RAPPORT GÉNÉRAL DU LOGICIEL AJUSTE II: Théorie et application

Rapport de recherche R-421

RAPPORT GÉNÉRAL DU LOGICIEL *AJUSTE-II* : THÉORIE ET APPLICATION

Rapport rédigé pour

Projet de partenariat Hydro-Québec / CRSNG
Modélisation des événements hydrologiques extrêmes
à partir de lois statistiques

par

Luc Perreault
Bernard Bobée
Pierre Legendre

Institut National de la Recherche Scientifique, INRS-Eau
2800, Einstein, CP 7500, Sainte-Foy, Québec, G1V 4C7

Projet conjoint
Hydro-Québec / CRSNG - IDR0120996

Rapport de recherche No R-421

Juillet 1994

TABLE DES MATIÈRES

LISTE DES TABLEAUX

LISTE DES FIGURES

1 INTRODUCTION

1.1 Problématique

Les activités d'Hydro-Québec dans le domaine de l'aménagement et de la réfection des centrales hydroélectriques impliquent un grand nombre d'études concernant les débits extrêmes de crue. Ces études sont requises pour la conception des évacuateurs, des barrages et des dérivations provisoires. La planification et le dimensionnement de ces ouvrages hydrauliques reposent donc sur une estimation adéquate des événements extrêmes de crue. En effet, une surestimation des crues peut entraîner un sur-dimensionnement des ouvrages hydrauliques et conduire à des coûts de construction supplémentaires. Une sous-estimation des crues peut causer des défaillances d'ouvrages conduisant à des inondations qui se traduisent par des dégâts matériels importants et parfois par des pertes de vies humaines.

Un des outils privilégié par les hydrologues pour estimer les débits extrêmes de crue est l'analyse de fréquence des crues (flood frequency analysis). Cette approche a pour objectif l'utilisation des mesures d'événements hydrologiques extrêmes passés pour estimer les probabilités futures d'occurrence (inférence statistique). Les quatre étapes principales de cette procédure sont :

- la sélection d'un échantillon de mesures de débits extrêmes satisfaisant certaines hypothèses statistiques de base (chapitre 5);
- le choix d'un modèle paramétrique considéré comme une approximation de la distribution théorique inconnue pouvant représenter adéquatement un échantillon donné (loi de probabilité, chapitres 2 et 5);
- l'ajustement du modèle aux données à l'aide de la méthode d'estimation la plus adéquate compte tenu des objectifs visés (estimation des paramètres de la loi, chapitres 3 et 4);
- la détermination des événements extrêmes x_T de période de retour T (quantiles de la loi) pour faire une inférence statistique (chapitres 3 et 4).

Plusieurs lois de probabilité ont été utilisées en hydrologie comme modèle afin de représenter les débits extrêmes. Mentionnons en particulier les lois de la famille gamma :

gamma (G), log-gamma (LG), Pearson Type 3 (P3), log-Pearson Type 3 (LP3) et gamma généralisée (GG). Bobée et Ashkar (1991) ont présenté une synthèse des principales méthodes d'estimation des paramètres des lois G, LG, P3, LP3 et GG développées à l'INRS-Eau et ailleurs dans le monde. Un logiciel d'ajustement automatique de ces distributions, incluant les méthodes d'estimation des paramètres des lois de la famille gamma, accompagne cette synthèse. Ce logiciel, nommé HFA (*AJUSTE-I* pour la version française), a été essentiellement conçu pour des fins de recherches et ne comprend que les lois de la famille gamma. Il ne répond pas à tous les besoins pratiques requis par des études hydrologiques menées par une compagnie comme Hydro-Québec (plus grande variété de lois, tests d'adéquation supplémentaires, etc.). C'est pourquoi, dans le cadre d'un projet de partenariat financé par Hydro-Québec et le Conseil de Recherche en Sciences Naturelles et en Génie (CRSNG), une nouvelle version du logiciel (*AJUSTE-II*) répondant mieux aux besoins d'Hydro-Québec a été développée.

1.2. Le logiciel *AJUSTE-II*

Le logiciel *AJUSTE-II* est une version améliorée de HFA. Une amélioration majeure apportée au logiciel concerne sa structure et sa convivialité. *AJUSTE-II* a été écrit de façon modulaire en utilisant le langage C++ orienté objet et la librairie XVT, alors que HFA utilisait le Fortran. La librairie XVT a permis de développer *AJUSTE-II* de façon à ce qu'il puisse être utilisé dans les environnements graphiques *Windows* (DOS) et *X-Windows* (UNIX).

Plusieurs améliorations en ce qui a trait aux méthodes statistiques ont aussi été apportées. HFA contient un grand nombre de méthodes d'estimation des paramètres pour chacune des lois de la famille gamma. Par exemple, dix méthodes sont disponibles pour la loi log-Pearson Type 3. Si pour des études théoriques il est intéressant d'employer un logiciel contenant l'ensemble des méthodes disponibles, en pratique, il n'est pas nécessaire de disposer d'autant de méthodes. Nous n'avons donc retenu pour les lois de la famille gamma, suite à une étude comparative (Messaoudi, 1994), que les méthodes d'estimation des paramètres les plus efficaces :

- d'un point de vue descriptif (interpolation);
- d'un point de vue prédictif (extrapolation).

Suite à l'utilisation pratique de HFA, et à de nouveaux développements théoriques, nous avons ajouté quelques lois souvent utilisées en hydrologie, ainsi que les trois types de lois de Halphen (Perreault *et al.*, 1994). Plusieurs tests ont également été ajoutés pour vérifier les hypothèses de base et l'adéquation du modèle. Le Tableau 1.1 présente les principales différences entre les logiciels HFA et ***AJUSTE-II*** concernant les lois de probabilité, les méthodes d'estimation et les tests statistiques considérés.

Tableau 1.1. Contenu comparatif des logiciels HFA et ***AJUSTE-II***.

	HFA[a]	**AJUSTE II (méthodes d'estimation retenues[b])**
Lois	log-gamma	
	gamma	gamma (MXV, MOM)
	Pearson Type 3	Pearson Type 3 (MXV, MOM)
	log-Pearson Type 3	log-Pearson Type 3 (MOM, SAM, WRC)
	gamma généralisée	gamma généralisée (MOM, MM1)
		gamma inverse (MXV)
		GEV (MXV, MOM, MMP)
		Gumbel (MXV, MOM)
		Weibull (MXV, MOM)
		Halphen (Types A, B et B^{-1}) (MXV)
		normale (MXV)
		log-normale à 2 paramètres (MXV, MOM)
		log-normale à 3 paramètres (MXV, MOM)
		exponentielle à 2 paramètres (MXV)
Tests		
Indépendance	Wald-Wolfowitz	Wald-Wolfowitz
Homogénéité	Mann-Withney	Wilcoxon
Stationnarité		Kendall
Valeurs singulières	Grubbs et Beck	Discordance pour chaque loi
Adéquation		Khi-deux pour chaque loi

[a] Les méthodes d'estimation associées aux lois de la famille gamma incluses dans HFA sont décrites en détail dans Bobée et Ashkar (1991).

[b] MXV=Maximum de vraisemblance, MOM=Méthode des moments, WRC=Méthode des moments sur la série des logarithmes des observations, SAM=Sundry averages method, MM1=Méthode qui utilise les deux premiers moments centrés de l'échantillon et le premier moment centré de l'échantillon des logarithmes, MMP=Méthode des moments pondérés.

Finalement, en ce qui concerne les aspects pratiques, ***AJUSTE-II*** présente les principales améliorations suivantes par rapport à HFA :

- la taille des échantillons est seulement limitée par la mémoire disponible alors que dans HFA la taille des échantillons était limitée à 200 observations. L'usager a aussi la possibilité de désactiver certaines données qu'il croit douteuses ou qu'il peut vouloir éliminer. Enfin, l'usager peut considérer plusieurs échantillons dans la même session de travail;
- pour les opérations fréquemment utilisées, certaines touches de raccourci ont été définies;
- une aide contextuelle est disponible en tout temps. Cette aide est autant technique (pour l'utilisation du logiciel) que théorique (pour guider l'usager dans ses décisions);
- l'usager a la possibilité de faire imprimer un tableau synthèse de tous les ajustements effectués lors d'une session de travail;
- l'usager a la possibilité de modifier les caractéristiques des graphiques et d'effectuer des agrandissements (zoom) sur certaines parties du graphique;

Le présent manuel vise à décrire les méthodes statistiques incluses dans le logiciel ***AJUSTE-II*** pour effectuer une analyse hydrologique de fréquence. Il est complété par un guide de l'usager et un manuel du programmeur (Perron, 1994a, 1994b).

Le chapitre II présente les principales notions statistiques de base nécessaires à l'emploi du logiciel ***AJUSTE-II***. On y donne entre autres les définitions de la fonction de densité de probabilité (f.d.p.), des moments et coefficients adimensionnels, ainsi que des événements x_T de période de retour T.

Dans le chapitre III, une description des méthodes d'estimation des paramètres de lois de probabilité est faite et les principales propriétés de chacune d'elles y sont données. Une section traite aussi de l'estimation des quantiles (événements x_T de période de retour T). Le chapitre IV est consacré aux mesures de précision des estimateurs issus des différentes méthodes d'estimation. On y présente le calcul des variances asymptotiques de ces estimateurs et des quantiles estimés. Il sera de plus question, dans ce chapitre, de la construction d'intervalles de confiance asymptotiques.

Au chapitre V, une introduction aux tests statistiques permettant de valider les hypothèses de base du modèle et d'en apprécier l'adéquation est donnée. Les notions de base des tests

d'hypothèses sont d'abord présentées pour ensuite décrire les tests inclus dans le logiciel ***AJUSTE-II***. Enfin, le chapitre VI est consacré à la description d'une analyse hydrologique de fréquence complète appliquée aux débits maximums annuels de la rivière Harricana.

