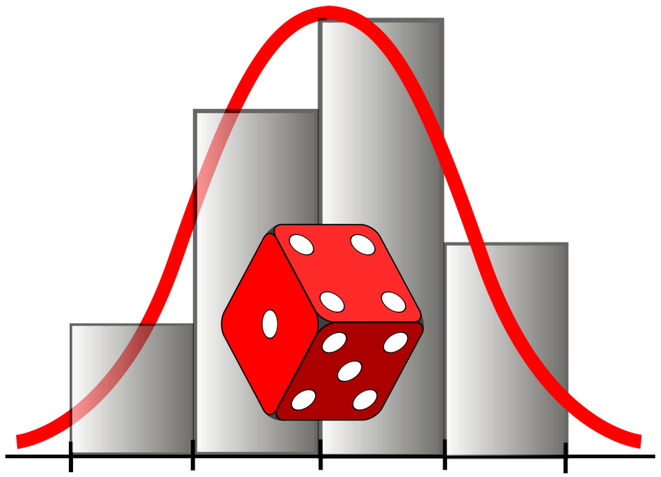

# Chapitre 4 : introduction à la théorie des valeurs extrêmes.
# Partie 2 : théorie des valeurs extrêmes

Risques hydrologiques et aménagement du territoire

Christophe Ancey

# Chapitre 4 : théorie des valeurs extrêmes



Une introduction...

- Historique
- Forme générique des valeurs extrêmes
- Formes particulières : Gumbel, Fréchet, Weibull
- Maxima annuels
- Méthodes à seuil
  - méthode du renouvellement
  - loi de Pareto
- Alternatives
  - loi de log-Pearson III
  - loi de mélange

# Historique

- Années 1920 : fondation des arguments asymptotiques par Ronald Fisher et Leonard Tippett, deux mathématiciens anglais
- Années 1940 : théorie asymptotique développée par Boris Gnedenko, un étudiant de Andrei Kolmogorov, puis Richard von Mises
- Années 1950 : Emil Gumbel, un mathématicien allemand émigré aux États-Unis, unifia les approches en montrant notamment que toutes les lois utilisées jusque lors pour décrire des valeurs extrêmes constituaient des cas particuliers d'une loi générale
- Années 1970 : travaux de James Pickands sur les lois limites

# Historique (2)

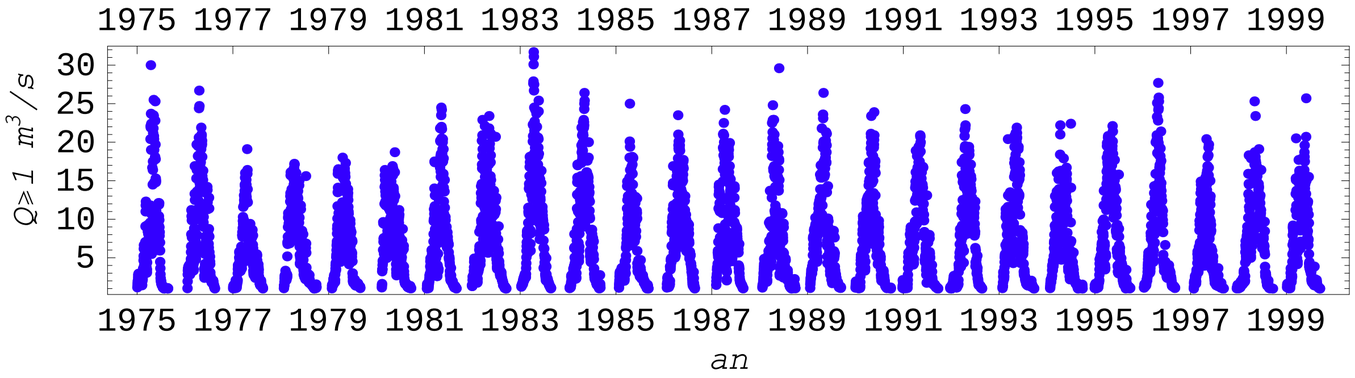
- Années 1980 : travaux de Leadbetter (entre autres) avec l'extension de la théorie aux processus aléatoires stationnaires
- Années 1990 : extension de la théorie des valeurs extrêmes aux processus à plusieurs variables aléatoires notamment en statistique financière, développement et application des techniques d'inférence (maximum de vraisemblance, inférence bayesienne)
- Années 2000 : développement de nouveaux champs tels que l'interpolation spatiale des valeurs extrêmes, la prise en compte de la non-stationnarité, etc.

# Lois de valeurs extrêmes

Considérons une variable aléatoire $X$ distribuée selon une loi $P(X)$. On définit $M_n$ la valeur maximale sur un bloc de $n$ valeurs : $M_n = \max\{X_i\}_{1 \leq i \leq N}$. On s'intéresse à la manière dont est distribuée cette nouvelle variable. La distribution de $M_n$ est donnée par

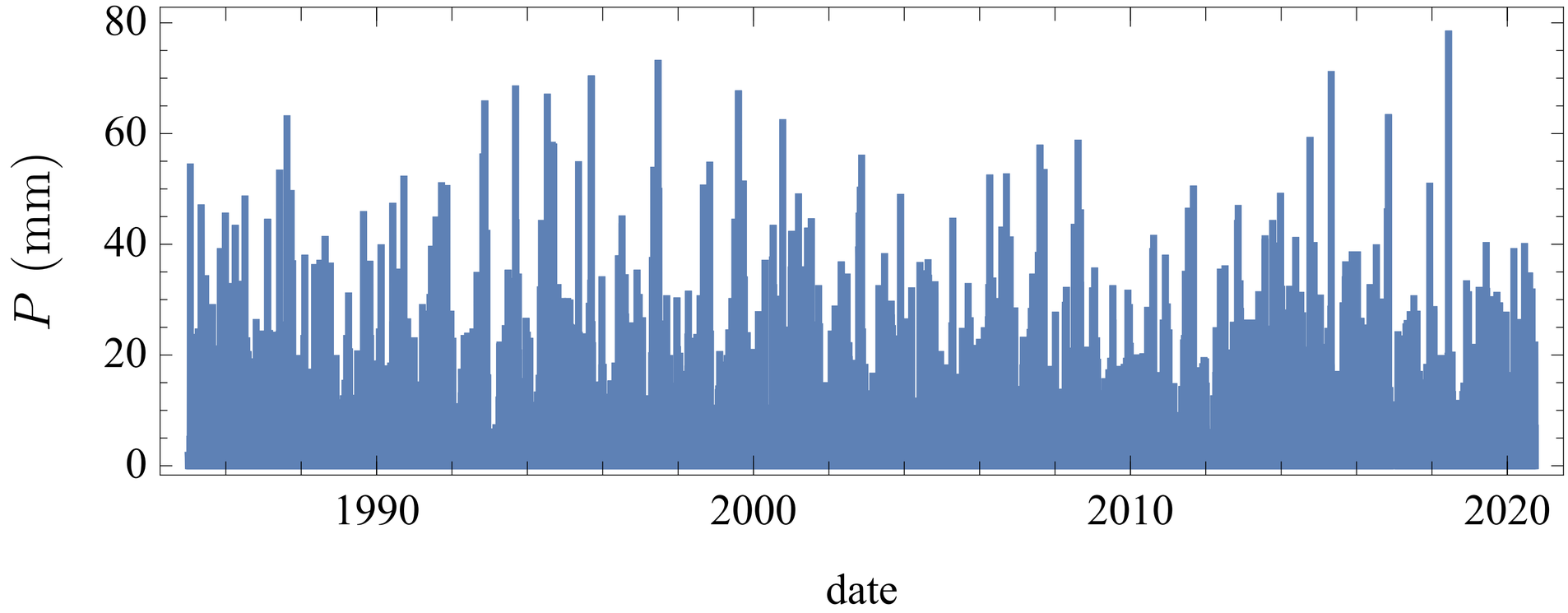
$$\text{prob}(M_n \leq x) = [\text{prob}\{X_i < x\}]^n = P(x)^n$$

Exemple de série temporelle : débits journaliers sur la Lonza supérieurs à 1 m$^3$/s



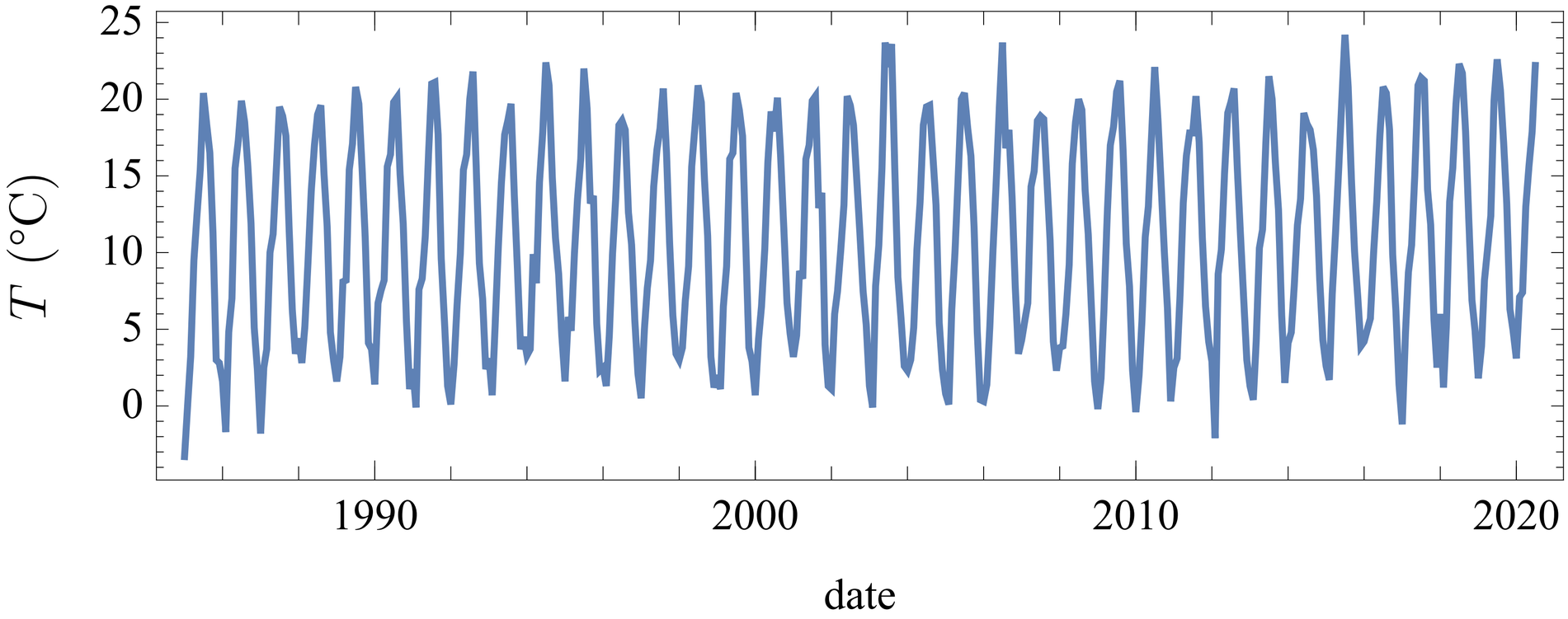
# Lois de valeurs extrêmes (2) : exemple

