

# Introduction aux probabilités et statistiques

MIGUEL MUÑOZ ZUNIGA



# Table des matières

## Probabilités

mesure de probabilité

Dénombrement et probabilité

Probabilité conditionnelle et  
indépendance

Exemple du jeu Monty Hall

Simulation de variables aléatoires

Statistique descriptive

Statistique Inférentielle

Approche Bayésienne

## Variables aléatoires

## Subsection 1

mesure de probabilité

# Origines

Selon le dictionnaire Larousse

- ▶ **Hasard** : Circonstance de caractère imprévisible
- ▶ **Incertain** : Qui n'est pas établi avec exactitude, connu avec certitude
- ▶ **Aléatoire** : Soumis au hasard, dont le résultat est incertain

# Origines

- ▶ Origine du calcul de probabilités : jeu de hasard. De l'arabe "az-zahr" :
- ▶ 17ième siècle : **mathématisation** par notamment Fermat et Pascal
- ▶ 18ième siècle : **intérêt croissant** avec comme contributeurs : Laplace, Poisson, Gauss, Poincaré, Borel, Fréchet, Levy, Kolmogorov, Khintchine, ...
- ▶ 1933 : **axiomatisation** moderne des probabilités par Kolmogorov

# Origines

- ▶ Origine du calcul de probabilités : jeu de hasard. De l'arabe "az-zahr" : dé à jouer
- ▶ 17ième siècle : **mathématisation** par notamment Fermat et Pascal
- ▶ 18ième siècle : **intérêt croissant** avec comme contributeurs : Laplace, Poisson, Gauss, Poincaré, Borel, Fréchet, Levy, Kolmogorov, Khintchine, ...
- ▶ 1933 : **axiomatisation** moderne des probabilités par Kolmogorov

# Univers

- ▶ L'univers,  $\Omega$ , est **l'ensemble des possibles**, des éventualités.
- ▶ La notion d'éventualité dépend de ce qui **intéresse l'expérimentateur**.

Pour le lancé d'un dé équilibré,

- ▶ si on s'intéresse au résultat de la face supérieure, on prendra pour univers

$$\Omega = \{1, 2, 3, 4, 5, 6\}$$

- ▶ si on s'intéresse à la parité du numéro de cette face, on prendra

$$\Omega = \{Pair, Impair\} := \{0, 1\}$$

# Univers

L'univers peut être

- ▶ Fini :

$$\{pile, face\}$$

$$\{Ironchain, Steelchain, Polyester, PEHD\}$$

$$\{1, \dots, n\}$$

- ▶ Infini dénombrable :

$$\mathbb{N}, \mathbb{Z}, \mathbb{N}^2 \dots$$

- ▶ Infini non-dénombrable :

$$[0, 1], \mathbb{R}, \mathbb{R}^2, \dots$$

# Les évènements de l'univers

Ensemble	Événement
L'ensemble vide $\emptyset$	Événement impossible
L'univers $\Omega$	Événement certain
Un singleton $\{\omega\}$ où $\omega \in \Omega$	Un événement élémentaire
Un sous-sensemble $A$ de $\Omega$	Un événement
$\omega \in A$	Le résultat $\omega$ est une réalisation possible de $A$
$A \subset B$	si $A$ est réalisé, alors $B$ est réalisé
Le complémentaire $\Omega \setminus A$ de $A$ dans $\Omega$	Événement contraire de $A$
$A \cap B$	Réalisation simultanée de $A$ et $B$
$A \cap B = \emptyset$	Les événements $A$ et $B$ sont incompatibles
$A \cup B$	Réalisation de $A$ ou $B$
$(A_i)_{i \in I}$ une partition dénombrable de $\Omega$	$(A_i)_{i \in I}$ est un système complet d'événements

(tableau extrait d'un cours de Jean-Etienne Rombaldi : <https://www-fourier.ujf-grenoble.fr/~rombaldi>)

# Les tribus de l'univers

- ▶ A présent nous allons définir une **mesure des évènements de l'univers**, pour cela il faut s'assurer que nous travaillerons avec des **objets mesurables**
- ▶ Pour un univers donné nous travaillerons sur une **tribu** c'est à dire l'ensemble  $\mathcal{B}$  tel que :
  - ▶  $\Omega \in \mathcal{B}$
  - ▶ stable par complémentaire
  - ▶ stable par union dénombrable d'événement

# Mesure de probabilité

Soit  $\Omega = \{pile, face\}$  et  $\mathcal{B} = \{\emptyset, pile, face, \Omega\}$

*Proba(pile) = ?*

# Mesure de probabilité

Soit  $\Omega = \{pile, face\}$  et  $\mathcal{B} = \{\emptyset, pile, face, \Omega\}$

$$Proba(pile) = 1/2$$

# Mesure de probabilité

Soit  $\Omega = \{pile, face\}$  et  $\mathcal{B} = \{\emptyset, pile, face, \Omega\}$

$$Proba(pile) = 1/2$$

mais pourquoi ?

# Mesure de probabilité

Soit  $\Omega = \{pile, face\}$  et  $\mathcal{B} = \{\emptyset, pile, face, \Omega\}$

$$Proba(pile) = 1/2$$

mais pourquoi ?

- ▶ Idée d'une probabilité : **intuition, expériences**
  - ▶ Exemple : Bridge nombre de possible mains  $52!/(13!)^4 \sim 10^{30}$  :  
Nous n'avons pas testé si toutes ces mains étaient équiprobables et nous ne le ferons pas.
- ▶ Besoin d'outils (statistique) plus avancés pour avoir accès à (une approximation de) la mesure de probabilité théorique.
- ▶ Apparaît comme un modèle que l'on souhaite proche d'une réalité dont seul l'expérience nous permet de mesurer la qualité

# Mesure de probabilité

## Approche fréquentiste

- ▶ Considérons une expérience aléatoire et un évènement aléatoire  $A \in \mathcal{B}$ . Intuitivement la probabilité que l'évènement  $A$  survienne peut être définie comme la fréquence de réalisation de cet évènement.
- ▶ Si je répète l'expérience un grand nombre  $n$  de fois et que je note  $n(A)$  le nombre de fois où  $A$  s'est réalisé on a naturellement envie d'écrire que

$$Proba(A) = \lim_{n \rightarrow +\infty} \frac{n(A)}{n}$$

# Mesure de probabilité

## Approche modélisation (Bayésienne)

- ▶ On ne peut pas toujours répéter une expérience
  - ▶ Impossible
    - ▶ probabilité qu'il y ait une vie extra-terrestre
    - ▶ probabilité que le soleil se lève demain
    - ▶ probabilité que je rate mon train du 23/03/2017 de 8h00
  - ▶ Peu raisonnable
    - ▶ probabilité qu'une cuve de central nucléaire se casse
    - ▶ probabilité que je me casse une jambe en sautant du 2-ième étage
    - ▶ heureusement il existe la simulation (potentiel retour fréquentiste)
- ▶ La probabilité vient modéliser une expérience (passé, multiple...) ou tout simplement est un choix subjectif
- ▶ elle peut aussi intégrer à la fois des données expérimentales et des choix a priori (approche bayésienne)

# Mesure de probabilité

Soit  $\Omega$  l'univers des possibles et  $A, B \subset \Omega$  deux évènements disjoints.

- ▶ **Mesure** de probabilité :

$$\mathbb{P} : \mathcal{B} \rightarrow [0, 1]$$

Telle que

- ▶  $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$
- ▶  $\mathbb{P}(\Omega) = 1$

## Subsection 2

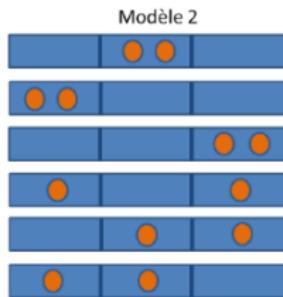
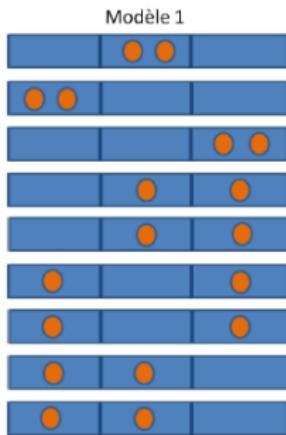
### Dénombrément et probabilité

# Des cellules et des balles

- ▶ Comment disposer  $r$  balles indiscernables dans  $n$  cellules ?
- ▶ On note  $r_k \geq 0$  le nombre de balles dans la cellule  $k$ , tel que

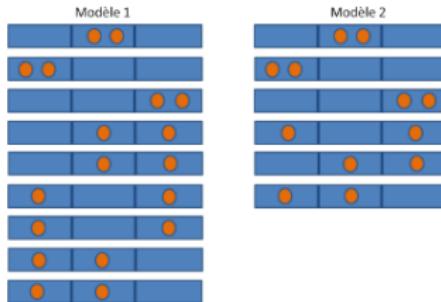
$$r_1 + \dots + r_n = r$$

Considérons l'exemple où  $n = 3$  et  $r = 2$



# Des cellules et des balles

Considérons l'exemple où  $n = 3$  et  $r = 2$



Probabilité que  $r_1 = 0, r_2 = 1, r_3 = 1$  ?

► Modèle 1

- Nombre de possibilités :  $n^r = 3^2 = 9$
- Nombre de cas favorable :  $\frac{r!}{r_1!r_2!r_3!} = \frac{2!}{0!1!1!} = 2$
- probabilité :  $2/9$

► Modèle 2

- Nombre de possibilités :  $C_{n+r-1}^r = \frac{4!}{2!2!} = 6$
- Nombre de cas favorables : 1
- probabilité :  $1/6$

# Des cellules et des balles

- ▶ Comment disposer  $r$  balles indiscernables dans  $n$  cellules avec la contrainte de n'avoir pas plus d'une balle par cellule ?



Probabilité que  $r_1 = 0, r_2 = 1, r_3 = 1$  ?

- ▶ Nombre de possibilités :  $C_n^r = \frac{3!}{2!1!} = 3$
- ▶ Nombre de cas favorables : 1
- ▶ probabilité : 1/3

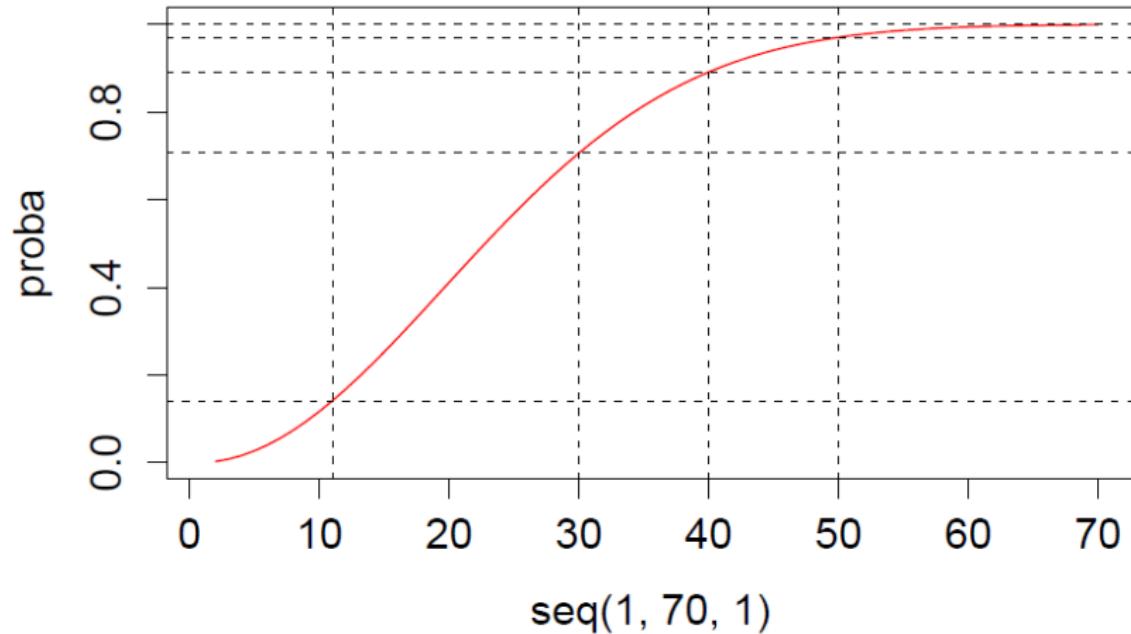
# Des cellules et des dés

- ▶ Sur 5 lancés de dé, quelle est la probabilité d'obtenir 5 fois de suite la face 6 ?

# Des cellules et des anniversaires

- ▶ Quelle est la probabilité que dans un groupe de 30 personnes au moins 2 d'entre eux aient le même jour d'anniversaire ?

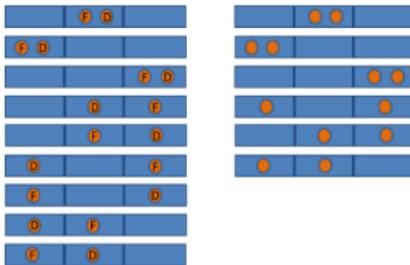
# Des cellules et des anniversaires



# Des cellules et des accidents

Delphine et Frédéric partent en weekend de 3 jours à la montagne pour faire de la randonnée (les yeux bandés !). Sachant qu'ils ne survivront pas (pas très optimal tout cela),

- ▶ Quelle est la probabilité que Delphine meurt le même jour ou avant Frédéric ?
- ▶ Quelle est la probabilité que les deux randonneurs meurent des jours différents ?



