Chapitre 1. Notions de probabilités et de statistiques

En hydrologie, nous ne connaissons jamais la population totale mais nous disposons d'un échantillon non exhaustif tiré de cette population. C'est à partir de cet échantillon que nous choisirons la loi de probabilité adéquate et d'en calculer le mieux que possible ses paramètres statistiques **Laborde, (2003)**.

1.1. RAPPELS STATISTIQUES

I.1.1. PARAMETRES DE L'ECHANTILLON

I.1.1.1. Paramètre de tendance centrale

La moyenne arithmétique :

$$\overline{X} = \frac{1}{N}.\sum_{i=1}^{N} X_i$$

Avec N le nombre d'observation composant l'échantillon

I.1.1.2. Paramètre de dispersion

a) L'intervalle total de l'échantillon exprime la différence entre les valeurs extrêmes.

$$R = X_{max} - X_{min}$$

b) Déviation moyenne : c'est la moyenne arithmétique des valeurs absolues de la différence de chaque Xi de la moyenne $\overline{X}$.

$$D = \frac{1}{N}\sum_{i=1}^{N}\left|X_{I.} - \overline{X}\right|$$

c) La variance :

$$S^2 = \frac{1}{N}\sum_{i=1}^{N}\left(X_i - \overline{X}\right)^2$$

- L'écart type :

$$S = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(X_i - \overline{X}\right)^2}$$

d) Le coefficient de variation :

$$Cv = \frac{S}{\overline{X}}$$

I.1.1.3. Paramètre de forme

Ces paramètre caractérisent l'aplatissement et l'asymétrie,

Mesure de dissymétrie : (selon Pearson)

$$Cs = \frac{\left[\frac{1}{N}\cdot\sum_{i=1}^{N}\left(X_i - m^{1/3}\right)\right]}{\left[\frac{1}{N}\cdot\sum_{i=1}^{N}\left(X_i - \overline{X}\right)^2\right]^{3/2}}$$

Si Cs est positif la distribution est étalée sur la droite, on observe la succession mode – médiane –moyenne la dissymétrie est dite positive.

Si Cs est négatif la distribution est étalée sur la gauche, on observe la succession moyenne – médiane –mode la dissymétrie est dite négative.

Si Cs est nul, la distribution n'est forcement symétrique.

I.1.2. TEST RELATIF A L'ECHANTILLONNAGE

Des tests statistiques sont généralement indispensables pour vérifier l'homogénéité, l'indépendance, la stationnarité et, parfois, la détermination des points singuliers (**Bobée**,1990).

Test de stationnarité et d' homogénéité de Mann et Whitney (1947).

Soit une série complète divisée en deux échantillons de taille n_1 et n_2, la date des séparations étant choisie à priori. Les deux échantillons sont classés par ordre croissant.

Nous considérons la statistique suivante :

$$V = R - n_1(n_2 + 1)/2$$

$$W = n_1 n_2 - V$$

Où :

R : la somme des rangs des éléments de premier échantillon (taille n_1) et V représentent le nombre de fois des individus du premier échantillon (n_1) qui existe dans le deuxième échantillons (taille n_2).

$n = n_1 + n_2$ avec $n_1 < n_2$ puis on classe l'échantillon global de taille n par ordre croissant.

La statistique U de Mann- Whitney , est définie par le minimum de (V, U), U est supposé suivre une loi normale de moyenne et de variance .

$$\overline{U} = n_1.n_2/2 \qquad , Var(U) = \left[\frac{n_1.n_2}{n(n-1)}\right].\left[\frac{n^3 - n_2}{12} - \sum T\right]$$

Avec : $T = (J^3 - J)/12$, où J est le nombre d'observation lié à un rang donné.

$\sum T$: représente toutes les observations qui sont liées à des rangs pour les deux échantillons, l'hypothèse d'homogénéité et de stationnarité pour un seuil de signification $\propto$ % stipule que l'indicateur :

$$\left|\frac{U - \overline{U}}{\sqrt{V(U)}}\right| \leq U_{\alpha/2}$$

<u>Test de singularité et Beck (1947)</u>

Nous considérons les deux valeurs extrêmes suivantes :

$$X_{min} = e^{(\overline{X}_n - K_n.S_n)} \quad \text{et} \quad X_{max} = e^{(\overline{X}_n + K_n.S_n)}$$

K_n : paramètre statistique de prubbs et Beck, tabulé en fonction de la taille n de l'échantillon et du seuil $\propto$ % . Pilon et als (1985), propose une expression approchée de K_n pour un seuil de $\propto = 10\ \%$.

