Probabilités et Statistiques appliquées à l'Hydrologie

Benjamin RENARD

benjamin.renard@irstea.fr

Dernière mise à jour : 06/10/2014

Sommaire

1 Données, échantillons et Statistiques descriptives

Les chroniques de débit issues de stations hydrométriques constitueront les données de base utilisées dans ce document. Dans un souci de simplicité, la plupart des exemples utilisent des débits journaliers, mais les méthodes statistiques présentées s'appliquent également à des débits à d'autres pas de temps (horaire, mensuel ou pas de temps variable). De manière plus générale, elles s'appliquent également à d'autres variables hydro-météorologiques (pluies, températures, vent, etc.) et sont également utilisées dans de nombreux autres domaines (assurance, finance, etc.).

La plupart des analyses statistiques ne s'effectuent pas directement sur la chronique initiale, mais plutôt sur des séries de **variables hydrologiques** qui sont extraites, par exemple, de la chronique de débits journaliers : on parle de **l'échantillonnage** des variables hydrologiques. Ces variables ont pour but de caractériser de manière quantitative le phénomène étudié (les crues, les étiages, la ressource).

L'objectif de cette section est tout d'abord de présenter quelques approches permettant de définir des variables hydrologiques en basses, moyennes et hautes eaux. L'objectif de cette présentation n'est pas d'être exhaustif (les possibilités sont infinies…), mais plutôt d'illustrer la grande diversité de variables que l'hydrologue peut être amené à étudier. Dans un second temps, nous présenterons quelques méthodes de **statistiques descriptives** qui visent à résumer l'information contenue dans les séries de variables.

1.1 Définition de variables hydrologiques

En général, une variable hydrologique est calculée en deux étapes :

- Pré-traitement : Application d'un filtre à la chronique initiale. Par exemple, il est fréquent de travailler à partir de la série des débits moyennés sur *d* jours.
- Calcul de la variable sur la série filtrée. Par exemple, on pourra calculer le minimum annuel, le nombre de jours passés sous un seuil de bas débit, etc.

1.1.1 Pré-traitements

Le pas de temps de la chronique initiale n'est pas forcément adapté à l'étude du phénomène d'intérêt. Par exemple, pour les bassins à crue lente, le débit de pointe à un pas de temps fin n'est pas forcément la variable la plus pertinente. De même, si l'on s'intéresse aux écoulements liés à la fonte nivale, on pourra souhaiter « gommer » les variations rapides de débit dues à des faibles pluies se superposant au signal de fonte.

Ces considérations conduisent parfois à filtrer la chronique initiale avant le calcul des variables. Les quatre filtres présentés dans ce document sont les suivants :

- Moyenne mobile calculée sur une fenêtre de *d* jours.
- Minimum mobile calculé sur une fenêtre de *d* jours.
- *Base Flow Separation* (BFS), qui vise à séparer le débit de base du débit de ruissellement.
- *Sequent Peak Algorithm* (SPA), qui calcule un déficit de volume cumulé.

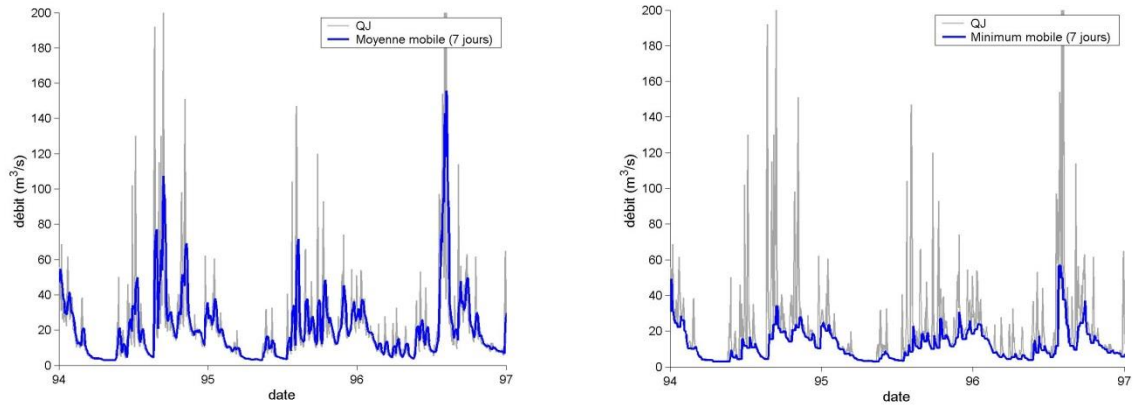
Les deux premiers filtres sont largement connus et sont utilisés en routine pour le calcul de débits caractéristiques. Les deux filtres suivants sont moins utilisés en pratique, mais sont présentés afin d'illustrer d'autres choix possibles pour le calcul de variables hydrologiques. Cette liste ne prétend évidemment pas être exhaustive, et d'autres pré-traitements existent ou peuvent être imaginés.

Dans le cas du filtrage par moyenne ou minimum mobile, la série filtrée S_t est obtenue à partir de la chronique journalière initiale q_t de la manière suivante :

$$\begin{aligned} &\text{Moyenne mobile sur } d \text{ jours: } S_t = \frac{1}{d}\sum_{k=t-d/2}^{t+d/2} q_k \\ &\text{Minimum mobile sur } d \text{ jours: } S_t = \min_{k\in[t-d/2;t+d/2]}\{q_k\} \end{aligned} \qquad (1\text{-}1)$$

Dans l'équation (1-1), la fenêtre de calcul est centrée sur le pas de temps courant *t*. Il est fréquent (notamment en statistiques) de calculer la moyenne ou le minimum sur une fenêtre [*t-d* ;*t*] précédant le pas de temps courant *t*. Ceci ne modifie que marginalement la série filtrée.

Ces deux filtres ont pour effet de lisser les hydrogrammes en limitant les variations de débit à l'échelle journalière (cf. exemples en Figure 1-1). Le débit moyen sur *d* jours est proportionnel à un volume, alors que le débit minimum sur *d* jours peut être interprété comme la valeur continûment dépassée sur *d* jours consécutifs. Ce dernier filtre est notamment utilisé pour l'étude des crues.



(a) (b)

Figure 1-1. Exemples de filtres appliqués à une chronique de débits journaliers. (a) Moyenne mobile sur 7 jours ; (b) minimum mobile sur 7 jours.

Les filtres BFS et SPA sont moins couramment utilisés pour le calcul de variables hydrologiques, et sont donc présentés plus en détail ci-dessous.

L'objectif du BFS est d'estimer le débit de base dans l'hydrogramme des débits journaliers. Plusieurs algorithmes existent dans la littérature. Nous présentons ici celui présenté par *Tallaksen and Van Lanen* (2004) :

- La série de débits journaliers est découpée en *n* blocs consécutifs (non-recouvrant) de *d* jours.
- Les valeurs minimales dans chaque bloc sont calculées, et sont notées $Qmin_1,\ldots Qmin_n$

- Pour $k = 2,\ldots,n\text{-}1$:
  - Si $w * Qmin_k < \min(Qmin_{k-1}, Qmin_{k+1})$, $Qmin_k$ est considéré comme un point pivot.
- La série filtrée est finalement obtenue en interpolant linéairement entre les points pivots.

L'algorithme BFS dépend de deux paramètres *d* et *w*. *Tallaksen and Van Lanen* (2004) recommandent des valeurs typiques égales à $d = 5$ jours et $w = 0.9$, qui donnent généralement des résultats très satisfaisants.

La Figure 1-2 présente un exemple d'application de ce filtre à une série de débits journaliers, pour un bassin à forte influence nivale (la Garonne à Montrejeau, 2100 km²). Ce filtre est particulièrement intéressant si l'on s'intéresse à la caractérisation de l'onde de fonte nivale, puisqu'il permet de gommer les variations de débit vraisemblablement dues à des pluies. Par exemple, le maximum annuel de la série filtrée par BSF est une variable plus robuste pour caractériser l'onde de fonte que le maximum annuel des débits bruts.

Signalons que le BFS est également à la base d'un indice couramment utilisé, nommé le *Base Flow Index* (BFI), définit comme le ratio entre le débit de base (estimé par l'algorithme BFS) et le débit total. Le BFI mesure ainsi la part des apports stockés (e.g. apports sous-terrains ou écoulements de fonte) par rapport aux débits issus du ruissellement rapide.

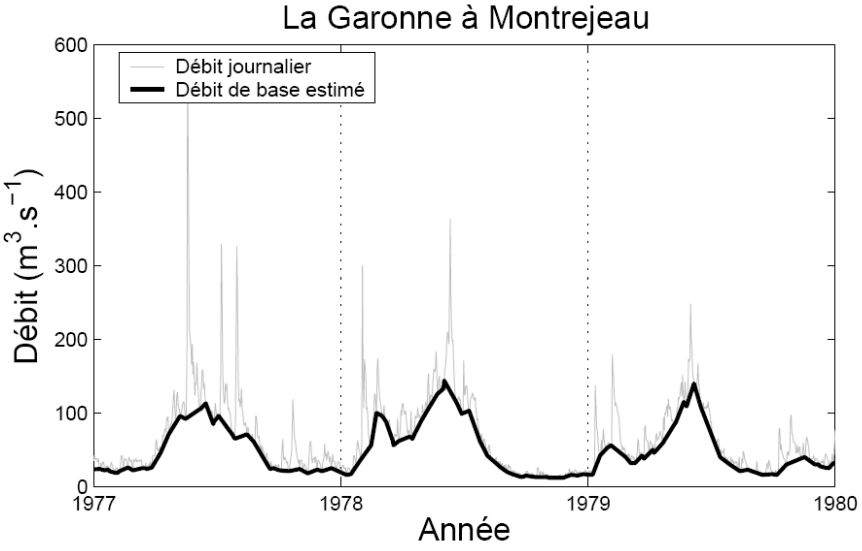


Figure 1-2. Application du filtre BFS pour un bassin à influence nivale : la Garonne à Montrejeau (2100 km²).

Le filtre SPA est quant à lui plutôt orienté vers l'étude des étiages. Selon *Tallaksen and Van Lanen* (2004), cet algorithme a été initialement développé pour le dimensionnement des réservoirs, mais il peut aussi être appliqué pour cumuler le déficit de volume d'une série de débits journaliers par rapport à un seuil de bas débit (que l'on appellera ici *qzero*).

Soient q_t le débit journalier et *qzero* un seuil de bas débit. La série filtrée par SPA est définie par :

$$S_t = \max(0; S_{t-1} + qzero - q_t) \qquad (1\text{-}2)$$

La Figure 1-3 illustre l'application du filtre SPA. Les périodes durant lesquelles la série filtrée est nulle correspond aux périodes de débits soutenus. Inversement, les événements d'étiages correspondent à des valeurs de SPA positives. Ce filtre possède donc la particularité intéressante d' « inverser » l'hydrogramme, les plus fortes valeurs correspondant aux étiages les plus sévères. L'inconvénient de ce filtre est sa sensibilité aux données manquantes, du fait de sa définition récursive.

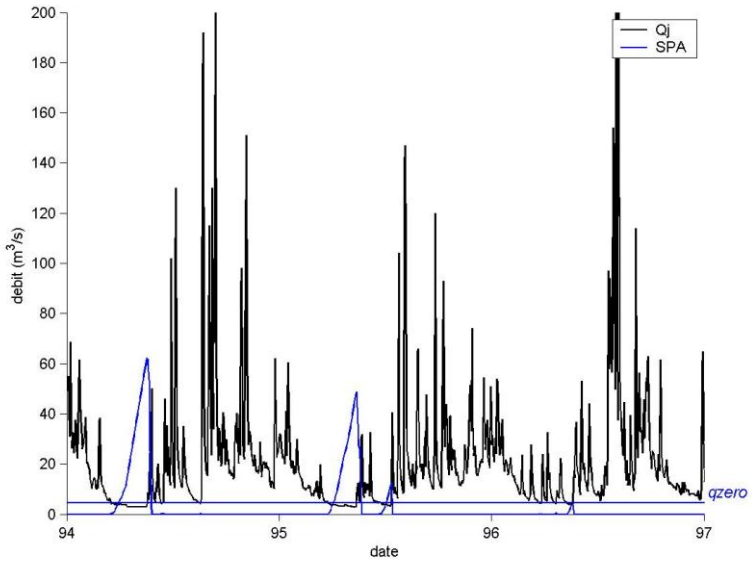


Figure 1-3. Exemple d'application du filtre SPA sur une chronique de débits journaliers.

En plus de l'utilisation de filtres, signalons également deux pré-traitements importants utilisés en pratique :

- Définition de **l'année hydrologique** : suivant le phénomène auquel on s'intéresse, l'utilisation de l'année civile n'est pas compatible avec la périodicité naturelle du phénomène. Typiquement, pour l'étude des crues pluviales, on définira l'année hydrologique comme débutant au premier septembre afin d'éviter de sectionner artificiellement la période hivernale des hautes eaux.
- **Saisonnalisation** : il arrive fréquemment de restreindre l'analyse à une période particulière de l'année. Par exemple, si l'on s'intéresse à la fonte nivale, on pourra se contenter d'étudier les débits printaniers et estivaux.

Dans la suite de cette section, ces pré-traitements seront omis par souci de simplicité : nous utiliserons systématiquement le terme « année », étant entendu que l'année pourra être remplacée par « l'année hydrologique » ou « la saison » suivant le phénomène étudié.

1.1.2 Variables pour les moyennes eaux

La variable la plus utilisée pour l'étude des moyennes eaux est le **module annuel** : il s'agit tout simplement de la moyenne des débits observés en une année donnée.

Il est également possible d'étudier des variables issues de la **courbe des débits classés** annuelle. Cette courbe est obtenue en classant, pour une année donnée, les débits par ordre décroissant (Figure 1-4). On peut alors lire sur la courbe la durée annuelle de dépassement d'un débit donné. Des variables descriptives des moyennes eaux peuvent être extraites de cette courbe en calculant, par exemple, le débit dépassé pendant la moitié de l'année (**débit**

annuel médian), ou plus généralement pendant *p*% de l'année (en choisissant une valeur de *p* pas trop éloignée de 50% si l'on s'intéresse aux moyennes eaux).

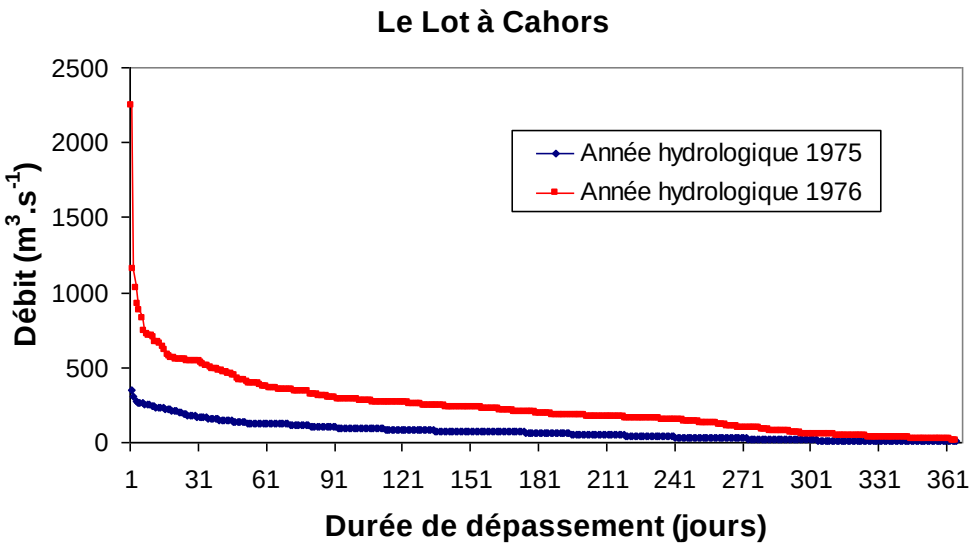


Figure 1-4. Exemples de courbes annuelles des débits classés pour le Lot à Cahors (9170 km²).

1.1.3 Hautes eaux

Nous allons nous intéresser aux deux principales techniques utilisées par les hydrologues pour l'étude des crues: l'échantillonnage par valeurs maximales annuelles (MAXAN) et l'échantillonnage par valeurs supérieures à un seuil (SUPSEUIL).

L'échantillonnage MAXAN consiste à sélectionner chaque année le débit observé le plus fort. L'échantillonnage SUPSEUIL consiste quant à lui à choisir un seuil de haut débit, puis à sélectionner les pointes des évènements dépassant ce seuil. Dans la pratique, plutôt qu'un seuil, on se donne en général un nombre d'évènements à sélectionner par an (en moyenne), et par itérations successives, on calcule le seuil conduisant à cet objectif.

Chacune de ces méthodes présente des avantages et des inconvénients, notamment dans l'optique d'une analyse statistique.

La mise en œuvre de l'approche MAXAN est très simple. En ne sélectionnant qu'un unique événement par an, on évite le risque de sélectionner deux valeurs de débits correspondant au même événement de crue (le risque de sélectionner, par exemple, un débit le 31 décembre 2000 et un autre le 1er janvier 2001 est limité par l'utilisation d'années hydrologiques). L'inconvénient est d'ignorer un certain nombre d'évènements lors des années où beaucoup de crues se sont produites (par exemple, année 1994-1995 dans la Figure 1-5), et inversement de prendre en compte des évènements peu importants lors des années peu actives (par exemple, année 1991-1992 dans la Figure 1-5). L'homogénéité de l'échantillon n'est donc pas optimale.

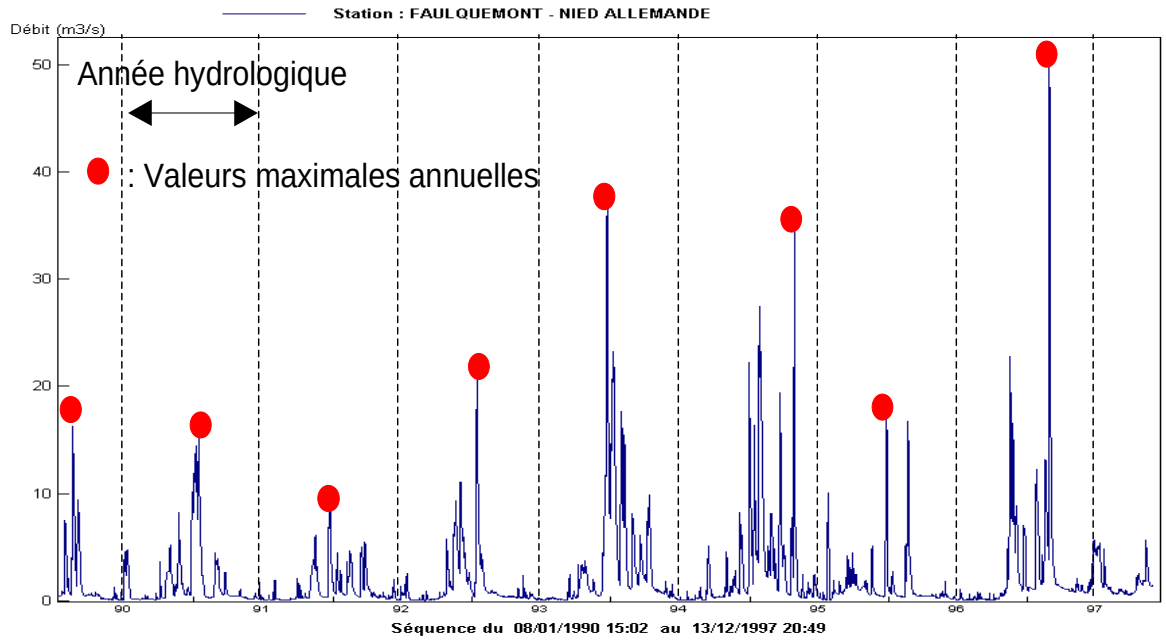


Figure 1-5. Illustration de l'échantillonnage par valeurs maximales annuelles pour la Nied Allemande à Faulquemont (187 km²).

L'approche SUPSEUIL est plus difficile à mettre en œuvre. En effet, il faut ajouter des contraintes d'indépendance afin de ne pas échantillonner plusieurs fois le même événement hydrologique (cf. exemple en Figure 1-7). On impose en général une contrainte d'espacement temporel minimal entre deux pointes sélectionnées, ainsi qu'une contrainte de redescente vers un débit de base. Bien choisies, ces contraintes permettent de garantir l'indépendance de l'échantillon (cf. Lang 1995 pour un état de l'art sur la technique d'échantillonnage SUP-SEUIL). Néanmoins, l'approche SUPSEUIL possède plusieurs avantages par rapport à l'échantillonnage MAXAN :

- Il est possible d'étoffer l'échantillon disponible pour l'analyse statistique en choisissant, en moyenne, plus d'un événement par an.
- Cet échantillon sera également plus homogène que celui fournit par la méthode MAXAN : ainsi, il sera possible de ne sélectionner aucune valeur pour les années peu actives (par exemple, année 1991-1992 dans la Figure 1-6), et inversement d'en sélectionner plusieurs pour les années ayant connu plusieurs événements importants (par exemple, année 1994-1995 dans la Figure 1-6).
- L'approche SUPSEUIL conduit à s'intéresser au processus d'occurrence des crues : on peut par exemple définir les variables « nombre d'événements sélectionnés chaque année », ou « durée entre deux événements successifs ».

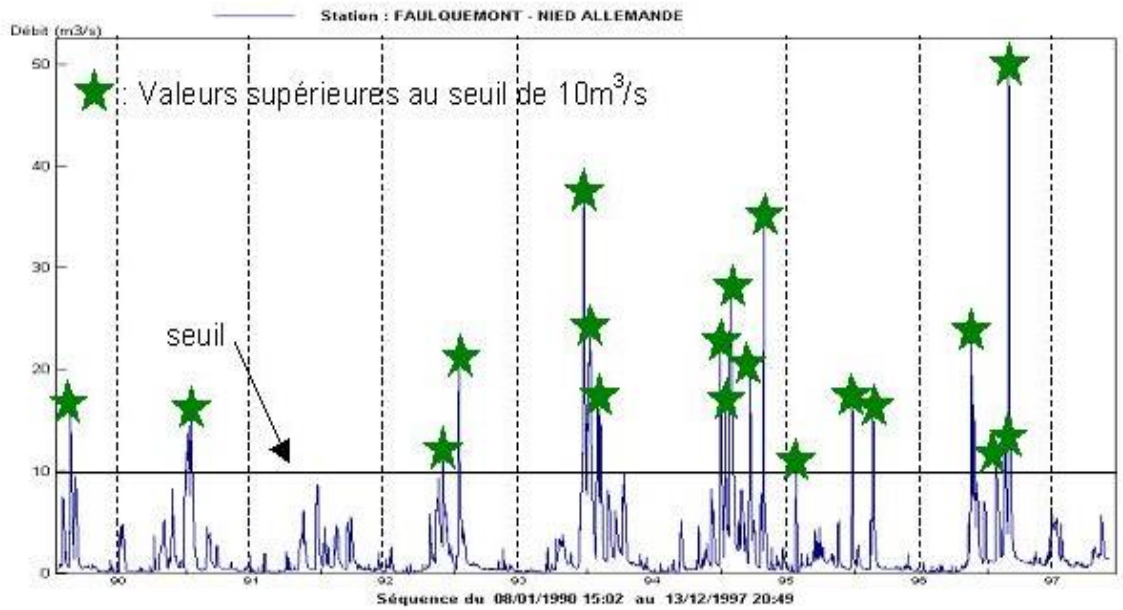


Figure 1-6. Illustration de l'échantillonnage par valeurs supérieures à un pour la Nied Allemande à Faulquemont (187 km²).

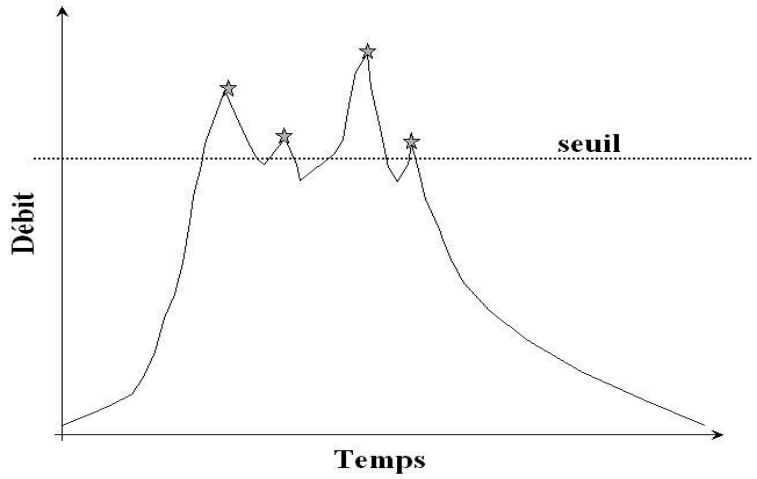


Figure 1-7. Illustration de la nécessité des contraintes d'indépendance pour l'échantillonnage SUPSEUIL.

Signalons que les approches MAXAN et SUPSEUIL peuvent s'appliquer sur des séries préalablement filtrées. On peut notamment appliquer les filtres moyenne / minimum mobile sur une durée *d* (avec par exemple *d* = durée caractéristique de crue). De plus, il existe de très nombreuses autres variables visant à caractériser les crues. Citons notamment :

- Variables issues de la courbe annuelle des débits classés, par exemple, débit dépassé 5% de l'année.
- Volume annuel cumulé au-dessus d'un seuil de haut débit.
- Variables décrivant les hydrogrammes de crue, par exemple, temps de montée / descente, asymétrie, aplatissement, etc.

1.1.4 Basses eaux

Les étiages sont des phénomènes physiques complexes qui ne peuvent pas être complètement décrits sur la base d'une unique valeur numérique. L'ouvrage de *Tallaksen and Van Lanen* (2004) fournit un panorama très complet des différents indices utilisables pour décrire les étiages. En général, les caractéristiques d'étiages les plus couramment étudiées sont la durée, le déficit de volume, et le pic (i.e. débits les plus bas).

Débit minimum annuel (MINAN)

Il correspond au pic de l'étiage, et constitue l'équivalent de l'approche MAXAN pour les basses eaux. Le débit minimum annuel est rarement calculé directement sur la chronique journalière (ou à pas de temps plus fin) initiale. On préfère généralement calculer le débit mensuel minimum (noté **QMNA**, correspondant au mois calendaire ayant le plus faible débit moyen), ou appliquer préalablement un filtre « moyenne mobile » sur *d* jours. La durée la plus fréquemment utilisée est de 30 jours, conduisant à la variable **VCN30**.

Variables relatives à un seuil de bas débit (INFSEUIL)

De manière similaire à l'approche SUPSEUIL utilisée pour l'étude des crues, on peut identifier des événements d'étiage comme des périodes durant lesquelles le débit est resté inférieur à un seuil donné. Néanmoins, contrairement à l'approche SUPSEUIL, il est très difficile de sélectionner plusieurs étiages indépendants au cours d'une même année. La Figure 1-8 illustre la définition de ces événements d'étiage : au cours d'une année hydrologique donnée, cinq événements sont identifiés dans cet exemple. Il est néanmoins discutable d'étudier ces cinq événements séparément : d'un point de vue hydrologique, il s'agit en fait du même événement d'étiage, entrecoupé par quelques remontées de débit dues à des pluies isolées. Pour cette raison, on préférera cumuler les caractéristiques de ces cinq événements sur l'année. Plus précisément, on peut calculer les variables suivantes:

- **Déficit de volume** : Il s'agit du déficit de volume par rapport au seuil de bas débit, cumulé sur la totalité de l'année hydrologique (aire en rouge sur la Figure 1-8).
- **Durée de l'étiage** : Elle est définie comme le nombre de jours où le débit est resté inférieur au seuil de bas débit (somme des durées des 5 événements d'étiage sur la Figure 1-8).

Variables basées sur la série filtrée par SPA

La série filtrée par SPA fournit un moyen naturel de définir des événements d'étiage et d'en déduire des variables caractéristiques. Par exemple, on pourra définir un événement d'étiage comme une période durant laquelle la série SPA est restée strictement positive. On peut alors aisément calculer la durée de l'événement, tandis que le maximum de la série SPA au cours de l'événement décrit le déficit de volume cumulé au pic d'étiage. Par rapport aux variables de durée et de déficit de volume présentées précédemment, l'avantage est que cette approche évite de sectionner un étiage qui s'étendrait sur plusieurs années, comme illustré en Figure 1-9 dans le cas d l'étiage de 1921.

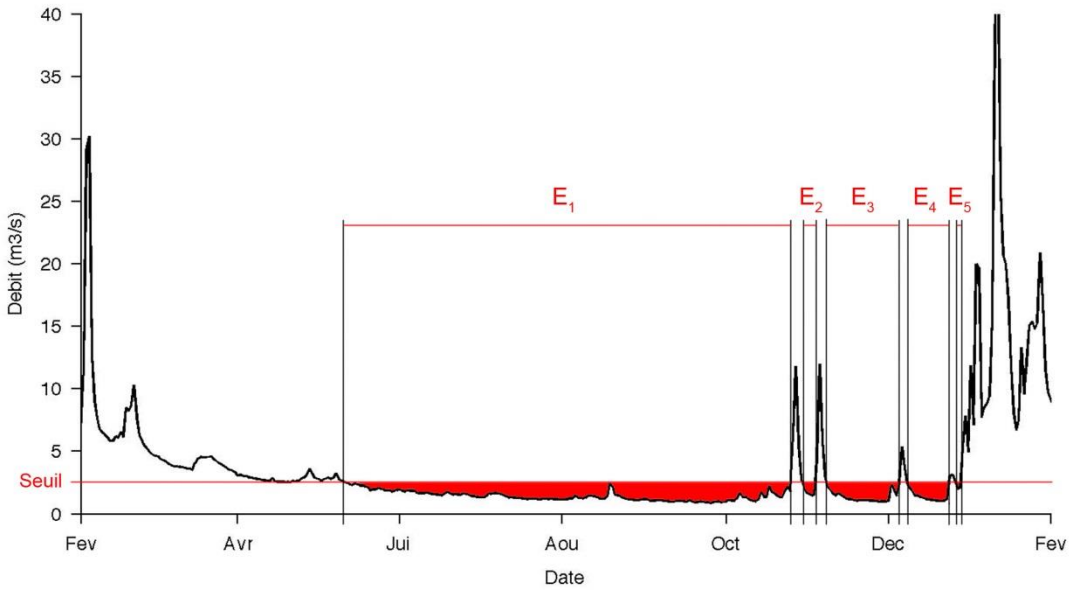


Figure 1-8. Identification des périodes d'étiages à partir d'un seuil de bas débit.

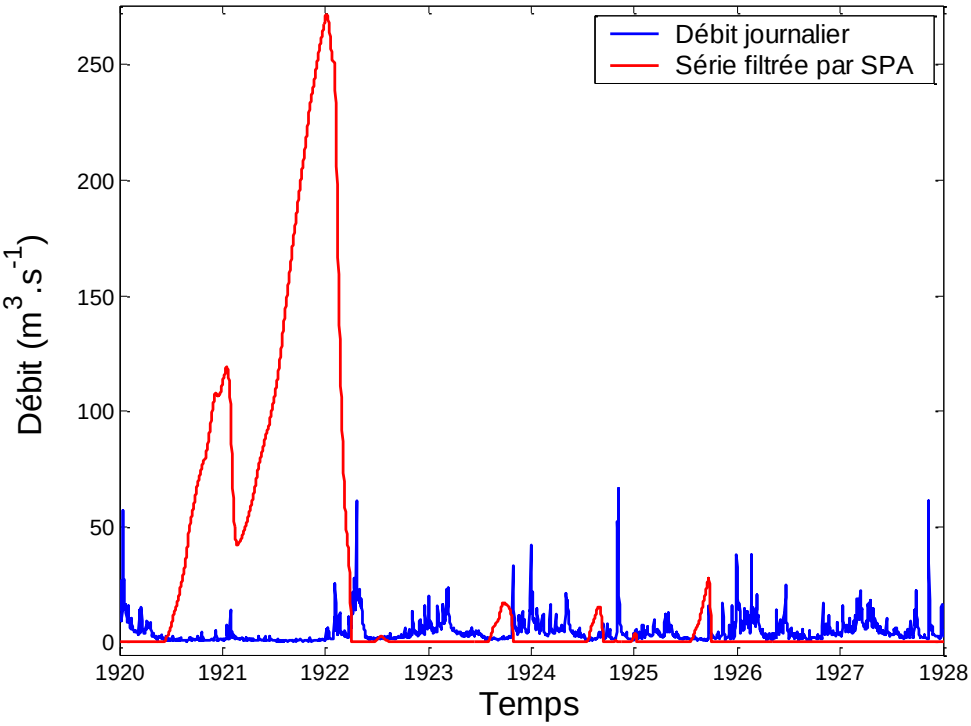


Figure 1-9. Illustration de l'intérêt du filtre SPA dans le cas d'un étiage pluriannuel (la Zorn à Waltenheim, 688 km²).