Précipitations journalières sur Lausanne



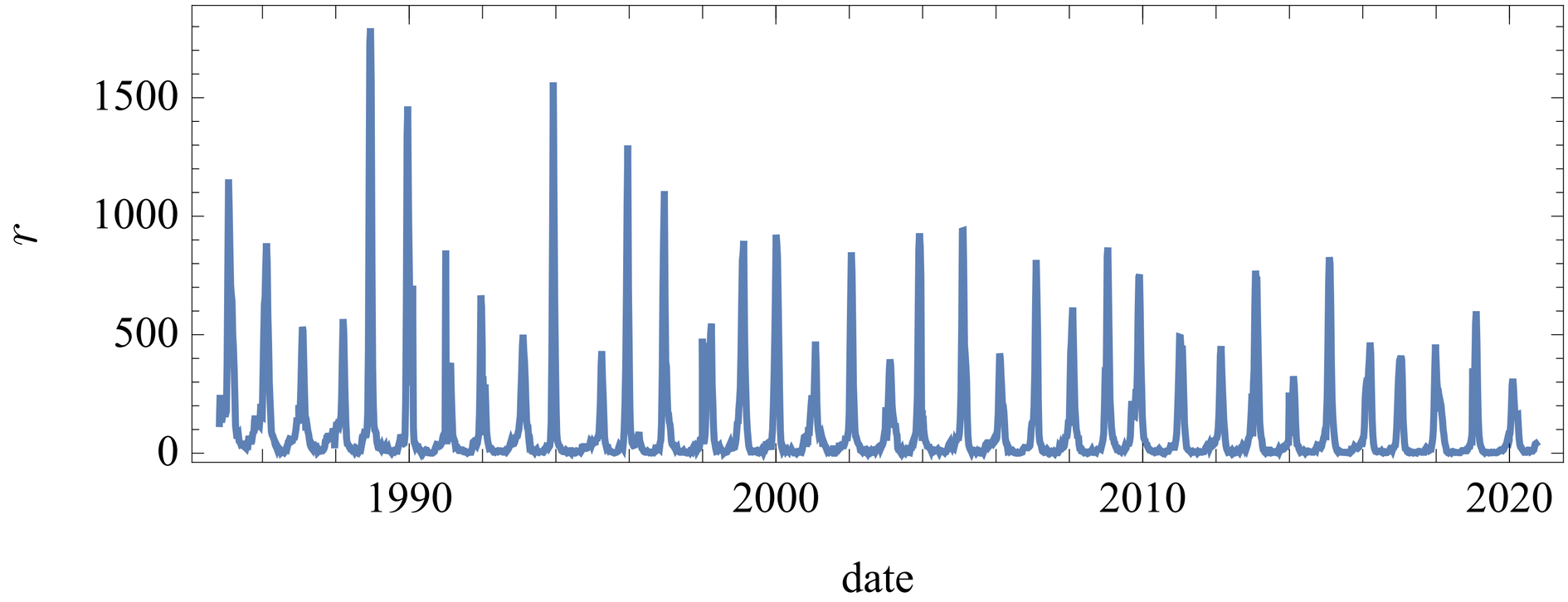
# Lois de valeurs extrêmes (3) : exemple

Températures moyennes mensuelles à Genève



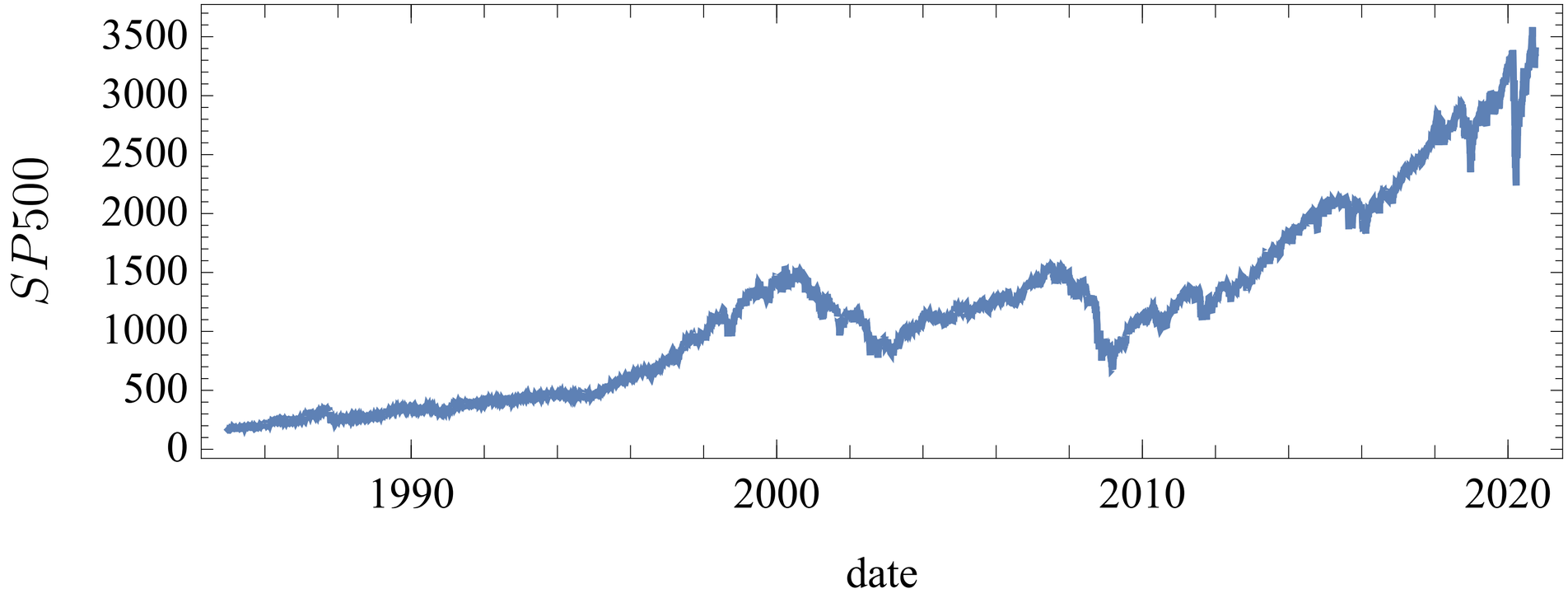
# Lois de valeurs extrêmes (4) : exemple

Exemple de la grippe (réseau Sentinelle pour la France) : évolution du taux d'incidence $r$ (nombre de malades pour $10^5$ hab.)



# Lois de valeurs extrêmes (5) : exemple

Indice boursier Standard & Poor's 500 (S&P500) aux États-Unis




Série non stationnaire !

# Lois de valeurs extrêmes (3)

Le problème est qu'en pratique $P$ n'est pas connu. Même s'il est possible de trouver une distribution empirique $\hat{P}$ qui approche $P$ raisonnablement bien, les erreurs s'additionnent de telle sorte que l'erreur d'estimation commise en substituant $P^n$ par $\hat{P}^n$ est généralement grande.

On a $\hat{P} = P(1+\epsilon)$ avec $\epsilon \ll 1$ l'erreur d'estimation que l'on suppose ici fixe. On a donc $\hat{P}^n = P^n(1+\epsilon)^n = P^n(1+n\epsilon+O(\epsilon))$, ce qui montre que l'erreur dans l'estimation de $P^n$ est $n\epsilon$. Comme $n$ est généralement grand, $n\epsilon$ n'est pas petit. Par exemple pour une chronique de débits journaliers, si l'estimation de $P$ est précise à $\epsilon = 0{,}1\,\%$ près, alors l'erreur sur le débit maximal annuel est précis à $n\epsilon = 36\,\%$ !

# Lois de valeurs extrêmes (4)

Argument asymptotique :

- Théorème central limite : la moyenne empirique $\bar{X} = \sum_{i=1}^{n} x_i/n$ tend en distribution vers la loi normale centrée

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \rightarrow \mathrm{N}(0,1)$$

- Similairement pour les maxima $M_n$, on peut trouver deux suites de nombres $a_n$ et $b_n$ telles que

$$Z = \frac{M_n - a_n}{b_n} \rightarrow H(z)$$

On montre qu'il existe une fonction $H(z) = \mathrm{prob}(Z \leq z)$.

# Lois de valeurs extrêmes (5)

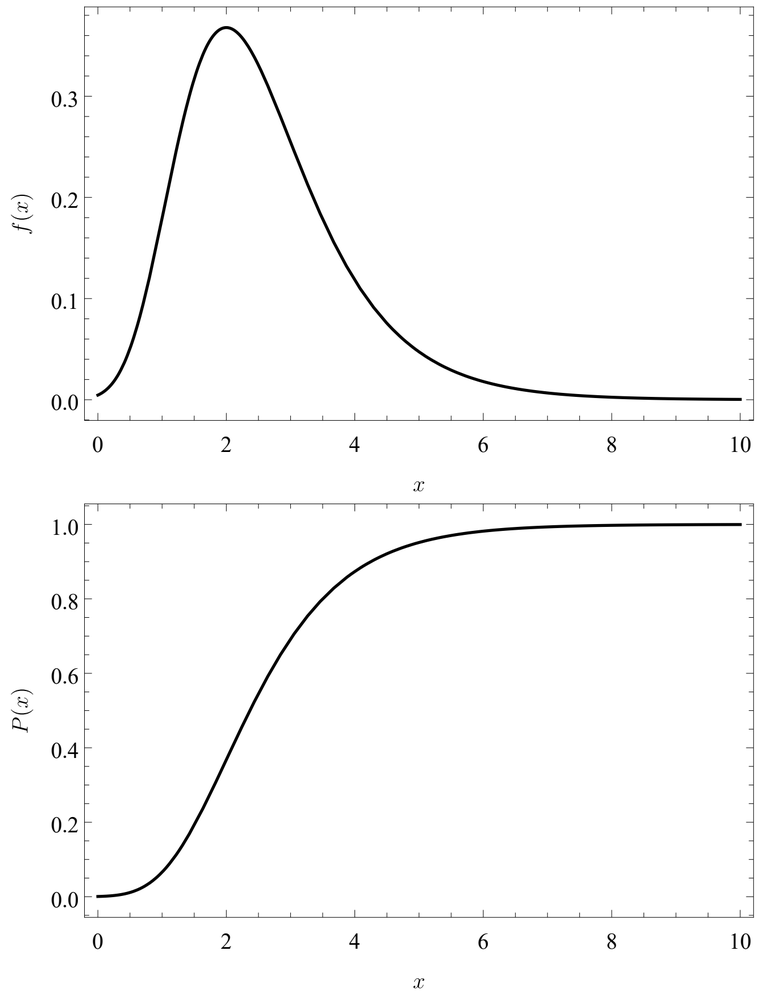
Sous réserve que $X$ vérifie quelques conditions, on montre que quand le nombre de blocs $N \to \infty$, alors les maxima sont distribués selon la loi :

$$P(x\,;\,\mu,\,\sigma,\,\xi) = \exp\left[-\left(1+\xi\frac{x-\mu}{\sigma}\right)_+^{-1/\xi}\right]$$

où $(x)_+ = \max(0,\,x)$.

On l'appelle *la distribution généralisée des valeurs extrêmes*, notée souvent GEV dans la littérature technique pour *Generalized Extreme Value*. Attention, le terme élevé à la puissance $-1/\xi$ peut être négatif.

