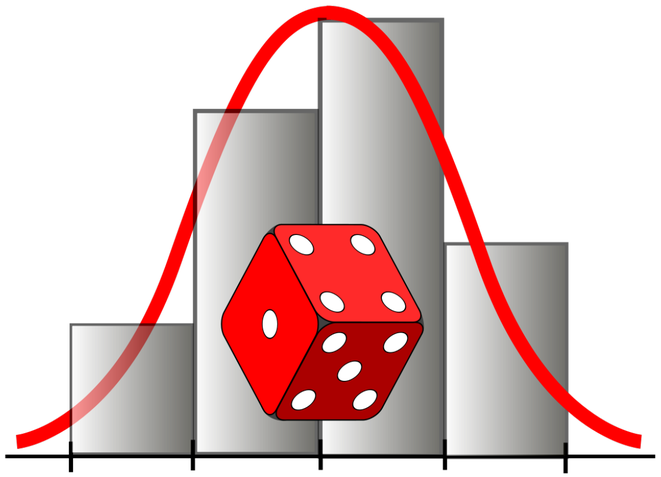
Chapitre 4 : introduction à la théorie des valeurs extrêmes
Partie 1 : rappel sur les probabilités

Risques hydrologiques et aménagement du territoire

Christophe Ancey

Chapitre 4 : théorie des valeurs extrêmes



Une introduction...

- Rappel :
  - probabilités : définition et propriétés
  - théorème de Bayes
- Ajustement de loi
  - principe de l'estimation de paramètres
  - stationnarité et indépendance
  - tests graphiques d'adéquation
  - méthode des moments
  - méthode du maximum de vraisemblance
  - inférence bayésienne
  - algorithme de Metropolis-Hastings

Qu'est ce qu'une probabilité ?

Deux approches :

- Approche classique ou fréquentiste. Théorie des jeux :

$$P = \frac{\text{nombre de cas favorables}}{\text{nombre de possibilités}}$$

Par extension

$$P = \lim_{n \to \infty} \frac{\text{nombre d'événements observés}}{\text{nombre total d'événements } n}.$$

- Approche bayésienne

$$P = \text{mesure du degré de croyance qu'un événement se produise.}$$

Définitions

Probabilité : à tout événement, on associe un nombre positif P compris entre 0 et 1. Soient E_1 et E_2 deux événements, on a :

- *complémentarité* de deux événements

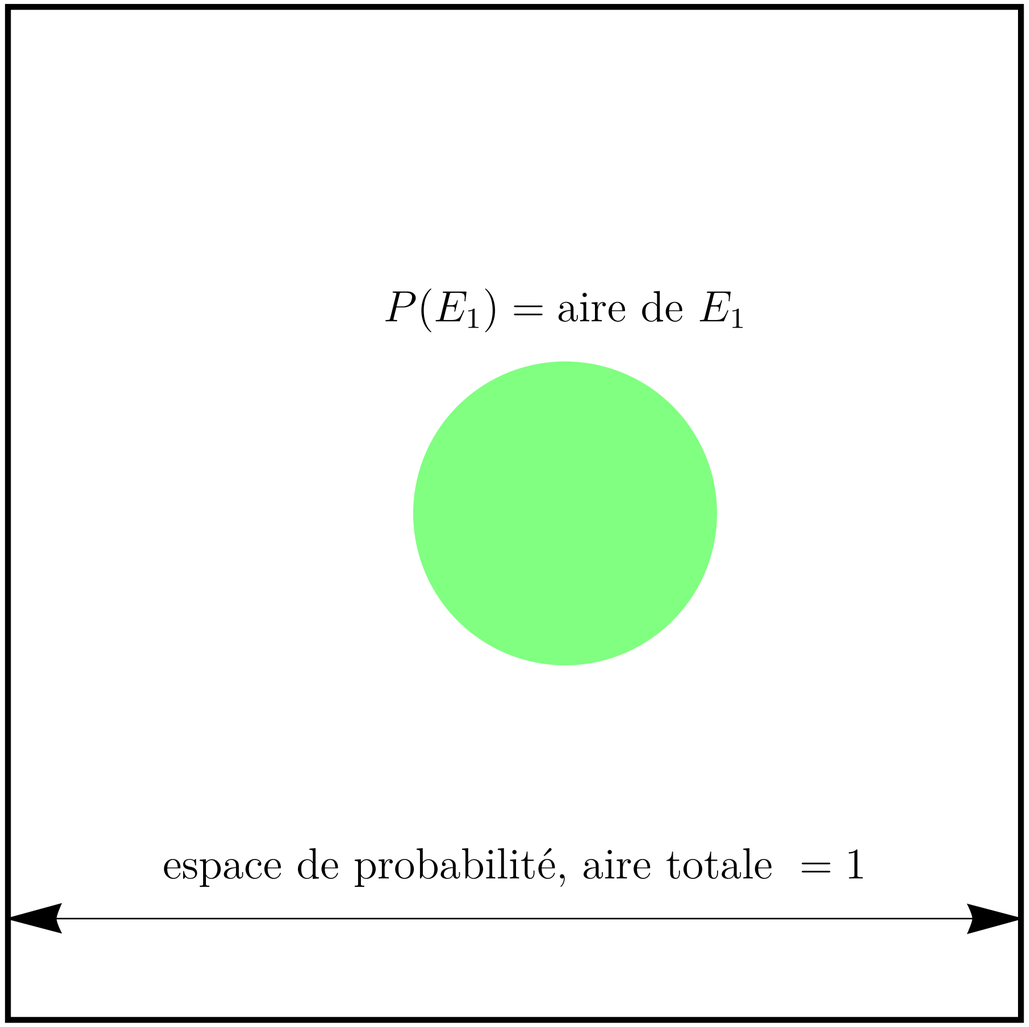
$$P(\text{non } E_1) = 1 - P(E_1).$$

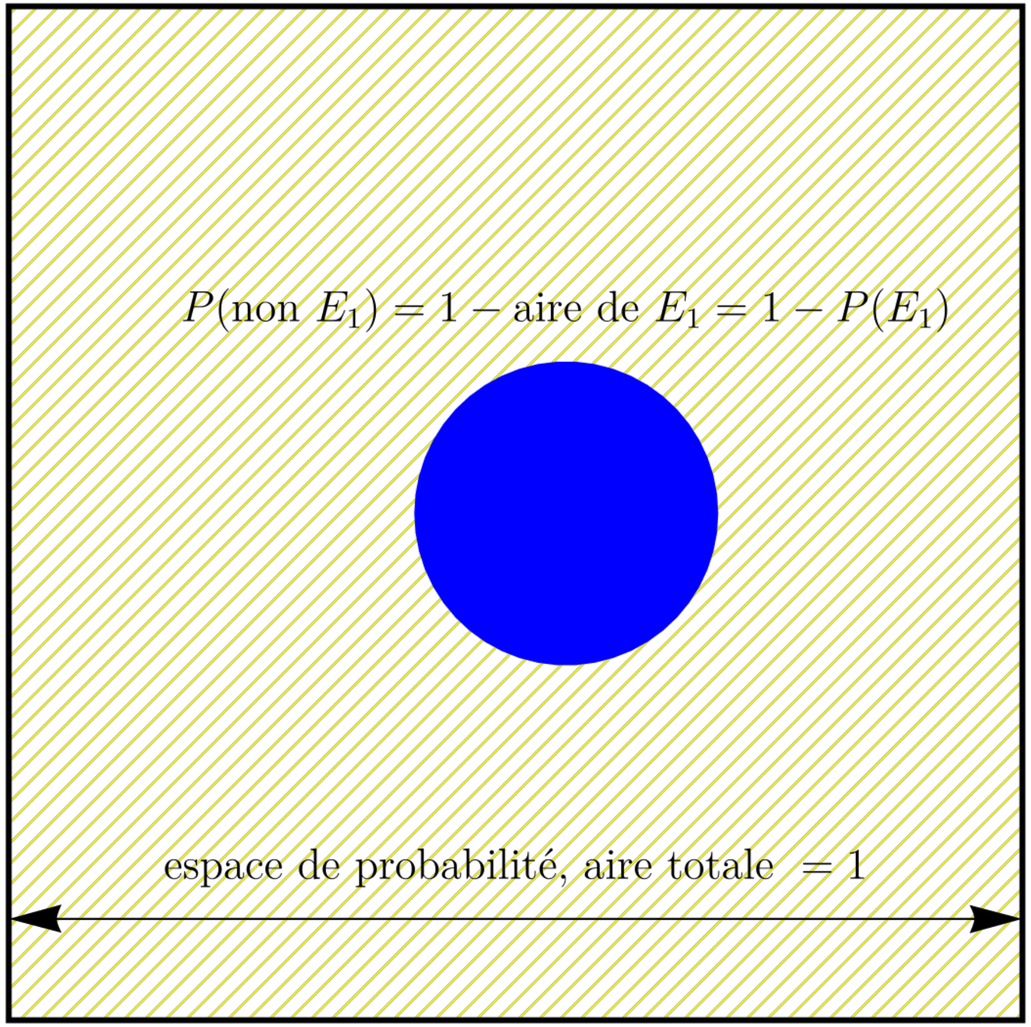
- *probabilité d'observer deux événements*. Deux notations :
  - $E_1 \bigcup E_2$: on observe E_1 ou E_2 (ou bien encore au moins une des propositions E_1 ou E_2 est vraie)
  - $E_1 \bigcap E_2$: on observe à la fois E_1 et E_2 (ou bien encore à la fois E_1 et E_2 sont vraies).

  La probabilité d'observer E_1 ou E_2 est égale à la somme des probabilités d'observer individuellement E_1 et E_2 moins la probabilité d'observer E_1 et E_2 ensemble (afin de ne pas compter deux fois le même événement)

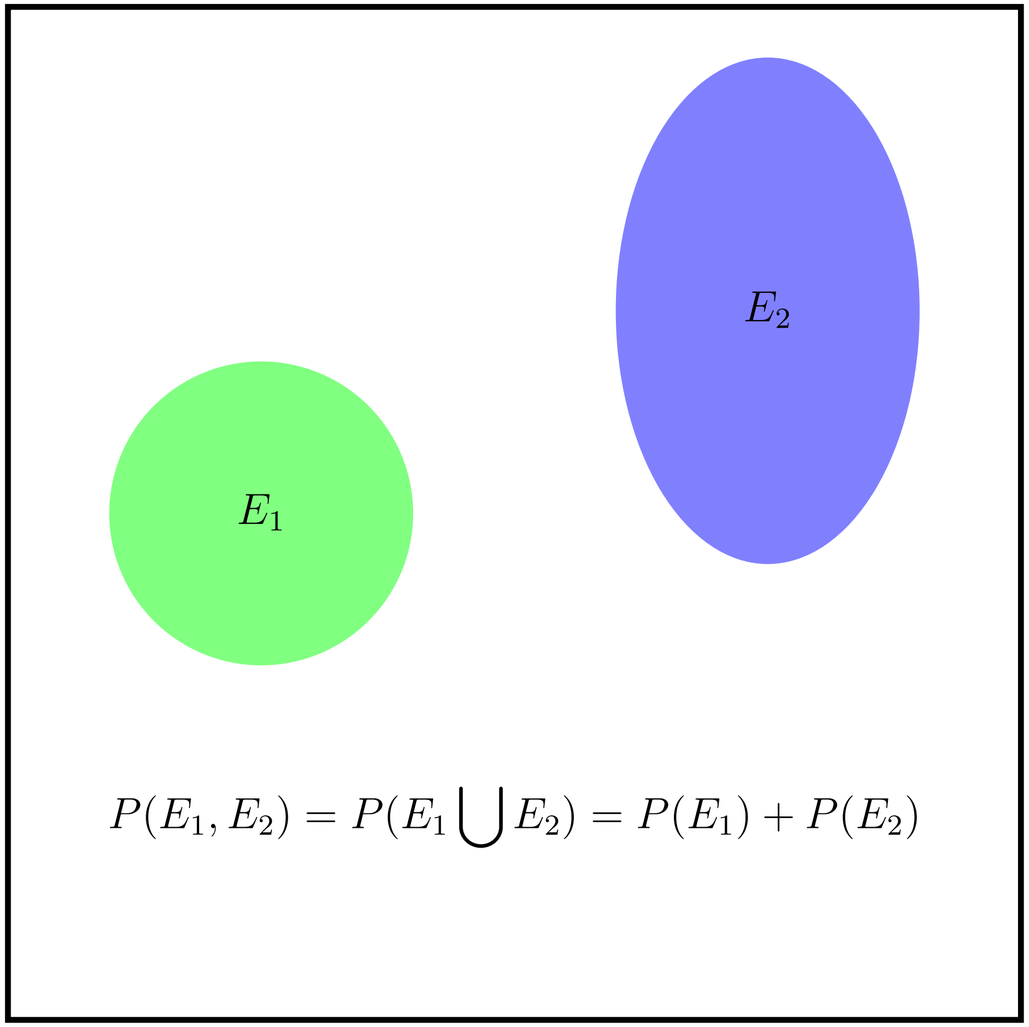
$$P(E_1 \bigcup E_2) = P(E_1) + P(E_2) - P(E_1 \bigcap E_2).$$

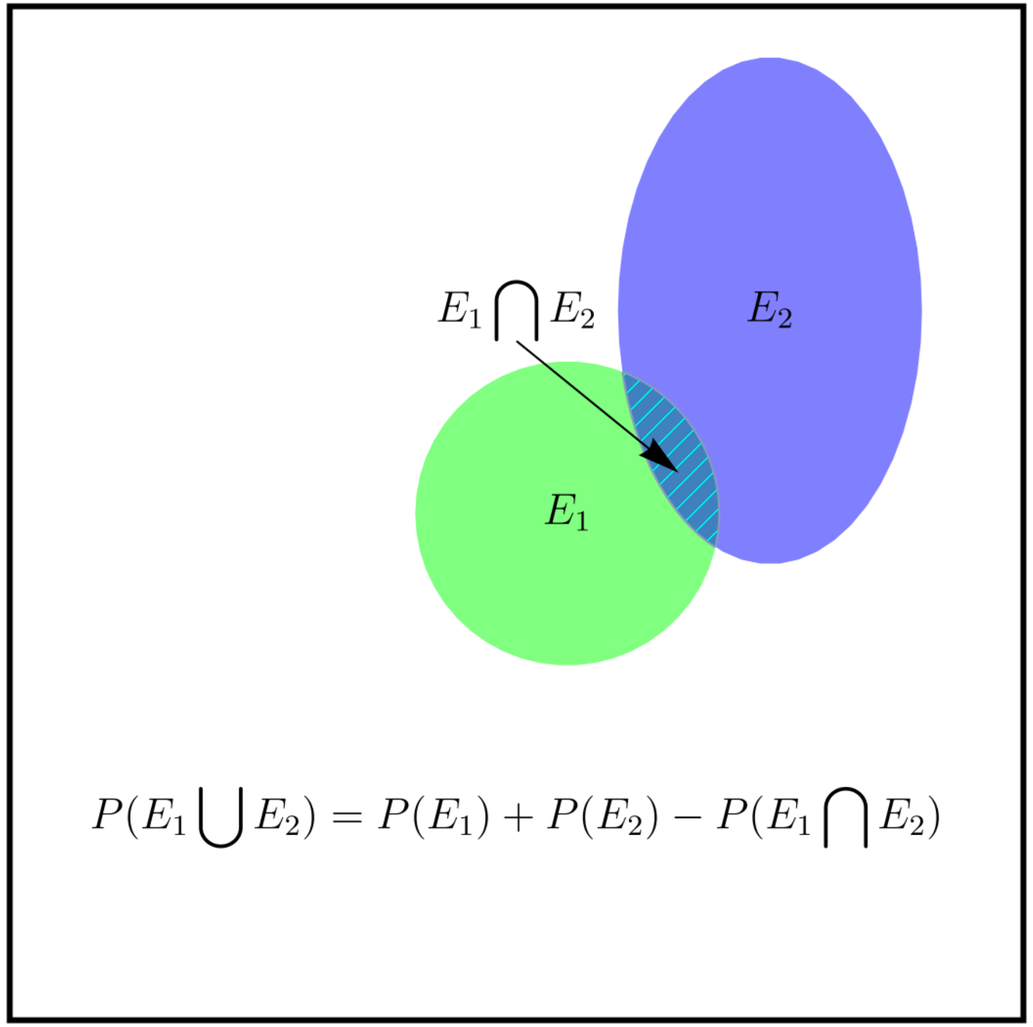
Définitions et propriétés



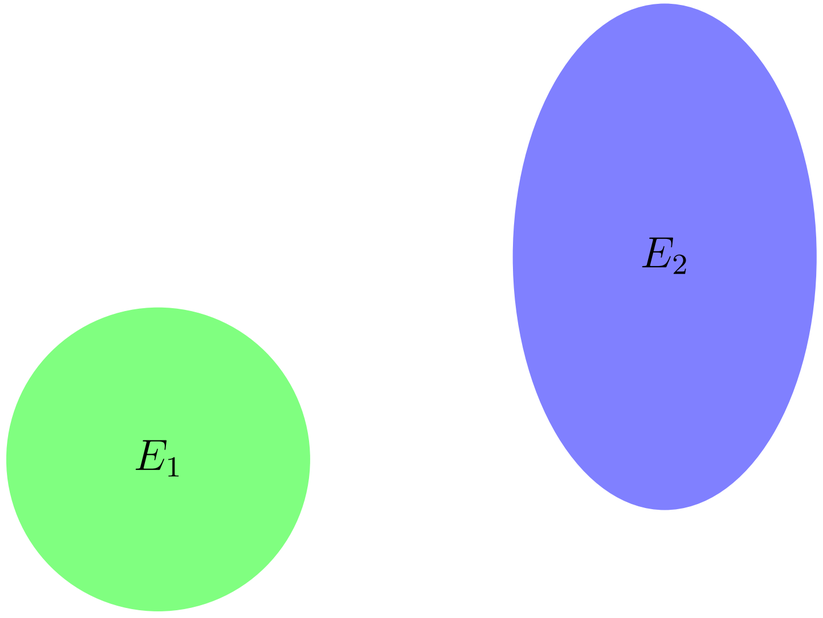


Définitions et propriétés (2)

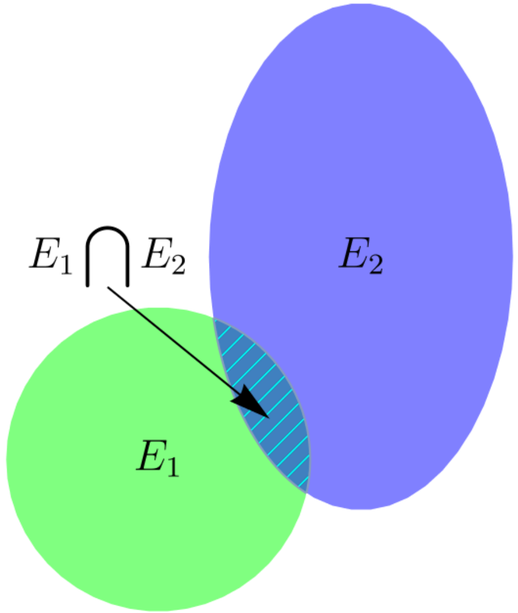




Définitions et propriétés (3)



$$P(E_1 \bigcap E_2) = P(E_1) \times P(E_2)$$



$$P(E_1 \bigcap E_2) = P(E_2)P(E_1|E_2) = P(E_1)P(E_2|E_1)$$

$$P(E_1|E_2) = \frac{P(E_1 \bigcap E_2)}{P(E_2)}$$

Définitions

- *Probabilité conditionnelle* : probabilité d'observer E_1 **sachant que** E_2 est observé. C'est la probabilité d'observer à la fois E_1 et E_2 sur la probabilité d'observer E_2 séparément.

$$P(E_1|E_2) = \frac{P(E_1 \bigcap E_2)}{P(E_2)}.$$

  Si les deux événements sont dits (statistiquement) *indépendants* alors :

  $P(E_1|E_2) = P(E_1)$

- *Première relation de Bayes* :

$$P(E_2|E_1) = \frac{P(E_1|E_2)P(E_2)}{P(E_1)}.$$

Théorème de Bayes

- *Seconde relation de Bayes* : en combinant avec la première règle de composition (étendue à une série de n événements complémentaires (exclusifs les uns par rapport aux autres : $\sum_{i=1}^{n} P(E_i) = 1$)), on obtient :

$$P(E_j|F) = \frac{P(F|E_j)P(E_j)}{P(F)},$$

où F désigne un événement quelconque et où

$$P(F) = \sum_{i=1}^{n} P(F|E_i)P(E_i).$$

- Pour des variables aléatoires continues (y et z deux variables aléatoires) :

$$P(z|y) = \frac{P(y|z)P(z)}{\int P(y|z)P(z)\mathrm{d}z}$$

Théorème de Bayes : exemple 1

Révérend Thomas Bayes

Comparez les énoncés :

- J'ai deux enfants. Quelle est la probabilité pour que les deux soient des fils ?
- J'ai deux enfants, dont un fils. Quelle est la probabilité pour que l'autre soit aussi un fils ?

Théorème de Bayes : exemple 1 (2)

Raisonnement par dénombrement. On a 4 possibilités équiprobables (25 %), qui sont

- A- une fille puis une fille
- B- un garçon puis une fille
- C- une fille puis un garçon
- D- un garçon puis un garçon

Dans le 1er énoncé, je déduis que la réponse est 25 %. Dans le 2nd énoncé, j'ai eu un fils. Du coup la combinaison A avec les deux filles est à retirer de ma liste des possibilités. Donc, si je regarde parmi les combinaisons qui me restent, seule la D satisfait au problème. La probabilité est donc 1/3.

Théorème de Bayes : exemple 1 (3)

Appelons F l'événement « l'enfant $E1 = g$ **et** $E2 = g$ » (g : garçon). On a $P(F) = 0{,}25$ (d'après le dénombrement précédent). Appelons (H) la probabilité que l'un **ou** l'autre des enfants soit un garçon. On a $P(H) = 0{,}75$.

$$P(F|G) = \frac{P(G|F)P(F)}{P(G)} = \frac{1 \times 0{,}25}{0{,}75} = \frac{1}{3}$$

Morale de l'histoire : il ne faut pas confondre la probabilité (dite marginale) $P(F)$ et la probabilité conditionnelle $P(F|G)$. L'information modifie la valeur des probabilités.