Concluons cette section en remarquant que la liste de variables d'étiage dressée dans ce document est loin d'être exhaustive. En particulier, on pourra encore une fois utiliser la courbe annuelle des débits classés pour extraire des variables descriptives de l'étiage (par exemple, débit dépassé 90% de l'année).

1.1.5 Variables de saisonnalité

Les variables présentées jusqu'ici sont des variables de quantité. On s'intéresse également parfois à la saisonnalité des événements hydrologiques. Par exemple, les dates de début et de fin d'étiage peuvent être importantes en termes de gestion.

Des variables de saisonnalité peuvent aisément être déduites des variables de quantité exposées précédemment, en utilisant par exemple la date du maximum ou du minimum annuel, les dates de dépassement de seuil, etc. D'autres variables peuvent être calculées en utilisant le concept de **centre de masse**. Ce concept a été introduit par *Stewart et al.* (2005) dans le cadre de l'étude de la saisonnalité de l'onde de fonte pour des cours d'eau à influence nivale. Dans cette publication, le centre de masse est défini comme la date à laquelle 50% du volume de fonte a été écoulé.

Cette définition peut aisément être généralisée d'une part pour l'étude d'autres types d'événements hydrologiques (crues ou étiages), d'autre part pour caractériser le début et la fin des événements (sans se limiter au « pic » de l'événement).

Dans le cas des étiages par exemple, on définira le centre de masse à *p%* comme la date à laquelle *p%* du déficit de volume annuel a été écoulé. En utilisant par exemple les centres de masse à 10%, 50% et 90%, on définit respectivement le début, le centre et la fin de la période d'étiage (Figure 1-10). On procèdera de même pour les crues, mais en considérant plutôt le volume écoulé au-dessus d'un seuil de haut débit.

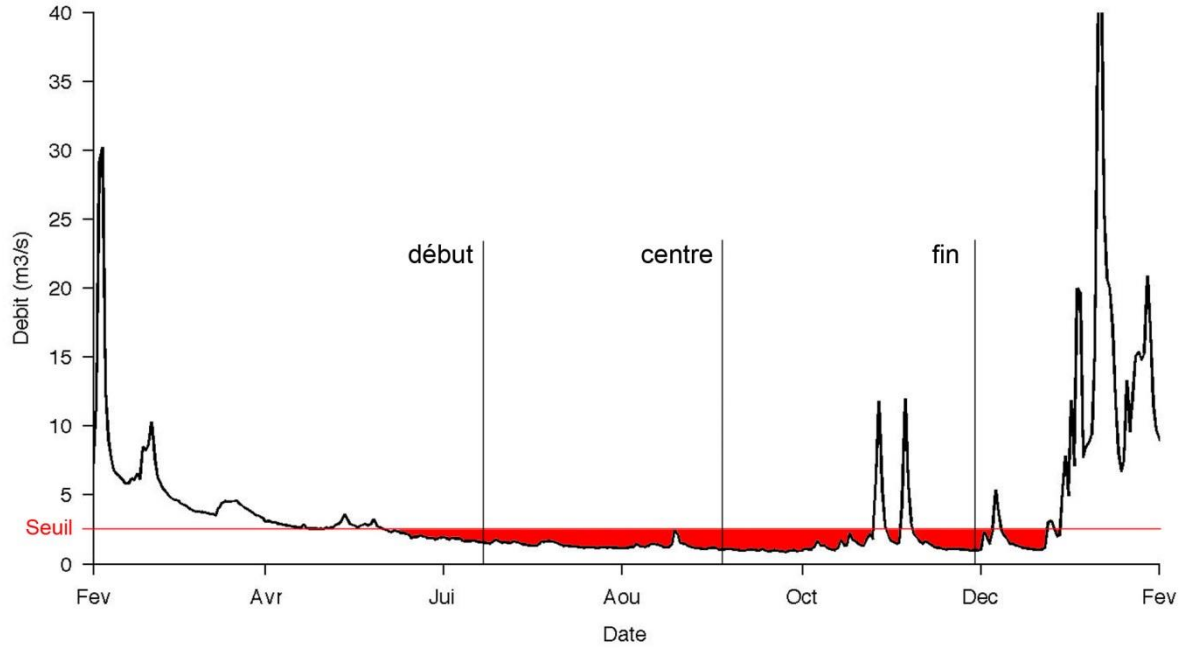


Figure 1-10. Exemple d'indices de saisonnalité : dates de début, centre et fin d'étiage.

1.1.6 Discussion : choix des variables hydrologiques

Dans certains cas le choix de la variable hydrologique à étudier est imposé. C'est le cas notamment pour les variables « réglementaires » (module ou QMNA pour les débits réservés ou les autorisations de prélèvement).

Si ce n'est pas le cas, le choix de la variable d'étude n'est ni anodin ni trivial. Par exemple, si l'on s'intéresse au régime des étiages, de nombreuses variables peuvent être calculées pour décrire la durée, l'intensité ou le déficit de volume des étiages (cf. section 1.1.4). Ainsi, ce que l'on appellera (de manière abusive car ambiguë) « l'étiage décennal », par exemple, correspondra à des phénomènes potentiellement bien différents suivant que l'on parle de

durée, de volume ou d'intensité. En conséquence, il peut s'avérer intéressant de décrire le phénomène qu l'on souhaite caractériser (crue ou étiages) par plusieurs variables hydrologiques.

Les principaux éléments guidant le choix du ou des variables à étudier sont les suivants :

- Objectifs de l'analyse. C'est certainement le premier aspect à évaluer : étant donnés les enjeux, quel aspect du phénomène hydrologique cherche-t-on à caractériser en priorité ? Par exemple, certains ouvrages seront vulnérables à des débits de pointe élevés, d'autre à des durées de submersion élevées.
- Propriétés du bassin versant à étudier. A titre d'illustration, un débit de pointe horaire est pertinent pour l'étude des crues d'un petit bassin cévenol, mais est moins adapté à l'étude des bassins à crues lentes comme la Somme par exemple, pour lesquels on cherchera plutôt à caractériser les crues en terme de volume ou de durée.
- Eviter la redondance d'information. D'un point de vue plus statistique, il n'est pas forcément utile de sélectionner plusieurs variables très corrélées. Ainsi, si l'on cherche à caractériser le phénomène d'étiage dans son ensemble, on cherchera à utiliser quelques variables qui présentent des niveaux de corrélation limités.

1.2 Description empirique des échantillons

Dans cette partie, nous allons nous intéresser à la description d'une série de données représentant une des variables hydrologiques présentées précédemment. L'objectif est notamment de présenter quelques graphiques classiques, et de résumer l'information contenue dans les données grâce à quelques grandeurs caractéristiques.

1.2.1 Représentations graphiques

Il existe une infinité de façons de représenter graphiquement un jeu de données : « camemberts », diagrammes en bâtons, courbes, nuages de point…. Tout dépend de la structure des données, mais également de l'information que l'on souhaite faire passer. Nous allons nous intéresser ici à deux représentations fondamentales pour les données quantitatives : l'**histogramme** et la **courbe des fréquences cumulées**.

L'histogramme consiste à représenter la **fréquence** (ou parfois **l'effectif**) de chaque valeur présente dans l'échantillon. Il est également possible de cumuler ces fréquences par ordre croissant des valeurs rencontrées : on peut ainsi tracer **la courbe des fréquences cumulées**, qui représente, à chaque abscisse, la proportion de valeurs de l'échantillon inférieures ou égales à cette abscisse.

Ces deux représentations sont primordiales, car elles peuvent être reliées aux notions probabilistes de densité et de fonction de répartition, dont nous parlerons prochainement.

Exemple 1 : Nous débutons par un exemple utilisant une variable hydrologique à valeurs entières. Le Tableau 1-1 donne le nombre annuel de crues supérieures à 24.7 $m^3.s^{-1}$ pour L'Ubaye à Barcelonnette (549 km²).

Année	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003
Nb	1	1	3	1	5	2	1	3	3	5	0	3	5	2

Tableau 1-1. Tableau de données : nombre de crues supérieures à 24.7 $m^3.s^{-1}$.

Les questions qui pourraient être posées à l'hydrologue sont du type : « Quelle est la probabilité de ne pas observer de crue au cours de l'année ? » ; « Quelle est la probabilité d'observer 6 crues ou plus au cours d'une même année ? » ; « Quel nombre annuel de crue a une probabilité 1/10 d'être dépassé ? ».

Avant de tenter de répondre à ces questions, la première représentation graphique consiste généralement à tracer la série chronologique des données à analyser (Figure 1-11).

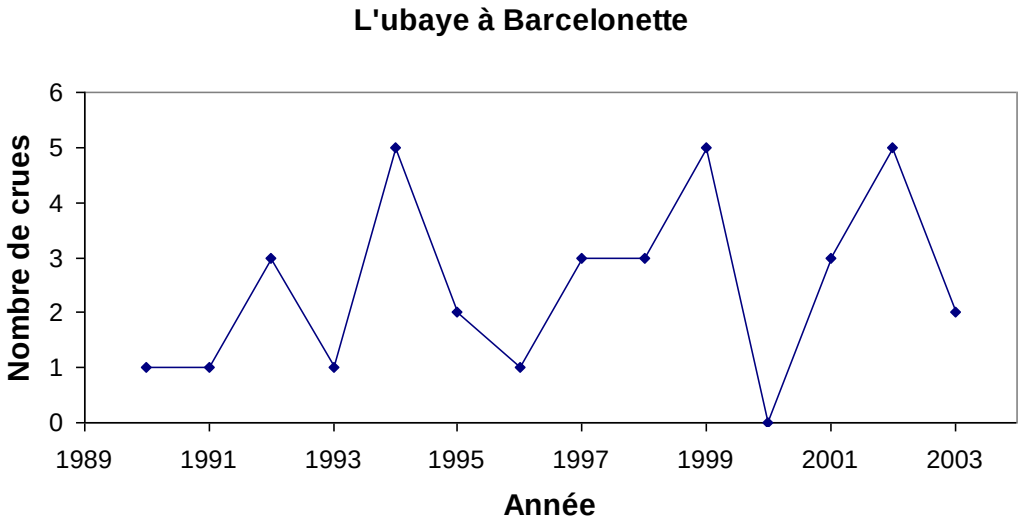


Figure 1-11. Série chronologique.

Dans un second temps, on s'intéresse aux fréquences d'observations des différentes valeurs. Pour cela, on calcule **l'effectif**, la **fréquence** et la **fréquence cumulée** associés à chaque valeur possible :

- L'effectif d'une valeur donnée est le nombre de fois où cette valeur a été observée. Par exemple, la valeur « 0 » est d'effectif 1.
- La fréquence est égale à l'effectif divisé par la taille de la série analysée. Ici, la fréquence de la valeur « 0 » vaut $1/14 \approx 0.07$.
- La fréquence cumulée est égale au cumul des fréquences associées à chaque valeur, lorsque ces valeurs sont classées par ordre croissant (et non par ordre chronologique !). Par exemple, fréquence(0) $\approx$ 0.07, fréquence(1) $\approx$ 0.29, donc fréquence cumulée(1) $\approx 0.07 + 0.29 \approx 0.36$

La notion de fréquence cumulée est extrêmement importante puisque nous verrons ultérieurement que les notions de quantiles et de période de retour y sont étroitement liées. En hydrologie, la fréquence cumulée est souvent appelée **fréquence au non-dépassement**: en effet, la fréquence cumulée d'une valeur est égale au pourcentage de données dans l'échantillon qui sont inférieures ou égales à cette valeur. Notons qu'il peut sembler arbitraire d'utiliser « inférieur ou égal » plutôt qu' « inférieur strict » dans la définition de la fréquence cumulée : nous reviendrons sur ce point ultérieurement.

Le calcul de **l'effectif**, de la **fréquence** et de la **fréquence cumulée** pour chaque valeur possible conduit au Tableau 1-2 ci-dessous.

Valeur	Effectif	Fréquence	Fréq. cumulée
0	1	0.07	0.07
1	4	0.29	0.36
2	2	0.14	0.5
3	4	0.29	0.79
4	0	0.00	0.79
5	3	0.21	1

Tableau 1-2. Effectifs, fréquences et fréquences cumulées.

L'histogramme des données (Figure 1-12) est alors obtenu en reportant en abscisses les valeurs possibles (par ordre croissant) et en ordonnées les fréquences (ou parfois les effectifs) correspondants. La courbe des fréquences cumulées (Figure 1-13) est obtenue en reportant en abscisses les valeurs possibles (par ordre croissant) et en ordonnées les fréquences cumulées.

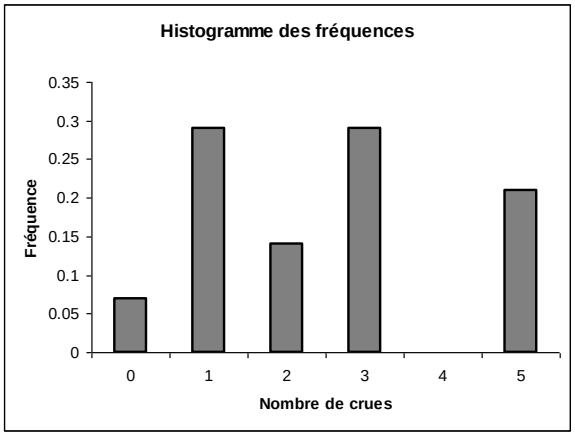


Figure 1-12. Histogramme des fréquences.

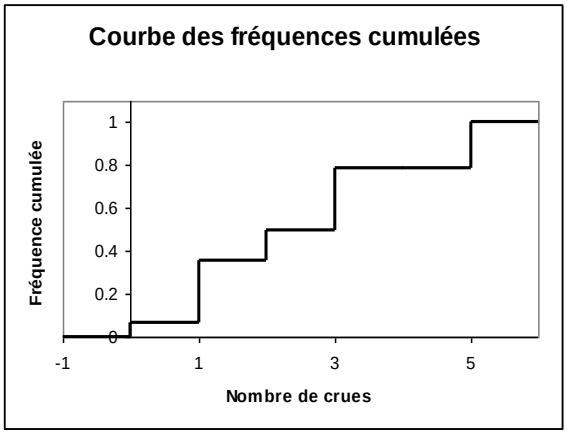


Figure 1-13. Courbe des fréquences cumulées.

Ces premiers calculs et graphiques élémentaires permettent d'apporter un premier élément de réponse aux questions posées à l'hydrologue, mais illustrent également les limites de cet exercice :

- « Quelle est la probabilité de ne pas observer de crue au cours de l'année ? » : la fréquence observée de la valeur « 0 » valant environ 0.07, on serait tenté de répondre que cette probabilité vaut environ 0.07.
- « Quelle est la probabilité d'observer 6 crues ou plus au cours d'une même année ? » : Au cours des 14 années d'observation, on n'a jamais observé plus de cinq crues. On serait donc tenté de répondre que la probabilité d'en observer 6 ou plus est nulle. Néanmoins, on peut légitimement douter de cette estimation : un événement qui n'a jamais été observé est-il pour autant impossible ?
- « Quel nombre annuel de crue a une probabilité 1/10 d'être dépassé ? » : on peut théoriquement répondre à cette question en recherchant la valeur dont la fréquence au dépassement vaut 0.1, ou de manière équivalente dont la fréquence au non-dépassement vaut 1-0.1 = 0.9. En observant le Tableau 1-2, on constate que la fréquence cumulée « 0.9 » n'est jamais atteinte : la valeur « 4 crues par an » a une fréquence cumulée égale à 0.79 (<0.9), tandis que la valeur « 5 crues par an » a une fréquence cumulée égale à 1 (>0.9). On serait donc tenté de répondre que le nombre annuel de crue ayant une probabilité 1/10 d'être dépassé vaut entre 4 et 5.

Exemple 2 : Ce second exemple correspond au cas plus courant d'une variable hydrologique continue, en l'occurrence le débit journalier maximum annuel (cf. section 1.1.3) pour l'Ariège à Foix (1340 km²). Les données sont fournies dans le Tableau 1-3, le tracé chronologique de cette série de valeurs est donné en Figure 1-14.

Année	1981	1982	1983	1984	1985	1986	1987	1988	1989	1990
MAXAN ($m^3.s^{-1}$)	252	412	126	167	121	129	174	88.5	111	162
Année	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000
MAXAN ($m^3.s^{-1}$)	178	264	146	116	146	342	82	162	218	101

Tableau 1-3. Maxima annuels de l'Ariège à Foix (1340 km²).

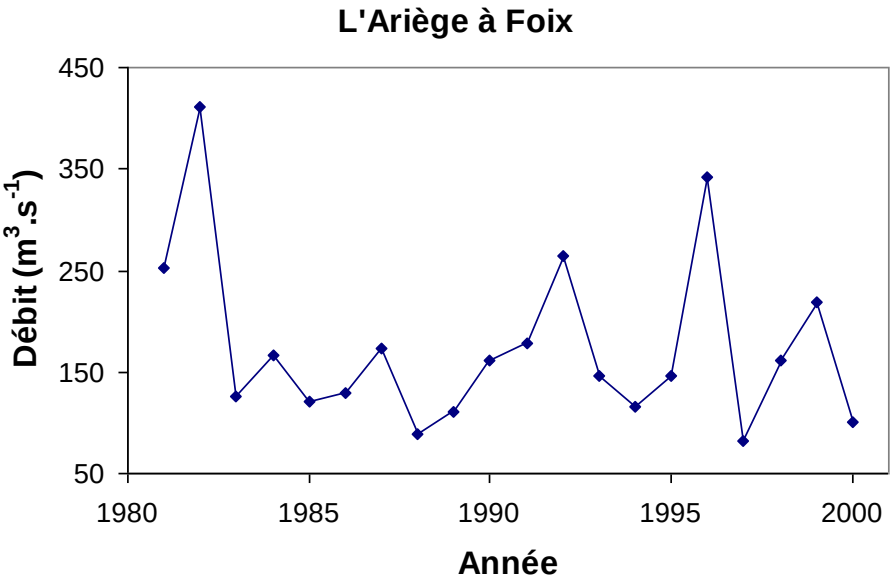


Figure 1-14. Série chronologique des maxima annuels.

Comme précédemment, les questions d'intérêt pourraient être « Quelle est la probabilité d'observer une crue supérieure à 300 $m^3.s^{-1}$? à 450 $m^3.s^{-1}$? », ou bien « Quel débit de crue a une probabilité 1/10 d'être dépassé ? une probabilité 1/100 ? ». Pour fournir des éléments de réponse, une approche similaire à celle présentée dans l'exemple précédent est utilisée, et conduit au Tableau 1-4.

Valeur	Effectif	Fréquence	Fréquence cumulée
82	1	0.05	0.05
88.5	1	0.05	0.1
101	1	0.05	0.15
111	1	0.05	0.2
116	1	0.05	0.25
121	1	0.05	0.3
126	1	0.05	0.35
129	1	0.05	0.4
146	1	0.05	0.45
146.5	1	0.05	0.5
162	1	0.05	0.55
163	1	0.05	0.6
167	1	0.05	0.65
174	1	0.05	0.7
178	1	0.05	0.75
218	1	0.05	0.8
252	1	0.05	0.85
264	1	0.05	0.9
342	1	0.05	0.95
412	1	0.05	1

Tableau 1-4. Effectifs, fréquences et fréquences cumulées.

On remarque dans le Tableau 1-4 que chaque valeur n'apparaît qu'une seule fois. Ceci ne pose pas de problème pour le tracé de la courbe des fréquences cumulées (Figure 1-15), la fréquence au non-dépassement de la i^{eme} valeur (par ordre croissant) étant alors égale à i/n, où n est la taille de l'échantillon (égale à 20 ici). Par contre, un histogramme dont toutes les ordonnées seraient égales à 1/20 ne serait pas très informatif… Pour y remédier, on effectue un regroupement des valeurs en classes, ce qui revient à « discrétiser » la variable continue. Le choix des classes reste arbitraire ; on essaiera en général de créer entre 5 et 15 classes de même étendue. L'histogramme des fréquences pour des classes d'étendue 80 m^3s^{-1}, en partant de 10 m^3s^{-1}, est donné en Figure 1-16. On peut noter la différence de représentation par rapport au précédent histogramme (Figure 1-12) : on a représenté des rectangles pleins pour bien signifier qu'il s'agit de la fréquence d'une classe, et non d'une valeur ponctuelle discrète. Pour être tout à fait précis, seule la Figure 1-16 est un histogramme, le terme de « diagramme en bâtons » est plutôt utilisé pour le cas discret (Figure 1-12).

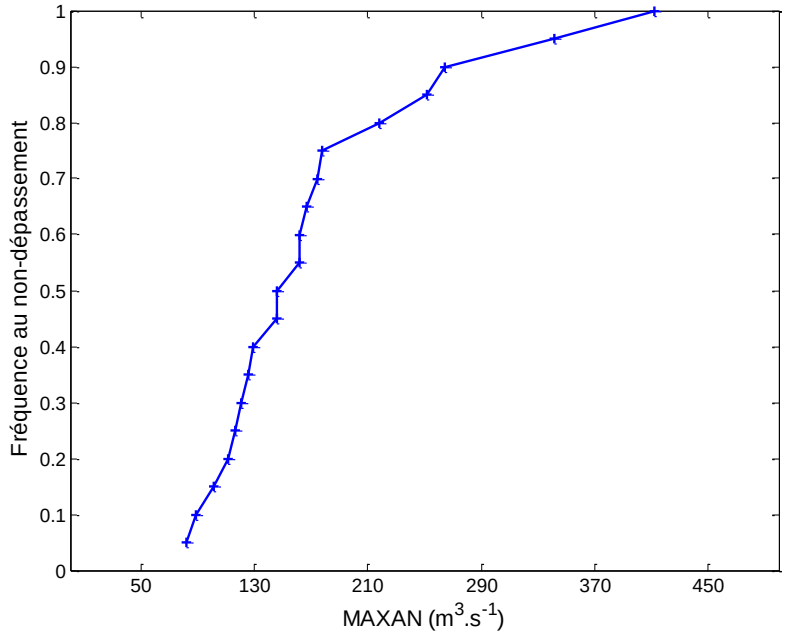


Figure 1-15. Courbe des fréquences cumulées.

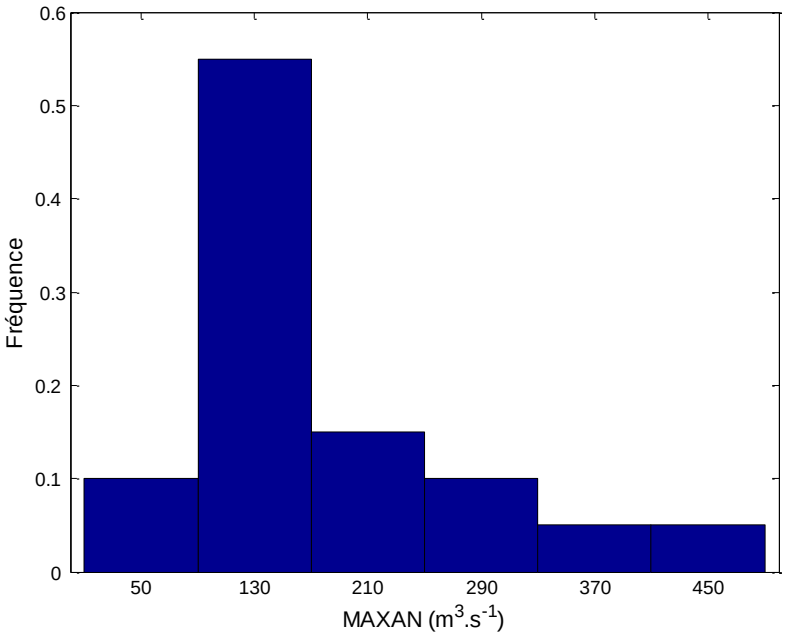


Figure 1-16. Histogramme des fréquences.

On peut alors apporter des éléments de réponse aux questions posées à l'hydrologue :

- « Quel débit de crue a une probabilité 1/10 d'être dépassé ? » : on recherche la valeur ayant une fréquence au non-dépassement égale à 1-1/10=0.9. On trouve 264 $m^3.s^{-1}$, ce qui fournit une première estimation.
- « Quel débit de crue a une probabilité 1/100 d'être dépassé ? » : on recherche la valeur ayant une fréquence au non-dépassement égale à 1-1/100=0.99. On peut alors donner la fourchette 342-412 $m^3.s^{-1}$.

- « Quelle est la probabilité d'observer une crue supérieure à 300 $m^3.s^{-1}$? » : en se basant sur la Figure 1-15, on trouve une fréquence au non-dépassement d'environ 0.92. La probabilité recherchée est donc de l'ordre de 0.08
- « Quelle est la probabilité d'observer une crue supérieure à 450 $m^3.s^{-1}$? » : en se basant sur la Figure 1-15, on serait tenté de répondre « zéro » puisque la valeur observée maximale est de 412 $m^3.s^{-1}$… mais on sent bien qu'il s'agit là d'une des limitations de l'approche basée sur la seule description empirique des données : il faudrait se doter d'un outil pour pouvoir **extrapoler** au-delà de la valeur la plus forte. Nous y reviendrons dans la suite de ce document.

1.2.2 Indicateurs de localisation

En plus des représentations graphiques présentées précédemment, on souhaite généralement résumer les données analysées sur la base de quelques indicateurs quantitatifs, renseignant sur l'ordre de grandeur des observations, leur variabilité, etc. Nous nous intéressons dans un premier temps aux indicateurs dits de localisation.

Notons $(x_i)_{i=1,\dots,n}$ les données observées que l'on souhaite analyser. Le premier indicateur calculé est généralement la **moyenne**, que nous noterons $\overline{x}$.

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad (1\text{-}3)$$

La médiane est la valeur qui sépare l'échantillon ordonné en deux sous-parties de même effectif. Par exemple, la médiane des valeurs 10, 15, 18, 19, 20 vaut 18. Si le nombre de valeurs est pair, on calculera la demi-somme entre les deux valeurs du milieu, soit une médiane de 16.5 pour l'échantillon des quatre premières valeurs ci-dessus. La médiane peut également être lue sur la courbe des fréquences cumulées : c'est la valeur correspondant à une fréquence cumulée de 0.5.

Notons que par rapport à la moyenne, la médiane est beaucoup moins sensible à la présence de valeurs extrêmes dans l'échantillon. Par exemple, si l'on remplace la valeur « 20 » par « 2000 » dans l'exemple ci-dessus, la médiane est inchangée, alors que la moyenne sera évidemment très différente. L'exemple classique pour illustrer la différence entre la moyenne et la médiane est le salaire des Français : la moyenne vaut environ 2000 €, la médiane 1600 €… (salaires nets en 2007 dans le secteur privé, source INSEE).

Un troisième indicateur est parfois utilisé, surtout pour les données discrètes, il s'agit du **mode**, qui est la valeur (pas forcément unique) la plus fréquente dans l'échantillon. Pour des données quantitatives continues, il faut (comme pour l'histogramme) procéder à un regroupement des individus. L'inconvénient est que le mode devient dépendant du regroupement arbitraire effectué.

Le dernier indicateur est particulièrement important dans le domaine de l'hydrologie : le **quantile** d'ordre ***p*** est la valeur de l'échantillon dont la fréquence cumulée vaut *p*. Le quantile peut donc être lu sur la courbe des fréquences cumulées, en faisant éventuellement une interpolation linéaire entre deux points. La médiane est ainsi le quantile d'ordre 0.5.

Notons enfin que l'on reporte fréquemment le **minimum** et le **maximum** d'un échantillon.

1.2.3 Indicateurs de dispersion

Les indicateurs de dispersion ont pour but de décrire la variabilité des données.

La **variance** permet de mesurer la façon dont les données se dispersent autour de la moyenne :

$$Var_x = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2 \tag{1-4}$$

On trouve parfois (notamment dans les calculettes ou les tableurs type Excel) une autre définition, qui n'est pas recommandée pour les petits échantillons :

$$Var_x^* = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2 \tag{1-5}$$

Ces deux définitions deviennent équivalentes lorsque *n* est grand.

On définit également **l'écart-type**, qui a l'avantage d'avoir la même dimension que les données :

$$s_x = \sqrt{Var_x} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2} \tag{1-6}$$

Le **coefficient de variation** est également utilisé pour comparer la variabilité de plusieurs séries de données dont les ordres de grandeurs ne sont pas comparables, mais n'est évidemment pas défini pour les données dont la moyenne est nulle. Il est généralement exprimé en pourcentage.

$$CV_x = \frac{s_x}{\bar{x}} \tag{1-7}$$

Enfin, **l'étendue** est également parfois utilisée.

$$Etendue_x = \max_{i=1,\ldots,n}\{x_i\} - \min_{i=1,\ldots,n}\{x_i\} \tag{1-8}$$

1.2.4 Autres indicateurs

Mentionnons finalement quelques indicateurs supplémentaires, qui n'ont pas de signification aussi claire que les indicateurs de localisation ou de dispersion, mais qui sont importants en Statistiques, notamment dans le cadre de la théorie de l'estimation.

Le **moment d'ordre k** est défini de la manière suivante :

$$m_k = \frac{1}{n}\sum_{i=1}^{n} {x_i}^k \tag{1-9}$$

Pour *k=1*, on reconnaît la moyenne. On définit de même le **moment centré d'ordre k** :

$$m_k' = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^k \tag{1-10}$$

On peut déduire de ces moments centrés deux indicateurs relatifs à la forme de l'histogramme :

$$\textbf{Asymétrie} \quad asym_x = \frac{m_3^{'}}{s_x^3} \tag{1-11}$$

$$\textbf{Aplatissement} \; apla_x = \frac{m_4^{'}}{s_x^4} \tag{1-12}$$

Comme son nom l'indique, l'asymétrie décrit le caractère symétrique de l'histogramme (l'asymétrie est nulle pour des données distribuées symétriquement). L'aplatissement (également appelé parfois **kurtosis**) décrit quant à lui le caractère plus ou moins « pointu » de l'histogramme.

En guise d'illustration, le Tableau 1-5 fournit un résumé statistique pour les données de l'Ariège à Foix étudiées en section 1.2.1. Fournir ce type de résumé lors de toute analyse statistique est une bonne pratique.

MAXAN - L'Ariège à Foix	
Moyenne	174.95
Médiane	154.25
Quantile 25%	116.00
Quantile 75%	178.00
Minimum	82.00
Maximum	412.00
Écart-type	85.35
Coefficient de variation	49%
Variance de l'échantillon	7284.97
Etendue	330.00
Kurtosis (Coefficient d'aplatissement)	2.24
Coefficient d'asymétrie	1.55
Effectif de l'échantillon	20

Tableau 1-5. Résumé statistique des valeurs maximales annuelles pour l'Ariège à Foix.

2 Principales distributions

Nous avons exposé dans la section 1.2 précédente un certain nombre de méthodes permettant de décrire une série d'observations et d'en extraire de l'information. Dans cette section, nous allons tenter d'aller au-delà d'une simple description empirique des données, en utilisant des **modèles probabilistes** pour décrire la **distribution** des observations.

La théorie des probabilités fournit un cadre général pour décrire les caractéristiques de variables « génératrices » de données. La puissance de cette théorie tient à sa généralité. En contrepartie, la définition même du terme « probabilité » fait appel à des notions mathématiques trop élaborées pour être présentées dans ce cours si l'on souhaite être totalement rigoureux du point de vue mathématique. Nous nous contenterons donc d'étudier une version « allégée » de la théorie probabiliste, qui sera néanmoins amplement suffisante en pratique.