# Lois de valeurs extrêmes (4)



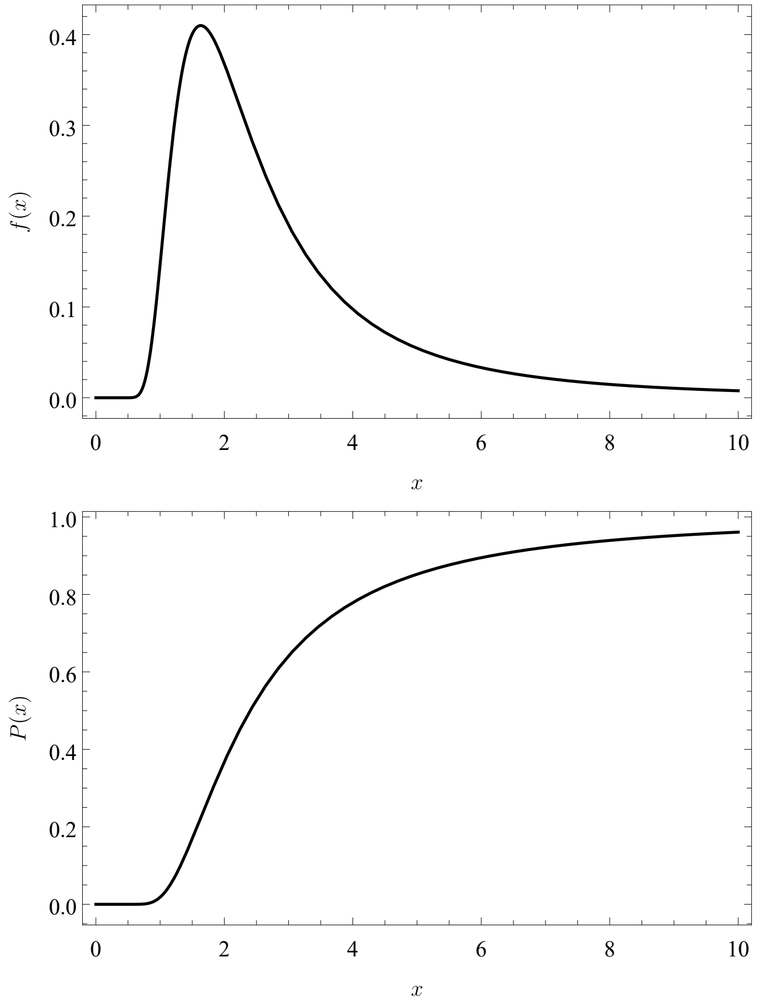

Trois distributions élémentaires :

1) La *loi de Gumbel* est une loi à deux paramètres définie sur $\mathbb{R}_+$, obtenue en faisant tendre $\xi$ vers 0 :

$$\mathrm{Gu}(x\,;\,\mu,\,\sigma) = \exp\left[-\exp\left(-\frac{x-\mu}{\sigma}\right)\right].$$

La moyenne est : $\mathbb{E}(X) = \mu + \sigma\gamma$ (avec $\gamma \approx 0{,}5772$ la constante d'Euler) ; la variance est : $\mathrm{var}(X) = \sigma^2\pi^2/6$.

# Lois de valeurs extrêmes (5)

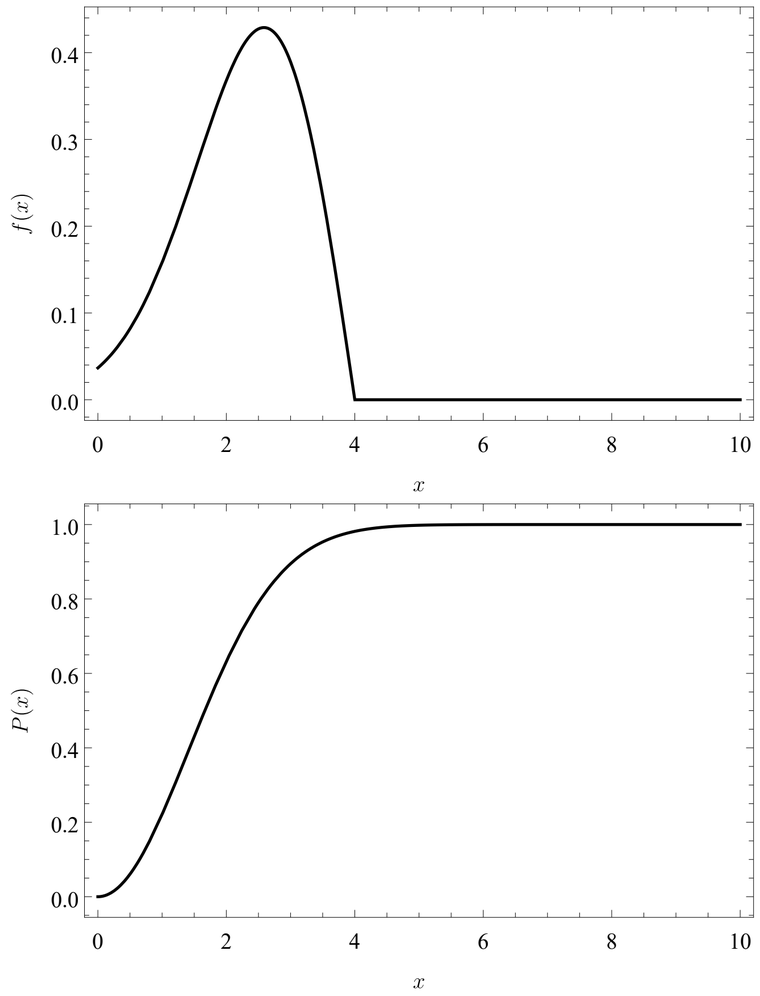



2) La *loi de Fréchet* est une loi à trois paramètres définie sur $]\mu - \sigma/\xi, +\infty[$, obtenue en prenant $\xi > 0$ :

$$\mathrm{Fr}(x\,;\,\mu,\,\sigma,\,\xi) = \exp\left(-\frac{1}{(1+\xi(x-\mu)/\sigma)^{1/\xi}}\right).$$

Une « queue » de distribution épaisse : les événements extrêmes sont bien plus fréquents que pour une distribution à queue mince (p. ex., loi normale).

# Lois de valeurs extrêmes (6)



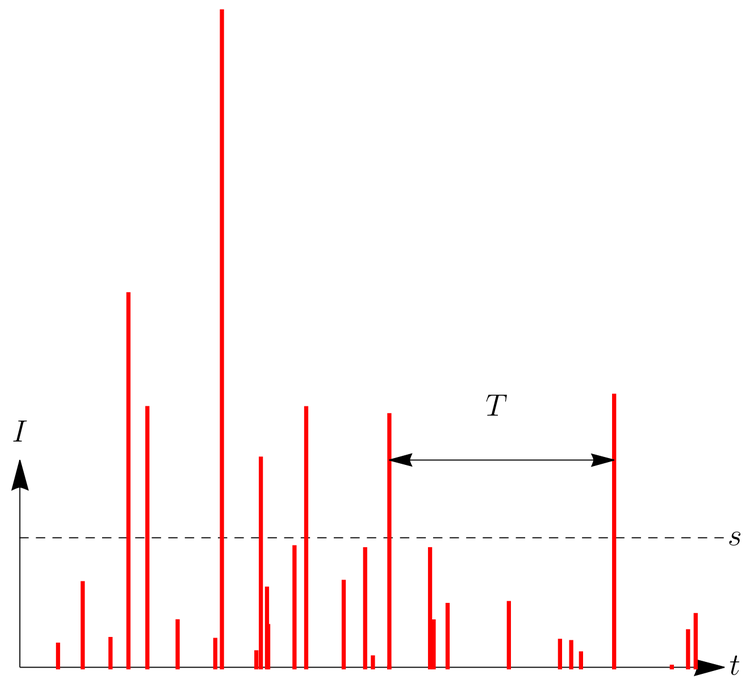

3) La *loi de Weibull* est une loi à trois paramètres définie sur $]-\infty,\ \mu+\sigma/|\xi|[$, obtenue en prenant $\xi<0$. On peut utiliser la même fonction de répartition que précédemment ou bien l'arranger un peu :

$$\mathrm{We}(x\,;\,\mu,\,\sigma,\,\xi)=\exp\left(-\left(|\xi|\frac{\mu+\sigma/|\xi|-x}{\sigma}\right)^{1/|\xi|}\right).$$

Existence d'une borne supérieure : les valeurs saturent...

# Période de retour

$T$ : intervalle de temps moyen entre deux événements, dont l'intensité atteint ou dépasse un certain seuil $s$




Un événement de période de retour $T$ a en moyenne une probabilité $1/T$ de se produire chaque année. Ainsi la crue centennale est :

- une crue qui se produit en moyenne tous les cent ans ;
- il y a en moyenne chaque année une probabilité de 1 % qu'une crue centennale ou plus rare se produise.

On relie la période de retour à la probabilité de dépassement $P(s) = \mathrm{prob}[I > s]$ ou de non-dépassement $P' = 1 - P$ :

$$T = \frac{1}{P} = \frac{1}{1 - P'}.$$

# Période de retour (2)

Relation entre période de retour $T$ (en années), probabilité de dépassement $P = 1/T$, et de non-dépassement $P' = 1 - P$

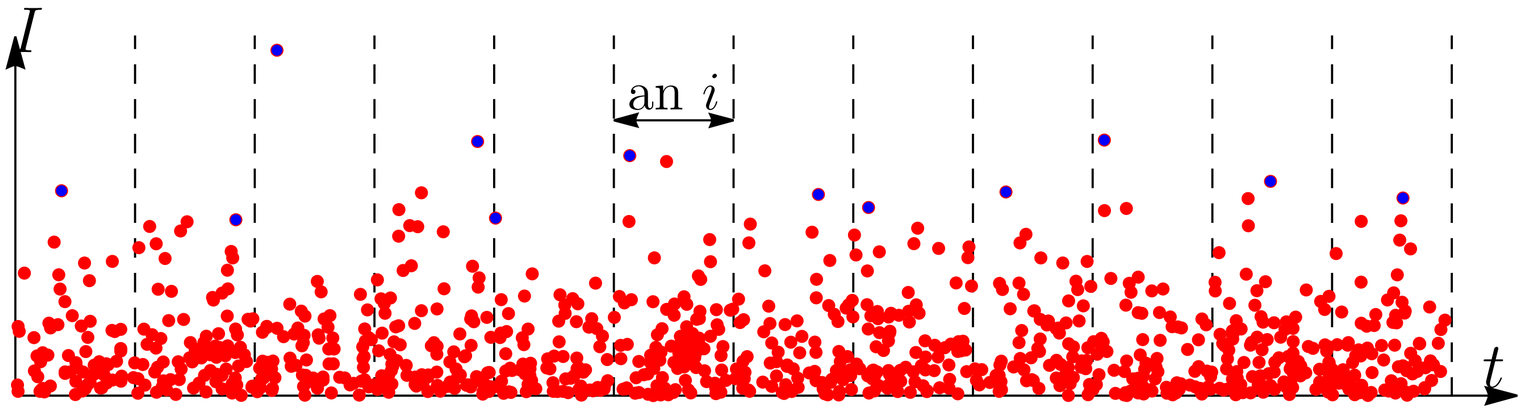
| $T$ (ans) | $P$ | $P'$ |
|---|---|---|
| $1$ | $1$ | $0$ |
| $10$ | $0{,}1$ | $0{,}9$ |
| $100$ | $0{,}01$ | $0{,}99$ |
| $1000$ | $0{,}001$ | $0{,}999$ |

# Maxima sur des pas de temps fixes

Problématique : on a des données et on cherche à ajuster une loi de valeurs extrêmes sur les maxima annuels de ces données

$$C = f(T),$$

avec $T$ la période de retour exprimée en années et $C$ le quantile étant la variable étudiée (chute de pluie, débit de pointe, etc.).



# Maxima sur des pas de temps fixes

Densité de probabilité, fonction de répartition $P$ (dépassement), quantiles $C$ pour les lois de valeurs extrêmes, période de retour comme $T = P^{-1}$

| fonction | $\xi \neq 0$ | $\xi = 0$ |
|---|---|---|
| densité | $\frac{1}{\sigma} e^{-\left(\frac{(c-\mu)\xi}{\sigma}+1\right)_+^{-1/\xi}} \left(\frac{(c-\mu)\xi}{\sigma}+1\right)_+^{-\frac{\xi+1}{\xi}}$ | $\frac{1}{\sigma} e^{\frac{\mu-c}{\sigma}-e^{\frac{\mu-c}{\sigma}}}$ |
| fonc. de répartition | $P = e^{-\left(\frac{(c-\mu)\xi}{\sigma}+1\right)^{-1/\xi}}$ | $P = e^{-e^{\frac{\mu-c}{\sigma}}}$ |
| quantile $(C(P))$ | $C = \mu - \frac{\sigma}{\xi}\left(1-(-\ln(1-P))^{-\xi}\right)$ | $C = \mu - \sigma \ln(-\ln(1-P))$ |
| quantile $(C(T))$ | $C = \mu - \frac{\sigma}{\xi}\left(1-\left(-\ln\left(1-\frac{1}{T}\right)\right)^{-\xi}\right)$ | $C = \mu - \sigma \ln\left(-\ln\left(1-\frac{1}{T}\right)\right)$ |

# Maxima sur des pas de temps fixes (2)

**Démarche** :

- On a un échantillon $\boldsymbol{x}$ de $n$ valeurs couvrant $n_a$ années.
- On classe les maxima annuels par ordre croissant $(C_i)_{1 \leq i \leq n_a}$. On note $\bar{C}$ la moyenne empirique de cet échantillon et $\mathrm{var}\, C$ sa variance.
- À chaque valeur de rang $i$, on affecte la probabilité empirique d'occurrence et la période de retour :

$$P_i = \frac{i - 0{,}28}{n_a + 0{,}28} \text{ et } T_i = \frac{1}{1 - P_i} = \frac{n_a + 0{,}28}{n_a - i + 0{,}56}.$$

- On ajuste une loi de valeurs extrêmes à l'aide de l'une des méthodes vues précédemment : méthode des moments, maximum du vraisemblance, ou inférence bayésienne.