Théorème de Bayes : exemple 2

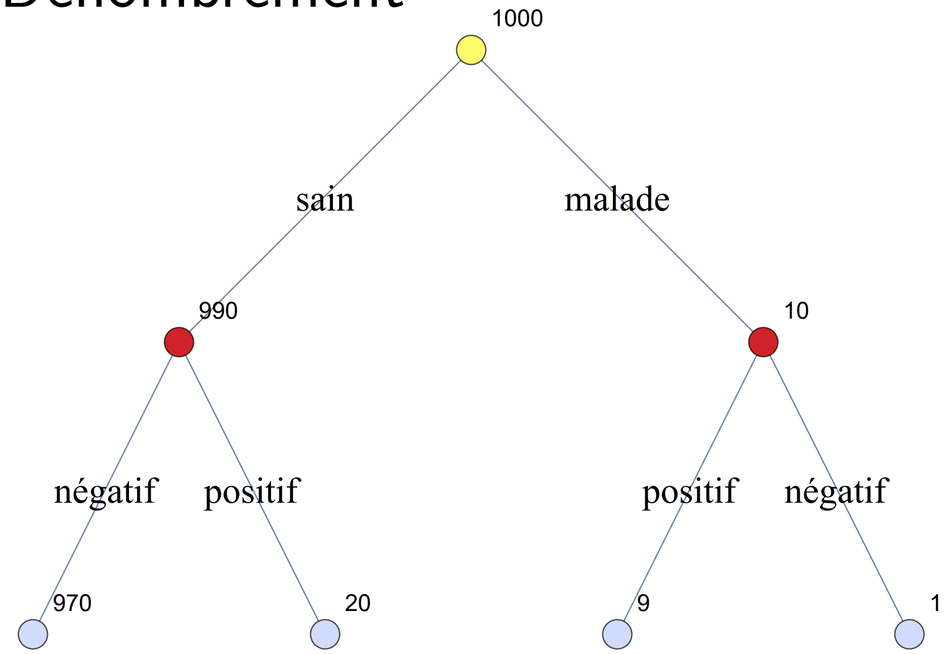
Données (chiffres d'octobre 2020) :

- Actuellement 0,55 % de la population suisse a la covid, mais disons que 1 % l'a
- Si vous passez un test au bout d'une semaine, le test PCR détectera la maladie dans 30 % des cas, 70 % des cas au bout de 2 semaines, 90 % au bout de 3 semaines, et pas plus après. Pour aller dans le sens de la sécurité, on peut supposer que le test donne un vrai positif dans 90 % des cas (donc 10 % de faux négatif) ;
- Si la personne n'a pas la covid, il y a une probabilité de 2 % que le test soit quand même positif (faux positif).

Vous passez le test. Il est positif. Quelle est la probabilité que vous soyez réellement malade ?

Théorème de Bayes : exemple 2 (2)

Dénombrement



Sur 1000 personnes :

- 990 pas malades, donc s'ils font le test, $0{,}02 \times 990 = 19{,}8 \sim 20$ des faux positifs et 970 négatifs
- 10 malades, donc s'ils font le test, $0{,}9 \times 10 = 9$ positifs, et 1 faux négatif.

Si on est positif, cela représente une population de 20 + 19 = 29 personnes. La probabilité pour qu'on soit positif et malade est donc : $P = 9/29 \approx 31$ %

Théorème de Bayes : exemple 2 (3)

Résolution avec le théorème de Bayes :

$$P(\text{malade}|\text{positif}) = \frac{P(\text{positif}|\text{malade})P(\text{malade})}{P(\text{positif})}$$

$P(\text{malade}) = 0{,}01$, $P(\text{sain}) = 1 - P(\text{malade}) = 0{,}99$, $P(\text{positif}|\text{malade}) = 0{,}9$, et $P(\text{positif}|\text{sain}) = 0{,}02$. Pour le dénominateur, la règle de composition des probabilités nous donne :

$$P(\text{positif}) = P(\text{positif}|\text{malade})P(\text{malade}) + P(\text{positif}|\text{sain})P(\text{sain}),$$

et $P(\text{positif}) = 0{,}9 \times 0{,}01 + 0{,}02 \times 0{,}99 = 0{,}0288$ donc on trouve :

$$P(\text{malade}|\text{positif}) = \frac{0{,}9 \times 0{,}01}{0{,}0288} = 31{,}25\ \%.$$

Théorème de Bayes : exemple 3

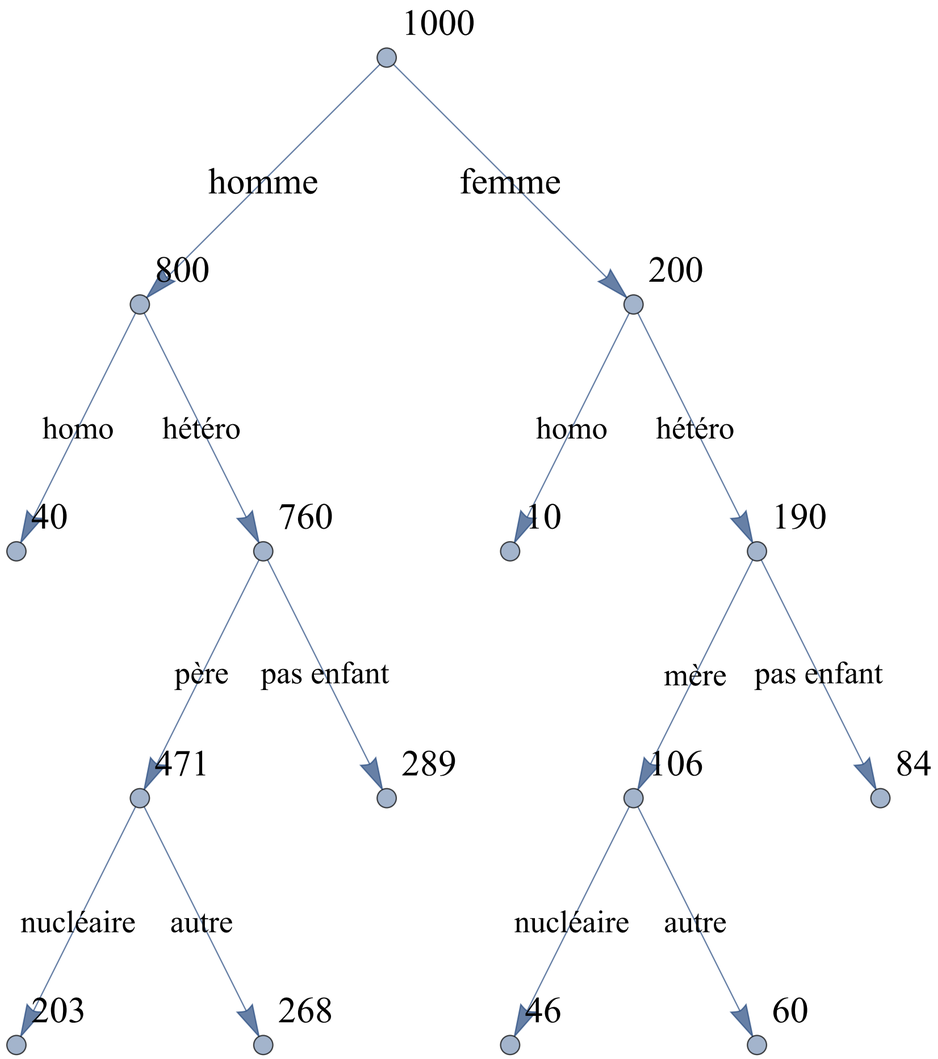
Selon Mikaela Wapman et and Deborah Belle, le langage véhicule des “biais de genre,” c’est-à-dire à des stéréotypes sur les sexes. Elles veulent le mettre en évidence à travers cette devinette posée à 472 étudiants :

“A father and his son are out driving and are involved in a terrible accident. The father is killed instantly, and the son is in critical condition. The son is rushed to the hospital and prepared for an operation that could save his life. The surgeon comes in, sees the patient, and exclaims, “I can’t operate, that boy is my son!’ How can this be? ”

Résultat du sondage : 30 % des étudiants considéraient que le chirurgien était la mère, et 67 % en déduisirent qu’il s’agissait d’un père (beau-père, père adoptif, second père d’un couple homosexuel). Dans le détail, 19 % des garçons ont pensé à la mère, contre 36 % des filles.

Belle, D., A.B. Tartarilla, M. Wapman, M. Schlieber, and A.E. Mercurio, “I Can’t Operate, that Boy Is my Son:´´ *Gender Schemas and a Classic Riddle, Sex Roles,* 1-11, 2021.

Théorème de Bayes : exemple 3 (2)



Que dit l'analyse bayésienne du problème ?

- 20 % des chirurgiens sont des femmes dans les services d'urgence américains
- On compte environ 5 % d'homosexuels, 62 % des hommes sont pères, et 56 % des femmes sont mères
- Un enfant a 43 % de chances d'être élevé dans une famille nucléaire, et 57 % de chances de vivre dans une famille recomposée ou un parent seul

La probabilité que le chirurgien soit une femme est $106/(106+268)=28$ % contre $268/(106+268)=72$ % pour un homme (il est alors le beau-père) ; si on inclut les gays, alors ce chiffre grimpe à 73 %.

Le résultat est conforme au mode de fonctionnement bayésien du cerveau. L'étude de Belle et *al.* ne permet pas de conclure à un "biais de genre."

Loi de probabilité

Convention. Variable aléatoire X, et sa valeur particulière x est le *quantile*.

Si la variable est discrète, la loi de probabilité fournit la probabilité d'observer dans quel état est le système :

$$P_X(X = x) = \mathrm{prob}(X \text{ prend la valeur } x).$$

C'est un nombre entre 0 et 1.

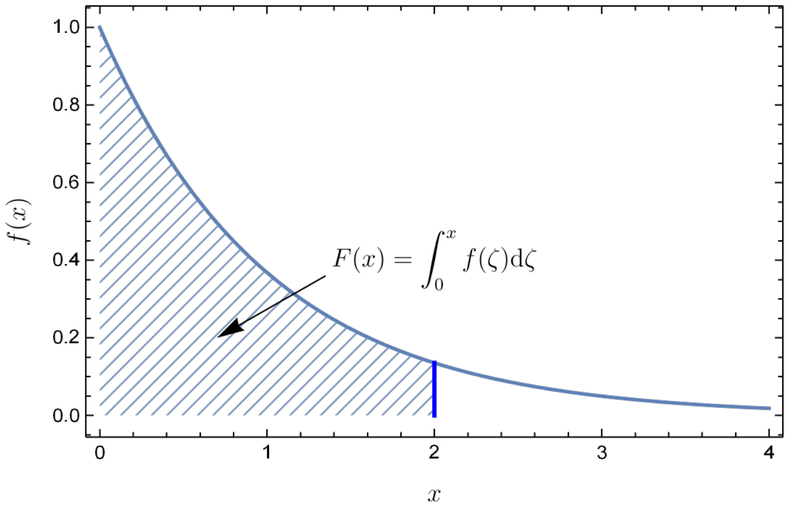
Si la variable est continue, on introduit la *densité de probabilité* $f(x)$ qui est la probabilité d'observer l'état du système dans un certain voisinage $\mathrm{d}x$:

$$f(x)\mathrm{d}x = P_X(x \leqslant X \leqslant x + \mathrm{d}x).$$

f est positive, mais peut prendre des valeurs > 1. On a :

$$f = \frac{\mathrm{d}P_X}{\mathrm{d}x}$$

Fonction de répartition



Pour une densité de probabilité f de support $[a, b]$, on appelle *fonction de répartition* F_X l'intégrale de f :

$$F_X(x) = P(X \leq x) = \int_a^x f(u)\mathrm{d}u$$

C'est la probabilité que la variable aléatoire ne dépasse par une valeur donnée x : *probabilité de non-dépassement*. La quantité complémentaire est la *probabilité de dépassement* :

$$1 - F_X(x) = P(X \geq x) = \int_x^b f(u)\mathrm{d}u.$$

Fonction de répartition (2)

À noter : $F_X(b) = P(X \leqslant b) = \int_a^b f(u)\mathrm{d}u = 1$. La fonction F_X doit tendre vers 1. Inversement $F_X(a) = P(X \leqslant a) = 0$. La fonction de répartition est une probabilité, donc comprise entre 0 et 1. On déduit

$$\mathrm{prob}[a \leq X \leq b] = \int_a^b f(x)\mathrm{d}x = F_X(b) - F_X(a).$$

Changement de variable : $x \to y = v(x)$ (g densité de probabilité de Y), la probabilité doit rester invariante :

$$f(x)\mathrm{d}x = P_X(x \leq X \leq x + \mathrm{d}x) = P_Y(y \leq Y \leq y + \mathrm{d}x) = g(y)\mathrm{d}y,$$

Et donc :

$$g(y) = f(x)\frac{\mathrm{d}x}{\mathrm{d}y} = f(x)|v'(x)|^{-1}.$$

Moyenne et variance

On appelle *moyenne* (ou espérance ou moment d'ordre 1) la moyenne arithmétique des différentes valeurs que X peut prendre, pondérées par leurs probabilités $m = \mathbb{E}(X) = \sum_i x_i P(X = x_i)$. On appelle *variance* (ou moment centré d'ordre 2) :

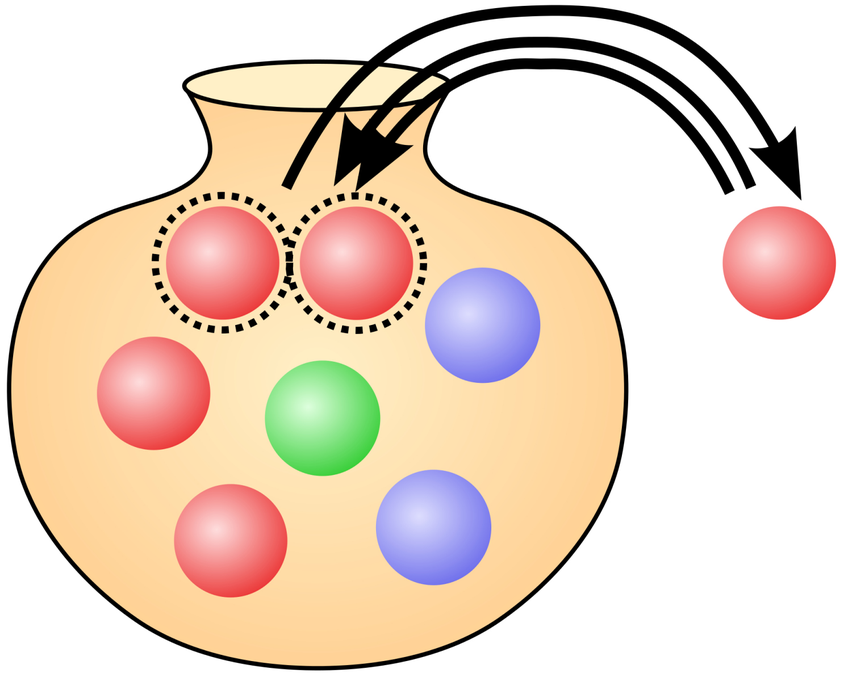
$$\sigma^2 = \mathrm{var}X = \sum_i (x_i - m)^2 P_X(X = x_i).$$

Les équivalents pour une variable continue sont :

$$m = \mathbb{E}(X) = \int_a^b x f(x) \mathrm{d}x = \int_a^b x \mathrm{d}P_X$$

$$\sigma^2 = \mathrm{var}X = \mathbb{E}[(X - m)^2] = \int_a^b (x - m)^2 f(x) \mathrm{d}x$$

Loi de Bernoulli



Modèle d'urne

Il s'agit d'une loi discrète à un paramètre p d'une variable N qui peut prendre deux valeurs (0 ou 1 par exemple) avec les probabilités p et $1-p$ respectivement. On parle aussi de modèle d'urne : si l'on place des boules noires et blanches et qu'il y a une proportion p de boules blanches, alors la probabilité de tirer au hasard une blanche est p.

La moyenne est : $\mathbb{E}(N) = p$; la variance est : $\mathrm{var} N = p(1-p)$.

Loi binomiale

Supposons que l'on répète m fois l'expérience de tirage de boule ; après chaque tirage, on replace la boule dans l'urne (pour que le nombre de boules soit identique). On note N le nombre de fois qu'une boule blanche est apparue dans cette séquence de m tirages. La probabilité que $N = k$ est :

$$\mathrm{Bin}(m,\, p)(k) = \mathrm{prob}(N = k) = C_m^k p^k (1-p)^{m-k}.$$

La moyenne est : $\mathbb{E}(N) = mp$; la variance est : $\mathrm{var}N = mp(1-p)$.

Exemple : crue centennale $p = 1/100$. On a une série de $m = 100$ années.

Probabilité d'observer 0, 1 et 2 crues ? $\mathrm{Bin}(m,\, p)(0) = 0{,}3660$, $\mathrm{Bin}(m,\, p)(1) = 0{,}3697$ et $\mathrm{Bin}(m,\, p)(2) = 0{,}1848$.

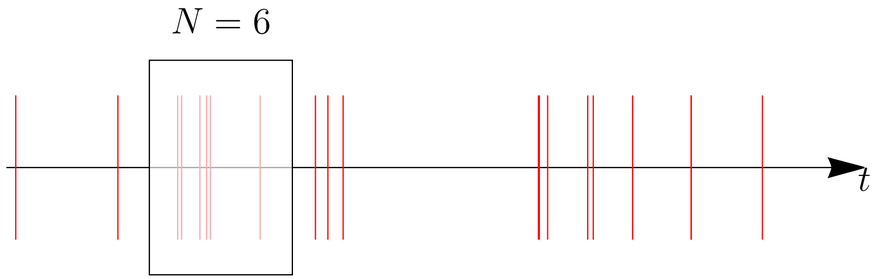
Loi binomiale négative

On appelle N le nombre de tirages qu'il faut réaliser pour obtenir un ensemble de k succès. On montre que :

$$\mathrm{Neg}(k,\, p)(i) = \mathrm{prob}(N = i) = C_{i-1}^{k-1} p^k (1-p)^{i-k}.$$

La moyenne est : $\mathbb{E}(N) = k(1-p)/p$; la variance est : $\mathrm{var} N = k(1-p)/p^2$. La variance est toujours supérieure à la moyenne. En pratique, cette loi peut se révéler utile en remplacement de la loi de Poisson pour décrire des processus hydrologiques instationnaires. En effet, la loi binomiale négative peut être vue comme une loi de Poisson dont le taux est lui-même aléatoire et distribué selon une loi gamma.

Loi de Poisson



Série temporelle d'événements distribués selon une loi de Poisson

Nombre d'événements se produisant dans un intervalle de temps fixé

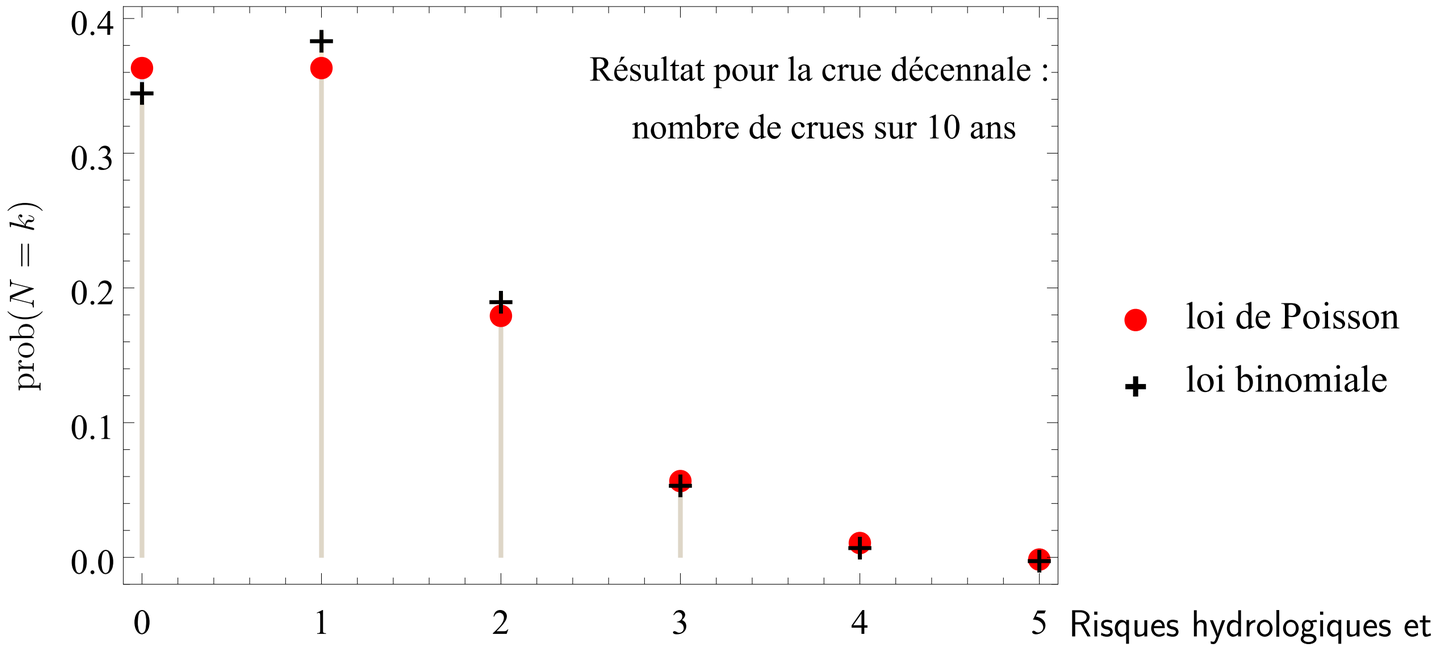
$$\mathrm{Po}(\lambda)(k) = \mathrm{prob}(N = k) = e^{-\lambda}\frac{\lambda^k}{k!}$$

La moyenne est : $\mathbb{E}(N) = \lambda$; la variance est : $\mathrm{var}N = \lambda$. Lois binomiale et de Poisson sont reliées. Si on décompte les événements sur l'intervalle T (entier) : $(m = T,\, p = \lambda/T)$, alors $\mathrm{Bin}(m,\, p) \to \mathrm{Po}(\lambda)$ quand $T \gg 1$.