2 NOTIONS STATISTIQUES DE BASE

Plusieurs processus hydrologiques peuvent être analysés et expliqués en termes de probabilités à cause de leur caractère aléatoire. Des méthodes statistiques sont disponibles pour organiser, présenter et réduire de tels ensembles de données de façon à faciliter leur interprétation et leur évaluation. L'une de ces approches, l'analyse hydrologique de fréquence (AHF), est souvent employée comme première phase des études de conception pour l'aménagement ou la réfection des centrales hydroélectriques. Cette méthode statistique a comme objectif principal d'établir la relation existant entre des événements extrêmes (crues, étiages, etc.) et leurs probabilités de dépassement ou de non-dépassement. L'AHF repose sur certaines notions statistiques de base qu'il est nécessaire de connaître afin d'appliquer cette approche. Nous les présentons brièvement dans ce chapitre. Plus de détails peuvent être obtenus dans des ouvrages de base en probabilité et en statistique (par exemple, Lehmann, 1983; Bickel et Doksum, 1977; Kendall et Stuart, 1987).

2.1 Variable aléatoire et modèle de loi de probabilité

2.1.1 Notion de variable aléatoire

Une **variable aléatoire**, que l'on note X, est le résultat d'un processus incertain. C'est une quantité (débit, précipitation, vent, etc.) qui ne peut être prédite avec certitude. On appelle **domaine** de la variable X, l'ensemble $\mathcal{D}$ des valeurs que peut prendre la variable aléatoire.

De telles variables peuvent être **continues** ou **discrètes**. La plupart des variables hydrologiques sont continues et sont analysées en utilisant des modèles de lois de probabilité continues. Par exemple, les valeurs de débits dans l'hydrogramme présenté à la Figure 2.1 peuvent être égales à tout nombre réel positif. L'ensemble des nombres réels positifs constitue donc le domaine $\mathcal{D}$, et cette variable aléatoire est continue. Les valeurs prises par une variable aléatoire continue sont non dénombrables. Certaines études hydrologiques peuvent nécessiter l'analyse de variables aléatoires discrètes. Les variables aléatoires discrètes admettent seulement des valeurs spécifiées par un domaine discret, c'est-à-dire un ensemble de valeurs connues et dénombrables. Par exemple, on peut s'intéresser au nombre de débits maximums annuels inférieurs à un seuil Q sur une période de temps donnée. La variable ainsi définie ne peut admettre comme valeurs que les entiers positifs compris entre 1

et le nombre total d'années n sur la période de temps considérée: $\mathcal{D}= \{1, 2, 3, ..., n\}$. Cette variable aléatoire est alors discrète.

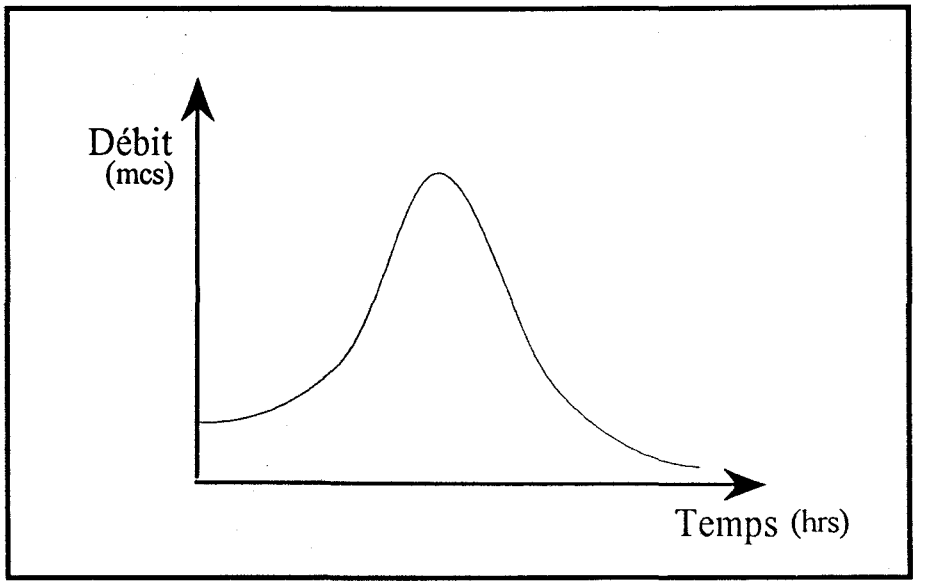


Figure 2.1. Variable aléatoire continue. Hydrogramme.

Les propriétés et les caractéristiques d'un processus hydrologique peuvent être décrites par la loi de probabilité de la variable aléatoire qui y est associée. Il est en effet important de connaître la probabilité d'occurrence de toutes les valeurs prises par la variable aléatoire qui nous intéresse. Il existe deux types de lois de probabilité selon la nature de la variable aléatoire : la loi de probabilité discrète et la loi de probabilité continue. Nous les présentons dans ce qui suit.

2.1.2 Loi de probabilité discrète

Considérons une variable aléatoire discrète X et l'ensemble $\mathcal{D} = \{x_1, x_2, ...\}$ des valeurs qu'elle peut prendre. Il est certainement intéressant de connaître la probabilité de chacun des événements possibles, c'est-à-dire la probabilité que la variable X prenne la valeur x_i pour tout i. On note cette probabilité $\text{Prob}\{X = x_i\}$. La fonction f définie par $f(x) = \text{Prob}\{X = x\}$ est appelée la **fonction de densité de probabilité discrète**, et possède les propriétés suivantes :

(i) $f(x) \geq 0$, pour tout x

(ii) $\sum_i f(x_i) = 1$

Exemple 2.1. *Soit $0 < \pi < 1$, alors la fonction f définie par*

$$f(x) = \begin{cases} \pi(1-\pi)^{x-1}, & si \quad x = 1, 2, 3, \ldots \\ 0, & \text{sinon} \end{cases} \tag{2.1}$$

est la fonction de densité de probabilité discrète appelée ***loi géométrique*** *de paramètre π. Cette loi de probabilité est utilisée en hydrologie, entre autres, pour déterminer la probabilité d'occurrence des durées des étiages (Mathier et al., 1992).*

La fonction f de l'Exemple 2.1 est bien une densité de probabilité car les conditions (i) et (ii) sont satisfaites. En effet, on a :

- $f(x) \geq 0$ puisque $0 < \pi < 1$
- $\sum_i f(x_i) = \pi \sum_{x=1}^{\infty} (1-\pi)^{x-1} = 1$, puisque $\sum_{x=1}^{\infty} (1-\pi)^{x-1} = \pi^{-1}$

La fonction de densité de probabilité discrète est représentée graphiquement par un diagramme en bâtons puisque les valeurs du domaine d'une variable aléatoire discrète sont dénombrables. La Figure 2.2 présente la densité d'une loi géométrique avec $\pi = 0{,}5$.

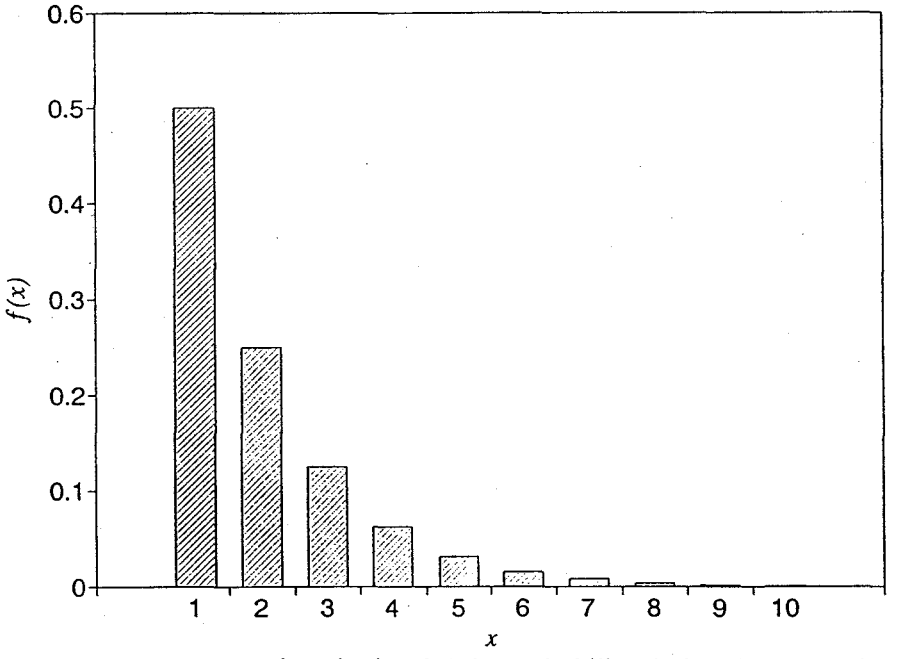


Figure 2.2. Fonction de densité de probabilité de la loi géométrique.

Nous avons discuté jusqu'ici du calcul de la probabilité de l'événement ponctuel $\{X = x\}$. On est souvent intéressé, toutefois, à déterminer la probabilité d'un événement plus large tel que $\{X \in D\}$[a] où D est un sous-ensemble du domaine $\mathcal{D}$. Si f est la densité de la variable aléatoire discrète X et que les événements dans D sont mutuellement exclusifs (c'est-à-dire, ne peuvent se réaliser simultanément), la probabilité $\text{Prob}\{X \in D\}$ est alors donnée par :

$$\text{Prob}\{X \in D\} = \sum_{x \in D} f(x) \tag{2.2}$$

Supposons que le sous-ensemble D est l'intervalle $]-\infty, t]$. Alors, la fonction $F(t)$ définie comme suit :

$$\begin{aligned} F(t) &= \text{Prob}\{X \in D\} \\ &= \text{Prob}\{X \le t\} \\ &= \sum_{x \le t} f(x) \end{aligned} \tag{2.3}$$

est appelée la **fonction de répartition discrète** de la variable aléatoire X (elle est aussi appelée quelquefois **fonction de distribution cumulée**). Cette fonction, qui donne la probabilité au non-dépassement de la valeur t, est très importante en hydrologie statistique comme on le verra plus loin dans ce chapitre.