$K_n = -3{,}62201 + 6{,}28446.n^{1/4} - 2{,}49835.n^{1/2} + 0{,}491436.n^{3/4} - 0{,}037911.n.$

Une valeur quelconque de X_i est considérée si :

$$\begin{cases} X_i < X_{min.} \\ \text{ou} \\ Xi > X_{max} \end{cases}$$

Test de stationnarité et d'indépendance de Wald et Wolfowizitz (1943)

Soit un échantillon de taille N, (x i, i = 1 , 2 ,……, n) le test de Wald-Wilfowitz
Considère :

$$R = \sum_{i=1}^{N-1} X_i . X_{i+1} X_1 . X_n$$

Supposons que R suit une loi normale de moyenne $\overline{R}$ et de variance Var (R) dont leurs expressions sont respectivement les suivantes :

$$\overline{R} = \frac{S_1^2 - S_2}{n-1} \quad ; \quad Var(R) = \frac{S_2^2 - S_4}{n-1} + \frac{S_1^4 - 4S_1^2 S_2 + 4S\, S_3 + S_2^2 - 2S_4}{(n-1)(n-2)} - \overline{R}^2$$

Nous calculons la statistique $U = \frac{R - \overline{R}}{\sqrt{var(R)}}$ qui suit une loi normale n(0,1).L'hypothèse de stationnarité et d'indépendance est vérifiée pour un seuil signification si l'indicateur :

$$\left| \frac{R - \overline{R}}{\sqrt{var(R)}} \right| \leq U_{\alpha/2}$$

Test d'Anderson

On pose :

$$R' = \left[\frac{\left(\sum_{i-1}^{n-1} X_i - X_{i+1} + X_n X_1 \right) - \left(\sum_{i}^{n} X_i \right)^2 / n}{\sum_{i}^{n} X_i^{\,2} - (\sum_{i}^{n} X_i)^2 / n} \right]$$

On peut écrire : $R' = \frac{R - \frac{S_1^2}{n}}{S_2 - \frac{S_1^2}{n}}$

De moyenne : $\overline{R}' = -\frac{1}{(n-1)}$

Et de variance : $\text{var}(R') = \frac{R' - \overline{R}'}{\sqrt{\text{var}(R')}}$

- La valeur statistique $U = \frac{R' - \overline{R}'}{\sqrt{\text{var}(R')}}$ suit une loi normale centrée réduite.

Si $U < U_{\alpha 2,}$, on dit que l'hypothèse Ho d'indépendance est vraie.

Test du cumul des résidus

C'est une méthode complexe, mais beaucoup plus efficace.

Soient $\overline{X}, \overline{Y}, Sx^2, Sy^2, r$ respectivement les moyennes, la variance, le coefficient de corrélation empirique des séries à vérifier Xt, X et Yt .

Le résidu ε_i est l'écart entre la valeur Xi et la valeur correspondante à Yi dans la régression linéaire de X et Y :

$$\varepsilon_i = X_i - \overline{X} - r\frac{S_x}{S_y}\left(Y_i - \overline{Y}\right)$$

Nous considérons la courbe chronologique cumulée des résidus successifs :

$$Z_i = \sum_{i=1}^{t} \varepsilon_i$$

Soient les cumuls partiels :

$$T_{t,t+1} = \sum_{i=1}^{t+m} \varepsilon_i$$

Soit l'ellipse de contrôle :

$$Z(m) = +U_{\alpha/2n}\sqrt{m(n-m)(n-1)\frac{S_{y.\sqrt{1-r^2}}}{n}}$$

où : $U_{\alpha}/_{2n}$ est la valeur de variable.

m la distance verticale entre deux observations dans la série chronologique normale centrée réduite dont la probabilité de dépassement est $\frac{X}{2n}(\alpha = 10\% ou 5\%)$

En déplaçant cette ellipse, dont le grand axe est parallèle à l'axe des abscisses, de façon que l'un de ses sommet coïncide avec les points successifs de la courbe du cumule des résidus, on détecte les écarts Zt , t + m qui sortent des limites de l'ellipse.

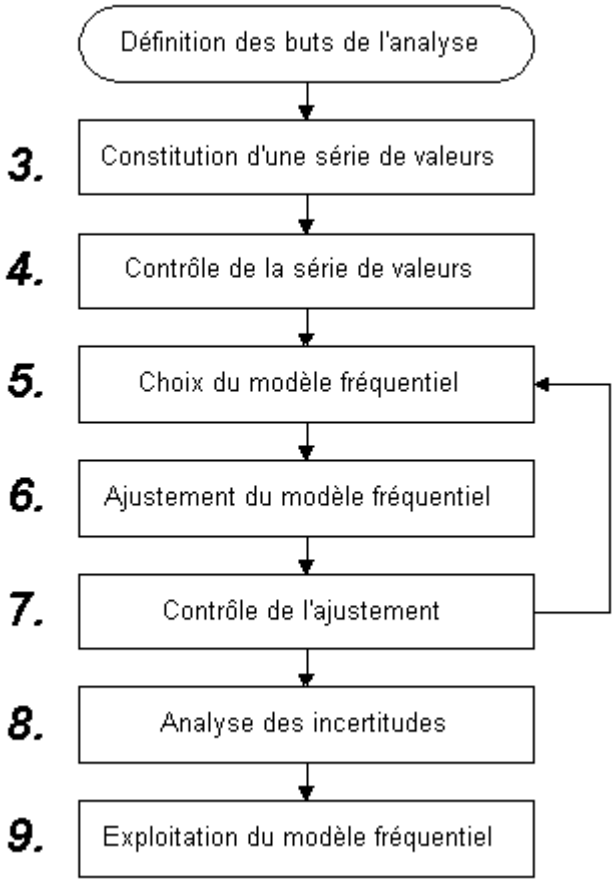
I.2. ANALYSE FREQUENTIELLE

I.2.1 Définition de l'analyse fréquentielle

L'analyse fréquentielle est une méthode statistique de prédiction consistant à étudier les événements passés, caractéristiques d'un processus donné (hydrologique ou autre), afin d'en définir les probabilités d'apparition future.