Cette seconde section est organisée de la manière suivante. Dans un premier temps, nous présenterons les principales motivations conduisant à s'intéresser au cadre probabiliste (section 2.1). Les notions de bases seront ensuite définies (section 2.2), avant de présenter plus en détail quelques distributions fréquemment utilisées en hydrologie : la loi de Gauss (section 2.3), la loi de Galton (section 2.4), et les lois de valeurs extrêmes (section 2.5). Nous conclurons par une rapide présentation de quelques méthodes pour l'estimation des paramètres de ces lois (section 2.7).

2.1 Motivations

Les approches empiriques présentées en section 1.2 permettaient déjà d'apporter des éléments de réponse aux questions fréquemment posées à l'hydrologue (par exemple, « Quelle valeur de débit a une probabilité 1% d'être dépassée chaque année ? » ; « Quelle est la probabilité de dépassement de la crue de décembre 2003 ? »). On peut alors s'interroger sur l'intérêt de recourir à des concepts mathématiques supplémentaires. Nous expliquons ici pourquoi il n'est pas raisonnable de se contenter de l'approche descriptive de la section 1.2.

2.1.1 Quelle formule pour la fréquence au non-dépassement ?

Nous avons déjà mentionné en section 1.2.1 le caractère arbitraire de la formule utilisée pour la fréquence au non-dépassement. Par définition, la fréquence au non-dépassement de la i^{eme} valeur de l'échantillon (lorsque celui-ci est classé par ordre croissant, et non dans l'ordre chronologique) est égale à i/n, où n est le nombre d'observations. Ceci correspond à la définition suivante : « La fréquence au non-dépassement d'une valeur donnée est égale à la fréquence des observations *inférieures ou égales* à cette valeur ». Mais si l'on remplace « inférieur ou égal » par « inférieur strict », la formule devient alors $(i-1)/n$. En plus de ces deux formules naturelles, la littérature regorge d'autres formules pour le calcul des fréquences au non-dépassement, comme par exemple $(i-0.5)/n$, $i/(n+1)$, etc. Il serait donc souhaitable de se donner un cadre méthodologique s'affranchissant de ce choix largement arbitraire.

2.1.2 Interpolation et extrapolation

Les exemples donnés en section 1.2.1 ont permis d'illustrer les limites d'une approche purement empirique pour **l'extrapolation**. Ainsi, si l'on se basait uniquement sur les fréquences au non-dépassement empiriques (avec la formule i/n), on conclurait que la plus forte valeur observée ne peut pas être dépassée ! L'utilisation d'une autre formule de fréquence empirique ne résoudrait que partiellement le problème, puisque l'extrapolation au-

delà de la valeur la plus forte resterait problématique. Mentionnons également que **l'interpolation** entre deux fréquences empiriques par de segments de droite n'est guère satisfaisante.

2.1.3 Variabilité d'échantillonnage et incertitudes

La plus forte motivation conduisant à l'utilisation de modèles probabilistes est liée à la **variabilité d'échantillonnage**. Ce terme générique désigne le fait que les observations de débit dont nous disposons ne constituent généralement qu'un faible échantillon de l'ensemble des réalisations possibles de la rivière. En conséquence, les estimations seront dépendantes des données effectivement observées.

Cette variabilité d'échantillonnage peut être illustrée en considérant les 20 maxima annuels présentés en Figure 2-1. On constate que deux événements forts ont été observés lors des dix premières années. La Figure 2-2 illustre l'impact de ces deux valeurs sur l'allure des courbes de fréquence cumulée calculées sur les périodes 1951-1960 et 1961-1970. Ainsi, si l'hydrologue doit répondre à la question « quel débit de crue a une probabilité au non-dépassement de 0.9 ? », une approche purement empirique le conduira à répondre « environ 3500 $m^3.s^{-1}$ » en utilisant les données de la première période, mais « environ 5000 $m^3.s^{-1}$ » sur la base de la seconde !

Précisons que l'utilisation de modèles probabilistes ne permet pas de s'affranchir totalement de cette difficulté : en effet, leur estimation est également soumise à la variabilité d'échantillonnage, mais dans une bien moindre mesure. De plus, l'approche probabiliste possède l'avantage majeur de permettre la quantification de cette variabilité, et des **incertitudes** qui en résultent.

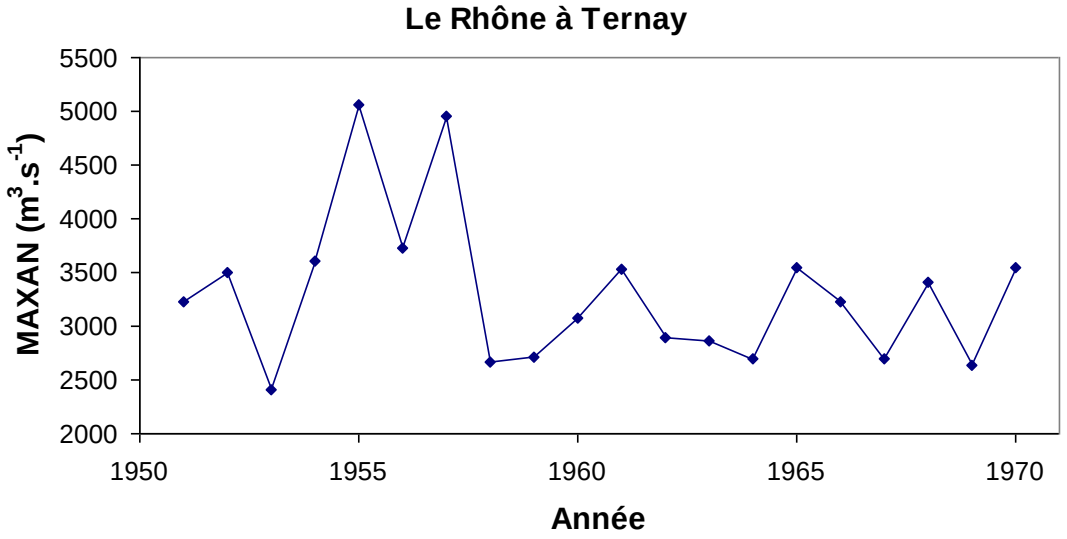


Figure 2-1. Valeurs maximales annuelles pour le Rhône à Ternay (50560 km²).

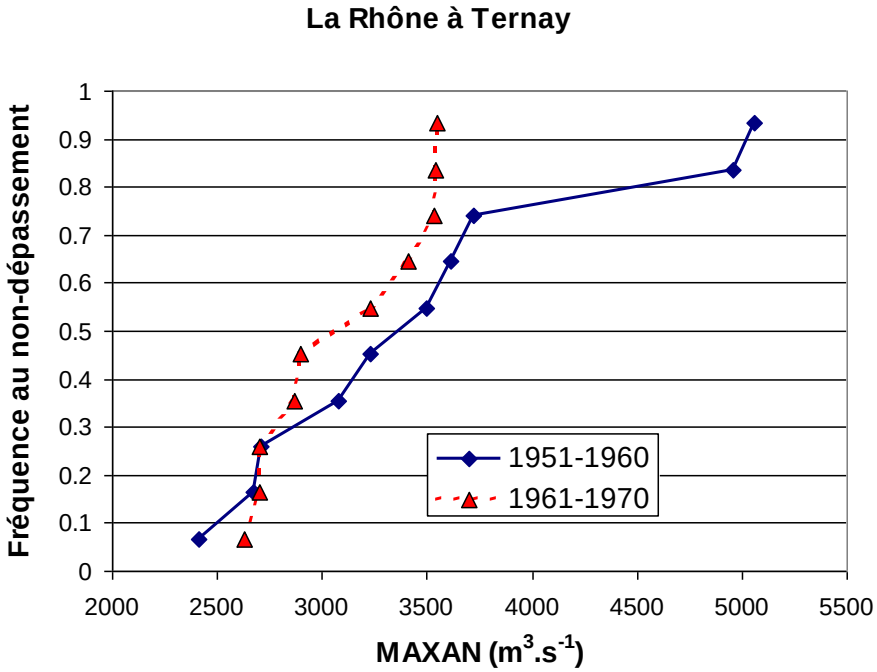


Figure 2-2. Illustration de la variabilité d'échantillonnage : comparaison des courbes de fréquence cumulées pour les périodes 1951-1960 et 1961-1970.

2.2 Définitions

Nous présentons dans cette section les concepts issus de la théorie des Probabilités qui seront utiles à l'hydrologue.

2.2.1 Probabilité

En termes intuitifs, une **probabilité** est un instrument de mesure du hasard : elle évalue le caractère probable (ou improbable) **d'événements** que l'on peut (ou que l'on choisit) de considérer comme **aléatoires**.

Une probabilité varie entre 0 et 1 : plus la probabilité est proche de 1, plus l'événement a des chances de se produire ; inversement, une probabilité proche de zéro dénote un événement ayant peu de chance de se produire. Un événement impossible a une probabilité nulle, un événement certain a une probabilité égale à 1.

La notion de probabilité peut également être formalisée de façon plus mathématique. Une **probabilité** sur un ensemble Ω est une fonction $P : \wp(\Omega) \mapsto [0;1]$ qui vérifie les trois axiomes suivants :

- ✓ $P(A) \geq 0, \forall A \subset \Omega$
- ✓ $P(\Omega) = 1$
- ✓ Soit (A_i) une suite de sous-ensembles de Ω deux à deux disjoints, c'est à dire $A_i \cap A_j = \varnothing, \forall i \neq j$. Alors $P(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$

Dans la définition ci-dessus, Ω représente l'ensemble des issues possibles de l'expérience aléatoire, il est appelé **univers** ou **ensemble fondamental.** Un sous-ensemble A de Ω est

appelé un **événement**. La notation $\wp(\Omega)$ représente l'ensemble des parties de Ω, c'est à dire l'ensemble de tous les sous-sembles possibles de Ω, ou encore l'ensemble des événements.
Les trois axiomes des Probabilités conduisent aux propriétés suivantes :

- ✓ $P(A) \in [0,1], \forall A \subset \Omega$
- ✓ $P(\varnothing) = 0$
- ✓ $A \subset B \Rightarrow P(A) \leq P(B)$
- ✓ $P(A^c) = 1 - P(A)$
- ✓ $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

Remarque 1 : la définition donnée ci-dessus est en fait abusive d'un point de vue strictement mathématique, car il peut exister des sous-ensembles de Ω pour lesquels la probabilité n'est pas définie (sous-ensembles non-mesurables). Il faut alors définir en ensemble d'évènements mesurables, ce qui requiert des outils mathématiques trop sophistiqués pour être évoqués dans ce cours.

Remarque 2 : Un événement impossible a une probabilité nulle, un événement certain a une probabilité égale à 1. La réciproque n'est cependant pas vrai du point de vue mathématique : des événements de probabilité nulle ne sont pas nécessairement impossibles, dans le sens où des sous-ensembles non vides de Ω peuvent être de probabilité nulle. Par exemple, la probabilité de « tirer au hasard » un nombre entier dans l'ensemble des nombres réels est nulle, bien que l'ensemble des nombres entiers ne soit pas vide…

2.2.2 Variable aléatoire

En termes intuitifs, une **variable aléatoire** permet de décrire de manière quantitative l'ensemble des résultats possibles d'une expérience aléatoire. Supposons par exemple que nous lancions deux dés, et que nous nous intéressions à la somme des deux faces observées. La variable aléatoire « X := Somme des deux dés » prend ainsi des valeurs entières comprises entre 2 et 12. Nous nous intéresserons au calcul de probabilité d'événements du type « la somme des deux dés vaut 4 » (noté $P(X = 4)$) ou « la somme des deux dés est strictement inférieure à 6 » (noté $P(X < 6)$).

D'un point de vue plus hydrologique, une variable aléatoire X représentera tout simplement la variable que nous souhaitons étudier, par exemple « Nombre de crues observées en une année » ou « débit maximum annuel » (cf. exemples en section 1.2). Il est par contre important de bien faire la différence entre la variable aléatoire X et les observations $(x_1,\ldots,x_n)$: la variable aléatoire X est un objet mathématique abstrait, que l'on considère comme le « générateur » des données observées $(x_1,\ldots,x_n)$. On dit d'ailleurs que $(x_1,\ldots,x_n)$ sont des **réalisations** de la variable aléatoire X. Les notations utilisées dans ce document utilisent des lettres capitales pour les variables aléatoires, des lettres minuscules pour leurs réalisations.

En termes plus mathématiques, une **variable aléatoire réelle** est une fonction définie sur l'univers Ω et à valeurs réelles, $X : \Omega \mapsto]-\infty;+\infty[$. Dans l'exemple des dés utilisé ci-dessus, l'ensemble Ω est égal à l'ensemble des couples *(a, b)*, où *a* et *b* sont des entiers entre 1 et 6. X sera la fonction somme :

$$X : (1,\ldots,6) \times (1,\ldots,6) \mapsto (2,\ldots,12)$$
$$(a,b) \mapsto a+b$$

Ainsi, l'événement « la somme des deux dés vaut 4 » correspond à l'ensemble des couples *(a, b)* pour lesquels $a+b=4$, soit $X^{-1}(\{4\})=\{(1, 3), (3, 1), (2, 2)\}$. Notons que nous utiliserons l'abus de notation « $X=4$ » au lieu de la notation « $X^{-1}(\{4\})$ ». Dans cet exemple précis, X ne prend pas ses valeurs dans l'ensemble de tous les nombres réels, mais seulement dans un sous-ensemble de l'ensemble des entiers : il s'agit d'une **variable aléatoire discrète**.

2.2.3 Distribution d'une variable aléatoire discrète

Soit X une variable aléatoire discrète, à valeurs dans un ensemble dénombrable ordonné $\{x_i, i=1,2,...\}$. La **loi de probabilité** de X est la donnée des quantités $f(x_i)=P(X=x_i), \forall i=1,2,...$. Cette loi de probabilité peut être décrite par un diagramme en bâtons (Figure 2-3).

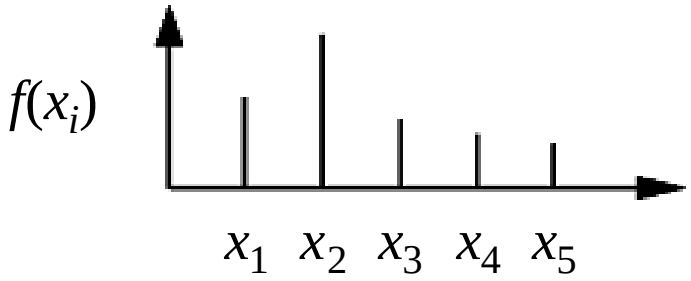


Figure 2-3. Loi de probabilité discrète.

Il existe évidemment une analogie entre cette représentation est celle présentée en section 1.2.1, où nous avions reporté les fréquences (en ordonnées) de chaque observation (en abscisses): nous aurons l'occasion de revenir à plusieurs reprises sur cette analogie fréquence / probabilité. L'objectif d'un modèle probabiliste est justement de remplacer les fréquences observées par des probabilités issues du modèle, pour les raisons exposées en section 2.1.

Nous définissons également la **fonction de répartition** F (Figure 2-4) de la variable aléatoire X de la manière suivante : $F(x_i)=P(X \leq x_i)$.

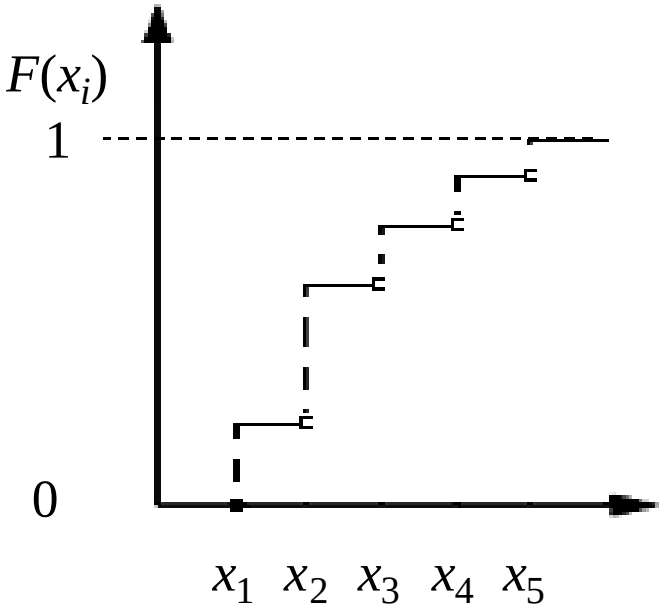


Figure 2-4. Fonction de répartition discrète.

Il est aisé de démontrer la relation suivante : $F(x_i) = \sum_{k=1}^{i} f(x_k)$. Autrement dit, la fonction de répartition est égale au cumul des probabilités individuelles de chaque valeur possible. On retrouve encore une fois l'analogie avec la courbe des fréquences cumulées : la fonction de répartition calculée en une valeur *x* est l'analogue de la fréquence au non-dépassement d'une observation *x*.

En corollaire, *F* est une fonction croissante, qui part de 0 et tend vers 1 en l'infini.

De ces deux définitions découlent quelques grandeurs caractéristiques de la variable aléatoire *X* (notez encore une fois l'analogie avec la section 1.2!) :

- ✓ L'**espérance**, $E(X) = \sum_{i=0}^{\infty} x_i \, f(x_i)$
- ✓ La **variance**, $Var(X) = \sum_{i=0}^{\infty} (x_i - E(X))^2 f(x_i)$, et **l'écart type** $\sigma(X) = \sqrt{Var(X)}$
- ✓ Le **quantile d'ordre *p***, noté x_p, qui vérifie $F(x_p) = p$
- ✓ Les **moments d'ordre *k***, $\mu_k = \sum_{i=0}^{\infty} x_i^k f(x_i)$. Si *k=1*, on reconnaît l'espérance.
- ✓ Les **moments centrés d'ordre *k***, $\mu_k' = \sum_{i=0}^{\infty} (x_i - E(X))^k f(x_i)$. Si *k=2*, on reconnaît la variance.

Remarque 1 : Les quantités faisant intervenir des sommes infinies peuvent ne pas exister.

Remarque 2 : Etant donné que la fonction de répartition *F* est en escalier, le quantile d'ordre *p* n'est pas forcément défini (si *p* tombe « entre les marches » de l'escalier).

2.2.4 Distribution d'une variable aléatoire continue

Les notions de loi de probabilité et de fonction de répartition décrites précédemment peuvent se généraliser au cas d'une variable continue, avec quelques différences liées au passage du

discret au continu. Nous allons commencer cette fois ci par définir la distribution d'une variable aléatoire continue *X* comme la donnée d'une **fonction de répartition** *F*, telle que $F(a) = P(X \le a)$. La **densité de probabilité** de *X* sera alors la fonction *f* telle que : $F(a) = \int_{-\infty}^{a} f(x)dx$. En d'autres termes, *f* est la dérivée de la fonction de répartition *F* (Figure 2-5).

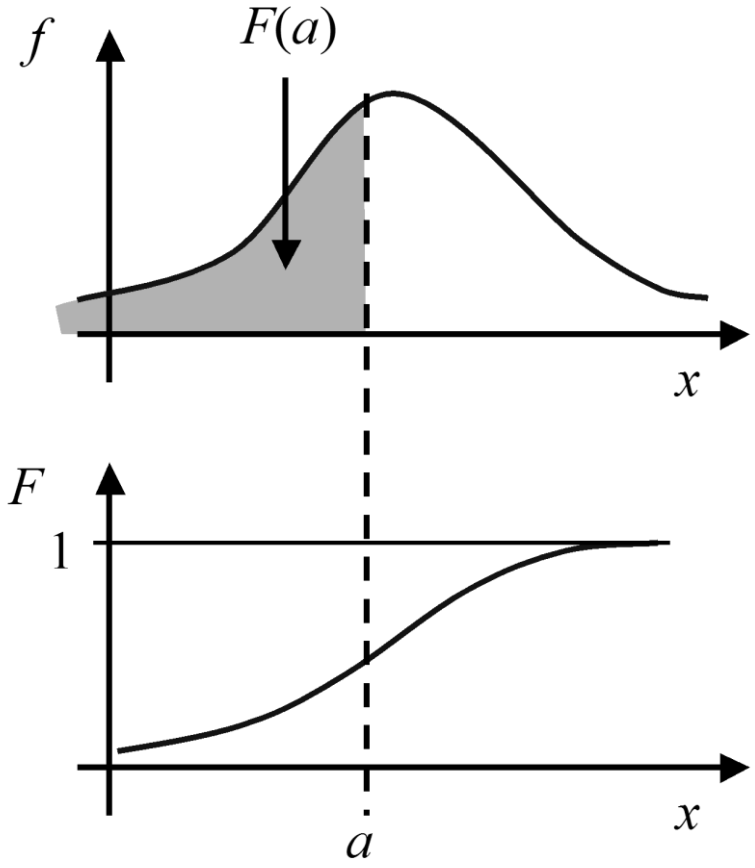


Figure 2-5. Densité et fonction de répartion continues.

Propriété : si *f* est continue, alors $P(a < X \le b) = \int_{a}^{b} f(x)dx = F(b) - F(a)$.

Dans le cas continu, on voit ainsi apparaître une analogie entre la probabilité d'un événement et l'aire sous la courbe de la densité. Cette analogie implique en particulier que $P(X = x_0) = 0, \forall x_0 \in \boldsymbol{R}$. C'est la raison pour laquelle il n'est pas possible de définir la loi de probabilité directement à partir des probabilités des éléments de Ω, ce qui conduit à raisonner sur des intervalles. On remarque encore une fois l'analogie avec les histogrammes dans le cas continu du chapitre précédent, où nous étions obligés de procéder à des regroupements en classes.

A partir de la densité de probabilité, il est possible de définir les mêmes grandeurs caractéristiques que dans le cas discret, en remplaçant les sommes par des intégrales :

- ✓ L'**espérance**, $E(X) = \int_{-\infty}^{+\infty} xf(x)dx$

- ✓ La **variance**, $\mathrm{Var}(X) = \int_{-\infty}^{+\infty} \left(x - E(X)\right)^2 f(x)dx$, et **l'écart type** $\sigma(X) = \sqrt{Var(X)}$
- ✓ Le **quantile d'ordre *p***, noté x_p, qui vérifie $F(x_p) = p$. De manière équivalente, le quantile d'ordre p est égal à l'inverse de la fonction de répartition prise en p, $x_p = F^{-1}(p)$.
- ✓ Les **moments d'ordre *k***, $\mu_k = \int_{-\infty}^{+\infty} x^k f(x)dx$. Si *k=1*, on reconnaît l'espérance.
- ✓ Les **moments centrés d'ordre *k***, $\mu'_k = \int_{-\infty}^{+\infty} (x - E(X))^k f(x)dx$. Si *k=2*, on reconnaît la variance.

Remarque : Encore une fois, rien ne garantit l'existence des intégrales infinies. De plus, certaines lois ne sont pas définies sur $\mathbb{R}$ tout entier, ces intégrales doivent alors être réduites aux supports de ces lois.

2.3 Distribution des valeurs « moyennes » : la loi de Gauss

2.3.1 Définition

La **loi de Gauss** (également appelée **loi normale**) est centrale en statistiques, puisqu'un grand nombre de phénomènes ont un comportement approximativement Gaussien. De plus, un théorème que nous verrons ultérieurement affirme qu'une moyenne (ou une somme) de variables aléatoires tend (presque) toujours en distribution vers une loi normale. Cette loi dépend de deux paramètres μ et σ représentant la moyenne et l'écart-type de la distribution. Elle est parfois utilisée en hydrologie pour modéliser des modules annuels.

La loi normale, notée *N(μ, σ)*, possède les propriétés suivantes :

$$
\begin{aligned}
&f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}[(x-\mu)/\sigma]^2 \right] \\
&F(x) = \int_{-\infty}^{x} f(t)dt \text{ (pas d'expression analytique)} \\
&\sigma > 0 \\
&E(X) = \mu \\
&Var(X) = \sigma^2
\end{aligned}
\qquad (2\text{-}1)
$$

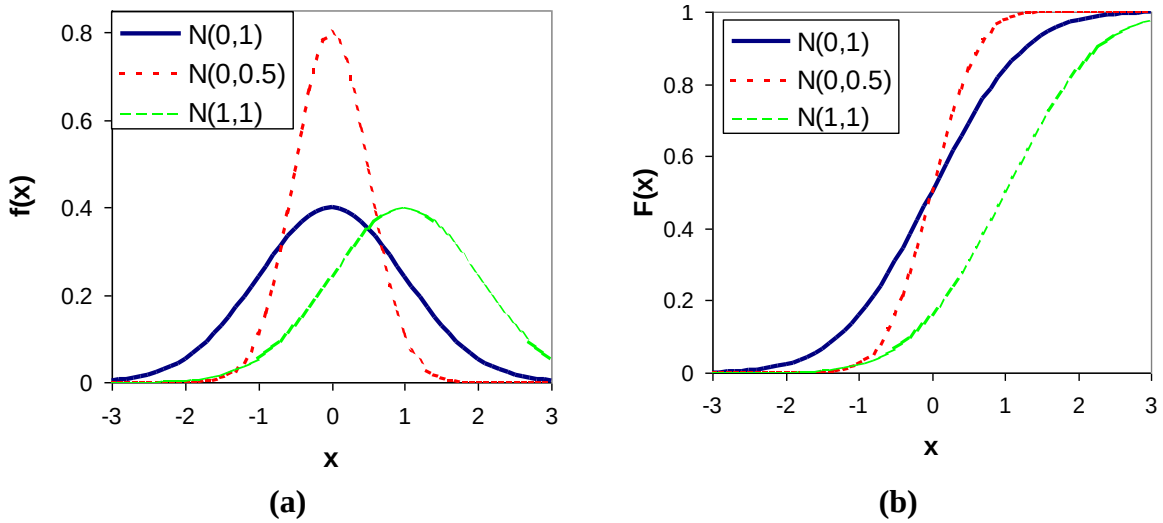


Figure 2-6. Exemples de densités (a) et de fonctions de répartition (b) pour la loi normale.

2.3.2 Justification théorique

Le rôle important joué par la loi de Gauss en statistiques appliquées peut s'expliquer par le théorème central limite. Ce théorème peut s'exprimer de la manière suivante. Soit $(X_1,...,X_n)$ un *n*-uplet de variables aléatoires indépendantes et de même distribution (on notera ***iid*** pour *indépendantes et identiquement distribuées*), d'espérance μ et d'écart-type σ. Alors :

$$\frac{X_1+...+X_n-n\mu}{\sigma\sqrt{n}} \xrightarrow[n\to\infty]{} N(0,1) \qquad (2\text{-}2)$$

Ce qui peut se réécrire de manière équivalente :

$$\frac{\bar{X}_n-\mu}{\sigma/\sqrt{n}} \xrightarrow[n\to\infty]{} N(0,1) \qquad (2\text{-}3)$$

La signification du théorème est la suivante : **quelle que soit la loi dont est issu un échantillon** (pourvu qu'espérance et variance existent), la statistique « moyenne empirique » suivra une loi normale, pourvu que *n* soit suffisamment grand. En pratique, une trentaine d'observations (indépendantes) est souvent jugée suffisante pour appliquer l'approximation Gaussienne.

La Figure 2-7 illustre le théorème central limite. Pour trois lois de probabilité bien distinctes (Normale, Uniforme et de Bernoulli), nous avons simulé des échantillons de diverses tailles *n*, et calculé la moyenne $\bar{X}_n$ de ces *n* valeurs. En répétant l'expérience un grand nombre de fois (1000 fois ici), on obtient les histogrammes en Figure 2-7 qui approximent la densité de probabilité de la moyenne $\bar{X}_n$. Nous observons alors que quelle que soit la forme de la distribution parente (première colonne), la distribution d'échantillonnage de $\bar{X}_n$ prend peu à peu une forme Gaussienne, comme le prédit le théorème central limite.

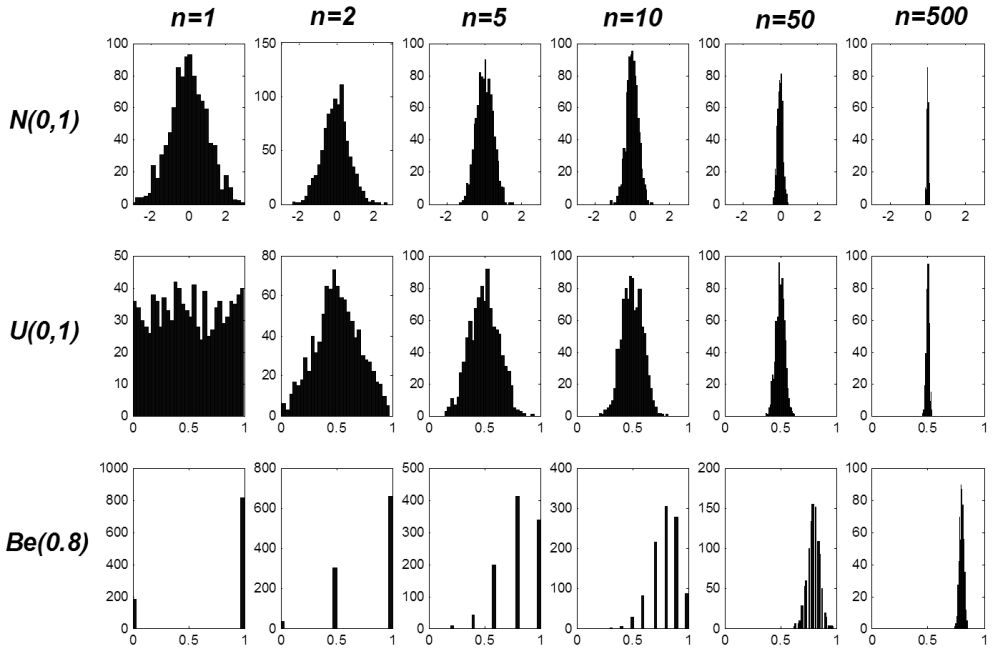


Figure 2-7. Illustration du théorème central limite.

2.4 Loi de Galton

La **loi de Galton**, également appelée **loi log-normale**, est une distribution dérivée de la loi normale de la manière suivante : la loi log-normale de paramètres μ et σ est la distribution de la variable exp(X), où X suit une loi normale de paramètres μ et σ. Inversement, le logarithme népérien d'une variable suivant une loi log-normale suit une loi normale. De par sa définition, une variable suivant une loi log-normale est strictement positive. Cette distribution est utilisée en hydrologie notamment pour modéliser des variables d'étiage ou parfois des modules annuels.

La loi log-normale, notée Log$N(\mu, \sigma)$, possède les propriétés suivantes :

$$f(x) = \begin{cases} \dfrac{1}{x\sigma\sqrt{2\pi}} \exp\left[-\dfrac{1}{2}[(\log(x) - \mu)/\sigma]^2 \right] & \text{si } x > 0 \\ 0 \text{ sinon} \end{cases}$$

$$F(x) = \int_{-\infty}^{x} f(t)dt \text{ (pas d'expression analytique)} \qquad (2\text{-}4)$$

$$\sigma > 0, x > 0$$

$$E(X) = \exp(\mu + \sigma^2/2)$$

$$Var(X) = \left[\exp(\sigma^2) - 1 \right] \exp(2\mu + \sigma^2)$$

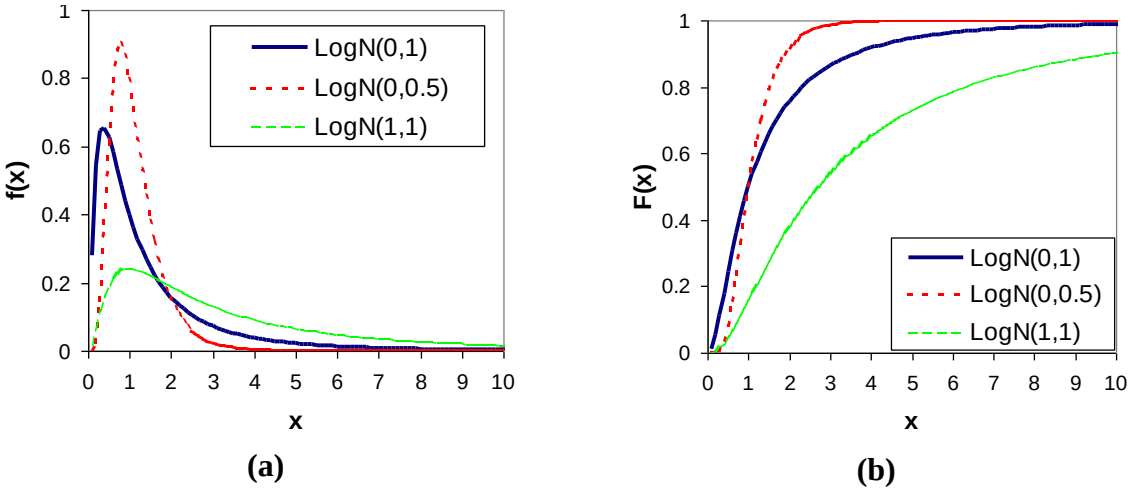


Figure 2-8. Exemples de densités (a) et de fonctions de répartition (b) pour la loi log-normale.