Exemple 2.2. *La fonction de répartition correspondant à la loi géométrique définie à l'Exemple 2.1 est donnée par :*

$$F(t) = \sum_{x=1}^{t} \pi(1-\pi)^{x-1}, \quad t \ge 0 \tag{2.5}$$

En évaluant cette somme (on utilise la formule d'une progression géométrique finie), on déduit que :

$$F(t) = 1 - (1-\pi)^t, \quad t \ge 0 \tag{2.6}$$

Ainsi, si on suppose que le nombre d'années consécutives où l'on a observé un débit inférieur à un seuil donné est distribué selon une loi géométrique, l'équation (2.6) peut

---

[a] Le symbole $\in$ se lit "appartient à".

nous permettre d'évaluer la probabilité que la durée d'un étiage soit inférieure ou égale à t années.

2.1.3 Loi de probabilité continue

Considérons une variable aléatoire continue X. Une telle variable peut prendre toutes les valeurs comprises dans un intervalle donné. Comme on l'a vu à la section 2.1.1, ces valeurs constituent le domaine $\mathcal{D}$ de la variable aléatoire et sont non dénombrables. Cette caractéristique nous permet de donner une définition plus mathématique d'une variable aléatoire continue : une variable aléatoire X est dite **continue** si $\text{Prob}\{X = x\} = 0$ pour tout x. Puisque pour une variable aléatoire continue on ne peut considérer les événements ponctuels, il est plus approprié d'introduire d'abord la fonction de répartition F plutôt que la fonction de densité de probabilité f.

La fonction F définie par $F(x) = \text{Prob}\{X \le x\}$ pour tout x dans le domaine $\mathcal{D}$ est appelée la **fonction de répartition continue** et répond aux axiomes suivants :

(i) $0 \le F(x) \le 1$, pour tout x

(ii) F est une fonction non-décroissante de x

(iii) $F(-\infty) = 0 \quad et \quad F(+\infty) = 1$

La fonction de répartition représente la somme des probabilités des valeurs de la variable comprises dans l'intervalle $(-\infty, x]$ et est utile pour calculer différentes probabilités associées à la variable aléatoire X. Par exemple, on déduit de la définition que :

$$\text{Prob}\{a < X \le b\} = F(b) - F(a), \quad a \le b \tag{2.7}$$

Toutefois, en pratique, les lois de probabilité continues sont généralement définies en terme de fonction de densité de probabilité. Une **fonction de densité de probabilité continue** est une fonction f admettant que des valeurs positives, et telle que :

$$\int_{-\infty}^{+\infty} f(x)dx = 1 \tag{2.8}$$

On peut maintenant aisément faire la correspondance entre la fonction de répartition F et la densité f. Ainsi, si f est une fonction de densité de probabilité, alors la fonction F définie par:

$$F(x) = \int_{-\infty}^{x} f(y)dy, \quad -\infty < x < +\infty \tag{2.9}$$

est la fonction de répartition correspondante car elle satisfait aux axiomes (i)-(iii).

Tandis que les probabilités correspondant aux valeurs d'une variable discrète sont représentées graphiquement par un diagramme en bâtons (Figure 2.2), la représentation graphique de la densité de probabilité d'une variable aléatoire continue prend plutôt la forme d'une courbe sans discontinuité. La Figure 2.3 donne les graphiques de la fonction de densité de probabilité f et de la fonction de répartition F d'une variable aléatoire continue quelconque. Cette figure illustre bien la correspondance entre les deux fonctions. En effet, à chaque point sur la courbe F correspond une aire A sous la courbe f.

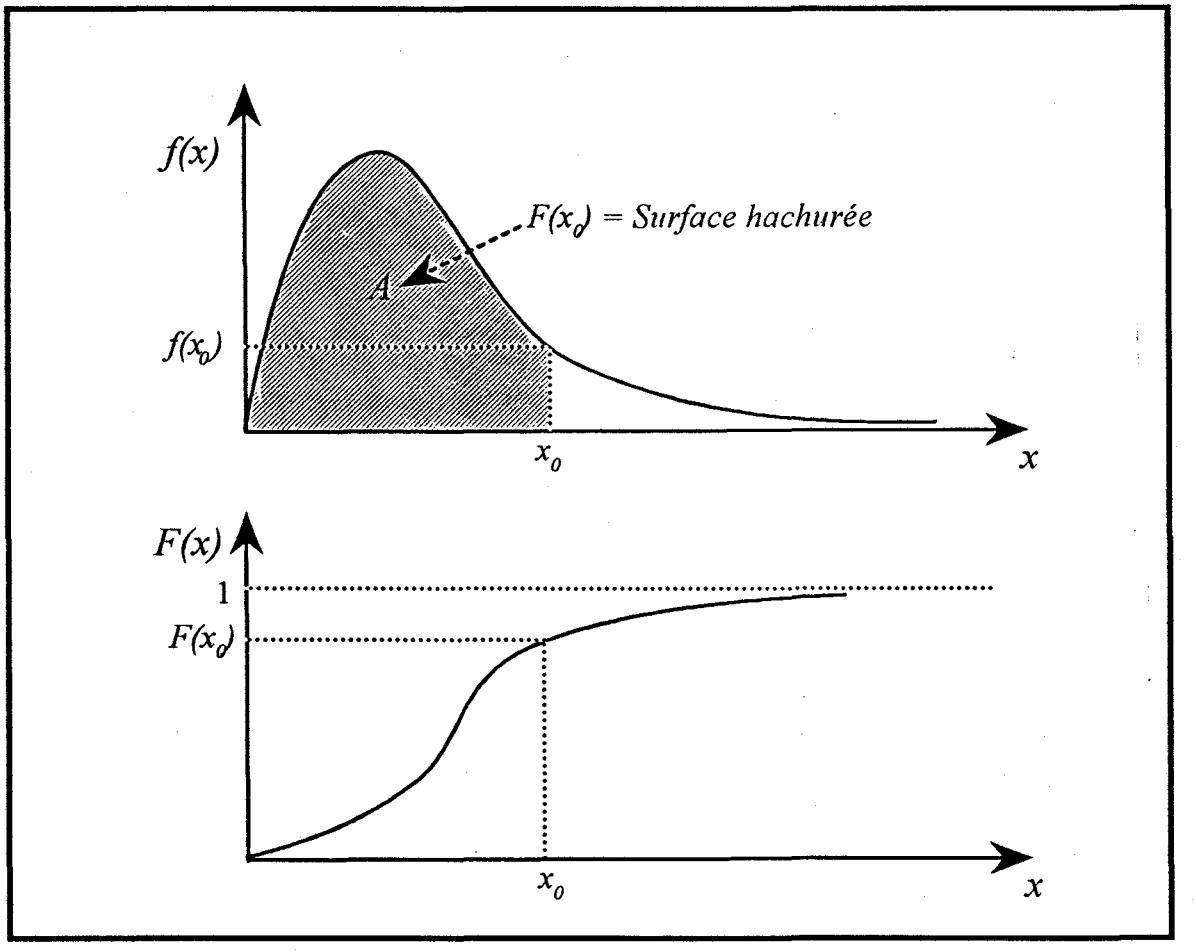


Figure 2.3. Fonction de densité de probabilité et fonction de répartition

On peut maintenant déduire des équations (2.7) et (2.9) que de façon générale :

$$\text{Prob}\{a < X \leq b\} = \int_a^b f(x)dx, \quad a \leq b \tag{2.10}$$

et que cette probabilité est représentée, pour une variable continue, par l'aire sous la courbe f lorsque x varie dans l'intervalle (a, b) (Figure 2.4).

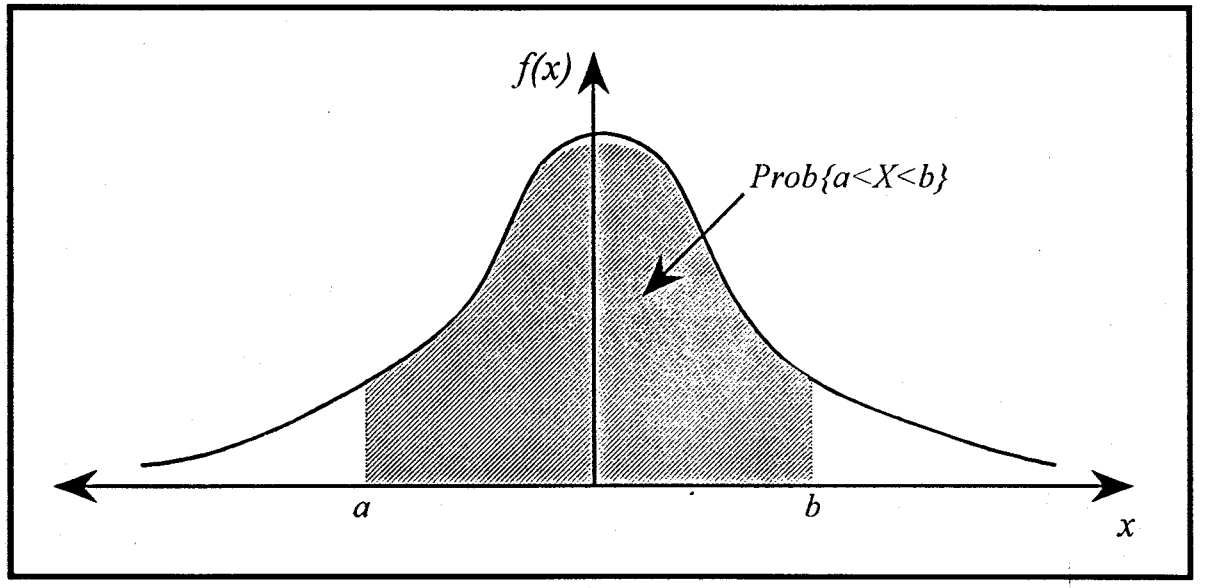


Figure 2.4. Représentation graphique de $\text{Prob}\{a < X \leq b\}$

Exemple 2.3. *Soit $\alpha > 0$ et $-\infty < u < +\infty$, alors la fonction f définie par*

$$f(x) = \frac{1}{\alpha}\exp\left[-\frac{x-u}{\alpha} - \exp\left(\frac{x-u}{\alpha}\right)\right], \quad -\infty < x < +\infty \tag{2.11}$$

est la fonction de densité de probabilité continue appelée loi Gumbel de paramètres α et u. Cette loi de probabilité, connue également sous le nom de loi des valeurs extrêmes de Type 1 (EV1), est utilisée en hydrologie particulièrement pour déterminer la probabilité d'occurrence des débits maximums annuels (Perreault et al., 1992a).