Cette prédiction repose sur la définition et la mise en œuvre d'un modèle fréquentiel, qui est une équation décrivant le comportement statistique d'un processus. Ces modèles décrivent la probabilité d'apparition d'un événement de valeur donnée.

L'analyse fréquentielle fait appel à diverses techniques statistiques et constitue une filière complexe qu'il convient de traiter avec beaucoup de rigueur. Ses diverses étapes peuvent être schématisées très simplement selon le diagramme suivant :



Principales étapes de l'analyse fréquentielle.

Ce sont essentiellement les étapes 5,6,7 et 8 qui sont développées dans ce document. Pour d'autres d'informations, prière de se référer au polycopié " Hydrologie fréquentielle, P. Meylan et A. Musy, EPFL, 1999 " duquel plusieurs graphiques de ce document ont été tirés.

Avant de commencer tout travail, il est primordial de formuler clairement les buts de l'analyse et d'adapter la démarche en conséquence. A cet égard, en hydrologie, l'un des critères essentiels est certainement l'échelle spatio-temporelle : étudier le comportement des crues dans un microbassin urbain (à très faible temps de concentration) avec des données de pluie au pas de temps mensuel n'aurait pas de sens ! L'inverse est tout aussi vrai : il est probablement inutile de disposer de pluies au pas de temps de la minute pour l'étude du bassin versant de l'Amazone !

La constitution d'échantillons, au sens statistique du terme, est un processus long, parsemé d'embûches, et au cours duquel de nombreuses erreurs, de nature fort différente, sont susceptibles d'être commises. Par ailleurs, il est indispensable, avant d'utiliser des séries de données, de se préoccuper de leur qualité et de leur représentativité. Le contrôle des données fera l'objet d'un chapitre qui sera traité ultérieurement dans ce cours.

I.2.2 Choix du modèle fréquentiel

La validité des résultats d'une analyse fréquentielle dépend du choix du modèle fréquentiel et plus particulièrement de son type. Diverses pistes peuvent contribuer à faciliter ce choix, mais il n'existe malheureusement pas de méthode universelle et infaillible.

I.2.2.1 Considérations théoriques

1. Loi normale

La loi normale se justifie, théoriquement par le théorème central-limite, comme la loi d'une variable aléatoire formée de la somme d'un grand nombre de variables aléatoires. En hydrologie fréquentielle des valeurs extrêmes, les distributions ne sont cependant pas symétriques, ce qui constitue un obstacle à son utilisation. Cette loi s'applique toutefois généralement bien à l'étude des modules annuels des variables hydro-météorologiques en climat tempéré.

2 Loi log-normale

La loi log-normale est préconisée par certains hydrologues dont V.-T. Chow qui la justifient en argumentant que l'apparition d'un événement hydrologique résulte de l'action combinée d'un grand nombre de facteurs qui se multiplient. Dès lors la variable aléatoire $X = X_1 \cdot X_2 \cdot \cdots \cdot X_r$ suit une loi log-normale. En effet le produit de r variables se ramène à la somme de r logarithmes de celles-ci et le théorème central-limite permet d'affirmer la log-normalité de la variable aléatoire.

3 Loi de Gumbel

E.-J. Gumbel postule que la loi double exponentielle, ou loi de Gumbel, est la forme limite de la distribution de la valeur maximale d'un échantillon de n valeurs. Le maximum annuel d'une variable étant considéré comme le maximum de 365 valeurs journalières, cette loi doit ainsi être capable de décrire les séries de maxima annuels.

Il est à remarquer que plus le nombre de paramètres d'une loi est grand, plus l'incertitude dans l'estimation est importante. Pratiquement il est par conséquent préférable d'éviter l'utilisation de lois à trois paramètres ou plus.

I.2.2 Utilisation des tests d'adéquation

Beaucoup d'auteurs utilisent les tests d'adéquation (voir paragraphe contrôle de l'ajustement) comme technique permettant de choisir le modèle fréquentiel approprié. Cependant il est à remarquer qu'un test statistique ne permet que de conclure au rejet, ou à l'acceptation, de

l'hypothèse nulle H_0 Il n'est pas en mesure de comparer plusieurs modèles fréquentiels et de choisir le meilleur.

I.2.3 Ajustement du modèle fréquentiel

Dans ce chapitre nous étudierons les techniques de l'ajustement ou du calage d'un modèle fréquentiel à une série de données : il s'agit de définir les paramètres de la loi retenue. Nous utiliserons comme support pédagogique la loi de Gumbel, fréquemment utilisée en hydrologie, pour modéliser les événements extrêmes, les pluies notamment.