Loi de Poisson (2)

Exemple : nombre de crues décennales par décennie. Par définition $\lambda = 1$. On a :

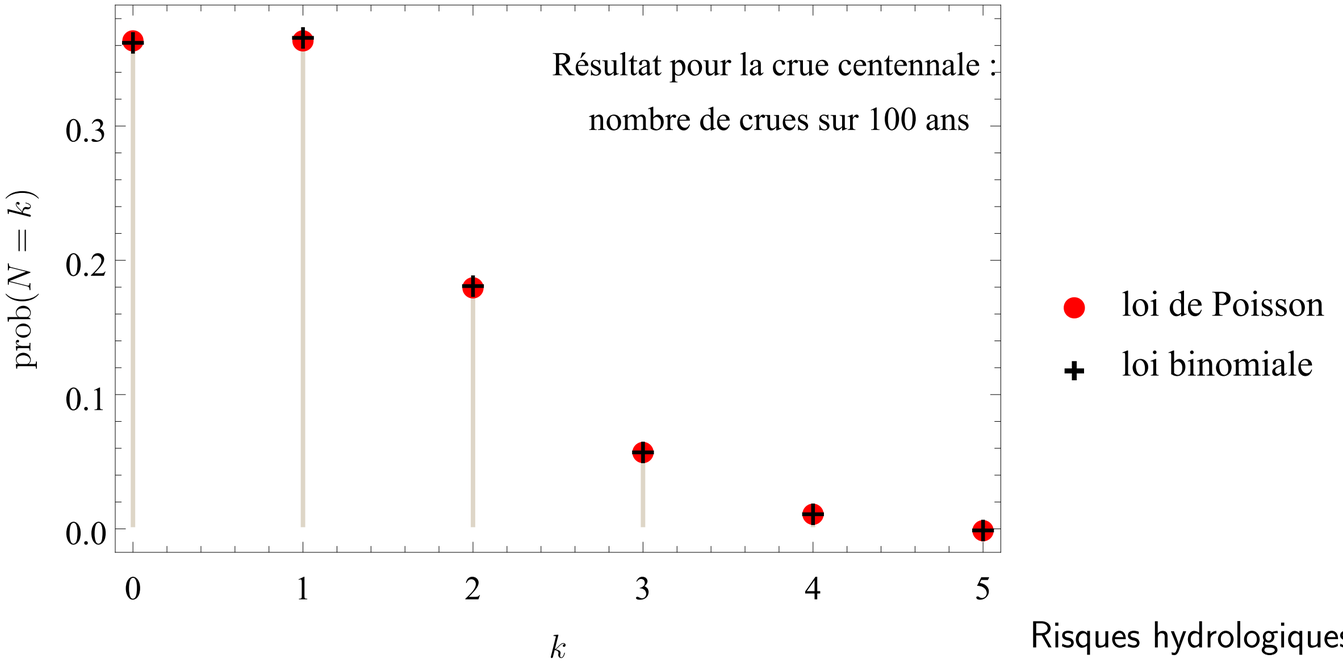
- 0 crue : $\mathrm{Po}(\lambda = 1)(0) = 0{,}36787$ (on avait $\mathrm{Bin}(10,\, 1/10)(0) = 0{,}34867$)
- 1 crue : $\mathrm{Po}(\lambda = 1)(1) = 0{,}36787$ (on avait $\mathrm{Bin}(10,\, 1/10)(1) = 0{,}38742$)
- 2 crues : $\mathrm{Po}(\lambda = 1)(2) = 0{,}18394$ (on avait $\mathrm{Bin}(10,\, 1/10)(2) = 0{,}19371$)



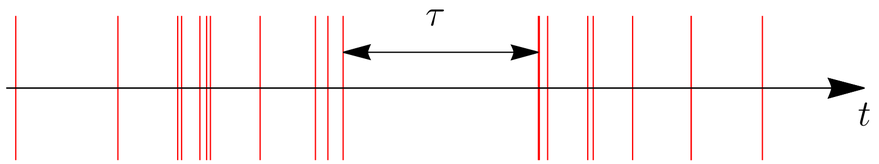
Loi de Poisson (3)

Exemple : nombre de crues centennales par siècle. Par définition $\lambda = 1$. On a donc :

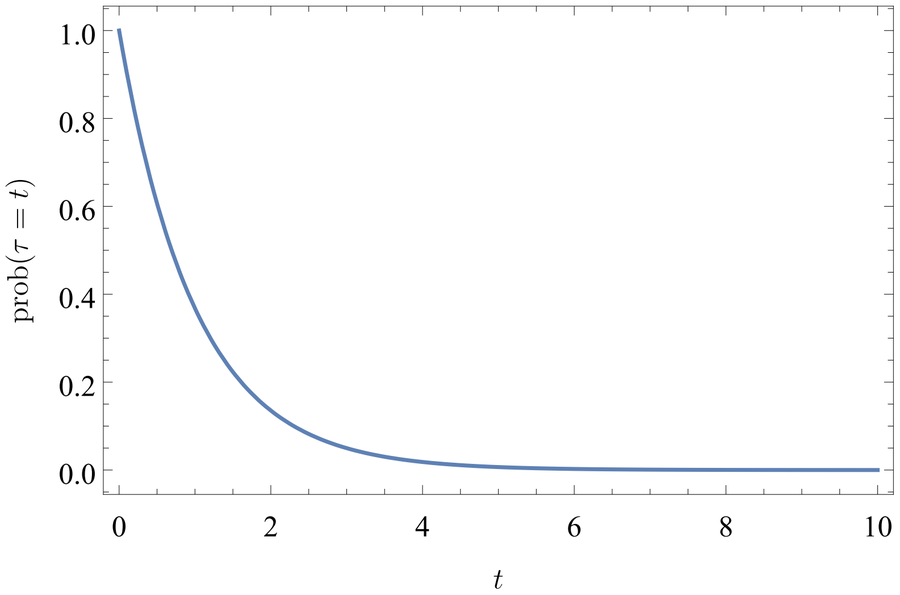
- 0 crue : $\mathrm{Po}(\lambda = 1)(0) = 0{,}36787$ (on avait $\mathrm{Bin}(100,\ 1/100)(0) = 0{,}36660$)
- 1 crue : $\mathrm{Po}(\lambda = 1)(1) = 0{,}36787$ (on avait $\mathrm{Bin}(100,\ 1/100)(1) = 0{,}36973$)
- 2 crues : $\mathrm{Po}(\lambda = 1)(2) = 0{,}18394$ (on avait $\mathrm{Bin}(100,\ 1/100)(2) = 0{,}18486$)



Loi exponentielle



Série temporelle d'événements distribués selon une loi de Poisson

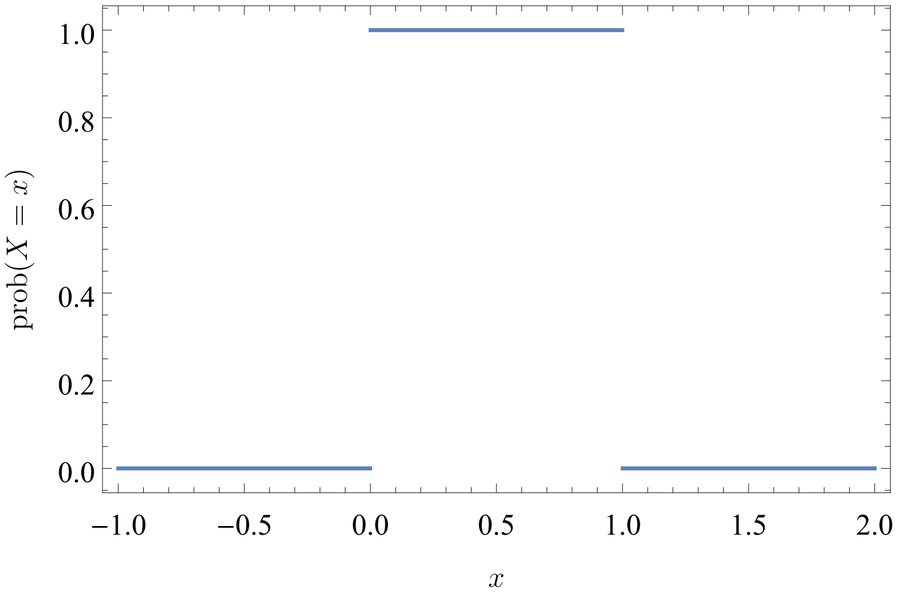


C'est une loi continue qui est le pendant de la loi de Poisson. On appelle τ le temps entre deux événements :

$$\mathrm{Exp}(\lambda)(\tau = t) = \lambda e^{-\lambda t}$$

La moyenne est : $\mathbb{E}(\tau) = 1/\lambda$; la variance est : $\mathrm{var}\tau = 1/\lambda^2$.

Loi uniforme



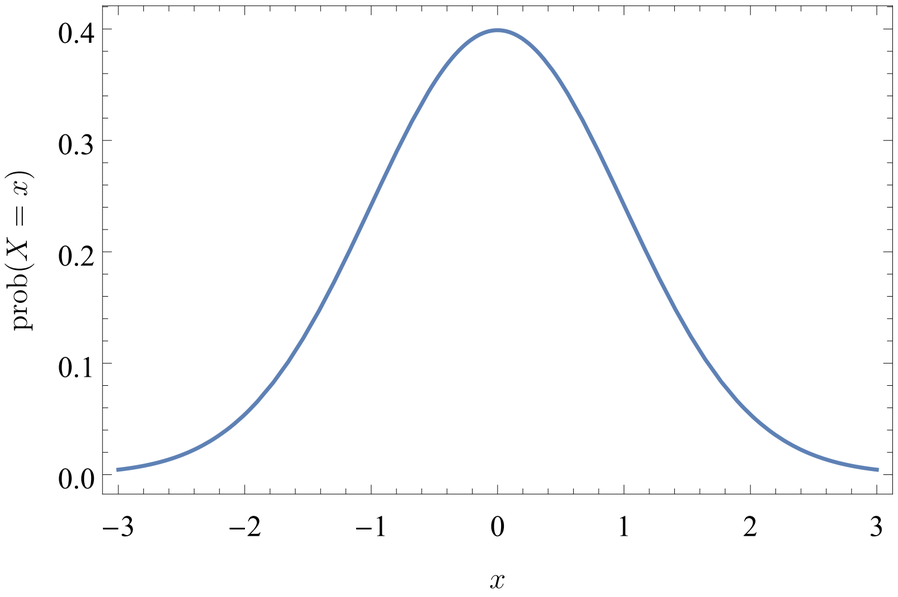
Loi uniforme sur [0, 1]

Loi continue définie sur un intervalle $[a,\, b]$ (aucun paramètre hormis les deux bornes). La densité de probabilité est constante :

$$\mathrm{U}[a,\, b](x) = \begin{cases} 0 \text{ si } x < a \\ \frac{1}{b-a} \text{ si } a \leq x \leq b \\ 0 \text{ si } x > b \end{cases}$$

La moyenne est : $\mathbb{E}(X) = 1$; la variance est : $\mathrm{var}X = 0$. Cette loi sert souvent à traduire l'absence d'information ou de connaissance.

Loi normale (Laplace-Gauss)



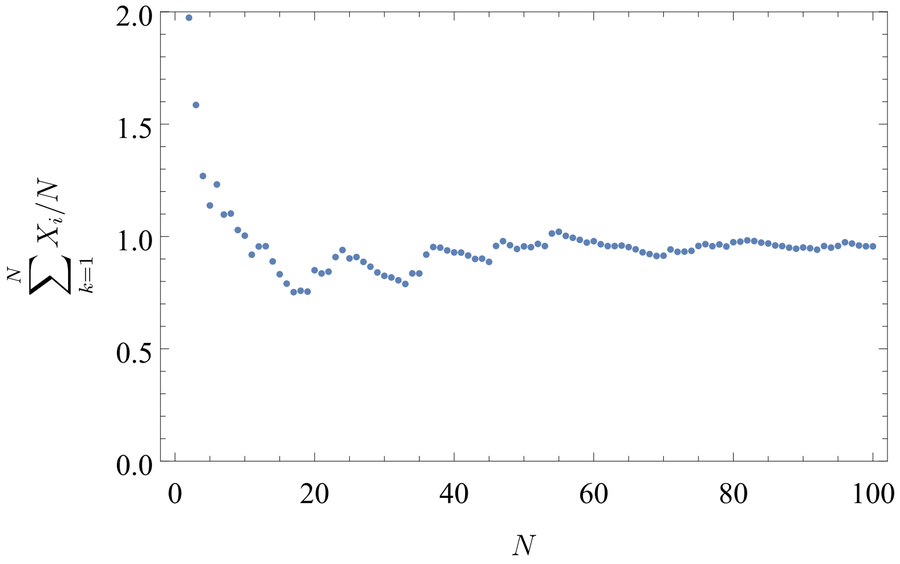
Loi normale $\mu = 0$ et $\sigma = 1$

Une variable X est distribuée selon une loi de Laplace-Gauss de moyenne μ et de variance σ^2 si :

$$\mathrm{No}(\mu,\, \sigma)(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}. \tag{1}$$

C'est une courbe symétrique en forme de cloche autour de la valeur moyenne. La moyenne est : $\mathbb{E}(X) = \mu$; la variance est : $\mathrm{var} X = \sigma^2$.

Loi faible des grands nombres



Somme de $X \sim \Gamma(1, 1)$

$(\mathbb{E}(X) = \mathrm{var}X = 1)$

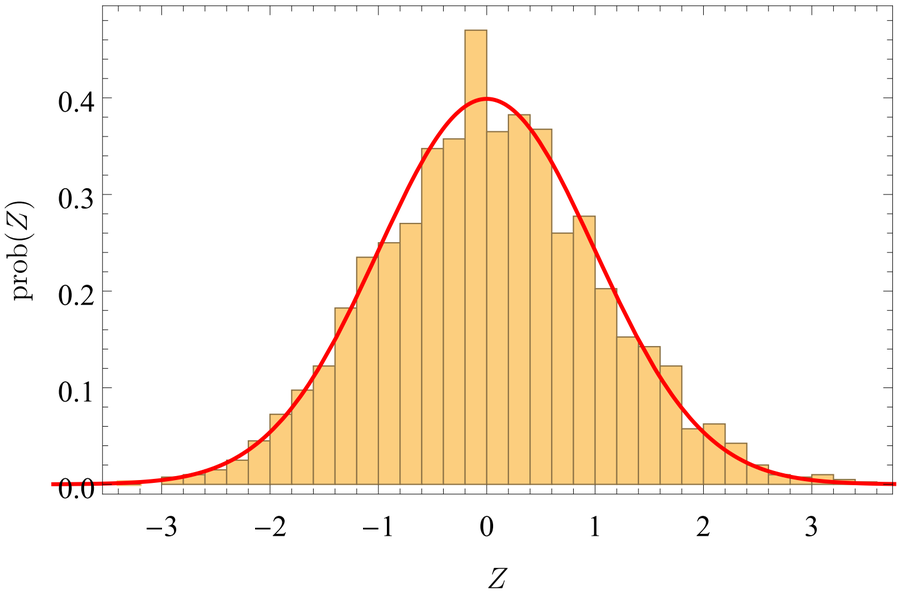
Soit X_1, X_2, $\ldots X_n$ une séquence de variables indépendantes distribuées selon une loi dont les deux moments μ et σ^2 sont finis, alors pour tout $\varepsilon >$, on a :

$$\mathrm{prob}\left(\left|\frac{X_1 + \cdots + X_n}{n} - \mu\right| > \varepsilon\right) \to 0 \text{ quand } n \to \infty.$$

Formulation forte :

$$\lim_{n\to\infty} \frac{X_1 + \cdots + X_n}{n} = \mu.$$

Théorème central de la limite



Histogramme de $Z_n = (X_n - n\mu)/(\sqrt{n}\sigma)$ et comparaison avec $\mathrm{No}(0,\,1)$

Soit $X_1,\, X_2,\, \ldots X_n$ une séquence de variables indépendantes distribuées selon une loi dont les deux moments μ et σ^2 sont finis, alors a :

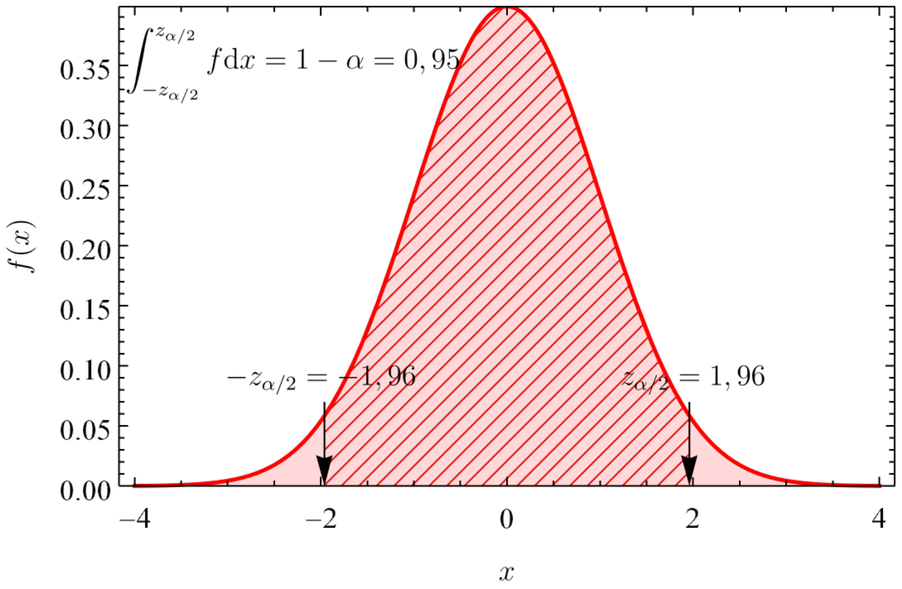
$$\frac{X_1 + \cdots + X_n - n\mu}{\sqrt{n}\sigma} \to \mathrm{No}(0,\,1).$$

Quand on fait des essais avec une variable aléatoire, on peut approcher sa moyenne en prenant :

$$\mu \approx \bar{X} \pm \frac{z\sqrt{Z_n^2}}{\sqrt{n}}$$

avec $z \sim \mathrm{No}(0,\,1)$. Théorème asymptotique de grande importance !

Intervalle de confiance



Loi normale centrée et intervalle de confiance à 95 %

L'intervalle de confiance définit la précision avec laquelle la moyenne empirique $\bar{X}$ tend vers la moyenne μ. On définit $Z_n^2 = (X_1^2 + \cdots + X_n^2)/(n-1) - \bar{X}_n^2$ (écart-type empirique). Soit α un réel (petit) et $z_{\alpha/2}$ le réel tel que :

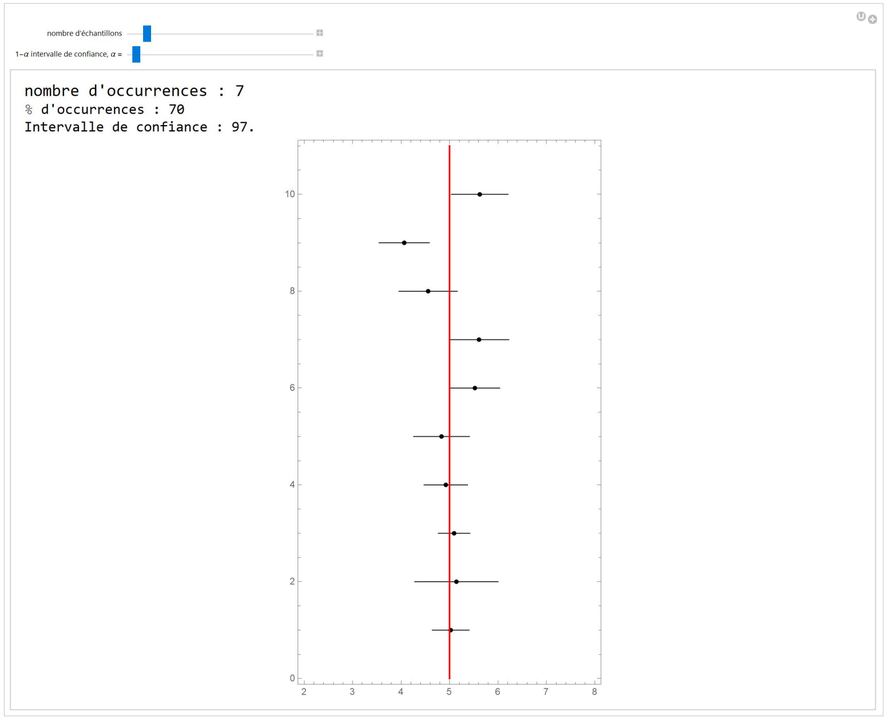
$$\int_{-z_{\alpha/2}}^{z_{\alpha/2}} \mathrm{d}x \frac{1}{\sqrt{2\pi}} e^{-x^2/2} = 1 - \alpha.$$

On pose :

$$T_1 = \bar{X} - \frac{z_{\alpha/2}\sigma}{\sqrt{n}} \qquad T_1' = \bar{X} - \frac{z_{\alpha/2}\sqrt{S_n^2}}{\sqrt{n}},$$

$$T_2 = \bar{X} + \frac{z_{\alpha/2}\sigma}{\sqrt{n}} \qquad T_2' = \bar{X} + \frac{z_{\alpha/2}\sqrt{S_n^2}}{\sqrt{n}},$$

Intervalle de confiance (2)



Jouer avec le notebook Mathematica IntervalleConfiance.nb

On a en théorie

$$\lim_{n\to\infty} \text{prob}[T_1 \leq \mu \leq T_2] = 1 - \alpha$$

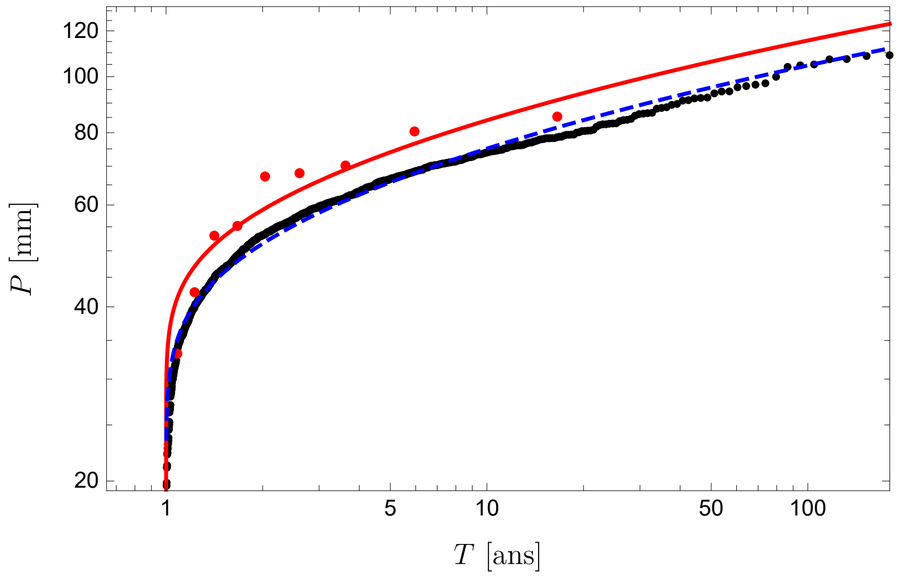
mais aussi de façon pratique

$$\lim_{n\to\infty} \text{prob}[T_1' \leq \mu \leq T_2'] = 1 - \alpha$$

Intervalle de confiance classique :

- à 5 % : $z_{\alpha/2} = \text{quantile}(0{,}975) = 1{,}96$
- à 10 % : $z_{\alpha/2} = \text{quantile}(0{,}95) = 1{,}64$
- à 30 % : $z_{\alpha/2} = \text{quantile}(0{,}85) = 1{,}03$

Ajustement de loi



On a des données. Que fait-on ?

- les données sont-elles indépendantes et distribuées selon la même loi ?
- quelle forme de loi puis-je a *priori* utiliser ?
- comment ajuster les paramètres de cette loi (problème d'*inférence*) ?
- comment vérifier la pertinence du choix d'une forme particulière de loi de probabilité ?
- quelle incertitude ou quelle confiance ai-je dans l'ajustement des paramètres ?

Principe de l'estimation

$x = \{$1,62334, 1,31887, 3,04122, 0,454991, 0,922461, 1,54628, 8,55486, 1,2709, 1,1607, 1,66493$\}$ avec ici $N = 10$

Supposons que j'aie un échantillon x de N valeurs x_i tirées selon une loi gamma $\Gamma(1,2)$ dont la moyenne et la variance sont $m = 2$ et $\sigma^2 = 4$. Je définis la moyenne et la variance empirique de la façon suivante

$$\hat{m} = \frac{1}{N}\sum_{i=1}^{N} x_i \text{ et } \hat{\sigma}^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \hat{m})^2$$

Je veux savoir si on peut préciser la précision et la robustesse d'une telle estimation (à noter le chapeau sur les variables). Les opérateurs introduits sont appelés des *estimateurs*.

