MNCS–Thème 3 : analyse de données, statistique descriptive et inférentielle

J. Lefrère

Université Pierre et Marie Curie

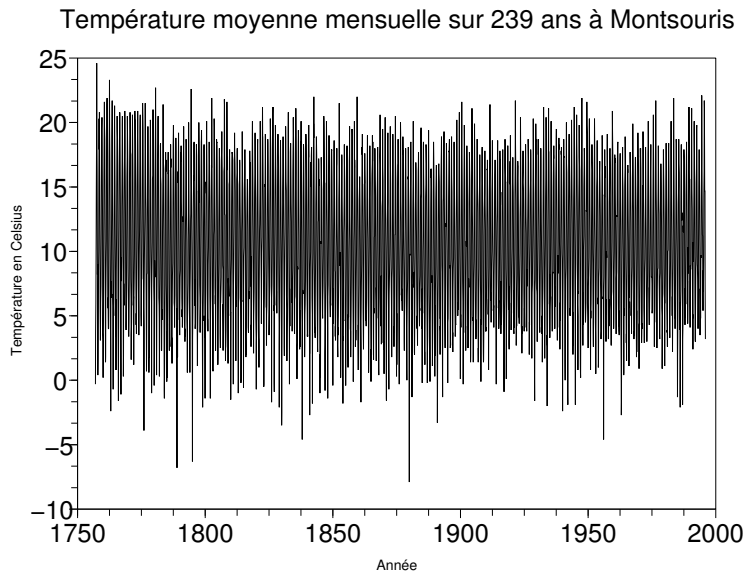
février 2015

Table des matières I

Table des matières II

Introduction

En physique, on est souvent amené à analyser de vastes séries de mesures, dont on cherche à synthétiser les propriétés statistiques.



- On fait alors appel à la **statistique descriptive** dont le but est de faire ressortir l'information contenue dans les données (par exemple par des méthodes de classification, d'analyse en composantes principales...) ;
- Souvent, on doit définir ces propriétés sans disposer de toutes les mesures possibles. On ne dispose dans ce cas que d'un échantillon des mesures et l'on cherche à estimer les propriétés de la variable aléatoire sous-jacente. On utilise alors les résultats de la **statistique inférentielle** qui s'appuie sur la théorie des probabilités.

Variables aléatoires

Une **variable aléatoire** (v.a.) X est un objet mathématique permettant de représenter une expérience ou une mesure dont le résultat n'est ou ne peut être connu exactement à l'avance.
On distingue deux types de variables aléatoires :

- variables **à valeurs discrètes** : tirage d'un dé, jet d'une pièce de monnaie, comptage d'événements (particules, photons, appels téléphoniques), etc.
- variables **à valeurs continues** : durée d'un trajet, durée de vie d'un composant, vitesse d'une molécule dans un gaz, température, etc.

Variables aléatoires discrètes

Dans le cas discret, à chaque valeur possible x_k de la variable X, on associe la probabilité de réalisation de l'événement $X = x_k$ soit $P(X = x_k) = p_k$.
Loi des grands nombres : cette probabilité peut être vue comme la fréquence d'occurrence de l'événement x_k dans la limite d'un nombre d'expériences infini.
P est appelé la **loi de probabilité de X** et vérifie $\sum_k P(X = x_k) = 1$.

Définir la **fonction de répartition F_X** de X par :

$$F_X(x) = P(X \leqslant x) \tag{1}$$

F_X est croissante, $\lim_{x \to -\infty} F_X(x) = 0$, $\lim_{x \to +\infty} F_X(x) = 1$ et $P(a < X \leqslant b) = F_X(b) - F_X(a)$.

Variables aléatoires continues

Dans le cas continu, la loi de probabilité peut admettre une **densité de probabilité** f_X (*pdf : probability density function*), et alors :

$$\boxed{P(x < X \leqslant x + dx) = f_X(x)\ \mathrm{d}x} \tag{2}$$

avec $\int_{-\infty}^{+\infty} f_X(x)\ \mathrm{d}x = 1$. N.-B. : dans le cas continu, $P(X = x) = 0$
La **fonction de répartition** (*cdf : cumulative density function*) est définie de la même manière que dans le cas continu, et dans le cas où f_X existe, F_X est la primitive de f_X : $F_X(-\infty) = 0$ et $F_X(+\infty) = 1$

$$\boxed{F_X(x) = \int_{-\infty}^{x} f_X(t)\ \mathrm{d}t} \tag{3}$$

$P(a < X \leqslant b) = F_X(b) - F_X(a) = \int_a^b f_X(t)\ \mathrm{d}t$.
N.-B. : Une loi discrète peut être représentée dans ce formalisme par une densité constituée de distributions de Dirac aux points x_k pondérées par leur probabilités p_k respectives.

$$f_X(x) = \sum_k p_k \delta(x - x_k) \tag{4}$$

La fonction de répartition présente alors des discontinuités p_k en ces points.

Moments d'une v.a.

On définit le **moment d'ordre** n d'une variable aléatoire X par :

$$m_n = \mathrm{E}\left(X^n\right) = \begin{cases} \displaystyle\sum_k x_k^n \, P(X = x_k) & \text{discret} \\ \displaystyle\int_{-\infty}^{+\infty} x^n f(x) \, \mathrm{d}x & \text{continu} \end{cases} \tag{5}$$

Moment d'ordre 1 = espérance mathématique $\mathrm{E}\left(X\right)$ (ou la moyenne)
On définit de même le **moment centré d'ordre** n par :

$$\mu_n = \mathrm{E}\left(\left[X - \mathrm{E}\left(X\right)\right]^n\right) = \begin{cases} \displaystyle\sum_k (x_k - m_1)^n \, P(X = x_k) & \text{discret} \\ \displaystyle\int_{-\infty}^{+\infty} (x - m_1)^n f(x) \, \mathrm{d}x & \text{continu} \end{cases} \tag{6}$$

Moment centré d'ordre 2 = **variance** $\mathrm{V}(X)$ de la v.a. X

$$\mathrm{V}(X) = \mathrm{E}\left([X - \mathrm{E}(X)]^2\right) = \mathrm{E}\left(X^2\right) - \mathrm{E}(X)^2$$

$\sigma = \sqrt{\mathrm{V}(X)} = \sqrt{\mu_2}$ est **l'écart type** (*standard deviation*) de X
σ mesure la dispersion des données autour de la moyenne.

- Variable **centrée** lorsque sa moyenne est nulle,
- Variable **réduite** lorsque de plus sa variance vaut 1.
- Variable centrée-réduite X' associée à X par : $X' = \dfrac{X - m_1}{\sigma}$

On définit aussi **les coefficients** sans dimension liés aux moments d'ordre 3 et 4 :

$$\begin{aligned} \textbf{d'asymétrie } (\textit{skewness}) \qquad & \gamma_1 = \frac{\mu_3}{\sigma^3} \\ \textbf{d'aplatissement} (\textit{kurtosis}) \qquad & \gamma_2 = \frac{\mu_4}{\sigma^4} - 3 \end{aligned} \tag{7}$$

Certains auteurs définissent ce coefficient comme $\frac{\mu_4}{\sigma^4}$; on parle alors d'*excess kurtosis* pour γ_2.

Le coefficient d'asymétrie (et plus généralement, tous les moments centrés d'ordre impair) est nul pour une variable dont la loi de probabilité est symétrique par rapport à la moyenne.

Le coefficient d'aplatissement mesure l'importance des « queues » de la loi de probabilité : il est positif quand les événements extrêmes sont plus probables que pour une variable gaussienne.

Statistiques d'ordre et quantiles d'une v.a.

Quantile (percentile ou fractile) d'ordre q $(0 < q < 1)$=
valeur x_q telle que $P(X \leqslant x_q) = q$, c'est-à-dire $x_q = F_X^{-1}(q)$
x_q ne dépend que de l'**ordre** entre les valeurs prises par la v.a.
$\Rightarrow$ moins sensible aux valeurs aberrantes isolées que les moments.

- **Mode** = valeur la plus probable (v.a. discrète) ou valeur où la densité de probabilité est maximale (v.a. continue)
- **Médiane** = valeur $x_{1/2}$ telle que $F_X(x_{1/2}) = 1/2$
  $\Leftrightarrow$ autant de valeurs à gauche qu'à droite de la médiane
  (ne pas confondre avec la moyenne ni avec le mode)
  utilisée pour éliminer des parasites isolés dans un signal (filtre médian)
- en statistique, on définit des **intervalles de confiance** à 90%, par $[x_{0.05},\ x_{0.95}]$ dans le cas symétrique pour encadrer des paramètres estimés.