1 Présentation de la loi de Gumbel

La distribution des valeurs extrêmes provenant de n'importe quelle distribution converge vers la loi des extrêmes généralisés (GEV). La distribution de cette loi s'exprime de la manière suivante :

$$F(x) = \exp\left(-\left(1 - c\frac{x-a}{b}\right)^{1/c}\right)$$

où a est le paramètre de position, b le paramètre d'échelle et c le paramètre de forme. 3 lois peuvent être distinguées en fonction des valeurs de c . Leurs caractéristiques sont résumées dans le tableau suivant :

c	type	nom	borne inférieure	borne supérieure
$c > 0$	III	$-X \approx$ Weibull	$-\infty$	$a + \frac{b}{c}$
$c = 0$	I	Gumbel	$-\infty$	$+\infty$
$c < 0$	II	Fréchet	$a + \frac{b}{c}$	$+\infty$

La fonction de répartition de la loi de Gumbel s'exprime de la manière suivante :

$$F(x) = \exp\left(-\exp\left(-\frac{x-a}{b}\right)\right)$$

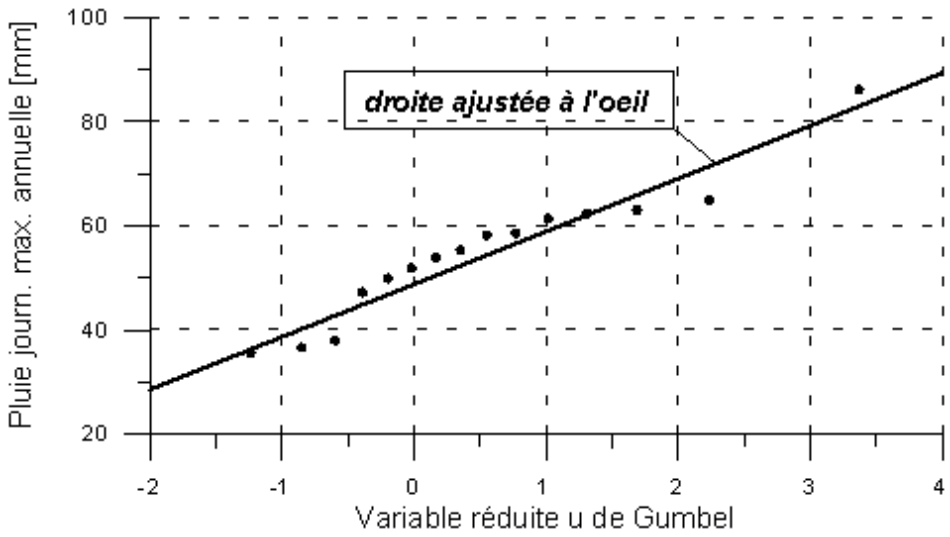
Posons la variable réduite suivante $u = \frac{x - a}{b}$. La distribution s'écrit alors comme suit : $F(x) = \exp\left(-\exp(-u)\right)$ et $u = -\ln\left(-\ln\left(F(x)\right)\right)$. L'avantage d'utiliser la variable réduite est que l'expression d'un quantile est alors linéaire. En effet pour trouver la valeur x_q d'un quantile, correspondant à la distribution $F(x_q) = q$, en fonction des deux paramètres a et b, il suffit d'utiliser la relation suivante :

$x_q = a + bu_q$

Techniques d'ajustement

1 Méthode graphique

Dans le cas d'un ajustement selon la loi de Gumbel, la méthode graphique repose sur le fait que l'expression d'un quantile correspond à l'équation d'une droite. En conséquence, les points de la série à ajuster peuvent être reportés dans un système d'axes $x - u$; il est alors possible de tracer la droite qui passe le mieux par ces points et d'en déduire les deux paramètres a et b définissant la loi. Le graphique ci-dessous montre un ajustement à l'œil. Dans la mesure où les points x_i sont connus (ils font partie de la donnée du problème), il suffit de définir les coordonnées u_i correspondant à chaque point pour pouvoir le positionner dans le graphique. Ces coordonnées se déterminent à partir de la relation inverse de la fonction de répartition qui donne u en fonction de la distribution $F(x)$. Il s'agit donc essentiellement d'estimer la probabilité de non-dépassement $F(x_i)$ qu'il convient d'attribuer à chaque valeur x_i.



Principe de la méthode d'ajustement graphique.

Il existe de nombreuses formules d'estimation de la fonction de répartition $\hat{F}(x)$ à l'aide de la distribution empirique. Elles reposent toutes sur un tri de la série par valeurs croissantes permettant d'associer à chaque valeur son rang r. Ces formules peuvent être résumées par une relation générale qui garantit la symétrie autour de la médiane :

$$\hat{F}\left(x_{[r]}\right) = \frac{r-\alpha}{n+1-2\alpha}$$

où n est la taille de l'échantillon, $x_{[r]}$ la valeur de rang r et α un coefficient compris entre 0 et 0.5. Le tableau ci-dessous présente quelques exemples de distributions empiriques :

Nom	α	**Formule**
Weibull	0	$\frac{r}{n+1}$
Cunnane	0.4	$\frac{r-0.4}{n+0.2}$
Gringorten	0.44	$\frac{r-0.44}{n+0.12}$
Hazen	0.5	$\frac{r-0.5}{n}$

Des simulations ont montré que pour la loi de Gumbel, il est judicieux utiliser la distribution empirique de Hazen.

L'ajustement graphique, bien qu'étant une méthode approximative, a le très grand avantage de fournir une représentation visuelle des données et de l'ajustement. Celle-ci constitue un aspect essentiel du jugement porté sur l'adéquation entre la loi choisie et les données traitées, quelle que soit la méthode d'ajustement utilisée.