# Maxima sur des pas de temps fixes (3)

Par exemple avec la méthode des moments et une loi de Gumbel :

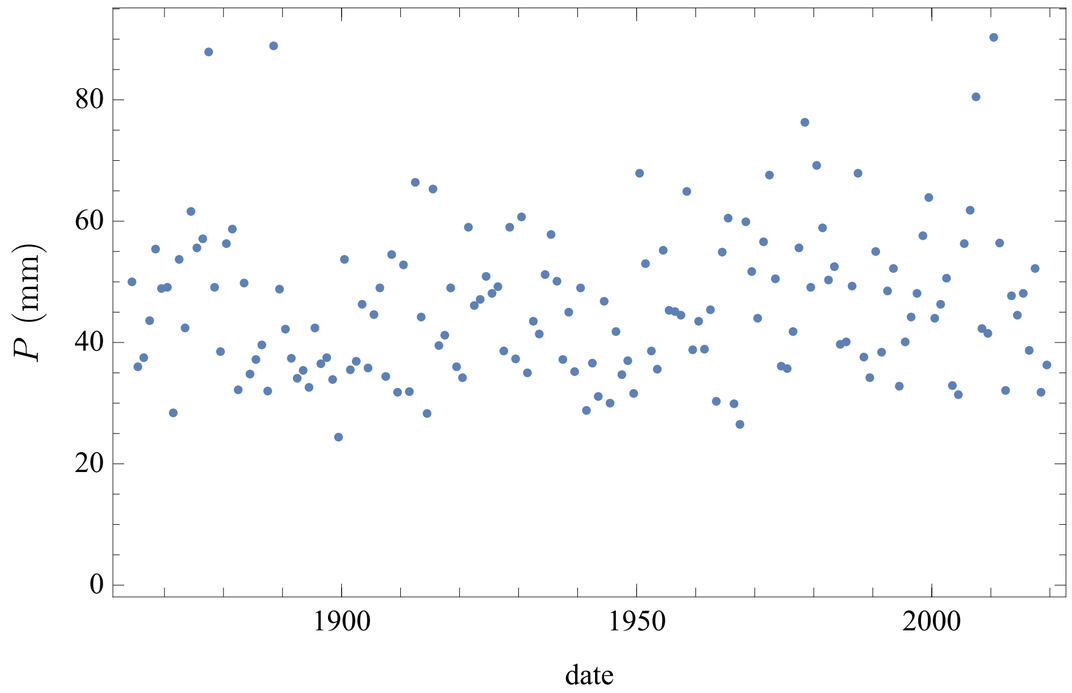
- On calcule par la méthode des moments les paramètres de la loi de Gumbel $\mathrm{Gu}[\mu,\, \sigma]$ ($\gamma \approx 0{,}577$ constante d'Euler) :

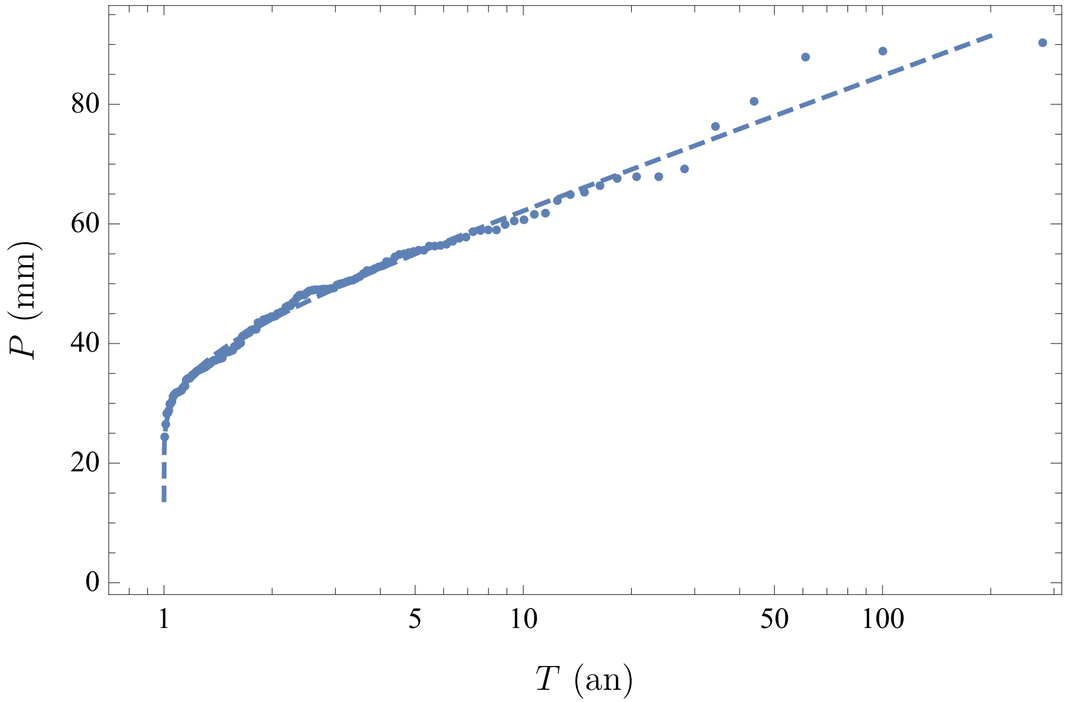
$$\sigma = \frac{\sqrt{6}}{\pi}\sqrt{\mathrm{var}C} \approx 0{,}7796\sqrt{\mathrm{var}C},$$
$$\mu = \bar{C} - \gamma\sigma \approx \bar{C} - 0{,}45\sqrt{\mathrm{var}C}.$$

- On reporte dans un diagramme $(T,\, C)$ la variation du quantile $C$ en fonction de la période de retour. On peut reporter à la fois les données $(T_i,\, C_i)_{1\leq i\leq n_a}$ et la loi de Gumbel ajustée $\mathrm{Gu}[T\,;\, \mu,\, \sigma]$ afin de vérifier visuellement l'adéquation de l'ajustement.

# Maxima sur des pas de temps fixes : exemple 1

Berne (1864–2019) : $P = 40{,}5 - 9{,}6 \log(-\log(1 - 1/T))$

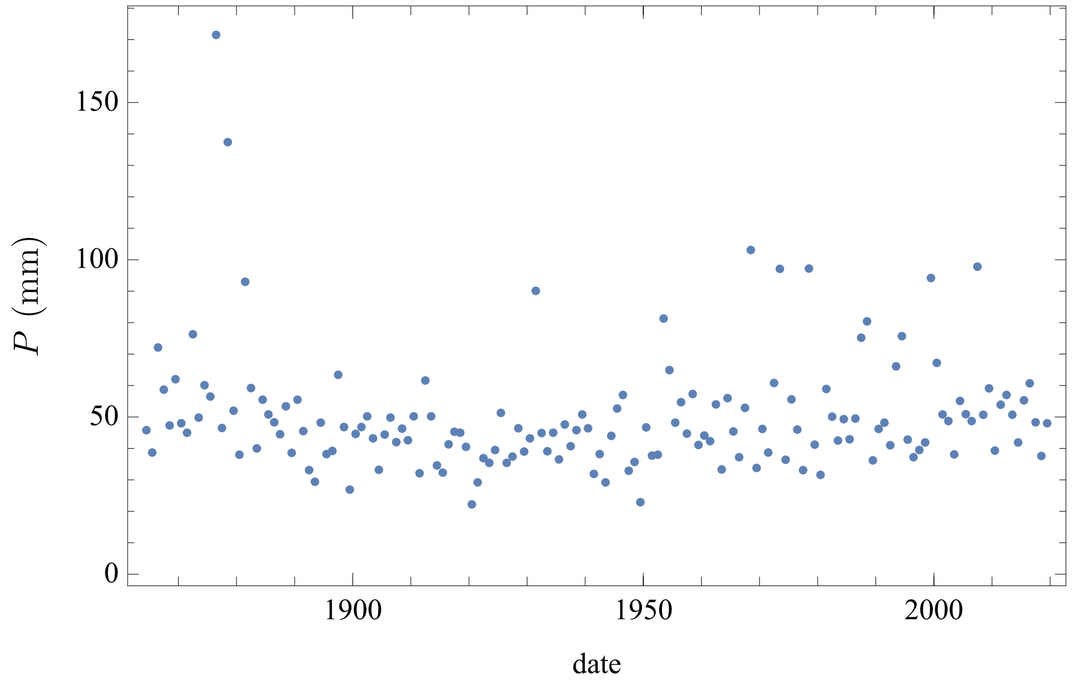


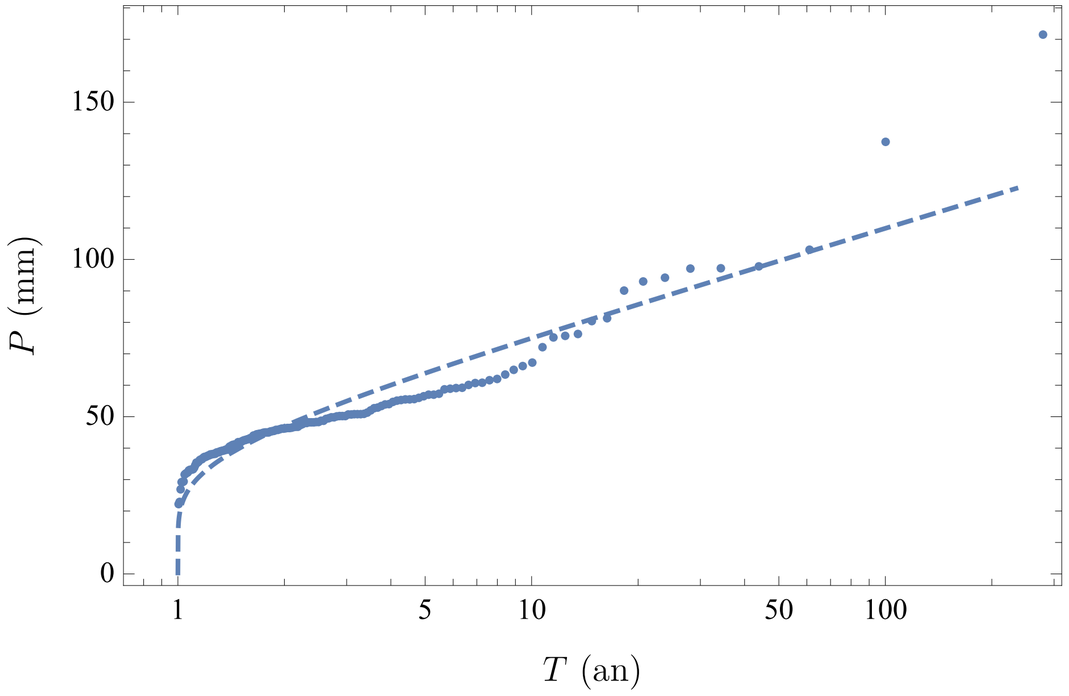



# Maxima sur des pas de temps fixes : exemple 2

Série des pluies sur Zurich/Fluntern (1864–2019) :

$$P = 40{,}5 - 9{,}6 \log(-\log(1 - 1/T))$$



# Maxima mensuels

La loi de composition des probabilités nous donne

$P_{an}(C) = \text{prob}[X < C \text{ sur une année}] = \text{prob}[X < C \text{ en janvier, } X < C \text{ en février, . .}$

$$P_{an}(C) = \prod_{i=1}^{12} \text{prob}[X < C \text{ sur le mois } i] = P_{mois}^{12}(C).$$