Fonctions caractéristiques et fonction génératrice I

La **première fonction caractéristique** Φ_X est la transformée de Fourier de la densité f_X :

$$\boxed{\Phi_X(u) = \mathrm{E}\left(e^{iuX}\right) = \int_{-\infty}^{+\infty} f_X(x) e^{iux} \, \mathrm{d}x} \tag{8}$$

À partir de la densité de probabilité, on peut calculer tous les moments d'une v.a. et il faut tous les connaître pour revenir à sa densité.

$$\Phi_X(u) = \sum_{k=0}^{\infty} \frac{i^k u^k}{k!} \mathrm{E}\left(X^k\right) \tag{9}$$

$$\frac{\mathrm{d}^n \Phi_X}{\mathrm{d}u^n}(u=0) = i^n \mathrm{E}\left(X^n\right) \quad \text{et} \quad \Phi_X(0) = 1 \tag{10}$$

$$\Phi_X(u) = 1 + iu\,\mathrm{E}\left(X\right) - \frac{1}{2}u^2 \mathrm{E}\left(X^2\right) + \cdots \tag{11}$$

Fonctions caractéristiques et fonction génératrice II

La deuxième fonction caractéristique Ψ_X est le logarithme de la première.

$$\boxed{\Psi_X(u) = \ln \Phi_X(u) = \ln \left[\mathrm{E}\left(e^{iuX}\right) \right]} \tag{12}$$

Ψ_X se développe selon les **cumulants** $K_n(X)$.

$$\frac{\mathrm{d}^n \Psi_X}{\mathrm{d}u^n}(u=0) = i^n K_n(X) \quad \text{et} \quad \Psi_X(0) = 0 \tag{13}$$

En particulier $K_1(X) = \mathrm{E}(X)$ et $K_2(X) = \mathrm{V}(X)$.
Si X est centrée, $K_3(X) = \mu_3(X)$ et $K_4(X) = \mu_4(X) - 3\mu_2(X)^2$
(tous deux nuls pour une gaussienne).

$$\Psi_X(u) = iu\,\mathrm{E}(X) - \frac{1}{2}u^2\,\mathrm{V}(X) + \cdots \tag{14}$$

Transformation linéaire de X :

$$\begin{aligned} Y = aX + b \quad &\Rightarrow \quad \Phi_{aX+b}(u) &&= \mathrm{E}\left(e^{iu(aX+b)}\right) = e^{iub}\Phi_X(au) \\ &\Rightarrow \quad \Psi_{aX+b}(u) &&= iub + \Psi_X(au) \end{aligned}$$

Fonctions d'une variable aléatoire I

Distribution de la fonction d'une v.a.

$Y = g(X)$ où g est une fonction certaine.
Déduire la **loi de** $g(X)$ de celle de X
Cas où g est **monotone** (croissante par exemple)

$$P(X \leqslant x) = F_X(x) = P(g(X) \leqslant g(x)) = F_Y(g(x))$$

$$\Rightarrow \quad F_Y(y) = F_X(g^{-1}(y)) \quad \Rightarrow \quad f_Y(y) = \frac{f_X(g^{-1}(y)}{g'(g^{-1}(y))}$$

Plus généralement

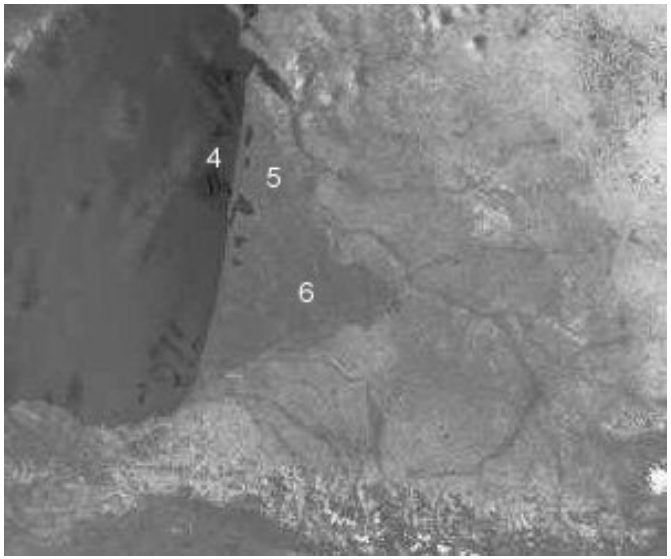
$$\boxed{f_Y(y) = f_X(x) \left| \frac{\mathrm{d}x}{\mathrm{d}y} \right|}$$

Sommer toutes les contributions si g non monotone.

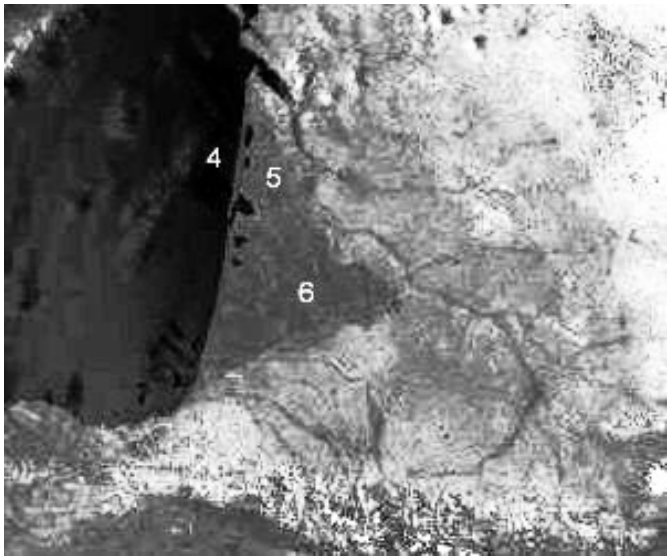
Application I

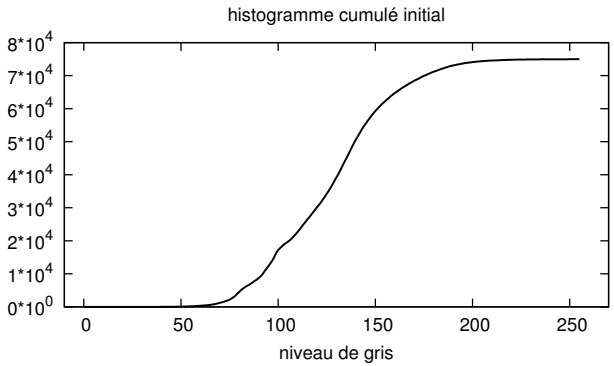
Égalisation d'histogramme en traitement d'image

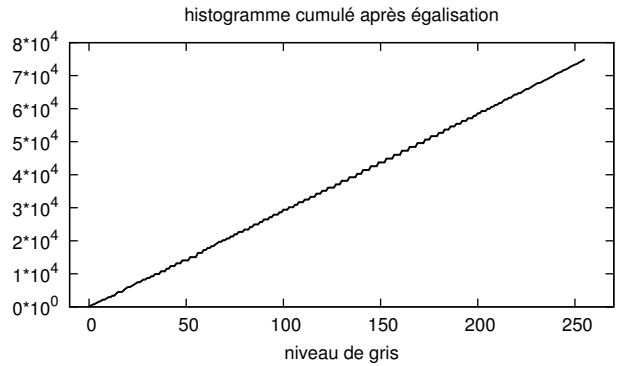
Rendre uniforme l'histogramme des niveaux de gris pour améliorer le contraste



nécessite de classer les données







Application II

Générateurs pseudo-aléatoires de loi imposée à partir d'un générateur uniforme
X uniforme sur $[0,1]$ donc $f_X(x) = 1$ et $F_X(x) = x$.
f_Y imposée donc F_Y par intégration, puis choisir $g = F_Y^{-1}$

Exemple : loi exponentielle de paramètre λ

générer la v.a. Y de densité

$$f_Y(y) = \lambda e^{-\lambda y} \quad \text{pour} \quad y \geqslant 0.$$

1. intégrer la densité de probabilité de Y

$$F_Y(y) = \int_0^y f_Y(y_1)\,\mathrm{d}y_1 = 1 - e^{-\lambda y} = x$$

2. inverser la fonction de répartition de Y

$$\Rightarrow Y = -\ln(1-X)/\lambda \quad \text{suit une loi exponentielle}$$

Moments de la fonction d'une v.a.