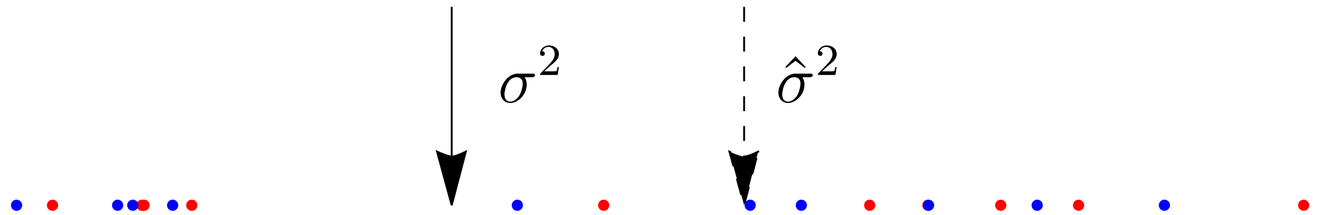
Principe de l'estimation (2)

Quand on a un estimateur $\hat{\theta}$ d'un paramètre théorique (inconnu) θ, on qualifie l'écart entre les deux valeurs avec le biais

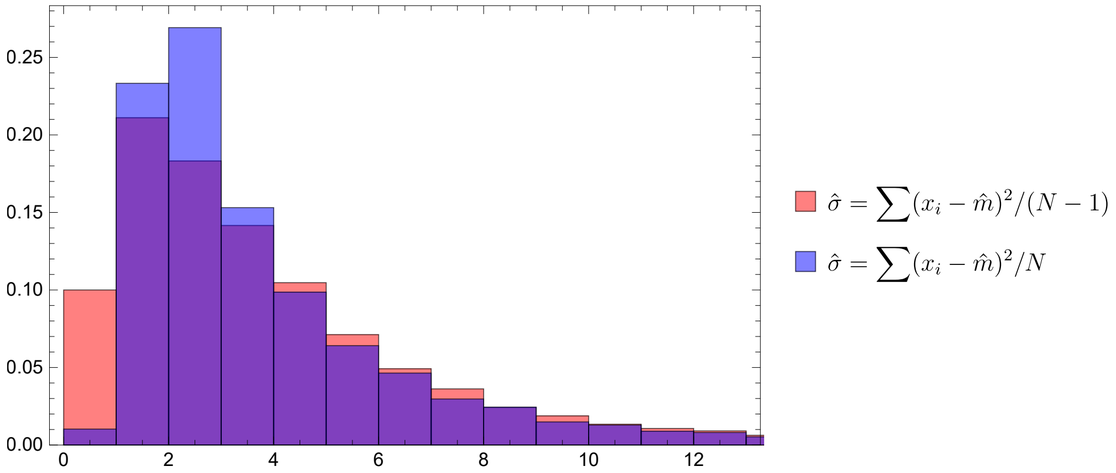
$$\text{Biais}(\hat{\theta}) = \mathbb{E}(\hat{\theta}) - \theta,$$

et l'*erreur quadratique moyenne*

$$\text{MSE}(\hat{\theta}) = \mathbb{E}\left((\hat{\theta} - \theta)^2\right).$$



Principe de l'estimation (3)



Les estimateurs sont optimisés pour réduire à la fois le biais et l'erreur quadratique moyenne.

Par exemple, pour la variance empirique

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

ou bien

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

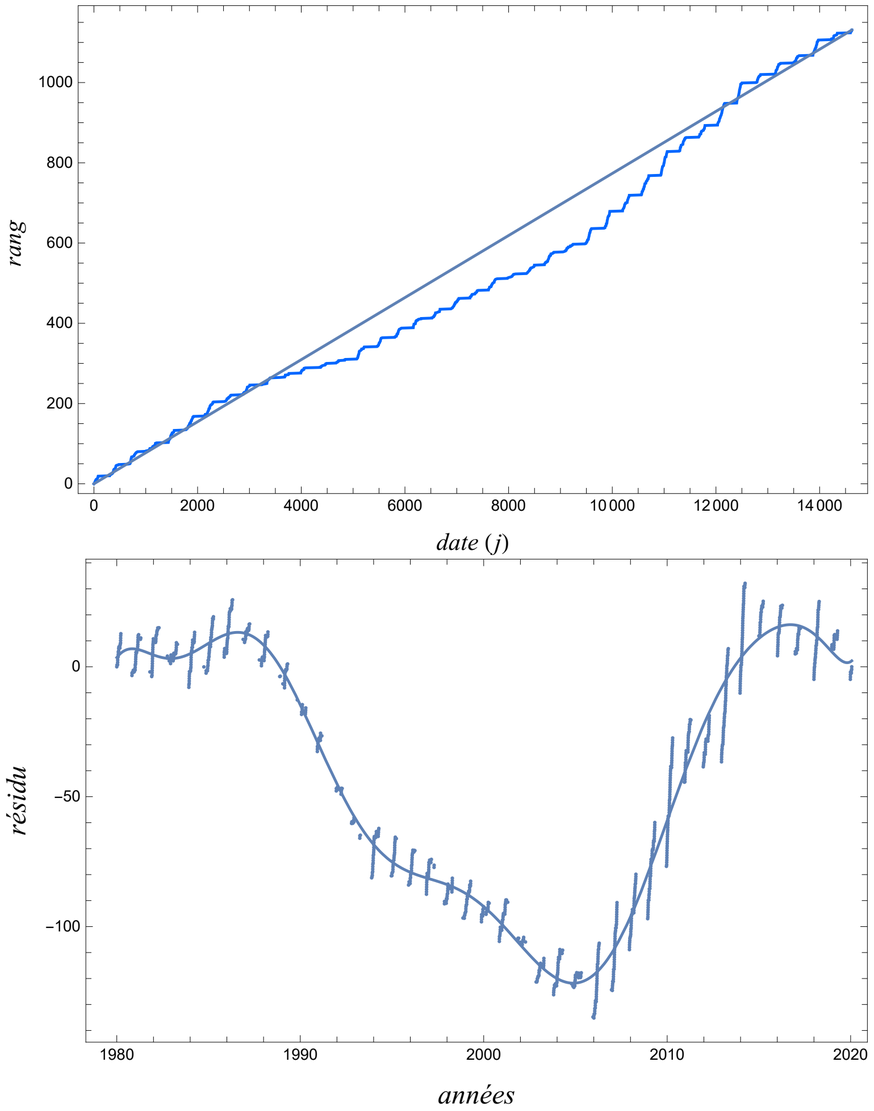
La première forme permet une estimation plus robuste de la variance.

Stationnarité et indépendance

Ajuster une loi de probabilité sur des données suppose que ces données sont bien issues d'une même loi, donc que la loi est unique, et que ces paramètres sont constants.

Indépendance : indépendance des mesures dans le temps des propriétés. Tests : fonction d'autocorrélation, autocorrélation partielle (modèle ARMA) pour les processus à une variable, et tests de Spearman ou Kendall pour les processus à deux variables.

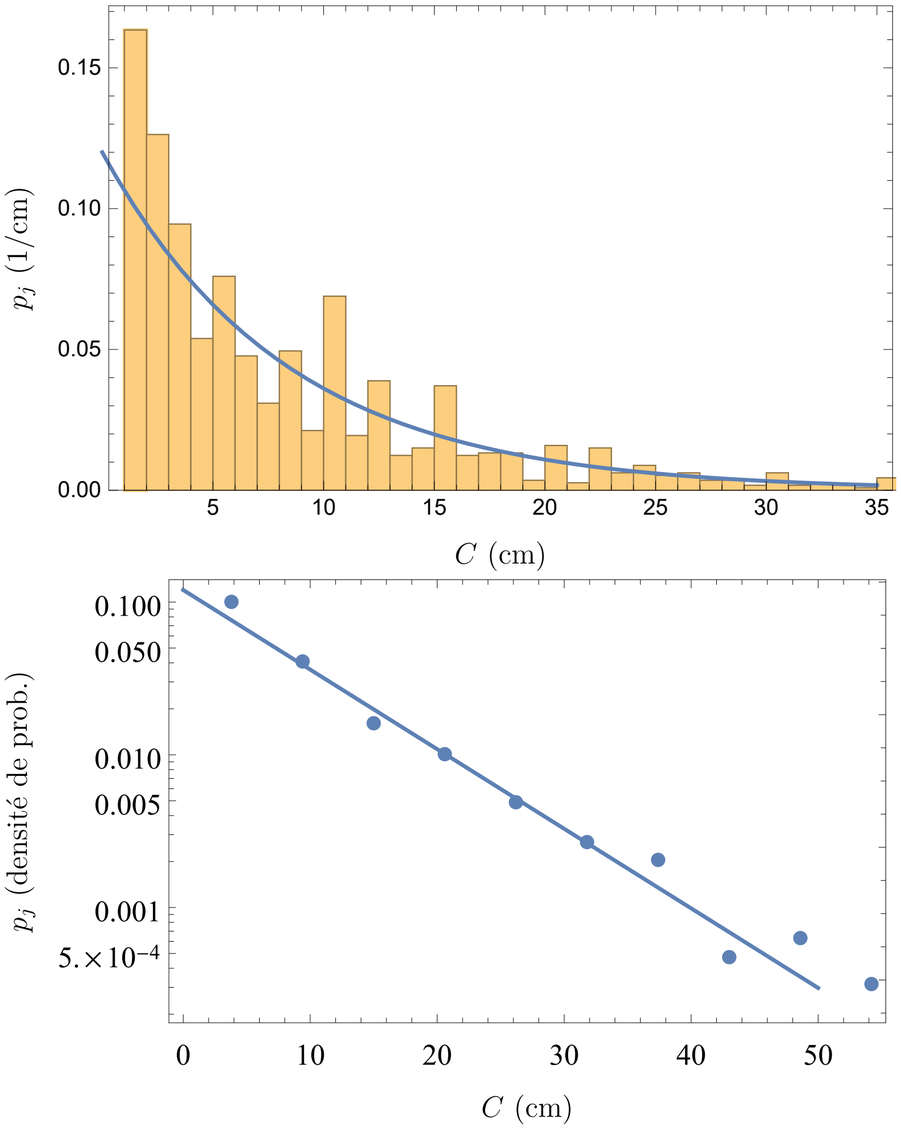
Stationnarité et indépendance (2)



On a N événements $(\text{date}_k)_{1 \leq k \leq N}$ ordonnés chronologiquement sur une durée T :

- report du i$^{\text{ème}}$ point $(\text{date}_i,\ i)$
- tracé de la droite théorique $y = xN/T$ (le point (0, 0) doit correspondre au 1$^{\text{er}}$ événement)
- tracé du résidu $\varepsilon_i = i - y$ (distance entre le i$^{\text{ème}}$ point et la droite théorique) en fonction de i

Densité de probabilité empirique

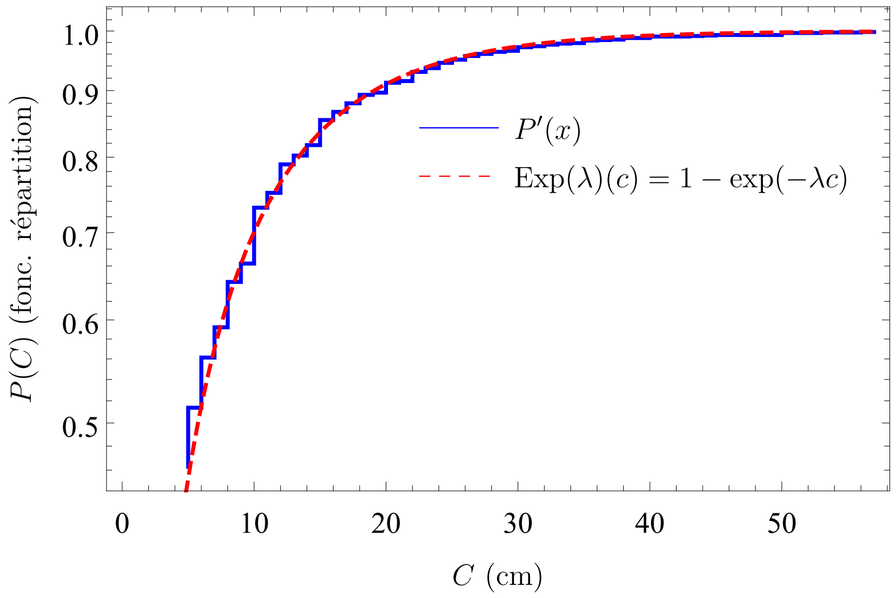


On a N événements $(x_i)_{1\leq i\leq N}$ en ordre croissant :

- partition de l'intervalle $[x_1,\, x_N]$ en n intervalles longueur $\delta = (x_N - x_1)/n$
- les bornes de ces intervalles $y_k = x_1 + (k-1)\delta$ $(1 \leq k \leq n+1)$.
- On compte le nombre m_j d'événements dans chaque intervalle
- La *densité de probabilité empirique* peut alors se définir comme :

$$p_j = \frac{m_j}{\delta N}.$$

Fonction de répartition empirique



On a N événements $(x_i)_{1 \leq i \leq N}$ en ordre croissant :

- probabilité de non-dépassement et dépassement empirique

$$P'_i = \frac{i}{N+1} \text{ et } P_i = 1 - \frac{i}{N+1} = \frac{N+1-i}{N+1}$$

- La *fonction de répartition empirique* (de non-dépassement) :

$$P'(x) = \begin{cases} 0 \text{ si } x < x_1 \\ \dfrac{i-1}{N+1} \text{ si } x_{i-1} \leq x < x_i \\ 1 \text{ si } x > x_N \end{cases}$$

Fonction de répartition empirique (2)

Quand la loi de probabilité est connue à l'avance, on peut modifier la fonction de répartition empirique :

$$P(x) = \begin{cases} 0 \text{ si } x < x_1 \\ \dfrac{i-a}{N+b} \text{ si } x_{i-1} \leq x < x_i \\ 1 \text{ si } x > x_N \end{cases}$$

où a et b sont choisis selon la loi :

- $a = -0{,}28$ et $b = 0{,}28$ pour une loi de Gumbel
- $a = 0{,}375$ et $b = 0{,}25$ pour une loi de Gauss-Laplace.

Test : diagramme QQ

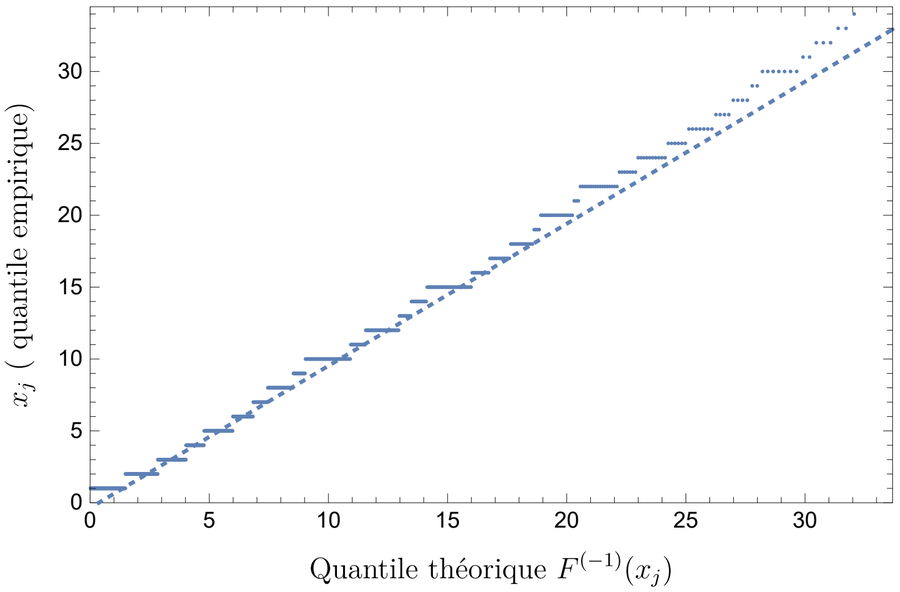


Diagramme de quantiles

On a N événements $(x_i)_{1\leq i\leq N}$ en ordre croissant :

- détermination de la fonction de répartition empirique F
- un *diagramme de quantile* le tracé des points dans un diagramme $(F^{-1}[i/(N+1)],x_i)$ pour $i = 1 \cdots N$
- Si F est un modèle raisonnable alors les points doivent se trouver alignés sur une droite diagonale (première bissectrice)

Test : diagramme PP

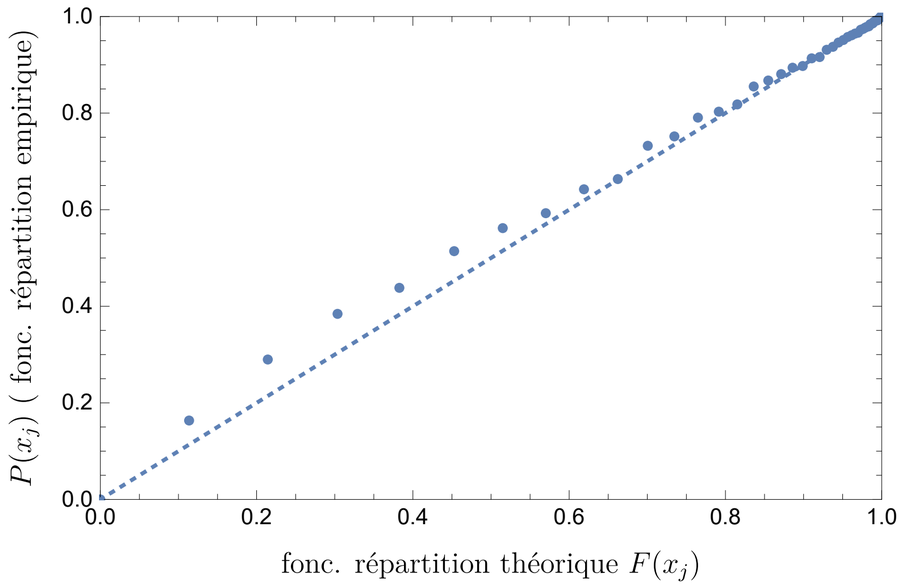


diagramme de probabilités

On a N événements $(x_i)_{1 \leq i \leq N}$ en ordre croissant :

- détermination de la fonction de répartition empirique F
- tracé des points dans un diagramme $(F(x_i),\, i/(N+1))$ pour $i = 1 \cdots N$
- si F est un modèle raisonnable alors les points doivent se trouver alignés sur une droite diagonale (première bissectrice)

Méthode d'ajustement

Soit une loi de distribution $f(x\,;\,\theta)$ où θ représente le ou les paramètre(s) à déterminer. On note p le nombre de paramètres : $p = \dim \theta$ et $[a,\, b] = \operatorname{supp} f$ le support de f (a ou b pouvant prendre des valeurs infinies). On désigne par F la fonction de répartition de cette loi. On dispose d'un jeu de n données $\boldsymbol{x} = (x_i)_{1 \leq i \leq n}$. De ce jeu, on cherche à obtenir une estimation des paramètres θ (on parle d'*inférence*) ; on note ici $\hat{\theta}$ cette estimation de θ. On va voir :

- méthode des moments
- méthode du maximum de vraisemblance
- inférence bayésienne

Alternatives : maximisation des espacements, L-moments (voir § 4.3 notes de cours)

Méthode des moments

Idée : si les moments de la loi sont finis ($< \infty$) et si on sait les calculer analytiquement :

$$M_k = \int_a^b x^k f(x) \mathrm{d}x$$

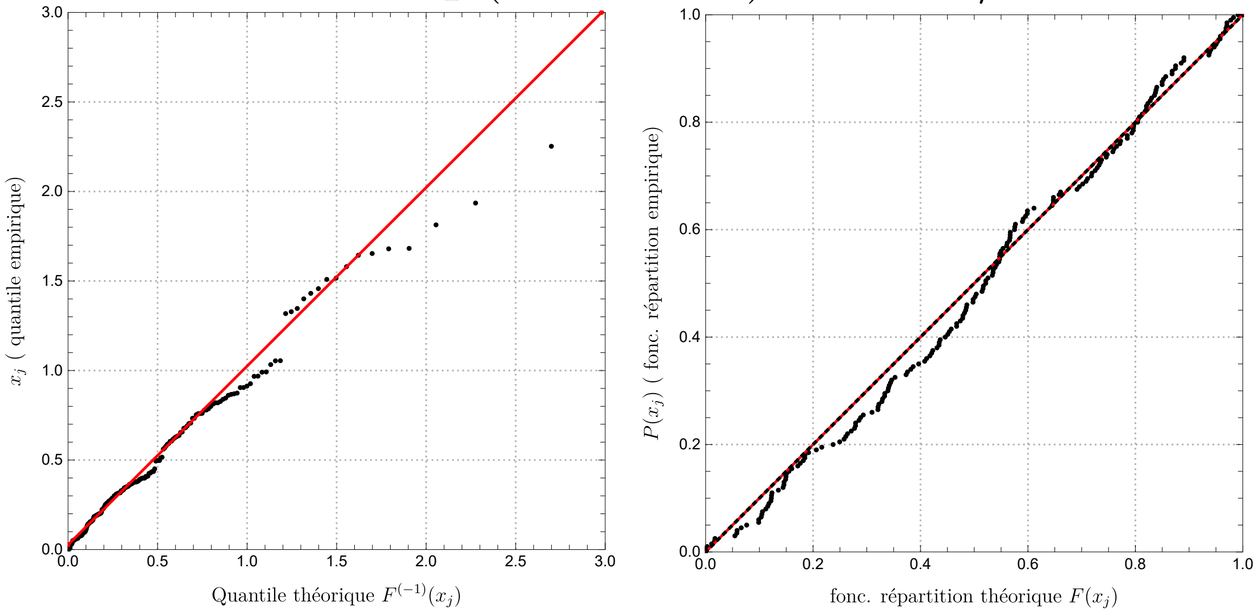
alors il suffit d'égaler les moments théoriques M_k et les moments empiriques $\hat{M}_k$ ($1 \leq k \leq p$). On aboutit alors à p équations (linéaires ou non) liant les p paramètres θ. On peut également utiliser les moments centrés :

$$m_k = \int_a^b (x - m)^k f(x) \mathrm{d}x,$$

avec $m = \mathbb{E}[f]$.