L'ajustement graphique est une approximation de la méthode statistique des moindres rectangles. Il est à remarquer cependant que, si un seul point parmi les données est fortement décalé par rapport aux autres, la méthode graphique est difficile à réaliser. En effet l'œil humain a de la peine à juger le poids à donner à ce point. Dans ce cas, des méthodes statistiques rigoureuses doivent être utilisées.

Méthode des moments

La méthode des moments consiste à égaler les moments échantillonnaux et les moments théoriques de la loi choisie. Soit $x_1, x_2, \cdots x_n$ l'échantillon de données à disposition. Posons $\hat{\mu} = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ et $\hat{\sigma}^2 = s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$ les estimateurs standard de la moyenne et de la variance. Les deux premiers moments théoriques de la loi de Gumbel s'expriment à partir des paramètres de position et d'échelle de la manière suivante :

$$\begin{cases} \mu = a + b\gamma \\ \sigma^2 = \frac{\pi^2}{6} b^2 \end{cases}$$ avec $\gamma = 0.5772$ (constante d'Euler).

On obtient donc les formules suivantes pour l'estimation par la méthode des moments :

$$\begin{cases} \hat{b} = \frac{\sqrt{6}}{\pi}\hat{\sigma} \\ \hat{a} = \hat{\mu} - \hat{b}\gamma \end{cases}$$

Méthode des L-moments

Le but de cette méthode est de réaliser un ajustement lorsque les moments classiques ne conviennent pas. Les deux paramètres a et b sont obtenus très simplement à partir des valeurs des deux premiers L-moments de la loi de Gumbel et des estimations calculées sur l'échantillon :

$$\begin{cases} \hat{b} = \frac{\hat{\lambda}_2}{\ln 2} \\ \hat{a} = \hat{\lambda}_1 - \hat{b}\gamma \end{cases}$$ avec $\gamma = 0.5772$ (constante d'Euler).

Méthode des moindres rectangles

La solution des moindres rectangles conduit à trouver la droite bissectrice des solutions classiques de la régression par moindres carrés de y en x d'une part et de x en y d'autre part. Cette méthode revient donc à minimiser la distance du point à sa projection orthogonale sur la droite de régression. Dans le cas de la loi de Gumbel l'axe x est remplacé par l'axe u de la variable réduite de Gumbel et l'axe y par celui de la variable hydrologique étudiée que nous notons ici x.

Nous obtenons par cette méthode les estimateurs suivants :

$$\begin{cases} \hat{b} = \dfrac{s_x}{s_u} \\ a = \bar{x} - \hat{b}\bar{u} \end{cases}$$

Méthode du maximum de vraisemblance

La vraisemblance offre une approche générale à l'estimation de paramètres inconnus à l'aide de données. Soit $x_1, x_2, \cdots x_n$ un échantillon provenant d'une loi $F_\theta(x)$, où θ est un paramètre inconnu qui peut être réel ou multivarié.

La fonction de vraisemblance, qu'il s'agit de maximiser, s'écrit :

$V = \prod_{i=1}^{n} f_\theta(x_i)$ où f est la densité de probabilité.

Souvent pour se simplifier le calcul, en remplaçant le produit par une somme, il est judicieux de maximiser le logarithme de la fonction de vraisemblance. On obtient dans le cas de la loi de Gumbel les estimateurs suivants :

$$\begin{cases} \hat{b} = \bar{x} - \dfrac{\sum_{i=1}^{n} x_i \exp\left(-\dfrac{x_i}{\hat{b}}\right)}{\sum_{i=1}^{n} \exp\left(-\dfrac{x_i}{\hat{b}}\right)} \\ \hat{a} = \hat{b} \ln\left(\dfrac{n}{\sum_{i=1}^{n} \exp\left(-\dfrac{x_i}{\hat{b}}\right)}\right) \end{cases}$$

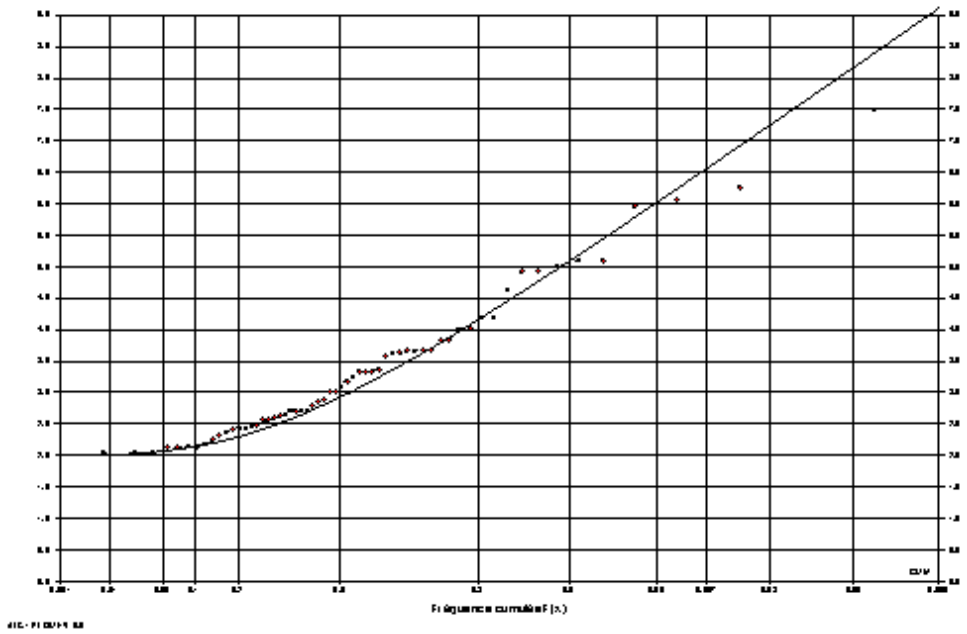
La première équation doit être résolue de façon itérative. Dans ce cas la solution de la méthode des moments peut par exemple être utilisée comme première approximation.