Approximation des **moments d'une fonction non-linéaire d'une v.a.** : développer $g(X)$ autour de la moyenne $m_1 = \mathrm{E}(X)$

$$g(X) = g(m_1) + \left(\frac{\mathrm{d}g}{\mathrm{d}X}\right)_{m_1} (X - m_1) + \frac{1}{2}\left(\frac{\mathrm{d}^2 g}{\mathrm{d}X^2}\right)_{m_1} (X - m_1)^2 + \cdots$$

En prenant l'espérance,

$$\mathrm{E}(g(X)) \approx g(m_1) + \frac{1}{2}\left(\frac{\mathrm{d}^2 g}{\mathrm{d}X^2}\right)_{m_1} \mathrm{V}(X) \approx g(m_1) = g(\mathrm{E}(X))$$

En négligeant le terme du second ordre dans la moyenne,

$$\mathrm{V}(g(X)) \approx \left(\frac{\mathrm{d}g}{\mathrm{d}X}\right)^2_{m_1} \mathrm{V}(X)$$

Distributions bivariées, conditionnement

Probabilité conjointe de A et B : $P[A \text{ et } B]$
X et Y de loi conjointe $F_{X,Y} = P[X \leqslant x \text{ et } Y \leqslant y]$
Densité de probabilité conjointe de X et Y :

$$f_{X,Y}(x,y) = \frac{\partial^2 F_{X,Y}}{\partial x \, \partial y}$$

Densité de probabilité marginale de Y :

$$f_Y(y) = \int_{-\infty}^{+\infty} f_{X,Y}(x,y) \, \mathrm{d}x$$

Probabilité conditionnelle de B sachant que A est vrai :

$$P[B|_A] = P[A \text{ et } B]/P[A]$$

Densité de probabilité conditionnelle de Y connaissant X :

$$\boxed{f_{Y|X}(y;x) = \frac{f_{X,Y}(x,y)}{f_X(x)}} \tag{15}$$

Variables aléatoires indépendantes (v.a.i.)

Événements **indépendants**

$$P[A \text{ et } B] = P[A] \times P[B] \quad \Longleftrightarrow \quad P[A|_B] = P[A]$$

X et Y **variables aléatoires indépendantes** (v.a.i.) si factorisation des cdf et pdf :

$$F_{X,Y}(x,y) = F_X(x) \times F_Y(y) \quad \Leftrightarrow \quad f_{X,Y}(x,y) = f_X(x) \times f_Y(y)$$

ou encore $f_{Y|X}(y;x) = f_Y(y)$

Moments bivariés

$\mathrm{E}\left(g(X,Y)\right) = \iint g(x,y) f_{X,Y}(x,y)\ \mathrm{d}x\ \mathrm{d}y$

Covariance (* pour complexe conjugué)

$$\mathrm{Cov}(X,Y) = \mathrm{E}\left((X - \mathrm{E}(X))(Y - \mathrm{E}(Y))^\star\right) = \mathrm{E}(XY^\star) - \mathrm{E}(X)\,\mathrm{E}(Y^\star)$$

Moments de v. a. indépendantes.

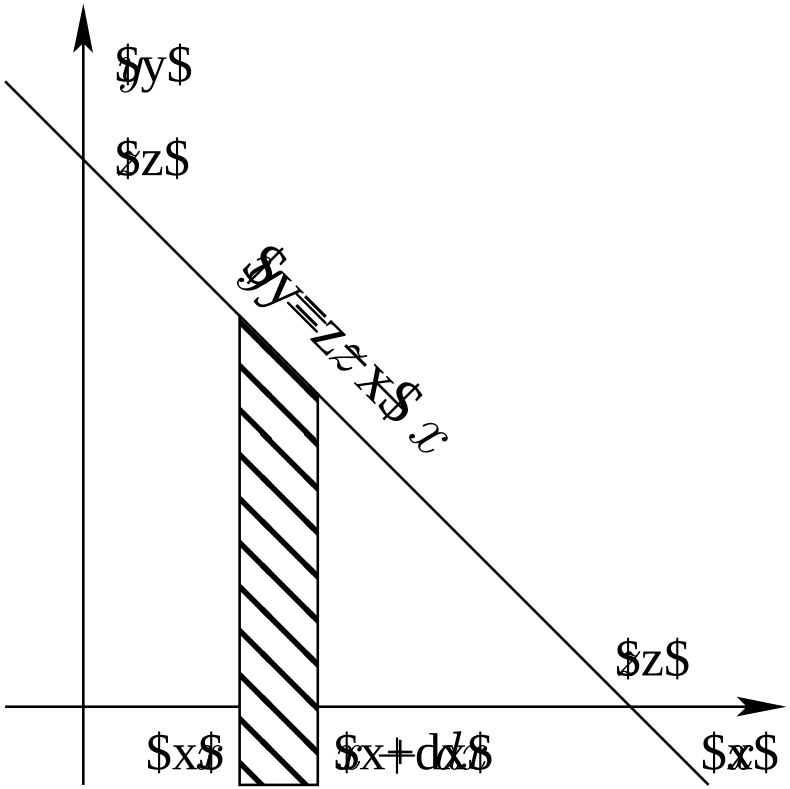
$\boxed{\mathrm{E}(XY) = \mathrm{E}(X)\,\mathrm{E}(Y)}$ et covariance nulle car

$\iint_{x,y} f_{X,Y}(x,y) xy\ \mathrm{d}x\ \mathrm{d}y = \left[\int_x x\, f_X(x)\,\mathrm{d}x\right]\left[\int_y y\, f_Y(y)\,\mathrm{d}y\right]$

Somme de variables aléatoires indépendantes

Loi de la somme de deux variables aléatoires : $Z = X + Y$

$$F_Z(z) = P[Z \leqslant z] = \int_{x=-\infty}^{+\infty} \int_{y=-\infty}^{z-x} f_{X,Y}(x,y)\,\mathrm{d}x\,\mathrm{d}y = \int_{x=-\infty}^{+\infty} F_{Y|_{X=x}}(z-x)\,f_X(x)\,\mathrm{d}x$$



Cas général : **produit de convolution**
$f_Z = f_{Y|_{X=x}} \star f_X$

Cas où X et Y sont **indépendantes** :

$$f_{X+Y} = f_X \star f_Y \quad \Longrightarrow \quad \Phi_{X+Y} = \Phi_X \Phi_Y$$

Les secondes fonctions caractéristiques s'ajoutent.

$$\boxed{\Psi_{X+Y} = \Psi_X + \Psi_Y}$$

Variables aléatoires discrètes I

- la loi discrète **uniforme** sur $1, 2, ..., n$:

$$\forall k \in \{1, 2, ..., n\},\ P[X = k] = \frac{1}{n}; \tag{16}$$

moyenne $m_1 = \frac{n+1}{2}$, variance $\sigma^2 = \frac{n^2-1}{12}$.

- **la loi de Bernoulli** (ou de pile ou face) de paramètre p :

$$P[X = 1] = p \text{ et } P[X = 0] = 1 - p; \tag{17}$$

moyenne $m_1 = p$, variance $\sigma^2 = p(1-p)$,
$\Phi(u) = p\mathrm{e}^{iu} + 1 - p \approx 1 + piu - pu^2/2 + \cdots$
$\Psi(u) = \ln\left(1 + p(\mathrm{e}^{iu} - 1)\right) \approx piu - p(1-p)u^2/2 + \cdots$

Variables aléatoires discrètes II

- **la loi binomiale de paramètres n et p,**
  obtenue en sommant n variables de pile ou face indépendantes :

$$P(X = k) = C_n^k \, p^k \, (1-p)^{n-k}, \text{ pour } 0 \leqslant k \leqslant n. \tag{18}$$

  moyenne $m_1 = np$, variance $\sigma^2 = np(1-p)$.
  $\Psi(u) = n \ln\left(1 + p(\mathrm{e}^{iu} - 1)\right) \approx npiu - np(1-p)u^2/2 + \cdots$

- **la loi de Poisson** de paramètre λ :

$$P[X = k] = \frac{\lambda^k}{k!} \exp(-\lambda) \quad k \geqslant 0 \tag{19}$$

  moyenne $m_1 = \lambda$, variance $\sigma^2 = \lambda$.
  $\Phi_X(u) = \exp\left[\lambda\left(\mathrm{e}^{iu} - 1\right)\right]$
  $\Psi_X(u) = \lambda\left(\mathrm{e}^{iu} - 1\right) = \lambda\left(iu + u^2/2 - iu^3/6 + \cdots\right) \quad K_n(X) = \lambda \ \ \forall n$

Variables aléatoires discrètes III

Applications des secondes fonctions caractéristiques

- La somme de deux v.a.i. de Poisson de paramètres λ_1 et λ_2 est une v.a. de Poisson de paramètre $\lambda_1 + \lambda_2$.
- La loi de Poisson est la limite de la loi binomiale pour $n \to \infty$ et $p \to 0$ avec $np = \lambda$. $\to$ loi des événements rares
  En développant la seconde fonction caractéristique

$$\Psi(u) = n \ln\left(1 + \frac{\lambda}{n}(\mathrm{e}^{iu} - 1)\right) = n\left[\frac{\lambda}{n}(\mathrm{e}^{iu} - 1) - \left(\frac{\lambda}{n}\right)^2 \cdots\right] \to \lambda(\mathrm{e}^{iu} - 1)$$

Variables aléatoires continues I

- la loi **uniforme** sur $[0\,;a]$, de densité de probabilité :

$$f(x) = \frac{1}{a} \text{ pour } 0 \leqslant x \leqslant a; \tag{20}$$

moyenne $m_1 = a/2$, variance $\sigma^2 = a^2/12$.