Méthode des moments : exemple 1

diagrammes PP et QQ sur un échantillon de 200 valeurs tirées $\mathrm{Exp}(x;\ \lambda = 2)$: $\hat{\lambda} = 1/\bar{x} = 2{,}09$



L'estimateur de λ est obtenu en égalant moyennes théorique et empirique $\hat{\lambda} = n/\sum x_i = 1/\bar{x}$

Considérons la loi exponentielle dont la densité de probabilité est

$$\mathrm{Exp}(x;\ \lambda) = \lambda e^{-\lambda x},$$

donc la moyenne théorique (espérance) est :

$$\mathbb{E}(X) = \int_{\mathbb{R}_+} \lambda x e^{-\lambda x} \mathrm{d}x$$

$$\mathbb{E}(X) = \left[-\frac{e^{-x\lambda}(x\lambda + 1)}{\lambda} \right]_0^{\infty} = \frac{1}{\lambda}.$$

Méthode des moments : exemple 2

Si $x \sim f(x\ ;\ \mu, \sigma, \xi)$ avec f loi de valeurs extrêmes, les trois premiers moments sont :

$$\mathbb{E}[X] = \int_{\mathbb{R}_+} x f(x, \mu, \sigma, \xi) \mathrm{d}x = \mu + \frac{\sigma}{\xi}\left(\Gamma(1-\xi) - 1\right),$$

$$\mathrm{Var}X = \int_{\mathbb{R}_+} (x - \bar{X})^2 f(x, \mu, \sigma, \xi) \mathrm{d}x = \frac{\sigma^2}{\xi^2}\left(\Gamma(1-2\xi) - \Gamma^2(1-\xi)\right),$$

$$\begin{aligned}\mathrm{Skew}X &= \frac{\int_{\mathbb{R}_+} (x - \bar{X})^3 f(x, \mu, \sigma, \xi) \mathrm{d}x}{(\mathrm{Var}X)^{3/2}} \\ &= \frac{-\Gamma(1-3\xi) + 3\Gamma(1-2\xi)\Gamma(1-\xi) - 2\Gamma^3(1-\xi)}{(\Gamma(1-2\xi) - \Gamma^2(1-\xi))^{3/2}},\end{aligned}$$

où $\Gamma(x) = \int_0^\infty t^{x-1} e^{-t} \mathrm{d}t$ la fonction gamma. Ces moments définis que pour $\xi < 1/2$.

Méthode des moments : exemple 3

On trouve que pour la loi de Gumbel, la moyenne et la variance théoriques sont données par

$$\bar{X} = \mu + \sigma\gamma \text{ et } \mathrm{Var}X = \frac{\sigma^2\pi^2}{6},$$

avec $\gamma \approx 0{,}577$ la constante d'Euler

Méthode du maximum de vraisemblance

Idée : soit $\boldsymbol{x}$ un échantillon de n valeurs x_i tirées de $f(\cdot;\ \theta)$. La probabilité d'observer $\boldsymbol{x}$ est

$$\mathrm{prob}(\boldsymbol{x}|\theta) = \prod_{k=1}^{n} f(x_k\ ;\ \theta).$$

Au lieu de regarder cette expression comme une fonction de $\boldsymbol{x}$, θ, on la définit comme une fonction $L(\theta)$ que l'on appelle la *vraisemblance* de l'échantillon $\boldsymbol{x}$:

$$L(\theta) = \prod_{k=1}^{n} f(x_k\ ;\ \theta).$$

On emploie souvent la log-vraisemblance $\ell = \ln L$:

$$\ell(\theta) = \ln L = \sum_{k=1}^{n} \ln f(x_k\ ;\ \theta)$$

Méthode du maximum de vraisemblance (2)

Le principe de maximum de vraisemblance affirme que les valeurs de θ ajustées à l'échantillon sont celles qui maximalisent la fonction $L(\theta)$. Si $\hat{\theta}$ est un maximum de L, alors on a :

$$\left.\frac{\partial L(\theta)}{\partial \theta_i}\right|_{\theta=\hat{\theta}} = 0 \text{ pour } 1 \leq i \leq p.$$

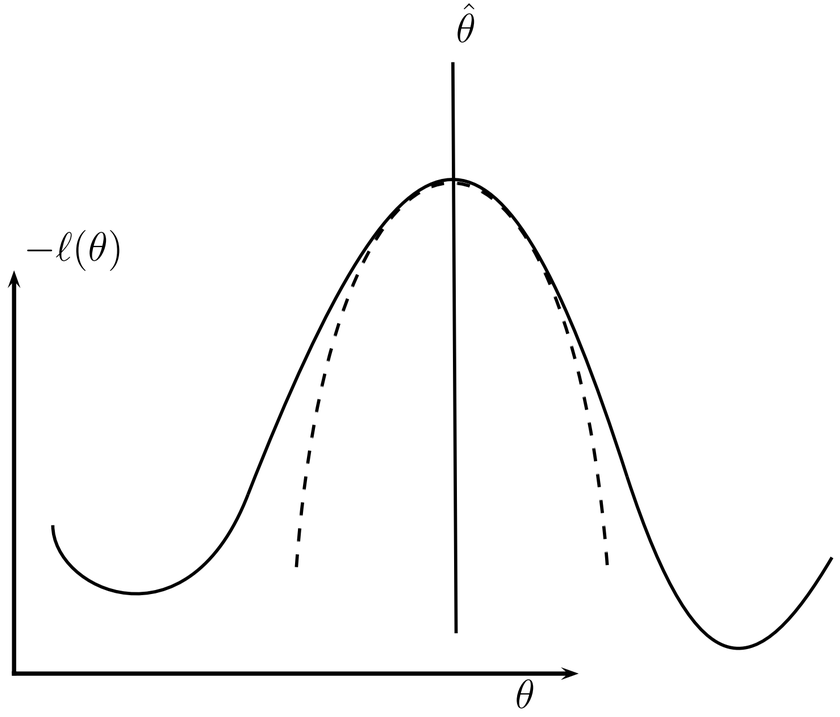
En résolvant ce système, on trouve les valeurs estimées de θ. Pour certaines lois, une solution analytique générale existe ; dans la plupart des cas, il faut procéder à une résolution numérique pour déterminer le maximum de L.

Méthode du maximum de vraisemblance (3)

Commentaires :

- l'estimateur du maximum de la vraisemblance $\hat{\theta}$ peut ne pas exister ou quand il existe, il peut ne pas être unique ;
- la vraisemblance n'est pas la densité de probabilité de θ ;
- la méthode du maximum de vraisemblance est intéressante car elle est rapide (par rapport à l'inférence bayesienne) et permet également de calculer des intervalles de confiance ;
- attention la méthode du maximum de la vraisemblance ne marche pas pour $\xi < -1$ dans le cas de la loi des valeurs extrêmes, mais ce cas ne se rencontre pas en hydrologie.

Méthode du maximum de vraisemblance (4)



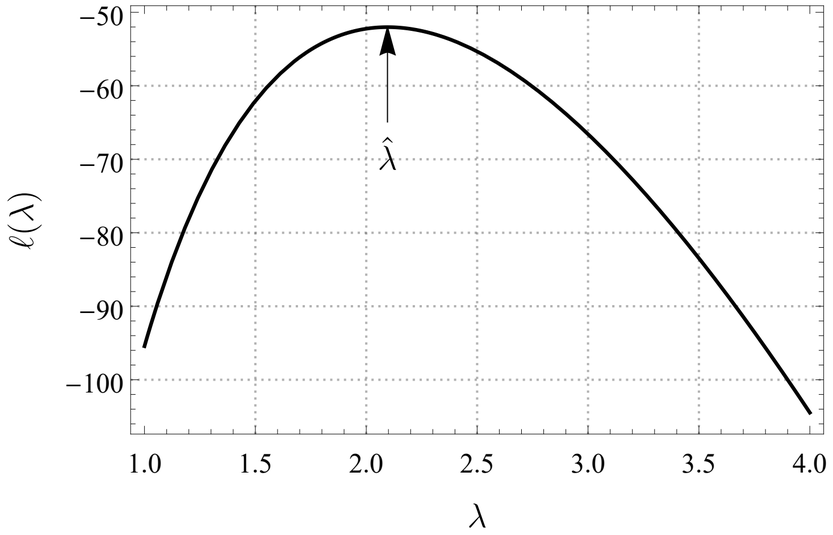
Approximation par une quadratique

Localement autour du pic de vraisemblance, la courbe a généralement une forme parabolique. Comme $\ell'(\hat{\theta}) = 0$, un développement limité à l'ordre 2 donne

$$\ell(\theta) \approx \ell(\hat{\theta}) + \frac{1}{2}\ell''(\hat{\theta})(\theta - \hat{\theta})^2,$$

- Plus il y a de données, plus le pic sera effilé (ℓ'' est plus grand), plus « certaine » sera la détermination de θ.
- La courbure $\ell''(\hat{\theta})$ influe sur la précision de l'estimation. On l'appelle l'*information observée*.
- La valeur du pic (le maximum de vraisemblance) est de moindre importance.

Méthode du maximum de vraisemblance : exemple



Approximation par une quadratique

Estimateur $\hat{\lambda}$ du paramètre λ d'une loi exponentielle

$$\mathrm{Exp}(x\ ;\ \lambda) = \lambda e^{-\lambda x},$$

La log-vraisemblance d'un échantillon $\boldsymbol{x}$ est :

$$\ell(\lambda) = n \ln \lambda - \lambda \sum x_i.$$

L'estimateur de λ est obtenu en recherchant $\ell'(\lambda) = 0$, soit

$$\hat{\lambda} = \frac{n}{\sum x_i} = \frac{1}{\bar{x}}.$$

Dans ce cas particulier, méthodes des moments et du maximum de vraisemblance donnent la même chose.

Intervalle de confiance

Soit un échantillon $\boldsymbol{x}$ de n valeurs $\boldsymbol{x} = x_1, \dots x_n$ tirées de $f(\cdot\,;\,\theta^0)$ où θ^0 est le paramètre à estimer. ℓ est la log-vraisemblance de l'échantillon et $\hat{\theta}$ est l'estimateur du maximum de vraisemblance. Alors pour n suffisamment grand, on a :

$$\sqrt{I_A(\theta^0)}(\hat{\theta} - \theta^0) \sim \mathrm{No}(0,\,1)$$

soit encore

$$\frac{\hat{\theta} - \theta^0}{1/\sqrt{I_A(\theta^0)}} \sim \mathrm{No}(0,\,1)$$

où l'on a défini *l'information attendue*

$$I_A(\theta^0) = \mathbb{E}\left(-\frac{\partial^2}{\partial\theta^2}\ell(\theta^0|\boldsymbol{x})\right),$$

Intervalle de confiance (2)

Forme équivalente

$$\hat{\theta} \sim \mathrm{No}(\theta^0,\, I_A(\theta^0)^{-1}).$$

L'estimateur $\hat{\theta}$ se comporte comme une variable aléatoire normale centrée sur θ^0, avec pour variance I_A^{-1}. L'intervalle de confiance pour θ^0 à $1-\alpha$ % est

$$1-\alpha = \mathrm{prob}\left[z_{\alpha/2} \leq \sqrt{I_A(\theta^0)}(\hat{\theta}-\theta^0) \leq z_{1-\alpha/2}\right],$$

où z_β est le β-quantile de la loi normale ($\mathrm{prob}(z_\beta)=\beta$), ou encore

$$1-\alpha = \mathrm{prob}\left[\hat{\theta} - z_{1-\alpha/2} I_A^{-1/2}(\theta^0) \leq \theta^0 \leq \hat{\theta} - z_{\alpha/2} I_A^{-1/2}(\theta^0)\right]$$

Par exemple, pour un intervalle de confiance à 95 % (soit $\alpha = 0{,}05$ et $z_{0{,}975} = 1{,}96$) :

$$\theta^0 \in [\hat{\theta} - 1{,}96 I_O^{-1/2}(\hat{\theta}),\, \hat{\theta} + 1{,}96 I_O^{-1/2}(\hat{\theta})].$$

Intervalle de confiance (3)

Généralisation : soient $x_1 \dots x_n$ des réalisations indépendantes d'une distribution $f(\cdot\,;\,\theta)$ où $\theta = (\theta_i)_{1 \leq i \leq d}$ désigne l'ensemble des d paramètres de f, ℓ la log-vraisemblance, et $\hat{\theta}$ l'estimateur du maximum de vraisemblance. Alors pour n suffisamment grand, on a :

$$\hat{\theta} \sim \mathrm{No}(\theta_0,\, \boldsymbol{I}_A(\theta^0)^{-1}),$$

où l'on introduit la *matrice d'information attendue*

$$\boldsymbol{I}_A(\theta) = \begin{bmatrix} e_{1,1} & \cdots & e_{1,d} \\ \vdots & e_{i,j} & \vdots \\ e_{d,1} & \cdots & e_{d,d} \end{bmatrix}, \text{ avec } e_{i,j} = -\mathbb{E}\left(\frac{\partial^2 \ell(\theta)}{\partial\theta_i \partial\theta_j}\right).$$

La matrice $\boldsymbol{I}_A$ mesure la courbure de la surface « log-vraisemblance ».

Intervalle de confiance (4)

Remarques :

- Par construction, intervalle de confiance symétrique par rapport à la valeur estimée $\hat{\theta}$. On peut faire mieux...
- La théorie demande de calculer la moyenne des dérivées d'ordre 2 de la log-vraisemblance, ce qui impliquerait en pratique d'avoir un grand nombre d'échantillons. En pratique donc, on substitue la matrice d'information $\boldsymbol{I}_A$ par l'information observée $\boldsymbol{J}_A$:

$$\boldsymbol{J}_A(\hat{\theta}) = \begin{bmatrix} \frac{\partial^2 \ell(\theta)}{\partial \theta_1^2} & \cdots & \frac{\partial^2 \ell(\theta)}{\partial \theta_1 \partial \theta_d} \\ \vdots & \frac{\partial^2 \ell(\theta)}{\partial \theta_i \partial \theta_j} & \vdots \\ \frac{\partial^2 \ell(\theta)}{\partial \theta_d \partial \theta_1} & \cdots & \frac{\partial^2 \ell(\theta)}{\partial \theta_d^2} \end{bmatrix}.$$

Intervalle de confiance (5)

Approximation à l'aide de la fonction déviance :

$$D(\theta_0) = 2(\ell(\hat{\theta}) - \ell(\theta_0))$$

où $\ell = \log L$ est la log-vraisemblance et θ_0 la « bonne » valeur. On montre que pour n suffisamment grand, on a :

$$D(\theta_0) \sim \chi^2_1$$

Ce théorème se généralise à des fonctions à d paramètres :

$$D(\theta_0) \sim \chi^2_d$$

où χ_d est la loi du χ^2 à d paramètres.

Intervalle de confiance (6)

Intervalle de confiance : on définit c_β le β-quantile ($\text{prob}[Z \leq c_\beta] = \beta$, avec $Z \sim \chi^2_d$), alors on a $I_\alpha = \{\theta \text{ tel que } D(\theta) \leq c_{1-2\alpha}\}$ qui est un $(1-2\alpha)$ intervalle de confiance ($\beta = 1-2\alpha$). Puisque $D \sim \chi^2_d$, la définition du quantile implique

$$\text{prob}(D \leq c_\beta) = \beta,$$

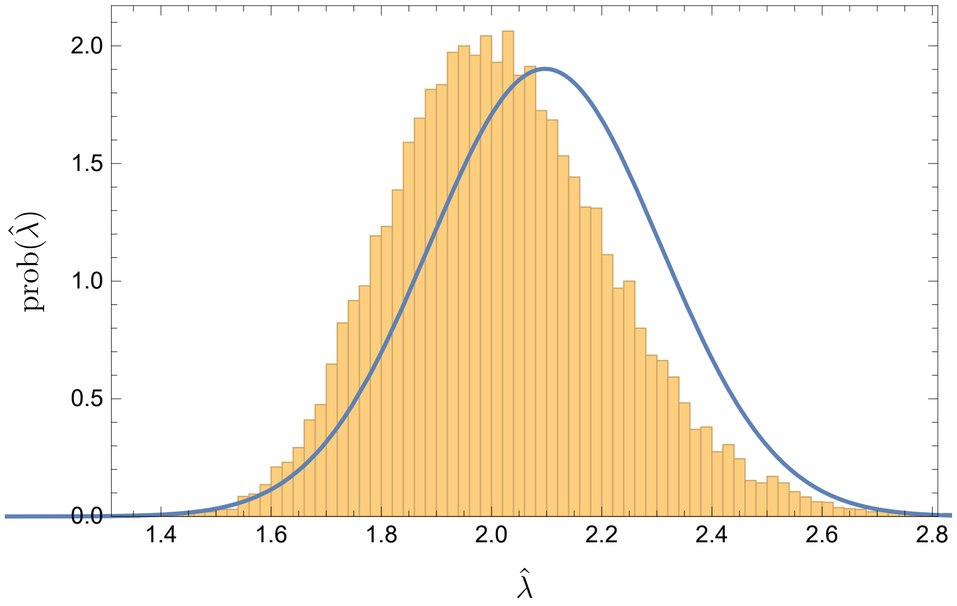
donc $D \leq c_\beta$ peut être interprété comme l'intervalle où il est probable de trouver θ_0, avec un niveau de confiance de β, ce qui veut également dire que

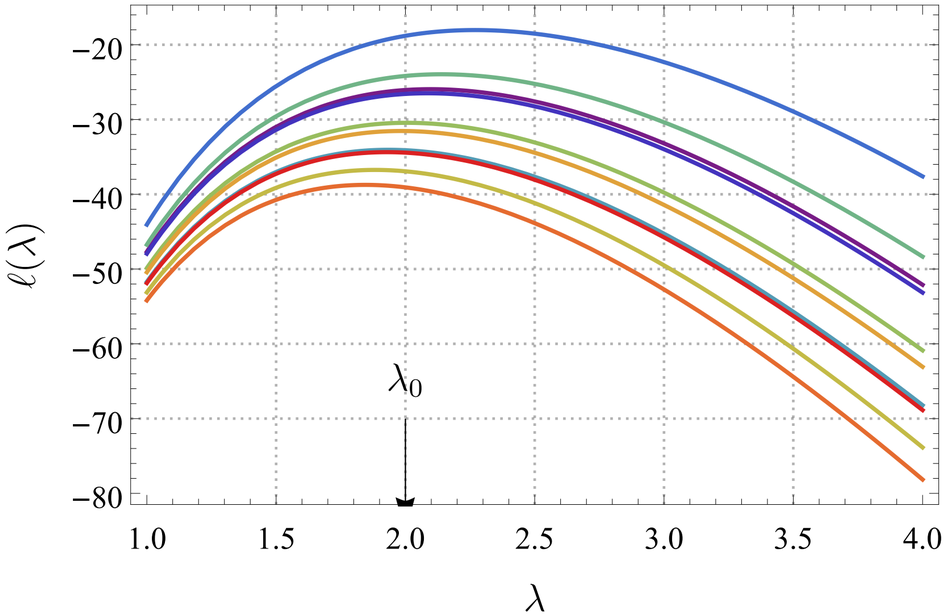
$$\theta \text{ tel que } \ell(\theta) \geq \ell(\hat{\theta}) - \frac{1}{2}c_\beta$$

est le β-intervalle de confiance pour le paramètre recherché θ_0. Pour un intervalle de confiance à 95 %, on a $\beta = 0{,}95$, soit $c_\beta = 3{,}84$. On cherche les valeurs de θ telles que $D(\theta) = 3{,}84$ (graphiquement en retranchant 3,84/2 = 1,92 à la valeur maximale).

Exemple 1

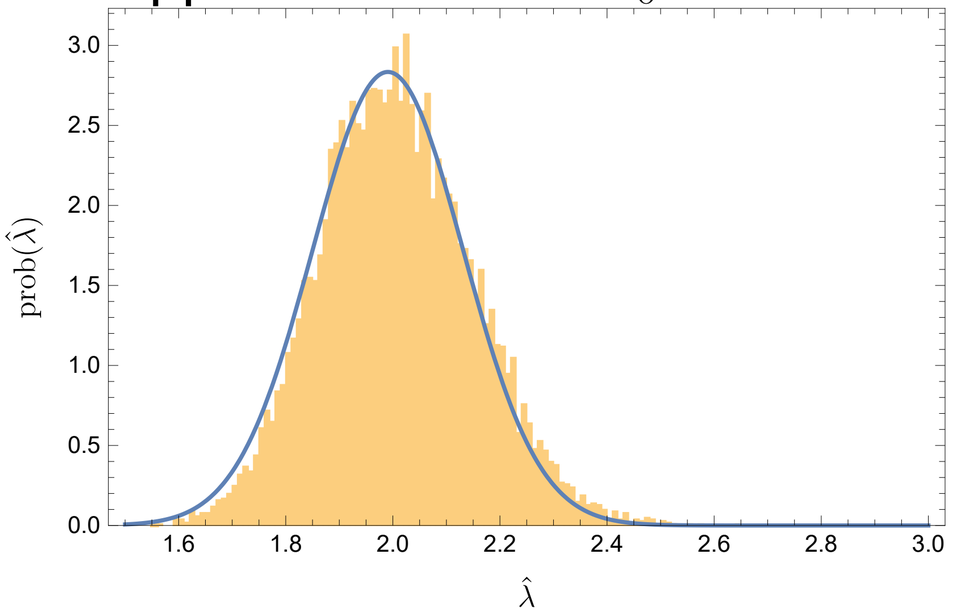
On tire 10 000 échantillons de $n = 100$ valeurs de $\mathrm{Exp}(\lambda_0)$ avec $\lambda_0 = 2$. La log-vraisemblance est $\ell(\lambda\ ;\ \boldsymbol{x}) = n \ln \lambda - \lambda \sum_{i=1}^{n} x_i$ et la courbure est $\ell''(\lambda) = -n/\lambda^2$. Si on prend, p. ex., le premier échantillon on a : $\hat{\lambda}_1 = 1/0{,}47 = 2{,}09$ et $I_O = -\ell''(\hat{\lambda}_1) = n/\lambda_1^2 = 22{,}7$. La loi normale de moyenne λ_1 et variance I_O^{-1} fournit une approximation de λ_0

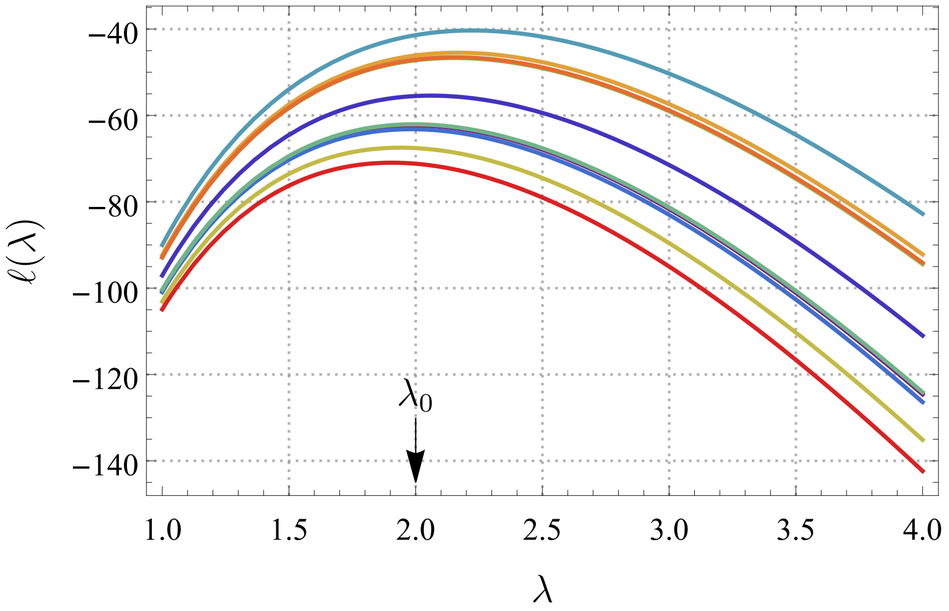




Exemple 1 (2)

On tire 10 000 échantillons avec maintenant $n = 200$ valeurs de $\mathrm{Exp}(\lambda_0)$ avec $\lambda_0 = 2$. Si on prend le premier échantillon on a : $\hat{\lambda}_1 = 1/0{,}502 = 1{,}99$ et $I_O = -\ell''(\hat{\lambda}_1) = n/\lambda_1^2 = 50{,}46$. La loi normale de moyenne λ_1 et variance I_O^{-1} fournit une approximation de λ_0