# Des cellules et des particules

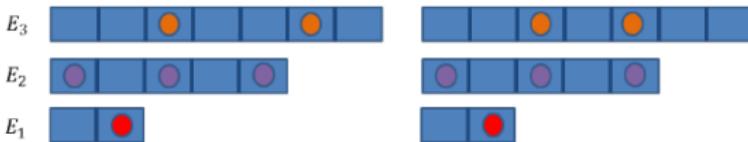
- ▶ Soit  $r$  particules indiscernables et  $n$  positions/états.
- ▶ Il y a  $n^r$  arrangements possible
- ▶ On note  $r_k \geq 0$  le nombre de particules dans l'état  $k$ , tel que

$$r_1 + \dots + r_n = r$$

# Des cellules et des particules

- ▶ Statistique de Maxwell-Boltzmann (très haute température ou faible concentration)
  - ▶ les  $n^r$  arrangements ont la même probabilité
  - ▶

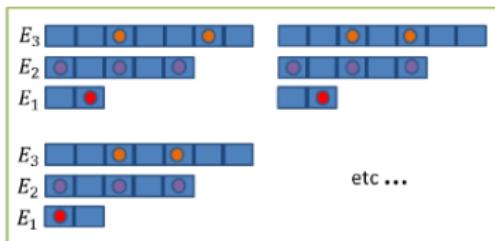
$$\frac{r!}{r_1! \dots r_n!} n^{-r}$$



# Des cellules et des particules

- ▶ Statistique de Bose-Einstein (photons, bosons, atomes (pair))
  - ▶ On ne compte que les arrangements distinguables
  - ▶

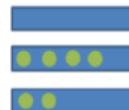
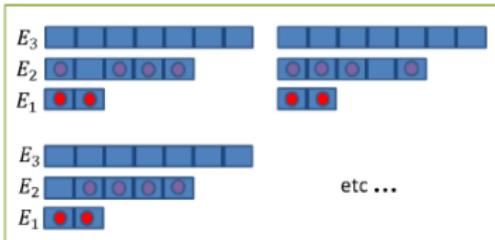
$$1/C_{n+r-1}^r$$



$$n = 3$$

$$r = 6$$

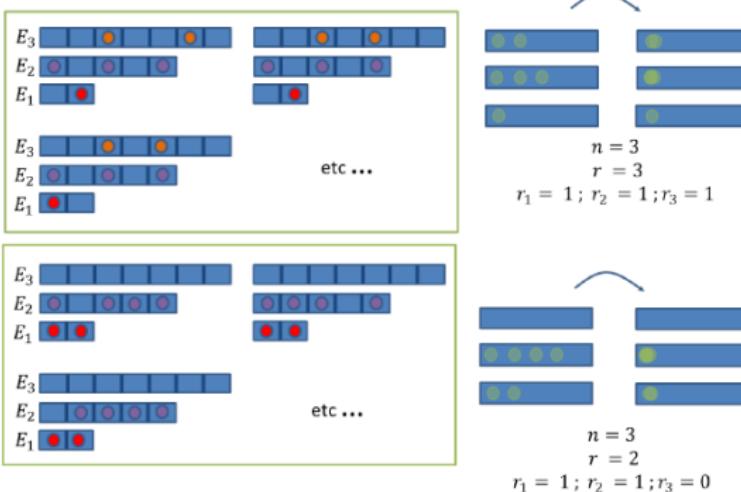
$$r_1 = 1; r_2 = 3; r_3 = 2$$



# Des cellules et des particules

- ▶ Statistique de Fermi-Dirac (electrons, neutrons, protons)
  - ▶ pas plus d'une particule dans le même état ( $r \leq n$  et  $r_i = 0$  ou 1)
  - ▶

$$1/C_n^r$$



### Subsection 3

Probabilité conditionnelle et indépendance

# Probabilité Conditionnelle

- Soit  $A$  et  $B$  deux évènements de l'univers  $\Omega$  de cardinal fini.

Par définition

$$\mathbb{P}(A \cap B) = \frac{|A \cap B|}{|\Omega|} = \frac{|A \cap B|}{|B|} \frac{|B|}{|\Omega|} = \frac{|A \cap B|}{|B|} \mathbb{P}(B)$$

Naturellement

$$\mathbb{P}(A|B) := \frac{|A \cap B|}{|B|} = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}$$

## Formule de Bayes

- ▶ Pour  $\Omega$  quelconque on définit la probabilité conditionnelle comme

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}$$

- ▶ ou encore

$$\mathbb{P}(A \cap B) = \mathbb{P}(A|B) \times \mathbb{P}(B)$$

## Formule de Bayes

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(B|A)\mathbb{P}(A)}{\mathbb{P}(B)}$$

# Indépendance d'évènements

- Deux évènements  $A, B$  de  $\Omega$  sont dit indépendant si et seulement si

$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$$

## Formule des probabilités totales

- Soit  $\Omega$  quelconque et soit un système exhaustif (partition) de  $\Omega$  d'évènements de probabilités non nuls :

$$(A_i)_{i \in I}$$

avec  $I$  fini ou dénombrable alors pour  $B$  un évènement quelconque

$$\mathbb{P}(B) = \sum_{i \in I} \mathbb{P}(B|A_i)\mathbb{P}(A_i)$$

# Formule de Bayes

On peut également écrire

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(B|A)\mathbb{P}(A)}{\mathbb{P}(B)} = \frac{\mathbb{P}(B|A)\mathbb{P}(A)}{\mathbb{P}(B|A)\mathbb{P}(A) + \mathbb{P}(B|A^c)\mathbb{P}(A^c)}$$

Plus généralement

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(B|A)\mathbb{P}(A)}{\sum_{i \in I} \mathbb{P}(B|A_i)\mathbb{P}(A_i)}$$

## Subsection 4

Exemple du jeu Monty Hall

# Monty Hall Problem

- ▶ Supposons faire face à 3 portes.
- ▶ Derrière 1 porte se cache 1 million d'euros et derrière les 2 autres des chèvres.
- ▶ Le joueur sélectionne 1 porte.
- ▶ Le présentateur ouvre 1 porte sans aucune préférence (la porte au million n'est évidemment jamais ouverte par le présentateur).

Faut-il changer de porte ou garder sa porte ?

# Super Monty Hall problem simulation

- ▶ Supposons  $k$  portes. Derrière 1 porte se cache 1 millions d'euros et derrière les  $k - 1$  autres des chèvres.
- ▶ Le joueur peut sélectionner  $n$  portes.
- ▶ Le présentateur ouvre  $m$  portes sans aucune préférence (la porte au million n'est évidemment jamais ouverte par le présentateur).

Faut-il changer de groupe de porte ? Implémentez une version simulée du problème général et répondez à la question pour  
 $(k, n, m) = (10, 5, 2)$  via une simulation numérique du problème.

# Table des matières

Probabilités

Variables aléatoires

Densité et expérance  
conditionnelle

Distributions Classiques

Types de convergence

Exercices d'application du TCL

Simulation de variables aléatoires

Statistique descriptive

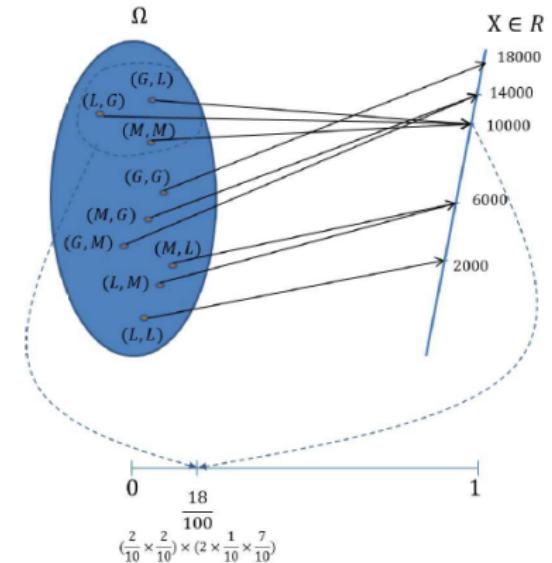
Statistique Inférentielle

Approche Bayésienne

# Variable aléatoire

- ▶ L'assureur de Delphine et Frédéric constate qu'il y a deux accidents par jour (en moyenne) qui peuvent être :
  - ▶ Léger (L) : coût de 1000 euros
  - ▶ Moyen (M) : coût de 5000 euros
  - ▶ Grave (G) : coût de 9000 euros
- ▶ Les deux accidents seront supposés indépendants l'un de l'autre
- ▶ Il veut connaître la probabilité d'avoir à débourser un certains montant
- ▶  $X$  est la variable (aléatoire) correspondant au montant à débourser en une journée

# Variable aléatoire



$$\mathbb{P}(L) = \frac{7}{10}, \mathbb{P}(M) = \frac{2}{10}, \mathbb{P}(G) = \frac{1}{10}$$

$$\mathbb{P}(X = 2000) = \frac{49}{100}, \mathbb{P}(X = 6000) = \frac{28}{100}, \mathbb{P}(X = 10000) = \frac{18}{100}, \mathbb{P}(X = 14000) = \frac{4}{100}, \mathbb{P}(X = 18000) = \frac{1}{100}$$

# Variable aléatoire

- ▶ **Variable aléatoire** a.k.a fonction mesurable

- ▶ discrète :

$X : \Omega \rightarrow S \sim \{1, \dots, n\}$  telle que

$$X^{-1}(\text{partie de } S) \in \mathcal{B}$$

- ▶ discrète :

$X : \Omega \rightarrow \mathbb{N}$  telle que

$$X^{-1}(\text{partie de } \mathbb{N}) \in \mathcal{B}$$

- ▶ continue :

$X : \Omega \rightarrow \mathbb{R}$  telle que  $\forall a \in \mathbb{R}$ ,

$$X^{-1}(]-\infty, a]) \in \mathcal{B}$$

# Caractérisation d'une v.a. discrète

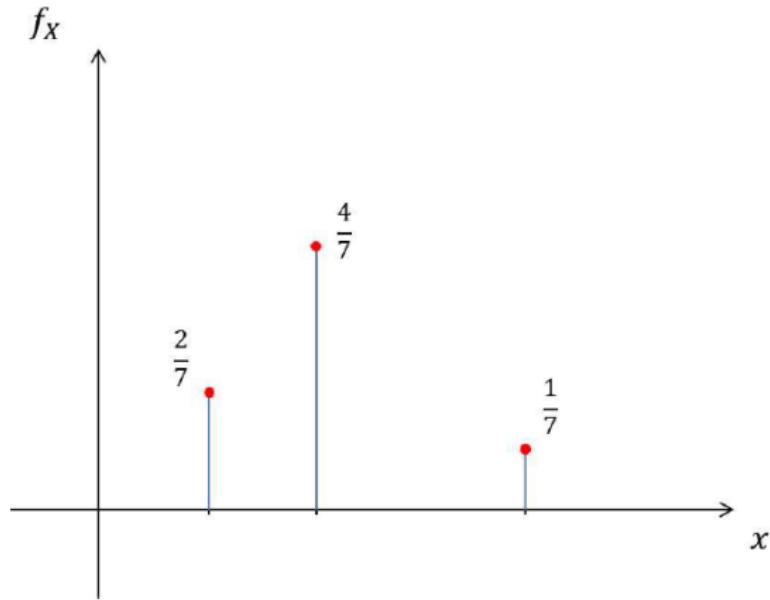
- ▶ **densité de distribution :**

$$\begin{aligned}f_X(x) &= \mathbb{P}_X(\{x\}) \\&= \mathbb{P}(X = x) \\&= \mathbb{P}(X^{-1}(x)) \\&= \mathbb{P}(\{\omega \in \Omega, X(\omega) = x\})\end{aligned}$$

- ▶ **densité de distribution (equiprobabilité) :**

$$f_X(x) = \mathbb{P}(X = x) = \frac{\text{Card}(\{\omega \in \Omega, X(\omega) = x\})}{\text{Card}(\Omega)}$$

## Caractérisation d'une v.a. discrète

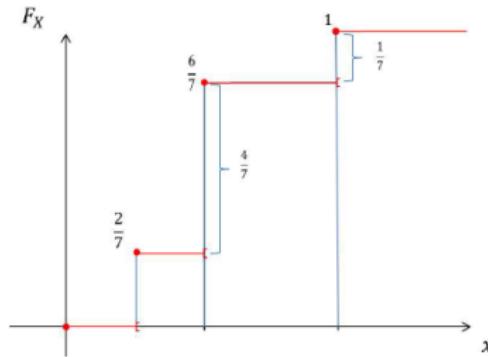


# Caractérisation d'une v.a. discrète

## ► Fonction de répartition

►  $F_X : \mathbb{R} \rightarrow [0, 1]$  such that

$$\begin{aligned} F_X(x) &= \mathbb{P}(X \leq x) \\ &= \sum_{i:x_i \leq x} f_X(x_i) \\ &= \sum_{i:x_i \leq x} \mathbb{P}(X = x_i) \end{aligned}$$



# Caractérisation d'une v.a. continue

- ▶ **Fonction de répartition**

- ▶  $F_X : \mathbb{R} \rightarrow [0, 1]$  such that

$$F_X(x) = \mathbb{P}(X \leq x) = \mathbb{P}(X^{-1}(]-\infty, x]))$$

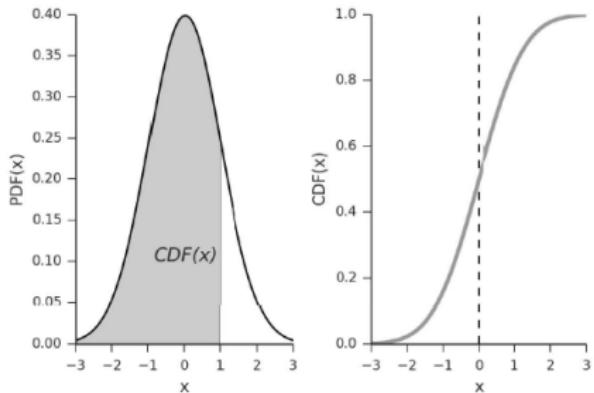
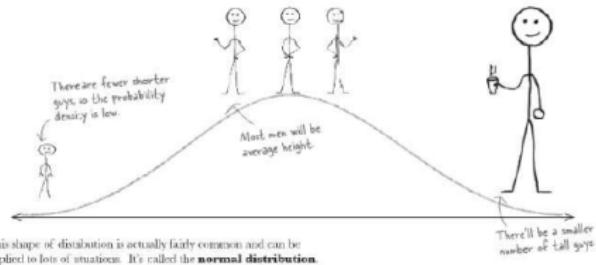
- ▶ **Densité de distribution** (elle n'existe pas toujours)

$$\begin{aligned} f_X(x) &= \lim_{\epsilon \rightarrow 0} \frac{\mathbb{P}(x \leq X \leq x + \epsilon)}{\epsilon} \\ &= \lim_{\epsilon \rightarrow 0} \frac{F_X(x + \epsilon) - F_X(x)}{\epsilon} \\ &= F'_X(x) \end{aligned}$$

- ▶ ou encore

$$F_X(x) = \int_{-\infty}^x f_X(t) dt$$

# Caractérisation d'une v.a. continue



Sources images : Google image

## Caractérisation d'une v.a. continue

- ▶ On notera en minuscule la **réalisation d'une variable aléatoire** : c'est à dire la valeur prise par cette dernière à l'issue d'une expérience
- ▶  $X$  est la variable aléatoire et  $x$  une réalisation de la variable aléatoire  $X$  ou encore  $x_1, \dots, x_n$  sont  $n$  réalisations de la variable aléatoire

Exemple :

- ▶  $X$  : résultat d'un tirage pile ou face. On suppose les conditions expérimentales inchangés pendant  $n = 4$  répétitions de l'expérience :

$$x_1 = 1 \quad x_2 = 1 \quad x_3 = 0 \quad x_4 = 1$$

## Caractérisation d'une v.a. continue

- ▶ On peut également considérer des **copies**  $X_1, \dots, X_n$  d'une variable aléatoire  $X$  : on dira quelles sont **identiquement distribuées** (i.d.)
- ▶ l'indice permet de prendre en compte la répétition de l'expérience : le résultat de la  $i$ -ième expérience est aléatoire :  $X_i$ , et après expérience nous aurons observé  $x_i$
- ▶ Les expériences peuvent être i.d. ou non mais aussi **indépendante ou non.**

# Indépendance de variables aléatoires continues

- Soient  $X$  et  $Y$  deux variables aléatoires de densité jointe  $f_{X,Y}$  et de densités marginales  $f_X$  et  $f_Y$  alors  $X$  et  $Y$  sont indépendantes si

$$f_{X,Y}(x, y) = f_X(x)f_Y(y)$$

- Plus généralement si une suite  $X_1, \dots, X_n$  de variables aléatoires sont mutuellement indépendantes si

$$f_{X_1, \dots, X_n}(x_1, \dots, x_n) = \prod_{i=1}^n f_{X_i}(x_i)$$

# Espérance et variance d'une v.a.

## ► Espérance d'une v.a. X

$$\mathbb{E}(X) = \int_{\mathbb{R}} xf_X(x)dx$$

$$\mathbb{E}(h(X)) = \int_{\mathbb{R}} h(x)f_X(x)dx$$

## ► Linéarité de l'espérance

$$\mathbb{E}(X + Y) = \mathbb{E}(X) + \mathbb{E}(Y)$$

## ► Variance d'une v.a. X

$$Var(X) = \mathbb{E}\left((X - \mathbb{E}(X))^2\right)$$

# Variance et Covariance

## ► Covariance

$$Cov(X, Y) = \mathbb{E}((X - \mathbb{E}(X))(Y - \mathbb{E}(Y))) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)$$