- loi **exponentielle** de paramètre λ :

$$f(x) = \lambda \exp(-\lambda x) \text{ si } x \geqslant 0 \tag{21}$$

moyenne $m_1 = 1/\lambda$, variance $\sigma^2 = 1/\lambda^2$, $m_n = n!/\lambda^n$.
Cette distribution est étalée vers la droite : $\gamma_1 = 2$ et $\gamma_2 = 6$.

$$\Phi(u) = \frac{\lambda}{\lambda - iu} = \frac{1}{1 - iu/\lambda} = \sum_{k=0}^{\infty} \left(\frac{iu}{\lambda}\right)^k = \sum_{k=0}^{\infty} \frac{(iu)^k}{k!} m_k$$

Variables aléatoires continues II

- la loi **gaussienne** (ou normale) d'espérance m et de variance σ^2 :

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{(x-m)^2}{2\sigma^2}} \tag{22}$$

$\mu_{2k+1} = 0$ et $\gamma_2 = 0$. Fonction de répartition sans expression analytique obtenue à partir de la « **fonction d'erreur** » ($\mathrm{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2}\, \mathrm{d}t$) :

$$F(x) = \frac{1}{2}\left[1 + \mathrm{erf}\left(\frac{x-m}{\sigma\sqrt{2}}\right)\right] \tag{23}$$

moyenne $m_1 = m$, variance $\mathrm{V}(X) = \sigma^2$
$\Phi(u) = \exp(ium - u^2\sigma^2/2)$
$\Psi(u) = ium - u^2\sigma^2/2$ polynôme de degré 2 (cumulants d'ordre > 2 nuls)
Conséquence : la somme de deux v.a.i. de Gauss est une v.a. de Gauss (les moyennes et les variances s'ajoutent).

Variables aléatoires continues III

- **loi du** χ^2_n à n degrés de liberté ; c'est la loi de la variable :

$$\sum_{k=1}^{n} X_k^2 \tag{24}$$

où les variables X_k sont des variables gaussiennes centrées-réduites indépendantes.
Fonction de répartition

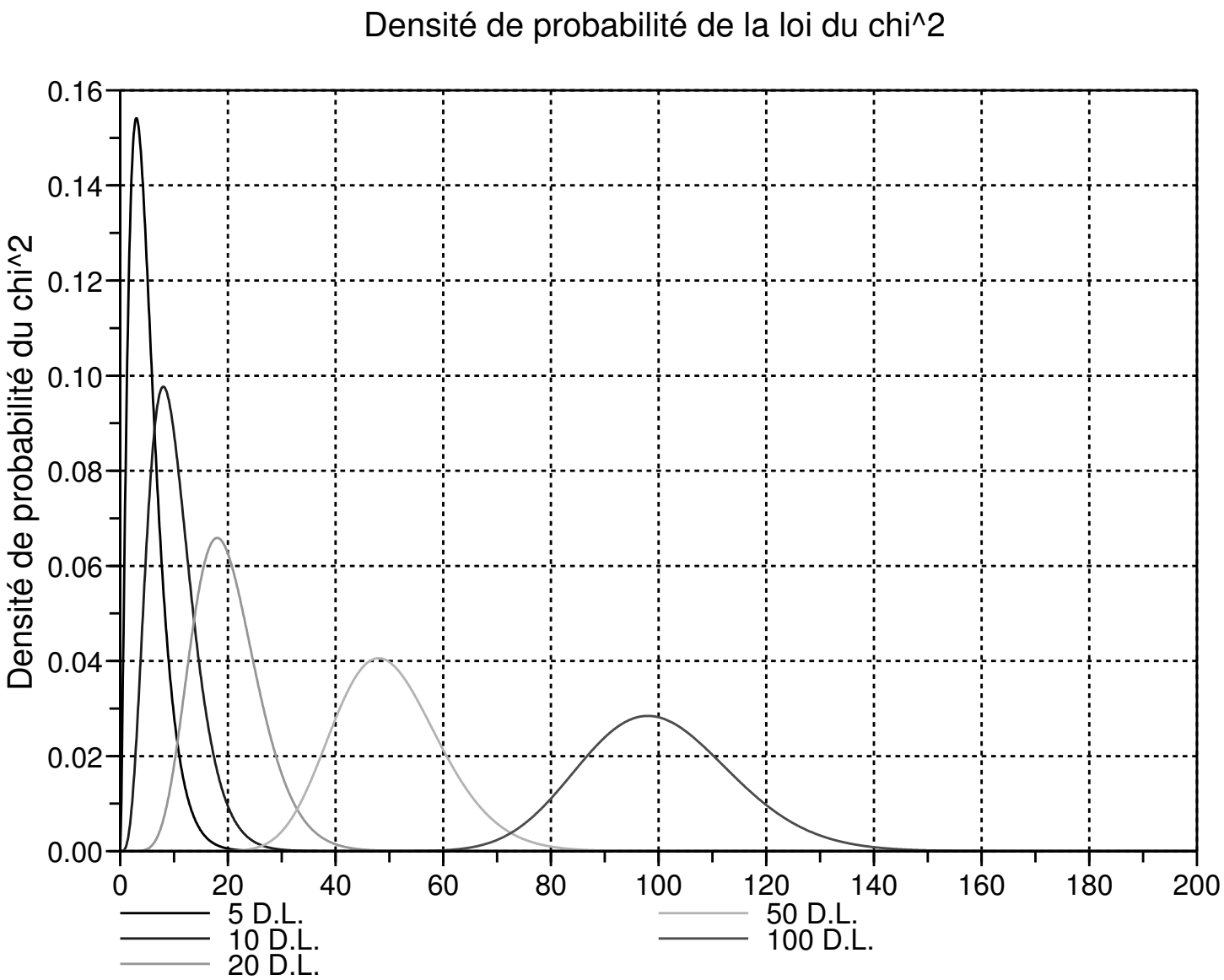
$$F(x) = P\left(\sum_{k=1}^{n} X_k^2 < x\right) = \Gamma\left(\frac{n}{2},\, \frac{x}{2}\right) \tag{25}$$

où Γ est la fonction gamma incomplète :

$$\Gamma(a,\, x) = \frac{\int_0^x e^{-t}\, t^{a-1}\, \mathrm{d}t}{\int_0^\infty e^{-t}\, t^{a-1}\, \mathrm{d}t} \tag{26}$$

moyenne $m_1 = n$, variance $\mathrm{V}\,(X) = 2n$.

Densité de la loi du χ^2



La question de l'estimation

On se place ici dans le cas où l'on dispose d'un ensemble de données (par exemple, une série d'observations d'un même phénomène en physique) et l'on cherche à estimer les propriétés de la variable aléatoire (le phénomène physique) dont les données constituent une réalisation.
Si les n données sont des tirages **indépendants** de la variable aléatoire, elles constituent un **échantillon de taille** n.
Définir un **estimateur** d'un paramètre statistique de la loi que suit la variable aléatoire, c'est se donner une méthode de calcul approché de ce paramètre en fonction des n tirages. **L'estimateur est alors lui-même une variable aléatoire**, puisque sa valeur dépend de la réalisation dont on dispose.
Il s'agit de choisir un estimateur :

- **sans biais** (sans erreur systématique)
- **de faible variance** (présentant le minimum de dispersion)
- **convergent** : qui tend vers le paramètre à estimer (souvent au sens de la moyenne quadratique) lorsque la taille de l'échantillon tend vers l'infini.

Estimateur de la moyenne

L'estimateur $\overline{X}$ de la moyenne est calculé selon :

$$\overline{X} = \frac{1}{n}\sum_{k=1}^{n} X_k \tag{27}$$

L'estimation $\overline{x}$ est une réalisation de la variable aléatoire $\overline{X}$, fonction des n réalisations indépendantes de X dont on cherche à estimer les propriétés : m sa moyenne et σ^2 sa variance. $\overline{X}$ est un estimateur :

- sans biais : $\forall n,\ \mathrm{E}\left(\overline{X}\right) = m$;

$$\mathrm{E}\left(\overline{X}\right) = \mathrm{E}\left(\frac{1}{n}\sum_{k=1}^{n} X_k\right) = \frac{1}{n}\sum_{k=1}^{n}\mathrm{E}\left(X_k\right) = \frac{1}{n} n\,\mathrm{E}\left(X\right) = \mathrm{E}\left(X\right)$$

- de variance : $\mathrm{V}\left(\overline{X}\right) = \dfrac{\sigma^2}{n}$ (car les X_k sont indépendants)

$$\mathrm{V}\left(\overline{X}\right) = \mathrm{V}\left(\frac{1}{n}\sum_{k=1}^{n} X_k\right) = \frac{1}{n^2}\,\mathrm{V}\left(\sum_{k=1}^{n} X_k\right) = \frac{1}{n^2}\sum_{k=1}^{n}\mathrm{V}\left(X_k\right) = \frac{\mathrm{V}\left(X\right)}{n}$$

- convergent : $\overline{X} \underset{n\to\infty}{\longrightarrow} m$.