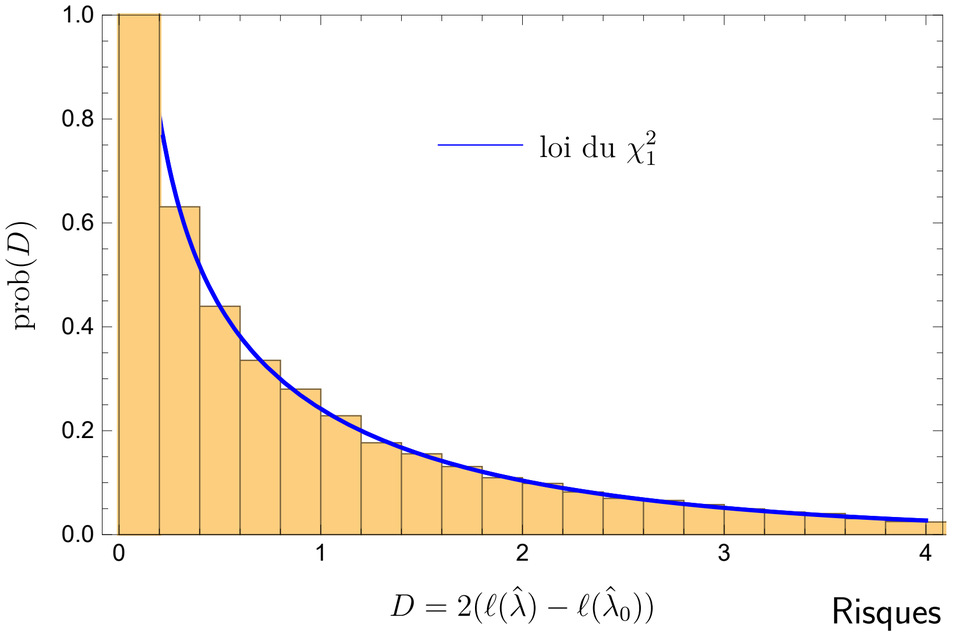


Exemple 1 (3)

Pour la loi exponentielle, la déviance d'un échantillon est

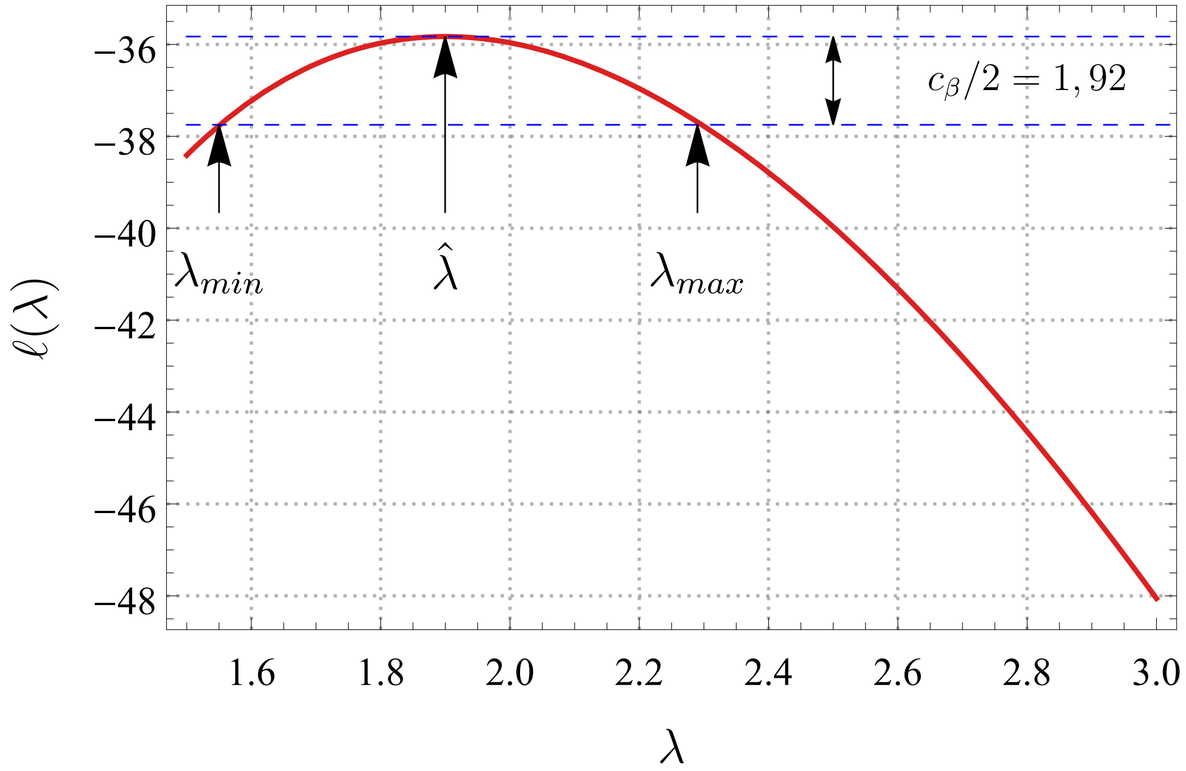
$$D(\lambda_0) = 2(\ell(\hat{\lambda}) - \ell(\lambda_0)) = 2n\left(\ln\frac{\lambda_0}{\bar{x}} - 1 + \lambda_0\bar{x}\right),$$

On vérifie en simulant 10 000 échantillons de 100 valeurs que D est bien distribué selon χ^2_1



Exemple 1 (4)

Avec la loi du χ^2_1, il est simple de trouver l'intervalle de confiance : il suffit de retrancher 1,92 au pic de ℓ (pour l'intervalle de confiance à 95 %).



Exemple 1 (5)

Résumé des intervalles de confiance à 95 % trouvés pour un échantillon de $n = 100$ valeurs :

- approximation de ℓ par une loi normale :
  - $\lambda_{min} = \text{quantile}(\text{No}(\hat{\lambda}_1,\ \hat{\lambda}_1/\sqrt{n}))(0{,}025) = 1{,}52$
  - $\lambda_{max} = \text{quantile}(\text{No}(\hat{\lambda}_1,\ \hat{\lambda}_1/\sqrt{n}))(0{,}975) = 2{,}27$
- approximation de la déviance par une loi du χ^2_1 :
  - $D(\hat{\lambda}_{min}) = 3{,}84 \Rightarrow \hat{\lambda}_{min} = 1{,}55$
  - $D(\hat{\lambda}_{max}) = 3{,}84 \Rightarrow \hat{\lambda}_{max} = 2{,}29$

Inférence bayésienne

Idée : le théorème de Bayes dit que si on a des observations $\boldsymbol{x} = (x_i)_{1 \leq i \leq n}$ et qu'on veut déterminer les paramètres θ d'une loi de probabilité f, alors

$$\pi(\theta|\boldsymbol{x}) = \frac{f(\boldsymbol{x}|\theta)\pi(\theta)}{\int \mathrm{d}\theta\, f(\boldsymbol{x}|\theta)\pi(\theta)}.$$

Interprétation : un apport d'information $\boldsymbol{x}$ modifie notre connaissance *a priori* de la valeur possible de θ, qui est encodée via la probabilité dite *prior* $\pi(\theta)$. La distribution *a posteriori* de θ connaissant l'information $\boldsymbol{x}$ est notée $\pi(\theta|\boldsymbol{x})$; elle est proportionnelle au prior $\pi(\theta)$ et à la fonction $f(\boldsymbol{x}|\theta)$, qui est la *vraisemblance* $L(\theta) = f(\boldsymbol{x}\,;\,\theta) = \prod f(x_i\,;\,\theta)$. En résumé on a :

$$\pi(\theta|\boldsymbol{x}) \propto L(\theta)\pi(\theta)$$

Le dénominateur (compliqué) ne va pas nous embêter...

Inférence bayésienne (2)

Supposons qu'on étudie un phénomène aléatoire et qu'on sache que le phénomène soit correctement traduit sous forme d'une probabilité f de θ :

- Si on n'a pas d'idée sur la valeur de ce paramètre, on dit qu'on n'a pas d'information a priori : θ peut prendre n'importe quelle valeur, p. ex. sur un intervalle $[a, b]$. Sa probabilité est donc (loi uniforme) :

$$\pi(\theta) = \frac{1}{b-a}$$

- Au contraire, on peut avoir une idée assez précise de la valeur. Un expert dira, par exemple, que θ doit proche de la valeur θ_0 un certain pourcentage ε près. On pourra alors employer (p. ex.) la loi normale

$$\pi(\theta) = \mathrm{No}(\theta, \varepsilon^2)$$

Inférence bayésienne (3)

Dans les deux cas, le prior π encode notre connaissance (notre méconnaissance) de θ. C'est une connaissance subjective.

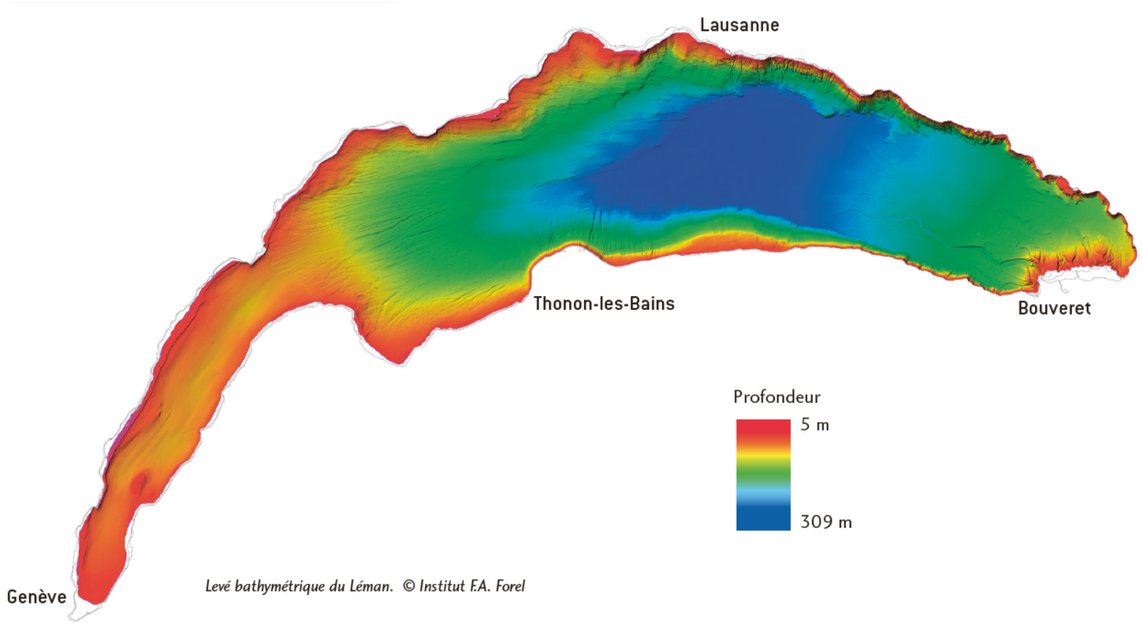
Supposons maintenant que l'on ait de nouvelles données $\boldsymbol{x} = (x_i)_{1 \leq i \leq n}$. Comment ces données modifient-elles notre connaissance de θ ? D'après le théorème de Bayes, la connaissance *a posteriori* (posterior) est

$$\pi(\theta|\boldsymbol{x}) = \frac{L(\theta)\pi(\theta)}{Z} \text{ avec } Z = \int \mathrm{d}\theta f(\boldsymbol{x}|\theta)\pi(\theta) \text{ et } L = f(\boldsymbol{x}|\theta)$$

Formule simple, mais que généralement on ne sait pas calculer. Il faut donc simuler...

L'idée est de simuler un échantillon $\boldsymbol{\theta}$ de valeurs à partir du posterior $\pi(\theta|\boldsymbol{x})$, et tracer l'histogramme pour voir comment évolue notre connaissance.

Inférence bayésienne (4)



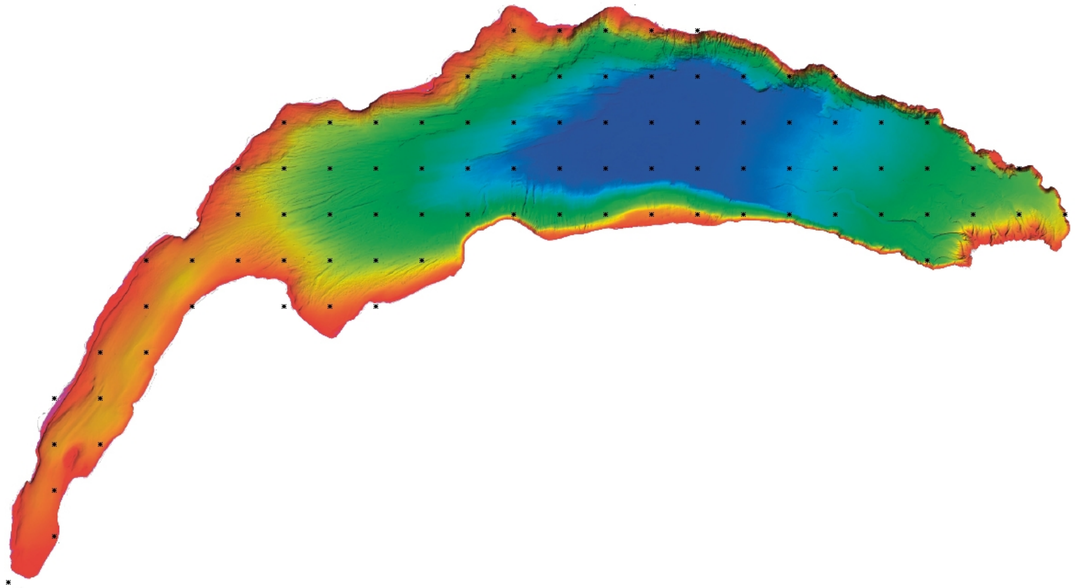
Tirer un échantillon de valeurs aléatoires d'une loi de probabilité : exercice difficile.

Analogie : comment trouver le volume du lac Léman ?

Deux stratégies :

- approche déterministe
- approche stochastique

Inférence bayésienne (5)



Grille à pas fixe

Approche déterministe : on se fixe une grille et on va de point en point.

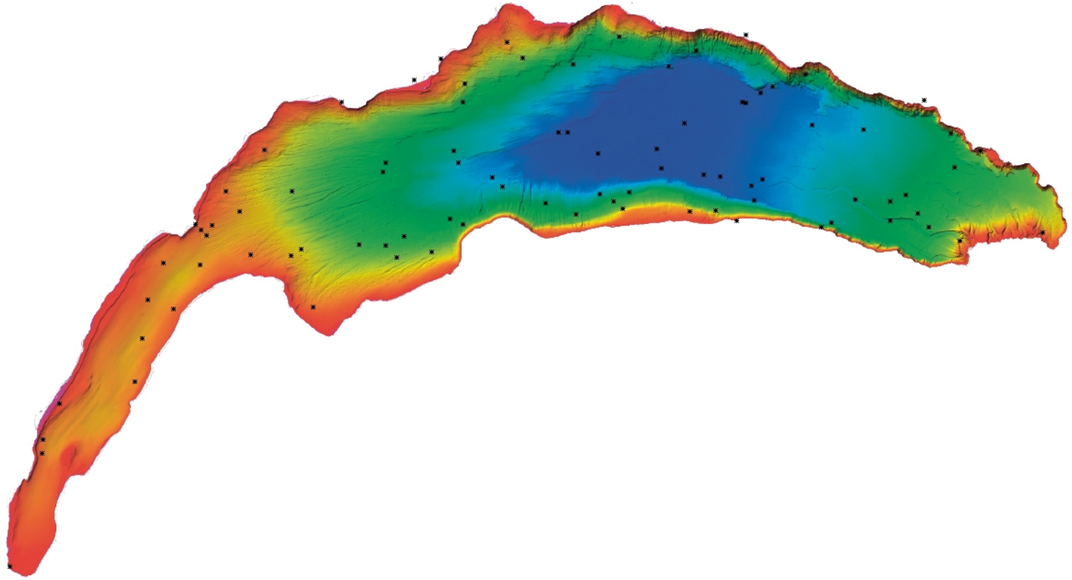
Avantages :

- nombre d'opération fixé à l'avance
- simple à mettre en œuvre

Inconvénients :

- des mesures apportent peu de chose
- coût exploration prohibitif si précision requise

Inférence bayésienne (6)



Grille aléatoire très lâche

Approche stochastique : on explore aléatoirement la surface du lac.

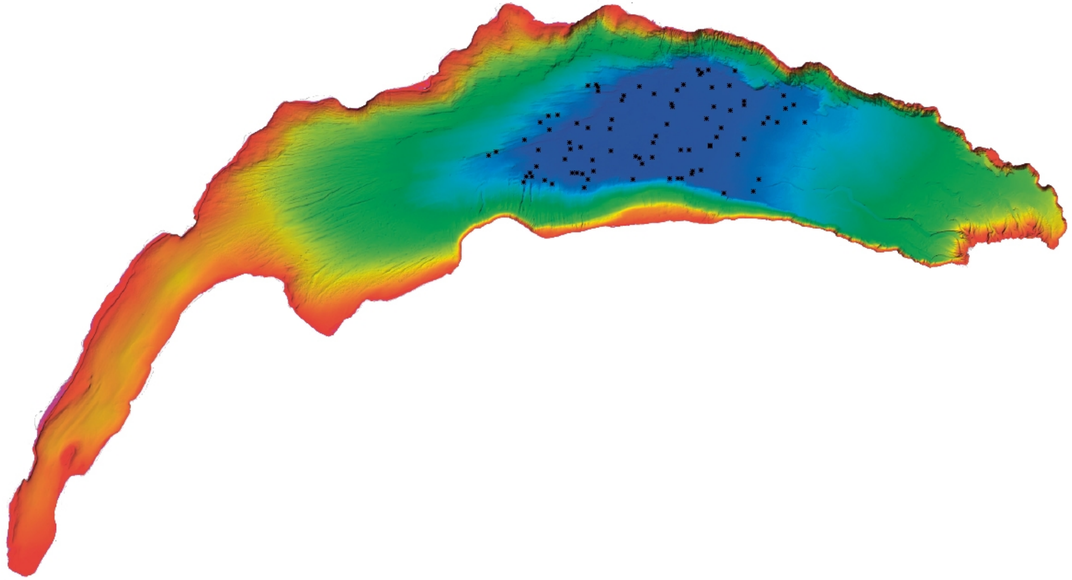
Avantages :

- coût d'exploration que l'on peut optimiser
- précision qui peut être très bonne avec un nombre limité

Inconvénients :

- nombre d'opération non fixé à l'avance
- pas si simple à mettre en œuvre

Inférence bayésienne (7)



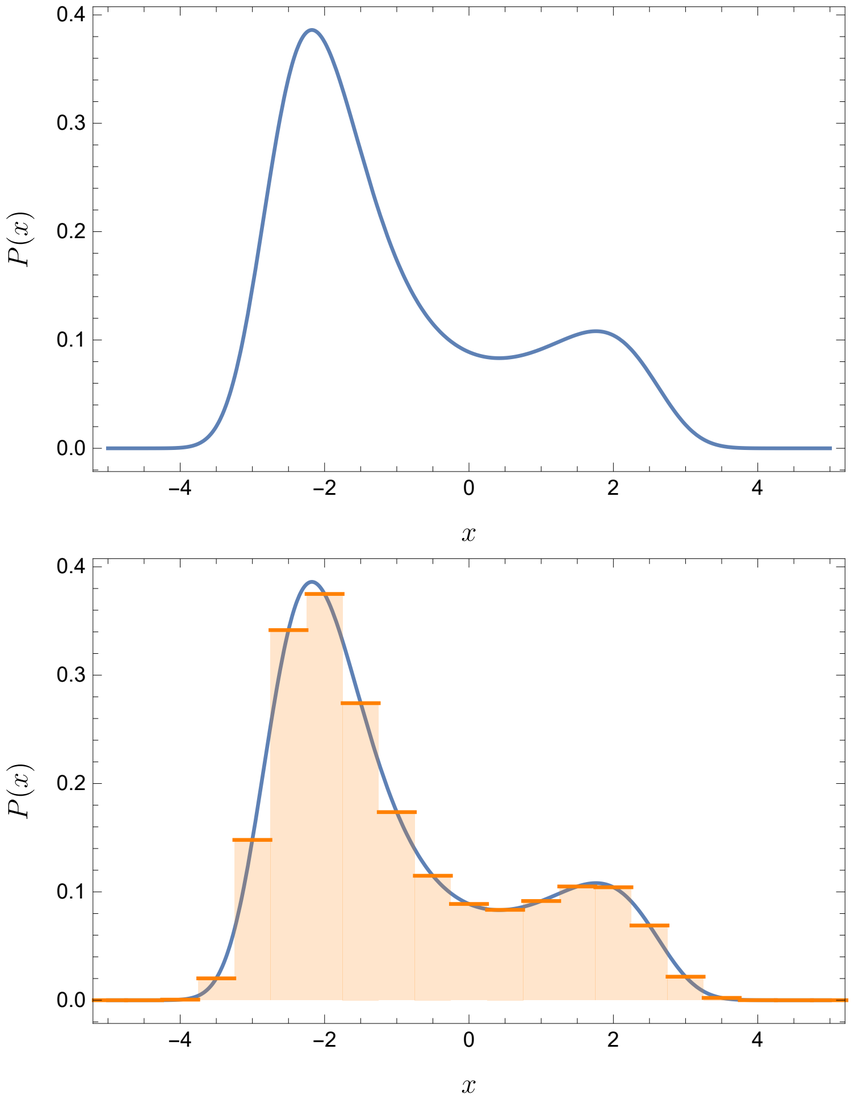
Grille aléatoire très concentrée sur les grandes profondeurs

Approche stochastique : optimisation de l'exploration. On peut régler la taille des sauts...

On a intérêt :

- rester là où il y a de la profondeur
- ne pas oublier de sauter de temps à autre vers une zone a priori peu intéressante... minimum local ou global ?