Selon l'équation (2.9), la fonction de répartition correspondante est donnée par :

$$F(x) = \int_{-\infty}^{x} \frac{1}{\alpha}\exp\left[-\frac{y-u}{\alpha} - \exp\left(\frac{y-u}{\alpha}\right)\right]dy, \quad -\infty < x < +\infty \tag{2.12}$$

En évaluant cette intégrale (Perreault et al. 1992a), on en déduit que :

$$F(x) = \exp\left[-\exp\left(\frac{x-u}{\alpha}\right)\right], \quad -\infty < x < +\infty \tag{2.13}$$

Ainsi, si on suppose que les débits maximums annuels sont distribués selon une loi Gumbel, l'équation (2.13) peut nous permettre d'évaluer la probabilité que le débit maximum soit inférieur ou égal à x.

Il existe une infinité de formes de fonction de densité de probabilité. La Figure 2.5 indique quelques formes classiques. En particulier, la loi Gumbel de l'Exemple 2.3 est caractérisée par une asymétrie positive.

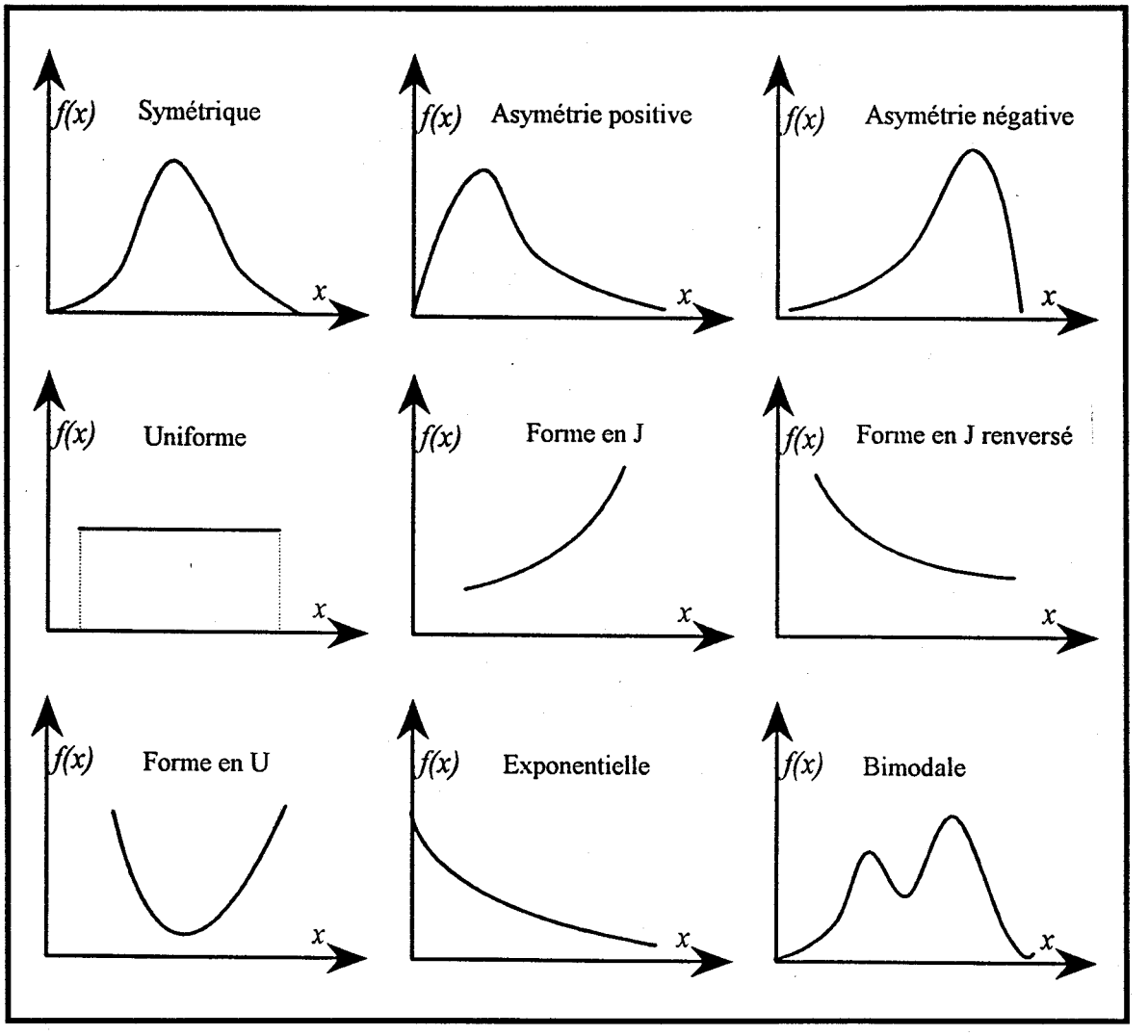


Figure 2.5. Formes classiques de fonctions de densité de probabilité continues

Il est important de mentionner que pour plusieurs lois de probabilité on ne peut calculer explicitement la fonction de répartition. On doit alors utiliser une méthode d'intégration numérique.

Étant donné que le logiciel ***AJUSTE-II*** a été conçu pour l'analyse de données hydrologiques continues, les notions statistiques de base présentées dans le reste du chapitre 2 ne concernent que les modèles de loi de probabilité continues. De plus, l'abréviation f.d.p. signifiant "fonction de densité de probabilité" sera utilisée dans les sections qui suivent .

2.2 Espérance mathématique, moments et coefficients théoriques

Un des concepts les plus utiles lorsqu'on utilise un modèle de loi de probabilité est l'espérance mathématique. Soit X une variable aléatoire possédant une f.d.p. $f(x)$, et considérons $u(X)$ une fonction telle que l'intégrale :

$$\int_{-\infty}^{+\infty} u(x) f(x) dx \tag{2.14}$$

existe (possède une solution finie : $\neq \pm\infty$). Cette intégrale est appelée **espérance mathématique** de $u(X)$ et est notée $E\{u(X)\}$. L'espérance mathématique est toujours une fonction des paramètres de la loi de probabilité de la variable aléatoire X.

L'application élémentaire de l'espérance mathématique concerne le cas où $u(X) = X$, on obtient alors :

$$E\{X\} = \int_{-\infty}^{+\infty} x f(x) dx \tag{2.15}$$

qui est la moyenne théorique de la variable aléatoire X : c'est une mesure de tendance centrale. Une autre application importante est le calcul de la variance $Var\{X\}$ de la variable X qui en caractérise la dispersion. La variance de X est définie comme l'espérance mathématique de $(X - E\{X\})^2$:

$$Var\{X\} = \int_{-\infty}^{+\infty} (x - E\{X\})^2 f(x) dx \tag{2.16}$$

Une mesure de dispersion équivalente et souvent utilisée en pratique est **l'écart-type** qui est tout simplement la racine carrée de la variance. La moyenne, la variance et l'écart-type d'une variable aléatoire sont généralement notés μ, σ^2 et σ respectivement.

La notion d'espérance mathématique conduit à la définition des moments d'une variable aléatoire qui sont utiles pour construire des indicateurs de la forme de la fonction de densité de probabilité et pour estimer les paramètres de la loi de probabilité (chapitre 3). Le **moment non-centré d'ordre *r*** que l'on note $\mu'_r(X)$ est obtenu en calculant l'espérance mathématique de $u(X) = X^r$:

$$\mu'_r(X) = E\{X^r\} = \int_{-\infty}^{+\infty} x^r f(x)dx \tag{2.17}$$

Le moment non-centré d'ordre 1, c'est-à-dire $\mu'_1(X)$, correspond donc à la moyenne théorique de la variable aléatoire X distribuée selon la loi F[b] . Le **moment centré d'ordre *r*** de la variable X, noté $\mu_r(X)$, est obtenu en calculant l'espérance mathématique de $u(X) = (X - E\{X\})^r$:

$$\mu_r(X) = E\{(X - E\{X\})^r\} = \int_{-\infty}^{+\infty} (x - E\{X\})^r f(x)dx \tag{2.18}$$

On remarque que le moment centré d'ordre 2, $\mu_2(X)$, est la variance théorique de la variable aléatoire X distribuée selon la loi F. Souvent, les moments centrés sont plus difficiles à calculer que les moments non-centrés. On peut toutefois déduire le moment centré $\mu_r(X)$ d'ordre r à partir des moments non-centrés d'ordre inférieur. En effet, il suffit d'appliquer la relation suivante (Kendall et Stuart, 1987) :

$$\mu_r(X) = \sum_{j=0}^{r} \binom{r}{j} \mu'_r(X) [-\mu'_1(X)]^j \tag{2.19}$$

où $\binom{r}{j} = \frac{r!}{j!(r-j)!}$ est le nombre de combinaisons de j éléments parmi r éléments et où ! est le symbole factoriel.

Pour quelques lois de probabilité, les intégrales qui définissent les moments non-centrés et centrés ne convergent pas (aucune solution). Les moments dans ce cas n'existent qu'à certaines conditions qui dépendent de l'ordre r. On peut montrer (Kendall et Stuart, 1987) que si $\mu'_r(X)$ existe, alors $\mu'_s(X)$ existe pour $s < r$; et que si $\mu'_r(X)$ n'existe pas, alors $\mu'_s(X)$ n'existe pas si $s > r$.