## ► Propriétés de la variance

$$Var(a + X) = Var(X)$$

$$Var(aX) = a^2Var(X)$$

$$Var(X + Y) = Var(X) + Var(Y) + 2cov(X, Y)$$

## ► Matrice de covariance

$$Var(\mathbf{X}) = [Cov(X_i, X_i)]_{i,j}$$

# Covariance et corrélation

## ► Covariance

$$Cov(X, Y) = \mathbb{E}((X - \mathbb{E}(X))(Y - \mathbb{E}(Y)))$$

## ► Corrélation de Pearson

$$\rho = \frac{Cov(X, Y)}{\sigma_X \sigma_Y}$$

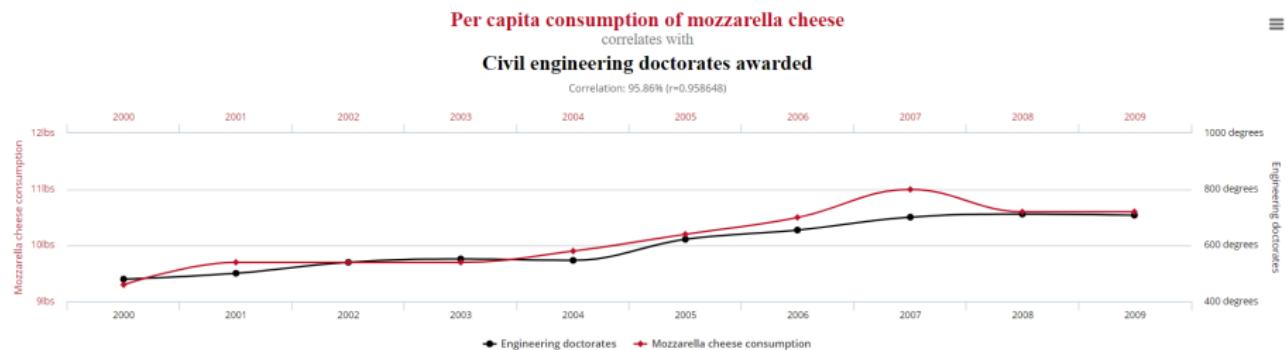
# Covariance et indépendance

## Cas général

- ▶ Indépendance  $\Rightarrow \text{Cov}(X, Y) = 0$
- ▶  $\text{Cov}(X, Y) = 0 \not\Rightarrow$  indépendance

## Cas de variables Gaussiennes

Indépendance  $\Leftrightarrow \text{Cov}(X, Y) = 0$



<https://www.tylervigen.com/spurious-correlations>

# Covariance et indépendance

Exercice 1 : Soit  $X$  une variable aléatoire normal centrée réduite et  $Y = X^2$ . Générer un échantillon de  $X$  et  $Y$ , le graphe 2d de l'échantillon et estimer/calculer la covariance du couple.

Exercice 2 : Imaginer deux variables aléatoires

## Autres caractéristiques d'un couple de v.a.

- ▶ Coefficient de corrélation de Spearman
- ▶ Mesure de Kendal

# Moments

- On appelle moments d'ordre  $r$  d'une variable aléatoire la quantité

$$\mathbb{E}(X^r) = \int x^r f_X(x) dx$$

- version centrée :

$$\mathbb{E}[(X - \mathbb{E}(X))^r]$$

- version centrée réduite :

$$\mathbb{E}\left[\left(\frac{X - \mathbb{E}(X)}{\sigma}\right)^r\right]$$

# Moments

## ► Coefficient d'assymétrie (Skewness)

$$\mathbb{E}\left[\left(\frac{X - \mathbb{E}(X)}{\sigma}\right)^3\right]$$

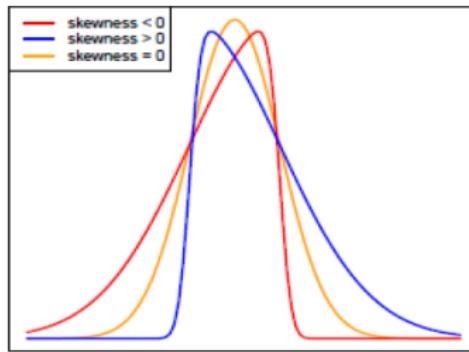


FIGURE – Exemples de distributions et leurs coefficients d'assymétries

# Moments

- ▶ Coefficient d'aplatissement des extrêmes (Kurtosis)

$$\mathbb{E}\left[\left(\frac{X - \mathbb{E}(X)}{\sigma}\right)^4\right]$$

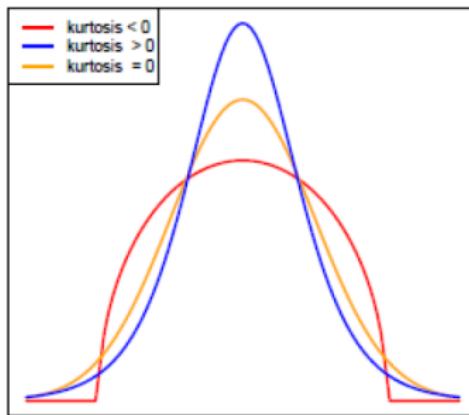


FIGURE – Exemples de distribution et leurs Kurtosis. Les Kurtosis sont centrés par rapport au Kurtosis d'une loi normale qui vaut 3.

# Quantiles

- ▶ On appelle quantile d'ordre  $\alpha \in [0, 1]$  d'une variable aléatoire  $X$  le scalaire  $q_\alpha$  tel que

$$\mathbb{P}(X \leq q_\alpha) = F_X(q_\alpha) \geq \alpha$$

- ▶ La médiane est le quantile d'ordre  $\alpha = 1/2$
- ▶ le premier et troisième quartiles sont respectivement  $q_{0.25}$  et  $q_{0.75}$

## Subsection 1

Densité et expérance conditionnelle

## Extension formule de Bayes au cas de variables aléatoires continues

- Soient  $X$  et  $Y$  deux variables aléatoires de densité jointe  $f_{X,Y}$  et de densités marginales  $f_X$  et  $f_Y > 0$  alors on définit la densité conditionnelle de  $X|Y$  comme

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

$$\left( \mathbb{P}(A|B) = \frac{\mathbb{P}(B|A)\mathbb{P}(A)}{\sum_{i \in I} \mathbb{P}(B|A_i)\mathbb{P}(A_i)} \right)$$

Attention aux hypothèses pour que ces quantités soient bien définie.

# Espérance conditionnelle

- ▶ Espérance conditionnelle à un évènement

$$\mathbb{E}(X|B) = \frac{1}{\mathbb{P}(B)} \mathbb{E}(X \mathbf{1}_B)$$

- ▶ Espérance conditionnelle à une variable aléatoire

$$\mathbb{E}(X|Y = y) = \int x f_{X|Y}(x|y) dx = \frac{1}{f_Y(y)} \int x f_{X,Y}(x,y) dx = \varphi(y)$$

et

$$\mathbb{E}(X|Y) = \varphi(Y)$$

Attention aux hypothèses pour que ces quantités soient bien définies.

# Espérance conditionnelle

- ▶ Cas d'une variable aléatoire carré intégrable i.e.  $\mathbb{E}(X^2) < \infty$  On peut dans ce cadre définir l'espérance conditionnelle comme

$$\mathbb{E}(X|Y) = \arg \min_{\varphi \in \mathcal{M}} \mathbb{E}[(X - \varphi(Y))^2]$$

où  $\mathcal{M}$  est l'ensemble des fonctions mesurables de  $\mathbb{R}^d \rightarrow \mathbb{R}$

# Espérance conditionnelle

- ▶ *Exemple de l'analyse de sensibilité*
- ▶ *Exemple d'une somme aléatoire à nombre de termes aléatoire*

## Subsection 2

### Distributions Classiques

# Introduction

- ▶ Quelques **densités de distribution classique** : uniforme, triangulaire, Gaussienne, Chi-2, Student, Fisher...
- ▶ Distribution **Gaussienne**

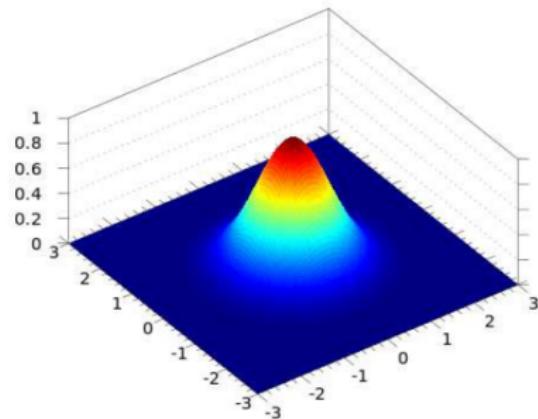
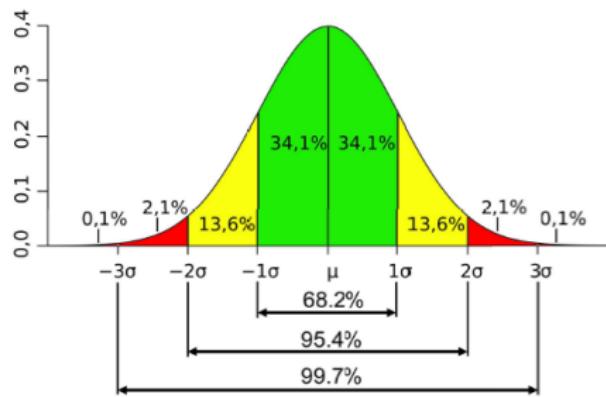
$$\mathcal{N}(m, \sigma^2) \sim f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(\frac{-(x-m)^2}{2\sigma^2}\right)$$

$$f(\mathbf{x}) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left(\frac{-(\mathbf{x}-\mathbf{m})^t \Sigma^{-1} (\mathbf{x}-\mathbf{m})}{2}\right)$$

- ▶ **Chi-2**  $\sim \sum_{i=1}^k X_i^2$  with  $X_i \sim$  Gaussienne
- ▶ **Student**  $\sim \frac{\text{Gaussienne}}{\text{Chi-2}}$
- ▶ **Fisher**  $\sim \frac{\text{Chi-2}}{\text{Chi-2}}$

# La Loi Gaussienne a.k.a. Loi Normale

Personne ne peut ignorer la loi (Normal) !



Source des images :

[http://www.muelaner.com/wp-content/uploads/2013/07/Standard\\_deviation\\_diagram.png](http://www.muelaner.com/wp-content/uploads/2013/07/Standard_deviation_diagram.png)

[https://en.wikipedia.org/wiki/Gaussian\\_function](https://en.wikipedia.org/wiki/Gaussian_function)

# Loi Uniforme discrète

- $X \sim \mathcal{U}(\{1, \dots, n\})$  alors  $\mathbb{P}(X = k) = 1/n$

$$\mathbb{E}(X) = (n + 1)/2 \quad Var(X) = (n^2 - 1)/12$$

## Loi de Bernoulli

- $X \sim \mathcal{B}(p)$  : prend valeur 1 ou 0 avec respectivement probabilité  $p$  et  $1 - p$  :

$$E(X) = p \quad Var(X) = p(1 - p)$$

# Loi Binomiale

- $X \sim \mathcal{B}(n, p)$  : définit comme la somme de  $n$  répétition indépendante d'une Bernoulli

$$X = \sum_{i=1}^n X_i \quad \mathbb{P}(X = k) = C_n^k p^k (1-p)^{n-k}$$

$$\mathbb{E}(X) = np \quad Var(X) = np(1-p)$$

- Exemple d'un sondage binaire

# Loi de Poisson

- $X \sim \mathcal{P}(\lambda)$  : loi du nombre d'occurrences d'évènements "rares", sans mémoire et dans un intervalle de temps donné.

$$\mathbb{P}(X = k) = \frac{e^{-\lambda} \lambda^k}{k!} \quad \mathbb{E}(X) = Var(X) = \lambda$$

- Exemple du nombre de personnes dans une file, nombre d'appels à un standard

# LoiMultinomiale

- ▶ Répartition des résultats de  $n$  répétition d'une expérience ayant  $m$  résultats possibles.
- ▶ Exemple d'un lancé de  $n$  dés

$$\mathbb{P}(N_1 = n_1, \dots, N_m = n_m) = \frac{n!}{n_1! \dots n_m!} p_1^{n_1} \dots p_m^{n_m}$$

# Loi uniforme continue

$\mathcal{U}([a, b])$

- ▶ Valeurs équiprobables sur un intervalle
- ▶ 2 paramètres :  $a$  et  $b$
- ▶ Exemple en modélisation : connaissance d'un intervalle de variation pour un paramètre

Fonction de répartition	$F(x) = \frac{x-a}{b-a}$
Fonction de densité	$f(x) = \frac{1}{b-a}$
Moyenne	$\frac{a+b}{2}$
Variance	$\frac{(b-a)^2}{12}$
Asymétrie	0
Aplatissement	1,8

```
# import uniform distribution
from scipy.stats import uniform
```

# Loi Exponentielle

## Loi Exponentielle

- ▶ densité de probabilité décroissant exponentiellement
- ▶ 1 paramètre (moyenne) ou 2 si on considère décalage  $\gamma$
- ▶ Exemples : temps d'attente, durée de vie de systèmes sans usure i.e. la proportion de matériels défaillants est chaque année la même.