Estimateur de la variance

Si la moyenne n'est pas connue, un estimateur naturel de la variance est :

$$S'^2 = \frac{1}{n}\sum_{k=1}^{n}\left(X_k - \overline{X}\right)^2$$

Mais l'emploi de la moyenne empirique en fait un estimateur biaisé :

$$\sum(X_k - m)^2 = \sum(X_k - \overline{X})^2 + 2(\overline{X} - m)\sum(X_k - \overline{X}) + n(\overline{X} - m)^2$$

$$\sum(X_k - \overline{X}) = 0 \quad \Rightarrow \quad n\,\mathrm{V}\left(X\right) = n\,\mathrm{E}\left(S'^2\right) + n\,\mathrm{V}\left(\overline{X}\right)$$

$$\mathrm{E}\left(S'^2\right) = \frac{n-1}{n}\,\sigma^2$$

On lui préfère donc l'estimateur sans biais S :

$$S^2 = \frac{1}{n-1}\sum_{k=1}^{n}\left(X_k - \overline{X}\right)^2 \tag{28}$$

Cet estimateur est aussi convergent.

Estimateurs des moments d'ordre supérieur

Les coefficients d'asymétrie et d'aplatissement sont estimés respectivement selon :

$$\Gamma_1 = \frac{\frac{1}{n}\sum_{k=1}^{n}(X_k - \overline{X})^3}{S^3} \qquad \Gamma_2 = \frac{\frac{1}{n}\sum_{k=1}^{n}(X_k - \overline{X})^4}{S^4} - 3 \tag{29}$$

Les estimateurs associés sont asymptotiquement $(n \to \infty)$ non-biaisés et convergents.

Estimateurs des lois et fonction de répartition

On obtient des informations sur la loi de probabilité de la variable aléatoire que l'on étudie en construisant l'**histogramme empirique** des fréquences d'occurrence des différentes valeurs observées. Le problème pratique, dans le cas continu, consiste à trouver un compromis entre des intervalles de valeurs (ou classes) assez larges permettant d'obtenir un histogramme d'allure régulière, mais suffisamment petits pour ne pas trop réduire l'information.
La fonction de répartition est quant à elle approchée par la **fonction de répartition empirique** F_n^* : soient $(x_1, x_2, \ldots, x_n)$ les différentes valeurs observées **classées par ordre croissant**, alors :

$$F_n^*(x) = \begin{cases} 0 & \text{si } x < x_1 \\ \dfrac{k}{n} & \text{si } x_k \leqslant x < x_{k+1} \\ 1 & \text{si } x \geqslant x_n \end{cases} \tag{30}$$

On peut démontrer que la fonction de répartition empirique converge presque sûrement vers la fonction de répartition de la variable dont on possède un échantillon.

Théorème de la limite centrale I

$X_1, X_2, \cdots, X_n$ n variables aléatoires indépendantes de même loi (échantillon de taille n) possédant une moyenne m et une variance σ^2.

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \quad \text{moyenne empirique}$$

$$\mathrm{E}\left(\overline{X}_n\right) = \mathrm{E}\left(X\right) \quad \text{et} \quad \mathrm{V}\left(\overline{X}_n\right) = \frac{\sigma^2}{n} \to 0 \text{ si } n \to \infty$$

Soit Y_i la variable centrée associée à X_i
Soit Z_n la variable centrée réduite associée à $\overline{X}_n$

$$Z_n = \frac{\overline{X}_n - m}{\sigma/\sqrt{n}} = \frac{1}{\sigma\sqrt{n}}\sum_{i=1}^{n} Y_i$$

$$\Psi_{Z_n}(u) = \Psi_{\sigma\sqrt{n}Z_n}(u/\sigma\sqrt{n}) = n\Psi_{Y_i}(u/\sigma\sqrt{n})$$

Théorème de la limite centrale II

Comme Y_i est centrée, $\Psi_{Y_i}(v) = -\sigma^2 v^2/2 + v^2\varepsilon(v)$, donc

$$\Psi_{Z_n}(u) = n\left[-\frac{\sigma^2}{2}\frac{u^2}{n\sigma^2} + \frac{u^2}{n\sigma^2}\varepsilon\left(\frac{u}{\sigma\sqrt{n}}\right)\right]$$

$$\Psi_{Z_n}(u) = -\frac{u^2}{2} + \frac{u^2}{\sigma^2}\varepsilon\left(\frac{u}{\sigma\sqrt{n}}\right) \to -\frac{u^2}{2} \quad \text{si} \quad n \to \infty$$

$\Psi_{Z_n} \to$ deuxième fonction caractéristique d'une gaussienne centrée-réduite.
$\overline{X}_n$ tend vers une gaussienne de moyenne m et sa variance tend vers 0 comme $1/n$
Rôle prépondérant des gaussiennes dans les **phénomènes additifs**
(physique macroscopique = somme de contributions microscopiques indépendantes)

Application : générateurs pseudo-aléatoires quasi-gaussiens

Somme de n v.a.i. uniformes $\implies$ convolution des distributions :
densité de la somme = polynôme de degré $n-1$.
Convergence rapide : $n = 12$ pour simplifier le calcul (variance unité).

Autre application : la loi du χ^2_n tend vers une gaussienne pour n grand.

Un cas de non application de la limite centrale : la loi de Cauchy

Loi de Cauchy : celle du rapport de 2 gaussiennes centrées réduites indépendantes
Densité de probabilité

$$f_X(x) = \frac{1}{\pi}\frac{1}{1+x^2}$$

Fonction de répartition

$$F_X(x) = \frac{1}{2} + \frac{1}{\pi}\arctan x$$

Fonctions caractéristiques

$$\Phi_X(u) = \exp(-|u|) \quad \text{et} \quad \Psi_X(u) = -|u|$$

Mais aucun moment de la loi de Cauchy n'existe ! Mais mode et médiane nuls.

La moyenne $\overline{X}_n$ de n v.a.i. de Cauchy suit aussi une loi de Cauchy quel que soit n.