Considérons maintenant une loi de Gumbel (exprimée en non-dépassement) que l'on ajusterait sur les maxima mensuels

$$C = \mu - \sigma \ln(-\ln P_{mois}).$$

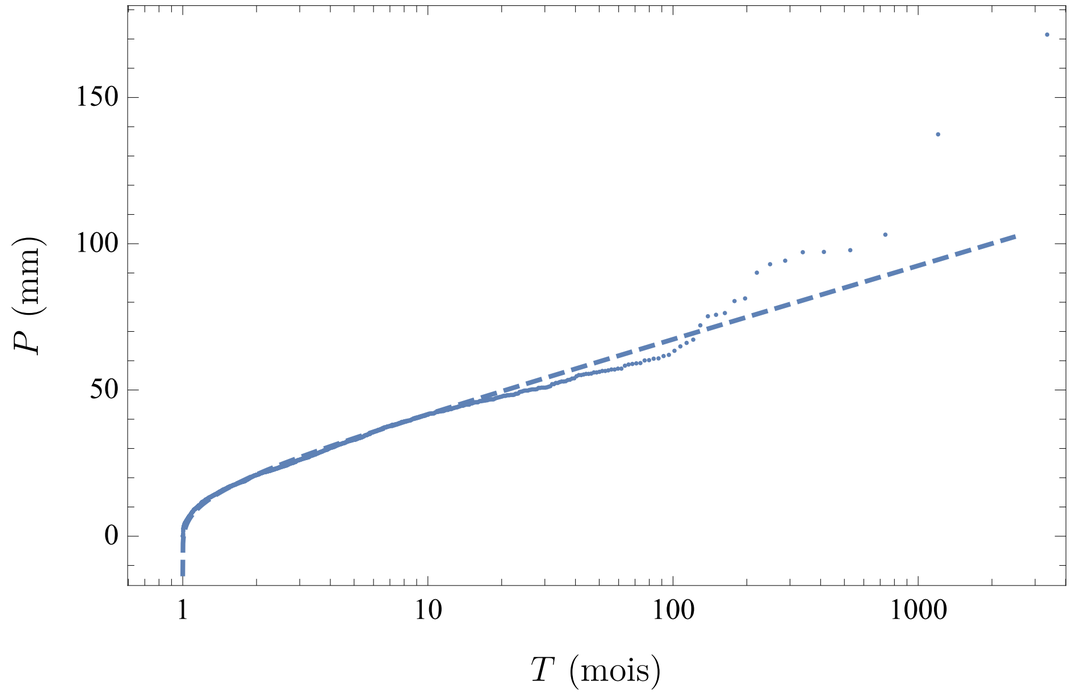
Pour repasser à une relation exprimée en années, on utilise la relation $P_{an}(C) = P_{mois}^{12}(C)$ et $P_{an} = 1 - 1/T$ ; on déduit

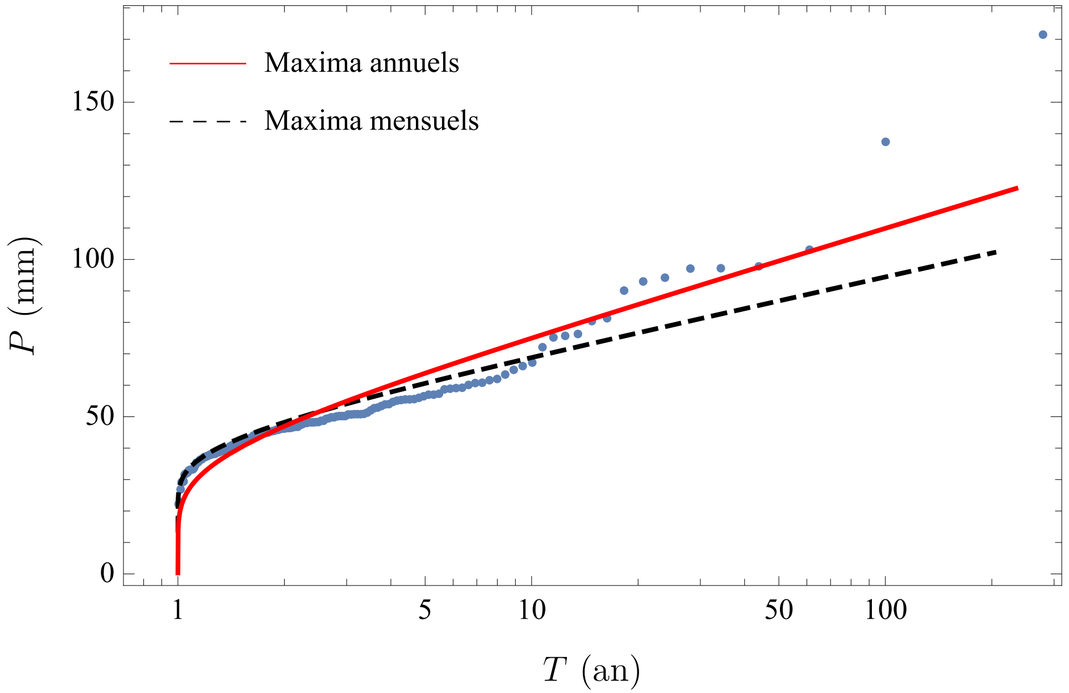
$$C = \mu - \sigma \ln(-\ln(P_{mois})) = \mu - \sigma \ln(-\ln(1 - T^{-1})^{1/12}) \approx \mu + \sigma \ln 12 + \sigma \ln T.$$

# Maxima mensuels : exemple

Série des maxima des pluies journalières sur Zurich/Fluntern (1864–2019) :

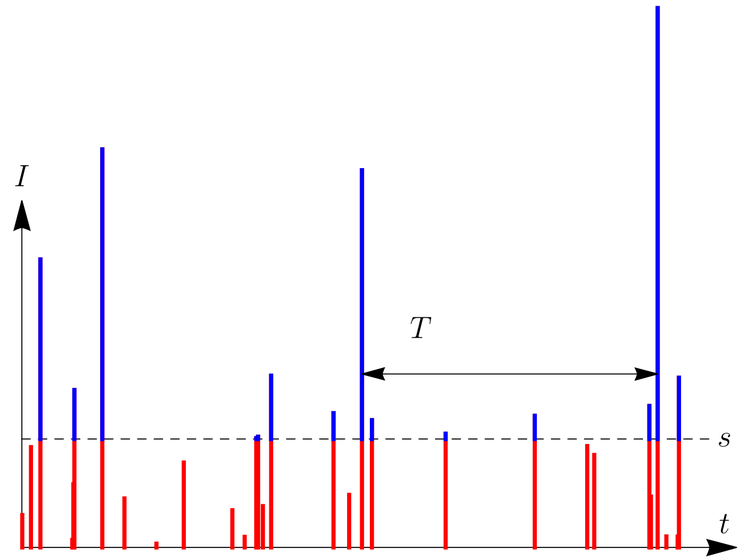
$$P_{mens} = 17{,}1 - 10{,}9 \log(-\log(1 - 1/T_{mois})) \Rightarrow P_{an} = 44{,}2 - 10{,}9 \log(-\log(1 - 1/T))$$



# Méthodes à seuil

Extrême : toute valeur dépassant à seuil

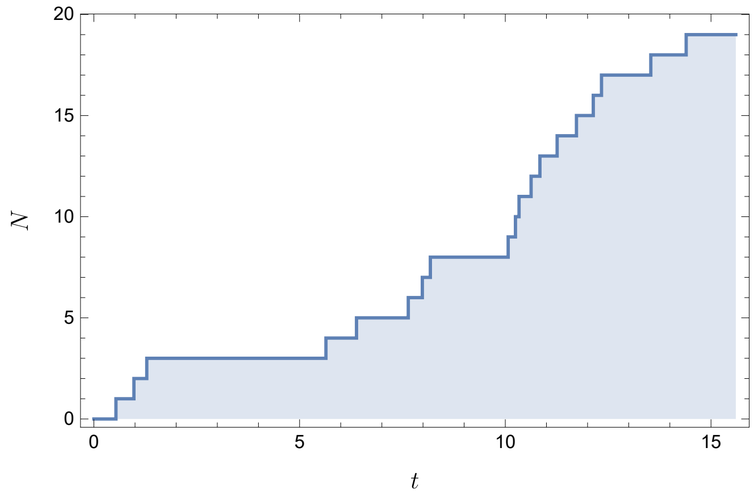



Méthode sur des pas de temps fixes : quantité d'information limitée... On peut définir un extrême comme une valeur dépassant un certain seuil. Deux grandes familles :

- théorie des valeurs extrêmes : méthode de dépassement du seuil (POT : peak over threshold). Loi de Pareto
- méthode du renouvellement : construction des « défaillances » (intensité, fréquence)

# Méthode du renouvellement

Extrême : toute valeur dépassant un seuil $s$




Processus de Poisson : $N(t)$ subit un saut au temps $t_i$ et augmente de $+1$. Le délai entre deux sauts est distribué selon la loi exponentielle.

Le processus de Poisson décrit, par exemple, l'arrivée (peu fréquente) de clients, d'électrons, de demandes d'impression...

On peut généraliser en s'intéressant à des temps non exponentiels et des intensités de saut différentes de 1.

# Méthode du renouvellement (2)

Quand on examine la distribution statique de la variable aléatoire $C$ au-dessus d'un seuil $s$, il y a deux éléments à prendre en compte :

- la fréquence d'occurrence $f(k)$ ou le temps $T_i$ entre deux événements ;
- l'intensité des phénomènes $G(c|s)$ conditionnée par l'existence du seuil.

Modèle le plus simple : $f$ est la loi de Poisson (autrement dit le temps entre deux événements est distribué exponentiellement) et l'intensité est aussi exponentielle : un modèle qui marche bien pour l'Europe occidentale hors la zone méditerranéenne

# Méthode du renouvellement (3)

$\text{prob}[C \leq c | C > s] =$

$\text{prob}[$au cours de l'année, il y a 0 chute $C > s]+$

$\text{prob}[$au cours de l'année, il y a 1 chute telle que $C > s$ et $C \leq c]+$

⋮

$\text{prob}[$au cours de l'année, il y a $k$ chutes telles que $C > s$ et $C \leq c]+$

⋮

Comme les événements sont indépendants

$$P(c|s) = \text{prob}[C \leq c | C > s] = \sum_{k=0}^{\infty} f(k) G(c|s)^k$$

# Méthode du renouvellement (4)

$$\text{prob}[C \leq c | C > s] = \sum_{k=0}^{\infty} \text{prob}[\text{au cours de l'année, il y a } k \text{ chutes d'intensité } C \text{ telles que } C > s \text{ et } C \leq c].$$

Après simplification, on trouve

$$\text{prob}[C \leq c | C > s] \approx 1 - \lambda \left(1 - G(c|s)\right)$$

Comment ajuster $f$ et $G$ sur des données ?