Méthode de Monte Carlo

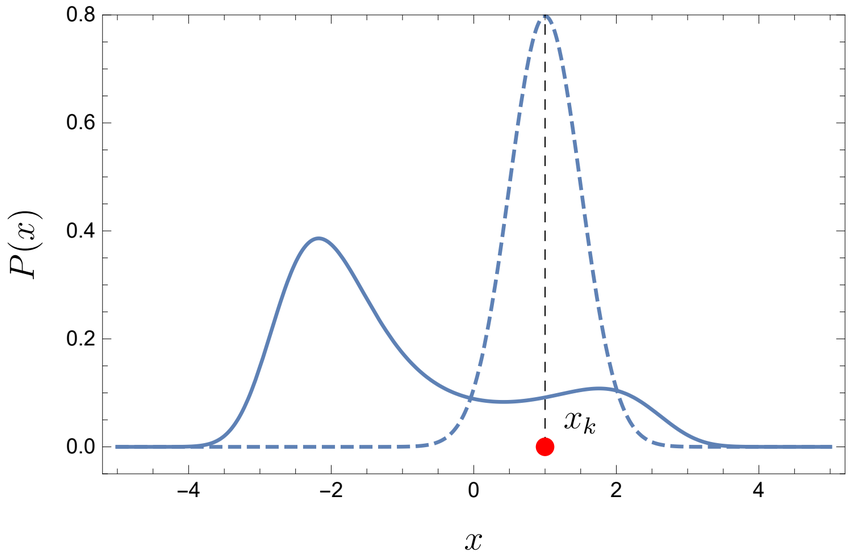


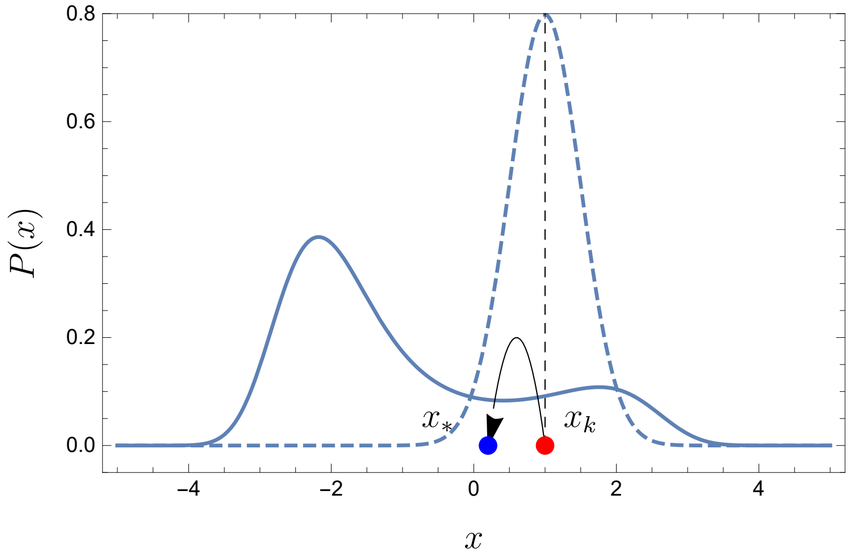
But : simuler une loi de probabilité $P(x)$

Idée : on simule un échantillon de N valeurs x_i. L'histogramme des valeurs de x_i permet d'approcher P

Technique : on ne sait pas tirer un échantillon de valeurs depuis P, mais on le sait faire avec une loi de probabilité simple... on va donc employer une telle loi dite *loi instrumentale* Q, et accepter/rejeter les valeurs. Une loi normale $Q = \mathrm{No}(x_i, \sigma^2)$ est intéressante (σ taille typique des sauts)

Algorithme de Metropolis-Hastings





Supposons qu'on ait déjà simulé k valeurs. On va tirer une valeur x_* depuis Q.

Doit-on accepter cette valeur ou non ? Pour cela on calcule le rapport

$$r = \frac{P(x_*)Q(x_k\ ;\ x_*,\, \sigma^2)}{P(x_k)Q(x_*\ ;\ x_k,\, \sigma^2)}$$

Note : si on prend une loi instrumentale symétrique, alors $Q(x_k\ ;\ x_*,\, \sigma^2) = Q(x_*\ ;\ x_k,\, \sigma^2)$ et donc alors

$$r = P(x_*)/P(x_k)$$

Avantage : pas de calcul du dénominateur Z

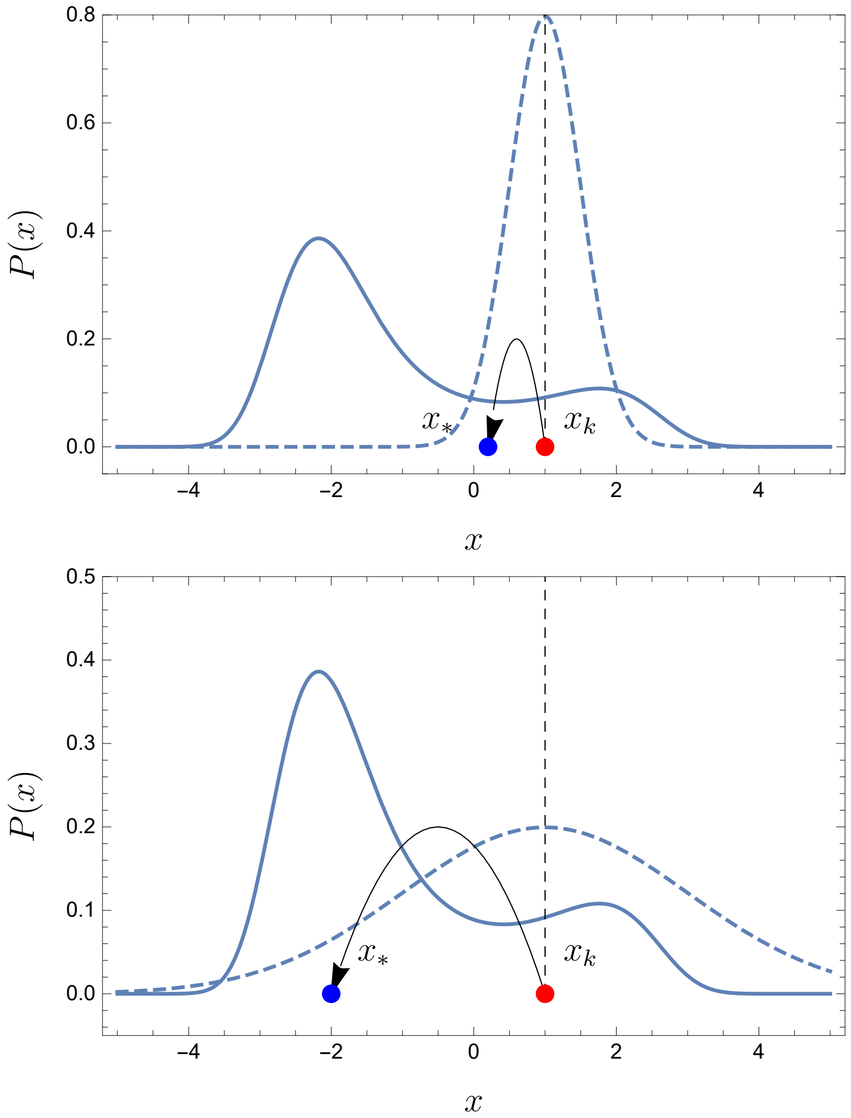
Algorithme de Metropolis-Hastings (2)

- On accepte la valeur x_* si $r \geq 1$.
- Sinon, si $r < 1$, on l'accepte avec une probabilité r. En pratique, cela veut dire que l'on tire un nombre aléatoire u de la loi uniforme sur [0, 1] :
  - si $r \geq u$ on accepte x_* et donc on pose $x_{k+1} = x_*$
  - si $r < u$ on n'accepte pas la valeur et on pose $x_{k+1} = x_k$

On répète la procédure autant de fois que nécessaire pour obtenir un échantillon de taille suffisante.

Questions : comment fixer la loi instrumentale ? Comment fixer la taille des sauts ?

Algorithme de Metropolis-Hastings (3)



Dilemme :

- Si on fait de petits sauts, on explore uniquement les maxima locaux et on risque de louper le maximum global. Il faut tolérer des sauts de plus grande taille.
- Si on fait de grands sauts, on explore tout le domaine, mais on passe beaucoup de temps à l'explorer sans s'attarder là où il y a des choses intéressantes (le pic de probabilité). Il faudrait opter pour de petits sauts:

Règle empirique : l'algorithme est bien réglé quand le taux moyen d'acceptation est compris entre 0,25 et 0,50.

Algorithme de Metropolis-Hastings (4)

Un algorithme simple à coder avec Mathematica, Matlab...

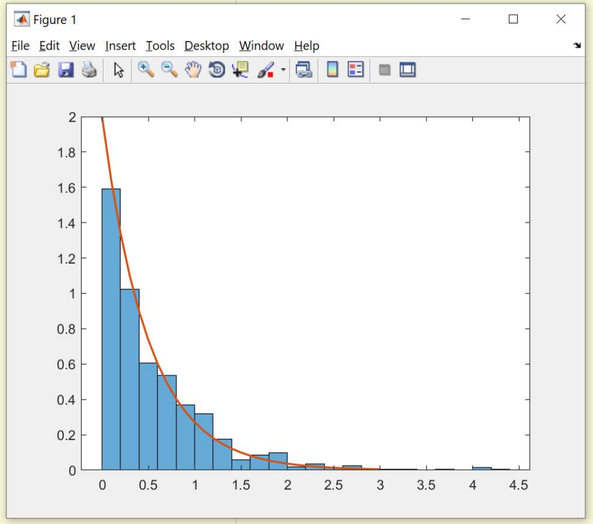
```
In[353]:= x0 = 1
          Q[x_, s_] := NormalDistribution[x, s]
          f[x_] := PDF[ExponentialDistribution[2], x]
          sample = {}
          taux = {}
          Do[
           xs = RandomReal[Q[x0, 1]];
           r = f[xs] / f[x0];
           If[r ≥ 1, x0 = xs; AppendTo[taux, 1]];
           If[r < 1,
            u = RandomReal[];
            If[r > u, x0 = xs; AppendTo[taux, 1],
             AppendTo[taux, 0]
            ]
           ];
           AppendTo[sample, x0],
           {i, 1000}]
```

Editor - D:\hastings.m

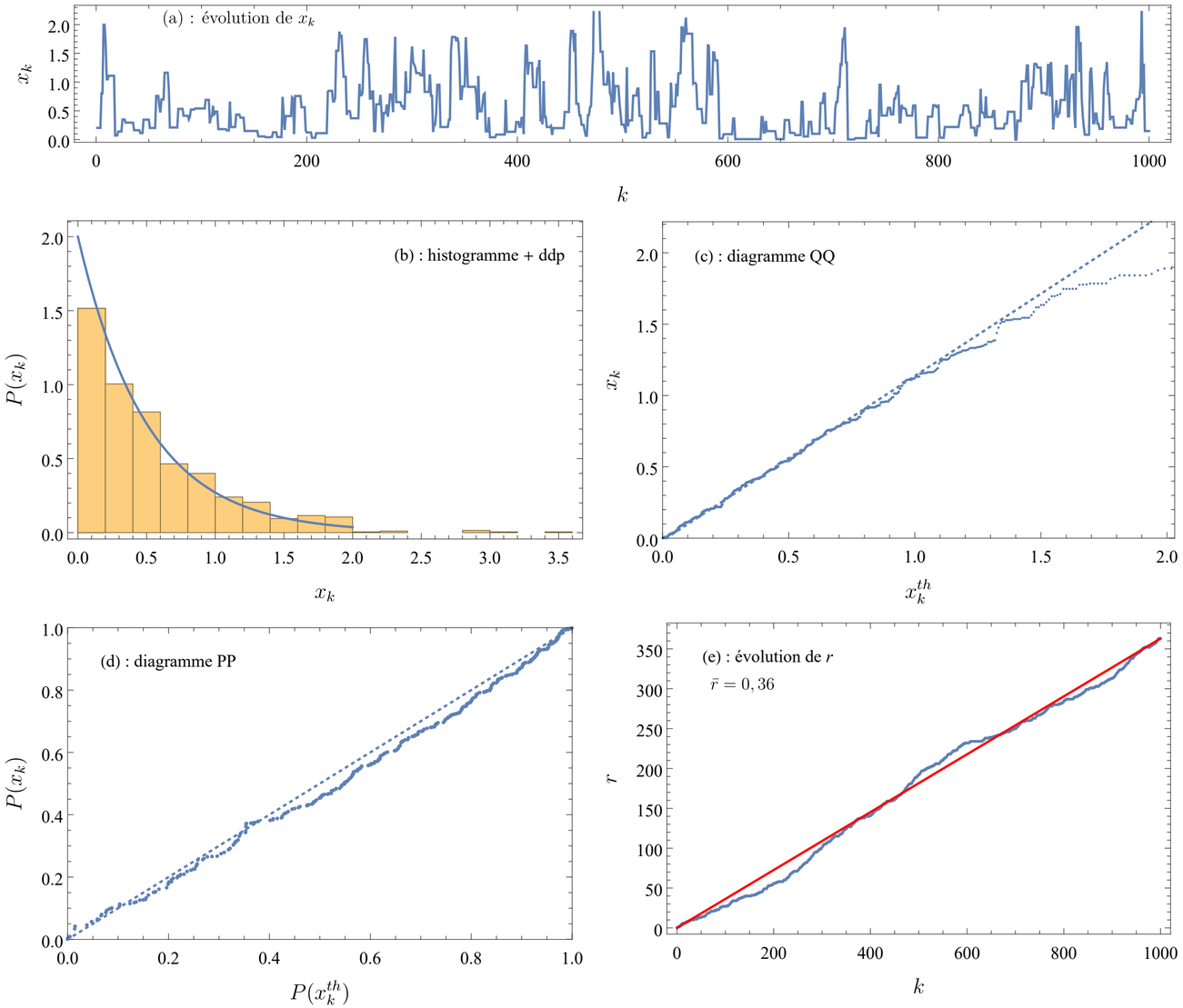
benchmarkCyclic.m | hastings.m

This file can be opened as a Live Script. For more information, see Creating Live Scripts.

```
clear all;
close all;
x0 = 1;
sample = [];
taux = [];
step = 1;


for i=1:1000
    xs  =randn*step+x0;
    r = expdist(xs)/expdist(x0);
    if r>1
        x0 = xs;
        taux = [taux, 1];
    end
    if r<= 1
        u = rand;
        if r>u
            x0 = xs;
            taux = [taux, 1];
        elseif r<=u
            taux = [taux, 0];
        end
    end
    sample = [sample,x0];
end


histogram(sample,'Normalization','pdf')
hold on
x = 0:0.1:3;
lambda = 2
f = exp(-x*lambda)*lambda;
plot(x,f,'LineWidth',1.5)

%% exponential distribution
function e = expdist(x)
lambda = 2 ;
if x<0
    e = 0;
elseif x>0
    e = exp(-x*lambda)*lambda;
end
end
```



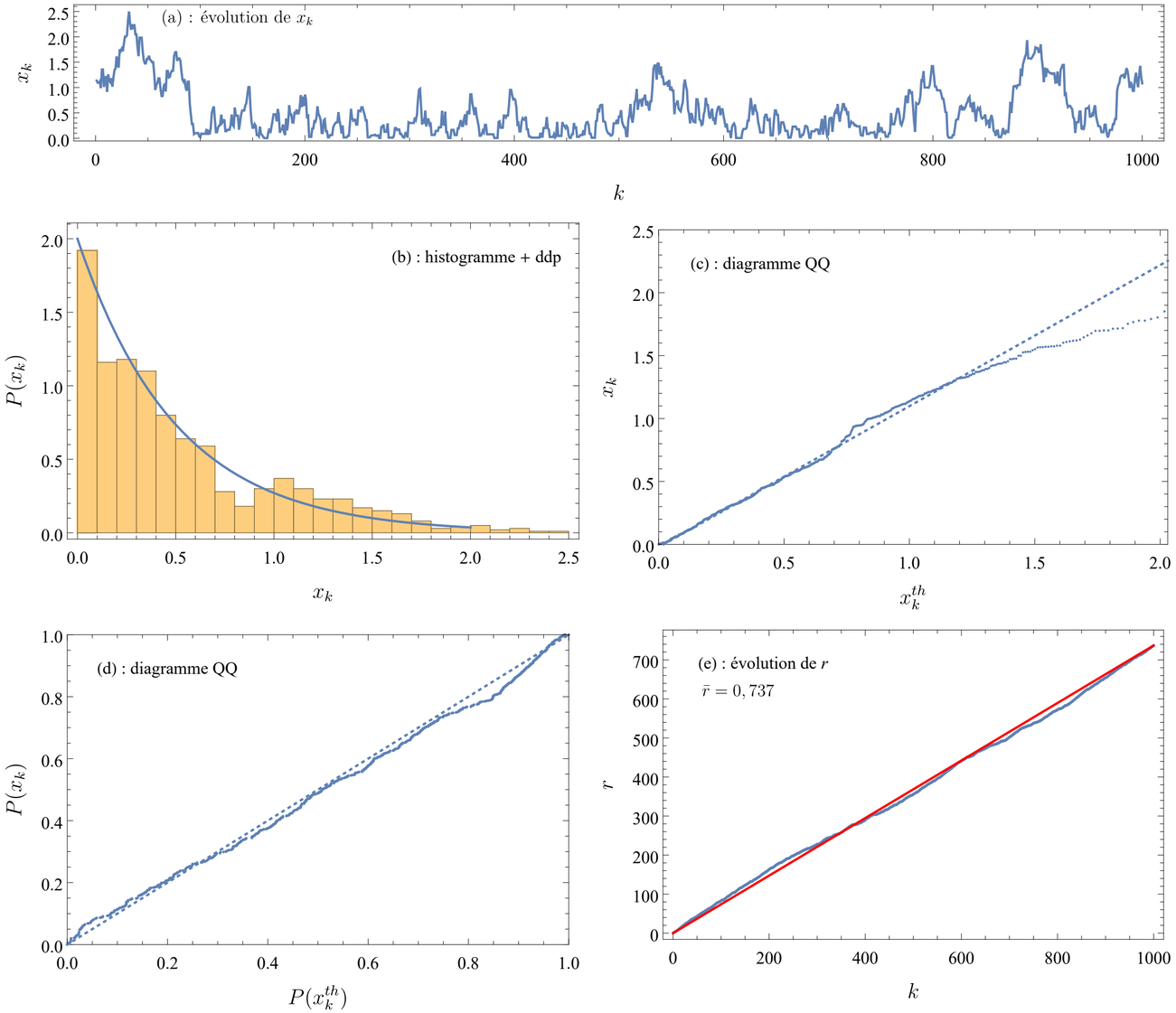
Exemple 1: simuler la loi exponentielle



- Simulation de la loi exponentielle avec $\lambda = 2$
- 1000 valeurs simulées
- loi instrumentale : loi normale avec $\sigma = 1$

Un taux d'acception moyen de 0,36. Diagrammes PP et QQ bons

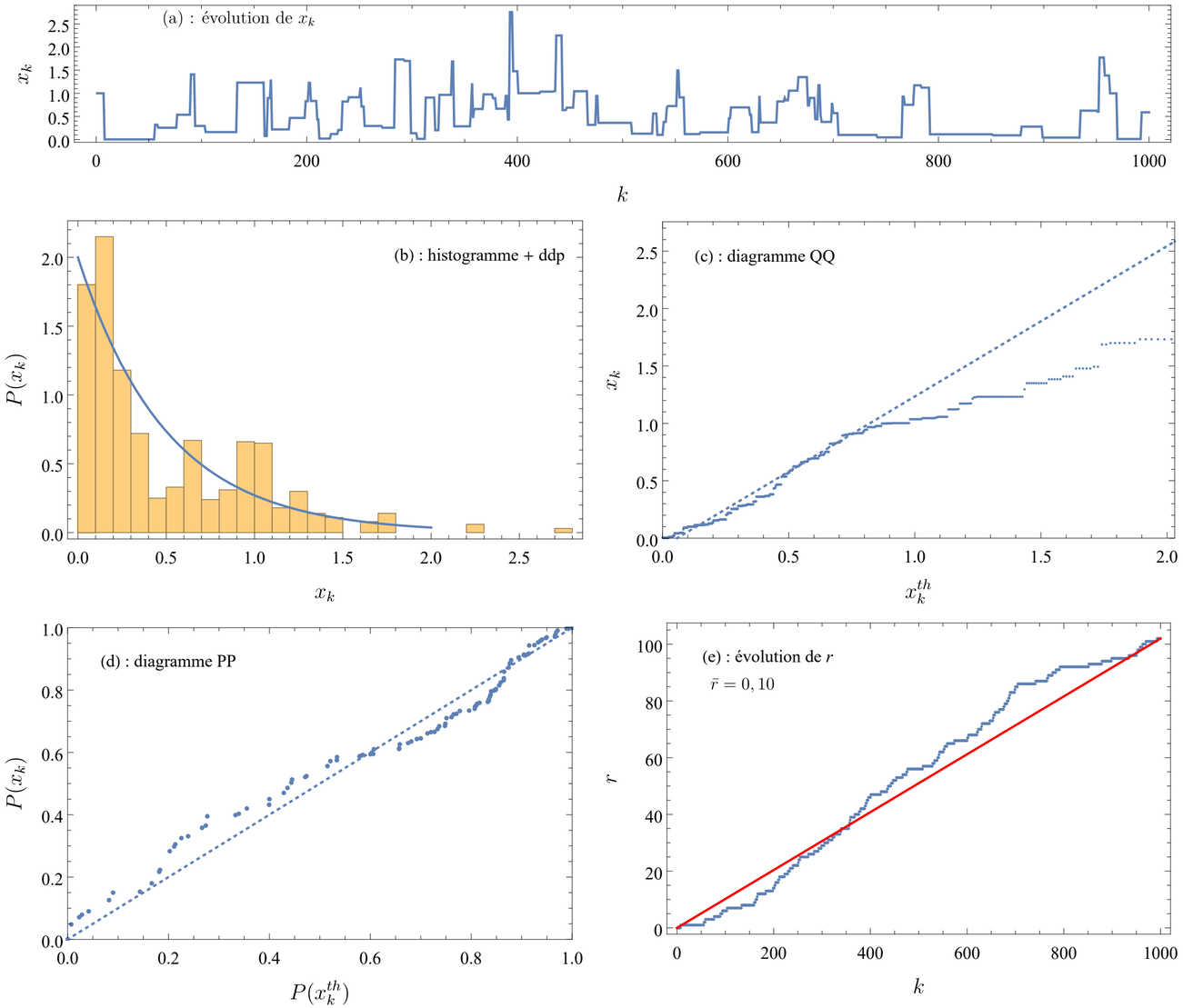
Exemple 1: simuler la loi exponentielle (2)



- Simulation de la loi exponentielle avec $\lambda = 2$
- 1000 valeurs simulées
- loi instrumentale : loi normale avec $\sigma = 0{,}2$ (petit pas)