---

[b] Dans ce qui suit, nous confondons en écriture la loi et sa fonction de répartition F.

Exemple 2.4. *Considérons une variable aléatoire X distribuée selon une loi de probabilité dont la f.d.p. est définie par*

$$f(x) = \frac{c}{\alpha}\left(\frac{x}{\alpha}\right)^{c-1} \exp\left[-\left(\frac{x}{\alpha}\right)^{c}\right], \quad 0 < x < +\infty \tag{2.20}$$

où α et c sont positifs.

Cette loi de probabilité continue est appelée ***loi Weibull*** *de paramètres α et c, et est parfois utilisée en hydrologie pour déterminer la probabilité d'occurrence des débits extrêmes (Perreault et al., 1992b).*

Selon l'équation (2.17), le moment non-centré d'ordre r est donné par :

$$\mu'_r(X) = \int_0^{+\infty} y^r \frac{c}{\alpha}\left(\frac{y}{\alpha}\right)^{c-1} \exp\left[-\left(\frac{y}{\alpha}\right)^{c}\right] dy, \quad 0 < x < +\infty \tag{2.21}$$

En évaluant cette intégrale (Perreault et al. 1992b), on déduit que :

$$\mu'_r(X) = \alpha^r \Gamma\left(1+\frac{r}{c}\right) \tag{2.22}$$

où $\Gamma(.)$ désigne la fonction gamma. En particulier, la moyenne d'une variable aléatoire qui suit une loi Weibull est donnée par :

$$E\{X\} = \mu'_1(X) = \alpha\Gamma\left(1+\frac{1}{c}\right) \tag{2.23}$$

L'argument de la fonction gamma doit être positif. Le moment non-centré d'ordre r $\mu'_r(X)$ existe alors si $(1+r/c) > 0$, c'est-à-dire lorsque $r > -c$. Si on utilise la relation (2.19), on peut déterminer les moments centrés. En particulier, le moment centré d'ordre 2 est donné par :

$$Var\{X\} = \mu_2(X) = \alpha^2\left[\Gamma\left(1+\frac{2}{c}\right) - \Gamma^2\left(1+\frac{1}{c}\right)\right] \tag{2.24}$$

Comme on l'a mentionné précédemment, une des principales utilisations des moments est la construction d'indices qui caractérisent la forme de la fonction de densité de probabilité. L'indice le plus utilisé en hydrologie statistique est le coefficient d'asymétrie *Cs* qui est obtenu en divisant le moment centré d'ordre 3 par l'écart-type élevé au cube :

$$Cs = \frac{\mu_3}{\mu_2^{3/2}} = \frac{\mu_3}{\sigma^3} \tag{2.25}$$

La façon dont ce coefficient caractérise la loi de probabilité est illustrée à la Figure 2.6. Le signe de *Cs* indique le type d'asymétrie de la f.d.p.. Si la f.d.p. est symétrique, le coefficient d'asymétrie est nul.

Le coefficient d'aplatissement *Ck*, quoique peu utilisé en hydrologie, donne une idée de la concentration des probabilités dans les extrémités de la f.d.p.. Ce coefficient est généralement utilisé pour distinguer différentes lois symétriques. Le coefficient d'aplatissement est fonction du moment centré d'ordre 4 et est donné par :

$$Ck = \frac{\mu_4}{\mu_2^2} = \frac{\mu_4}{\sigma^4} \tag{2.26}$$

Enfin, il peut être utile de considérer une mesure de dispersion normalisée (adimensionnelle) de la variable aléatoire, le coefficient de variation *Cv*, défini par :

$$Cv = \frac{\mu_2^{1/2}}{\mu_1} = \frac{\sigma}{\mu} \tag{2.27}$$

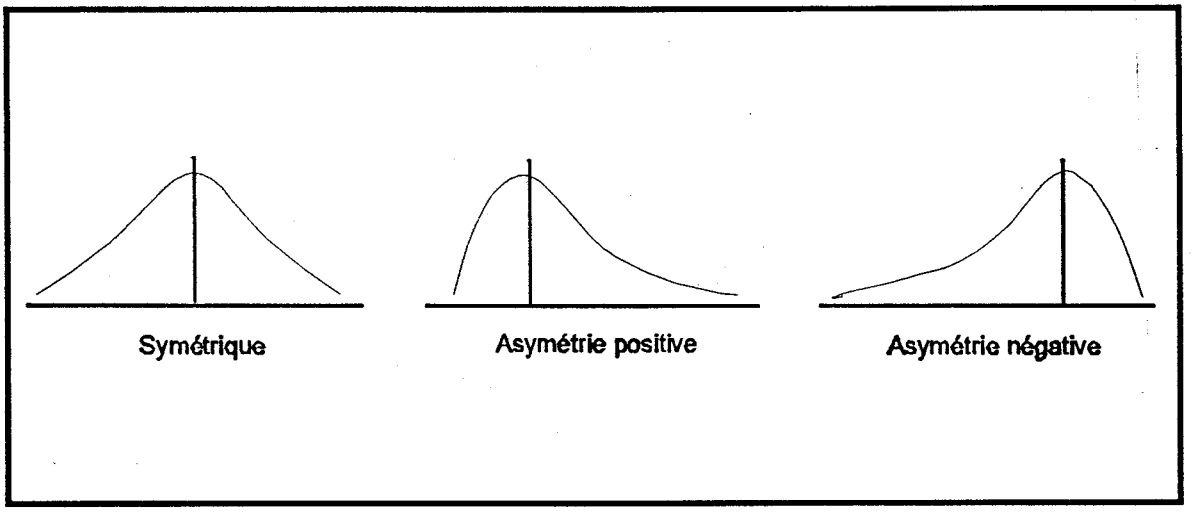


Figure 2.6. L'effet du coefficient d'asymétrie sur la f.d.p..

Exemple 2.5. *Considérons une variable aléatoire X distribuée selon une loi de probabilité dont la f.d.p. est définie par*

$$f(x) = \frac{1}{\alpha} \exp\left\{-\frac{x-m}{\alpha}\right\}, \quad m < x < +\infty \tag{2.29}$$

où α et m sont positifs. Cette loi de probabilité continue est appelée ***loi exponentielle*** *de paramètres α et m. Cette loi a été utilisée en hydrologie pour déterminer la probabilité d'occurrence des volumes d'eau en période d'étiage (Mathier et al., 1991).*

En utilisant l'équation (2.17) et la relation (2.19), on obtient les moments suivants :

$$\begin{aligned} \mu_1'(X) &= m + \alpha \qquad \mu_2(X) = \alpha^2 \\ \mu_3(X) &= 2\alpha^3 \qquad \mu_4(X) = 9\alpha^4 \end{aligned} \tag{2.30}$$

et on en déduit aisément les coefficients de variation, d'asymétrie et d'aplatissement :

$$Cv = \frac{\alpha}{m+\alpha}, \qquad Cs = 2, \qquad Ck = 9 \tag{2.31}$$

On remarque en particulier que la valeur Cs=2 est indicative d'une assez forte asymétrie positive. Cette caractéristique est très apparente lorsqu'on examine la f.d.p. de la loi exponentielle tracée à la Figure 2.7. Cette propriété constitue l'une des principales justifications de son utilisation pour modéliser les volumes d'eau en période d'étiage. En effet, il est raisonnable de croire qu'une forte proportion de faibles volumes soit observée en régime sec. La probabilité correspondant aux faibles volumes représentée par la f.d.p. est donc plus grande pour les valeurs peu élevées du volume.

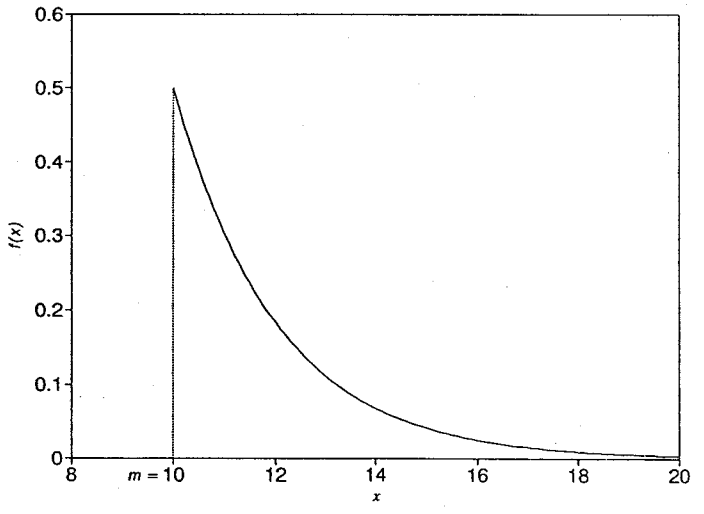


Figure 2.7. Fonction de densité de probabilité de la loi exponentielle.

2.3 Quantile x_T de période de retour T

L'objectif principal de l'analyse hydrologique de fréquence est d'établir la relation existant entre des événements hydrologiques extrêmes (crues, étiages, etc.) et leurs probabilités de dépassement ou de non-dépassement. On entend ici par **événement hydrologique extrême**, une situation qui pourrait engendrer un risque. En hydrologie la façon usuelle d'associer une probabilité à un tel événement est de lui faire correspondre une **période de retour** T. Évidemment, l'événement hydrologique extrême et l'interprétation de la période de retour dépendent du type de risque que l'on désire étudier et de la variable aléatoire considérée.