Lorsque la taille de l'échantillon est faible, la méthode du maximum de vraisemblance fournit une estimation biaisée des paramètres. Il s'agit, dans ce cas, d'utiliser la correction proposée par Fiorentino et Gabriele.

Contrôle de l'ajustement

1 Examen visuel de l'ajustement

L'examen visuel du graphique représentatif de l'ajustement réalisé, même s'il peut paraître rudimentaire, reste un des bons moyens pour juger de la qualité d'un ajustement et devrait toujours constituer un préambule à tout test statistique. La figure ci-dessous en présente un exemple.



Ajustement de la série tronquée des débits de pointe [m3/s] du Nozon à Orny (1923-1931) à une loi exponentielle.

1.Le test chi-carré de K. Pearson

Ce test est appliqué dans une situation où l'on observe la répartition de n objets dans I classes. Il est utilisé pour tester l'hypothèse que la répartition des données s'effectue selon une distribution théorique. On se pose donc la question de l'adéquation d'une distribution théorique à des données.

Pour tester l'adéquation d'une répartition théorique, on dispose de deux éléments. D'une part, n observations réparties dans I cellules. Cela se résume par :

n_1	n_2	...	n_I

où n_i est le nombre d'observations dans la i ème cellule.

D'autre part, une distribution théorique qui fixe la probabilité p_i($i = 1,2,\dots I$) de chaque cellule.

p_1	p_2	...	p_I

Le score du test de Pearson est une mesure de la distance entre la répartition empirique et la loi théorique. Elle se base sur la répartition des n objets selon la loi théorique :

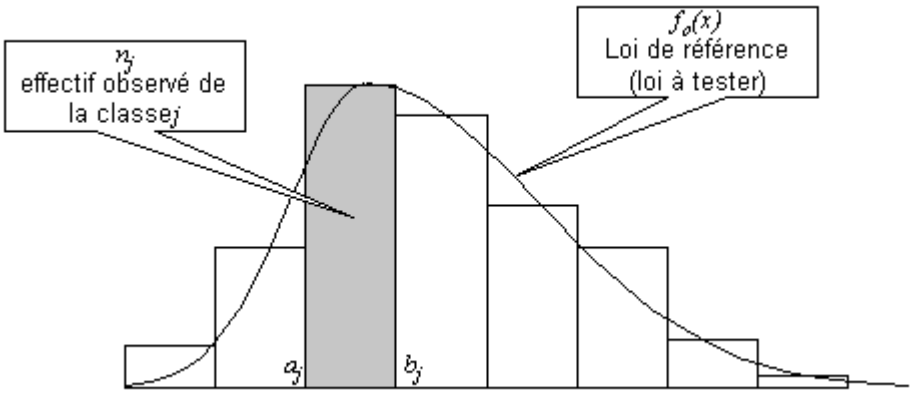
np_1	np_2	...	np_I

Ensuite, on calcule :

$$\text{Pearson}_{obs} = \sum_{i=1}^{I} \frac{(n_i - np_i)^2}{np_i}$$

L'hypothèse nulle que l'on teste avec le test de Pearson est H_0 : " La distribution théorique est la vraie distribution sous-jacente aux données ". On peut démontrer que : la distribution de la statistique de Pearson sous l'hypothèse H_0 est bien approchée par une loi χ^2_{I-1} (chi-carré avec $I-1$ degrés de liberté), si le nombre espéré np_i est suffisamment grand (≥ 5).

On rejette donc l'hypothèse nulle si $\text{Pearson}_{obs} > q\chi^2_{I-1}(95\%)$ où $q\chi^2_{I-1}(95\%)$ est le 95%-quantile d'une loi χ^2_{I-1}. La figure ci-dessous illustre le principe de ce test.



Principe du test de chi-carré.

Dans le cas où la variable aléatoire considérée est continue, il faut discrétiser, ce qui introduit un élément d'ambiguïté. Pour le cas continu il existe un autre test qui utilise la distribution empirique et qui, en règle générale, est plus puissant que le test de Pearson.