Fonction de répartition	$F(x) = 1 - \exp[-\lambda(x - \gamma)]$
Fonction de densité	$f(x) = \lambda \exp[-\lambda(x - \gamma)]$
Moyenne	$\frac{1}{\lambda} + \gamma$
Variance	$\frac{1}{\lambda^2}$
Asymétrie	2
Aplatissement	9

# Loi Weibull

- ▶ généralisation de la loi exponentielle
- ▶ les fortes valeurs restent probables
- ▶ 2 paramètres existe avec 3 paramètres
- ▶ Exemple : durée de vie d'un matériel vieillissant : les défaillances sont rares les premières années puis deviennent de plus en plus fréquentes

Fonction de répartition	$F(x) = 1 - \exp\left[-\left(\frac{x}{\lambda}\right)^k\right]$
Fonction de densité	$f(x) = \left(\frac{k}{\lambda}\right)\left(\frac{x-\theta}{\lambda}\right)^{k-1} \exp\left[-\left(\frac{x-\theta}{\lambda}\right)^k\right]$
Moyenne	$\lambda\Gamma\left(1 + \frac{1}{k}\right) + \theta$
Variance	$\lambda^2\left(\Gamma\left(1 + \frac{2}{k}\right) - \mu^2\right)$
Asymétrie	$\frac{1}{\sigma^3}\left(\lambda^3\Gamma\left(1 + \frac{3}{k}\right) - 3\mu\sigma^2 - \mu^3\right)$
Aplatissement	$\frac{1}{\sigma^4}\left(\lambda^4\Gamma\left(1 + \frac{4}{k}\right) - 4\delta\mu\sigma^3 - 6\mu^2\sigma^2 - \mu^4\right)$

# Loi Normale ou Gaussienne

- ▶ densité de probabilité symétrique autour de la valeur moyenne
- ▶ moyenne est aussi le mode i.e. la valeur la plus probable
- ▶ 2 paramètres :  $\mu$  et  $\sigma^2$
- ▶ Exemples : impacts des boulets de canon (Jouffret, 1872), incertitude de mesure

Fonction de répartition	$F(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^x e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} ds = \Phi\left(\frac{x-\mu}{\sigma}\right)$
Fonction de densité	$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} = \frac{1}{\sigma} \varphi\left(\frac{x-\mu}{\sigma}\right)$
Moyenne	$\mu$
Variance	$\sigma^2$
Asymétrie	0
Aplatissement	3

# Lois Log-Normale

- ▶ variables positives et asymétriques
- ▶ Exemples : poids, salaires, résolution d'un instrument (sources d'erreur = multiplication d'un grand nombre de petits facteurs indépendants)

Fonction de répartition	$F(x) = \Phi\left(\frac{\ln(x-\gamma)-\lambda}{\zeta}\right)$
Fonction de densité	$f(x) = \frac{1}{\zeta \sqrt{2\pi} (x-\gamma)} e^{-\frac{1}{2} \frac{[\ln(x-\gamma)-\lambda]^2}{\zeta^2}}$
Moyenne	$\exp\left(\lambda + \frac{\zeta^2}{2}\right) + \gamma$
Variance	$(\mu - \gamma)^2 \left( \exp(\zeta^2) - 1 \right)$
Asymétrie	$\sqrt{\exp(\zeta^2) - 1} \left( \exp(\zeta^2) - 2 \right)$
Aplatissement	$\exp(4\zeta^2) + 2\exp(3\zeta^2) + 3\exp(2\zeta^2) - 3$

# Loi du Chi-2



Fonction de répartition	$F(x) = \frac{\gamma(\nu/2, x/2)}{\Gamma(\nu/2)}$
Fonction de densité	$f(x) = \frac{2^{-\nu/2}}{\Gamma(\nu/2)} x^{\nu/2-1} \exp\left(-\frac{x}{2}\right)$
Moyenne	$\nu$
Variance	$2\nu$
Asymétrie	$\sqrt{\frac{8}{\nu}}$
Aplatissement	$\frac{12}{\nu}$

# Loi de Student



Fonction de répartition	$F(x) = F_{\text{Student}} \left( \nu, \frac{x-\mu}{\sigma} \right)$
Fonction de densité	$f(x) = \frac{1}{\sigma \sqrt{\nu \pi}} \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})} \left( 1 + \frac{(x-\mu)^2}{\nu} \right)^{-\frac{\nu+1}{2}}$
Moyenne ( $1 < k$ )	$\mu$
Variance ( $2 < k$ )	$\frac{\nu}{\nu-2} \sigma^2$
Asymétrie ( $3 < k$ )	0
Aplatissement ( $4 < k$ )	$\frac{6}{\nu-4} + 3$

# Loi Gamma

## Loi Gamma



Fonction de répartition	$F(x) = \frac{\gamma(k, \lambda x)}{\Gamma(k)}$
Fonction de densité	$f(x) = \frac{\lambda}{\Gamma(k)} (\lambda(x-\gamma))^{k-1} \exp[-\lambda(x-\gamma)]$
Moyenne	$\frac{k}{\lambda}$
Variance	$\frac{k}{\lambda^2}$
Asymétrie	$\frac{2}{\sqrt{k}}$
Aplatissement	$\frac{3(k+2)}{k}$

### Subsection 3

#### Types de convergence

# Convergence presque sûre

- Une suite de v.a.  $X_n$  converge presque sûrement vers une variable aléatoire  $X$  ssi

$$\mathbb{P}(\{\omega \in \Omega, X_n(\omega) \rightarrow X(\omega)\}) = 1$$

# Lemme de Borrel-Cantelli

- ▶ Si  $\forall \epsilon > 0$

$$\lim_{n \rightarrow \infty} \sum_{i=1}^n \mathbb{P}(|X_n - X| > \epsilon) < \infty$$

alors  $X_n$  converge presque sûrement vers  $X$

- ▶ la réciproque est vraie uniquement lorsque les  $X_i$  sont indépendantes.

# Convergence en probabilité

- ▶ Une suite de v.a.  $X_n$  converge en probabilité vers une variable aléatoire  $X$  ssi

$$\forall \epsilon > 0 \lim_{n \rightarrow \infty} \mathbb{P}(|X_n - X| > \epsilon) = 0$$

# Convergence en loi

- Une suite de v.a.  $X_n$  converge en loi vers une variable aléatoire  $X$  ssi

$$\lim_{n \rightarrow \infty} F_{X_n}(x) = F_X(x)$$

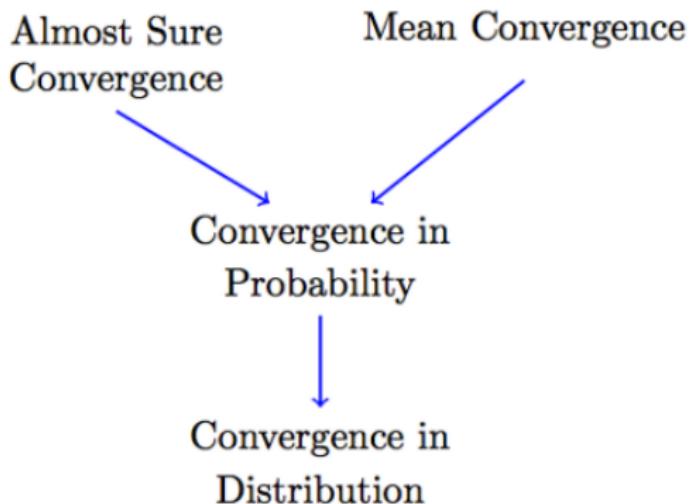
$\forall x$  où  $F$  est continue.

## Convergence en norme $L_r$

- ▶ Une suite de v.a.  $X_n$  converge en moyenne d'ordre  $r$  ou encore en norme  $L_r$  vers une variable aléatoire  $X$  si  $\mathbb{E}|X_n|^r < \infty$  et

$$\lim_{n \rightarrow \infty} \mathbb{E}[(X_n - X)^r] = 0$$

# Implications directes entre types de convergences



# Mapping theorem

- ▶ Pour toute fonction continue  $g$ , la convergence en probabilité, en loi et presque sûrement de  $X_n$  vers  $X$  implique la convergence de  $g(X_n)$  vers  $g(X)$ .

# Fonction caractéristique

- ▶ La fonction caractéristique d'un vecteur aléatoire est donnée par

$$\phi(t) = \mathbb{E}(e^{i\langle t, \mathbf{X} \rangle})$$

- ▶ Propriétés, Théorème de convergence de Levy

# Loi Forte des Grands Nombres

- Soit  $(X_i)_{i \geq 1}$  une suite de variables aléatoires i.i.d. selon  $\mathbb{X}$  alors

$$\frac{1}{N} \sum_{i=1}^N X_i \xrightarrow[N \rightarrow +\infty]{p.s.} \mathbb{E}(X)$$

# Théorème Central Limite

- Soit  $(X_i)_{i \geq 1}$  une suite de variables aléatoires i.i.d. selon  $\mathcal{X}$  tel que

$$\mathbb{E}(X^2) < \infty$$

alors

$$\frac{\sqrt{N}}{\sigma} \left[ \frac{1}{N} \sum_{i=1}^N X_i - \mathbb{E}(X) \right] \xrightarrow[N \rightarrow +\infty]{\mathcal{L}} Z \sim \mathcal{N}(0, 1)$$

## Subsection 4

### Exercices d'application du TCL

# Exercices

Supposons modélisé le temps qu'une machine mets pour effectuer une  $i$ -ème simulation numérique par une variable aléatoire  $X_i$  de moyenne  $\mathbb{E}(X_i) = 10$  minutes et de variance  $Var(X_i) = 2$  minutes. Les temps entre chaque simulation sont indépendant.

Quel est la probabilité qu'une machine effectue 40 simulations en 7 heures ?

## Exercices

Le nombre d'accident dans une certaine ville est modélisé par une variable aléatoire de Poisson avec un taux moyen de 10 accidents pour jour. On suppose que les nombres d'accidents sur des jours différents sont indépendants. Quel est la probabilité qu'il y ait plus de 3800 accidents sur une année (prendre pour base 365 jours par an) ?

## Exercices

Dans un système de communication chaque mot de passe est constitué de 1000 bits. Du bruit pouvant s'insérer dans les canaux de communication on mesure que chaque bit a une probabilité d'erreur estimé à 0.1. Les erreurs sur les bits sont supposés indépendantes. Des codes correcteurs d'erreurs étant utilisés sur ce système, chaque mot de passe peut être décodé de manière fiable si moins de 125 erreurs sont présentes sinon le décodage échoue. Quel est la probabilité que le décodage échoue ?

# Table des matières

Probabilités

aléatoires

Approche par transformation

Approche par acceptation/rejet

Variables aléatoires

Statistique descriptive

Simulation de variables aléatoires

Statistique Inférentielle

Introduction

Générateur de nombres

Approche Bayésienne

## Subsection 1

### Introduction

# Introduction à la simulation

- ▶ On souhaite étudier un phénomène **sans mener des expériences réelles** : trop coûteux, trop dangereux, trop lent
- ▶ Les expériences/phénomènes sont quantifiés/mis en équations
- ▶ **On simule** la réalité de manière déterministe ou stochastique
- ▶ Nous devons **être capable d'avoir des réalisations des v.a.** sans faire d'expériences réelles

# Introduction à la simulation

Exemple : la loi d'Arrhenius

$$k = a \exp\left(-\frac{e}{rT}\right)$$

avec  $k$  la vitesse d'une réaction chimique,  $E$  l'énergie d'activation,  $T$  la température et  $A, R$  des constantes,

# Introduction à la simulation

$$k = a \exp\left(-\frac{e}{rT}\right)$$

- ▶ Température non connue exactement : c'est une variable aléatoire ainsi que  $k = f(\theta, T)$  avec  $\theta = (a, e, r)$
- ▶ Simuler la vitesse de réaction pour des valeurs données de  $\theta$  : besoin d'avoir de réalisations de la v.a. température

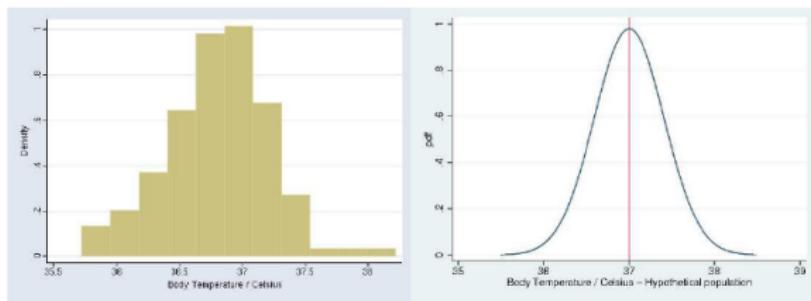
Quantités d'intérêt :

- ▶ vitesse moyenne
- ▶ si l'on a des données de vitesses on peut caler les autres paramètres en prenant en compte les incertitudes sur  $T$ ...

# Introduction à la simulation

Supposons la réaction chimique ayant lieu dans le corps humain

- ▶ Modélisation aléatoire de la température



Histogramme de la température corporelle de 130 personnes (gauche)  
distribution Gaussienne sous-jacente supposée (droite)

moyenne : 37 et écart-type : 0.407

Source des images :

<https://ww2.amstat.org/publications/jse/v4n2/datasets.shoemaker.html>

[https://en.wikibooks.org/wiki/Introduction\\_to\\_Medical\\_Statistics/Analysis\\_of\\_a\\_single\\_sample](https://en.wikibooks.org/wiki/Introduction_to_Medical_Statistics/Analysis_of_a_single_sample)

# Simulation d'une v.a.

"Aléa fort" :

- ▶ Utiliser des **processus physique** (chaotique, quantique) : lancer une pièce, un dé, Interroger des passants sur leur taille, prendre leur température
- ▶ Aléa de très **bonne qualité**
- ▶ Difficile/**Lent** à générer

# Simulation d'une v.a.

"Aléa faible" : Bon pour la **simulation intensive** (mais pas pour la cryptographie)

- ▶ Utiliser des procédés mathématiques/informatiques : par essence **déterministe**
- ▶ Génération de suite de nombres se **comportant statistiquement comme des variables aléatoires**
- ▶ Facile et **rapide** à générer

## Subsection 2

Générateur de nombres aléatoires

# Génération de nombres aléatoires

- ▶ Objectif premier : générer des nombres aléatoires **uniformément distribués** sur  $[0, 1]$
- ▶ Utilisation de **transformations** pour générer d'autres lois

Générer la suite déterministe d'entiers

$$x_{n+1} = ax_n + c \bmod b$$

Puis utiliser comme nombre pseudo aléatoire

$$u_n = x_n/b$$

On souhaite une suite de plus **grande période**

# Génération de nombres aléatoires

- ▶ Fonction rand de Scilab utilise les paramètres

$$a = 843314861 \ c = 453816693 \ \text{et} \ b = 2^{31}$$

- ▶ Anciennement Matlab utilisait le générateur **mcg16807** de période  $2^{31} - 2$  avec

$$a = 7^5 \ c = 0 \ \text{et} \ b = 2^{31} - 1$$

- ▶ Aujourd'hui Matlab/Python utilise le générateur **mt19937ar** de période  $2^{19937} - 1$  (basé sur le Mersenne Twister)

## Subsection 3

Approche par transformation

# Méthode de simulation

## Approche par transformation

- Générateur uniforme

Pour une fonction de répartition  $F$  définie sur  $\mathbb{R}$ , on définit son inverse généralisée par

$$F^{-1}(u) = \inf\{x; F(x) \geq u\}$$

Alors, si  $U$  est uniforme sur  $[0, 1]$ , la variable aléatoire  $F^{-1}(U)$  est de fonction de répartition  $F$  car

$$P[F^{-1}(U) \leq x] = P[U \leq F(x)] = F(x)$$

- Cette méthode nécessite de connaître l'inverse généralisée de la fonction de répartition.

# Méthode de simulation

## Approche par transformation

- Méthode de Box Müller

Si  $U_1$  et  $U_2$  sont deux variables indépendantes uniformes sur  $[0, 1]$ , alors

$$Y_1 = \sqrt{-2 \ln U_1} \cos(2\pi U_2)$$

$$Y_2 = \sqrt{-2 \ln U_1} \sin(2\pi U_2)$$

sont des variables iid distribuées suivant une loi  $\mathcal{N}(0, 1)$ .

