(x, y)dy \neq 0.$

II.2. Lois conditionnelles dans le cas continu



Exemple 2. $Y = X + \varepsilon$ où $X \sim N(\mu_X, \sigma_X^2)$ et $\varepsilon \sim N(0, \sigma^2)$ sont indépendantes

$$\mathbf{X, Y \text{ indépendantes} \Leftrightarrow P_{Y|X=x} = P_Y \Leftrightarrow f_{Y\,|\,X\,=\,x}(y) = \frac{f_{(X,\,Y)}(x, y)}{f_X(x)} = f_Y(y)}$$

FORMULE DE BAYES.

$$f_{X\,|\,Y\,=\,y}(x) = \frac{f_{Y\,|\,X\,=\,x}(y) \times f_X(x)}{f_Y(y)}$$

sachant que $f_Y(y) = \int_{-\infty}^{+\infty} f_{Y\,|\,X\,=\,x}(y) \times f_X(x)\,dx$ **est simplement une constante de normalisation.**

II.3. Espérance conditionnelle

On a vu que l'information apportée par $X = (X_1, \ldots, X_p)$ modifie (si non indépendance) les probabilités que l'on a sur les valeurs prises par Y avec la notion de loi conditionnelle.

Il est tout autant naturel de se demander en quoi la connaissance des variables $X_1, \ldots, X_p$ réduit ou non l'incertitude que l'on a sur Y si l'on mesure cette incertitude en terme de dispersion usuelle, i.e. de variance.

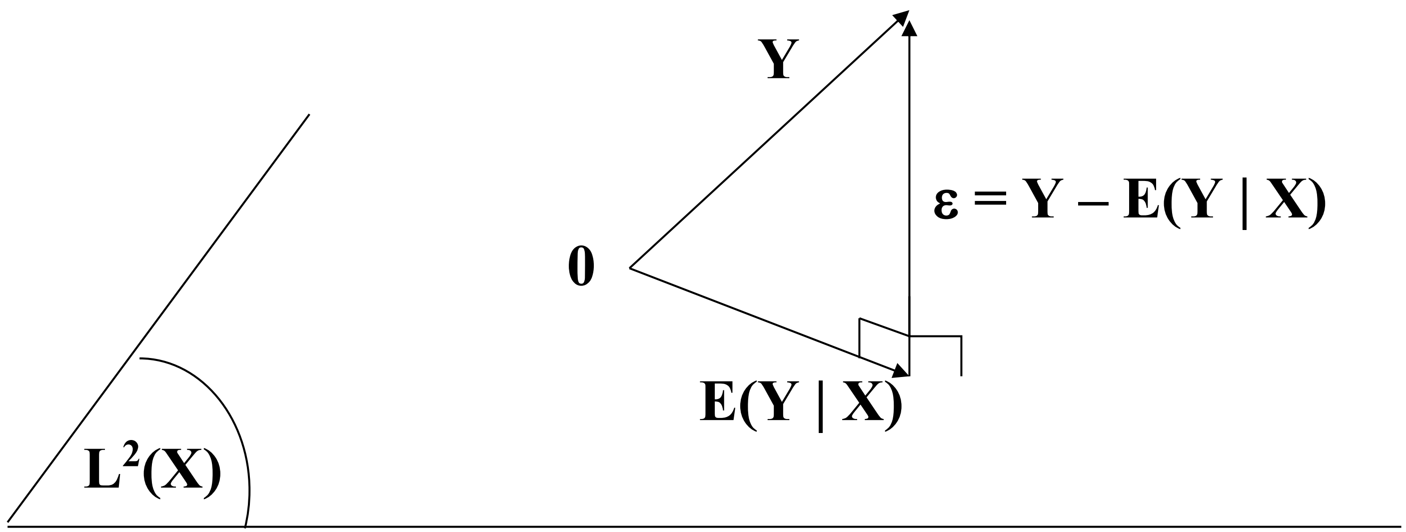
☺☺☺ On introduit pour cela l'espace vectoriel $L^2(P)$ des v.a. admettant une variance et muni du produit scalaire suivant :

$$<U \mid V> = E(UV)$$

Définition de l'*espérance conditionnelle* E(Y | X) :