Lors du dimensionnement d'un ouvrage hydraulique, par exemple, il est important d'évaluer le risque de défaillance (débordement ou inondation, par exemple). On s'intéresse alors particulièrement à la variable aléatoire X "débit maximum annuel" et à l'événement $\{X > x_c\}$ où x_c est un débit maximum critique pour l'ouvrage hydraulique. Si nous considérons le débit maximum annuel X distribué selon une loi de probabilité possédant une f.d.p. $f(x;\underline{\theta})$, où $\underline{\theta} = (\theta_1, \theta_2, \ldots, \theta_p)$ est le vecteur des p paramètres de la loi, la probabilité que l'événement $\{X > x_c\}$ survienne est donnée selon l'équation (2.9) par :

$$p = \text{Prob}\{X > x_c\} = 1 - F(x_c;\underline{\theta}) \tag{2.32}$$

Considérons maintenant les deux événements mutuellement exclusifs suivants :

$$E = \{X > x_c\} \quad \text{et} \quad \overline{E} = \{X \le x_c\} \tag{2.33}$$

où E représente ici un échec (ou une défaillance, par exemple un débordement), et $\overline{E}$ un succès. Ces deux événements définissent une expérience de Bernoulli telle que :

$$\text{Prob}\{E\} = p \quad \text{et} \quad \text{Prob}\{\overline{E}\} = 1 - p \tag{2.34}$$

Si Z est le nombre d'années s'écoulant entre deux événements $E = \{X > x_c\}$, on peut montrer (Benjamin et Cornell, 1970) que Z est une variable aléatoire distribuée selon une loi géométrique dont la f.d.p. est donnée à l'Exemple 2.1 avec $\pi = p$, et que :

$$\mu_1'(Z) = E\{Z\} = \frac{1}{p} = T \tag{2.35}$$

$E\{Z\} = T$ peut être interprétée comme le temps moyen (calculé sur une longue période) entre deux événements $E = \{X > x_c\}$ et T est appelée la **période de retour** de l'événement

$E = \{X > x_c\}$. Si on définit la valeur critique en fonction de la période de retour T en posant $x_T = x_c$, on a, à partir de l'équation (2.32) :

$$\text{Prob}\{X > x_T\} = 1 - F(x_T; \underline{\theta}) = \frac{1}{T} \tag{2.36}$$

La valeur x_T est appelée le **quantile de période de retour T** et est une fonction de T ainsi que des paramètres $\underline{\theta}$ de la loi de probabilité F. Plus précisément, ce quantile peut s'écrire mathématiquement de la façon suivante :

$$x_T = F^{-1}\left(\theta_1, \theta_2, \ldots, \theta_p, 1 - \frac{1}{T}\right) \tag{2.37}$$

où F^{-1} est la fonction inverse de la fonction de répartition F. Le terme mathématique "fonction inverse" est employé ici pour signifier qu'au lieu d'affecter une probabilité à une valeur donnée x, comme le fait la fonction de répartition F, la fonction inverse F^{-1} attribue plutôt une valeur x à une probabilité donnée.

Ainsi, le débit maximum annuel x_T possède une période de retour T si l'événement $E = \{X > x_T\}$ survient en moyenne chaque T années. La réciproque de T, $1/T$, est donc la probabilité au dépassement de cet événement. Par exemple, le débit de période de retour T=100, le débit centennal x_{100}, a une probabilité de 1% d'être dépassé. Il est important de noter que la période de retour ne permet pas de déterminer le moment où surviendra l'événement. La période de retour ne signifie pas, par exemple, que le débit centennal sera observé systématiquement à toutes les cent années, mais bien qu'en moyenne, sur une très longue période, il se réalisera une fois sur cent.

Exemple 2.6. *Soit X une variable aléatoire distribuée selon une loi Gumbel de paramètres α et u. Selon l'Exemple 2.3, la fonction de répartition correspondante est donnée par :*

$$F(x; u, \alpha) = \exp\left[-\exp\left(\frac{x-u}{\alpha}\right)\right], \quad -\infty < x < +\infty \tag{2.38}$$

Le quantile x_T de période de retour T de cette variable aléatoire est tel que (Équation 2.37):

$$1 - \exp\left[-\exp\left(\frac{x_T - u}{\alpha}\right)\right] = \frac{1}{T} \tag{2.39}$$

si l'événement extrême considéré est $E = \{X > x_T\}$. En appliquant la transformée logarithmique deux fois de chaque côté de l'équation et en isolant x_T, on a que :

$$x_T = u - \alpha \ln\left[-\ln\left(1-\frac{1}{T}\right)\right] \tag{2.40}$$

Ainsi, si on suppose que les débits maximums annuels sont distribués selon une loi Gumbel, l'équation (2.40) permet d'évaluer, si les paramètres α et u sont connus, les débits x_T de périodes de retour T.

Si on s'intéresse maintenant à la gestion des ressources en eau lors d'une sécheresse par exemple, on considérerait plutôt la variable aléatoire Y "débit minimum annuel" et l'événement extrême $\{Y \le y_c\}$ où y_c est un débit minimum critique. Si l'on considère le débit minimum annuel Y distribué selon une loi de probabilité possédant une f.d.p. $g(x;\underline{\theta})$, où $\underline{\theta}$ est le vecteur des paramètres de la loi, la probabilité que l'événement $\{Y \le y_c\}$ survienne est alors donnée selon l'équation (2.9) par :

$$1-p = \text{Prob}\{Y \le y_c\} = G(y_c;\underline{\theta}) \tag{2.41}$$

où $G(.)$ est la fonction de répartition de Y. Une période de retour T peut donc être affectée à des événements extrêmes minimums de façon tout à fait analogue à celle des événements extrêmes maximums. Un débit minimum annuel y_T possède alors une période de retour T si l'événement $\overline{E} = \{Y \le y_T\}$ survient en moyenne chaque T années et cet événement, qui dans le cas d'un étiage correspond à une défaillance, est considéré ici comme un échec dans l'expérience de Bernoulli (Équation 2.34). Ainsi, le débit minimum y_T est tel que sa probabilité au **non-dépassement** (1-p) est égale à $1/T$. Plus précisément, ce quantile s'exprime de la manière suivante :

$$y_T = G^{-1}\left(\theta_1, \theta_2, \ldots, \theta_p, \frac{1}{T}\right) \tag{2.42}$$

où G^{-1} est la fonction inverse de la fonction de répartition G.

Exemple 2.7. *Soit Y une variable aléatoire distribuée selon une loi exponentielle de paramètres α et m dont la f.d.p. est donnée à l'Exemple 2.5. Selon les équations (2.9) et (2.29), on peut déduire la fonction de répartition correspondante :*

$$G(x) = 1 - \exp\left[-\left(\frac{y-m}{\alpha}\right)\right], \quad m < y < +\infty \tag{2.43}$$

Si l'on s'intéresse aux débits extrêmes minimums, le quantile y_T de période de retour T de cette variable aléatoire est tel que (Équations 2.41, 2.42) :

$$1-p=1-\exp\left[-\left(\frac{y_T-m}{\alpha}\right)\right]=\frac{1}{T} \tag{2.44}$$

l'événement extrême considéré étant $\overline{E}=\{Y\le y_T\}$. *En appliquant la transformée logarithmique une fois de chaque côté de l'équation et en isolant* y_T, *on obtient que :*

$$y_T=m-\alpha\ln\left(1-\frac{1}{T}\right) \tag{2.45}$$

Ainsi, si on suppose que les débits minimums annuels sont distribués selon une loi exponentielle, l'équation (2.45) nous permet d'évaluer, pour des valeurs connues des paramètres α *et m, le débit d'étiage de période de retour T.*

Si l'on ne peut exprimer sous forme explicite les fonctions de répartition F et G, on doit utiliser une méthode numérique pour déduire les quantiles de période de retour T. L'approche retenue dans ***AJUSTE-II*** est la méthode de Newton-Raphson (Burden et Faires, 1981).

2.4 Lois de probabilité dans le logiciel *AJUSTE-II*

Plusieurs lois de probabilité sont susceptibles de représenter adéquatement les différentes variables hydrologiques (débits de crues, débits d'étiages, précipitation, volume, etc.). Le logiciel ***AJUSTE-II*** comprend 15 lois de probabilité qui donnent un large choix pour modéliser les données hydrologiques. Le Tableau 2.1 donne pour chaque distribution :

- l'expression de la fonction de densité de probabilité $f(x)$;
- l'expression de la fonction de répartition $F(x)$ (s'il existe une forme explicite);
- le domaine $\mathcal{D}$ de la variable aléatoire correspondante;
- l'expression des moments non-centrés $\mu'_r(X)$;

Le problème du choix du modèle de loi de probabilité sera discuté au chapitre 5.