- Loi de Poisson

Si  $X_i \sim \mathcal{E}(\lambda)$  et  $N \sim \mathcal{P}(\lambda)$  alors

$$P[N = k] = P[X_1 + \dots + X_k \leq 1 < X_1 + \dots + X_{k+1}]$$

## Subsection 4

Approche par acceptation/rejet

# Méthode de simulation

## Approche par rejet

- Beaucoup de lois sont difficiles à simuler **directement** avec les méthodes précédentes
- Il y a certaines applications où la loi à simuler  $f$  est connue **à une constante multiplicative près** (méthodes Bayésiennes)
- ☞ Une solution est de simuler à l'aide d'une loi de proposition  $g$  **plus simple** et d'utiliser un algorithme **d'acceptation-rejet**

# Méthode de simulation

## Approche par rejet

Soit une loi d'intérêt de densité  $f$  et une **loi de proposition** de densité  $g$  telle que

$$f(x) \leq Mg(x)$$

sur le support de  $f$ . Alors, on peut simuler suivant  $f$  avec l'algorithme suivant

- 1) **Générer**  $X \sim g$  et  $U \sim \mathcal{U}([0, 1])$
- 2) **Accepter**  $Y = X$  si

$$U \leq \frac{f(X)}{Mg(X)}$$

- 3) **Retourner en 1) si rejet**

# Méthode de simulation

## Approche par rejet

$$\begin{aligned} P[X \text{ accepté}] &= P\left[U \leq \frac{f(X)}{Mg(X)}\right] = E\left[\mathbb{I}_{\{U \leq \frac{f(X)}{Mg(X)}\}}\right] \\ &= E\left[E\left[\mathbb{I}_{\{U \leq \frac{f(X)}{Mg(X)}\}}\right] | X\right] \\ &= E\left[\frac{f(X)}{Mg(X)}\right] \\ &= \int \frac{f(x)}{Mg(x)} g(x) dx = \frac{1}{M} \end{aligned}$$

# Méthode de simulation

## Approche par rejet

$$\begin{aligned} P[X < x | X \text{ accepté}] &= \frac{P[X < x, X \text{ accepté}]}{1/M} \\ &= MP \left[ X < x, U < \frac{f(X)}{Mg(X)} \right] \\ &= ME \left[ \mathbb{I}_{\{X < x, U \leq \frac{f(X)}{Mg(X)}\}} \right] \\ &= ME \left[ E \left[ \mathbb{I}_{\{X < x, U \leq \frac{f(X)}{Mg(X)}\}} \right] | X \right] \\ &= ME \left[ \mathbb{I}_{\{X < x\}} \frac{f(X)}{Mg(X)} \right] \\ &= \int_{-\infty}^x \frac{f(x)}{g(x)} g(x) dx = F(x) \end{aligned}$$

# Méthode de simulation

## Approche par rejet

- Cet algorithme permet de simuler une densité connue à une const. multiplicative près, e.g.  $f(\theta|x) \propto f(x|\theta)\pi(\theta)$
- La probabilité d'acceptation est  $1/M$  donc la valeur de  $M$  règle l'efficacité (vitesse) de l'algorithme
- Problème pour des densités à queues lourdes. Par exemple, on ne peut simuler une loi de Cauchy avec une loi de proposition normale (mais on peut faire l'inverse !)
- Utilisable pour un grand nombre de lois :  $\mathcal{N}(0, 1)$ ,  $\mathcal{G}a(a, b)$ , lois normales tronquées, ...

# Méthode MCMC

Voir approche Bayésienne et simulation de la loi a posteriori

# Table des matières

Probabilités

Variables aléatoires

Simulation de variables aléatoires

Statistique descriptive  
Résumés statistiques  
Histogramme

Statistique Inférentielle  
Approche Bayésienne

## Subsection 1

### Résumés statistiques

# Statistique Descriptive

- ▶ Description synthétique des données (représentation graphique, résumé statistique, tableaux,...)
- ▶ Analyse de données (classification, analyse factorielle,...)
- ▶ **Pas de modèle probabiliste dans cet étape**

# Résumés statistiques

## ► Moyenne

$$\hat{m}_1 = \frac{1}{n} \sum_{i=1}^n x_i$$

## ► Variance

$$\hat{\sigma}^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - m_1)^2$$

## ► Ecart-type

$$\hat{\sigma} = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - m_1)^2}$$

# Résumés statistique

## ► Covariance Empirique

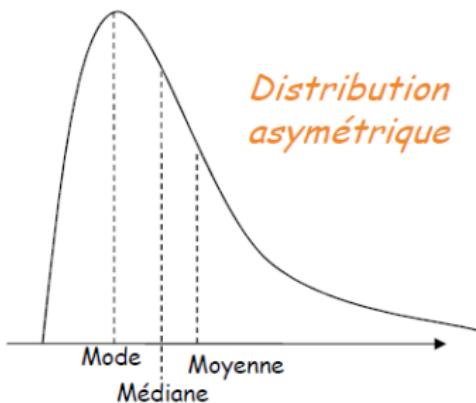
$$\frac{1}{n} \sum_{i=1}^n (x_i - \hat{m}_1(\mathbf{x}))(y_i - \hat{m}_1(\mathbf{y}))$$

## ► Corrélation de pearson

$$\frac{\sum_{i=1}^n (x_i - \hat{m}_1(\mathbf{x}))(y_i - \hat{m}_1(\mathbf{y}))}{\sqrt{\sum_{i=1}^n (x_i - \hat{m}_1(\mathbf{x}))^2 \sum_{i=1}^n (y_i - \hat{m}_1(\mathbf{y}))^2}}$$

# Résumés statistiques

- ▶ Skewness, Kurtosis : estimation standard ("plug-in") biaisé
- ▶ Médiane, quartiles : estimation à partir des rangs
- ▶ Mode (cas discret) : utiliser fréquence
- ▶ Mode (cas continue) : utiliser approximation par histogramme ou par maximisation de l'approximation de la densité à noyau



# Résumés statistiques

- ▶ Etendue

$$\max(x_1, \dots, x_n) - \min(x_1, \dots, x_n)$$

- ▶ Ecart absolu moyen

$$\frac{1}{n} \sum_{i=1}^n |x_i - m_1|$$

- ▶ Intervalle inter-quartile

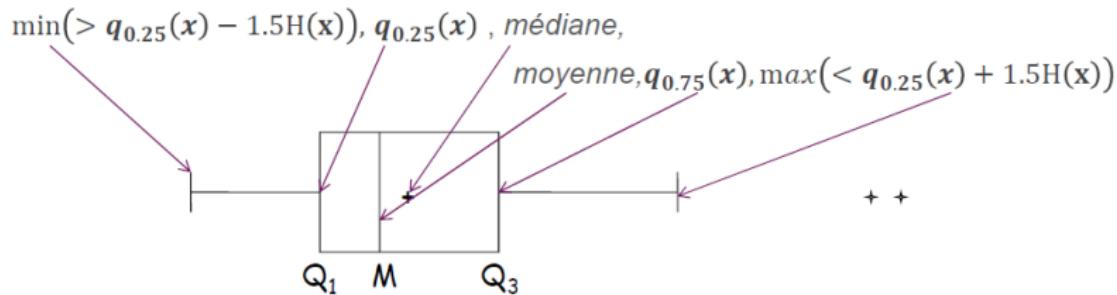
$$\hat{q}_{0.75} - \hat{q}_{0.25}$$

- ▶ Coefficient de variation

$$\delta = \frac{\hat{\sigma}}{\hat{m}_1}$$

# Résumés statistiques

## ► Boîte à moustache (Boxplot de Tukey)

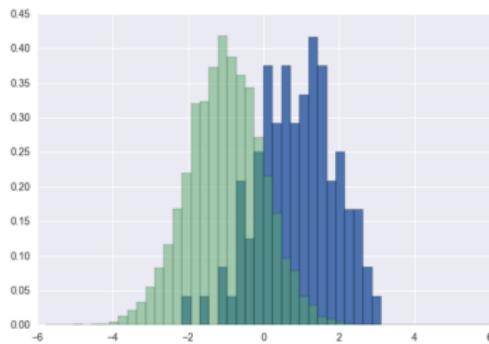


## Subsection 2

### Histogramme

# Histogramme

- ▶ représentation de la distribution des données (approximation d'une possible densité sous-jacente)
- ▶ L'ensemble des valeurs de la variable est découpé en classes représentées en abscisse.
- ▶ Le nombre de données (et/ou le pourcentage) appartenant à chaque classe est indiqué en ordonnée



# Table des matières

Probabilités

Variables aléatoires

Simulation de variables aléatoires

Statistique descriptive

Statistique Inférentielle

Motivation et notion  
d'estimateur

Propriétés des estimateurs

Quelques estimateurs  
paramétriques classiques

Modèle du plongeoir

Un estimateur  
non-paramétriques

Approche Bayésienne

## Subsection 1

### Motivation et notion d'estimateur

# Des données et des expériences

- ▶ **Données** disponibles : mais les phénomènes sous-jacent ne sont pas connus avec certitudes (compléxité, erreurs de mesures)
- ▶ On modélise ces incertitudes par des variables aléatoires et on utilise les **outils statistiques** associés pour étudier les caractéristiques du phénomène aléatoire : moyenne, variance etc...
- ▶ On **réitère des expériences** pour obtenir plus d'information sur l'aléa : plus de réalisations de la v.a.

# Statistique Inférentielle

- ▶ Etendre les propriétés d'un échantillon à toute une population
- ▶ Proposition de **modèles probabilistes**
- ▶ **Estimation de paramètres** du modèle (moyenne, variance,...)

# Modélisation probabiliste

- ▶ Moyen à large nombre de données : inférence paramétrique ou non-paramétrique
- ▶ Peu de données : élicitation par expert, méthode inverse, méthode du maximum d'entropie, Wilks, bootstrap, inégalités probabilistes robustes, ...

# Estimateur

- ▶ Soit  $\theta^*$  un paramètre à estimer.
- ▶ Soit  $(X_i)_{i=1,\dots,n}$  une suite de variables aléatoires i.i.d.
- ▶ On appelle estimateur statistique du paramètre  $\theta$  une variable aléatoire de la forme

$$\theta_n = r(X_1, \dots, X_n)$$

- ▶ Une réalisation de l'estimateur est donnée par

$$\hat{\theta}_n = r(x_1, \dots, x_n)$$

où les  $x_i$  sont des réalisations des  $X_i$ .

*Exemple de la moyenne*

# Estimateur

*Exemple de la moyenne, variance, fonction de répartition,...*

## Subsection 2

### Propriétés des estimateurs

# Propriétés des estimateurs

- ▶ Biais

$$b^2(\theta_n) = \mathbb{E}(\theta_n) - \theta^*$$

- ▶ Variance

$$Var(\theta_n) = \mathbb{E}\left[\left(\theta_n - \mathbb{E}(\theta_n)\right)^2\right]$$

- ▶ Précision/Erreur quadratique moyenne

$$\mathbb{E}\left[\left(\theta_n - \theta^*\right)^2\right] = b^2(\theta_n) + Var(\theta_n)$$

- ▶ Convergent/consistant

$$\theta_n \xrightarrow{p.s.} \theta^*$$

- ▶ Un estimateur sans biais et de variance asymptotiquement nulle est convergent

# Propriétés des estimateurs

- ▶ Estimateur de variance minimale
- ▶ Estimateur convergent
- ▶ Propriété : un estimateur sans biais et de variance asymptotiquement nulle est convergent
- ▶ Vitesse de convergence ?
- ▶ Comment trouver un bon estimateur ?

# Moyenne et variance d'un estimateur d'une variance

- Soit  $X_1, \dots, X_n$  une suite i.i.d. de v.a. de loi  $X$  de moyenne  $m$  et de variance  $\sigma^2$ . Supposons  $m$  connu alors un estimateur de la variance est

$$R_1 = \frac{1}{n} \sum_{i=1}^n (X_i - m)^2$$

- Moyenne de l'estimateur :

$$\mathbb{E}(R_1) = \sigma^2$$

- Variance de l'estimateur :

$$Var(R_1) = \frac{\mu^4 - \sigma^4}{n}$$

# Moyenne et variance d'un estimateur d'une variance

- Soit  $X_1, \dots, X_n$  une suite i.i.d. de v.a. de loi  $X$  de moyenne  $m$  et de variance  $\sigma^2$ . Supposons  $m$  inconnu alors un estimateur de la variance est

$$R_2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2$$

- Moyenne de l'estimateur :

$$\mathbb{E}(R_2) = \frac{n-1}{n} \sigma^2$$

- Variance de l'estimateur :

$$Var(R_2) = \frac{n-1}{n^3} \left[ (n-1)\mu^4 - (n-3)\sigma^4 \right]$$

## Moyenne et variance d'un estimateur d'une variance

- Soit  $X_1, \dots, X_n$  une suite i.i.d. de v.a. de loi  $X$  de moyenne  $m$  et de variance  $\sigma^2$ . Supposons  $m$  inconnu alors un estimateur de la variance est

$$R_3 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$$

- Moyenne de l'estimateur :

$$\mathbb{E}(R_3) = \sigma^2$$

- Variance de l'estimateur :

$$Var(R_3) = \frac{1}{n} \left[ \mu^4 - \frac{n-3}{n-1} \sigma^4 \right]$$

## Moyenne et variance d'un estimateur d'une variance

- ▶ Pour un écart type plus petit que la moyenne et un  $n$  pas trop grand on constate que l'estimateur biaisé à une variance plus petite que celui dé-biaisé.

$$Var(R_2) \leq Var(R_1) \leq Var(R_3)$$

# Propriétés des estimateurs

- ▶ Minimisation EQM difficile
- ▶ Chercher estimateur sans biais de variance minimale
- ▶ Pourquoi sans biais ?

## Propriétés des estimateurs

- ▶ Minimisation EQM difficile
- ▶ Chercher estimateur sans biais de variance minimale
- ▶ Pourquoi sans biais ? → difficile à estimer
- ▶ Il est possible de trouver des estimateurs biaisés plus précis que le meilleur estimateur sans biais.
- ▶ Parfois introduire un léger biais dans un estimateur initialement sans biais peut conduire à une réduction significative de sa variance, au point de provoquer une diminution de son EQM

# Propriétés des estimateurs

- ▶ Pour une suite de variables aléatoires  $X_1, \dots, X_n$  i.i.d. dont la loi dépend d'un paramètre inconnu  $\theta$  alors on écrit sa distribution de densité jointe comme

$$p_{X_1, \dots, X_n}(x_1, \dots, x_n | \theta) = \prod_{i=1}^n p_{X_i}(x_i | \theta)$$

- ▶ Trouver un estimateur sans biais et de variance minimale ne sera possible que si
  - ▶ **l'information** sur  $\theta$  contenue dans l'échantillon est suffisamment riche
  - ▶ On dispose d'une statistique **exhaustive** pour  $\theta$  (à définir dans la suite)

# Propriétés des estimateurs

- ▶ Information : une réalisation de l'échantillon aléatoire apporte une certaine information sur  $\theta$  : la répartition de ses valeurs donne une information sur la loi de  $X$ , qui dépend de  $\theta$ . Elle doit être suffisante - levier : taille de l'échantillon
- ▶ L'estimation réduit  $n$  valeurs en une seule (cas uni-dimensionnel). Perte d'une partie de l'information. A priori la connaissance de la seule estimation ne permet pas de remonter à l'échantillon tout entier. Sauf si l'estimateur est exhaustivité (définition bientôt) :
- ▶ Affaiblissement de l'information sur la loi
  - ▶ Via le sondage
  - ▶ Via la construction d'un estimateur
  - ▶ La perte doit être minimale pour construire un estimateur précis

# Vraisemblance d'un échantillon

- Soit  $X_1, \dots, X_n$  une suite de variables aléatoires de densité jointe  $f_{X_1 \dots X_n}$ .
- La vraisemblance de l'estimateur  $\hat{\theta}_n$  sur la base d'un ensemble de réalisations  $(x_i)_{i=1,\dots,n}$  est

$$f_{X_1 \dots X_n}(x_1, \dots, x_n | \theta_n = \hat{\theta}_n) = \prod_{i=1}^n p_{X_i}(x_i | \theta_n = \hat{\theta}_n)$$

- Log-vraisemblance

## Score et Information de Fisher

- ▶ Score : dérivée de la log-vraisemblance. Moyenne du score ?
- ▶ Information de Fisher : variance du score

## Score et Information de Fisher

- ▶ L'information de Fisher mesure l'information apportée par un échantillon sur le paramètre  $\theta$ . Une information de Fisher proche de zero indique un échantillon peu informatif.
- ▶ La valeur du score mesure la sensibilité de la vraisemblance à la valeur de  $\theta$ . Si le score est faible, la vraisemblance est peu sensible à de petites variations du paramètre : les observations n'arrivent pas à s'accorder entre elles sur la direction du changement à apporter à la valeur de  $\theta$  pour augmenter la vraisemblance de l'échantillon. On doit donc s'attendre à ce que l'échantillon contienne peu d'information sur la vraie valeur de ce paramètre.
- ▶ En moyenne,  $\theta$  fixé, le score est nul. Si sa variance (information de Fisher) est très petite alors, presque tous les jeux de données auront alors un score proche de 0 (l'espérance du score), et donc presque tous les échantillons ne contiendront qu'une faible quantité d'information sur la valeur réelle de  $\theta$ .