2.5 Distributions des valeurs extrêmes : maxima

Les distributions des valeurs extrêmes jouent un rôle central en hydrologie, notamment pour l'estimation des crues. Inversement, l'hydrologie a été une des motivations premières au développement d'une théorie des valeurs extrêmes, puisque les travaux précurseurs d'Emil Julius Gumbel avaient pour ambition de fournir un cadre théorique aux problèmes d'estimation des crues. Dans cette section, nous nous focalisons sur les distributions utilisées pour modéliser les maxima annuels (méthode d'échantillonnages MAXAN, section 1.1.3). Une autre section sera consacrée à l'étude des dépassements de seuil (méthode d'échantillonnage SUPSEUIL, section 1.1.3).

2.5.1 Définitions

La **loi de Gumbel**, notée $Gu(\mu, \lambda)$, est la première loi des valeurs extrêmes à avoir été utilisée pour l'étude des maxima annuels. Il s'agit d'une loi à deux paramètres, le **paramètre de position** μ et **le paramètre d'échelle** λ. Ce dernier paramètre est également appelé le **gradex** en hydrologie francophone. La loi de Gumbel possède les propriétés suivantes :

$$f(x) = \frac{1}{\lambda}\exp\left(-\frac{x-\mu}{\lambda} - \exp\left(-\frac{x-\mu}{\lambda}\right)\right)$$

$$F(x) = \exp\left[-\exp\left(-\frac{x-\mu}{\lambda}\right)\right] \qquad (2\text{-}5)$$

$$\lambda > 0$$

$$E(X) = \mu + \gamma\lambda, \text{ avec } \gamma = 0.5772 \text{ (constante d'Euler-Mascheroni)}$$

$$Var(X) = \frac{\pi^2}{6}\lambda^2$$

La **loi généralisée des valeurs extrêmes**, notée $GEV(\mu, \lambda, \xi)$, généralise la loi de Gumbel en lui ajoutant un troisième paramètre ξ appelé le **paramètre de forme**. Elle possède les propriétés suivantes :

$$f(x)=\begin{cases}\dfrac{1}{\lambda}\left(1-\dfrac{\xi(x-\mu)}{\lambda}\right)^{\frac{1}{\xi}-1}\exp\left(-\left(1-\dfrac{\xi(x-\mu)}{\lambda}\right)^{\frac{1}{\xi}}\right) \text{ si } 1-\dfrac{\xi(x-\mu)}{\lambda}>0\\ 0 \text{ sinon}\end{cases}$$

$$F(x)=\exp\left(-\left(1-\frac{\xi(x-\mu)}{\lambda}\right)^{\frac{1}{\xi}}\right)$$

$$\lambda>0,\ \xi\neq 0,\ 1-\frac{\xi(x-\mu)}{\lambda}>0 \qquad (2\text{-}6)$$

$$E(X)=\begin{cases}\mu+\dfrac{\lambda}{\xi}(1-\Gamma(\xi+1)) \text{ si } \xi>-1\\ \text{n'existe pas sinon}\end{cases}$$

$$Var(X)=\begin{cases}\left(\dfrac{\lambda}{\xi}\right)^2\left(\Gamma(2\xi+1)-\Gamma^2(\xi+1)\right) \text{ si } \xi>-1/2\\ \text{n'existe pas sinon}\end{cases}$$

$\Gamma(t)=\int_0^{+\infty}x^{t-1}e^{-x}dx$ est la fonction gamma.

La Figure 2-9 présente les densités de probabilité de trois lois GEV, qui possèdent les mêmes paramètres de position et d'échelle mais diffèrent par leur paramètre de forme ξ. Si $\xi<0$, la loi GEV est dite « de type Fréchet » : il s'agit d'une distribution bornée à gauche (i.e. elle possède une borne inférieure), que l'on qualifie de « distribution à queue lourde » pour signifier que la densité chute lentement vers zéro en l'infini. Si $\xi>0$, la loi GEV est dite « de type Weibull » : elle est bornée à droite (i.e. elle possède une borne supérieure). Enfin, lorsque $\xi\to 0$, la densité de la loi GEV tend vers la densité d'une loi de Gumbel. On pourra donc considérer que la loi de Gumbel est un cas particulier de la loi GEV, obtenu en prenant un paramètre de forme nul. Notons qu'il s'agit là d'un léger abus de langage, car les formules de l'équation (2-6) ci-dessus ne sont pas définies pour $\xi = 0$.

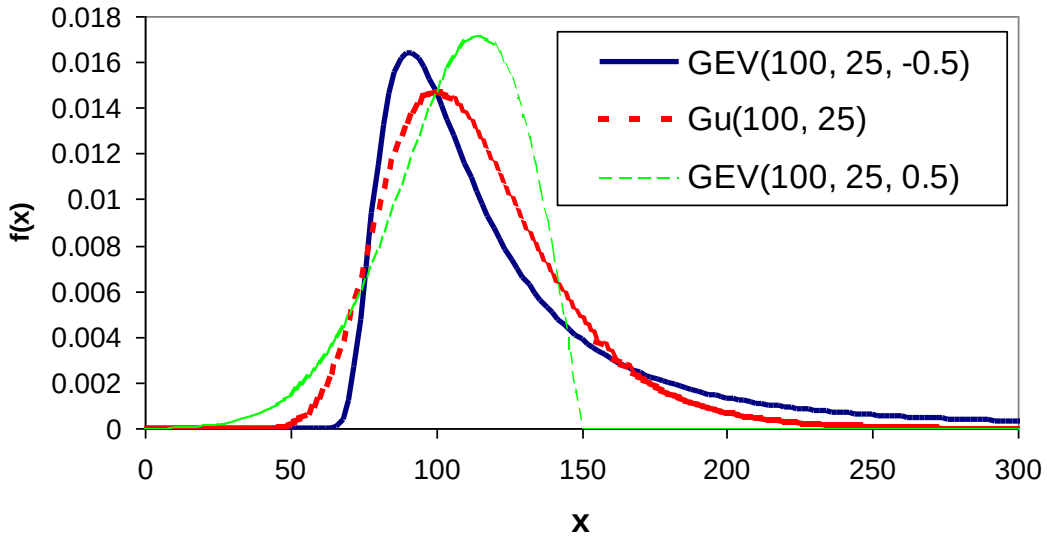


Figure 2-9. Densités de probabilités de lois GEV.

Le paramètre de forme joue un rôle majeur dans l'estimation de quantiles de crues, car il gouverne leur comportement asymptotique. A titre d'illustration, nous avons reporté en Figure 2-10 les fonctions de répartitions des trois lois GEV utilisées précédemment. La ligne horizontale correspond à la probabilité 0.95. L'intersection avec les fonctions de répartition donne donc la valeur du quantile d'ordre 0.95 : suivant la valeur du paramètre de forme, ce quantile varie entre environ 140 (pour ξ=0.5), environ 175 (pour ξ=0), et jusqu'à plus de 250 (pour ξ<0).

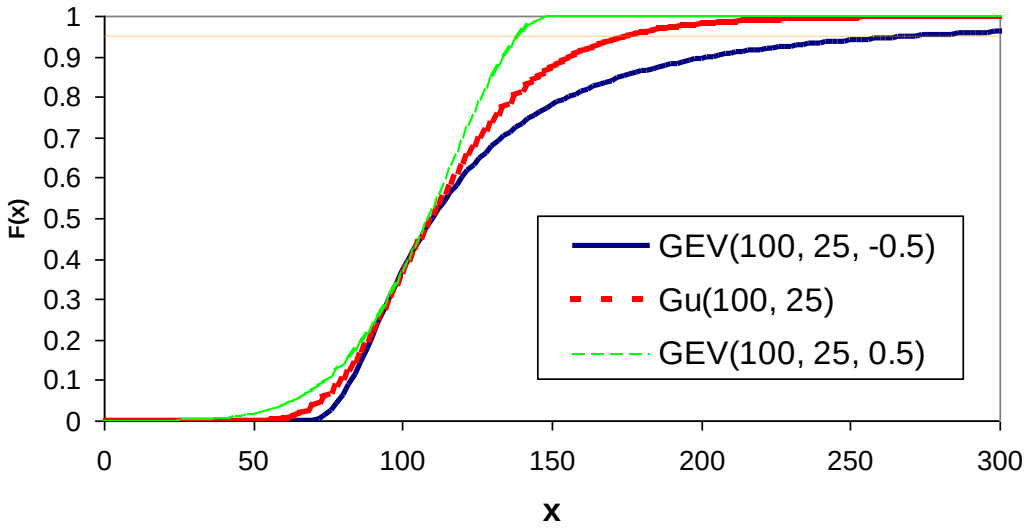


Figure 2-10. Fonctions de répartitions de lois GEV.

2.5.2 Justification théorique

De même que le rôle central de la loi normale pour les valeurs « moyennes » était du à l'existence d'un théorème central limite, la loi GEV trouve sa justification dans un théorème limite appelé le théorème des valeurs extrêmes.

Le théorème des valeurs extrêmes est, en quelque sorte, l'équivalent du théorème central limite, mais où la statistique « moyenne empirique » $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ est remplacée par la statistique $M_n = Max(X_1, ..., X_n)$. La formulation est un peu délicate, nous le retiendrons sous la forme suivante : si la distribution de M_n converge, alors c'est forcément vers une loi généralisée des valeurs extrêmes $GEV(\mu, \lambda, \xi)$.

La Figure 2-11 permet d'illustrer cette convergence pour les lois *N(0,1)*, *U(0,1)* (loi uniforme sur [0 ;1]) et de Cauchy, dont les maximums convergent vers une GEV où le paramètre ξ est respectivement nul (=loi de Gumbel), positif (=loi de Weibull) et négatif (=Loi de Fréchet). Cette figure a été obtenue de manière analogue à la Figure 2-7 : pour les trois lois de probabilité (Normale, Uniforme et de Cauchy), nous avons simulé des échantillons de diverses tailles *n*, et calculé le maximum M_n de ces *n* valeurs. En répétant l'expérience un grand nombre de fois (1000 fois), on obtient les histogrammes de la Figure 2-11.

Ce théorème justifie l'utilisation de ces lois en hydrologie des valeurs extrêmes, pour décrire le comportement probabiliste des crues. Notons qu'encore une fois, la loi limite du maximum (la loi GEV) ne dépend pas de la loi parente dont est issu l'échantillon.

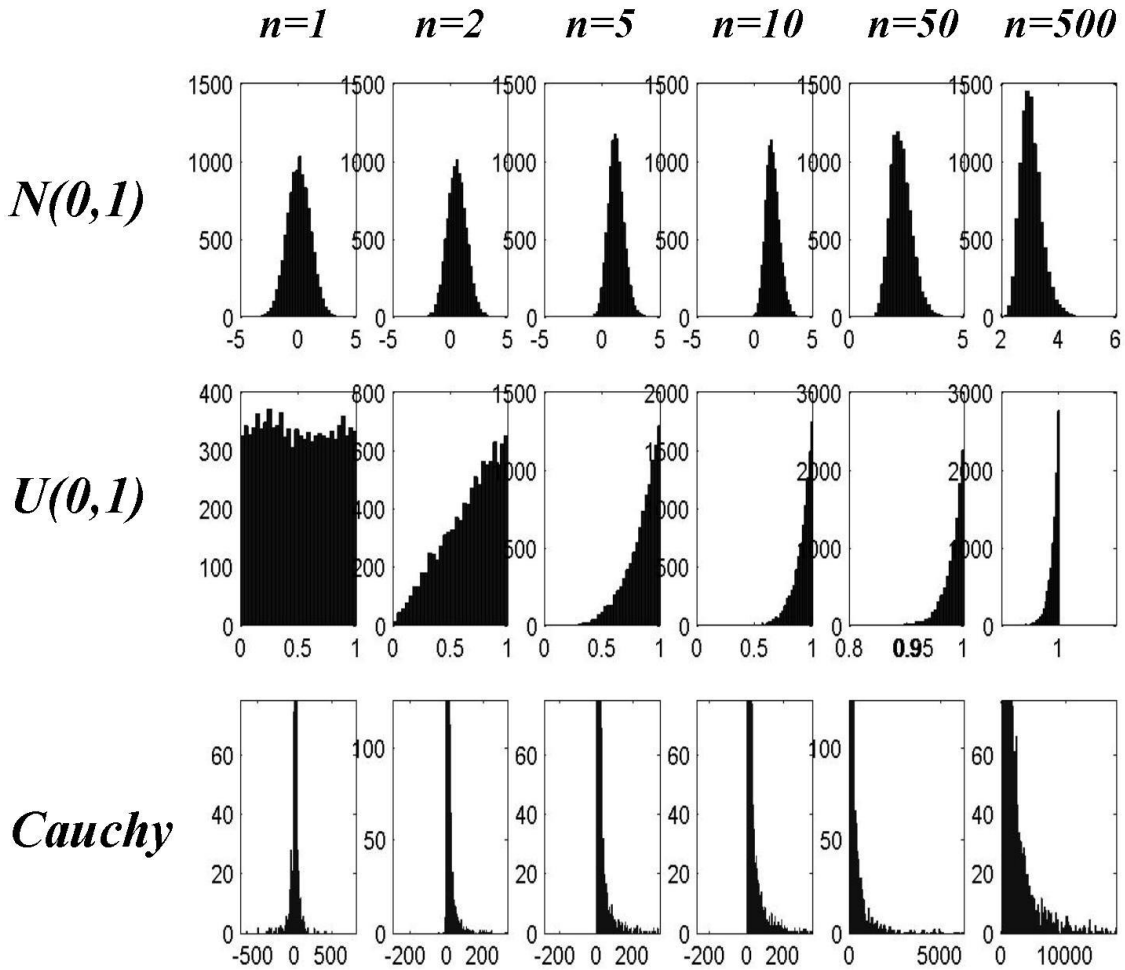


Figure 2-11. Illustration du théorème des valeurs extrêmes.

2.6 Distributions des valeurs extrêmes : dépassements de seuil

De même que la loi GEV et son cas particulier, la loi de Gumbel, sont des candidats naturels pour modéliser les maxima annuels, nous verrons dans cette section qu'il existe des distributions naturelles pour modéliser les dépassements de seuils obtenus par la méthode d'échantillonnage SUPSEUIL : la loi de Pareto généralisée, et son cas particulier la loi exponentielle.

2.6.1 Définitions

La **loi exponentielle**, notée $Exp(x_0, \lambda)$, est une distribution à deux paramètres : le **seuil** x_0, et le **paramètre d'échelle** λ (également appelé **gradex** en hydrologie francophone).

$$
\begin{aligned}
f(x) &= \begin{cases} \frac{1}{\lambda} e^{-\frac{x-x_0}{\lambda}} & \text{si } x > x_0 \\ 0 \text{ sinon} \end{cases} \\
F(x) &= \begin{cases} 1 - e^{-\frac{x-x_0}{\lambda}} & \text{si } x > x_0 \\ 0 \text{ sinon} \end{cases} \\
E(X) &= x_0 + \lambda \\
Var(X) &= \lambda^2
\end{aligned}
\qquad (2\text{-}7)
$$

La **loi de Pareto généralisée**, notée $GPD(x_0, \lambda, \xi)$, généralise la loi exponentielle en lui ajoutant un troisième paramètre ξ appelé le **paramètre de forme**. Elle possède les propriétés suivantes :

$$
\begin{aligned}
f(x) &= \begin{cases} \frac{1}{\lambda}\left(1 - \frac{\xi(x-x_0)}{\lambda}\right)^{\frac{1}{\xi}-1} & \text{si } x \geq x_0 \\ 0 \text{ sinon} \end{cases} \\
F(x) &= \begin{cases} 1 - \left(1 - \frac{\xi(x-x_0)}{\lambda}\right)^{\frac{1}{\xi}} & \text{si } x \geq x_0 \\ 0 \text{ sinon} \end{cases} \\
&\lambda > 0, \xi \neq 0, 1 - \frac{\xi(x-x_0)}{\lambda} > 0 \\
E(X) &= \begin{cases} \frac{\lambda}{1+\xi} + x_0 & \text{si } \xi > -1 \\ \text{n'existe pas sinon} \end{cases} \\
Var(X) &= \begin{cases} \frac{\lambda^2}{(1+\xi)^2(1+2\xi)} & \text{si } \xi > -1/2 \\ \text{n'existe pas sinon} \end{cases}
\end{aligned}
\qquad (2\text{-}8)
$$

La Figure 2-12 et la Figure 2-13 présentent respectivement les densités de probabilité et les fonctions de répartition de trois lois GPD. Le rôle du paramètre de forme est similaire à ce que l'on observait pour la loi GEV (section 2.5.1) : Si $\xi < 0$, la loi GPD est bornée à gauche et possède une queue lourde. Si $\xi > 0$, la loi GPD est bornée à droite. Enfin, lorsque $\xi \to 0$, la densité de la loi GPD tend vers la densité d'une loi exponentielle, ce qui conduit à considérer la loi exponentielle comme un cas particulier de la loi GPD, obtenu en prenant un paramètre de forme nul.

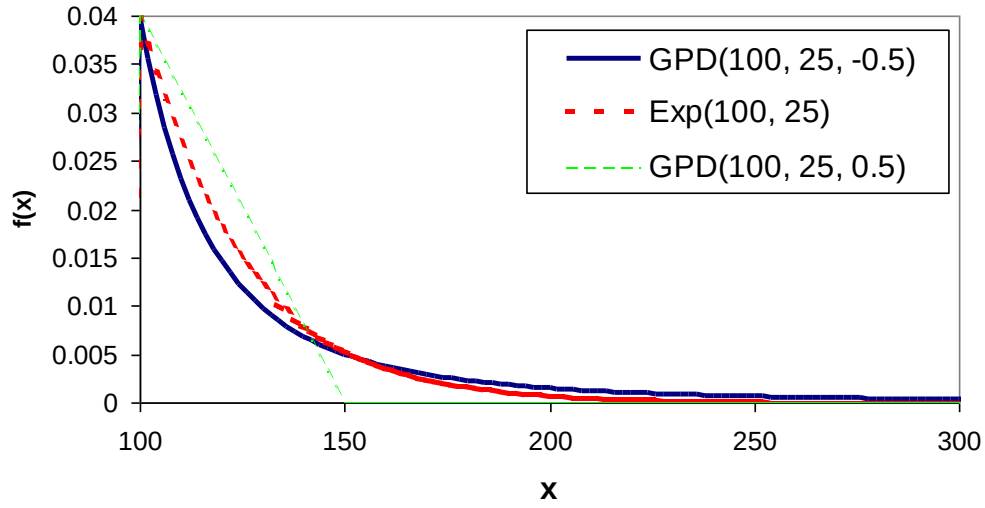


Figure 2-12. Densités de probabilités de lois GPD.

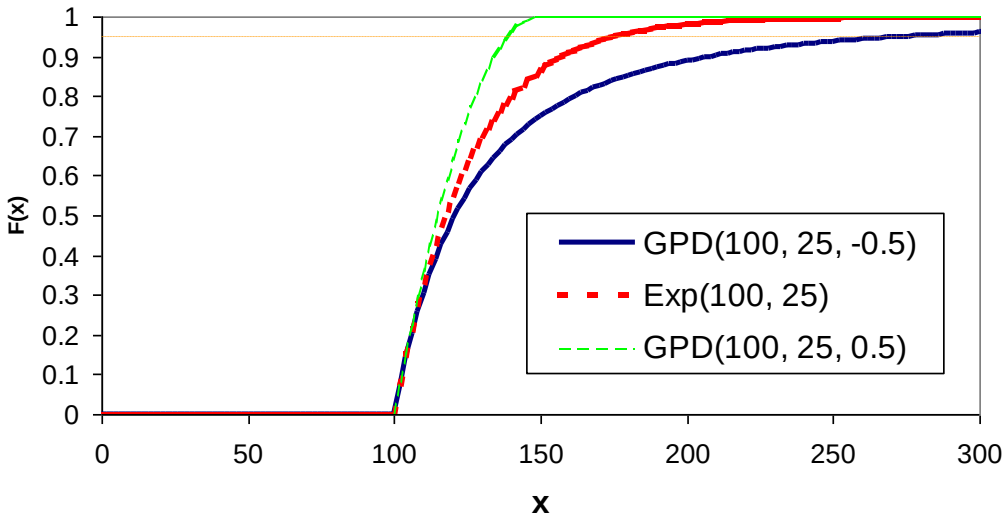


Figure 2-13. Fonctions de répartitions de lois GPD.

Par rapport à l'échantillonnage par la méthode MAXAN, l'échantillonnage SUPSEUIL présente l'avantage intéressant de pouvoir s'intéresser au processus d'occurrence des crues. Ainsi, si l'on considère la Figure 1-6, il est possible de dénombrer le nombre annuel d'événements de crue. Cette variable, que nous noterons N, est en général modélisée à l'aide d'une loi de Poisson notée $Pois(\theta)$. Le paramètre θ de la loi de Poisson est appelé le **taux d'occurrence**. La loi de Poisson possède les caractéristiques suivantes :

$$
\begin{aligned}
&P(N=k) = \mathrm{e}^{-\theta}\,\frac{\theta^k}{k!} \\
&P(N\le k) = \mathrm{e}^{-\theta}\sum_{i=0}^{k}\frac{\theta^i}{i!} \\
&k \text{ entier positif}, \theta > 0 \\
&E(N) = \theta \\
&Var(N) = \theta
\end{aligned}
\qquad (2\text{-}9)
$$

La Figure 2-14 illustre la loi de probabilité de la loi de Poisson.

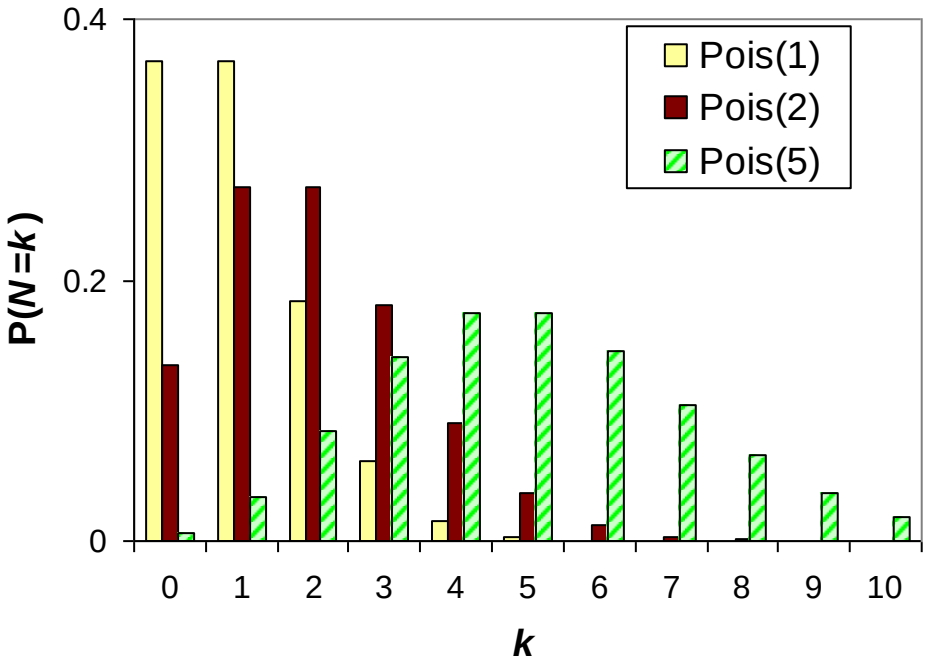


Figure 2-14. Lois de probabilités de lois de Poisson.

2.6.2 Justification théorique

L'utilisation du couple loi de Poisson + loi GPD pour la modélisation des dépassements de seuil (échantillonnage SUPSEUIL) est justifiée par une relation forte avec l'utilisation de la loi GEV pour les maxima (échantillonnage MAXAN). En effet, il est possible de démontrer que si le processus d'occurrence suit une loi de Poisson $Pois(\theta)$ et le processus de dépassement du seuil une loi $GPD(x_0, \lambda_{GPD}, \xi_{GPD})$ (échantillonnage SUPSEUIL), ces deux variables étant supposées indépendantes, alors le maximum annuel suit une loi $GEV(\mu, \lambda_{GEV}, \xi_{GEV})$ (MAXAN). De plus, les paramètres des distributions sont liés par les relations suivantes :

$$\begin{aligned} \xi_{GEV} &= \xi_{GPD} = \xi \\ \lambda_{GEV} &= \lambda_{GPD}.\theta^{-\xi} \\ \mu &= x_0 + \frac{\lambda_{GPD}(1-\theta^{-\xi})}{\xi} \end{aligned} \qquad (2\text{-}10)$$

Le choix entre les deux approches (modélisation SUPSEUIL par Poisson + GPD ou modélisation MAXAN par GEV) dépend de plusieurs facteurs qui sont généralement spécifiques au bassin étudié. Lang (1995) propose une comparaison poussée entres ces deux approches, et suggère que la modélisation SUPSEUIL devient plus avantageuse si l'on peut sélectionner au moins 2 événements par an en moyenne. Il convient néanmoins de ne pas sélectionner trop d'événements, car la modélisation par une loi GPD n'est valide que pour des seuils suffisamment élevés. Une bonne pratique est d'utiliser les deux approches simultanément sur le même bassin et de comparer les estimations.

2.7 Estimation des paramètres

2.7.1 Principes de l'estimation

La théorie de l'estimation est un domaine important et vaste des Statistiques. Nous n'en aborderons que quelques aspects, en expliquant comment mesurer la qualité d'un estimateur, et en présentant deux méthodes d'estimation très utilisées en pratique.

Considérons un échantillon d'observations $(x_1,...,x_n)$ *iid*, réalisation d'une variable aléatoire *X*. Il est fréquent d'avoir une idée *a priori* sur la loi dont devrait être issu cet échantillon, par expérience ou en considérant un des théorèmes limites des sections précédentes (ex. : la crue maximale annuelle suit une loi GEV). Nous allons donc supposer que l'échantillon est issu d'une certaine loi, de densité f paramétrée par un certain nombre de paramètres que nous noterons sous forme vectorielle $\boldsymbol{\theta}$. Nous noterons $f(x|\boldsymbol{\theta})$ cette densité. Par exemple, pour une loi normale, $\boldsymbol{\theta}=(\mu,\sigma)$. La question est donc à présent d'**estimer** la valeur des paramètres à partir des données observées. Pour cela, nous allons faire appel à des statistiques $T_n = g(X_1,...,X_n)$ particulières, appelées **estimateurs**. Nous réclamerons que ces estimateurs satisfassent à un certain nombre de critères garantissant leur **qualité**. Parmi ces critères, citons :

- ✓ La **convergence** : si T_n est un estimateur d'un paramètre α, obtenu à partir d'un échantillon de taille *n*, alors il est souhaitable que $T_n \xrightarrow[n\to\infty]{} \alpha$ (intuitivement, quand l'échantillon tend vers la population, on retrouve la vraie valeur)
- ✓ **L'absence de biais** : $E(T_n)=\alpha$.
- ✓ **La précision** : $E\left((T_n-\alpha)^2\right)$, **l'erreur quadratique moyenne**, est minimale.

Notons qu'il n'est pas forcément possible de trouver un estimateur qui satisfasse à ces trois critères simultanément.

2.7.2 Estimation par la méthode des moments

L'idée de cette méthode est la suivante : si les paramètres sont bien estimés, alors il devrait y avoir adéquation entre les caractéristiques observées (ou empiriques) et les caractéristiques théoriques. Nous allons rechercher cette adéquation sur les moments, en général le premier

moment non centré (c'est à dire la moyenne) et le second moment centré (c'est à dire la variance), plus éventuellement des moments d'ordre supérieur pour les distributions possédant plus de deux paramètres. Plus précisément, supposons que nous ayons *p* paramètres $\theta_1,...,\theta_p$ à estimer, alors leurs estimateurs $\hat{\theta}_1,...,\hat{\theta}_p = \hat{\boldsymbol{\theta}}$ seront solutions du système :

$$\begin{cases} E_{\hat{\theta}}(X) = m_1 \\ Var_{\hat{\theta}}(X) = m_2^{'} \\ ..... \\ E_{\hat{\theta}}\left(\left(X - E_{\hat{\theta}}(X)\right)^p\right) = m_p^{'} \end{cases} \quad (2\text{-}11)$$

La notation $E_{\hat{\theta}}$ a pour but de bien montrer que les moments théoriques sont des fonctions des paramètres à estimer.

Exemple 1 : Estimateurs des paramètres d'une loi normale

Soit $(x_1,...,x_n)$ *iid*, un échantillon réalisation d'une loi normale *N(μ,σ)*. L'espérance vaut donc μ et la variance σ^2. Les estimateurs des moments sont donc définis par :

$$\begin{aligned} &\hat{\mu} = m_1 = \bar{x} \\ &\hat{\sigma}^2 = m_2 = Var_x \Rightarrow \hat{\sigma} = s_x \end{aligned} \quad (2\text{-}12)$$

En d'autres termes, les estimateurs des moments sont tout simplement égaux à la moyenne et à l'écart-type empiriques.

Exemple 2 : Estimateurs des paramètres d'une loi uniforme

La loi uniforme sur l'intervalle $[0;\theta]$, notée U(θ), est une distribution dont la densité est constante sur l'intervalle $[0;\theta]$, et nulle ailleurs (cf. Figure 2-15). Cette distribution possède les propriétés suivantes :

$$\begin{aligned} &f(x) = \begin{cases} 1/\theta \text{ si } x \in [0;\theta] \\ 0 \text{ sinon} \end{cases} \\ &\theta > 0 \\ &E(X) = \theta/2 \\ &Var(X) = \theta^2/12 \end{aligned} \quad (2\text{-}13)$$

On s'intéresse à l'estimation de la borne supérieure θ. Si l'on observe des réalisations $(x_1,...,x_n)$ issues de cette loi, l'estimateur des moments du paramètre θ s'obtient simplement de la manière suivante :

$$E(X) = \hat{\theta}/2 = \bar{x} \Rightarrow \hat{\theta} = 2\bar{x} \quad (2\text{-}14)$$

Il s'agit néanmoins d'un très mauvais estimateur, pour des raisons très intuitives. Supposons par exemple que l'on ait observé les valeurs 1, 2, 3, et 14. La moyenne de ces valeurs vaut 5,

donc l'estimateur des moments de θ vaut 10 : cette estimation est tout simplement incompatible avec les observations, puisque l'on a observé la valeur 14 !

Pour information, cet exemple n'est pas uniquement didactique, mais est inspiré d'un problème réel. Durant la seconde guerre mondiale, les forces alliées ont cherché à estimer le nombre de tanks que possédaient les forces allemandes. Pour cela, les alliés ont utilisé les numéros de série observés sur des tanks capturés : si l'on admet que les tanks sont capturés « au hasard », alors la probabilité d'observer un numéro de série quelconque est constante et égale à $1/\theta$, où θ est le nombre total de tank que l'on cherche à estimer (loi uniforme discrète).