$$\Psi_{\overline{X}_n}(u) = n\Psi_X(u/n) = -n|u/n| = -|u| = \Psi_X(u)$$

Échantillon de la loi de Cauchy

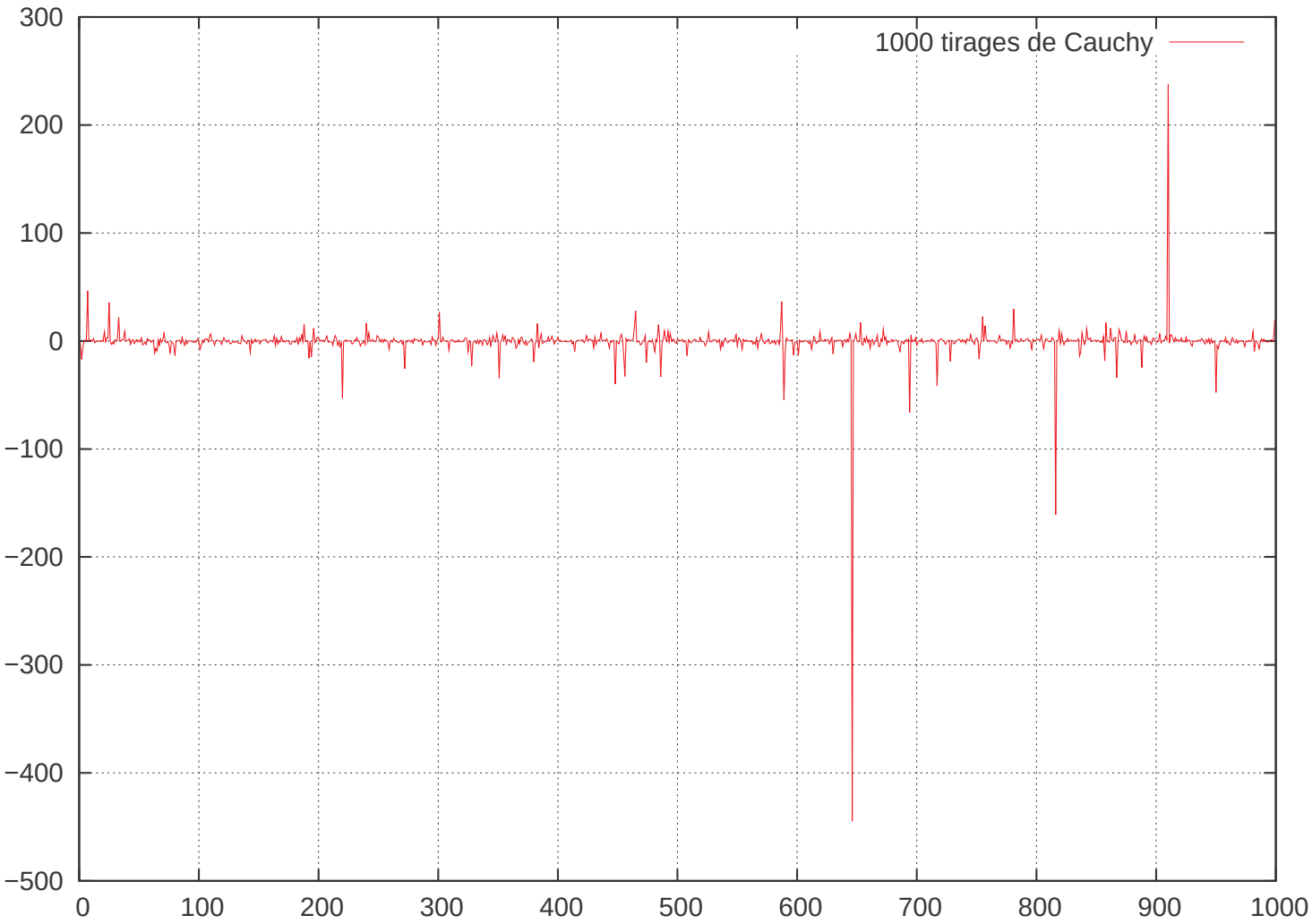


FIGURE : Un échantillon de 1000 tirages selon la loi de Cauchy

Introduction aux tests d'hypothèses I

Un **test** est un mécanisme de décision permettant de trancher entre deux hypothèses :

- une hypothèse dite « nulle » notée H_0
- et une hypothèse alternative notée H_1

au vu des résultats d'un échantillon d'observations.
Nous ne présentons ici que deux tests relatifs à la **loi de probabilité** que suit la variable étudiée, ou **tests d'adéquation** (*goodness-of-fit*) :

- le **test du** χ^2 et
- le **test de Kolmogorov-Smirnov**.

De nombreux autres tests existent, permettant de comparer par exemple

- la **moyenne** de deux échantillons (**test de Student**)
- ou leur **variance** (**test de Fisher**).

Introduction aux tests d'hypothèses II

Décision \ Hypothèse	H_0	H_1
acceptation D_0	$1-\alpha$	β
rejet (région critique) D_1	α	$1-\beta$

Stratégie de décision : partition de l'espace des observations entre région d'acceptation (D_0) et région critique (D_1).
Performances déterminées par :

- $\alpha = P[D_1|_{H_0}]$ risque de **première espèce** (rejet erroné ou fausse alarme)
- $\beta = P[D_0|_{H_1}]$ risque de **deuxième espèce** (acceptation erronée)
- $1-\beta = P[D_1|_{H_1}]$ **puissance** du test (rejet justifié)

Maximum de vraisemblance a posteriori

Minimiser la somme des probabilités d'erreur :

$$\min \quad (\beta + \alpha) \iff \min \quad P[D_0|_{H_1}] + P[D_1|_{H_0}]$$

$$\min \quad \int_{D_0} p(\vec{x}|_{H_1})\,\mathrm{d}\vec{x} + 1 - \int_{D_0} p(\vec{x}|_{H_0})\,\mathrm{d}\vec{x}$$

$$\min \quad \int_{D_0} [p(\vec{x}|_{H_1}) - p(\vec{x}|_{H_0})]\,\mathrm{d}\vec{x}$$

Choisir le domaine D_0 de façon à minimiser l'intégrale en n'intégrant que des termes négatifs. La décision est donc :

$$\boxed{\vec{x} \in D_0 \iff V(\vec{x}) = \frac{p(\vec{x}|_{H_0})}{p(\vec{x}|_{H_1})} > 1}$$

Comparaison du **rapport de vraisemblance** $V(\vec{x})$ à 1.

Stratégie bayésienne

Si on connait les probabilités a priori $p_0 = P[H_0]$ et $p_1 = P[H_1]$,
la probabilité d'erreur est :

$$p_e = p_0 P[D_1|_{H_0}] + p_1 P[D_0|_{H_1}] = p_0\alpha + p_1\beta$$

Mais on affecte parfois des coûts différents à ces deux risques

$$\min \quad p_0\, c_{10} P[D_1|_{H_0}] + p_1\, c_{01} P[D_0|_{H_1}]$$

La décision est donc :

$$\boxed{\vec{x} \in D_0 \iff V(\vec{x}) = \frac{p(\vec{x}|_{H_0})}{p(\vec{x}|_{H_1})} > \frac{p_1\, c_{01}}{p_0\, c_{10}}}$$

Comparaison du **rapport de vraisemblance** $V(\vec{x})$ à un seuil.

Cas non paramétrique

L'hypothèse H_1 est simplement la négation de H_0 et ne permet en général pas de calculer des probabilités.
Stratégie de Neyman-Pearson :

- choisir un risque acceptable de rejet α
- et maximiser la puissance $1-\beta$ du test (minimiser β)

Comparaison du rapport de vraisemblance $V(\vec{x})$ à un seuil déterminé par α.

Test du χ^2 sur l'histogramme I

On cherche à évaluer l'écart D entre l'histogramme construit à partir de l'échantillon et la loi de probabilité qu'est censée suivre la variable aléatoire.
Dans le cas continu, cela impose de discrétiser les observations en classes, ce qui peut faire préférer le test de Kolmogorov-Smirnov présenté plus loin.
On considère ainsi que l'**histogramme empirique** a été construit en utilisant k classes d'effectifs $N_1, N_2, \ldots, N_k$ avec la condition :

$$\sum_{i=1}^{k} N_i = N \quad \text{où } N \text{ est le nombre total d'observations} \tag{31}$$

L'hypothèse que l'on cherche à tester est :

H_0 : la variable aléatoire X dont on possède un échantillon suit une loi de probabilité donnée, pour laquelle la probabilité associée à chaque classe est $p_1, p_2, \ldots, p_k$.

L'hypothèse alternative est :

H_1 : la variable aléatoire X suit une autre loi de probabilité.

Test du χ^2 sur l'histogramme II

N_i = effectif d'une classe = somme de N v.a.i. de type pile ou face :

- 1 si le tirage tombe dans la classe i
  (probabilité p_i donnée par la loi et la classe)
- 0 sinon (probabilité $1 - p_i$)

N_i suit une loi binomiale de **moyenne** Np_i et de **variance** $Np_i(1-p_i) \approx Np_i$
Mais les N_i liés par $\sum_1^k N_i = N$

On calcule alors la distance (ou « statistique ») D permettant de mesurer l'**écart quadratique pondéré** par l'inverse de la variance de l'histogramme à celui de la loi de probabilité testée :

$$D = \sum_{i=1}^{k} \frac{(N_i - Np_i)^2}{Np_i} \tag{32}$$

Test du χ^2 sur l'histogramme III

On démontre que D suit une **loi du** χ^2_ν, à $\nu = k - 1 - c$ degrés de liberté, dans la limite où l'effectif de chaque classe est suffisamment grand,
$\Longrightarrow$ éliminer les classes d'effectif inférieur à 5 par exemple.

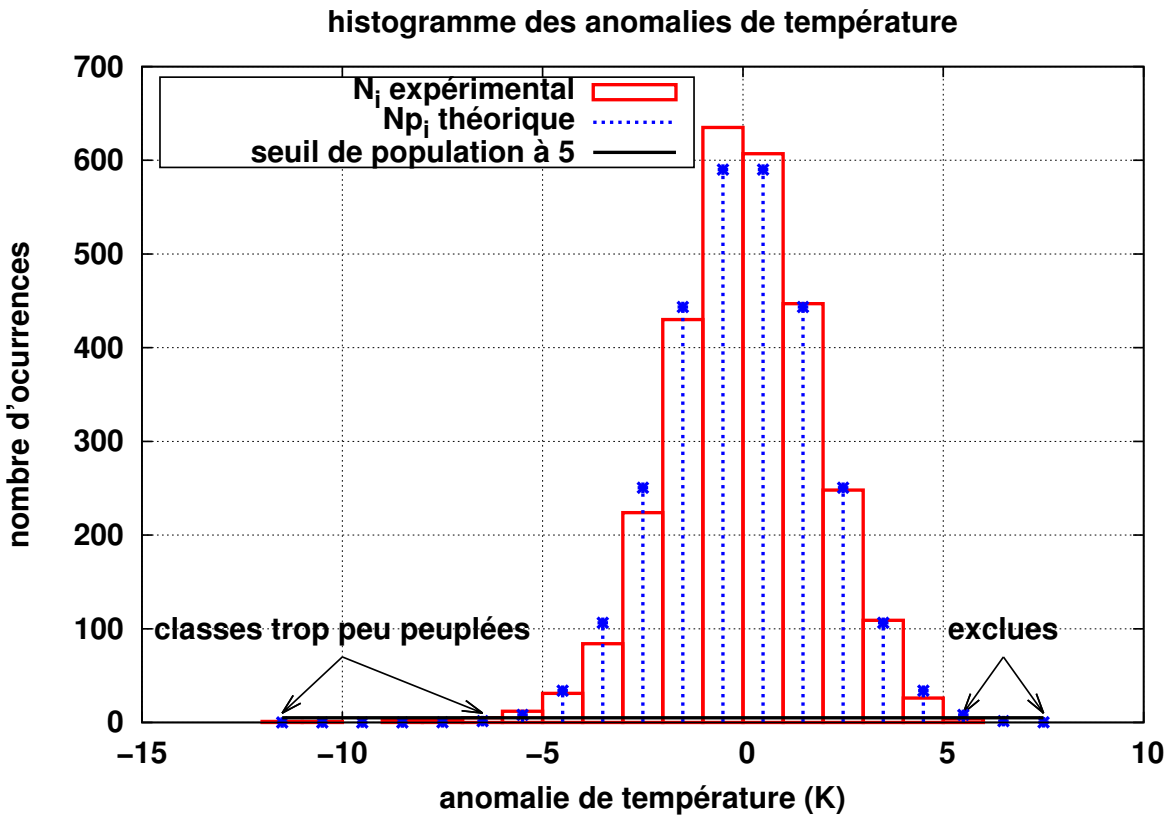
Nombre de degrés de liberté $\nu = k - 1 - c$