# Propriétés Information de Fisher

- ▶ Autre formulation
- ▶ Additivité
- ▶ cas multi-dimensionnel

# Statistique exhaustive

## Statisticien S1 :

- ✓ dispose d'un jeu de données obtenu par sondage.
- ✓ Peut construire la valeur de  $t$  à partir de ce jeu de données
- ✓ Peut construire une estimation de  $\theta$  à partir de ce jeu de données.

## Statisticien S2 :

- ✓ ne dispose pas du jeu de données
- ✓ s'est fait donner  $t$  par S1
- ✓ Connait  $g_\theta$

Pour disposer d'autant d'information que S1, S2 doit être capable de tirer une réalisation de l'échantillon aléatoire. Comme il dispose de  $t$ , il faut qu'il puisse tirer dans la loi conditionnelle de cet échantillon sachant  $t$ , mais celle-ci dépend généralement de  $\theta$ .

$$S1 \text{ et } S2 \Leftrightarrow k_\theta(x_1, \dots, x_n / T = t) \text{ ne dépend pas de } \theta$$

Alors seulement, S2 sera capable de tirer une réalisation de l'échantillon aléatoire et de construire une estimation de  $\theta$  à partir de ce jeu de données.

# Statistique exhaustive

- Un estimateur est exhaustif ssi  $\theta_n = r(x_1, \dots, x_n)$  contient toute l'information des données sur le paramètre à estimer.

$$p(x_1, \dots, x_n | \theta_n, \theta^*) = p(x_1, \dots, x_n | \theta_n)$$

Autre caractérisation (critère de factorisation)

$$p(x_1, \dots, X_n | \theta_n, \theta^*) = h(X_1, \dots, x_n)g(\theta_n, \theta^*)$$

- Exemple loi normale à moyenne connue et variance inconnue, loi Poisson

# Information de Fisher et statistique exhaustive

- ▶ Dégradation de l'information par une statistique de l'échantillon

$$I_n(\theta) \geq I_r(\theta)$$

avec égalité si la statistique est exhaustive

- ▶ Le montrer.

# Famille exponentielle et statistique exhaustive

## Théorème de Darmois

- ▶ On suppose que le domaine de définition de  $X$  ne dépend pas de  $\theta$
- ▶ Une condition nécessaire et suffisante pour que l'échantillon aléatoire admette une statistique exhaustive est que la distribution de  $X$  appartienne à la famille exponentielle, i.e.

$$p(x|\theta) = \exp \left[ a(x)\alpha(x) + b(x) + \beta(\theta) \right]$$

## Statistique exhaustive

- ▶ Les estimateurs les plus précis de  $\theta$  (en particulier les estimateurs sans biais de variance minimale) sont des statistiques exhaustives ou des fonctions de celles-ci (sous hypothèse d'existence de cette statistique exhaustive).
- ▶ il faut en plus que l'information contenue dans l'échantillon sur le paramètre soit suffisante

## Estimateur sans biais de variance minimale

- ▶ Il est fréquent qu'un paramètre admette plusieurs, voire une infinité d'estimateurs sans biais. *Exemple de la loi Normale.*
- ▶ De tous les estimateurs sans biais, le meilleur (au sens de EQM) est celui qui a la plus faible variance. On l'appelle "Estimateur sans biais de Variance Minimale".
- ▶ L'identification et la qualité d'un estimateur de  $\theta$  sans biais et de variance minimale est lié à l'information contenue dans l'échantillon et à l'existence d'une statistique exhaustive.

## Estimateur sans biais de variance minimale

- ▶ Unicité : s'il existe un estimateur sans biais de variance minimale alors il est unique p.s.
- ▶ Théorème de Rao-Blackwell Si un estimateur sans biais n'est pas de variance minimale, il est possible de l'améliorer si l'on dispose d'une statistique exhaustive. Le Théorème ne garantit cependant pas que le nouvel estimateur 'amélioré' soit de variance minimale.
- ▶ Inégalité de Cramér-Rao Permet d'établir, sous condition de régularité, une borne inférieure de la variance d'un estimateur sans biais.
- ▶ Conditions sous lesquelles la borne est atteinte. L'estimateur de variance minimale est alors celui ayant la variance de la borne de Cramer- Rao, il s'appelle estimateur efficace.

# Estimateur sans biais de variance minimale

## Théorème de Rao-Blackwell

- Soit  $\theta_n$  un estimateur sans bias et  $R$  une statistique exhaustive alors

$$Var(\mathbb{E}(\theta_n/R)) \leq Var(\theta_n)$$

- Inversement si on dispose d'un estimateur sans biais fonction d'une statistique exhaustive, on n'est pas sûrs qu'il soit de variance minimale.

# Borne de Cramér-Rao

Aussi connue sous le nom d'inégalité de Fréchet-Darmois-Cramér-Rao

- ▶ Hypothèse (H) : le domaine de définition de  $X$  ne dépend pas de  $\theta$  et l'information de Fisher existe alors pour tout estimateur sans biais

$$Var(\theta_n) \geq \mathbb{E} \left[ \left( \frac{\partial}{\partial \theta} \ln p(X_1, \dots, X_n | \theta) \right)^2 \right]^{-1}$$

# Borne de Cramér-Rao

## Interprétation.

- ▶ La valeur de la borne est fonction de l'information que peut contenir l'échantillon sur le paramètre : plus grande est l'information sur la valeur du paramètre, plus précises seront les prédictions d'un estimateur sans biais dont la variance est égale à la borne de Cramér-Rao.
- ▶ Inversement, si la variance du score (information de Fisher) est très petite, et donc si presque tous les échantillons ne contiennent que peu d'information sur la valeur du paramètre , on ne peut pas espérer d'un estimateur sans biais qu'il soit précis, c'est à dire qu'il ait une faible variance.

## Extension

- ▶ Cas d'une fonction  $k$  du paramètre  $\theta$ , la borne devient

$$\frac{(k'(\theta))^2}{I_n(\theta)}$$

## Estimateur efficace

- ▶ Sous hypothèse ( $H$ ), un estimateur efficace est un estimateur sans biais dont la variance est égale à la borne inférieure de Cramer-Rao et s'il existe il est unique p.s.

## Estimateur efficace

- ▶ sous hypothèse (H), la borne de Cramér-Rao n'est atteinte que si la loi de  $X$  appartient à la famille exponentielle.
- ▶ Dans ce cas il n'existe qu'une seule fonction de  $\theta$  qui puisse être estimée efficacement :

$$k(\theta) = -\frac{\beta'(\theta)}{\alpha'(\theta)}$$

- ▶ L'estimateur est donné par

$$\theta_n = \frac{1}{n} \sum_{i=1}^n a(X_i)$$

de variance minimale :

$$Var(\theta_n) = \frac{k'(\theta)}{n\alpha'(\theta)}$$

- ▶ Exemple loi de Poisson

## Estimateur efficace

- ▶ Hypothèse : Le domaine de définition de  $X$  ne dépend pas de  $\theta$  et l'information de Fisher existe alors pour tout estimateur sans biais
- ▶ Un estimateur efficace est un estimateur sans biais dont la variance est égale à la borne inférieure de Cramer-Rao et s'il existe il est unique p.s.

## Subsection 3

Quelques estimateurs paramétriques classiques

# Méthode des moments

- ▶ Soit  $\theta^*$  un paramètre à estimer et  $(x_1, \dots, x_n)$  une réalisation de la suite i.i.d. de v.a.  $(X_1, \dots, X_n)$  correspondantes dont la distribution dépend de  $\theta^*$
- ▶ On note  $\mathbf{m}(\theta) = (m_1(\theta), \dots, m_k(\theta))$  où  $m_j(\theta) = \mathbb{E}(X^j)$
- ▶ Soit  $\hat{\mathbf{m}} = (m_1, \dots, m_k)$  où  $\hat{m}_j = \frac{1}{n} \sum_{i=1}^n x_i^j$  est l'estimateur empirique du  $j$ -ième moment.
- ▶ l'estimateur par la méthode des moments est la solution  $\hat{\theta}$  du système d'équations

$$\hat{\mathbf{m}} = \mathbf{m}(\theta)$$

- ▶ *Exemple de la loi de Poisson*

# Estimateur du maximum de Vraisemblance

- ▶ Estimateur du maximum de vraisemblance :

$$\hat{\theta}_n = \arg \max_{\theta} f_{X_1, \dots, X_n}(X_1, \dots, X_n | \theta)$$

- ▶ Pour une réalisation  $(x_1, \dots, x_n)$  l'estimation du maximum de vraisemblance est

$$\hat{\theta}_n = \arg \max_{\theta} f_{X_1, \dots, X_n}(x_1, \dots, x_n | \theta)$$

- ▶  $\hat{\theta}_n$  est une réalisation de  $\theta_n$

# M-estimateur

## M-estimateur

- ▶ Par définition un M-estimateur est un estimateur qui minimise

$$\mathbb{E}(\rho(\theta_n, \theta^*))$$

où  $\rho$  est une fonction convexe.

- ▶ Exemple :

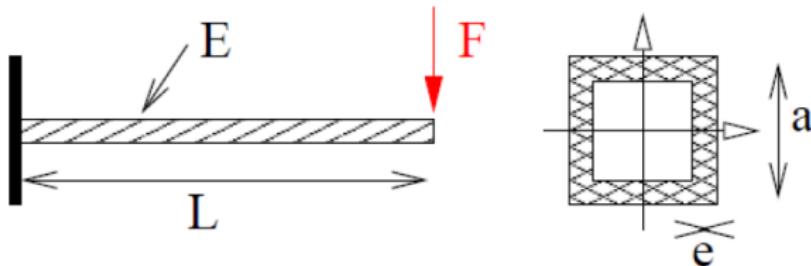
$$\mathbb{E}(|\theta_n - \theta^*|)$$

## Subsection 4

### Modèle du plongeoir

# Modèle Plongeoir

On s'intéresse à la déviation verticale d'une poutre encastrée en une extrémité et soumise en l'autre extrémité à une charge ponctuelle verticale.



La déviation verticale de l'extrémité (ou flèche) de la poutre s'exprime comme suit :

$$y(E, F, L, I) = \frac{FL^3}{3EI}$$

où :

- E est le module d'Young du matériau
- F est la charge ponctuelle appliquée
- L est la longueur de la poutre
- I est le moment quadratique, que l'on peut écrire :

$$I = \frac{a^4 - (a - e)^4}{12}$$

# Modèle Plongeoir

12

Les valeurs utilisées pour les études déterministes sont :

$$F = 300 \text{ N}$$

$$E = 3.0\text{e}9 \text{ Pa}$$

$$L = 2.5 \text{ m}$$

$$I = 4.0\text{e-}6 \text{ m}^4$$

ce qui correspond au point nominal  $(30000, 3.0\text{e}7, 250, 400)$  lorsque la longueur est exprimée en cm et non en m. Ces valeurs correspondent à une planche en plastique, longue de 2.5m, à l'extrémité de laquelle une charge de 30kg est posée. On pourrait imaginer par exemple un plongeoir de piscine en matière plastique, sur lequel se tient un enfant de 30kg.

Les incertitudes qui entachent les variables ( $E$ ,  $F$ ,  $L$ ,  $I$ ) sont dues, par exemple, à :

$E$  : fabrication de l'acier (incertitude stochastique),

$F$  : incertitude de mesure de la charge (incertitude épistémique),

$L$  et  $I$  : défaut géométrique de fabrication (incertitude stochastique).

## Modèle Plongeoir

On considère que l'on dispose de données sur  $F$ , issues de mesures, et l'on souhaite ajuster un modèle paramétrique sur ces données, afin notamment de l'utiliser pour l'étape de propagation des incertitudes. Les données sont contenues dans le fichier sampleF.csv.

- ▶ Calculer la moyenne et l'écart-type de l'échantillon, le minimum, le maximum. De combien de données dispose-t-on ?
- ▶ On souhaite visualiser l'histogramme de l'échantillon dont on dispose. A partir de l'allure de l'histogramme, commencez à proposer une(des familles de lois qui pourraient correspondre.
- ▶ Pour différentes familles de lois jugées "possibles", estimez les paramètres correspondant le mieux aux données. Par la méthode des moments ou la méthode du maximum de vraisemblance.

## Subsection 5

Un estimateur non-paramétriques

# Estimation non-paramétrique d'une densité

- ▶ Estimateur

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^n K\left(\frac{X_i - x}{h}\right)$$

- ▶ Estimation

$$\frac{1}{nh} \sum_{i=1}^n K\left(\frac{x_i - x}{h}\right)$$

- ▶  $K$  est une fonction dite noyau symétrique et d'intégrale égale à 1.
- ▶  $h$  est un paramètre de lissage aussi dit "fenêtre". Important pour la qualité de l'estimation
- ▶ exemple de noyau : Uniforme, triangulaire, Epanechnikov, Gaussien
- ▶ Mesure de qualité : Erreur quadratique intégrée moyenne (EQIM)

$$\mathbb{E} \left[ \int \left( \hat{f}(x) - f(x) \right)^2 dx \right]$$

## Estimation non-paramétrique d'une densité multivariée

- ▶ Supposons les données multivariées i.e. que chaque  $x_i \in \mathbb{R}^d$  est la réalisation d'un vecteur aléatoire  $X_i$  également à valeur dans  $\mathbb{R}^d$ . Alors l'estimateur de la distribution multivariée est donnée pour tout  $x \in \mathbb{R}^d$  par

$$\hat{f}(x) = \frac{1}{n|H|^{1/2}} \sum_{i=1}^n K(H^{-1}(X_i - x))$$

- ▶ L'estimation est

$$\frac{1}{n|H|^{1/2}} \sum_{i=1}^n K(H^{-1}(x_i - x))$$

- ▶ Le noyau est ici  $K(x) = 1/(2\pi)^{d/2} \exp(-x^t x/2)$  et l'EQIM est défini comme précédemment avec l'intégrale sur  $\mathbb{R}^d$

# Approximation de la loi bivariée Gaussienne

- Soit  $(X_1, X_2)$  un vecteur aléatoire suivant la loi normale bivariée  $\mathcal{N}(\mu, \Sigma)$  où

$$\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} = \begin{pmatrix} 2 \\ -2 \end{pmatrix}$$

et

$$\Sigma = \begin{pmatrix} \Sigma_{1,1} & \Sigma_{1,2} \\ \Sigma_{2,1} & \Sigma_{2,2} \end{pmatrix} = \begin{pmatrix} 2 & 0.9 \\ 0.9 & 1 \end{pmatrix}$$

- Simuler un échantillon de taille  $n$  de ce vecteur et estimer à partir de ce dernier la densité jointe. Dessiner la densité jointe approximée et comparer à la vraie distribution. Estimer EQIM pour plusieurs taille  $n$  de l'échantillon. Dessiner graphe de EQIM en fonction de  $n$ .