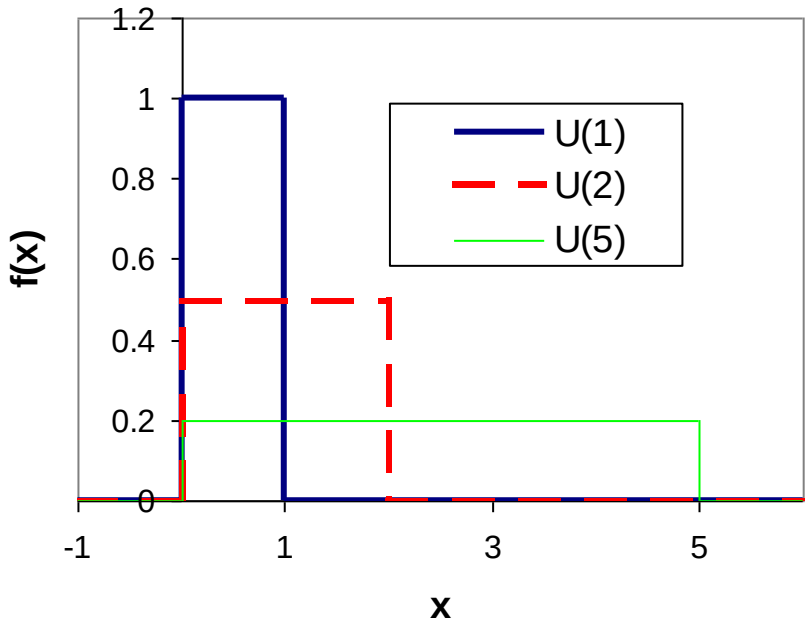


Figure 2-15. Densités de lois uniformes.

2.7.3 Estimation par la méthode du maximum de vraisemblance

Cette méthode consiste, étant donné un échantillon $(x_1,...,x_n)$ *iid*, de choisir comme estimateur de $\theta_1,...,\theta_p$ les valeurs qui rendent l'échantillon le plus « probable » possible. Plus précisément, nous appellerons **vraisemblance** de l'échantillon $(x_1,...,x_n)$ la fonction:

$$L(x_1,...,x_n \mid \boldsymbol{\theta}) = \prod_{i=1}^{n} f(x_i \mid \boldsymbol{\theta}) \qquad (2\text{-}15)$$

La vraisemblance d'un échantillon *iid* est donc simplement le produit des densités de probabilités calculées en chaque observation. Les valeurs $(x_1,...,x_n)$ ayant été observées et étant donc connues, la vraisemblance est vue comme une fonction des paramètres $\theta_1,...,\theta_p$ que l'on cherche à estimer. Il faut donc trouver les valeurs $\hat{\theta}_1,...,\hat{\theta}_p$ qui maximisent cette vraisemblance.

Dans la pratique, on aura souvent intérêt à maximiser la **log-vraisemblance**,

$$LogL(x_1,...,x_n \mid \boldsymbol{\theta}) = \sum_{i=1}^{n} \log\Big(f(x_i \mid \boldsymbol{\theta})\Big) \tag{2-16}$$

Pour ce faire, il faut résoudre *p* équations aux dérivées partielles :

$$\frac{\partial LogL}{\partial \theta_i} = 0, \forall i = 1,..., p \tag{2-17}$$

Exemple 1 : Estimation du paramètre d'une loi exponentielle *Exp(0,λ)*

Soient $(x_1,...,x_n)$ des observations issues de cette loi exponentielle, dont la densité s'écrit :

$$f(x \mid \lambda) = \frac{1}{\lambda} e^{-x/\lambda} \tag{2-18}$$

La Log-vraisemblance vaut donc :

$$LogL(x_1,...,x_n \mid \lambda) = \sum_{i=1}^{n} \log\Big(f(x_i \mid \lambda)\Big)$$

$$= \sum_{i=1}^{n} \log\left(\frac{1}{\lambda} e^{-x_i/\lambda} \right)$$

$$= n \log\left(\frac{1}{\lambda} \right) - \sum_{i=1}^{n} \frac{x_i}{\lambda}$$

$$= -n \log(\lambda) - \sum_{i=1}^{n} \frac{x_i}{\lambda}$$

D'où :

$$\frac{\partial LogL}{\partial \lambda} = -\frac{n}{\lambda} + \sum_{i=1}^{n} \frac{x_i}{\lambda^2} = 0$$

$$\Leftrightarrow -n + \sum_{i=1}^{n} \frac{x_i}{\lambda} = 0$$

$$\Leftrightarrow \sum_{i=1}^{n} \frac{x_i}{\lambda} = n$$

$$\Leftrightarrow \lambda = \frac{1}{n} \sum_{i=1}^{n} x_i$$

L'estimateur du maximum de vraisemblance correspond ici simplement à la moyenne empirique des observations.

Exemple 1 : Estimation du paramètre d'une loi uniforme

Reprenons l'exemple présenté en section 2.7.2. La vraisemblance d'un échantillon $(x_1,...,x_n)$ s'écrit :

$$L(x_1,...,x_n \mid \theta) = \begin{cases} \prod_{i=1}^{n} (1/\theta) = 1/\theta^n \text{ si } x_i \le \theta \ \forall i = 1,...,n \\ 0 \text{ sinon} \end{cases}$$

La fonction $1/\theta^n$ étant décroissante, le maximum de vraisemblance est atteint pour $\hat{\theta} = \max_{i=1,\ldots,n} \{x_i\}$. Par rapport à l'estimateur des moments calculé en section 2.7.2, l'estimateur du maximum de vraisemblance possède l'avantage de ne pas conduire à des estimations incompatibles avec les observations. Précisons néanmoins que dans le cas d'une loi uniforme, cet estimateur reste biaisé (il a tendance à sous-estimer la vraie valeur de θ), et des estimateurs plus efficaces ont été proposés dans la littérature.

2.7.4 Quel estimateur choisir ?

Il n'existe malheureusement pas de réponse universelle à cette question… Dans certains cas, les deux estimateurs présentés plus haut coïncident et sont « optimaux ». Dans d'autres cas, l'un aura des propriétés que l'autre n'aura pas, et vice-versa. En général, l'estimateur du maximum de vraisemblance est utilisé pour les grands échantillons, car la théorie permet d'en déterminer les propriétés asymptotiques (loi, variance). En contrepartie, les calculs détaillés ci-dessus sont souvent inextricables à la main dans des cas plus compliqués, il faut donc recourir à des méthodes d'optimisation numérique, qui ne garantissent d'ailleurs pas la convergence vers un minimum absolu.

Signalons enfin qu'il existe d'autres méthodes d'estimation utilisées dans la pratique hydrologique. On peut notamment citer la méthode des L-Moments (Hosking and Wallis 1997) et les estimations Bayesiennes (Gelman et al. 1995). L'exposition de ces approches dépassant le cadre de cette formation, le lecteur intéressé est invité à consulter les références indiquées pour une description détaillée.

2.7.5 Formulaire

Le Tableau 2-1 donne les estimateurs des moments et du maximum de vraisemblance pour les principales distributions présentées en section 2.3-2.6.

Distribution	**Exemple de variable hydrologique**	**Méthode des moments**	**Maximum de vraisemblance**
Normale N(μ,σ)	Module annuel	$\hat{\mu} = \bar{x}$ $\hat{\sigma} = s_x$	idem moments
Log-Normale LogN(μ,σ)	QMNA	$\hat{\mu} = \log(\bar{x}) - \frac{1}{2}\log(1+\frac{s_x^2}{\bar{x}^2})$ $\hat{\sigma} = \sqrt{\log(1+\frac{s_x^2}{\bar{x}^2})}$	$\hat{\mu} = \overline{\log(x)}$ $\hat{\sigma} = s_{\log(x)}$
Poisson *Pois*(θ)	Nombre annuel d'évènements de crue	$\hat{\theta} = \bar{x}$	idem moments
Gumbel *Gu*(μ,λ)	Débit MAXAN	$\begin{cases} \hat{\lambda} = \frac{\sqrt{6}}{\pi} s_x \\ \hat{\mu} = \bar{x} - \gamma\hat{\lambda}, \text{ où } \gamma = 0.5772 \end{cases}$	Méthode numérique
GEV(μ, λ, ξ)	Débit MAXAN	$\begin{cases} -\frac{\hat{\xi}}{\lvert\hat{\xi}\rvert}\frac{\left[\Gamma(3\hat{\xi}+1) - 3\Gamma(\hat{\xi}+1)\Gamma(2\hat{\xi}+1) + 2\Gamma^3(\hat{\xi}+1)\right]}{\left[\Gamma(2\hat{\xi}+1) - \Gamma^2(\hat{\xi}+1)\right]^{3/2}} = asym_x \\ \hat{\lambda} = \lvert\hat{\xi}\rvert s_x \left[\Gamma(2\hat{\xi}+1) - \Gamma^2(\hat{\xi}+1)\right]^{-1/2} \\ \hat{\mu} = \bar{x} - \frac{\hat{\lambda}}{\hat{\xi}}\left[1 - \Gamma(\hat{\xi}+1)\right] \end{cases}$ La première équation est résolue numériquement.	Méthode numérique. Préférable car l'existence des moments n'est pas assurée.
Exponentielle *Exp*(x_0, λ)	Débit SUPSEUIL	x_0 supposé connu (valeur du seuil) $\hat{\lambda} = \bar{x} - x_0$	idem moments
Pareto Généralisée *GP*(x_0, λ, ξ)	Débit SUPSEUIL	x_0 supposé connu (valeur du seuil) $\hat{\lambda} = \frac{1}{2}(\bar{x} - x_0)\left[\frac{(\bar{x} - x_0)^2}{s_x^2} + 1\right]$ $\hat{\xi} = \frac{1}{2}\left[\frac{(\bar{x} - x_0)^2}{s_x^2} - 1\right]$	Méthode numérique. Préférable car l'existence des moments n'est pas assurée.

Tableau 2-1. Formulaire des estimateurs pour les principales lois utilisées en hydrologie. $\bar{x}$ désigne la moyenne empirique des observations, s_x l'écart-type empirique et $asym_x$ le coefficient d'asymétrie empirique.

3 Calcul de débits caractéristiques

Les sections 1 et 2 ont permis de mettre en place tous les outils statistiques nécessaires au calcul de débits caractéristiques. Cette section a pour but d'illustrer l'application de ces outils à des problématiques hydrologiques. En particulier, nous décrirons les différences entre l'analyse réalisée par l'hydrologue et l'analyse statistique classique présentée dans les sections précédentes. De plus, nous insisterons sur les pièges à éviter dans le calcul de débits caractéristiques, et surtout dans leur interprétation.

Cette troisième section est organisée de la manière suivante. La section 3.1 définit le concept de **période de retour** qui est abondamment utilisé en hydrologie (et plus généralement, dans de nombreux domaines liés à la gestion du risque). La section 3.2 résume les différentes étapes précédant le calcul de débits caractéristiques. Enfin, la section 3.3 discute du **contrôle** et de la **validation** des estimations effectuées.

3.1 Quantiles et périodes de retour

3.1.1 Définition

Les sections précédentes ont permis de définir le quantile d'ordre p d'une distribution (par exemple, la distribution des débits de crue MAXAN), où p est la probabilité au non-dépassement. Ces quantiles sont des descripteurs du comportement des débits extrêmes sur un site bien précis. Dans le domaine de l'hydrologie, et plus généralement dans celui de la gestion des risques, la notion de probabilité au non-dépassement est généralement remplacée par la notion de **période de retour**.

A l'origine, ce concept est issu de la théorie du renouvellement (Cox 1966), et est défini comme l'espérance mathématique du temps d'attente jusqu'à la prochaine panne d'un composant. Appliquée au contexte des quantiles de crue, par exemple, la période de retour T relative à un débit x est donc définie comme l'espérance du temps d'attente jusqu'au prochain dépassement de x. Appliquée au contexte des quantiles d'étiage, la période de retour T relative à un débit x est définie comme l'espérance du temps d'attente jusqu'au prochain non-dépassement de x.

En hydrologie des crues, l'interprétation donnée à la notion de période de retour est la suivante : la période de retour T d'un événement est la durée moyenne qui sépare deux crues supérieures ou égales à cet événement. Inversement, une crue de période de retour T est une crue qui, en moyenne, est égalée ou dépassée toutes les T unités de temps. Nous verrons cependant ultérieurement que, dans certaines situations (cas non-stationnaire), cette interprétation n'est pas équivalente à la définition originelle de la période de retour.

En général, l'unité de temps est l'année, et des relations simples lient les probabilités de non-dépassement et les périodes de retour. Pour simplifier, on pourra retenir qu'une valeur de période de retour T ans à une probabilité $1/T$ d'être dépassée chaque année, soit une probabilité de non-dépassement de $1\text{-}1/T$. La crue décennale (T=10 ans) a ainsi une probabilité annuelle de dépassement de 0.1, soit une probabilité de non-dépassement de 0.9.

De manière plus mathématique, la relation liant probabilité au non-dépassement et période de retour est la suivante :

$$1 - p = \frac{1}{T} \qquad (3\text{-}1)$$

Notons qu'une correction doit être apportée à la relation ci-dessus dans le cas de méthodes d'échantillonnage ne conduisant pas à une unique valeur par an. Par exemple, dans le cas de l'échantillonnage SUPSEUIL conduisant à θ valeurs par an en moyenne, cette relation devient :

$$1-p=\frac{1}{\theta T} \qquad (3\text{-}2)$$

Enfin, concluons en remarquant que les définitions données ci-dessus ne sont valables que pour des variables conduisant à des valeurs élevées pour les événements à risque (variables de crue par exemple). Typiquement, pour les variables d'étiage, l'étiage décennal a une probabilité de non-dépassement de 0.1 (et non de 0.9 !).

3.1.2 Discussion

Il est bien important d'avoir conscience que malgré l'utilisation du terme « période », il n'y a pas de périodicité déterministe dans l'occurrence des événements hydrologiques : il est par exemple tout à fait possible d'avoir plusieurs crues de période de retour 10 ans au cours de la même année. Il faut donc bien garder à l'esprit que la période de retour n'est rien d'autre qu'une probabilité transformée.

Afin d'illustrer les pièges à éviter dans l'interprétation des périodes de retour, le lecteur est encouragé à réfléchir aux questions suivantes (les réponses seront données ultérieurement) :

Q1. Quelle est la probabilité d'observer une crue décennale (ou plus) cette année, sachant qu'il n'y en a pas eu depuis 20 ans ?

Q2. Quelle est la probabilité d'observer une crue décennale (ou plus) cette année, sachant qu'il y en a eu une l'année dernière ?

Q3. Quelle est la probabilité d'observer au moins une crue décennale (ou plus) en 10 ans ?

Q4. Je suis gérant d'un aménagement à risque, qui aura une durée de vie de 80 ans. L'aménagement est dimensionné pour résister à une crue millénale, au-delà, il sera endommagé. Quelle est la probabilité que l'aménagement soit endommagé au cours de sa vie ?

Q5. Je suis préfet d'une région dans laquelle 10 communes sont inondables par toute crue supérieure à la décennale. En admettant que les crues surviennent de manière indépendante sur chaque commune, quelle est, chaque année, la probabilité de devoir faire face à au moins une inondation ?

Q6. Je suis gérant d'un parc de 10 aménagements à risque, qui auront tous une durée de vie de 80 ans. Tous les aménagements sont dimensionnés pour résister à une crue millénale, au-delà, ils seront endommagés. Quelle est la probabilité qu'au moins un de mes aménagements soit endommagé au cours de sa vie ?

Signalons également la difficulté qu'il peut exister à interpréter une période de retour dans un contexte non-stationnaire, c'est à dire si la distribution des observations évolue avec le temps. La stationnarité des phénomènes hydrologiques reste aujourd'hui une hypothèse posée par la grande majorité des méthodes de prédétermination, mais elle pourrait être remise en cause en cas de changements marqués dans le bassin versant (occupation du sol par exemple), ou en cas d'impact du changement climatique sur le régime des rivières. Dans ce contexte, l'interprétation « une crue décennale est dépassée en moyenne tous les 10 ans » n'a plus guère de sens, puisque la fréquence de dépassement va elle-même dépendre du temps ! Le problème vient ici de l'utilisation d'une moyenne temporelle (« tous les 10 ans »), qui n'a de sens que si la distribution des crues n'évolue pas dans le temps. Par contre, la définition originale d'une période de retour (« espérance du temps d'attente jusqu'au prochain dépassement ») reste

valable en un temps *t* donné, car elle fait appel à une espérance mathématique (et non à une moyenne temporelle). De même, il est possible de calculer, en un temps *t* donné, une probabilité annuelle de dépassement (cf. Renard 2008, pour plus de détails).

Concluons cette section en donnant les réponses aux questions posées en début de section :

Q1. Réponse: 0.1. Les crues surviennent indépendamment d'une année sur l'autre, le fait qu'il n'y en ait pas eu depuis longtemps ne change rien à la probabilité d'en observer une cette année.

Q2. Réponse: 0.1, pour les mêmes raisons !

Q3. Réponse: Environ 0.65. Pour effectuer ce calcul, il est plus simple de calculer la probabilité de ne pas observer de crue décennale. Chaque année, cette dernière probabilité vaut 0.9. Les crues survenant de manière indépendante chaque année, la probabilité de ne pas observer de crue décennale sur 10 ans vaut donc $0.9^{10} \approx 0.35$. La probabilité d'en observer au moins une est donc de 1-0.35=0.65.

Q4. Réponse: Environ 0.077. En suivant un raisonnement identique à celui ci-dessus, cette probabilité vaut $1\text{-}0.999^{80}$

Q5. Réponse: Environ 0.65, le calcul étant le même qu'en Q3. Dit autrement, la période de retour de l'événement *« il y a une crue décennale quelque part sur au moins une de mes 10 communes* » vaut environ 1/0.65, soit environ… un an et demi !

Q6. Réponse: Environ 0.55. Par un raisonnement similaire aux précédents, la formule est $1\text{-}(0.999^{80})^{10}$

Bien que parfois un peu trop simplistes (en particulier, l'aléa n'est généralement pas indépendant spatialement), ces exemples permettent d'illustrer le fait que l'exposition au risque est généralement répétée dans le temps et dans l'espace, ce qui demande d'aller au-delà de la simple notion de période de retour locale afin de réellement appréhender le risque.

3.2 Calcul des débits caractéristiques

3.2.1 Formulaire

La définition d'une période de retour donnée en section 3.1 implique que les débits caractéristiques peuvent être calculés simplement comme des quantiles d'ordre *p* de la distribution de la variable étudiée. Par exemple :

- ✓ Pour la variable « MAXAN », la valeur de période de retour 10 ans (resp. *T* ans) est égale au quantile d'ordre 0.9 (resp. 1-1/*T*) de la distribution ajustée aux maxima annuels observés.
- ✓ Pour la variable « SUPSEUIL » avec 2 événements par an en moyenne, la valeur de période de retour 10 ans (resp. *T* ans) est égale au quantile d'ordre 0.95 (resp. 1-1/2*T*) de la distribution ajustée aux dépassements de seuils observés.
- ✓ Pour la variable d'étiage « Débit minimum annuel », la valeur de période de retour 10 ans (resp. *T* ans) est égale au quantile d'ordre 0.1 (resp. 1/*T*) de la distribution ajustée aux minima annuels observés.

Ces quantiles peuvent être calculés en inversant la fonction de répartition de la variable étudiée. En effet, un quantile q_p vérifie, par définition, $F(q_p)=p$, soit $q_p = F^{-1}(p)$

Pour les principales distributions utilisées dans ce document, les formules suivantes permettent de calculer les quantiles d'ordre *p*:

- ✓ Loi Normale: pas d'expression analytique. Les solutions de l'équation $F(q_p)=p$ doivent être approchées par des méthodes numériques. La plupart des logiciels proposent des fonctions pour cela.
- ✓ Loi Log-Normale: comme ci-dessus, pas d'expression analytique.
- ✓ Loi de Gumbel : $q_p = \mu - \lambda \log(-\log(p))$
- ✓ Loi GEV: $q_p = \mu + \frac{\lambda}{\xi}\left(1-(-\log(p))^{\xi}\right)$ (3-3)
- ✓ Loi Exponentielle: $q_p = x_0 - \lambda \log(1-p)$
- ✓ Loi de Pareto Généralisée: $q_p = x_0 + \frac{\lambda}{\xi}(1-(1-p)^{\xi})$

Dans la pratique, on calcule ces quantiles en remplaçant les paramètres par leur estimation (cf. formulaire dans le Tableau 2-1).

3.2.2 Résumé : mise en œuvre de la chaîne de traitement pour le calcul des débits caractéristiques

Le calcul d'un débit caractéristique, depuis la récupération de la chronique hydrologique jusqu'aux calculs statistiques, peut finalement se résumer aux étapes suivantes :

1. **Récupération de la chronique de données**. Il sera bénéfique à cette étape d'effectuer une critique préliminaire des données de base. Pour cela, il faut prendre le temps de visualiser la chronique (ce qui permet parfois de détecter des erreurs de saisie), de prendre connaissance des méta-données décrivant l'historique de la station hydrométrique, voire de discuter avec le producteur des données. En cas de doute, des techniques statistiques plus poussées peuvent être mises en oeuvre (quelques exemples seront données ultérieurement en section 5.2.2).
2. **Choix d'une variable hydrologique et création de l'échantillon**. Dans certains cas, la variable à étudier est imposée (réglementairement par exemple). Si ce n'est pas le cas, le choix de la variable d'étude n'est ni anodin ni trivial. Par exemple, si l'on s'intéresse au régime des étiages, de nombreuses variables peuvent être calculées pour décrire la durée, l'intensité ou le déficit de volume des étiages (cf. section 1.1). Ainsi, ce que l'on appellera (de manière abusive car ambiguë) « l'étiage centennal » correspondra à des phénomènes potentiellement bien différents suivant que l'on parle de durée, de volume ou d'intensité. En conséquence, il peut s'avérer intéressant de caractériser le phénomène qu l'on souhaite caractériser (crue ou étiages) par plusieurs variables hydrologiques.
3. **Choix d'une distribution**. La distribution à utiliser dépend évidemment de la variable étudiée. Dans certains cas, la théorie probabiliste guide le choix de la distribution (théorème central limite ou théorème des valeurs extrêmes, cf. section 2). Dans d'autres cas, l'expérience de l'hydrologue le conduira à favoriser telle ou telle distribution (en fonction de la variable étudiée, du contexte régional hydrologique et climatique, etc.). Il s'avère parfois nécessaire de tester plusieurs distributions et de sélectionner celle qui semble la plus en adéquation avec les observations (ce qui n'est d'ailleurs pas toujours évident à évaluer).

4. **Estimation des paramètres de la distribution**. Une fois la distribution choisie, il convient d'en estimer les paramètres. Là encore, plusieurs estimateurs coexistent et le choix d'un estimateur particulier n'est pas trivial (cf. section 2.7). On pourra utiliser plusieurs estimateurs afin d'apprécier la sensibilité des résultats au choix de l'estimateur.
5. **Calcul des quantiles**. Les débits caractéristiques sont finalement calculés à l'aide des quantiles présentés dans la section précédente 3.2.1.

Précisons qu'une fois le calcul des débits de référence effectué, le travail de l'hydrologue n'est pas terminé… En particulier il devra tout d'abord **contrôler** et **valider** ses estimations, afin d'en vérifier la cohérence avec les observations. Ceci fera l'objet de la prochaine section 3.3. Dans la mesure du possible, il devra également tenter d'apprécier, voire de quantifier les **incertitudes** entourant ses estimations. Le traitement de l'incertitude sera discuté en section 4.

3.3 Contrôle et Validation

Le calcul de débits caractéristiques repose sur un certain nombre d'hypothèses, dont une des principales est que les observations sont des réalisations d'une distribution choisie *a priori*. Même si le choix d'une distribution repose parfois sur des bases théoriques solides (par exemple, théorème des valeurs extrêmes), il convient de vérifier que la distribution est en adéquation avec les observations. L'objet de cette section est de discuter la manière d'effectuer un tel jugement. Nous commencerons par le **contrôle**, qui consiste à comparer la distribution estimée avec les observations ayant été utilisées pour effectuer cette estimation. Nous parlerons dans un second temps de **validation**, qui vise à mettre la distribution estimée à l'épreuve de nouvelles observations (non utilisées pour effectuer l'estimation).

3.3.1 Contrôle : la courbe des quantiles (graphique d'ajustement)

Pour juger de la qualité de l'ajustement, les hydrologues représentent en général sur un même graphique la fonction de répartition théorique (donnée par la distribution estimée) et la fonction de répartition empirique (directement calculée à partir des données ayant servi à l'estimation). Cependant, par rapport aux représentations classiques que nous avons vues précédemment (cf. sections 1.2 et 2), il existe quelques originalités. Nous allons illustrer ces différences à partir du jeu de données suivant, représentant les crues supérieures à 72 $m^3.s^{-1}$ (échantillonnage SUP-SEUIL) entre 1960 et 2002 pour l'Ubaye à Barcelonnette (549 km^2) :

QJX classé	Année	jour
72.6	1985	158
73.5	1973	125
74	1979	152
75.6	1994	310
75.8	1984	174
77.7	1986	141
79.9	1977	165
83	1972	158
84.1	1983	160
84.7	1994	267
91	1983	136
98.1	2000	165
100	2001	151
111	2000	289
112	1978	162
120	1963	320

Tableau 3-1. Tableau de données.

Nous allons ajuster une loi exponentielle à cet échantillon. Les formules du Tableau 2-1 donnent $\hat{\lambda} = \overline{X}_s - x_0 = 88.3 - 72 = 16.3$. On peut donc à présent tracer, sur un même graphique, la fonction de répartition estimée et la fonction de répartition empirique (c'est à dire la courbe des fréquences cumulées, dont nous ne représentons que les points observés) :

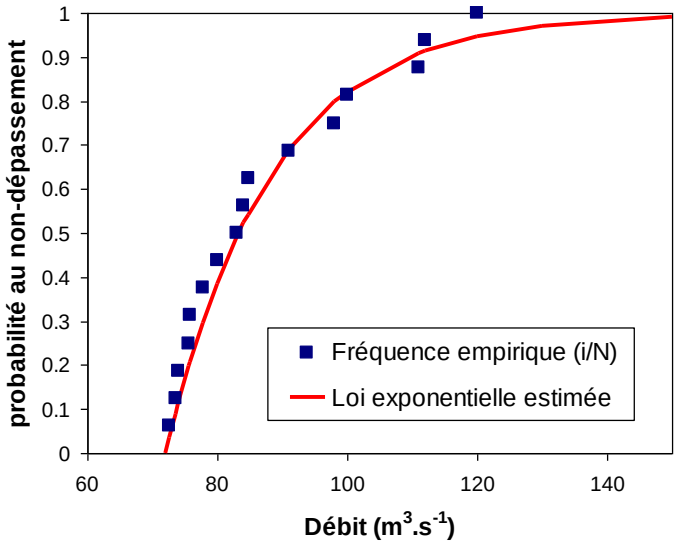


Figure 3-1. Fonction de répartition empirique et théorique.

En hydrologie, il est fréquent que les axes soient inversés, c'est à dire que l'on portera en abscisse la fréquence cumulée (probabilité de non-dépassement), et en ordonnée le débit correspondant :

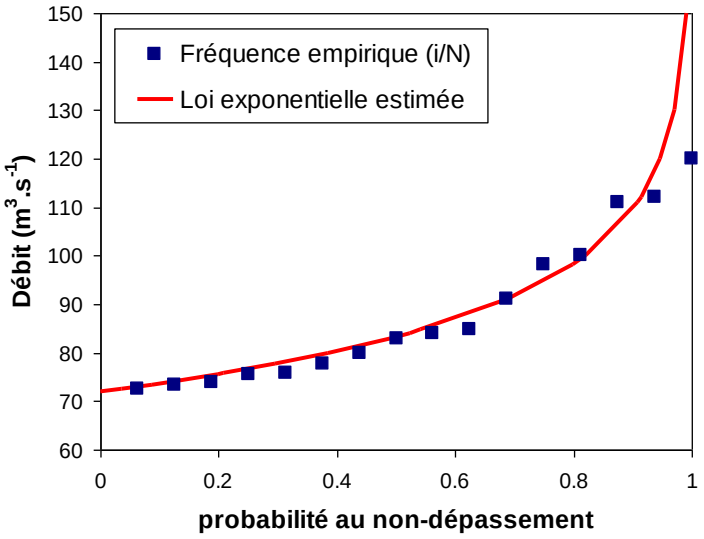


Figure 3-2. Fonction de répartition empirique et théorique, version 2.

Le seconde originalité est liée au calcul des fréquences cumulées empiriques : nous avons vu en section 2.1.1 que la formule *i/N* pour le calcul des fréquences cumulées n'était pas l'unique solution. En hydrologie, on lui préfère souvent les formules suivantes : $\frac{i-0.3}{N+0.4}$ ou $\frac{i-0.5}{N}$ (formule de Hazen).

De manière importante, notons que le choix de la formule de fréquence empirique n'a absolument aucun effet sur l'estimation des paramètres, puisque les estimateurs ne dépendent pas des fréquences empiriques.

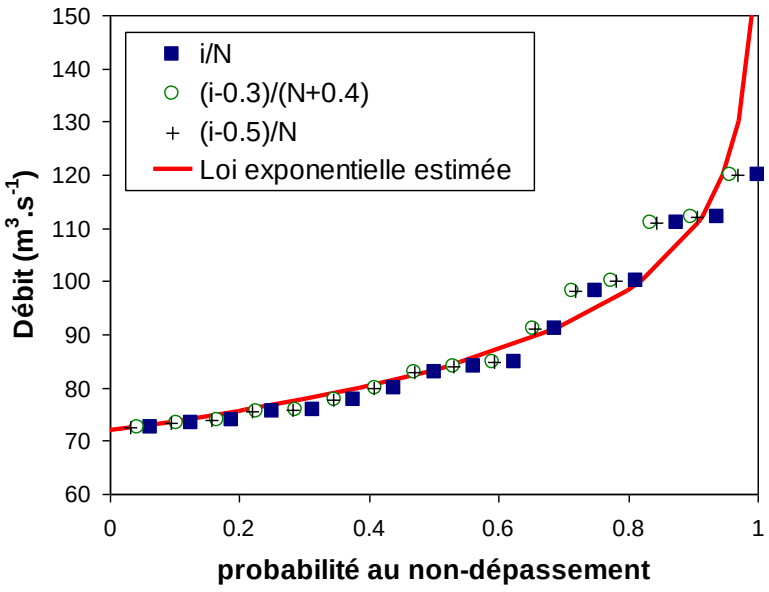


Figure 3-3. Fonction de répartition empirique et théorique, version 3.

Enfin, il est d'usage de linéariser la figure ci-dessus, en effectuant une transformation sur l'axe des abscisses. Pour un échantillonnage SUPSEUIL, étant donné la définition de la loi exponentielle, $F(x)=\begin{cases}1-e^{-\frac{x-x_0}{\lambda}} & \text{si } x>x_0 \\ 0 \text{ sinon}\end{cases}$, on tracera les débits en fonction de la variable $-\log(1-p)$. Pour un échantillonnage MAXAN, le changement de variable consistera à reporter en abscisses la variable $-\log(-\log(p))$.

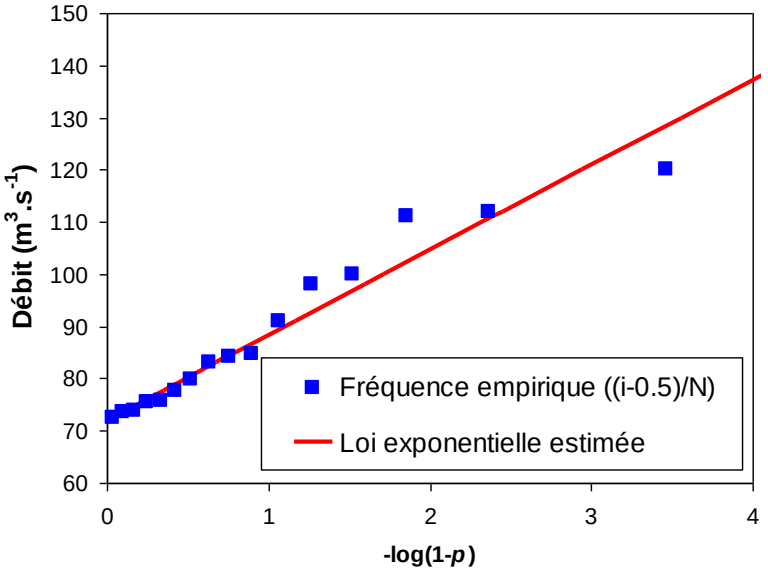


Figure 3-4. Fonction de répartition empirique et théorique, version 4.