3 Le test de Kolmogorov-Smirnov

Le test de Kolmogorov-Smirnov consiste à mesurer, pour une variable aléatoire continue, la plus grande distance entre la distribution théorique $F_0(x)$ et la distribution expérimentale $F(x)$. Nous avons donc $H_0 : F(x) = F_0(x) \; \forall x$ et $H_1 : F(x) \neq F_0(x)$ pour au

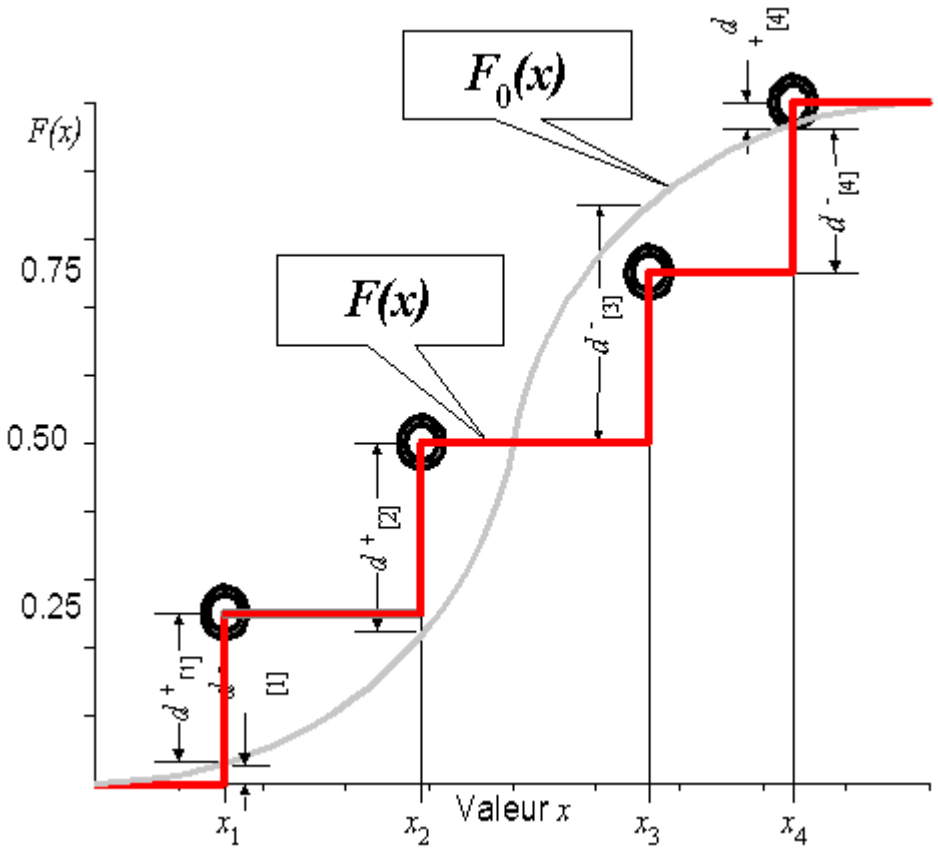
moins une valeur de x. La distribution empirique, ou observée, se calcule, dans la théorie de Kolmogorov-Smirnov, par la relation classique :

$$\hat{F}\left(x_{[r]}\right) = \frac{r}{n}$$

On définit alors la statistique d comme suit :

$$\begin{cases} d^+ = \text{Max}\left\{\frac{\text{r}}{\text{n}} - F_0\left(x_{[r]}\right)\right\}, \forall r = 1,2,\ldots n \\ d^- = \text{Max}\left\{F_0\left(x_{[r]}\right) - \frac{r-1}{n}\right\}, \forall r = 1,2,\ldots n \\ d = Max\left\{d^+; d^-\right\} \end{cases}$$

La statistique d est tabulée dans plusieurs ouvrages. Le principe de ce test est illustré dans la figure suivante :



Principe du test de Kolmogorov-Smirnov.

4 Test d'Anderson–Darling

Le test d'Anderson-Darling consiste à comparer la distribution théorique $F_0(x)$ à la distribution expérimentale $F(x)$ en calculant la statistique suivante :

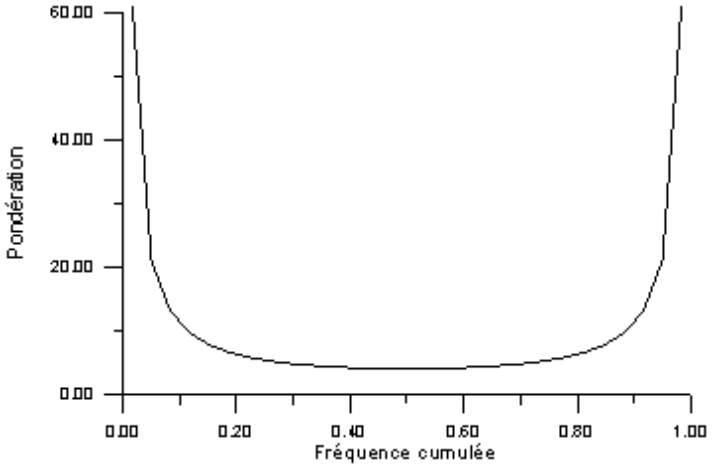
$$\int_{-\infty}^{+\infty}[F(x)-F_0(x)]^2 w(x)dF(x)$$

où $w(x)$ est une fonction de pondération.

Le cas standard d'Anderson-Darling correspond à la fonction de pondération suivante :

$$w(x)=\frac{1}{F_0(x)[1-F_0(x)]}$$

qui permet de donner plus d'influence aux faibles et fortes fréquences. Cela conduit à la statistique notée A^2.



Fonction de pondération du test A^2 d'Anderson-Darling.

En modifiant la fonction de pondération en $w(x)=\frac{1}{1-F_0(x)}$ on obtient un test sensible au comportement pour des fréquences rares. Cette procédure de test peut son se révéler particulièrement utile lorsqu'on s'intéresse, comme c'est généralement le cas en hydrologie, aux valeurs extrêmes.