# Méthode du renouvellement (5)

On considère que l'on a un jeu de $n_d$ données couvrant $n_a$ années ; parmi ces $n_d$ données, il y a $n_s$ valeurs qui dépassent le seuil $s$. Si comme on l'a suggéré plus haut, on choisit une loi de Poisson $f(k) = \mathrm{Po}(k;\lambda) = \lambda^k e^{-\lambda}/k!$ et $G(C|s) = \mathrm{Exp}(x - s;\mu) = 1 - \exp[-\mu(C - s)]$, alors on montre par la méthode du maximum de vraisemblance que :

$$\lambda = \frac{n_s}{n_a} \text{ et } \mu = \frac{1}{\bar{C} - s},$$

avec $\bar{C} = \sum_{i=1}^{n_s} c_i/n_s$ la moyenne des $n_s$ valeurs de $C$ dépassant $s$. On déduit la loi $C(T)$ :

$$C = s - \frac{1}{\mu}\ln\left(-\frac{1}{\lambda}\ln\left(1 - \frac{1}{T}\right)\right) \approx s + \frac{\ln\lambda}{\mu} + \frac{1}{\mu}\ln T - \frac{1}{2\mu T}$$

# Loi de Pareto

On considère une série de valeurs *iid* $X_1$, $X_2$, etc., tirées de $F$. Sélectionner des événements extrêmes revient à se fixer un seuil $s$ assez élevé et à retenir toutes les valeurs de $X$ qui dépassent $s$. La probabilité conditionnelle est alors pour $y > 0$

$$H(y) = \text{prob}[X > s + y | X > s] = \frac{1 - F(s + y)}{1 - F(s)}$$

Quand on possède un nombre suffisant de données et pour $s$ suffisamment grand, alors, $H$ peut être approché par une *distribution généralisée de Pareto* :

$$G(x) = 1 - \left(1 + \frac{\hat{\xi} x}{\hat{\sigma}}\right)^{-1/\hat{\xi}},$$

avec $\hat{\xi}$ et $\hat{\sigma}$ les deux paramètres de la loi (le troisième implicite est $s$).

# Loi de Pareto (2)

On peut relier ces paramètres à leurs équivalents dans la loi des valeurs extrêmes $\hat{\xi} = \xi$ et $\hat{\sigma} = \sigma + \xi(s - \mu)$. Lois de Pareto et des valeurs extrêmes sont duales. Le comportement de $G$ est entièrement dicté par le signe de $\hat{\xi}$ :

- Si $\hat{\xi} < 0$, les quantiles associés à la loi de Pareto généralisée sont bornés par $s - \hat{\sigma}/\hat{\xi}$
- Si $\hat{\xi} = 0$, la distribution tend vers une loi exponentielle de paramètre $1/\hat{\sigma}$

$$G(x) = 1 - \exp\left(-\frac{x}{\hat{\sigma}}\right)$$

- Si $\hat{\xi} > 0$, les quantiles croissent indéfiniment vers l'infini (pendant du comportement de la loi de Fréchet)

# Loi de Pareto : choix du seuil

Le problème principal est la détermination du seuil $s$ :

- Si $s$ est trop petit, les valeurs ne sont pas extrêmes et on ne peut pas espérer que la densité de probabilité de l'échantillon s'approche d'une loi de Pareto
- Si $s$ est trop grand, il y a peu de données dans l'échantillon et la variance de l'estimateur est grande

Une méthode pour optimiser le choix de $s$ est de regarder la moyenne des excès $Y = X - s$ pour les valeurs dépassant le seuil $X > s$ :

$$\mathbb{E}[Y] = \int_0^\infty y G'(y) \mathrm{d}y = \frac{\sigma_s}{1-\xi}$$

Le coefficient $\sigma_s$ dépend du choix de $s$.

# Loi de Pareto : choix du seuil (2)

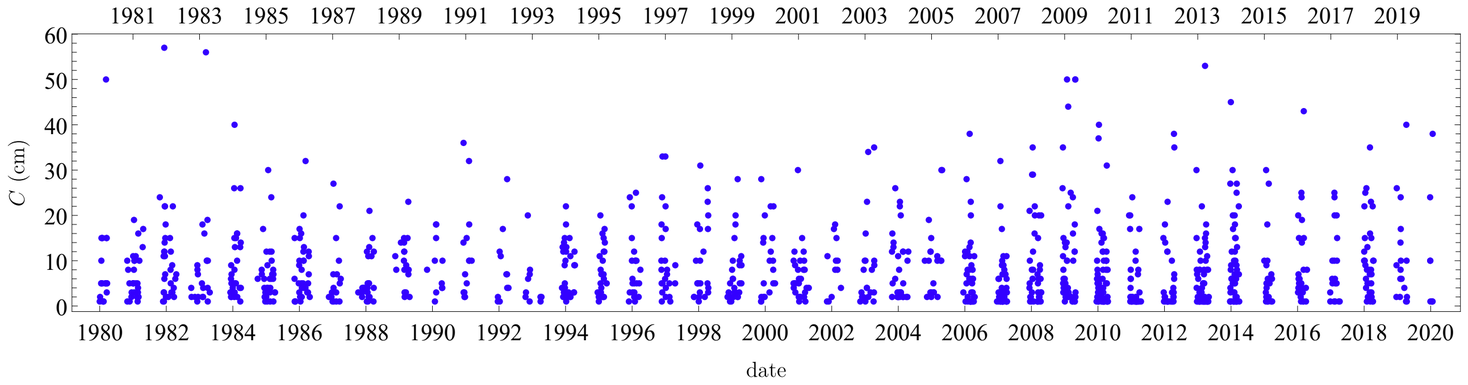
Donc, pour tout $s > s_0$ ($s_0$ premier seuil de définition), on trouve

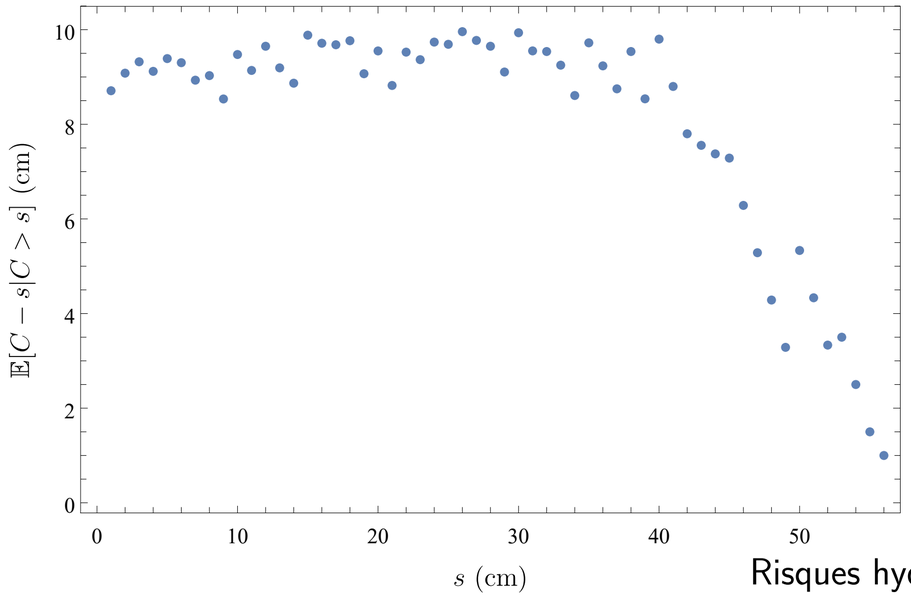
$$\mathbb{E}[X - s|X > s] = \frac{\sigma_s}{1-\xi} = \frac{\sigma_{s_0} + \xi(s - s_0)}{1-\xi}$$

Pour tout seuil $s > s_0$, alors $\mathbb{E}[X - s|X > s]$ doit être une fonction linéaire de $s$. Si on trace la courbe $\mathbb{E}[X - s|X > s] = f(s)$ et de rechercher le domaine sur lequel la fonction $f$ est linéaire. La connaissance de ce domaine linéaire permet également de déterminer la valeur de $\xi$. Notamment si $\mathbb{E}[X - s|X > s] = f(s)$ ne varie pas quand $s$ croît (domaine linéaire horizontal), alors $\xi \approx 0$ et un modèle de Gumbel est bien adapté à décrire les extrêmes de l'échantillon.

# Loi de Pareto : exemple

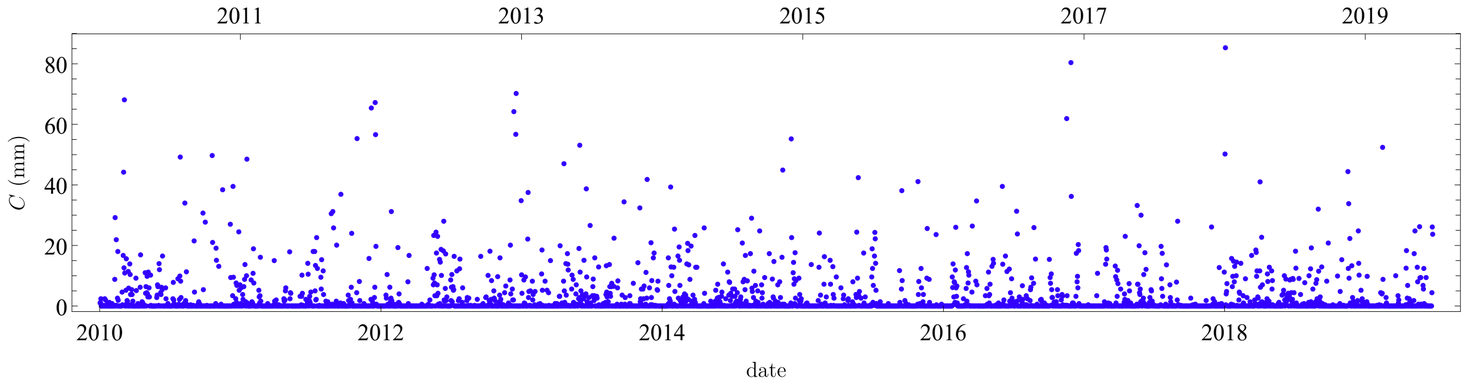
Chutes de neige journalières sur les Orres : Gumbel ?

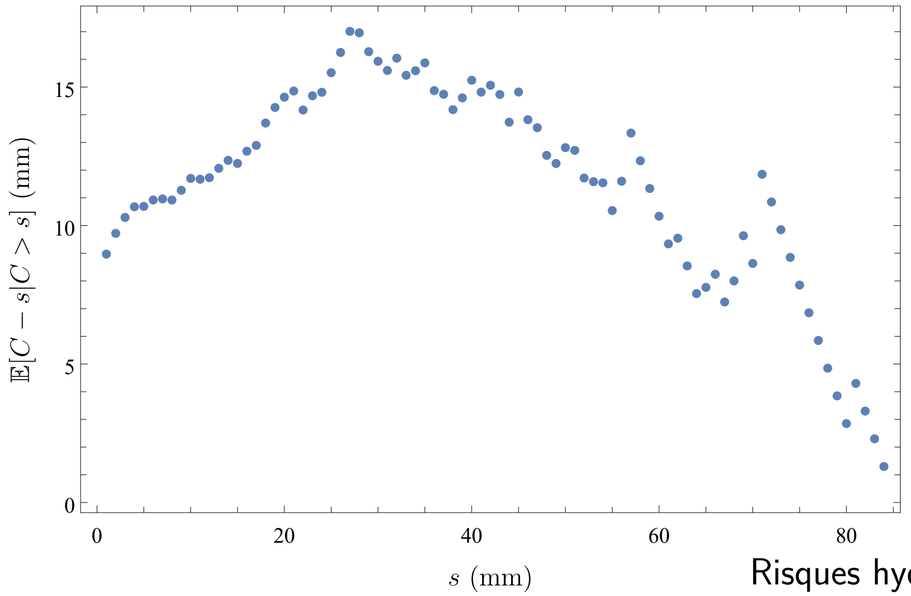


# Loi de Pareto : exemple 2

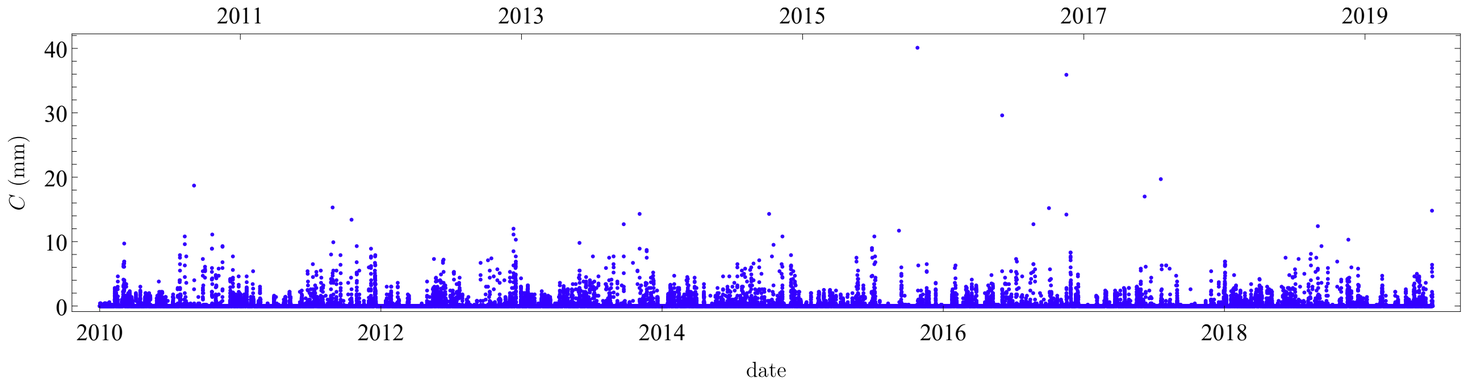
Précipitations journalières sur les Orres : Weibull ?