C'est sur ce dernier graphique (que l'on appellera en hydrologie la **courbe des quantiles**) que se juge l'adéquation entre distribution estimée et données. La représentation linéarisée permet de zoomer sur les valeurs les plus fortes, ce qui facilite le jugement sur les plus forts quantiles. Il convient néanmoins de rester prudent lors de l'interprétation de cette courbe. En particulier, il faut éviter de trop se focaliser sur les derniers points, car le positionnement de ces points en fréquence empirique est très incertain (cf. exemple en section 2.1.3). Il n'est donc pas anormal qu'un ou deux points s'éloignent de la droite théorique : cela peut être un simple effet de la variabilité d'échantillonnage. Par contre, on sera plus attentif aux différences de forme entre les courbes empirique et théorique. En particulier la présence d'une courbure (qui ne se limite pas à un unique point !) est souvent symptomatique d'un manque d'adéquation entre la distribution estimée et les observations.

3.3.2 Validation : motivations

Une fois que l'étape de contrôle a permis de juger que la distribution estimée **décrit** les données de manière satisfaisante, il reste à vérifier qu'elle sera également capable de **prédire** (au sens de prédéterminer) de nouvelles observations. Mais pourquoi cela est-il nécessaire ? Si l'étape de contrôle précédente s'est avérée satisfaisante, cela ne suffit-il pas à prouver le pouvoir prédictif de la distribution estimée ?

Un contre-exemple va permettre de se convaincre du contraire. Pour quatre bassins versants que nous nommerons BV1 à BV4, des échantillons de 50 années de données MAXAN sont utilisés. Deux distributions différentes, que nous nommerons D1 et D2, sont ajustées à ces

données. La Figure 3-5 illustre ces ajustements (l'axe des abscisses est ici gradué en période de retour). Pour le moment, la localisation de ces bassins versants et les distributions utilisées ne sont pas révélées.

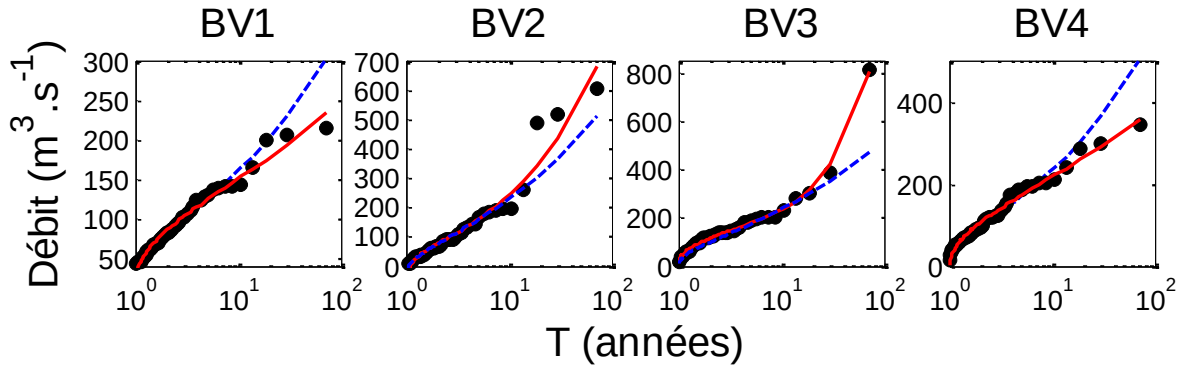


Figure 3-5. Courbes des quantiles pour les distributions D1 (trait plein rouge) et D2 (tirets bleus) pour quatre bassins versants.

Sur la base des graphiques de contrôle de la Figure 3-5, la distribution D1 semble fournir une très bonne description des observations. En particulier, l'ajustement aux fréquences empiriques les plus élevées est excellent. La distribution D2, quant à elle, donne un résultat acceptable, malgré quelques points qui s'éloignent de la courbe théorique.

Afin de procéder à la **validation** de ces distributions, 450 nouvelles valeurs MAXAN sont utilisées pour chaque bassin versant (à ce stade, le lecteur pourra légitimement se demander si ces bassins versants ne sont pas situés dans un monde imaginaire étant donné la taille des chroniques). La Figure 3-6 montre les fréquences empiriques de ces nouvelles observations. Un point extrêmement important ici est que les distributions D1 et D2 dans la Figure 3-6 n'ont pas été réajustées à ces 450 nouvelles observations : il s'agit des mêmes distributions qu'en Figure 3-5, ajustées sur les 50 valeurs initiales.

Les résultats de la Figure 3-6 peuvent à première vue laisser perplexe : alors que la distribution D1 donnait des résultats quasi-parfaits en contrôle, on observe en validation de très fortes différences avec les fréquences empiriques (notamment sur BV3 et BV4). Inversement, les performances de la distribution D2 restent acceptables. Que s'est-il passé ? La distribution des données aurait-elle changé entre la période de contrôle et la période de validation ?

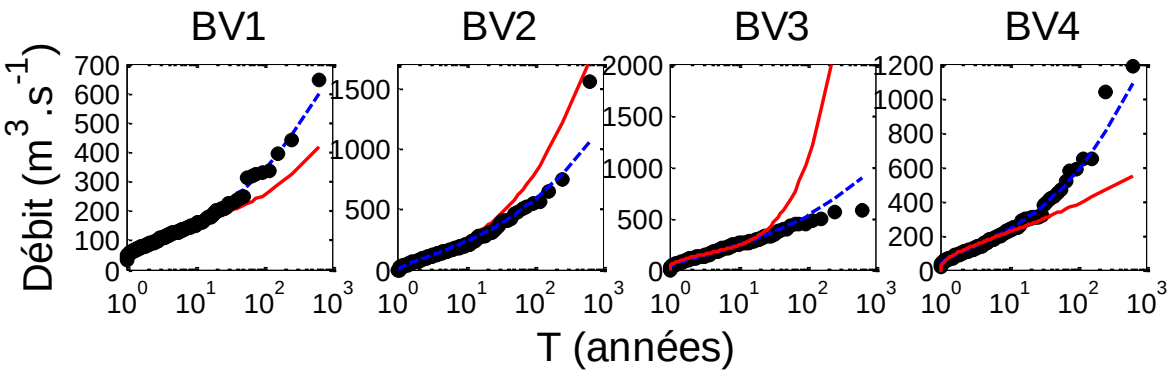


Figure 3-6. Courbes des quantiles pour les distributions D1 (trait plein rouge) et D2 (tirets bleus) pour quatre bassins versants. Les fréquences empiriques (points noirs) correspondent à des données de validation, non utilisées pour l'ajustement des distributions.

Il est temps à présent de révéler l'origine de ces données : elles ont toutes été simulées à partir de la distribution D2, en contrôle comme en validation. En d'autres termes, les différences avec les fréquences empiriques qui apparaissent pour la distribution D2 en Figure 3-5 et Figure 3-6 sont uniquement dues à la variabilité d'échantillonnage.

La distribution D1, quant à elle, correspond à l'ajustement d'un polynôme de degré 5 aux fréquences empiriques (points noirs) de la Figure 3-5. La flexibilité d'un tel polynôme permet de s'ajuster très fidèlement aux fréquences empiriques. Malheureusement, cet ajustement est en fait *trop* fidèle : comme signalé précédemment, les fréquences empiriques des plus fortes valeurs sont très incertaines, et peuvent donner une estimation très biaisée de la vraie probabilité au non-dépassement. Par exemple, les 50 maxima annuels observés sur le bassin BV3 et utilisés pour l'ajustement de la distribution D1 sont représentés en Figure 3-7. Une valeur exceptionnelle de 800 $m^3.s^{-1}$ est observée. La fréquence empirique de cette valeur est égale à $\frac{i-0.3}{N+0.4}=\frac{50-0.3}{50+0.4}\approx 0.986$, soit une période de retour empirique d'environ 72 ans. Puisque l'on connaît la vraie distribution des données, on peut calculer la vraie période de retour de cette valeur, qui est d'environ 475 ans ! Le caractère exceptionnel de cette valeur est d'ailleurs plus apparent lorsqu'on observe la totalité des données du bassin BV3 (Figure 3-8). Ceci illustre qu'il n'est pas nécessaire (ni même recommandable) de vouloir à tout prix s'ajuster au plus près du débit le plus fort.

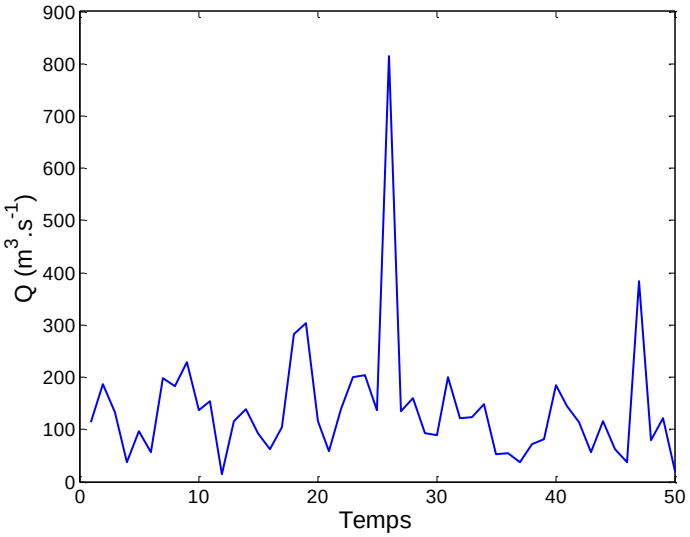


Figure 3-7. Maxima annuels sur le bassin BV3 (données utilisées pour l'estimation).

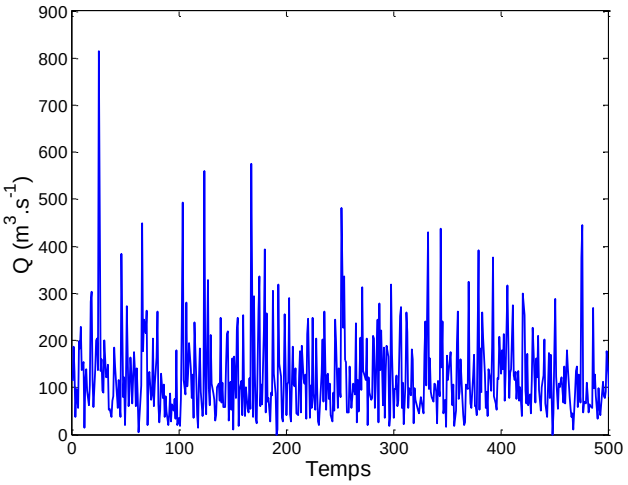


Figure 3-8. Maxima annuels sur le bassin BV3 (données utilisées pour l'estimation + données de validation).

Cet exemple synthétique permet d'illustrer la différence entre le **pouvoir descriptif** et le **pouvoir prédictif** d'une distribution. Le pouvoir descriptif se rapporte à la capacité d'une distribution de correctement décrire les observations qui ont été utilisées pour l'estimation. C'est une qualité nécessaire : une distribution incapable de correctement décrire les données utilisées pour l'estimation n'a guère de chance de mieux décrire de nouvelles observations. Ce n'est cependant pas une qualité suffisante, comme le montre l'exemple ci-dessus : une distribution peut avoir un très bon pouvoir descriptif mais aucun pouvoir prédictif ! Ce type de distribution est généralement appelé distribution **sur-paramétrée**, pour désigner le fait que l'on a utilisé trop de paramètres (6 dans le cas du polynôme de degré 5 de la distribution D1) pour améliorer le pouvoir descriptif de la distribution, au détriment de son pouvoir prédictif.

3.3.3 Stratégies de validation

Les outils utilisables en validation sont en fait identiques à ceux que l'on utilise à l'étape de contrôle, mais avec une différence capitale : les données de validation ne doivent pas avoir été

utilisées pour l'estimation. C'est sur ces données, et sur ces données seulement, que l'on peut juger le pouvoir prédictif d'une distribution.

La question n'est donc pas de trouver de nouveaux outils, mais plutôt de savoir où trouver ces données de validation. En pratique, on souhaite généralement utiliser toutes les données disponibles pour l'estimation des paramètres, afin d'obtenir des estimateurs aussi fiables que possible. Cette approche ne laisse malheureusement pas de données disponibles pour la validation. On pourra donc recommander la stratégie suivante : même si l'estimation finale est basée sur toutes les données disponibles, on pourra chercher à évaluer le pouvoir prédictif d'une méthode en excluant volontairement certaines données de l'estimation, et en les préservant pour la validation. Par exemple, si l'on dispose de 50 années de données, on pourra effectuer l'estimation sur 25 années tirées au hasard, et vérifier que cette estimation reste acceptable pour les 25 autres années.

Signalons l'existence d'un projet de recherche national qui vise justement à effectuer cette validation pour les principales méthodes de prédétermination utilisées en France (https://extraflo.cemagref.fr/). Cette évaluation est basée sur l'utilisation de nombreuses longues séries de données, permettant une validation plus poussée des méthodes existantes.

4 Incertitudes

Le calcul de débits caractéristiques est affecté par de nombreuses sources d'incertitude, qui peuvent devenir très importantes pour les périodes de retour élevées. Il est donc important de chercher à quantifier cette incertitude, afin de pouvoir juger de la confiance que l'on peut accorder aux estimations.

Malheureusement, la quantification des incertitudes réclame bien souvent des outils statistiques et probabilistes trop avancés pour être présentés en détail dans le cadre de cette formation. En conséquence, l'objectif de cette section n'est pas de présenter un catalogue de méthodes immédiatement applicables en pratique. Nous privilégions plutôt une discussion sur l'origine des incertitudes, et nous nous contentons au niveau méthodologique de donner quelques pistes que le lecteur intéressé pourra creuser.

Cette section est organisée de la manière suivante. Nous commençons par décrire les principales sources d'incertitudes en section 4.1. Les différentes approches pour tenter de quantifier les incertitudes sont ensuite discutées en section 4.2. Enfin, nous étudions plus en détail les méthodes probabilistes pour quantifier l'incertitude d'échantillonnage en section 4.3.

4.1 Les différentes sources d'incertitudes en hydrologie

4.1.1 Incertitudes liées à la mesure

Contrairement à d'autres variables hydro-météorologiques comme les pluies, les températures ou le vent, les chroniques de débit ont la particularité d'être constituées de mesures indirectes : il s'agit en fait de chroniques de hauteurs d'eau, qui sont transformées en chroniques de débit via la courbe de tarage (Figure 4-1). Cette dernière est construite à partir de mesures ponctuelles de débits (les jaugeages) effectuées par exemple au moulinet, par ADCP ou par dilution. Cette particularité induit plusieurs sources d'incertitudes spécifiques.

La première source d'incertitude est liée aux mesures effectuées lors des jaugeages. La mesure de débit peut notamment être entachée d'erreurs non négligeables, surtout en crue et/ou en étiage. Par exemple, en basses eaux, les faibles vitesses d'écoulement sont parfois difficiles à mesurer. Le jaugeage d'une crue comporte également de nombreuses difficultés opératoires, liées aux conditions d'écoulement turbulentes ou à la présence de débris dans la rivière. Notons que la mesure de hauteur effectuée lors du jaugeage peut également être entachée d'erreurs, que l'on considère généralement comme moins importantes que celles commises pour la mesure du débit.

La seconde source d'incertitude est liée à l'établissement de la courbe de tarage. Le choix de la formule mathématique de la courbe de tarage est une première difficulté (des considérations hydraulique permettent généralement de guider cette étape). L'ajustement de la courbe choisie réclame ensuite d'estimer un certain nombre de paramètres, ce qui induit une nouvelle source d'incertitude. Enfin, rappelons que les couples hauteurs/débits utilisés pour cette estimation sont eux-mêmes potentiellement affectés d'erreurs non négligeables.

Ces incertitudes sont généralement particulièrement forte dans le domaine de l'extrapolation en crue et en étiage). Par nature, les événements extrêmes sont rares, il n'est donc évidemment pas possible de jauger une crue décennale tous les ans ! De plus, effectuer un jaugeage en crue est une opération délicate, voire périlleuse (pour le matériel et les hommes). Ainsi, les débits très faibles ou très forts peuvent ne jamais avoir été jaugés, ce qui conduit,

lors de la survenue d'un tel événement, à calculer le débit par une extrapolation de la courbe de tarage éloignée du domaine jaugé, d'où un important risque d'erreur.

Enfin, l'instabilité de certaines relations hauteur-débit est une autre source d'incertitude. Dans de tels cas, on est amené à revoir fréquemment la courbe de tarage, alors que son établissement réclame un nombre suffisant de jaugeages sur une gamme de débits variée, ce qui peut réclamer plusieurs années ! Les périodes de validité de chaque courbe sont alors elles-même incertaines.

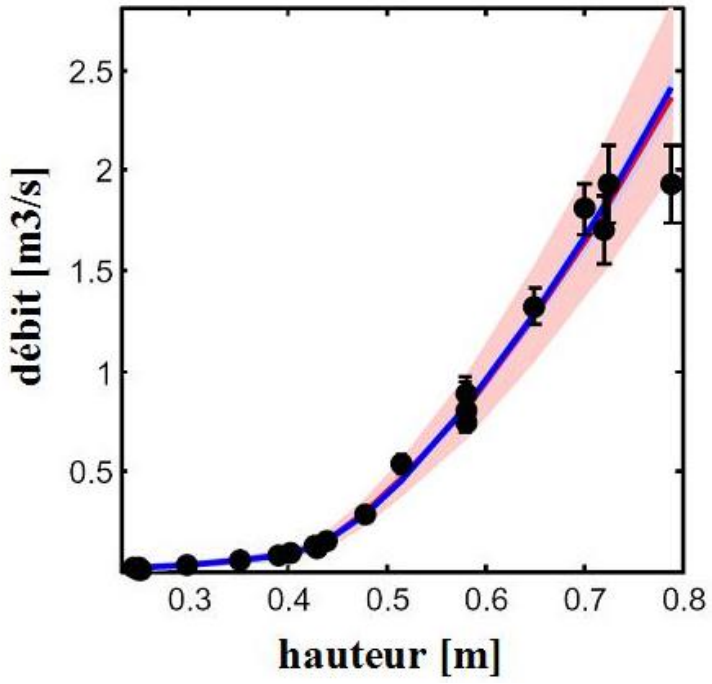


Figure 4-1. Exemple de courbe de tarage (le ruisseau de Charbonnières à Charbonnières-les-bains, 23 km²)

4.1.2 Incertitude d'échantillonnage

Comme discuté en section 2.1.3, toute estimation statistique est affectée par la variabilité d'échantillonnage, due au fait que l'échantillon analysé ne reflète qu'une partie de l'ensemble des réalisations possibles de la rivière, et que sa représentativité n'est donc pas absolue. En conséquence, l'estimation des débits caractéristiques est nécessairement entachée d'incertitudes importantes, surtout si les périodes de retour visées sont grandes par rapport à la taille des séries disponibles.

La Figure 4-2 fournit une illustration de la variabilité d'échantillonnage, sur la base de 1000 années de données simulées à partir d'une loi de Gumbel. La première période de 50 ans mis en évidence dans cette figure a été relativement peu active, alors que la troisième a vu plusieurs dépassements de la valeur centennale. Il est évident que les débits caractéristiques estimés sur une période ou sur l'autre seront donc différents.

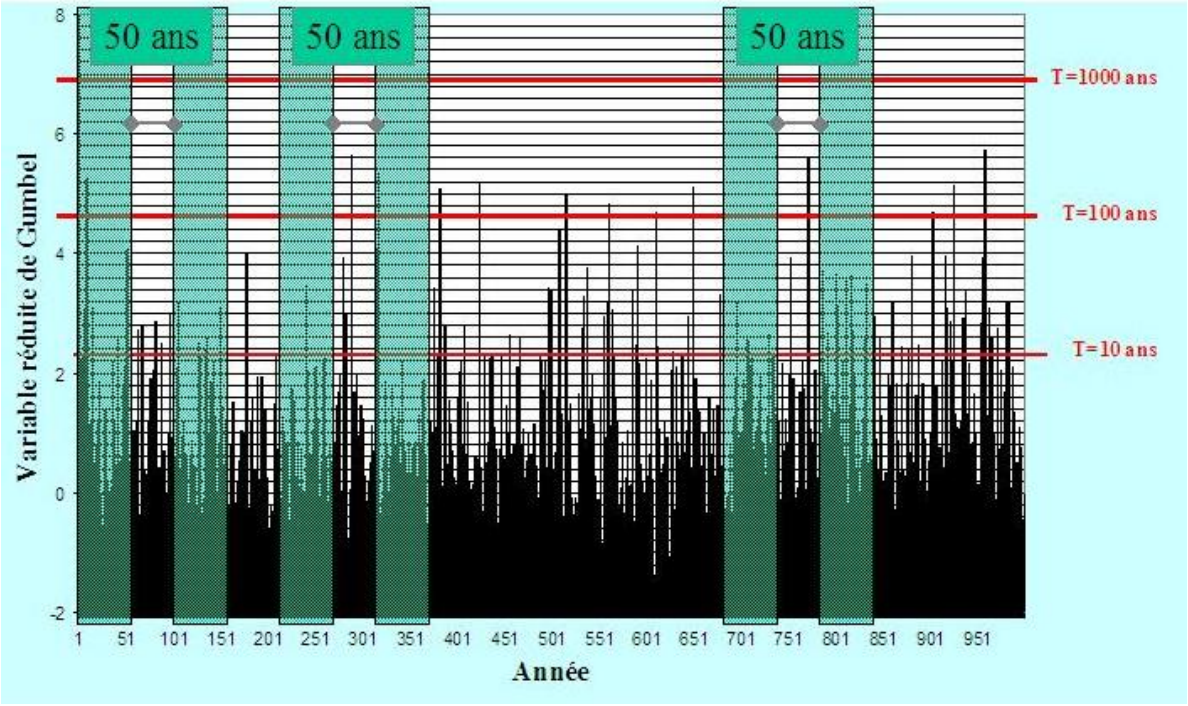


Figure 4-2. Illustration de la variabilité d'échantillonnage.

4.1.3 Incertitudes liées aux hypothèses de modélisation

L'utilisation d'un modèle probabiliste implique de faire un certain nombre d'hypothèses sur les propriétés statistiques des données analysées. Bien qu'il soit possible d'évaluer la pertinence de ces hypothèses (en particulier via le contrôle et la validation), il demeure toujours un doute quant à leur véracité. Par exemple, il est impossible de prouver formellement que tel jeu de données suit une loi de Gumbel : tout au plus peut-on vérifier que cette hypothèse n'est pas incompatible avec les observations.

Les principales hypothèses faites lors de l'estimation probabiliste des débits de référence sont les suivantes :

- **Choix d'une distribution**. Comme expliqué ci-dessus, il est possible de vérifier la pertinence du choix effectué dans le domaine des observations. Ceci est par contre bien plus difficile au-delà, de sorte qu'un doute ne peut que subsister sur la pertinence de la distribution dans le domaine de l'extrapolation. De plus, il est fréquent de constater que plusieurs distributions s'ajustent de manière également satisfaisante aux observations, mais divergent dans le domaine de l'extrapolation. Ceci indique que l'incertitude liée au choix de la distribution peut être importante en extrapolation.
- **Hypothèse d'indépendance**. Cette hypothèse est généralement considérée comme raisonnable pour les variables de crue (à condition d'avoir réalisé l'échantillonnage avec prudence, cf. section 1.1.3). Elle peut par contre être remise en question pour les variables d'étiage voire de moyennes eaux lorsque le contrôle par les nappes d'accompagnement est important. En effet, l'inertie induite par ces nappes crée un effet de mémoire inter-annuelle (la hauteur de nappe de cette année dépend de la hauteur de l'année dernière), qui à son tour rend les données dépendantes. Un exemple d'une telle situation est donné en Figure 4-3. Des études ont permis de mettre en évidence qu'ignorer cette dépendance conduit notamment à une sous-estimation forte de l'incertitude d'échantillonnage. Signalons également l'existence actuelle d'un débat

dans la communauté hydrologique sur l'existence possible de dépendances à longue portée (agissant sur plusieurs dizaines d'années) dans les variables hydro-météorologiques.

- **Hypothèse d'équi-distribution**. L'hypothèse que les observations sont toutes issues d'une unique distribution peut être remise en cause pour plusieurs raisons.
    - Mélange de populations : Les données analysées pourraient être le résultat de phénomènes physiques différents. Par exemple, sur un bassin donné, les crues peuvent résulter de précipitations intenses méditerranéennes liées à des flux de sud, ou des précipitations plus longues liées à des flux d'ouest. Ces deux types de crue ne proviennent donc pas de la même distribution (voir par exemple la méthode SCHADEX Paquet et al. 2006).
    - Variabilité climatique : La distribution de la variable hydrologique considérée peut être impactée par la variabilité climatique à grande échelle (par exemple, phénomène El Niño, oscillation nord-atlantique, oscillation pacifique). Si l'impact de cette variabilité a été clairement mis en évidence dans certaines régions (dans le Pacifique par exemple), son rôle dans l'hydrologie des rivières françaises reste largement incompris.
    - Non-stationnarité : la distribution de la variable hydrologique considérée pourrait évoluer dans le temps, par exemple sous l'effet possible d'un changement d'occupation du sol ou du changement climatique.

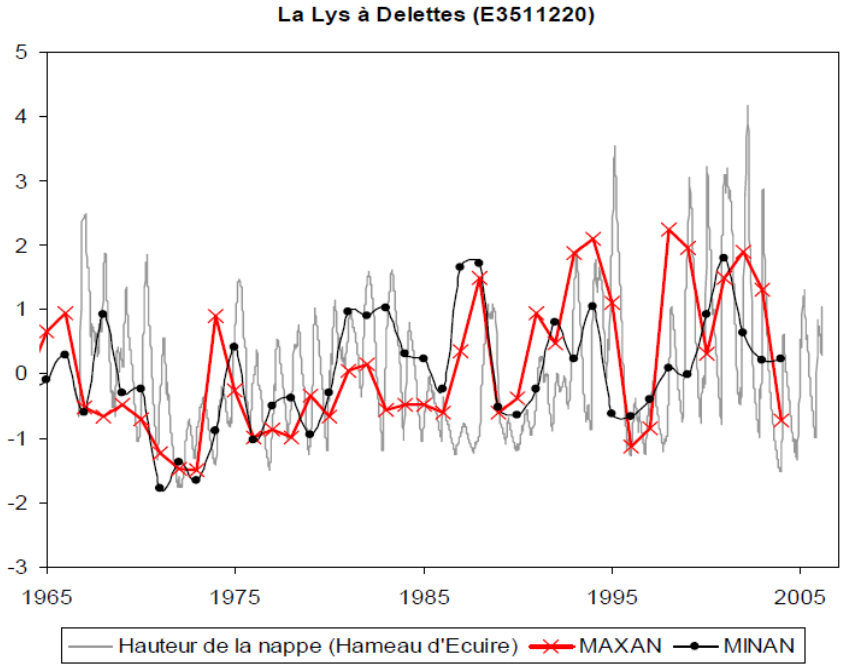


Figure 4-3. Valeurs centrées-réduites des hauteurs mensuelles de nappe, des minima annuels journaliers et des maxima annuels journaliers.

4.2 Quantification des incertitudes : discussion

Ayant accepté l'existence d'incertitudes dans la chaîne de traitements conduisant au calcul de débits caractéristiques, il convient de tenter d'en quantifier l'importance. Cette section discute de quelques approches permettant d'effectuer cette quantification.

4.2.1 Utilisation de scénarios

L'utilisation de scénarios est l'approche la plus simple pour visualiser l'impact d'une incertitude sur le résultat final. Cette approche est généralement utilisée pour les incertitudes de modélisation : elle permet de reporter les résultats obtenus sous plusieurs hypothèses de modélisation, et d'apprécier l'impact de ces hypothèses sur le résultat final.

La Figure 4-4 illustre la comparaison de trois scénarios sur la distribution à utiliser pour modéliser un jeu de données (Loi de Gumbel, loi GEV, loi Log-Normale). Etant donnée la forte incertitude affectant les périodes de retour estimées empiriquement (positionnement horizontal des points noirs), il semble difficile sur la base de cette figure de se prononcer sur la « meilleure » distribution à utiliser. Pourtant, on observe très clairement la forte divergence des estimations en extrapolation, avec une crue millénale estimée qui peut varier du simple au double suivant la distribution choisie. Ce constat suggère que l'incertitude liée au choix du modèle probabiliste peut être extrêmement forte en extrapolation.

Notons que si l'utilisation de scénarios permet de visualiser simplement les incertitudes de modélisation, elle ne quantifie aucunement le caractère plus ou moins probable des différents scénarios. De plus, il est possible que la réalité ne corresponde à aucun des scénarios envisagés.

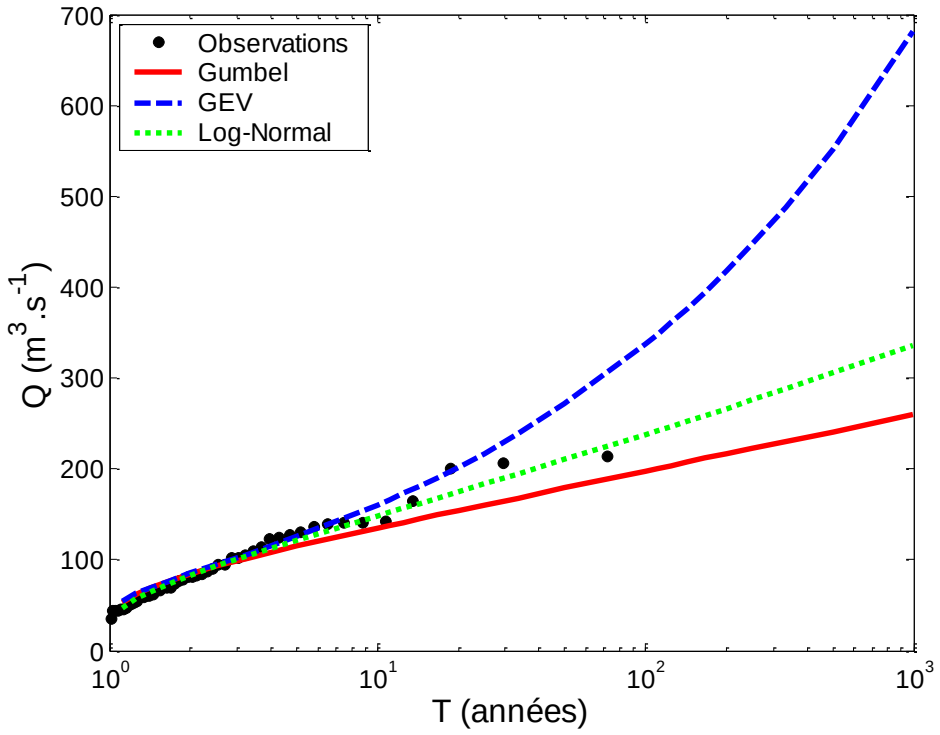


Figure 4-4. Exemple d'utilisation de scénarios : trois scénarios de distributions ajustées aux observations.

Concluons cette section par une des approches par scénarios les plus célèbres du moment : il s'agit des projections de réchauffement global fournies par le GIEC (Figure 4-5 ; cf. http://www.ipcc.ch/ pour plus d'information). Cet exemple est un peu plus complexe que le précédent, puisque deux types de scénarios sont ici croisés : scénarios sur les émissions futures de gaz à effet de serre (GES) et scénarios sur le modèle climatique utilisé. Comme précédemment, cette approche permet de visualiser les incertitudes entourant ces projections – et de se rendre compte qu'elles sont énormes ! – mais ne permet pas d'évaluer la

vraisemblance des différents scénarios. Selon toute vraisemblance, la réalité ne correspondra d'ailleurs à aucun de ces scénarios, mais se situera « quelque part » entre ces différentes projections.