Tableau 2.1. Lois de probabilité incorporées dans le logiciel ***AJUSTE-II***

Loi	Fonction de densité de probabilité	Fonction de répartition	Domaine	Moments non-centrés
gamma	$f(x)=\frac{\alpha^{\lambda}}{\Gamma(\lambda)}x^{\lambda-1}e^{-\alpha x}$	Forme non-explicite	$x>0$	$\mu'_r(X)=\frac{1}{\alpha^r}\frac{\Gamma(\lambda+r)}{\Gamma(\lambda)}$
Pearson Type 3	$f(x)=\frac{\alpha^{\lambda}}{\Gamma(\lambda)}(x-m)^{\lambda-1}e^{-\alpha(x-m)}$	Forme non-explicite	$x>m$	$\mu'_r(X-m)=\frac{1}{\alpha^r}\frac{\Gamma(\lambda+r)}{\Gamma(\lambda)}$
log-Pearson Type 3	$f(x)=\frac{\alpha^{\lambda}}{x\Gamma(\lambda)}(\ln x-m)^{\lambda-1}e^{-\alpha(\ln x-m)}$	Forme non-explicite	$x>e^m$	$\mu'_r(X)=e^{mr}\left(1-\frac{r}{\alpha}\right)^{-\lambda}$
gamma généralisée	$f(x)=\frac{\lvert s\rvert\alpha^{s\lambda}}{\Gamma(\lambda)}x^{s\lambda-1}e^{-(\alpha x)^s}$	Forme non-explicite	$x>0$	$\mu'_r(X)=\frac{1}{\alpha^r}\frac{\Gamma(\lambda+r/s)}{\Gamma(\lambda)}$
gamma inverse	$f(x)=\frac{\alpha^{\lambda}}{\Gamma(\lambda)}\left(\frac{1}{x}\right)^{\lambda+1}e^{-\alpha/x}$	Forme non-explicite	$x>0$	$\mu'_r(X)=\frac{\alpha^r}{\prod_{i=1}^{r}(\lambda-i)}$
GEV	$f(x)=\frac{1}{\alpha}\left[1-\frac{k}{\alpha}(x-u)\right]^{1/k-1}\exp\left\{-\left[1-\frac{k}{\alpha}(x-u)\right]^{1/k}\right\}$	$F(x)=\exp\left\{-\left[1-\frac{k}{\alpha}(x-u)\right]^{1/k}\right\}$	$x>u+\alpha/k,\ si\ k<0$ $x<u+\alpha/k,\ si\ k>0$	$\mu'_r(X)=u+\frac{\alpha}{k}\left[1-\Gamma(1+k)\right]$
Gumbel	$f(x)=\frac{1}{\alpha}\exp\left[-\frac{x-u}{\alpha}-\exp\left(\frac{x-u}{\alpha}\right)\right]$	$F(x)=\exp\left\{-\exp\left[-\frac{x-u}{\alpha}\right]\right\}$	$-\infty<x<+\infty$	Aucune forme générale pour le moment d'ordre r (cf. Perreault *et al.*, 1992a)
Weibull	$f(x)=\frac{c}{\alpha}\left(\frac{x}{\alpha}\right)^{c-1}\exp\left[-\left(\frac{x}{\alpha}\right)^c\right]$	$F(x)=1-\exp\left\{-\left(\frac{x}{\alpha}\right)^c\right\}$	$x>0$	$\mu'_r(X)=\alpha^r\Gamma\left(1+\frac{r}{c}\right)$

Tableau 2.1. Lois de probabilité incorporées dans le logiciel ***AJUSTE-II*** (suite).

Loi	Fonction de densité de probabilité	Fonction de répartition	Domaine	Moments non-centrés
normale	$f(x)=\frac{1}{\sigma\sqrt{2\pi}}\exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$	Forme non-explicite	$-\infty<x<+\infty$	Forme non-explicite pour le moment non-centré. Moment centré : $\mu_r(X)=\sigma^r\frac{2^{r/2}\Gamma\left[\frac{r+1}{2}\right]}{\sqrt{\pi}}$ si r est pair. Si r est impair le moment est nul.
log-normale à 2 paramètres	$f(x)=\frac{1}{x\sigma\sqrt{2\pi}}\exp\left\{-\frac{(\ln x-\mu)^2}{2\sigma^2}\right\}$	Forme non-explicite	$x>0$	$\mu'_r(X)=\exp\left\{r\mu+\frac{1}{2}r^2\sigma^2\right\}$
log-normale à 3 paramètres	$f(x)=\frac{1}{(x-m)\sigma\sqrt{2\pi}}\exp\left\{-\frac{[\ln(x-m)-\mu]^2}{2\sigma^2}\right\}$	Forme non-explicite	$x>m$	$\mu'_r[(X-m)]=\exp\left\{r\mu+\frac{1}{2}r^2\sigma^2\right\}$
exponentielle	$f(x)=\frac{1}{\alpha}\exp\left\{-\frac{x-m}{\alpha}\right\}$	$F(x)=1-\exp\left\{-\frac{x-m}{\alpha}\right\}$	$x>m$	$\mu'_r(X-m)=\alpha^r\Gamma(1+r)$
Halphen Type A	$f(x)=\frac{1}{2m^{\upsilon}K_{\upsilon}(2\alpha)}x^{\upsilon-1}\exp\left[-\alpha\left(\frac{x}{m}+\frac{m}{x}\right)\right]$	Forme non-explicite	$x>0$	$\mu'_r(X)=\frac{m^rK_{\nu+r}(2\alpha)}{K_{\nu}(2\alpha)}$
Halphen Type B	$f(x)=\frac{2}{m^{2\nu}ef_{\nu}(\alpha)}x^{2\nu-1}\exp\left[-\left(\frac{x}{m}\right)^2+\alpha\left(\frac{x}{m}\right)\right]$	Forme non-explicite	$x>0$	$\mu'_r(X)=\frac{m^ref_{\nu+r/2}(\alpha)}{ef_{\nu}(\alpha)}$
Halphen Type B^{-1}	$f(x)=\frac{2m^{2\upsilon}}{ef_{\upsilon}(\alpha)}x^{-2\upsilon-1}\exp\left[-\left(\frac{m}{x}\right)^2+\alpha\left(\frac{m}{x}\right)\right]$	Forme non-explicite	$x>0$	$\mu'_r(X)=\frac{m^ref_{\nu-r/2}(\alpha)}{ef_{\nu}(\alpha)}$

3 ESTIMATION

3.1 Problématique

Si l'on connaît les paramètres qui caractérisent exactement la forme d'une loi de probabilité F donnée, il est possible de générer une série de valeurs numériques $x_1, x_2, x_3, \ldots, x_n$ d'une variable aléatoire distribuée selon cette loi. Une telle série de taille infinie ($n \to +\infty$) constituerait la **population** de toute variable aléatoire provenant d'une loi F avec cet ensemble de paramètres. Connaissant la loi et ses paramètres, on peut alors déduire directement, à l'aide des expressions présentées au chapitre 2, les moments de la population (moments théoriques) et les quantiles x_T ou y_T de période de retour T.

Les données hydrologiques (débits, précipitations, etc.) sont le résultat de plusieurs processus physiques et sont en général sujettes à toutes sortes d'erreurs. De plus, les séries de données hydrologiques disponibles sont de taille n finie. En pratique, il est alors difficile, et souvent impossible, d'identifier exactement la forme de la loi F de ces observations, c'est-à-dire l'ensemble des paramètres qui caractérisent la distribution des probabilités d'occurrence de la variable aléatoire. Ainsi, les moments théoriques et les quantiles d'une variable aléatoire hydrologique demeurent inconnus. Cependant, une estimation des paramètres d'une loi F donnée peut être obtenue à l'aide d'une série finie de réalisations $x_1, x_2, x_3, \ldots, x_n$ de la variable aléatoire considérée qui constitue un **échantillon**. Une estimation des quantiles peut alors en être déduite. Afin de pouvoir ajuster une loi à cet échantillon (estimation des paramètres), on doit au préalable vérifier certaines conditions qui seront présentées au chapitre 5.

Le problème traité dans le présent chapitre se résume de la façon suivante. Considérons un processus hydrologique (par exemple, le débit de crue d'une rivière donnée) représenté par la variable aléatoire X distribuée selon la loi de probabilité $F(x;\underline{\theta})$ dont le vecteur de paramètres $\underline{\theta} = (\theta_1, \theta_2, \ldots, \theta_p)$ est inconnu. Supposons que ce processus s'est répété dans les mêmes conditions n fois de façon indépendante pour engendrer les variables aléatoires $X_1, X_2, X_3, \ldots, X_n$. Les variables aléatoires $X_1, X_2, X_3, \ldots, X_n$ sont alors indépendantes et identiquement distribuées selon la loi de probabilité $F(x;\underline{\theta})$ et forment l'**échantillon aléatoire** (par exemple, les débits maximums annuels des n dernières années). L'objectif des méthodes d'estimation est de définir p fonctions de l'échantillon aléatoire, $\hat{\theta}_1(X_1, X_2, \ldots, X_n), \ldots, \hat{\theta}_p(X_1, X_2, \ldots, X_n)$, telles que si $x_1, x_2, x_3, \ldots, x_n$ sont les

réalisations des n variables aléatoires (données numériques obtenues), alors $\hat{\theta}_1(x_1, x_2, \ldots, x_n), \ldots, \hat{\theta}_p(x_1, x_2, \ldots, x_n)$ sont de bonnes estimations de $\underline{\theta} = (\theta_1, \theta_2, \ldots, \theta_p)$.

Les paramètres ou toute autre caractéristique d'une loi de probabilité peuvent être estimés par différents estimateurs. Un problème important et souvent difficile à résoudre consiste à déterminer quel estimateur est préférable. C'est pourquoi, en théorie de l'estimation, on définit les propriétés que doit posséder un "bon" estimateur. Nous en présentons brièvement ici une synthèse tirée de Bickel et Doksum (1977).

1. Estimateur non-biaisé

*Théoriquement, un estimateur $\hat{\theta}$ d'une quantité θ est **non-biaisé** si sa moyenne théorique est égale à θ, c'est-à-dire si :*

$$E\{\hat{\theta}\} = \theta$$

Cette propriété est bien sûr souhaitable puisqu'elle signifie qu'en moyenne l'estimateur $\hat{\theta}$ conduit à une valeur très proche de la quantité théorique inconnue θ.

2. Estimateur convergent

*Un estimateur $\hat{\theta}$ d'une quantité θ, obtenu à partir d'un échantillon aléatoire de taille n, est **convergent** lorsque celui-ci tend en probabilité vers la valeur théorique θ, c'est-à-dire que pour une valeur ε aussi petite que l'on veut, on a que :*

$$\text{Prob}\{|\hat{\theta} - \theta| > \varepsilon\} = 0 \quad \text{lorsque} \quad n \to +\infty$$

Cette propriété signifie qu'à mesure que la taille d'échantillon augmente l'estimateur $\hat{\theta}$ fournit des valeurs qui se rapprochent de plus en plus de la quantité théorique θ.