# Table des matières

Probabilités

Variables aléatoires

Simulation de variables aléatoires

Statistique descriptive

Statistique Inférentielle

## Approche Bayésienne

Formule de Bays

Frequentiste et Bayesien

Choix du prior

Estimateur de Bays

Un exemple simple

Quelques conclusions

Rappels méthodes de simulations

Historique pour Monte-Carlo

Historique pour Monte-Carlo

Markov Chain

Méthode de Metropolis Hasting

## Subsection 1

### Formule de Bays

# Bays Formula

Bayes formula for random events :

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} = \frac{\mathbb{P}(B|A)\mathbb{P}(A)}{\mathbb{P}(B|A)\mathbb{P}(A) + \mathbb{P}(B|A^c)\mathbb{P}(A^c)}$$

[Thomas Bayes 1764]

Bayes formula for densities :

$$p(\theta|d) = \frac{L(d|\theta)\pi(\theta)}{\int L(d|\theta)\pi(\theta)d\theta} \propto L(d|\theta)\pi(\theta)$$

- ▶ Bayes formula offer us the density of  $\theta|D = d$  denoted  $p(\cdot|d)$  and called the **posterior distribution**.

## Subsection 2

Frequentiste et Bayesien

# Two point of view

## Frequentist point of view

- ▶ Quantitative measure of **uncertainty of observables** (repeatable)
- ▶ **Analysis** made only **on the data**  
⇒ actor : statistician

## Bayesian point of view

- ▶ Quantitative measure of **uncertainty of any unknown** (extraterrestrial life...)
- ▶ Bayesian inference consists in finding the distribution of the unknown **from observations and expertise**  
⇒ actors : statistician, expert
- ▶ Allows incorporation of imperfect information in the decision process

## Subsection 3

### Choix du prior

# Prior choice

## Prior joint distribution :

- ▶ The prior is the **joint distribution of the parameters** and not only the product of their marginals (independent case). **Not always simple to model.**

## conjugate prior :

- ▶ **posterior** belongs to the **same family** of parametrized distribution than the **prior**.

Case where analytical analysis and simulation are possible.

Bayesian approach was mostly limited to this case before introduction of numerical calculus (MCMC...)

# Prior choice

Informative prior :

- ▶ the prior is informative if it **impacts the posterior**. For instance use of previous study results as prior

Uninformative prior :

- ▶ uninformative : the prior **does not impact the posterior** :  
 $\mathcal{N}(\mu, +\infty)$ ,  $\mathcal{B}(1, 1)$ , Jeffrey's

Improper prior :

- ▶ **Robust** answer against possible **misspecifications of the prior**
- ▶ involve more general **penalization**
- ▶ preferred to vague priors
- ▶ require few hypothesis :

$$\int \pi(\theta) d\theta = +\infty \text{ and } \int L(d|\theta) \pi(\theta) d\theta < \infty$$

example :  $\pi(\theta) = 1_{\mathbb{R}^+}(\theta)$

# Prior choice

Transforming prior information into prior distribution : an art.

-  O'Hagan a., Buck C.E., Daneshkhah A., Eiser J.R., Garthwaite P.H., Jenkinson D.J., Oakley J.E., Rakow T, **Uncertain judgements : eliciting experts probabilities**. Statistics in Practice, John Wiley & Sons, Ltd, Chichester.
-  Cooke R.M. (1991). **Experts in uncertainty : opinion and subjective probability in science**. Environmental ethics and science policy, Oxford University Press, New York.

## Subsection 4

### Estimateur de Bays

# Bays Estimators

The posterior is the updated distribution of the parameters.

Useful to

- ▶ Graphically **represent** the current **knowledge** (marginals)
- ▶ Derive an **estimator** : mean, median, mode
- ▶ Derive a **credibility interval** :  $\mathbb{P}(\theta \in [a, b]) = 1 - \alpha$
- ▶ Derive **probability estimation** :  $\mathbb{P}(\theta > a)$

## Subsection 5

Un exemple simple

# A simple example



$$\mathbf{y} = \begin{pmatrix} 2 \\ 3 \end{pmatrix} \quad \theta^* = 3.5$$

- ▶ Error distribution

$$\Sigma = 4Id, \quad \epsilon \sim \mathcal{N}(0, \Sigma)$$

- ▶ Observations

$$\mathbf{d} = \mathbf{G}(\theta^*) + \epsilon = \mathbf{y}^{\theta^*} + \epsilon$$

- ▶ Prior

$$\pi \sim \mathcal{N}(3.55, 0.1) \text{ or } \pi \sim \mathcal{U}([a, b])$$

- ▶ Likelihood

$$L(\mathbf{d}|\theta) \sim \mathcal{N}(\mathbf{y}^\theta, \Sigma)$$

# A simple example

- ▶ Prior

$$\pi \sim \mathcal{N}(3.55, 0.1) \text{ or } \pi \sim \mathcal{U}([a, b])$$

- ▶ Likelihood

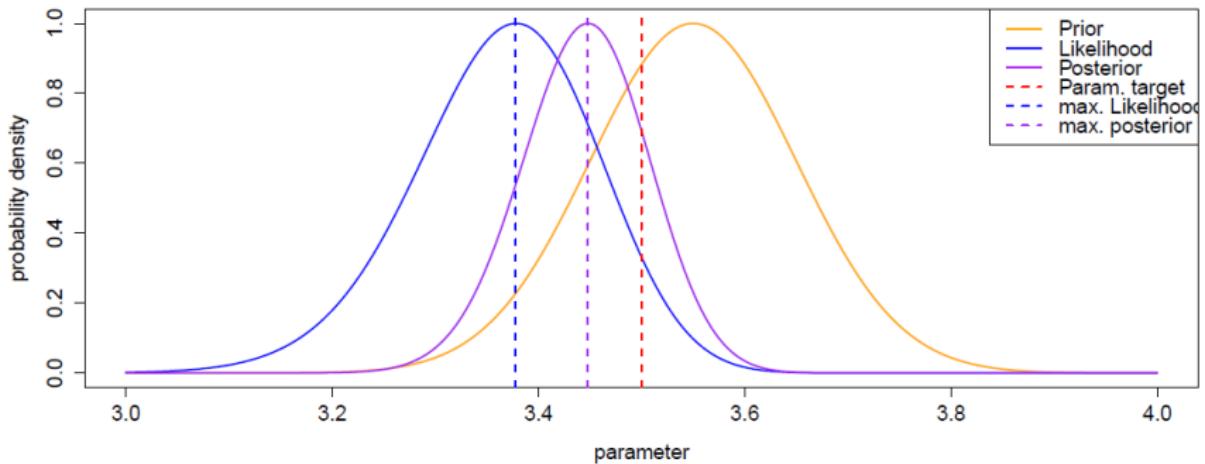
$$L(\mathbf{d}|\theta) \sim \mathcal{N}(\mathbf{y}^\theta, \Sigma)$$

- ▶ posterior

$$p(\theta|\mathbf{d}) \propto L(\mathbf{d}|\theta)\pi(\theta)$$

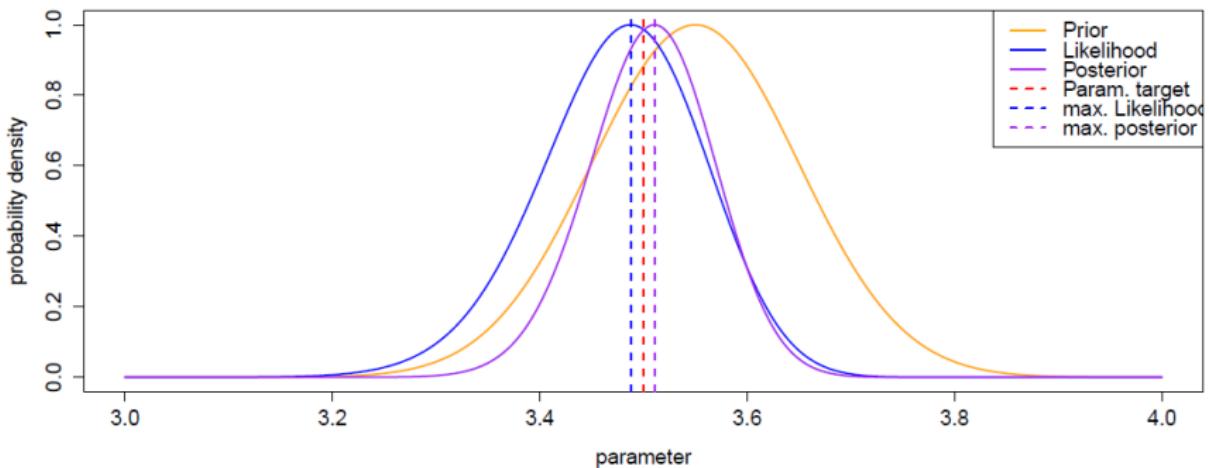
# A simple example

$$G(x^*) = (11.3137084989848, 46.7653718043597) \quad d = (16.9308918557838, 39.8700416899586)$$



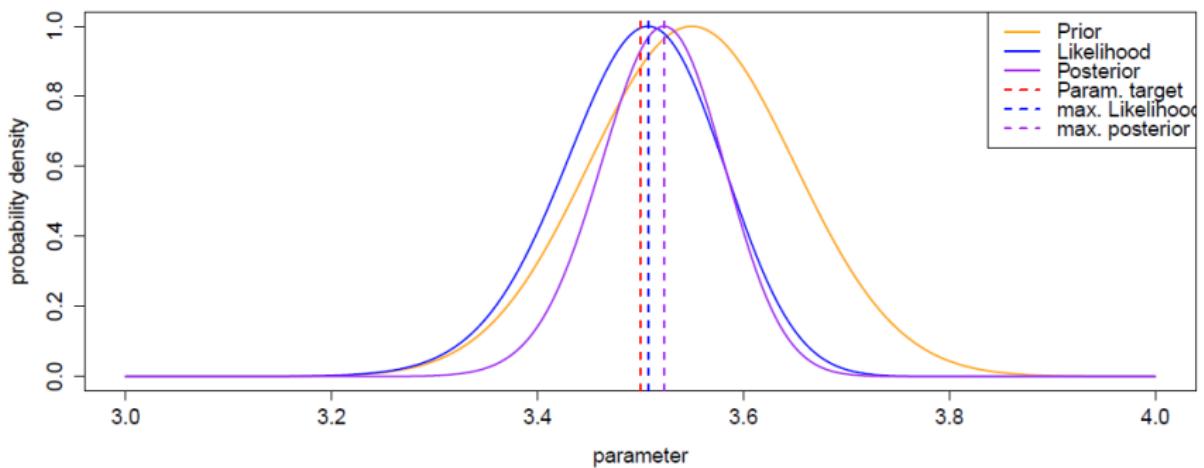
# A simple example

$$G(x^*) = (11.3137084989848, 46.7653718043597) \quad d = (5.93495468206409, 46.939222648288)$$



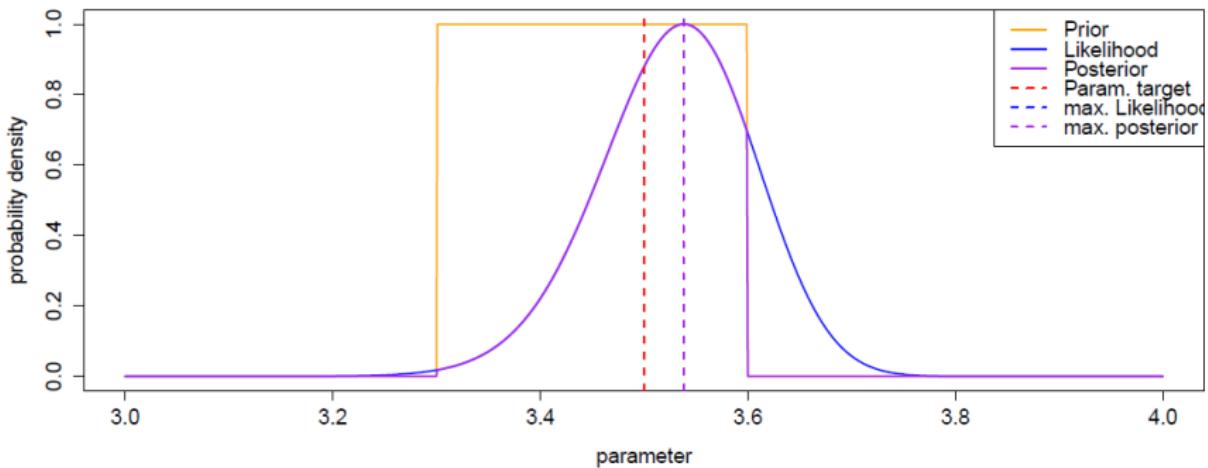
# A simple example

$$G(x^*) = (11.3137084989848, 46.7653718043597) \quad d = (11.736154068339, 47.0995401621255)$$



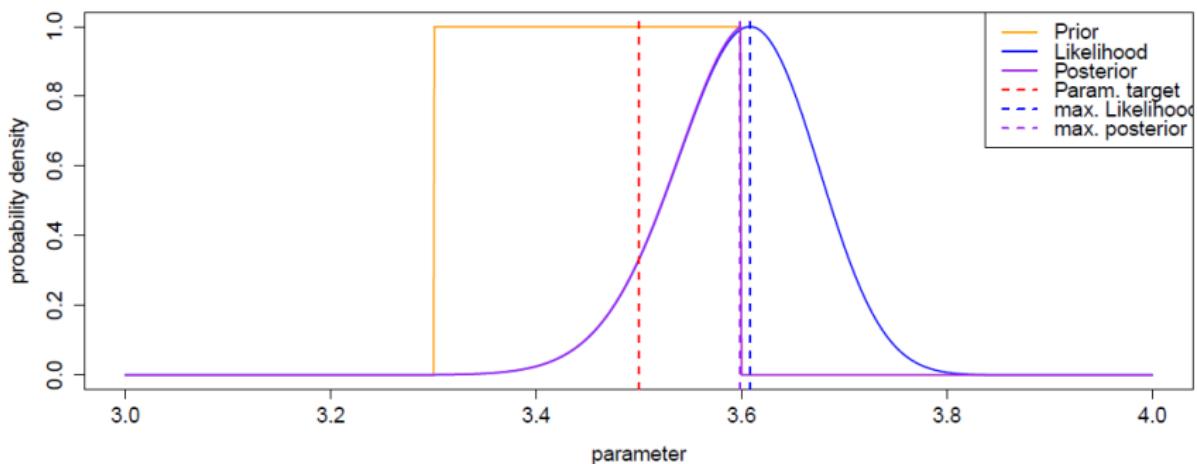
# A simple example

$$G(x^*) = (11.3137084989848, 46.7653718043597) \quad d = (8.75217722419654, 49.2147501153622)$$