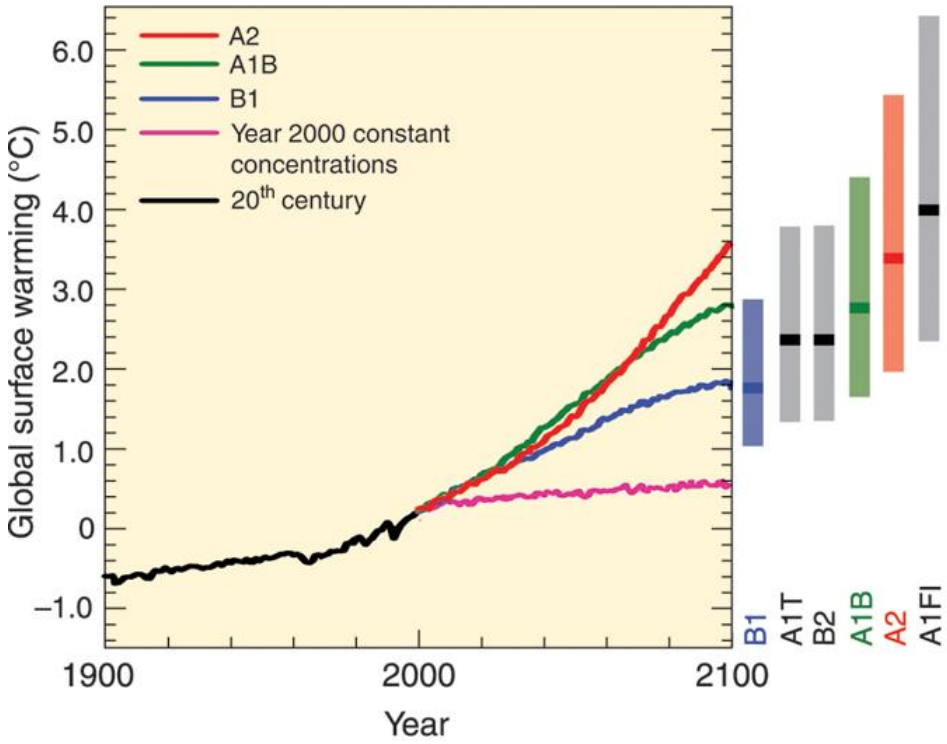


Figure 4-5. Exemple d'utilisation de scénarios : projections de réchauffement de la température globale annuelle. Deux types de scénarios sont croisés : scénarios sur les émissions futures de GES (courbes colorées dans la figure) et scénarios sur les modèles climatiques utilisés (boites verticales dans la marge droite, représentant l'étendue des projections pour plusieurs modèles). Figure disponibles sur http://www.ipcc.ch/

4.2.2 Etudes de sensibilité

Les études de sensibilité peuvent être considérées comme la version « continue » des études par scénarios. On les utilise lorsque l'on souhaite évaluer l'impact d'un facteur quantifiable sur les estimations. Par exemple, les études de sensibilité peuvent être utiles pour évaluer l'impact d'erreurs de mesure potentielles.

A titre d'illustration, nous utilisons les débits maxima annuels du Rhône à Beaucaire (95590 km²) sur la période 1920-2004. Cet exemple est intéressant car l'estimation des débits atteints lors de la grande crue de décembre 2003 a donné lieu a de nombreuses discussions. Une conférence de consensus a été organisée par le Préfet coordonnateur du bassin Rhône Méditerranée à la demande du Ministère de l'écologie et du développement durable. Les résultats de cette conférence peuvent être consultés à l'adresse http://ccbr.lyon.cemagref.fr/index.php. Notre objectif ici n'est aucunement de revenir sur les résultats de cette conférence, ni même d'émettre un quelconque avis sur le débit atteint lors de cette crue. Nous allons nous contenter d'étudier la sensibilité de la crue centennale estimée à la valeur de débit utilisée pour la crue de décembre 2003.

La Figure 4-6 présente les résultats de cette étude de sensibilité pour la crue centennale estimée à partir de deux distributions (GEV et Gumbel). Il apparaît que la crue centennale estimée par une loi de Gumbel n'est que peu sensible à la valeur utilisée pour la crue de

décembre 2003. Par contre, la crue centennale estimée par une loi GEV est beaucoup plus sensible à cette valeur (elle est également plus élevée). Précisons que comme dans le cas de l'étude par scénarios, cette étude de sensibilité ne permet pas d'émettre un quelconque jugement sur la valeur la plus « vraisemblable » pour la crue de décembre 2003.

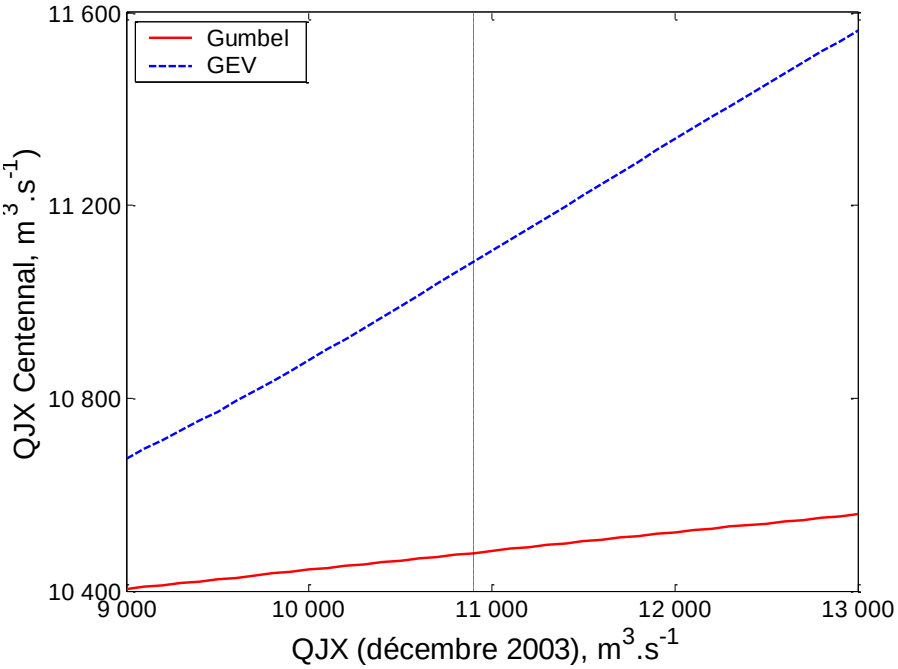


Figure 4-6. Sensibilité de la crue centennale estimée à la valeur de débit utilisée pour la crue de décembre 2003 sur le Rhône à Beaucaire. Le trait vertical en pointillé représente le débit donné dans la banque HYDRO.

4.2.3 Quantification probabiliste des incertitudes

La quantification probabiliste des incertitudes représente une approche plus complète que les approches par scénarios ou par étude de sensibilité. En effet, l'objectif d'une quantification probabiliste des incertitudes n'est pas seulement de visualiser la variation des estimations lorsque certains facteurs changent, mais également d'affecter une « vraisemblance » aux différentes estimations possibles. Dans le cas de l'estimation d'un débit décennal, par exemple, on ne se contentera pas d'évaluer la sensibilité de cette estimation, mais on cherchera à calculer une densité de probabilité des différentes estimations possibles du quantile décennal. La Figure 4-7 présente un exemple d'une telle représentation, où les quantiles décennal et centennal sont représentés sous la forme de distributions. On utilise également fréquemment des représentations basées sur des intervalles de confiance au niveau *p*%, qui sont des intervalles contenant *p*% de l'aire sous les densités de la Figure 4-7. Ceci permet de définir des courbes enveloppes au niveau de confiance *p*% autour de la courbe des quantiles (Figure 4-8).

Le principal attrait de la quantification probabiliste des incertitudes est qu'elle ouvre la possibilité de prendre en compte l'incertitude de manière explicite dans la prise de décision. En effet, il existe une théorie de la décision en contexte incertain (cf. par exemple Gilboa 2009), qui réclame généralement d'avoir été capable d'exprimer les incertitudes sous forme probabiliste. Ce type d'approche pour la prise de décision reste néanmoins peu utilisé en hydrologie.

La principale difficulté de la quantification probabiliste des incertitudes est qu'elle nécessite des outils statistiques sophistiqués. La possibilité de quantifier de façon probabiliste les incertitudes affectant l'estimation des débits de référence dépend en fait fortement de la source d'incertitude considérée :

- Il existe une théorie statistique bien développée pour la quantification probabiliste des **incertitudes d'échantillonnage**.
- La quantification des **incertitudes de mesure et de modélisation** est bien plus complexe à mettre en œuvre de manière probabiliste.

Dans la pratique hydrologique, les intervalles de confiance autour des quantiles estimés représentent en général la seule incertitude d'échantillonnage. Les incertitudes de mesure et de modélisation sont donc généralement ignorées. Des approches permettant d'intégrer également ces deux sources d'incertitudes existent, mais restent encore confinées au domaine de la recherche.

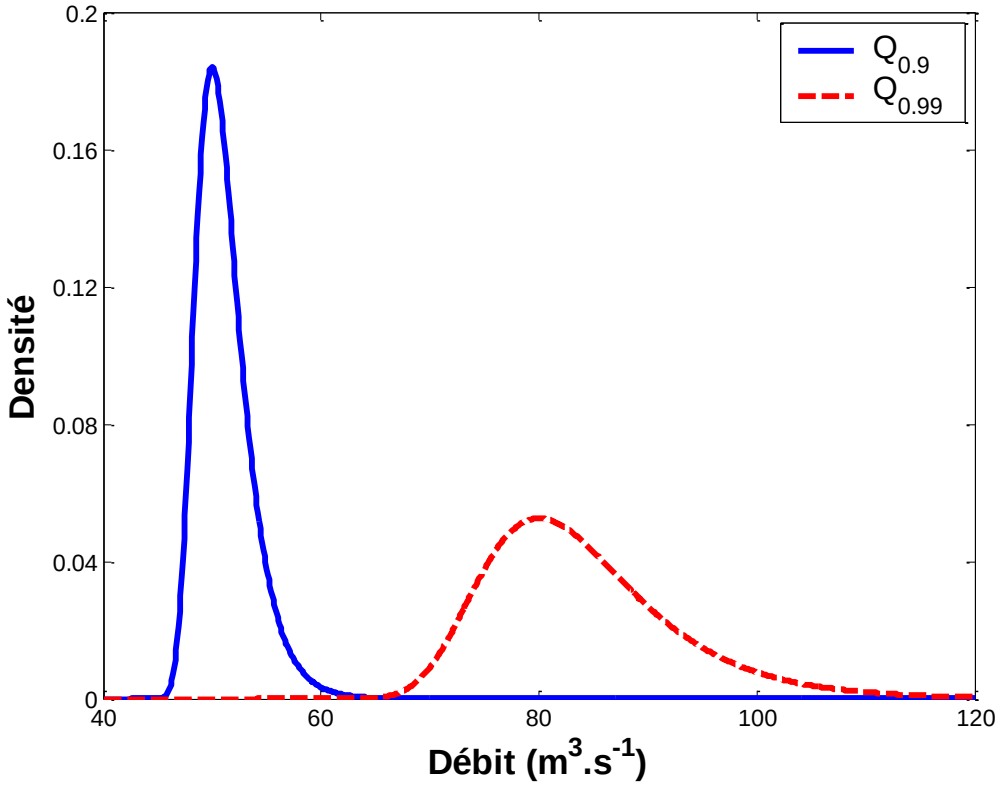


Figure 4-7. Représentation probabiliste des incertitudes d'échantillonnage affectant les débits décennal et centennal estimés.

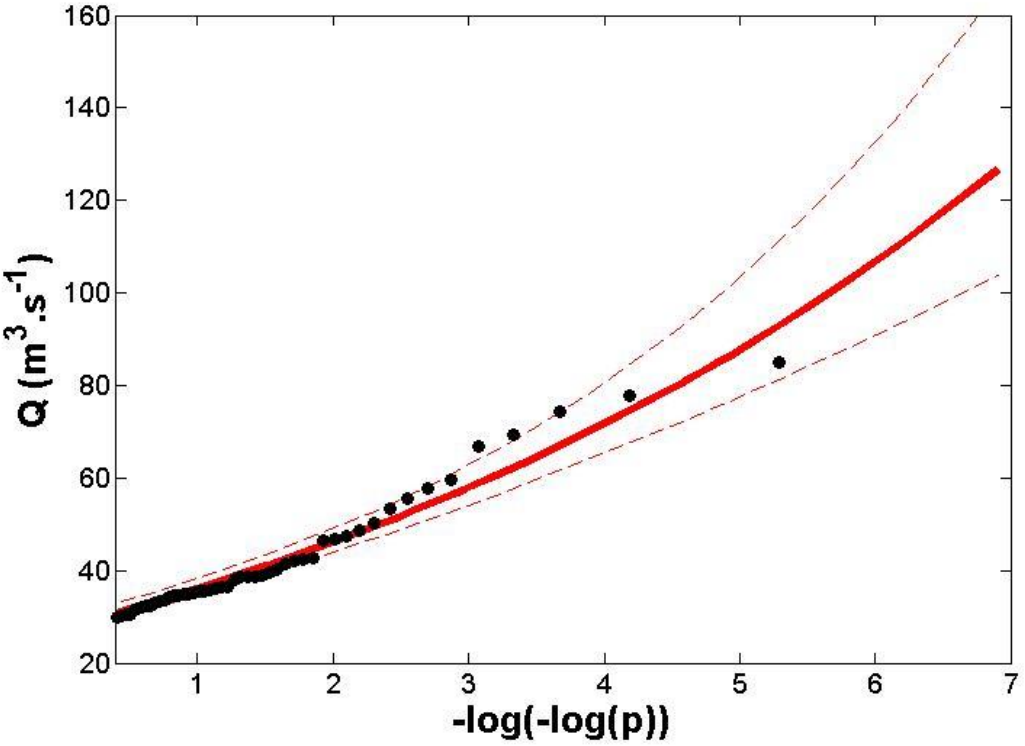


Figure 4-8. Représentation par des intervalles de confiance (ici à 90%) des incertitudes affectant les quantiles estimés.

4.3 Quelques techniques de quantification probabiliste de l'incertitude d'échantillonnage

Comme indiqué dans la section précédente, seule l'incertitude d'échantillonnage peut être quantifiée de manière aisée (tout est relatif…). Nous présentons donc dans cette section quelques approches pour calculer des intervalles de confiance qui reflètent l'incertitude d'échantillonnage (et seulement cette incertitude).

4.3.1 Intervalles de confiance : principe général

Les estimateurs des débits de référence sont soumis à la fluctuation d'échantillonnage dont nous avons longuement parlé précédemment (sections 2.1.3 et 4.1.2) : si l'on prend un autre échantillon, la valeur des estimations changera. Il est donc important de prendre en compte cette fluctuation afin de rendre plus interprétables les estimations ponctuelles.

De manière générale, supposons que $(X_1,...,X_n)$ soit un échantillon *iid* issu d'une certaine distribution $f(\theta)$. Notons $S = g(X_1,...,X_n)$ l'estimateur (d'un débit de référence, par exemple) pour lequel on souhaite calculer un intervalle de confiance. Supposons que l'on soit capable, connaissant la distribution $f(\theta)$ des observations, d'en déduire la distribution de l'estimateur *S* (cette distribution dépend donc aussi de θ). Un intervalle de probabilité au risque α (ou au niveau de confiance *1-*α), noté $I_\alpha(\theta)$, est un intervalle vérifiant la propriété :

$$P(S \in I_\alpha) = 1-\alpha$$

Dans la pratique, on choisira souvent un intervalle $[s^-,s^+]$ tel que :

$$P(S < s^-) = \alpha / 2$$

et

$$P(S > s^+) = \alpha / 2 \Leftrightarrow P(S \le s^-) = 1 - \alpha / 2$$

La principale difficulté dans le calcul d'intervalles de confiance réside dans la détermination de la distribution de l'estimateur *S*, que l'on appelle la **distribution d'échantillonnage** (car elle découle de la distribution de l'échantillon). Evidemment, cette distribution d'échantillonnage dépend à la fois du choix de la distribution des observations, mais également de la méthode d'estimation des paramètres. Des calculs explicites sont le plus souvent impossibles, et on a alors recours à des approximations asymptotiques, c'est à dire qui ne sont valables que pour des échantillons suffisamment grands. L'obtention pratique de ces approximations relève du travail du statisticien plutôt que de celui de l'hydrologue.

4.3.2 Quelques formules

Paramètres estimés par la méthode des moments

La formule suivante donne la variance du quantile q_p calculé à partir d'une loi à k paramètres estimés par la méthode des moments :

$$Var(\hat{q}_p) \approx \sum_{i=1}^{p} \left(\frac{\partial q_p}{\partial m_i} \right)^2 Var(m_i) + 2 \sum_{i=1}^{p} \sum_{j \neq i} \frac{\partial q_p}{\partial m_i} \frac{\partial q_p}{\partial m_j} Cov(m_i, m_j) \qquad (4\text{-}1)$$

Si on suppose que le quantile est asymptotiquement Gaussien et non biaisé, alors l'intervalle de confiance au niveau de confiance 1-α est de la forme $\hat{q}_p \pm u_{1-\alpha/2} \sqrt{Var(\hat{q}_p)}$, où $u_{1-\alpha/2}$ est le quantile de la loi normale *N(0,1)*.

L'application de la formule ci-dessus fournit par exemple les approximations suivantes :

- Loi Normale : $Var(\hat{q}_p) \approx \frac{s_x^2}{N} \left[1 + \frac{u_p^2}{2} \right]$
- Loi de Gumbel : $Var(\hat{q}_p) \approx \frac{s_x^2}{N} \left[1 + 1.14 w(p) + 1.10 w^2(p) \right]$ avec $w(p) = -\frac{\sqrt{6}}{\pi} \log(-\log(p)) + 0.45$ (4-2)
- Loi exponentielle : $Var(\hat{q}_p) \approx \frac{s_x^2}{N} \left[1 + 2z(p) + 2z^2(p) \right]$ avec $z(p) = \log(1-p)$

La Figure 4-9 présente une application de ces formules pour calculer l'intervalle de confiance à 90% du quantile estimé d'une loi exponentielle.

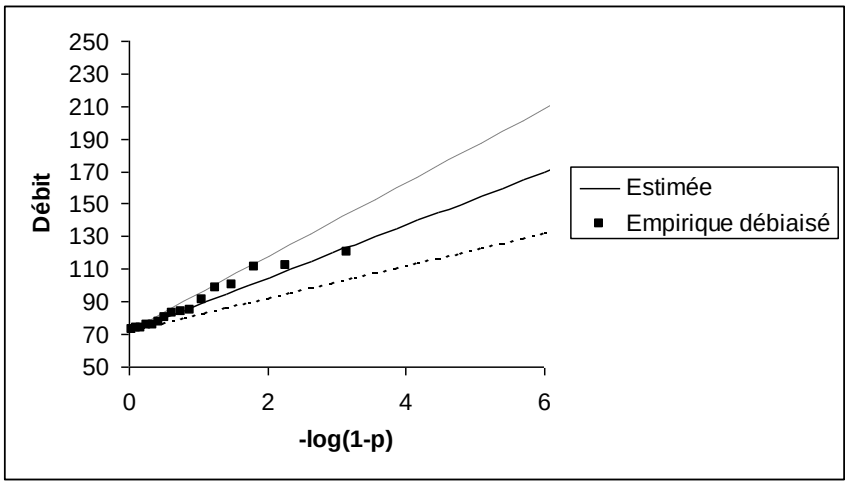


Figure 4-9. Intervalle de confiance à 90% pour les quantiles d'une loi exponentielle.

S'affranchir de l'hypothèse de normalité asymptotique des quantiles.

Cette hypothèse de normalité asymptotique des quantiles est à considérer avec précaution, surtout si l'effectif de l'échantillon n'est pas très important. Il existe des théorèmes, proche du théorème central limite, garantissant théoriquement cette normalité, mais certaines lois que nous utilisons peuvent invalider les hypothèses de ces théorèmes. En particulier, les lois GEV et de Pareto Généralisée peuvent avoir des moments infinis.

Plusieurs formules plus ou moins empiriques ont donc été établies pour calculer des intervalles de confiance plus réalistes. A titre d'illustration, on utilise parfois l'approximation suivante pour une loi de Gumbel :

Borne inférieure : $\hat{q}_p - h_1 s_x$

Borne supérieure : $\hat{q}_p + h_2 s_x$

Avec : $h_1 = \dfrac{A-B}{C}$ et $h_1 = \dfrac{A+B}{C}$

$$A = u_{1-(1-\alpha)/2} \frac{\sqrt{1+1.13t_p+1.1(t_p)^2}}{\sqrt{n}}$$

$$B = (u_{1-(1-\alpha)/2})^2 \frac{1.1t_p+0.57}{n} \qquad (4\text{-}3)$$

$$C = 1-\frac{1.1}{n}(u_{1-(1-\alpha)/2})^2$$

$$t_p = \frac{-\log(-\log(p))-0.577}{1.28}$$

$u_{1-(1-\alpha)/2}$ le quantile de la loi normale *N(0,1)*.

Notons que si *n* est grand, alors *B* devient négligeable devant *A*, et l'intervalle de confiance devient symétrique, signe de convergence vers la normalité.

La Figure 4-10 fournit un exemple d'intervalle de confiance à 90% calculé avec cette formule à partir d'un échantillon de 21 valeurs MAXAN. On voit clairement que cet intervalle n'est

pas symétrique, et que l'estimation de débits de périodes de retour 100 ou 1000 ans avec seulement 21 années de données est très imprécise, puisque les valeurs varient dans l'intervalle de confiance du simple au double.

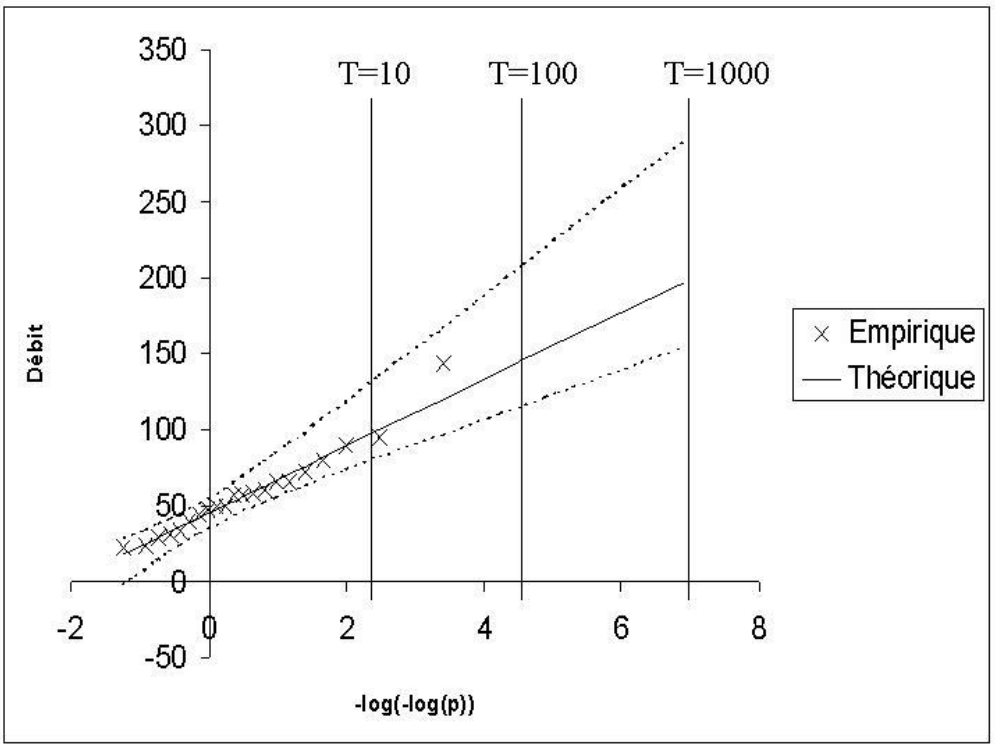


Figure 4-10. Intervalle de confiance à 90% pour les quantiles d'une loi de Gumbel

4.3.3 Méthodes de rééchantillonnage

Une alternative intéressante et simple à mettre en œuvre pour déterminer les intervalles de confiance est l'utilisation des méthodes de rééchantillonnage, le **Bootstrap** notamment, qui permettent de ne pas poser d'hypothèses *a priori* sur la distribution d'échantillonnage des quantiles. L'idée derrière ces méthodes de rééchantillonnage est la suivante : si l'on pouvait disposer de plusieurs échantillons, il nous serait possible d'estimer les caractéristiques de la distribution d'échantillonnage. Comme on ne dispose en pratique que d'un unique échantillon, de nouveaux échantillons synthétiques sont créés par tirage au sort avec remise dans les valeurs observées. Plus précisément, l'algorithme du Bootstrap, à partir d'un échantillon observé $(x_1,...,x_n)$, est le suivant :

- ✓ faire pour i=1,…, Nb
  - o tirer au sort avec remise n individus (certains individus apparaîtront plusieurs fois, d'autres aucune)
  - o calculer l'estimateur désiré $S_{obs}^{(i)}$ (par exemple, le quantile décennal) sur cet échantillon
- ✓ fin

Nous obtenons ainsi un nouvel échantillon $\left(S_{obs}^{(1)},...,S_{obs}^{(N_b)}\right)$ de **réplications Bootstrap**, dont on montre qu'il est issu d'une bonne approximation de la distribution d'échantillonnage. Un intervalle de confiance à 90% peut donc être directement déduit des réplications Bootstrap

$\left(S_{obs}^{(1)},...,S_{obs}^{(N_b)}\right)$ an calculant les quantiles empiriques de fréquence 0.05 et 0.95 issus de ces réplications.

La simplicité de cette méthode tient au fait qu'elle ne fait plus appel au cadre probabiliste pour approcher la distribution d'échantillonnage (même si la théorie probabiliste est nécessaire pour démontrer que la technique fonctionne….). Les méthodes de rééchantillonnage ont gagné en popularité ces dernières années du fait de leur simplicité et de la puissance accrue des moyens informatiques.

5 Autres techniques statistiques utiles

Cette dernière section est quelque peu à part dans l'organisation de ce cours : elle s'intéresse à deux techniques statistiques transversales, utilisées dans un grand nombre de situations, et qu'il convient donc de présenter. La section 5.1 décrit ainsi l'utilisation de **tests statistiques**, tandis que la section 5.2 présente la **régression linéaire** et quelques applications.

5.1 Tests

5.1.1 Principe général d'un test statistique

D'après Saporta (1990), « un test est un mécanisme qui permet de trancher entre deux hypothèses, dont une et une seule est vraie, au vu des résultats d'un échantillon ». Soient H_0 et H_1 ces deux hypothèses. Il existe alors 4 possibilités, de probabilités différentes, résumées dans le tableau suivant :

Vérité / Décision	H_0	H_1
H_0	$1-\alpha$	β
H_1	α	$1-\beta$

Tableau 5-1. Risques d'erreur.

H_0 est appelée **l'hypothèse nulle**, et H_1 l'**alternative**. Nous allons voir que ces hypothèses ne jouent pas un rôle symétrique, contrairement à ce que ce tableau pourrait laisser croire. α et β sont appelés les **risques de première et de seconde espèce**. $1-\beta$ est la **puissance** du test.

Ces définitions étant posées, comment effectue-t-on la prise de décision ?

La quasi-totalité des tests d'hypothèses suit le même schéma, qui peut se résumer comme suit :

1. Choix des hypothèses H_0 et H_1
2. Choix du risque de première espèce α
3. Choix de la statistique de test
4. Détermination de la loi de cette statistique sous H_0
5. Calcul de la zone de rejet
6. Calcul de la valeur expérimentale de la variable de test et décision
7. Si possible, calcul du risque de seconde espèce et de la puissance

5.1.2 Exemple détaillé

L'application d'un test est illustrée par un exemple détaillé pas à pas dans cette section. La Figure 5-1 présente les VCN7 (minima annuels de débits moyens sur 7 jours) calculés sur la

Seine à Paris Austerlitz (43800km²). On perçoit visuellement un changement dans la moyenne de ces observations, aux alentours de la date de 1974 qui correspond à la mise en service du lac-réservoir Marne. On cherche à savoir si la différence de moyenne est significative, ou bien si elle reste compatible avec la variabilité d'échantillonnage des données. Soient μ_1 et μ_2 les moyennes avant et après 1974. Les hypothèses à tester sont donc H0 : $\mu_1 = \mu_2$ contre H1 : $\mu_1 \neq \mu_2$. Le risque de première espèce est fixé à la valeur $\alpha = 10\%$.

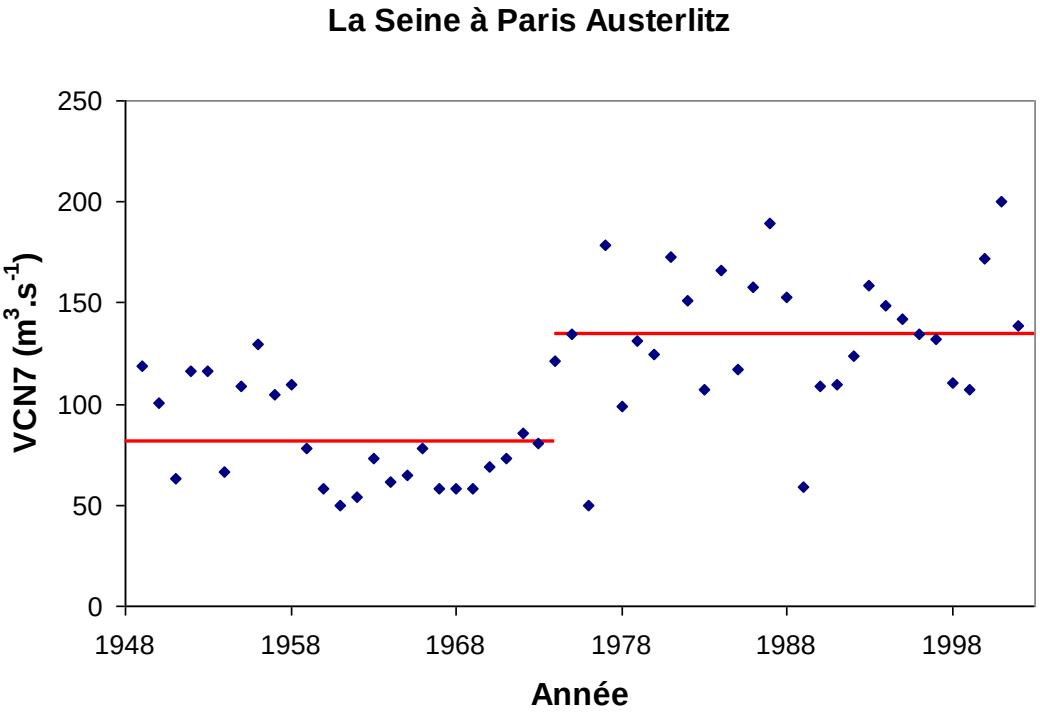


Figure 5-1. VCN7 de la Seine à Paris Austerlitz.

Une approche naturelle pour mettre en évidence la différence de moyenne avant et après l'installation du lac-réservoir est de calculer la différence des moyennes empiriques calculées avant et après 1974, $(\overline{x}_1 - \overline{x}_2)$. Si cette valeur s'éloigne de zéro, on aura tendance à conclure que la différence est significative.

Tout le problème réside donc dans la détermination de la valeur à partir de laquelle on peut conclure à une différence significative. Intuitivement, on peut deviner que la seule différence de moyenne ne suffit pas : il faut également prendre en compte la variabilité naturelle des données. Par exemple, une différence de moyenne de 50 $m^3.s^{-1}$ n'a pas la même signification si l'écart-type des données est de l'ordre de 5 $m^3.s^{-1}$ ou de 50 $m^3.s^{-1}$.