E(Y | X) est **la projection orthogonale de Y sur le sous-espace $L^2(X)$** de $L^2(P)$ défini par

$$L^2(X) = \{ Z = \varphi(X) \in L^2(P) \text{ avec } \varphi : \mathbb{R}^p \to \mathbb{R} \}.$$



Bien noter que E(Y | X) est une variable aléatoire!

Formules de l'espérance totale et de la variance totale

$$\mathbf{Y = E(Y \mid X) + \varepsilon} \qquad \text{avec} \quad \mathbf{E(\varepsilon) = 0} \text{ et :}$$

$$\mathbf{Cov(\varepsilon, \varphi(X)) = < \varepsilon, \varphi(X) > = 0 \text{ si } \varphi(X) \in L^2(X)}$$

En particulier : $\mathbf{E(E(Y \mid X)) = E(Y)}$ **et** $\mathbf{Var(Y) = Var(E(Y \mid X)) + Var(\varepsilon)}$
Dans cette décomposition, **ε est de variance minimale** : on ne peut pas réduire plus la variance résiduelle!

Définition. On définit le pourcentage de la variance de Y expliquée par X selon

$$\textbf{\% de variance expliquée} = 100 \times \frac{Var(E(Y \mid X))}{Var(Y)} \ \%$$

$$\textbf{\% de variance non expliquée ou résiduelle} = 100 \times \frac{Var(\varepsilon)}{Var(Y)} \ \%$$

Espérance conditionnelle comme moyenne de la loi conditionnelle. On a simplement

$$\mathbf{E(Y \mid X) = m(X)}$$

où la fonction m vérifie

$$\mathbf{m(x) = \int_{-\infty}^{+\infty} y \times f_{Y \mid X = x}(y)dy = E(Y \mid X = x)}$$

i.e. m (x) est la moyenne de la loi conditionnelle de Y sachant X = x.

De plus, on a **la formule de transfert conditionnel**

$$\mathbf{E(\, f(Y) \mid X) = \varphi(X) \text{ où } \varphi \text{ se calcule selon } \varphi(x) = \int f(y) f_{Y \mid X = x}(y)dy}$$

Exercice 7. Soit (X, Y) un vecteur gaussien bidimensionnel. On note **E_L(Y | X)** la projection orthogonale de Y sur $F = \mathrm{ev}\{1, X\} := \{\beta_0 + \beta_1 X : \beta_0, \beta_1 \text{ scalaires}\}$.

On dit encore que $E_L(Y \mid X)$ est la **régression linéaire** de Y sur X.