$+k$ nombre de classes retenues dans l'histogramme
(de population assez grande)

-1 car l'effectif total impose une relation entre les N_i

$-c$ le nombre de paramètres de la loi estimés à partir de l'échantillon
(par exemple : moyenne, écart-type, etc.)

Test du χ^2 sur l'histogramme IV



12 classes retenues
$\Rightarrow$ 9 degrés de liberté

FIGURE : Histogrammes expérimental et théorique

Test du χ^2 sur l'histogramme V

Question : **quand rejeter H_0 avec un risque r** (généralement faible) acceptable ?

- Estimer la moyenne et la variance via les moments empiriques pour identifier la gaussienne candidate.
- Fixer le nombre classes et leurs bornes (il faudrait qu'elle aient des populations proches). Calculer l'histogramme théorique.
- Calculer l'histogramme de l'échantillon et en déduire ν et d.
- Fixer un **risque** r (faible : 5% par ex.) de rejeter H_0 alors qu'elle est vraie ; en déduire un **seuil**, la valeur critique d_c définie par

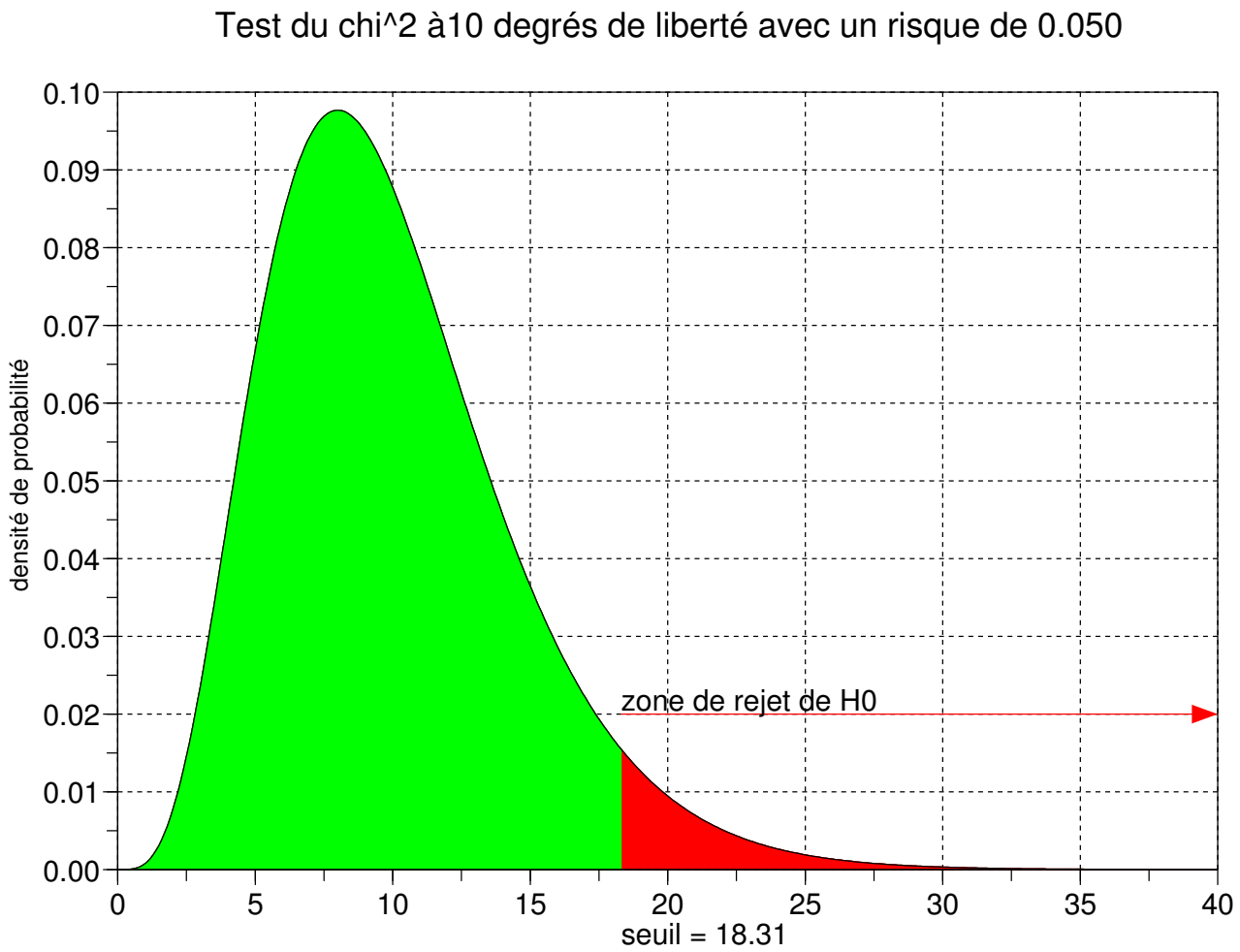
$$P_{\chi^2_\nu}[D \leqslant d_c] = 1 - r$$

  ce qui nécessite d'inverser la distribution cumulée du χ^2.
- Décider :
  - si $d > d_c$, rejeter l'hypothèse H_0 ;
  - si $d < d_c$, accepter l'hypothèse H_0 .

N.-B. : on ne démontre jamais que la variable aléatoire étudiée suit bien la loi de probabilité testée, mais seulement que l'échantillon dont on dispose est conforme avec l'hypothèse H_0.