Un taux d'acception moyen de 0,737 (trop fort). Diagrammes PP et QQ bons

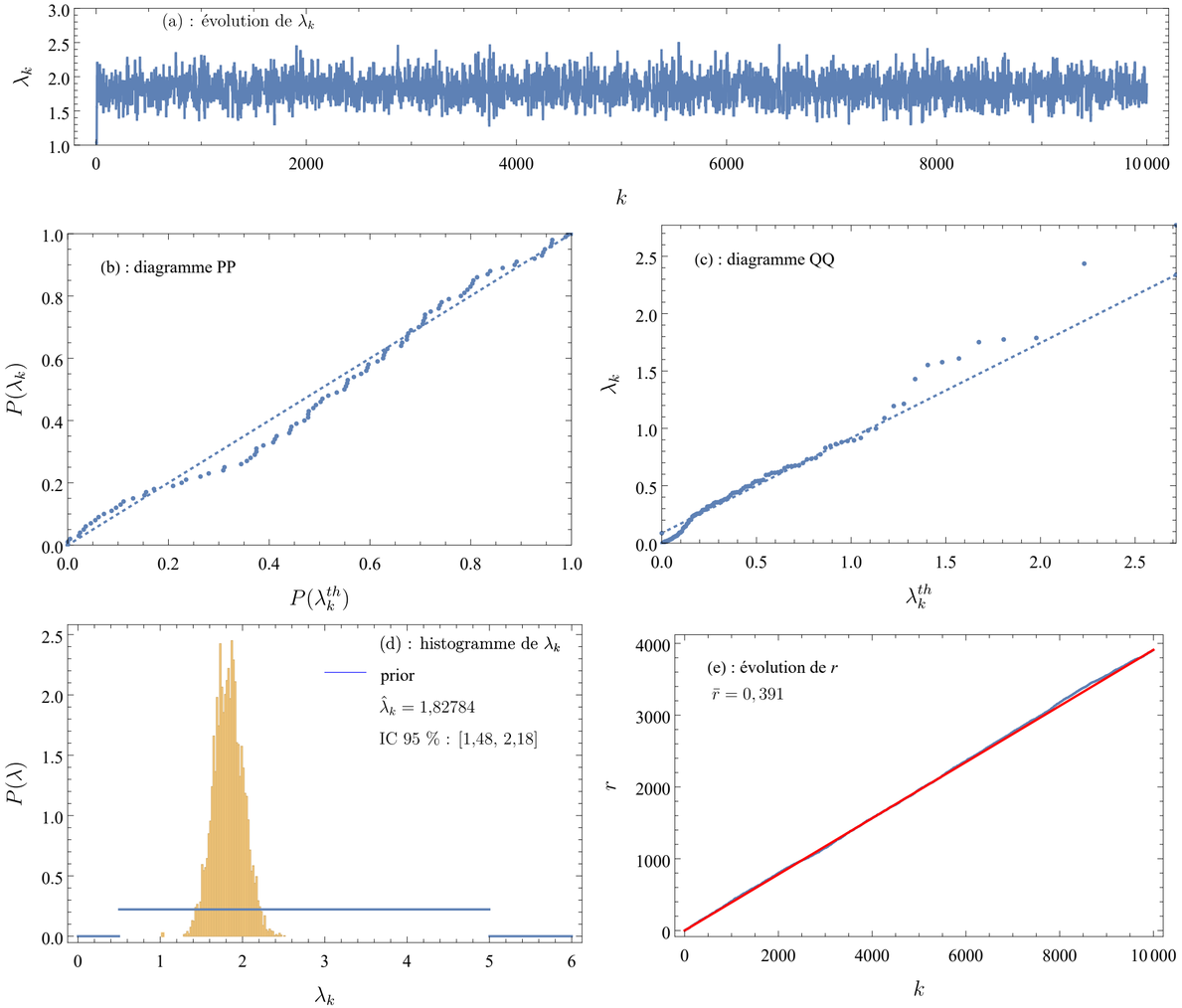
Exemple 1: simuler la loi exponentielle (3)



- Simulation de la loi exponentielle avec $\lambda = 2$
- 1000 valeurs simulées
- loi instrumentale : loi normale avec $\sigma = 4$ (grand pas)

Un taux d'acception moyen de 0,10. Diagrammes PP et QQ pas bons

Exemple 2 : caler un paramètre

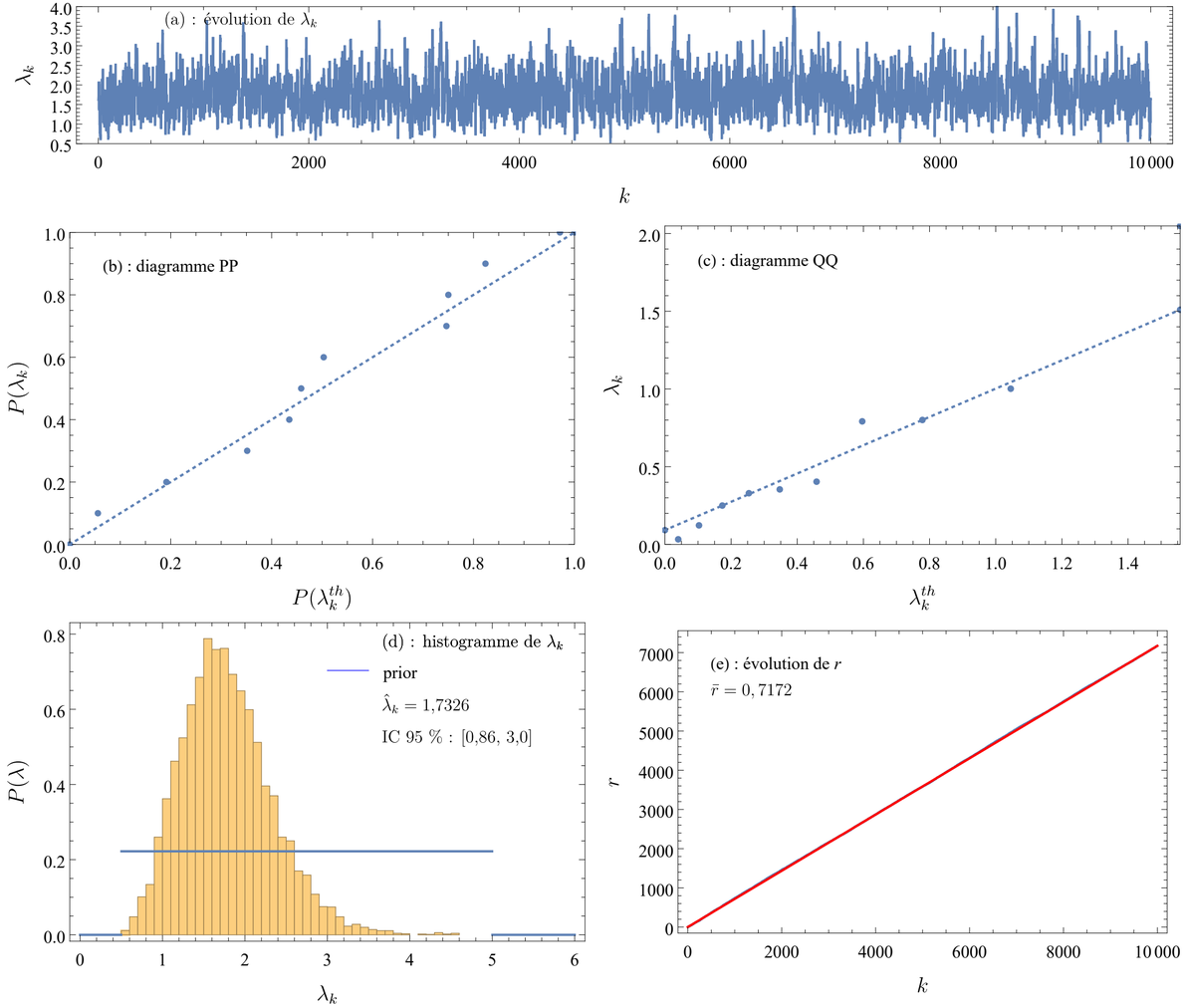


Supposons

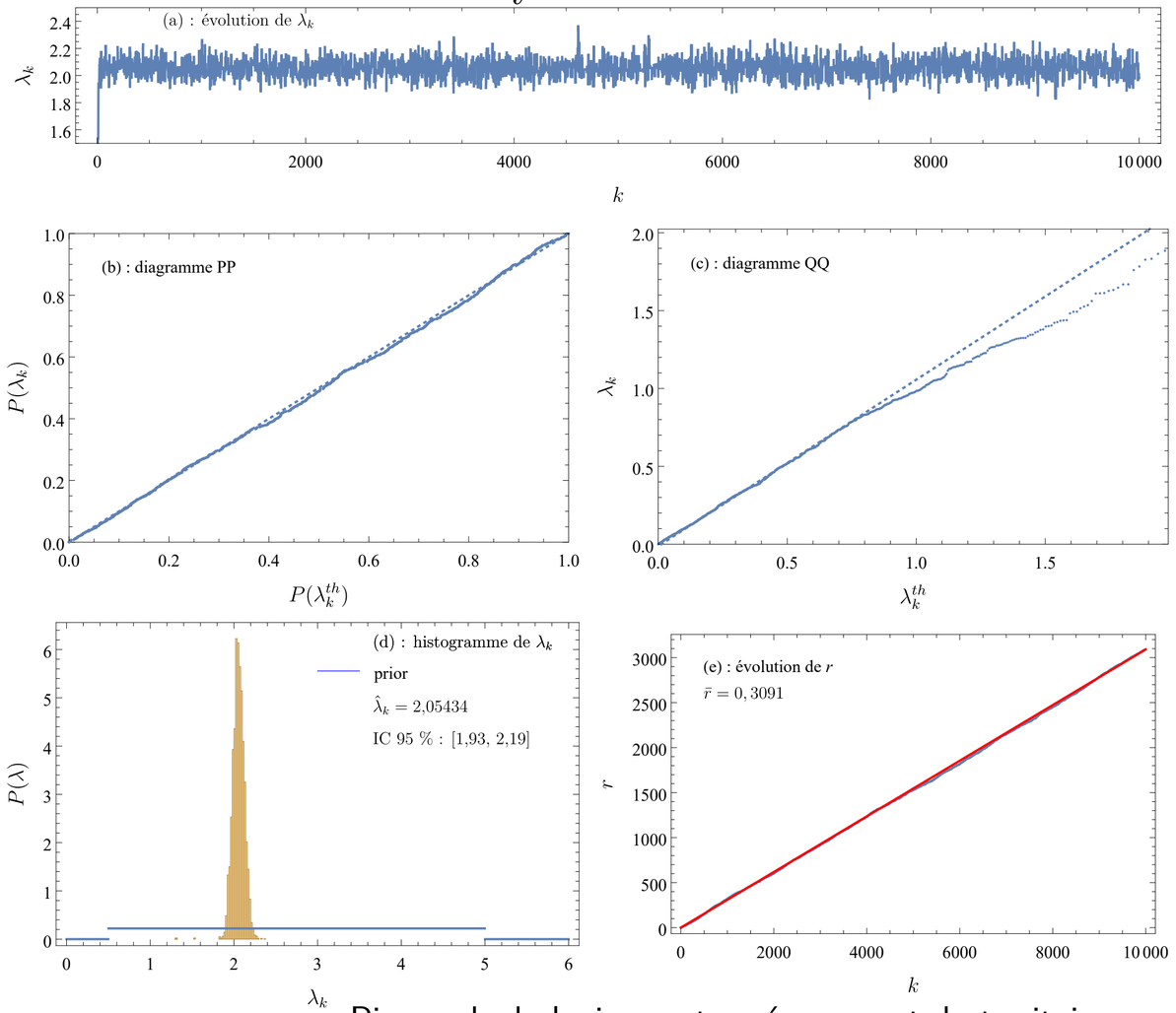
- on ait un échantillon de $n = 100$ valeurs x_i tirées de la loi exponentielle avec $\lambda = 2$. Que vaut $\hat{\lambda}$?
- 10 000 valeurs simulées λ_k. Loi instrumentale : loi normale avec $\sigma = 0{,}5$. Prior : loi uniforme sur $[1/2,\ 5]$
- On définit $\hat{\lambda} = \text{Quantile}(0{,}5) = 1{,}74$.
- Intervalles de confiance à 95 % : [1,44, 2,10]

Exemple 2 (2) : influence du nombre de données n

$n = 10$ valeurs x_i

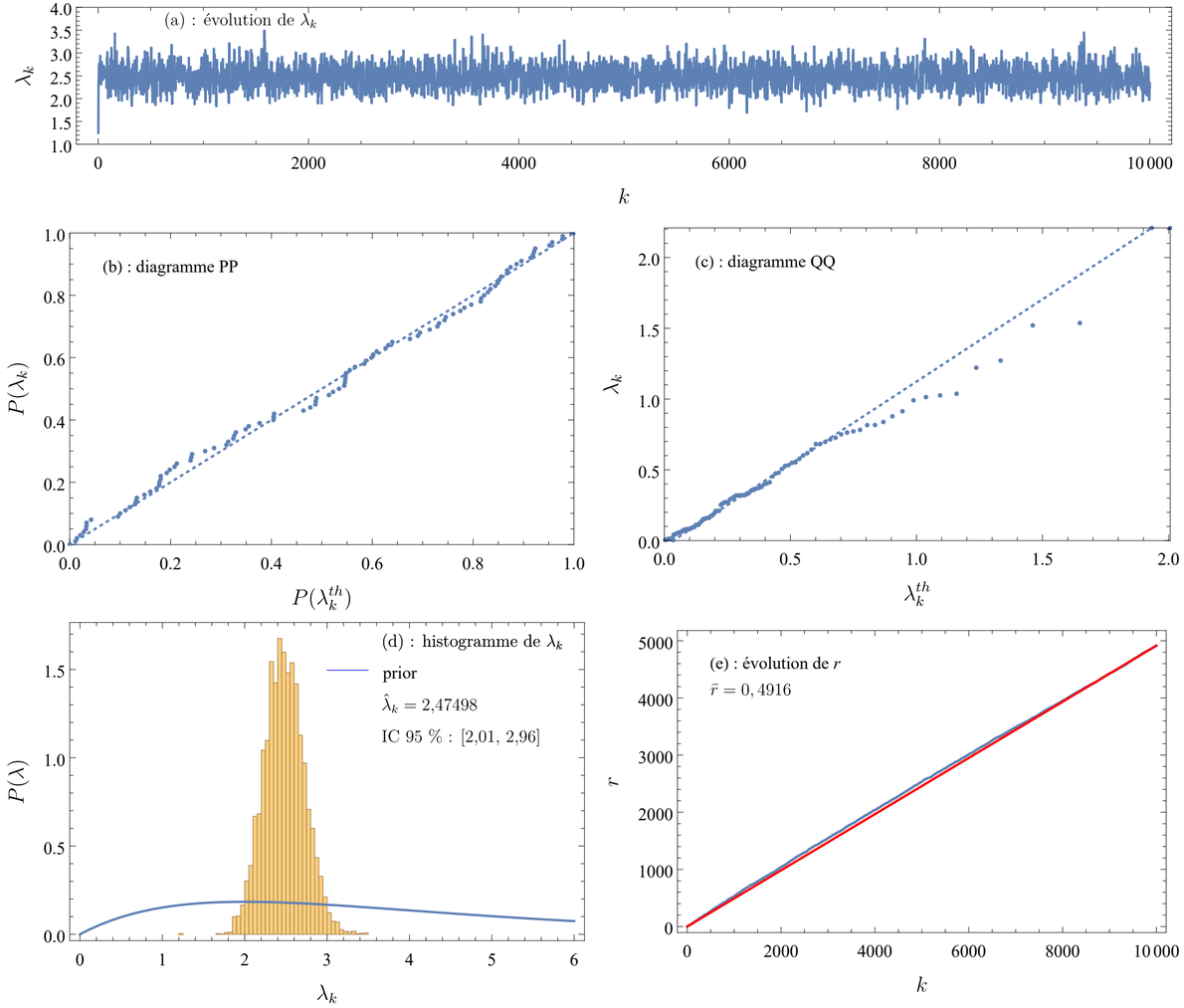


$n = 1000$ valeurs x_i

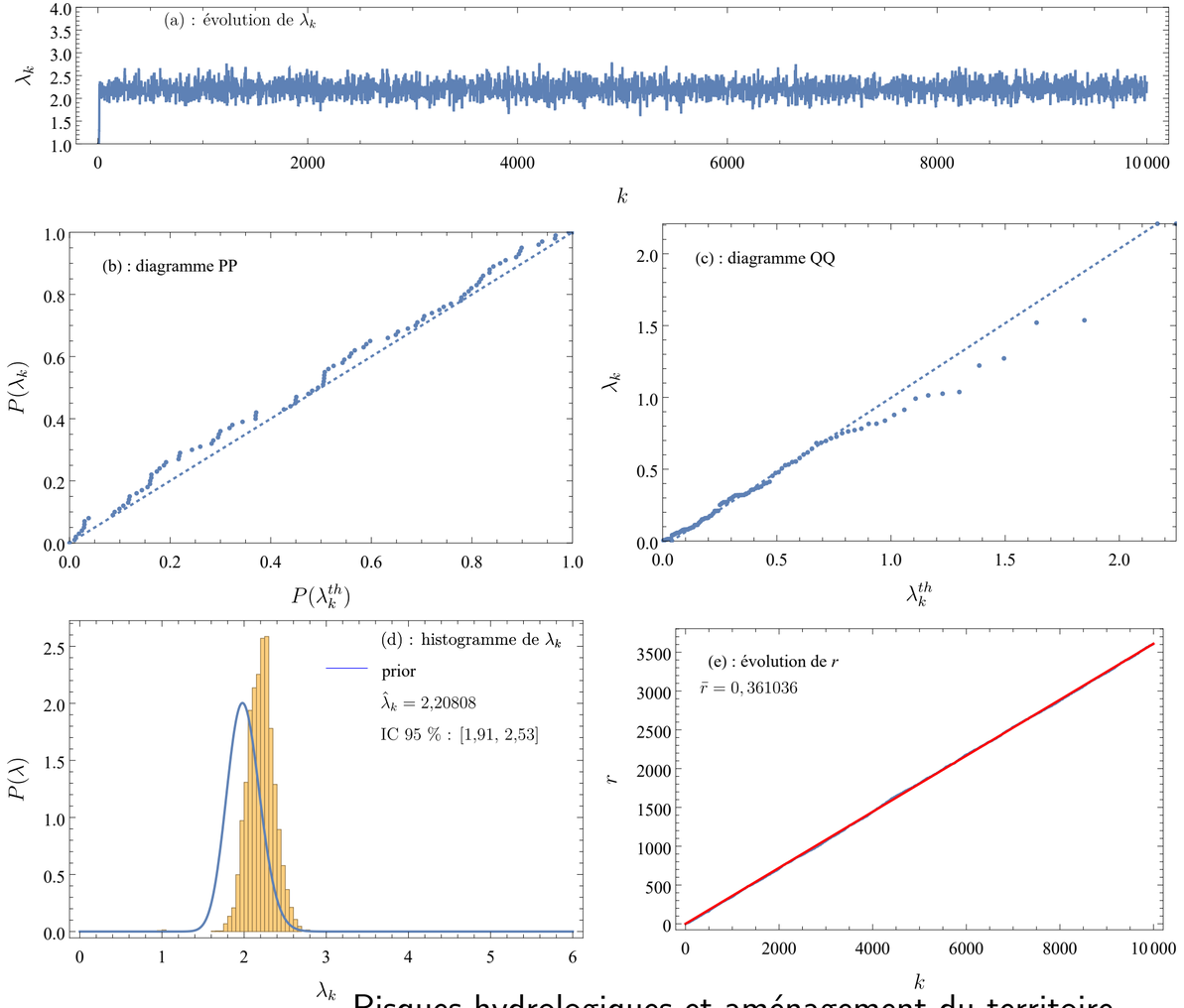


Exemple 2 (3) : influence du prior

$n = 100$ valeurs x_i et $\pi(\lambda) = \Gamma(2,2)$

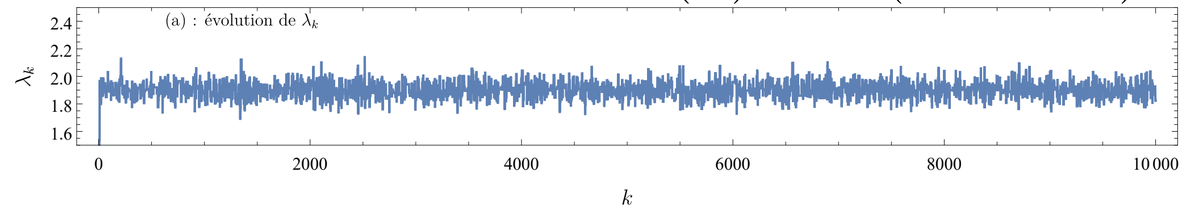


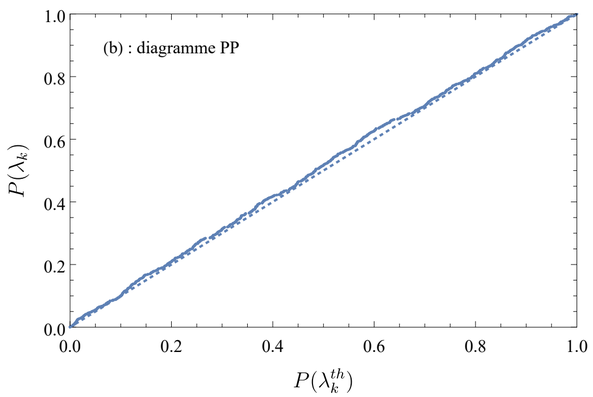
$n = 100$ valeurs x_i et $\pi(\lambda) = \Gamma(200,\ 0{,}2)$

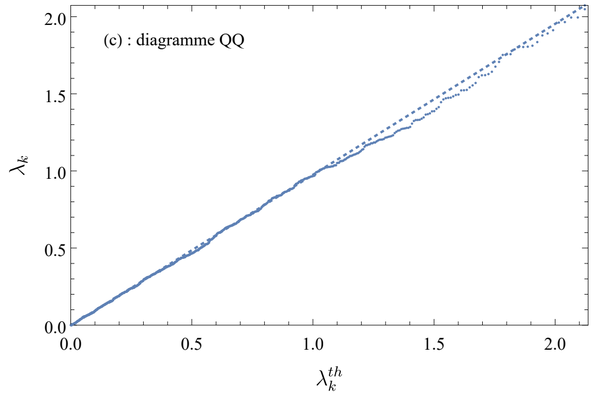


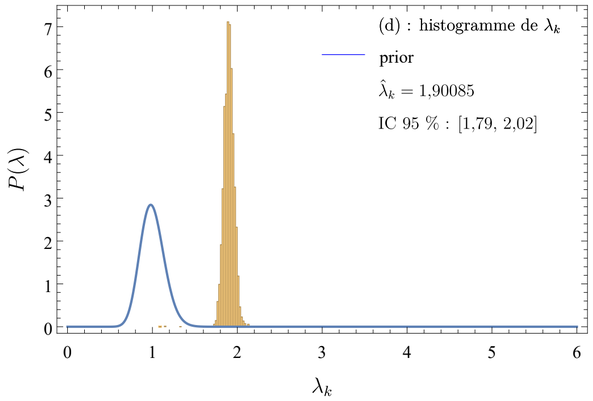
Exemple 2 (4) : et si le prior est mauvais ?

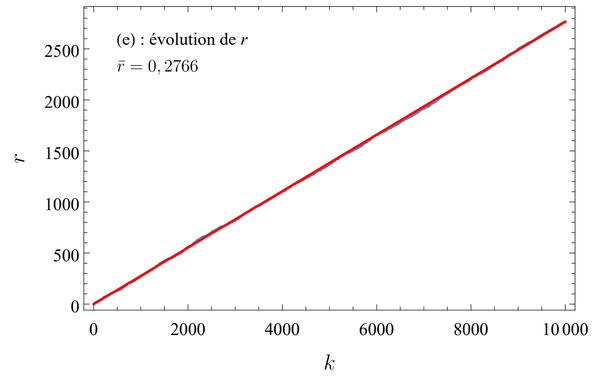
$n = 1000$ valeurs x_i et $\pi(\lambda) = \Gamma(100,\ 0{,}2)$











$n = 10$ valeurs x_i et $\pi(\lambda) = \Gamma(100,\ 0{,}2)$