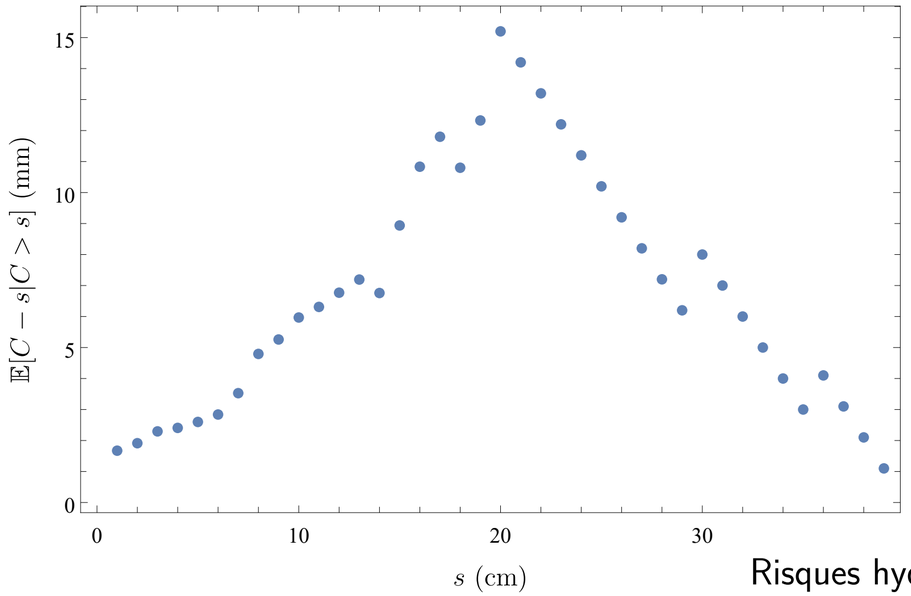


# Loi de Pareto : exemple 3

Précipitations horaires sur les Orres : Fréchet ?



# Passage de la loi de Pareto à une loi de valeurs extrêmes

Problème : comment introduire la période de retour dans la loi de Pareto ?

Asymptotiquement, on doit avoir le même comportement. Cela conduit à poser :

$$\xi = \xi_p$$

$$\mu \approx s + \frac{\sigma_p}{\xi_p}\left((n_o\zeta_s)^{\xi} - 1\right)$$

$$\sigma \approx \sigma_p(n_o\zeta_s)^{\xi}$$

# Passage de la loi de Pareto à une loi de valeurs extrêmes (2)

- Soit un échantillon $\boldsymbol{x}$ de $n_d$ données, sur $n_a$ années, avec $n_s$ valeurs dépassant $s$.
  On définit $\zeta_s$ et $n_o$

$$n_o = \frac{n_d}{n_a} \text{ et } \zeta_s = \frac{n_s}{n_d}.$$

  On obtient un nouvel échantillon de valeurs notées $(Y_i)_{1 \leq i \leq n_s}$ que l'on classe par ordre décroissant.

- Pour chaque valeur de rang $i$, on attribue la probabilité d'occurrence et la pseudo-période empiriques :

$$P_i = \frac{i}{n_s + 1} \text{ et } m_i = \frac{n_s + 1}{i}.$$

# Passage de la loi de Pareto à une loi de valeurs extrêmes (3)

- On cale les paramètres de Pareto $\xi_p$ et $\sigma_p$.
- La courbe $C = f(T)$ est la suivante

$$C(T) = \begin{cases} \text{si } \xi_p \neq 0, s + \frac{\sigma_p}{\xi_p}\left((Tn_o\zeta_s)^{\xi} - 1\right) \\ \text{si } \xi_p = 0, s + \sigma_p \ln(Tn_o\zeta_s) \end{cases}$$

# Intervalle de confiance des quantiles

Problématique : on a calé une loi de probabilité $C = F(T\ ;\ \boldsymbol{\theta})$ sur des données, on veut déterminer l'intervalle de confiance des quantiles $C$ pour une période de retour donnée.
On a montré que l'on pouvait estimer l'intervalle de confiance des paramètres $\boldsymbol{\theta} = (\mu,\ \sigma,\ \xi)$ de cette loi si on utilise la méthode du maximum de vraisemblance ou l'inférence bayésienne. En quoi cette connaissance de l'incertitude sur $\boldsymbol{\theta}$ peut-être utile ?
Deux approches possibles :

- développement de Taylor si on a utilisé la méthode du maximum de vraisemblance
- tirage des quantiles à partir du posterior si on a utilisé l'inférence bayésienne

# Intervalle de confiance des quantiles : Taylor

Cas général : si $Y$ est une variable aléatoire reliée à une autre variable aléatoire $X$ (de densité de probabilité $f$) par une fonction déterministe $v$ :

$$Y = v(X)$$

alors la densité de probabilité $g$ de $Y$ est

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

**Problème** : généralisable pour des fonctions multivariées, mais calcul non analytique...

# Intervalle de confiance des quantiles : Taylor (2)

**Solution approchée** : on calcule les moments de $Y$ à partir de ceux de $X$. Pour cela on se sert de développement de Taylor :

$$v(X) = v(m) + (X - m)v'(m) + \frac{1}{2}(x - m)^2 v''(m) + \cdots$$

où $m = \mathbb{E}(x)$.

De là, on peut appliquer l'opérateur moyenne

$$\mathbb{E}(Y) = \mathbb{E}(v(X)) = v(m) + \frac{1}{2}v''(m)\mathrm{Var}X$$

car l'opérateur moyenne est linéaire. Par exemple, si $a$ est une constante

$$\mathbb{E}(aX) = a\mathbb{E}(X) = am$$

# Intervalle de confiance des quantiles : Taylor (3)

Pour calculer la variance, on se sert de l'identité

$$\mathrm{Var}(aX + b) = a^2 \mathrm{Var} X,$$

qui se généralise à deux variables

$$\mathrm{Var}(aX + bY + c) = a^2 \mathrm{Var} X + b^2 \mathrm{Var} Y + 2ab\mathrm{Cov}(X,\, Y).$$

ou $n$ variables

$$\mathrm{Var}\left(\sum_{i=1}^{n} a_i X_i\right) = \sum_{i=1}^{n} a_i^2 \mathrm{Var} X_i + 2 \sum_{1 \leq i < j \leq n}^{n} a_i a_j \mathrm{Cov}(X_i,\, X_j).$$

# Intervalle de confiance des quantiles : Taylor (4)

**Application aux lois de valeurs extrêmes** :

$$C = F(T\ ;\ \boldsymbol{\theta}) = \mu - \frac{\sigma}{\xi}\left(1 - \left(-\ln\left(1 - \frac{1}{T}\right)\right)^{-\xi}\right)$$

On suppose que l'on a une estimation des paramètres $\hat{\boldsymbol{\theta}}$, de la variance empirique $\mathrm{Var}\boldsymbol{\theta}$, et de la covariance des paramètres $\mathrm{Cov}(\theta_i,\theta_k)$.
On introduit les variables intermédiaires pour le cas $\xi \neq 0$ :

$$C_\mu = \left.\frac{\partial C}{\partial \mu}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} = 1,$$

$$C_\sigma = \left.\frac{\partial C}{\partial \sigma}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} = \frac{1}{\hat{\xi}}\left(1 - \left(-\ln\left(1 - \frac{1}{T}\right)\right)^{-\hat{\xi}}\right),$$

$$C_\xi = \left.\frac{\partial C}{\partial \xi}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} = \frac{\hat{\sigma}\left(-\ln\left(1-\frac{1}{T}\right)\right)^{-\hat{\xi}}\left(\left(-\ln\left(1-\frac{1}{T}\right)\right)^{\hat{\xi}} - \hat{\xi}\ln\left(-\ln\left(1-\frac{1}{T}\right)\right) - 1\right)}{\hat{\xi}^2}.$$

# Intervalle de confiance des quantiles : Taylor (5)

**Calcul de la variance** :

$$\mathrm{Var}C = \nabla C \cdot \boldsymbol{V} \cdot \nabla C,$$

où $\nabla C = (\partial_\mu C, \partial_\sigma C, \partial_\xi C)$ est le gradient de $C$ par rapport à $\boldsymbol{\theta}$, évalué en $\hat{\boldsymbol{\theta}}$, et où la matrice $\boldsymbol{V}$ de variance-covariance (dont les éléments sont $\mathrm{Cov}(X_i, X_j)$ et en diagonale $\mathrm{Cov}(X_i, X_i) = \mathrm{Var}X_i$) est liée à la matrice d'information observée :

$$\boldsymbol{V} = \boldsymbol{I}_O^{-1}$$

L'intervalle de confiance à $\alpha$ % du quantile $C$ s'écrit donc

$$C(T) = \nabla C \cdot \boldsymbol{V} \cdot \nabla C \pm z_{\alpha/2}\sqrt{\mathrm{Var}C(T)}.$$

avec $z_{\alpha/2}$ le quantile de la loi normale associée à la probabilité $1 - \alpha/2$.

# Intervalle de confiance des quantiles : Bayes

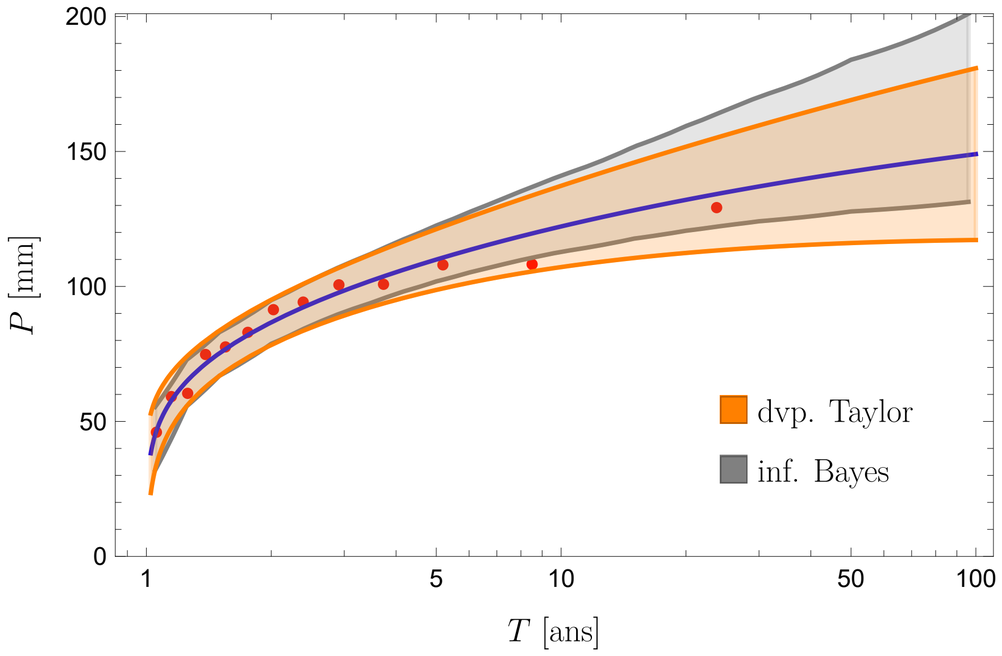
**Principe** : quand on se sert d'un algorithme de type Metropolis, on calcule la probabilité *a posteriori* d'observer le paramètre $\boldsymbol{\theta}$ à partir d'un jeu de données $\boldsymbol{d}$ et de la vraisemblance $\mathrm{prob}(\boldsymbol{d}|\boldsymbol{\theta})$ de ce jeu de données

$$\mathrm{prob}(\boldsymbol{\theta}|\boldsymbol{d}) = \frac{\mathrm{prob}(\boldsymbol{d}|\boldsymbol{\theta})\mathrm{prob}(\boldsymbol{\theta})}{\int \mathrm{d}\boldsymbol{\theta}\mathrm{prob}(\boldsymbol{d}|\boldsymbol{\theta})\mathrm{prob}(\boldsymbol{\theta})}.$$

En même temps que l'on simule l'échantillon $\boldsymbol{\theta}$, on peut calculer le quantile $C(T)$ associé à chaque valeur de l'échantillon *a posteriori* $\boldsymbol{\theta}$.