5 Analyse des incertitudes

A ce stade de l'analyse nous disposons d'un modèle fréquentiel $\hat{F}(x)$, obtenu après plusieurs étapes. On est donc en droit de se poser la question de sa fiabilité ou degré de confiance que l'on peut y accorder.

5.1 L'intervalle de confiance

L'incertitude liée au phénomène de la fluctuation d'échantillonnage peut être évaluée par la procédure classique de l'intervalle de confiance. La construction d'un tel intervalle peut-être effectuée par la méthode dite de l'erreur-type.

Dans ce cas, la construction de l'intervalle de confiance nécessite la connaissance de trois grandeurs :

1. L'estimation du quantile, qui est donnée par l'équation $\hat{x}_q = \hat{a} + \hat{b}u_q$.
2. L'erreur-type σ_{x_q}, dont la détermination sera développée ci-dessous.
3. La forme de la distribution d'échantillonnage, considérée dans la plupart des cas comme " normale ".

5.1.1 Erreur-type d'un quantile

$\hat{x}_q = \hat{\mu} + \frac{\sqrt{6}}{\pi} \cdot \hat{\sigma}(u_q - \gamma) = \hat{\mu} + K_q \cdot \hat{\sigma}$ Lorsque les paramètres a et b de la loi de Gumbel ont été estimés par la méthode des moments l'expression d'un quantile x_q peut s'écrire $\hat{x}_q = \hat{a} + \hat{b}u_q$. En substituant les estimations de a et b, on obtient

avec $\gamma = 0.5772$, constante d'Euler et K_q est appelé facteur de fréquence dans la formulation désormais classique d'un quantile aux USA : $K_q = \frac{\sqrt{6}}{\pi}(u_q - \gamma)$.

En utilisant les formules de calcul de la variance d'une fonction de variables aléatoires et en remplaçant la variance σ^2 par son estimation s^2 on trouve finalement la formule de Dick et Darwin :

$$\sigma_{x_q} = \frac{s}{\sqrt{n-1}}\sqrt{1 + 1.1396\,K_q + 1.1K_q^2}$$

où en introduisant u_q :

$$\sigma_{x_q} = \frac{s}{\sqrt{n-1}}\sqrt{0.7099 + 0.1165\,u_q + 0.6687\,u_q^2}$$

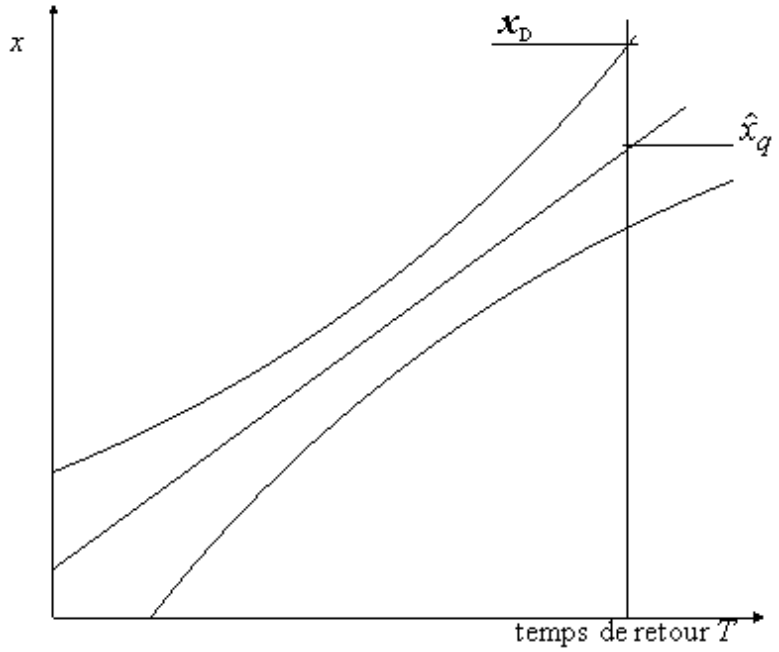
Lorsque les paramètres ont été estimés par la méthode du maximum de vraisemblance la procédure de calcul de l'erreur-type d'un quantile se base sur la méthode delta (méthode de linéarisation se basant sur le développement de Taylor). Pour la loi de Gumbel, on obtient :

$$\sigma_{x_q} = \frac{\hat{b}}{\sqrt{n}}\sqrt{1.1086 + 0.514 \cdot u_q + 0.6979 \cdot u_q^2}$$

Souvent la valeur de dimensionnement x_D (ou valeur de projet) à adopter est déterminée à partir de l'erreur-type par une relation telle que celle ci-:

$$x_D = \hat{x}_q + K\sigma_{\hat{x}_q}$$

où K est un facteur, communément nommé facteur de fréquence, dépendant de la forme de la loi de distribution d'échantillonnage et du niveau de confiance $1-\alpha$ désiré. Un tel intervalle de confiance est représenté par la figure ci-dessous.



Intervalle de confiance à $1-\alpha$ de la valeur de dimensionnement.

Référence bibliographique

https://echo2.epfl.ch/e-drologie/chapitres/annexes/AnalFrequ.html