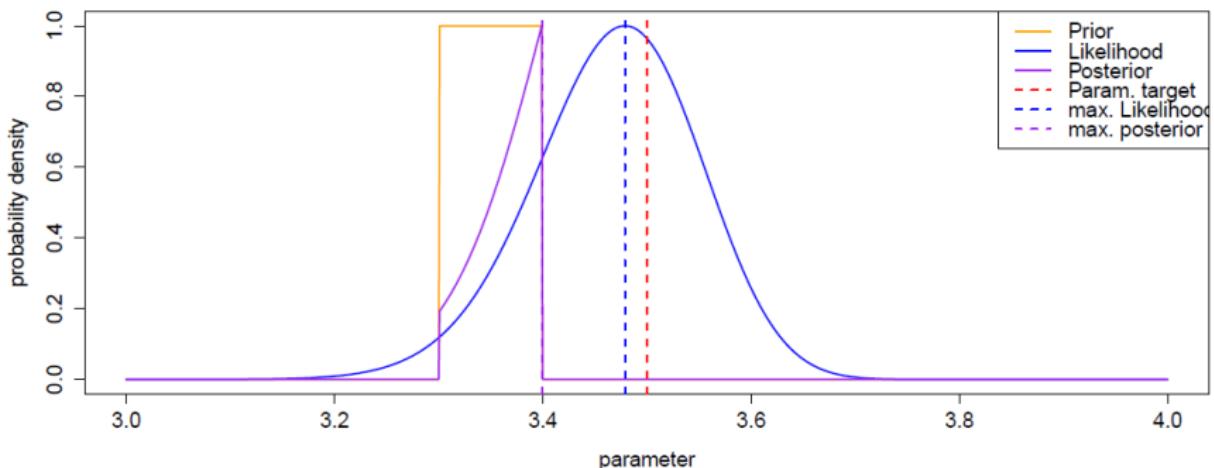
# A simple example

$$G(x^*) = (11.3137084989848, 46.7653718043597) \quad d = (9.34157225470381, 53.0832982128978)$$



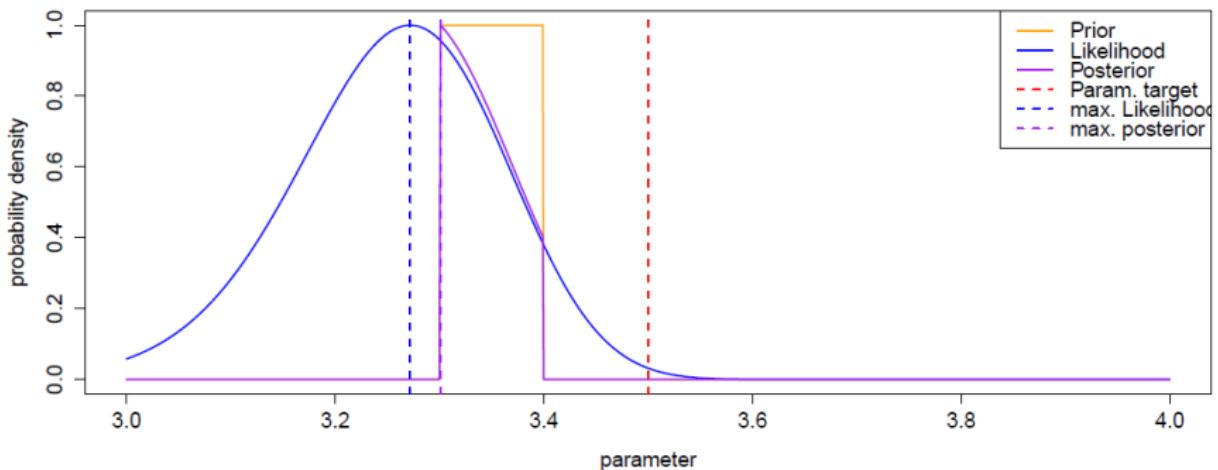
# A simple example

$$G(x^*) = (11.3137084989848, 46.7653718043597) \quad d = (13.6089681499255, 45.3184770072288)$$



# A simple example

$$G(x^*) = (11.3137084989848, 46.7653718043597) \quad d = (9.54729997729976, 36.4198580555248)$$



## Subsection 6

Quelques conclusions

# Some conclusions

Working hard and in interaction :

- ▶ Work has to be done on the **optimal choice of the data**, redundant information might be useful : use **design of experiment** techniques (tomorrow)  
⇒ expert + statistician
- ▶ Data has to be analyse and a **model has to be specified** (Likelihood)  
⇒ expert + statistician
- ▶ **Prior information** has to be determined and transposed into prior distribution  
⇒ expert + statistician

# Some conclusions

In practice how do we use the posterior :

- ▶ How do we **sample from the posterior** distribution in a general framework ?

Implementations :

- ▶ Bayesian methodology implemented in **CougarOpt** (normal likelihood and uniform priors) and **HubOpt** (more choices for likelihood and uniform priors)



Robert C., **The Bayesian Choice : from Decision-Theoretic Motivations to Computational Implementation** (2001),  
Springer-Verlag, New York

## Subsection 7

### Rappels méthodes de simulations

# Different simulation methods

- Inverse Cumulative Density Function
  - Specific methods for specific distributions (Box-Muller, Atkinson's Poisson,...)
  - In Bayesian context, conjugate methods leads to analytical posterior. In our case not possible to use.
  - Acceptation/Rejection: high computational time with high dimension
  - MCMC method: Metropolis-Hasting, Gibbs algorithms
- 

## Subsection 8

Historique pour Monte-Carlo

# Recipe for Monte-Carlo (MC)

- First commercial computer (~1942)
- 1946, physicists at Los Alamos Scientific Laboratory investigating radiation shielding and distance neutrons would likely travel through various materials (John Von Neumann, Stanislaw Ulam).
  - Idea : use simulated random experiments for calculus
- 1950 : gambling was illegal in most places, and the Monte Carlo casino was the most famous of the world. (Stanislaw Ulam's uncle would borrow money to gamble there)

Von Neumann : Let us call it MONTE CARLO !

## Subsection 9

Historique pour Monte-Carlo Markov Chain

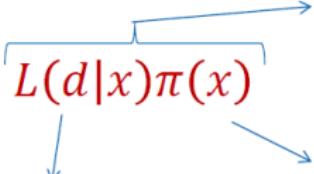
# Recipe for Monte-Carlo Markov Chain (MCMC)

- Monte Carlo Markov Chain (MCMC) invented soon after MC at Los Alamos.
- Metropolis et al. (1953) simulated a liquid in equilibrium with its gas phase
  - Idea : Metropolis Algorithm (1953)
  - Metropolis-Hasting generalization (1970), Metropolis-Hasting-Green generalization (1995)
- Gelfand and Smith (1990) made the wider Bayesian community aware of MCMC sampler.

# Objective

**Goal:** We would like to build a sampling (or sequence) which follows a particular distribution. In our case the posterior distribution.

$$X_n \sim p(x|d) \propto L(d|x)\pi(x)$$



Estimated via metamodel      Prior on parameters

## Subsection 10

### Méthode de Metropolis Hasting

# Motivating Metropolis-Hastings

**Goal:** For an arbitrary starting value  $X_1$ , an ergodic chain  $(X_n)$  is generated using a transition kernel with stationary distribution  $\Pi$

$$X_n \sim p(x|d) \propto \underbrace{L(d|x)\pi(x)}_{\rightarrow \Pi(x)}$$

- Insures the convergence in distribution of  $(X_n)$  to a random variable sampled from  $\Pi$ .
- For a large enough  $N_0$ ,  $X_{N_0}$  can be considered as distributed from  $\Pi$
- Produce a dependent sample  $X_{N_0}, X_{N_0+1} \dots$  which is generated from  $\Pi$ , sufficient for most approximation purposes.

**Problem:** How can one build a Markov chain with a given stationary distribution?

# Metropolis-Hastings algorithm

1. Initiate the Markov Chain with an « arbitrary » value  $X_1$ .

Simulate  $X_2, X_3 \dots$  such as :

2. Generate  $Y$  according to a transition kernel with distribution  $g(Y|X_i)$

3. Calculate the Hastings Metropolis ratio:

$$\alpha(Y, X_i) = \min\left(1, \frac{\Pi(Y) \times g(X_i|Y)}{\Pi(X_i) \times g(Y|X_i)}\right)$$

4. Simulate  $u \sim U([0,1])$

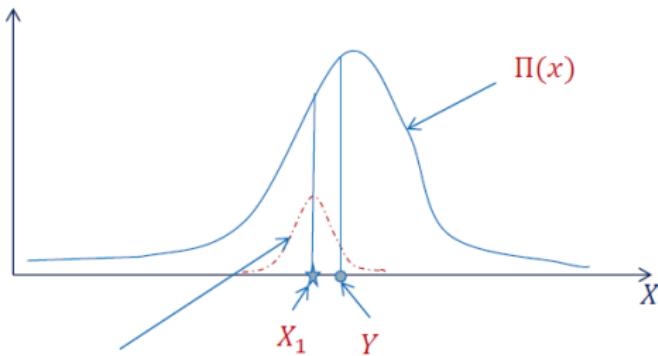
if  $u \leq \alpha(Y, X_i)$  then  $X_{i+1} = Y$  else  $X_{i+1} = X_i$

*Remark : point 4. is equivalent to :*

- 4b. Accept  $X_{i+1} = Y$  with probability  $\alpha(Y, X_i)$ , else  $X_{i+1} = X_i$ .

# Metropolis-Hastings algorithm

1. Initialization  $X_1$
2. Generate  $Y$  according to  $g(Y|X_i)$
3. Calculate ratio:  $\alpha(Y, X_i) = \min\left(1, \frac{\Pi(Y) \times g(X_i|Y)}{\Pi(X_i) \times g(Y|X_i)}\right)$
4. Accept  $X_{i+1} = Y$  with probability  $\alpha(Y, X_i)$ , else  $X_{i+1} = X_i$ .

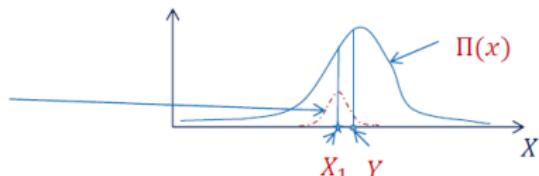


$$g(\cdot | X_1) = N(X_1, \sigma)$$

# Calibrate Metropolis-Hastings parameter « We don't want the porridge too cold or too hot »

- Transition distribution tuning

$$g(\cdot | X_1) = N(X_1, \sigma)$$



- If the variance is too big many proposed candidates will get far from current point, thus they will have small probability to be selected: the chain will be stuck for long periods.
- If the variance is too small the chain will need a long time to cover the whole distribution support.

“In small dimensions, aim at an average acceptance rate of 50%. In large dimensions, at an average acceptance rate of 25%.”

[Gelman, Gilks and Roberts, 1995]

“This rule is to be taken with a pinch of salt!”

[Christian Robert]

# Calibrate Metropolis-Hastings parameter « We don't want the porridge too cold or too hot »

- “In small dimensions, aim at an average acceptance rate of 50%. In large dimensions, at an average acceptance rate of 25%.”

[Gelman, Gilks and Roberts, 1995]

This rule gives the best performance for simulation of multivariate gaussian distributions.

- “This rule is to be taken with a pinch of salt!”

[Christian Robert]

« [Geyer and Thompson, 1995] came to a similar conclusion, that a 20% acceptance rate is about right, in a very different situation. They also warned that a 20% acceptance rate could be very wrong and produced an example where a 20% acceptance rate was impossible and attempting to reduce the acceptance rate below 70% would keep the sampler from ever visiting part of the state space. So the 20% magic number must be considered like other rules of thumb we toss around in statistics. »

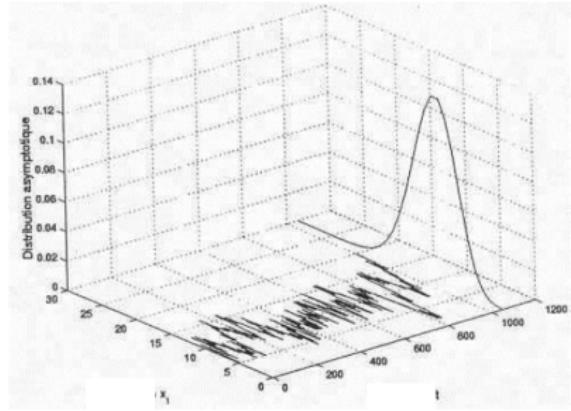
« There are many examples in the literature where they do fail. We keep repeating them because we want something simple to tell beginners and they are all right for many problems. »

# Metropolis-Hastings converges

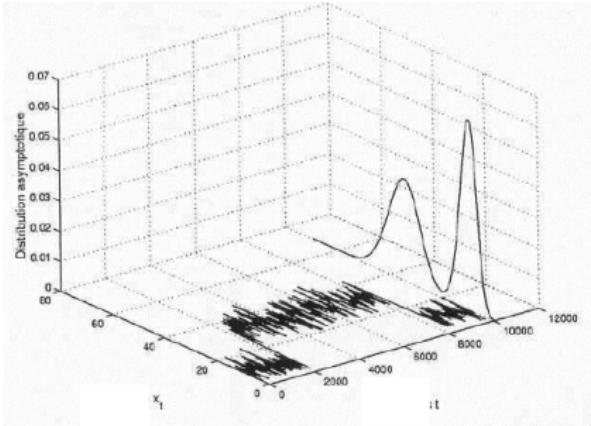
- **Existence theorem:** For any choice of transition density  $g$  the MH algorithm produces a markov chain which converges to the stationary goal distribution:  
 $\Pi(x) = p(x|d)$
- **The existence is not enough.** We need to ensure that the stationary distribution can be reached for any starting point !
- Fortunately the MH markov chain is ergodic (Harris positive irreducible and aperiodic):

*For any point  $x, y$  the chain is able to move from  $x$  to  $y$  in a finite number of steps and this number is not a multiple of some integer (aperiodicity: involved by the fact that  $X_{n+1} = X_n$  is possible).*

# Metropolis-Hastings diagnostic



MH covers well the whole distribution support



MH with localized jumps,  
seems stuck on local extremum.  
Convergence ensured but speed not  
geometric: can be slow.

# Metropolis-Hastings diagnostic

- **Starting point:**
  - Starting in the tail of distribution potentially involves a long period before leaving this area.
  - « *Any point you don't mind having in a sample is a good starting point* ».
  - Start at a mode.
- **Pseudo-convergence:** Convergence seems ok but in fact the algorithm has not run enough time.
- **Multistart heuristic:** start the chain from different points. Choice which you hope to cover the full distribution.
- **One long run vs multiple shorter runs ?**

« *One should start a run when the paper is submitted and keep running until the referees' report arrive* »

*Charles J. Geyer*