# Intervalle de confiance des quantiles : Bayes (2)

Pluie journalière à Selonnet et intervalles de confiance à 70 %




Commentaires sur la méthode de Taylor :

- facile à implémenter et automatiser :
- rapide
- ne marche pas que la variance est petite
- sinon il faut augmenter l'ordre du développement de Taylor

Commentaires sur la méthode bayésienne

- lenteur
- convergence à vérifier
- pas d'hypothèse sur le comportement des moments

# Comparaison et sélection de modèle

**Problématique** : on peut caler différentes lois sur le même jeu de données. Quel est le meilleur modèle ? Il existe plusieurs méthodes : critère d'information d'Akaike (AIC), facteur de Bayes, critère d'information bayésien (BIC), etc.

En hydrologie, on se sert du critère d'information d'Akaike $A$ :

$$A = 2k - 2\ln L,$$

avec $k$ le nombre de paramètres de la loi et $L$ la vraisemblance de la loi calée. Selon ce critère, le meilleur modèle est celui qui obtient le score $A$ le plus faible.

# Alternatives à la loi des valeurs extrêmes

**Historiquement** : importance de la loi de log-Pearson III dans le monde anglo-saxon (et en Suisse)

**Limites de la théorie des valeurs extrêmes** :

- il faut que la série de données soit stationnaire. Quid avec le réchauffement climatique ?
- il faut que les événements soient statistiquement caractérisés par la même loi de probabilité, or dans certains cas, les événements peuvent résulter de processus physiques différents et dès lors, ils sont issus de populations statistiques différentes.

# Loi de log-Pearson

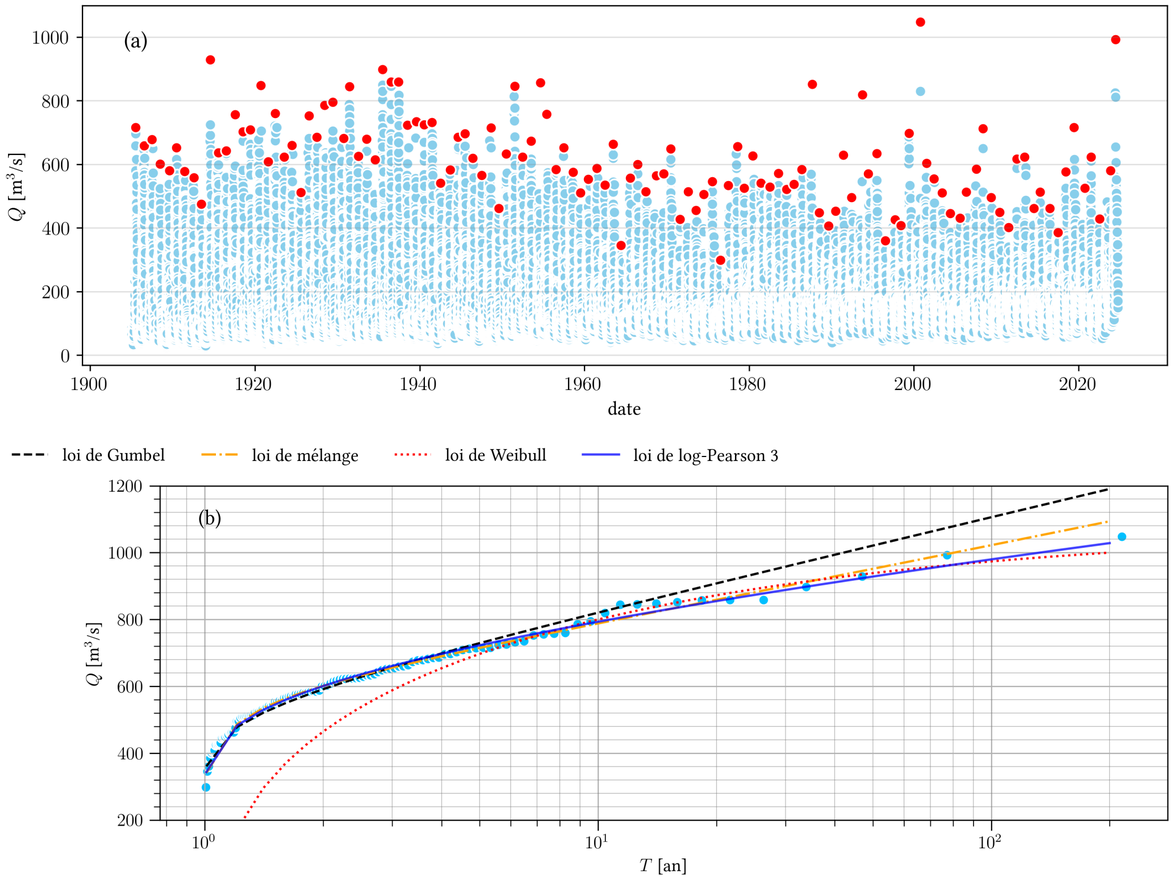
La loi de Pearson III est une loi de probabilité à trois paramètres $(m,\, \alpha,\, \lambda)$ :

$$g(x\ ;\ m,\, \alpha,\, \lambda) = \frac{|\alpha|}{\Gamma(\lambda)} \exp^{-\alpha(x-m)} (\alpha(x-m))^{\lambda-1} \text{ pour } x > m.$$

C'est une généralisation de la loi gamma. Lorsqu'une variable aléatoire $X$ est distribuée selon une loi de log-Pearson III, cela est équivalent à dire que $\ln X$ est distribué selon la loi de Pearson III. La densité de probabilité de $X$ est :

$$f(x\ ;\ m,\, \alpha,\, \lambda) = \frac{|\alpha|}{x\Gamma(\lambda)} \exp^{-\alpha(\ln x-m)} (\alpha(\ln x-m))^{\lambda-1} \text{ pour } x > e^m.$$

# Exemple de log-Pearson




Débits journaliers du Rhône à la Porte du Scex (Vouvry, VS) depuis le 1$^{er}$ janvier 1905. Les points rouges indiquent les maxi annuels. On a calé quatre lois de probabilité par la méthode du maximum de vraisemblance : loi de Gumbel, loi de Weibull, loi de log-Pearson III, et loi de mélange (combinant deux lois de Gumbel)

# Loi de mélange

Imaginons des pluies générées par deux mécanismes différents :

- type 1 : des dépressions atlantiques, dont l'intensité est décrite par une loi est $f_1$
- type 2 : des flux de sud (Méditerranée), dont l'intensité est décrite par une loi est $f_2$

alors on peut considérer que la densité de probabilité

$$f(x\ ;\ \theta) = \pi_1 f_1(x\ ;\ \theta_1) + \pi_2 f_2(x\ ;\ \theta_2)$$

avec les probabilité que l'événement soit de type 1 ou 2 vérifiant :

$$\pi_1 + \pi_2 = 1$$

Par ex., si $f_i$ sont des lois de Gumbel de paramètres $(\mu_i,\ \sigma_i)$, la loi de mélange est une loi à 5 paramètres ($\mu_1$, $\sigma_1$, $\mu_2$, $\sigma_2$, et $p = \pi_1$)

# Loi de mélange : calage

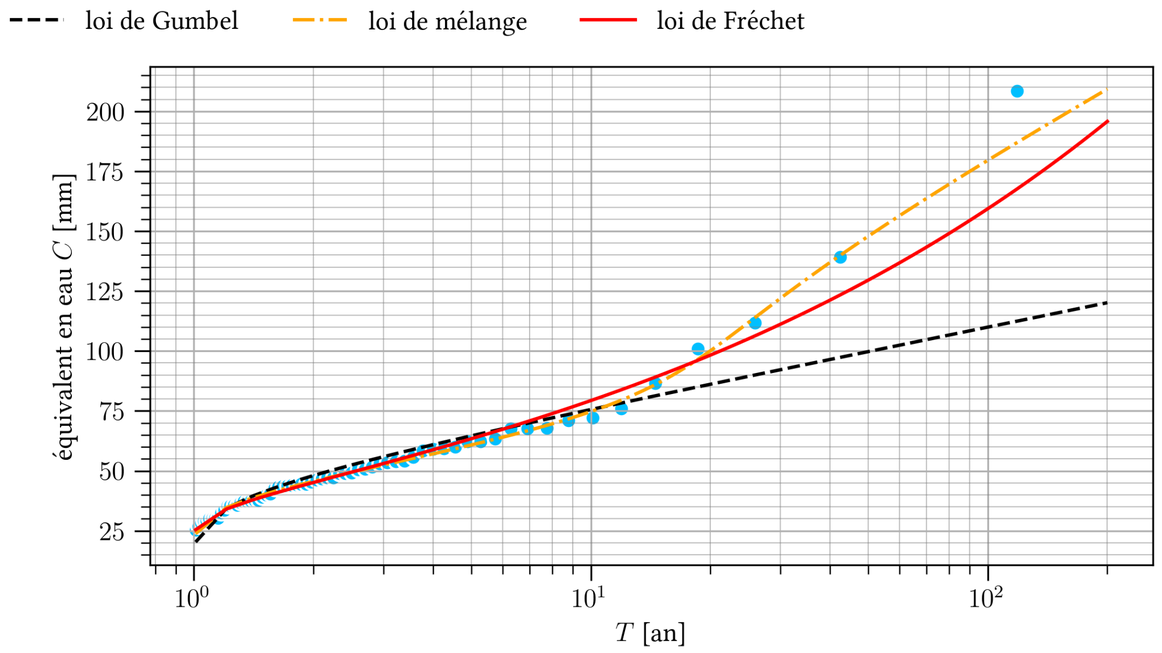
Techniques possibles : méthode des moments (marche mal) ou une version itérative de la méthode du maximum de vraisemblance *appelée espérance-maximisation* (EM)

**Principe de la méthode EM** :

- avoir une estimation de la variable latente (=non observée) $p = \pi_1$
- définir une vraisemblance dite complète comprenant les observations et les types de chaque observation (non connus)
- définir une vraisemblance d'observer un échantillon de valeurs conditionnellement à la connaissance du type de chaque observation
- réitérer le calcul

Voir § 4.6 dans les notes de cours

# Loi de mélange : exemple




Comparaison des lois de probabilité ajustées sur les maxima annuels des chutes de neige à Val-d'Isère (France) sur la période 1959–2023. Les données sont issues des données journalières du modèle Safran. La loi de mélange est la somme de deux lois de Gumbel.