Cette prise en compte de la variabilité des données est intégrée dans une **statistique de test**, qui dépendra à la fois de la différence des moyennes et de l'écart-type des données. Plus précisément, la statistique de test suivante est utilisée :

$$T = \frac{(\overline{X}_1 - \overline{X}_2)\sqrt{n_1 + n_2 - 2}}{\sqrt{(n_1 S_1^2 + n_2 S_2^2)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \qquad (5\text{-}1)$$

La formulation de cette statistique de test dérive de considérations théoriques : il est en effet possible de démontrer que cette statistique suit une loi de Student à $n_1 + n_2 - 2$ degrés de liberté sous les hypothèses suivantes : (i) H0 est vrai (les moyennes sont identiques) ; (ii) Les données avant et après 1974 sont issues des réalisations indépendantes d'une loi normale ; (iii) Les variances avant et après 1974 sont identiques.

La Figure 5-2 montre la densité de probabilité d'une telle loi de Student. La suite du raisonnement est la suivante : supposons que l'hypothèse H0 soit vraie. Alors, la valeur observée de la statistique *T* devrait être compatible avec cette distribution de Student. Si ce n'est pas le cas, on sera amené à rejeter l'hypothèse H0.

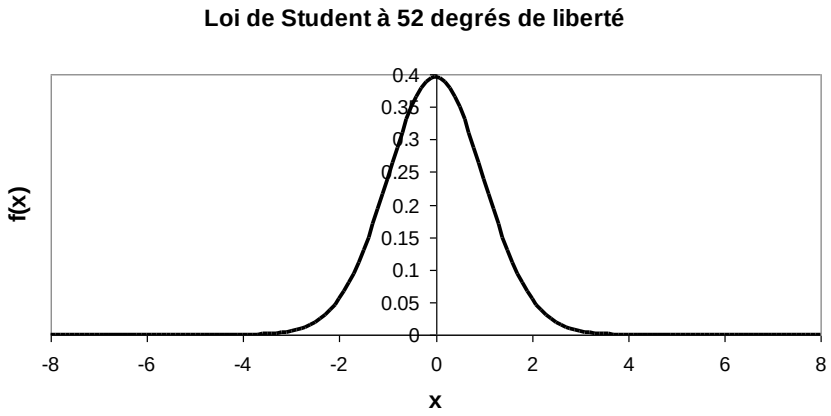


Figure 5-2. Densité d'une loi de Student à 52 degrés de liberté.

Reste à décider à partir de quelles valeurs on considérera qu'il y a incompatibilité entre la valeur observée et la distribution de Student. Pour calculer ces **valeurs critiques**, on fait intervenir le risque de première espèce α fixé au début de cet exemple à 10%. On va définir une **zone de rejet** correspondant aux valeurs rares de cette distribution de Student, de sorte que la probabilité d'appartenir à cette zone de rejet soit égale au risque de première espèce α=10%. On pourra pour cela définir les valeurs critiques comme les quantiles d'ordre 5% et 95% de la distribution de Student. L'intervalle entre ces quantiles définira la zone d'acceptation, le complémentaire la zone de rejet de l'hypothèse H0 (voir illustration en Figure 5-3).

La statistique de test (5-1) calculée sur les observations donne une valeur d'environ –6.4. Cette valeur est nettement dans la zone de rejet, elle n'est donc pas compatible avec la distribution de Student que devrait suivre la statistique de test si H0 était vraie. Avec un risque d'erreur α=10%, on peut donc rejeter H0 et affirmer que les moyennes avant et après 1974 sont significativement différentes.

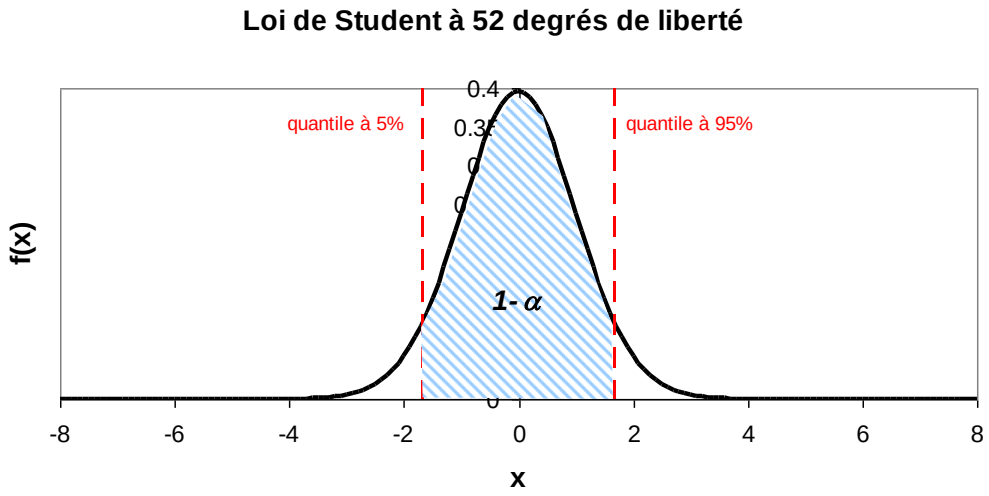


Figure 5-3. Illustration de la définition des zones d'acceptation et de rejet.

5.1.3 Formulaire

La quasi-totalité des tests statistiques suit le même schéma d'application. Nous nous contentons donc, dans cette section, de fournir les formules nécessaires à l'implémentation de divers tests utilisés en pratique.

Tests sur les paramètres d'une loi de Gauss

Condition d'application : l'échantillon *iid* est issu d'une loi normale $N(\mu,\sigma)$

H_0 : $\mu=m_0$ contre H_1 : hypothèse alternative ($\mu>m_0$, ou $\mu=m_1$, par exemple)

σ connu

Statistique de test : $\bar{X} \sim N(m_0, \frac{\sigma}{\sqrt{n}})$ sous H_0

H_0 : $\mu=m_0$ contre H_1 : hypothèse alternative ($\mu>m_0$, ou $\mu=m_1$, par exemple)

σ inconnu

Statistique de test : $T = \frac{\bar{X} - m_0}{S}\sqrt{n-1} \sim Student(n-1)$ sous H_0 ,

où $S = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2}$ est l'estimateur de l'écart type.

Note : ces deux tests restent applicables si les données ne sont pas Gaussiennes avec un échantillon d'effectif au moins 30.

H_0 : $\sigma=\sigma_0$ contre H_1 : hypothèse alternative

μ connu

Statistique de test : $\frac{\sum_{i=1}^{n}(X_i-\mu)^2}{\sigma_0^2} \sim \chi^2(n)$ sous H_0

H_0 : $\sigma=\sigma_0$ contre H_1 : hypothèse alternative

μ inconnu

Statistique de test : $\frac{nS^2}{\sigma_0^2} \sim \chi^2(n-1)$ sous H_0

Tests d'ajustement à une distribution

Les deux tests présentés ci-après ont pour but de vérifier que les données sont issues d'une distribution $F_0(x)$

Test de Kolmogorov

Conditions d'applications : $F_0(x)$ entièrement spécifiée (i.e. pas de paramètres) et continue.

H_0 : données issues de $F_0(x)$ contre H_1 : données issues d'une autre distribution

Statistique de test :

$D_n = \sup_x \left|F_n(x) - F_0(x)\right|$, où $F_n(x)$ est la fonction de répartition empirique

Sous H_0, la loi de D_n est tabulée

Test du χ^2

Conditions d'applications : $F_0(x)$ doit être discrète ou discrétisée. $p_1,\ldots,p_k$ les probabilités théoriques de chaque classe, et $N_1,\ldots,N_k$ les effectifs observés pour chaque classe.

H_0 : données issues de $F_0(x)$ contre H_1 : données issues d'une autre distribution

Statistique de test :

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

Loi sous H_0 :

si $F_0(x)$ est entièrement spécifiée, $D^2 \sim \chi^2(k-1)$ asymptotiquement

si $F_0(x)$ dépend de l paramètres, alors il faut estimer ces paramètres par maximum de vraisemblance à partir de la loi discrétisée, et $D^2 \sim \chi^2(k-1-l)$ asymptotiquement

Note : on admettra l'approximation asymptotique si $np_i>5$ pour toutes les classes (procéder éventuellement à des regroupements)

Tests de comparaison d'échantillons

Test des variances de Fischer

Conditions d'application : deux échantillons indépendants de tailles n_1 et n_2 et de lois normales $N(\mu_1, \sigma_1)$ et $N(\mu_2, \sigma_2)$.

$H_0 : \sigma_1 = \sigma_2$ contre $H_1 : \sigma_1 \neq \sigma_2$

Statistique de test : $F = \dfrac{n_1 S_1^2/(n_1-1)}{n_2 S_2^2/(n_2-1)} = \dfrac{S_1^{*2}}{S_2^{*2}}$, où S^{*2} est l'estimateur sans biais de la variance.

On mettra au numérateur l'échantillon conduisant à la variance estimée la plus grande.

Loi sous H_0 : $F \sim Fisher(n_1 - 1, n_2 - 1)$

Test des moyennes de Student

Conditions d'application : deux échantillons indépendants de tailles n_1 et n_2 et de lois normales $N(\mu_1, \sigma)$ et $N(\mu_2, \sigma)$ (écarts-types inconnus mais égaux).

$H_0 : \mu_1 = \mu_2$ contre H_1 : hypothèse alternative

Statistique de test : $T = \dfrac{(\bar{X}_1 - \bar{X}_2)\sqrt{n_1 + n_2 - 2}}{\sqrt{(n_1 S_1^2 + n_2 S_2^2)\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}} \sim Student(n_1 + n_2 - 2)$ sous H_0

Note : Si les deux échantillons sont suffisamment grands (quelques dizaines d'individus), le test de Student peut être appliqué même si les données ne sont pas gaussiennes ou ont des variances inégales. On dit que ce test est « **robuste** » (i.e. peu sensible au non-respect des conditions d'application).

Test des rangs de Wilcoxon

Conditions d'application : deux échantillons $(x_1, ..., x_{n_1})$ et $(y_1, ..., y_{n_2})$ indépendants de tailles $n_1 < n_2$. On mélange les échantillons et on note $rg(x_i)$ le rang d'une observation x_i dans cet échantillon mélangé

H_0 : Les échantillons sont issus d'une même population contre H_1 : hypothèse alternative

Statistique de test : $S = \dfrac{\sum_{i=1}^{n} rg(x_i) - \mu}{\sigma}$, où $\mu = \dfrac{n_1(n_1 + n_2 + 1)}{2}$ et $\sigma = \sqrt{\dfrac{n_1 n_2 (n_1 + n_2 + 1)}{12}}$

Loi sous H_0 : $S \sim N(0,1)$ dès que les deux échantillons sont d'effectifs supérieurs à 8

Test de corrélation

Conditions d'application : deux variables X et Y Gaussiennes dont on cherche à savoir si elles sont corrélées. Soit $\rho = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}$ le coefficient de corrélation.

H_0 *:* $\rho=0$ contre H_1 : Hypothèse alternative

Statistique de test : $R = \frac{\rho\sqrt{n-2}}{\sqrt{1-\rho^2}} \sim Student(n-2)$ sous H_0

Note : Ce test permet également de tester la nullité du paramètre *a* d'une régression $Y=aX+b$, puisque $\hat{a} = \frac{Cov(X,Y)}{Var(X)} = \rho\frac{\sigma_Y}{\sigma_X}$

Test de stationnarité

Test de Pettitt

Conditions d'application : échantillon $(x_1,...,x_n)$ dont on cherche à savoir s'il présente une rupture

H_0 *:* Echantillon stationnaire contre H_1 : Hypothèse alternative

Statistique de test : $S = \max_k(|U(k)|)$, où $U(k) = \sum_{i=1}^{k}\sum_{j=k+1}^{n} signe(x_i - x_j)$

Loi sous H_0 : $P(S \geq s_0) = 2\exp\left(\frac{-6s_0^2}{n^3+n^2}\right)$

Test de Spearman

Conditions d'application : échantillon $(x_1,...,x_n)$ dont on cherche à savoir s'il présente une tendance

H_0 *:* Echantillon stationnaire contre H_1 : Hypothèse alternative

Statistique de test : $Z = \sqrt{n-1}\left(1 - \frac{6\sum_{i=1}^{n}(rang(x_i)-i)^2}{n(n^2-1)}\right)$

Loi sous H_0 : $Z \sim N(0;1)$ asymptotiquement

Test de Mann-Kendall

Conditions d'application : échantillon $(x_1,...,x_n)$ dont on cherche à savoir s'il présente une tendance

H_0 *:* Echantillon stationnaire contre H_1 : Hypothèse alternative

Statistique de test : soit $S = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} signe(x_j - x_i)$. La statistique de test est :

$$Z = \begin{cases} \dfrac{S-1}{\sqrt{Var(S)}} \text{ si } S > 0 \\ 0 \text{ si } S = 0 \\ \dfrac{S+1}{\sqrt{Var(S)}} \text{ si } S < 0 \end{cases} \text{, avec } Var(S) = n(n-1)(2n+5)/18 .$$

Loi sous H_0 : $Z \sim N(0;1)$ asymptotiquement

Test de Buishand

Conditions d'application : échantillon Gaussien $(x_1,...,x_n)$ dont on cherche à savoir s'il présente une rupture

H_0 *:* Echantillon stationnaire contre H_1 : Hypothèse alternative

Statistique de test : $Z = \max_{k=1,..,n} \left(\dfrac{|U(k)|}{\sqrt{n}\sigma_X} \right)$, avec $U(k) = \sum_{i=1}^{k} (x_i - \overline{x})$

Loi sous H_0 : tabulée (cf. tableau ci-dessous)

n	α		
	0.1	0.05	0.01
10	1.05	1.14	1.29
20	1.1	1.22	1.42
30	1.12	1.24	1.46
40	1.13	1.26	1.5
50	1.14	1.27	1.52
100	1.17	1.29	1.55
∞	1.22	1.36	1.63

Tableau 5-2. Valeurs critiques de la statistique de Buishand

5.2 Etude conjointe de deux variables

5.2.1 La régression linéaire

Plaçons-nous dans la situation suivante : sur un échantillon, nous avons mesuré deux variables quantitatives, que nous noterons *X* et *Y*. Nous obtenons donc deux séries de valeurs $\boldsymbol{x}=(x_1,...,x_n)$ et $\boldsymbol{y}=(y_1,...,y_n)$. Les deux séries peuvent bien sûr être décrites indépendamment l'une de l'autre, mais nous allons ici nous intéresser à la liaison qui peut exister entre ces variables.

Exemple : Sur un échantillon de 10 petits cours d'eau, on a calculé la superficie du bassin versant (*X*) ainsi que la crue décennale (*Y*) :

BV	Superficie	Q10
1	16.2	5.77
2	21.4	5.37
3	38.7	7.29
4	42.3	12.66
5	55.8	11.09
6	62.1	25.31
7	64	20.18
8	68	39.27
9	73	39.59
10	94	28.30

Tableau 5-3. Tableau de données.

La manière la plus simple de représenter ces données est de projeter les couples (x_i,y_i) dans le plan : on obtient ainsi un **nuage de points**.

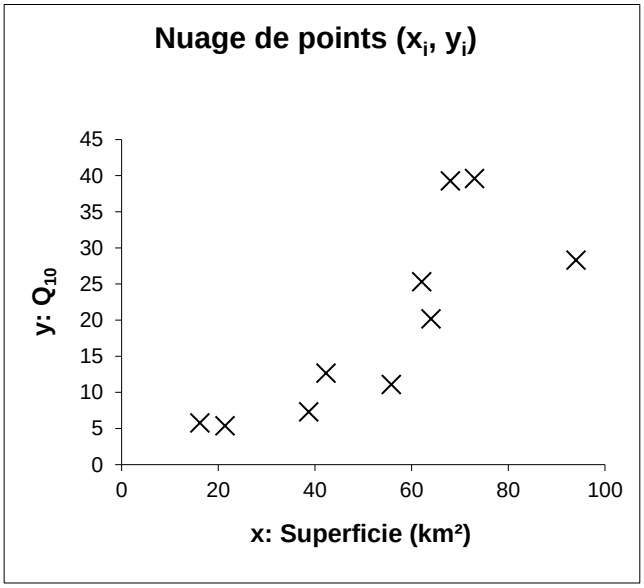


Figure 5-4. Nuage de points.

A la vue de ce graphique, il apparaît que les deux variables ont tendance à évoluer conjointement : les « grands » bassins versants ont des crues décennales plus importantes. C'est cette co-évolution que nous allons tenter d'étudier.

Commençons par définir la **covariance** entre ***x*** et ***y*** :

$$Cov(\boldsymbol{x},\boldsymbol{y}) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) \qquad (5\text{-}2)$$

Cette quantité mesure la manière avec laquelle *X* et *Y* varient « ensemble ». Comme pour la variance, on trouve parfois une définition alternative avec *1/n*. Le **coefficient de corrélation linéaire** est directement dérivé de la covariance :

$$r = \frac{Cov(\boldsymbol{x}, \boldsymbol{y})}{s_x s_y} \tag{5-3}$$

Il s'agit d'un coefficient adimensionnel, compris entre –1 et 1 : une valeur absolue proche de 1 sera la signature de deux variables liées linéairement, une valeur proche de zéro signifiera l'absence de relation linéaire. Il est important de noter que ceci n'interdit pas que les variables soient liées par un autre type de relation (polynomiale, sinusoïdale, …). Pour l'exemple représenté en Figure 5-4, les valeurs suivantes sont obtenues:

$$\begin{aligned} s_x &= 24 \\ s_y &= 13.17 \\ Cov(\boldsymbol{x}, \boldsymbol{y}) &= 255.01 \\ r &= 0.81 \end{aligned} \tag{5-4}$$

Ces chiffres confirment la liaison entre les variables. Nous pouvons essayer d'aller plus loin, en cherchant la droite la plus pertinente qui s'ajusterait au nuage de point, c'est en dire en évaluant une relation du type $Y=aX+b$. Evidemment cette relation ne peut pas être parfaite (tous les points ne sont pas alignés), nous introduisons donc des termes d'erreurs, ce qui nous donne la relation :

$$y_i = ax_i + b + e_i \quad \forall i = 1,...,n \tag{5-5}$$

Une « bonne » droite permettrait de minimiser ces erreurs. Nous allons donc définir un critère, dit des moindres carrés, construit à partir de la somme des carrés des erreurs :

$$S = \sum_{i=1}^{n} e_i^2 \tag{5-6}$$

Il s'agit donc à présent de trouver les valeurs de a et b minimisant S. Il est aisé de démontrer que ces valeurs sont données par :

$$\begin{aligned} \hat{a} &= \frac{Cov(\boldsymbol{x}, \boldsymbol{y})}{Var(\boldsymbol{x})} = r\frac{s_y}{s_x} \\ \hat{b} &= \bar{y} - \hat{a}\bar{x} \end{aligned} \tag{5-7}$$

La qualité de cet ajustement est mesurée par r^2 (qui varie dans [0,1]), qui mesure la part de variance expliquée par le modèle linéaire. Voici l'ajustement pour les données précédentes est donné en Figure 5-5.

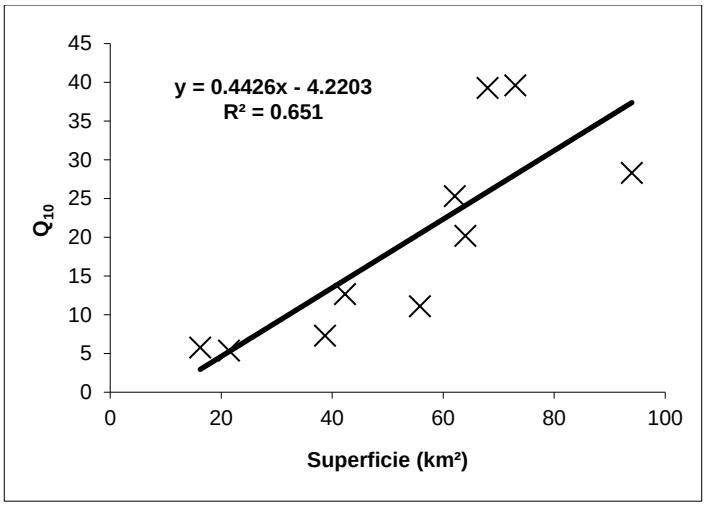


Figure 5-5. Ajustement de la droite de regression.

La démarche présentée dans ce cadre de recherche de relation linéaire entre variables peut être généralisée à toute forme de dépendance du type *Y=f(X)* :

- ✓ Choix d'une fonction *f*
- ✓ Calcul des paramètres optimaux de cette fonction, au sens d'un certain critère (moindres carrés, par exemple)
- ✓ Evaluation de la qualité de l'ajustement, par un indicateur du type $1-\frac{Var(\boldsymbol{e})}{Var(\boldsymbol{y})}$ (qui n'est plus égal à r^2 si *f* n'est pas affine)

Evidemment, la plupart des calculs se compliquent par rapport au cas linéaire, le recours à des méthodes d'optimisation numérique est souvent indispensable.

Une extension importante de la méthode consiste à intégrer plusieurs variables pour en « expliquer » une autre : dans le cas précédent, il pourrait ainsi être bénéfique d'intégrer la pluviométrie ou la nature et l'occupation du sol pour améliorer l'explication du débit décennal. On parle de **régression multiple**. Ce type d'approche est très important en hydrologie pour les problèmes de **régionalisation**, dont la description détaillée dépasse le cadre de cette formation.

5.2.2 Application au contrôle des données : Double Cumul et Ellipse de Bois

Dans cette section, nous présentons brièvement deux applications de la régression linéaire pour le contrôle des données hydrométriques.

La première application est la méthode du double cumul bien connue des hydrologues. Notons *X* et *Y* les débits issus de deux stations hydrométriques. On pourra notamment supposer que l'une des stations est la station dite de référence (supposée exempte de problèmes métrologiques), tandis que l'autre station est la station à contrôler. La méthode des doubles cumuls consiste simplement à comparer le cumul des valeurs observées aux deux stations. Une rupture de pente dans la relation liant ces deux cumuls indiquera un problème potentiel sur la station à contrôler.

La Figure 5-6 présente une application de cette approche basée sur les débits mensuels issus de deux stations hydrométriques de la Zorn, Saverne (185 km²) et Waltenheim (688 km²). On semble déceler une légère rupture de pente dans le premier tiers de la courbe. Il est cependant

difficile de conclure sur la base de ce seul graphe – en l'absence de la droite de régression, cette rupture de pente aurait très bien pu passer inaperçue…

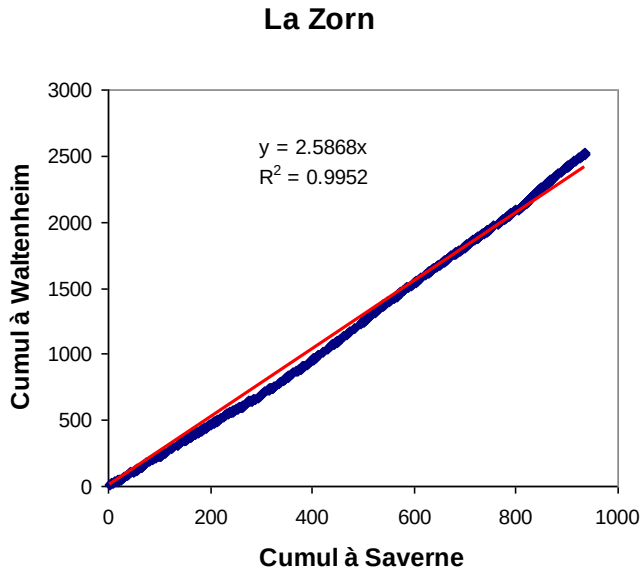


Figure 5-6. Double cumul des débits mensuels de la Zorn à Saverne et à Waltenheim, avec droite de régression.

Une extension intéressante de la méthode du double cumul a été proposée par Bois (1976). L'idée consiste à étudier le cumul des résidus de la régression linéaire entre les deux stations, plutôt que directement le cumul des données brutes.

La Figure 5-7 montre le résultat d'une régression linéaire entre les débits mensuels des deux stations de la Zorn, tandis que la Figure 5-8 décrit les résidus de cette régression. On semble percevoir à l'œil un changement dans la moyenne de ces résidus aux alentours de 1978, qui pourrait correspondre à la légère rupture de pente observée dans le graphe des doubles cumuls.

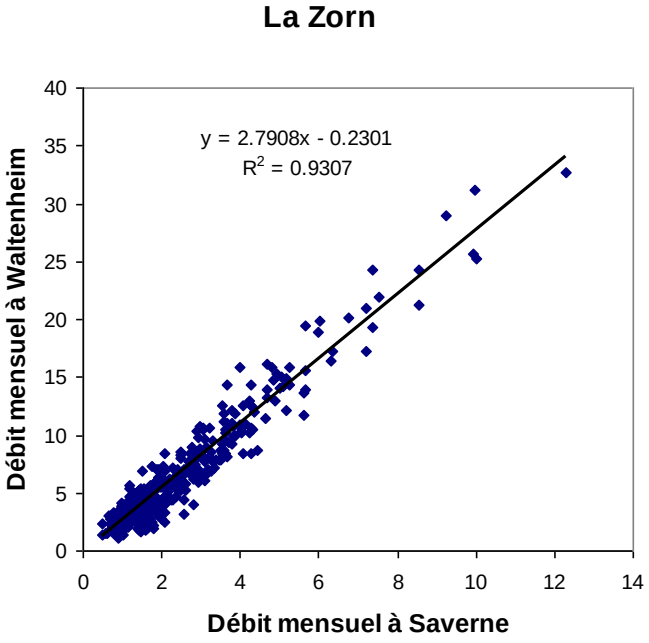


Figure 5-7. Régression linéaire entre les débits mensuels de la Zorn à Saverne et à Waltenheim.

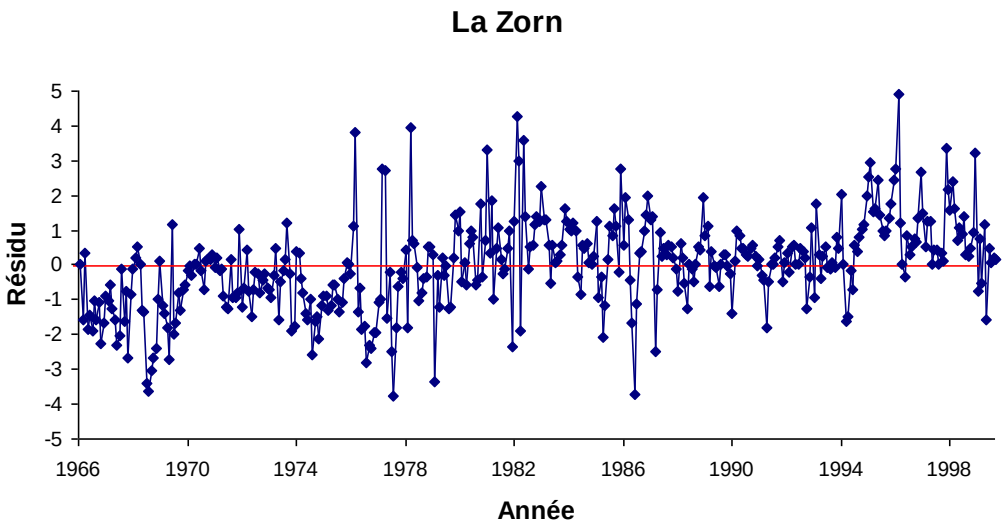


Figure 5-8. Résidus de la régression linéaire entre les débits mensuels de la Zorn à Saverne et à Waltenheim.

L'ellipse de Bois permet de mettre en évidence ce type de changement. Notons ε_t le résidu au pas de temps t, et définissons les résidus cumulés de la façon suivante :

$$E_0 = 0; E_t = \sum_{i=1}^{t} \varepsilon_i \qquad (5\text{-}8)$$

Bois (1976) a pu montrer que sous hypothèse de normalité des résidus, le k^{eme} résidu cumulé E_k suivait une loi normale de moyenne nulle et d'écart-type :

$$\sigma(E_k) = \sigma(\varepsilon)\sqrt{\frac{k(N-k)}{N-1}} \qquad (5\text{-}9)$$

On peut ainsi tracer les intervalles de confiance pour les résidu cumulés en chaque pas de temps, qui forment ainsi une **ellipse de confiance** à l'intérieur de laquelle la courbe des résidus cumulés devrait demeurer.

La Figure 5-9 montre la réalisation de cette ellipse au niveau de confiance 90% pour les données de la Zorn. Alors qu'une rupture de pente était difficilement décelable sur le graphe des doubles cumuls, il apparaît très clairement que la courbe du cumul des résidus sort très largement de l'ellipse de confiance, indiquant un possible problème dans les données. La pointe de la courbe des cumuls se situe aux alentours des années 1978-1982, ce qui indique une période où rechercher d'éventuelles explications à ce changement. Pour information, on retrouve dans les archives de la station de Waltenheim la pose d'un seuil en 1978 et le passage à des données limnigraphiques en 1983.

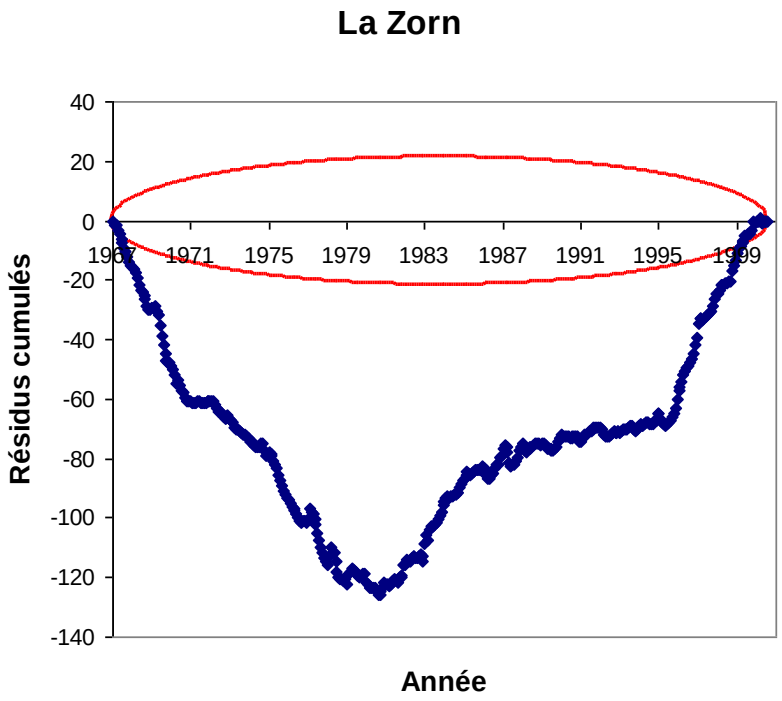


Figure 5-9. Ellipse de Bois pour les débits mensuels de la Zorn à Saverne et à Waltenheim.

6 Références bibliographiques

Bois, P. (1976). Contrôle de séries chronologiques corrélées par étude du cumul des résidus de la corrélation. 2e Journées Hydrologiques de l'ORSTOM.

Cox, D. R. (1966). Renewal Theory. Paris, Dunod.

Gelman, A., J. B. Carlin, et al. (1995). Bayesian data analysis, Chapman & Hall.

Gilboa, I. (2009). Theory of Decision under Uncertainty, Cambridge University Press.

Hosking, J. R. M. and J. R. Wallis (1997). Regional Frequency Analysis: an approach based on L-Moments. Cambridge, UK, Cambridge University Press.

Lang, M. (1995). Les chroniques en hydrologie: Modélisation comparée par un système de gestion de bases de données relationnel et orienté-objet. Traitements de base et intervalles de confiance des quantiles de crues. Techniques d'échantillonnage par la méthode du renouvellement. Grenoble, France, University Joseph Fourier**:** 296.

Paquet, E., J. Gailhard, et al. (2006). "Evolution de la méthode du gradex : approche par type de temps et modélisation hydrologique." La houille blanche **5**: 80-90.

Renard, B. (2008). "Détection et prise en compte d'éventuels impacts du changement climatique sur les extrêmes hydrologiques en France." La Houille Blanche(1): 109-117.

Saporta, G. (1990). Probabilités, analyse des données et statistiques, Technip.

Stewart, I. T., D. R. Cayan, et al. (2005). "Changes toward earlier streamflow timing across western North America." Journal of Climate **18**(8): 1136-1155.

Tallaksen, L. M. and H. A. J. Van Lanen (2004). Hydrological Drought: processes and estimation methods for streamflow and groundwater, Elsevier.