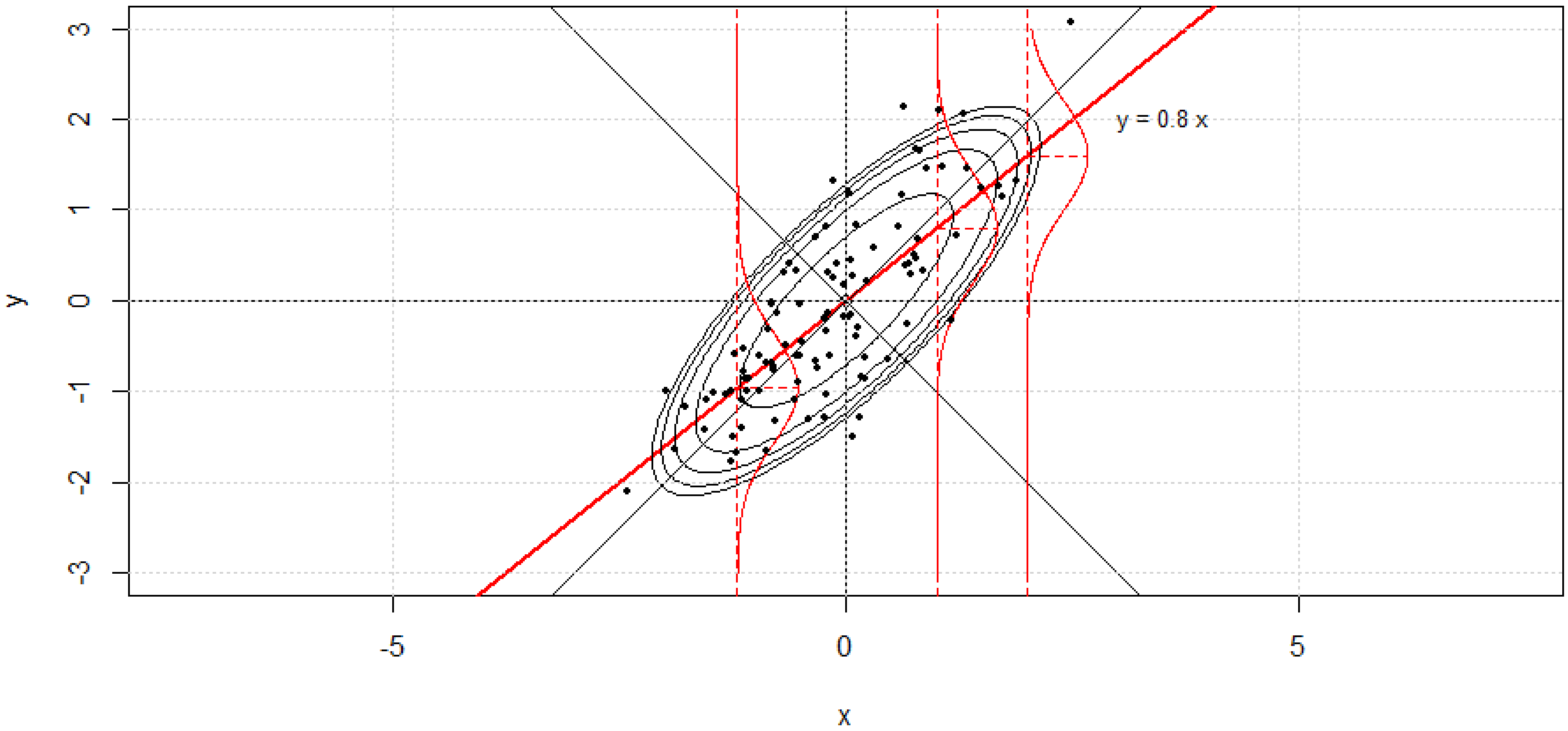
1. Calculer $E_L(Y \mid X)$.

2. On considère la décomposition orthogonale $Y = E_L(Y \mid X) + \varepsilon$ où donc

$$\varepsilon = Y - E_L(Y \mid X).$$

Montrer que X et ε sont non corrélées puis indépendantes.

3. En déduire les lois conditionnelles de Y sachant $X = x$ pour tout x réel et illustrer le résultat obtenu dans le plan $0xy$ en faisant apparaître les courbes d'iso-densité du vecteur gaussien (X, Y).


y = 0.8 x
x
y
-5
0
5
-3
-2
-1
0
1
2
3

On généralise l'exercice précédent :

Si $(X, Y) = (X_1, \ldots, X_p, Y)$ est un vecteur gaussien, alors :

- **$E(Y \mid X) = E_L(Y \mid X)$** est une **fonction linéaire (affine)** des X_k
- Le résidu ε est **indépendant** des X_k et de loi **$N(0, Var(\varepsilon))$ où**

$$Var(\varepsilon) = Var(Y) - Var(E(Y \mid X))$$

- **La loi conditionnelle de Y sachant $X = x$ est gaussienne de variance indépendante de x et égale à la variance du résidu ε**