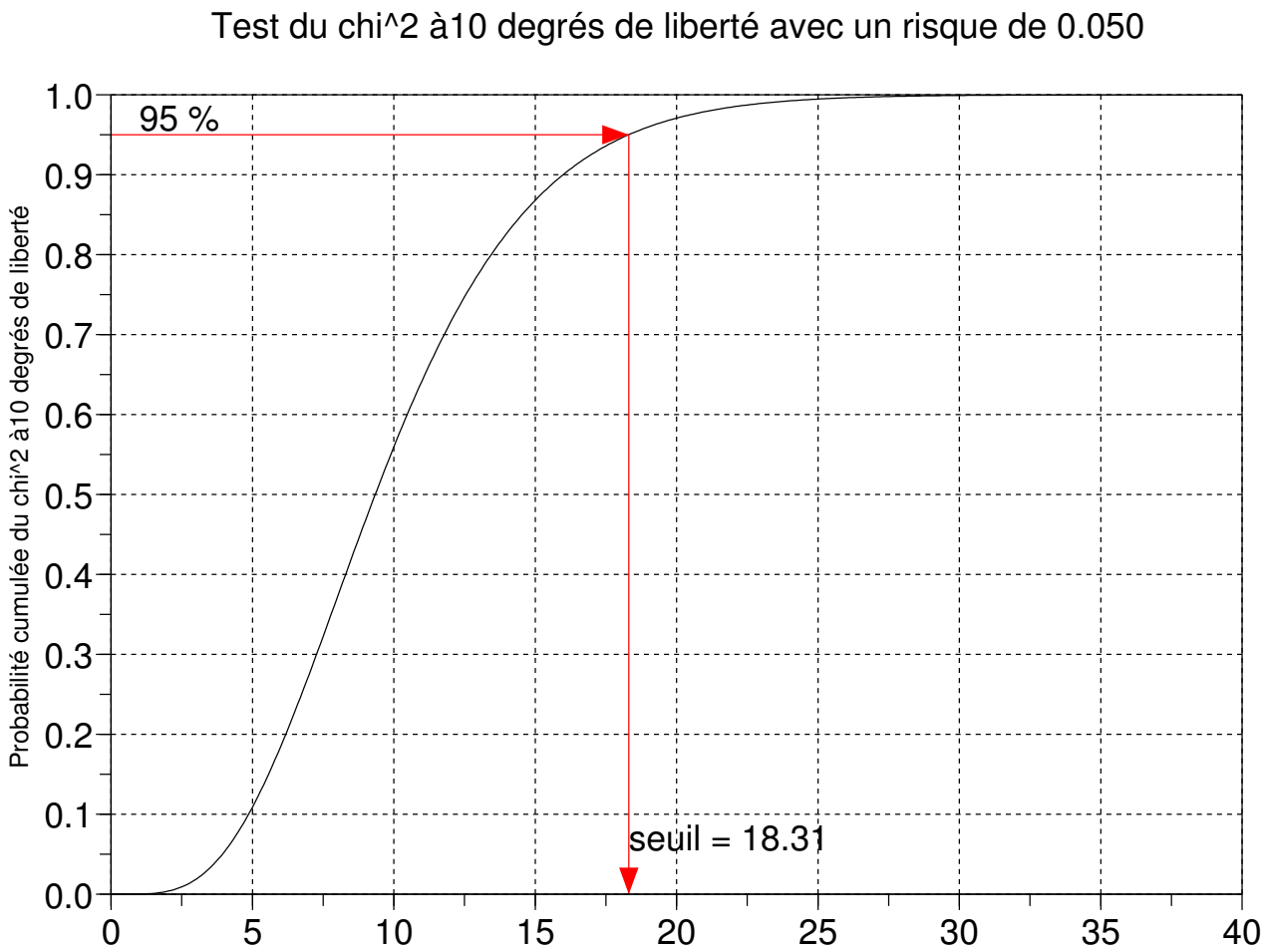
Test du χ^2 sur l'histogramme VI

Risque déduit du seuil : densité de probabilité du χ^2



Test du χ^2 sur l'histogramme VII

Risque déduit du seuil : inversion de la fonction de répartition cumulée du χ^2



Test de Kolmogorov-Smirnov I

Le **test de Kolmogorov Smirnov** ne nécessite pas de discrétiser les variables aléatoires continues (avantage par rapport au test du χ^2) :
$\Longrightarrow$ évite un choix arbitraire des classes de l'histogramme empirique.
Ce test travaille sur la fonction de répartition et l'on teste l'hypothèse :

H_0 : la variable aléatoire X dont on possède un échantillon suit une loi de probabilité donnée de fonction de répartition $F(x) = P(X \leqslant x)$.

L'hypothèse alternative est comme précédemment :

H_1 : la variable aléatoire X suit une autre loi de probabilité.

On mesure l'**écart maximal** d_N entre $F(x)$ et la fonction de répartition empirique $F_N^*(x)$:

$$\boxed{d_N = \sup |F_N^*(x) - F(x)|} \tag{33}$$

En pratique, il faut **classer** les données et, puisque la fonction de répartition empirique est **discontinue** aux valeurs x_k de l'échantillon, il faut calculer d_N comme :

$$d_N = \sup \left(|F_N^*(x_k) - F(x_k)| \; ; \; |F_N^*(x_{k-1}) - F(x_k)| \right) \tag{34}$$

Test de Kolmogorov-Smirnov II

La fonction de répartition de la variable aléatoire D_N dont d_N est une réalisation a pu être calculée (pour $N \to \infty$).
Elle est **indépendante de la loi de probabilité de la variable X** :

$$P\left[\sqrt{N} D_N < d_c\right] \underset{N\to\infty}{\longrightarrow} 1 + 2\sum_{k=1}^{\infty} (-1)^k \exp\left(-2\,k^2\,d_c^2\right) \tag{35}$$

On pourra utiliser la valeur limite dès que $N > 80$. Le calcul de la somme fournit :

$$P\left[D_N < \frac{1,223}{\sqrt{N}}\right] = 0,90\,; \quad P\left[D_N < \frac{1,358}{\sqrt{N}}\right] = 0,95\,; \quad P\left[D_N < \frac{1,629}{\sqrt{N}}\right] = 0,99$$

La comparaison de d_n calculé à partir de l'échantillon avec les valeurs limites précédentes permet donc de conclure, pour un certain risque, sur le rejet ou non de l'hypothèse nulle.

Test de Kolmogorov-Smirnov III

Cas où on estime des paramètres de la loi avant le test

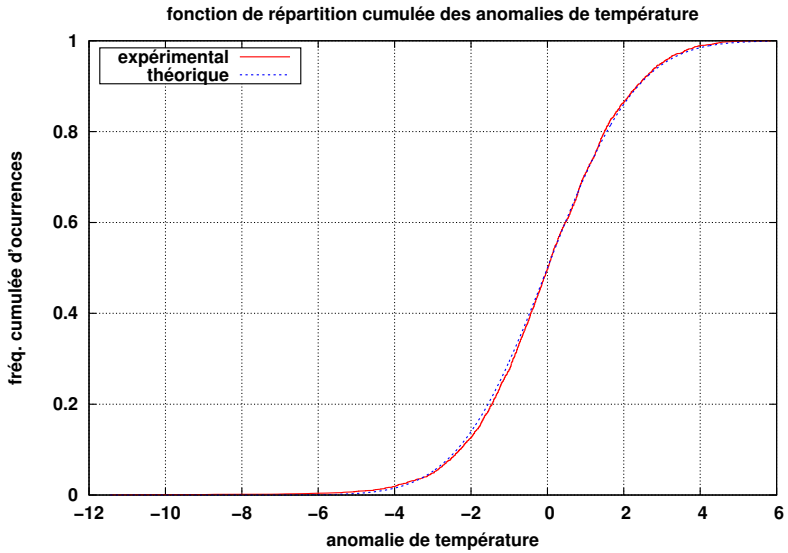
La limite (35) n'est exacte que lorsque la fonction de répartition $F(x)$ à tester est entièrement spécifiée *a priori*.

Si on a recours à l'échantillon pour estimer certains paramètres de $F(x)$ (tels que moyenne ou écart-type), les valeurs de distances critiques sont trop conservatives : le test pousse à conserver l'hypothèse nulle, alors qu'il faudrait la rejeter.

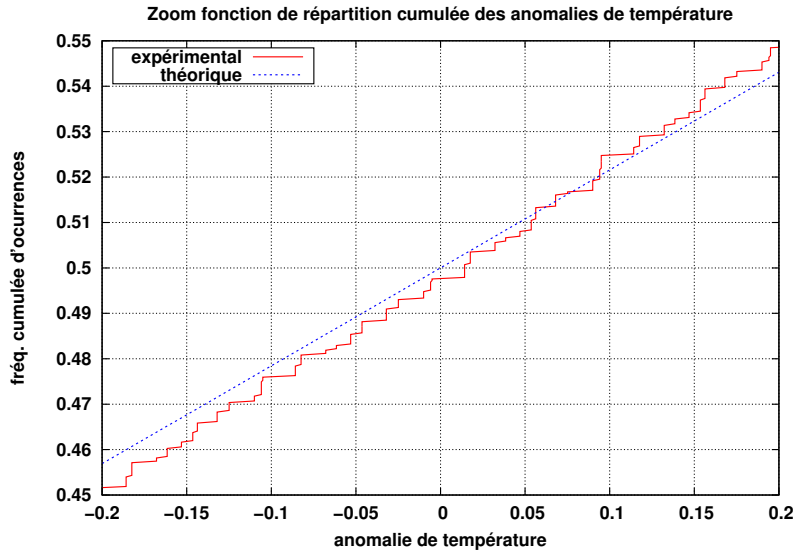
Des simulations numériques ont permis de calculer les valeurs de d_c adéquates, notamment lorsque l'on teste si l'échantillon peut être une réalisation d'une variable aléatoire gaussienne et qu'on estime la moyenne et l'écart-type à partir de l'échantillon.

$\Longrightarrow$ prendre $d'_c \approx 2d_c/3$, par exemple :

$$P\left(D_N < \frac{0,886}{\sqrt{N}}\right) = 0,95 \tag{36}$$



Allure générale



Zoom : on remarque les discontinuités de la répartition empirique

FIGURE : Répartition cumulée expérimentale et théorique pour le test de Kolmogoroff Smirnov

Quelques références bibliographiques

Jenkins, Gwilym et D. G. Watts, *Spectral analysis and its applications*, 525 pages (Holden-Day, 1968), ISBN 0-8162-4464-2.

Lejeune, Michel, *Statistique : la théorie et ses applications*, 434 pages (Springer, 2010), deuxième édition, ISBN 978-2-8178-0156-8.

Saporta, Gilbert, *Probabilité, Analyse des données et Statistique*, 622 pages (Technip, 2011), troisième édition, ISBN 978-2-7108-0980-7.