3. Estimateur efficace

*Cette propriété traduit le fait que l'estimateur doit avoir une variance faible. Lorsqu'il existe plusieurs estimateurs de θ, celui ayant la variance minimum est préféré, sur la base de ce critère, car sa distribution autour de θ est moins dispersée. Si de plus l'estimateur $\hat{\theta}$ est non-biaisé, sa distribution est centrée en θ. Lorsqu'il existe une valeur limite minimum de la variance (la borne de Cramer-Rao), l'estimateur conduisant à cette limite est un estimateur **efficace** (Bickel et Doksum, 1977). Généralement, la variance minimum est atteinte lorsque la taille d'échantillon tend vers l'infini et dans ce cas l'estimateur est **asymptotiquement efficace.***

Dans ce qui suit, nous décrivons les méthodes d'estimation des paramètres que l'on retrouve dans le logiciel ***AJUSTE-II***. L'application de ces méthodes suppose que l'on connaisse la loi F d'où proviennent les observations. Cependant, nous présentons, tout d'abord, le calcul des moments et des coefficients de l'échantillon. Ceux-ci sont des estimations des moments et coefficients théoriques de la loi dont on suppose que cet échantillon provient.

3.2 Estimation des moments et des coefficients

Considérons n variables aléatoires $X_1, X_2, X_3, \ldots, X_n$ indépendantes que l'on suppose identiquement distribuées selon une loi F **quelconque**. Soient $x_1, x_2, x_3, \ldots, x_n$ les réalisations de ces variables aléatoires constituant l'échantillon observé (par exemple, les valeurs numériques de débits maximums annuels des n dernières années). Le **moment non-centré d'ordre r de l'échantillon** est défini par:

$$m'_r = \frac{1}{n}\sum_{i=1}^{n} x_i^r \tag{3.1}$$

et est une estimation du moment non-centré théorique d'ordre r $\mu'_r(X)$. Pour $r = 1$, l'expression (3.1) correspond à la moyenne arithmétique que l'on note $\overline{x}$. De la même façon, **le moment centré d'ordre r de l'échantillon**, s'exprime comme suit :

$$m_r = \frac{1}{n}\sum_{i=1}^{n} (x_i - \overline{x})^r \tag{3.2}$$

et est une estimation du moment non-centré théorique d'ordre r $\mu_r(X)$. Pour $r = 2$, l'expression (3.2) correspond à la variance de l'échantillon que l'on note $\hat{\sigma}^2$.

Enfin, par analogie, on déduit les estimations des coefficients théoriques *Cv*, *Cs* et *Ck* introduits à la section 2.2 :

$$\hat{C}v = \frac{m_2^{1/2}}{m_1} = \frac{\hat{\sigma}}{\overline{x}}, \quad \hat{C}s = \frac{m_3}{m_2^{3/2}} = \frac{m_3}{\hat{\sigma}^3}, \quad \hat{C}k = \frac{m_4}{\hat{\sigma}^4} \tag{3.3}$$

On peut montrer, par exemple, que $E\{\overline{X}\} = \mu$ et donc que la moyenne arithmétique est un estimateur non-biaisé de la moyenne théorique (Kendall et Stuart, 1987). Toutefois, les estimateurs des moments centrés d'ordre élevé sont généralement biaisés. On peut dans certain cas corriger le biais. Ainsi, pour la variance de l'échantillon, un estimateur non-biaisé est donné par :

$$S^2 = \frac{n}{n-1}\hat{\sigma}^2 \tag{3.4}$$

où le dénominateur n - 1, au lieu de la valeur n, élimine le biais (Kendall et Stuart, 1987). Il est clair que pour des échantillons de très grande taille S^2 et $\hat{\sigma}^2$ donneront des valeurs semblables.

Le biais du coefficient d'asymétrie de l'échantillon est plus problématique puisque dans $\hat{C}s$ intervient la somme des écarts à la moyenne élevés au cube qui peuvent introduire de larges erreurs (Kirby, 1974). Plusieurs approximations d'un estimateur non-biaisé ont été proposées dans la littérature. En particulier :

$$\hat{C}_{s1} = \frac{\sqrt{n(n-1)}}{(n-2)}\hat{C}_s \tag{3.5}$$

$$\hat{C}_{s2} = \left(1+\frac{8.5}{n}\right)\hat{C}_{s1} \tag{3.6}$$

$$\hat{C}_{s3} = \left[\left(1+\frac{6.51}{n}+\frac{20.20}{n^2}\right)+\left(\frac{1.48}{n}+\frac{6.77}{n^2}\right)\hat{C}_s^2\right]\hat{C}_s \tag{3.7}$$

L'utilisation de $\hat{C}_{s1}$ a été recommandée le "United States Water Resources Council" (WRC, 1967), la correction $\hat{C}_{s2}$ été proposée par Hazen (1924), et C_{s3} par Bobée et Robitaille (1975). Ces estimateurs du coefficient d'asymétrie sont utilisés généralement en hydrologie lorsqu'on considère les lois à trois paramètres Pearson Type 3 et log-Pearson Type 3. Dans le logiciel AJUSTE-2, ces trois corrections peuvent être utilisées lors de l'estimation des paramètres de ces lois par la méthode des moments (Section 3.3.2). Notons qu'une étude menée par Wallis *et al.* (1985) a montré que dans la majorité des cas, le coefficient d'asymétrie corrigé $\hat{C}_{s2}$ conduit à de meilleurs résultats.

3.3 Estimation des paramètres d'une loi de probabilité

Il existe plusieurs méthodes pour estimer les paramètres d'une loi de probabilité donnée. Nous présentons, tout d'abord, les méthodes bien connues du maximum de vraisemblance et des moments. Par la suite, nous présenterons brièvement la méthode indirecte des moments, la méthode des moments mixtes, la méthode des moyennes ainsi que la méthode des moments pondérés.

3.3.1 Méthode du maximum de vraisemblance (MXV)

Considérons n variables aléatoires $X_1, X_2, X_3, \ldots, X_n$ indépendantes que l'on suppose identiquement distribuées selon une loi $F(x;\underline{\theta})$ **donnée** de f.d.p. $f(x;\underline{\theta})$. Soient $x_1, x_2, x_3, \ldots, x_n$ les réalisations de ces variables aléatoires qui forment l'échantillon observé. La probabilité que les événements $\{X_1 = x_1\}, \{X_2 = x_2\}, \ldots, \{X_n = x_n\}$ se réalisent simultanément est donnée par la fonction de densité de probabilité jointe. Puisque que l'on suppose que ces variables sont indépendantes et identiquement distribuées, la f.d.p. jointe est définie comme le produit des f.d.p. de chaque variable aléatoire $X_1, X_2, X_3, \ldots, X_n$ et s'exprime de la façon suivante :

$$f(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} f(x_i;\underline{\theta}) \tag{3.8}$$

Cette probabilité peut être considérée comme une fonction des paramètres $\underline{\theta} = (\theta_1, \theta_2, \ldots, \theta_p)$. Dans ce cas, elle est notée $L(\theta_1, \theta_2, \ldots, \theta_p)$ et est appelée **fonction de vraisemblance**. On peut se demander quelles valeurs de $\theta_1, \theta_2, \ldots, \theta_p$ maximiseraient la probabilité $L(\theta_1, \theta_2, \ldots, \theta_p)$ d'obtenir particulièrement cet échantillon observé $x_1, x_2, x_3, \ldots, x_n$. Ces valeurs maximisant $L(\theta_1, \theta_2, \ldots, \theta_p)$ seraient sans doute de bons estimateurs des paramètres puisqu'elles correspondraient à la plus grande probabilité d'obtenir cet échantillon. La méthode du maximum de vraisemblance s'appuie sur cette idée intuitive et le principe peut se formuler de la façon suivante :

> Supposons que pour l'échantillon observé $x_1, x_2, x_3, \ldots, x_n$ on peut déterminer p fonctions de ces observations, $\hat{\theta}_1(x_1, x_2, \ldots, x_n), \ldots, \hat{\theta}_p(x_1, x_2, \ldots, x_n)$, telles que lorsque chacun des paramètres $\theta_1, \theta_2, \ldots, \theta_p$ est remplacé par sa fonction correspondante, la fonction L soit maximum. Alors, les statistiques $\hat{\theta}_1(x_1, x_2, \ldots, x_n), \ldots, \hat{\theta}_p(x_1, x_2, \ldots, x_n)$ sont les **estimateurs du maximum de vraisemblance** de $\theta_1, \theta_2, \ldots, \theta_p$ respectivement.

En pratique, les estimateurs du maximum de vraisemblance sont obtenus par différentiation puisque la fonction de vraisemblance atteint son maximum lorsque toutes les dérivées partielles par rapport à chacun des paramètres sont nulles. Ainsi, pour déterminer les estimateurs du maximum de vraisemblance, il suffit de solutionner le système à p équations et p inconnus suivant :

$$\frac{\partial L(\theta_1, \theta_2, \ldots, \theta_p)}{\partial \theta_j} = 0, \qquad j = 1, 2, \ldots, p \tag{3.9}$$

En raison de la forme exponentielle de nombreuses lois de probabilité utilisées en hydrologie, il est souvent plus simple de maximiser le logarithme naturel de la fonction de vraisemblance $\ln L(\theta_1, \theta_2, \ldots, \theta_p)$ plutôt que la vraisemblance elle-même. Une ou l'autre des deux fonctions conduit au même maximum car :

$$\frac{\partial \ln L}{\partial \theta} = \frac{1}{L}\frac{\partial L}{\partial \theta} \tag{3.10}$$

Exemple 3.1. *Soit un échantillon aléatoire $X_1, X_2, X_3, \ldots, X_n$ provenant d'une loi normale de paramètres μ et σ^2 dont la f.d.p. est donnée par*

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}, \quad -\infty < x < +\infty \tag{3.11}$$

La fonction de vraisemblance s'exprime, en utilisant l'équation (3.8), de la façon suivante :

$$L(\mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2\right\} \tag{3.12}$$

En appliquant la transformation logarithmique, on obtient :

$$\ln L(\mu, \sigma^2) = -\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2 - \frac{n}{2}\ln(2\pi\sigma^2) \tag{3.13}$$

On en déduit ensuite les dérivées partielles par rapport à chacun des 2 paramètres μ et σ^2 :

$$\frac{\partial \ln L(\mu, \sigma^2)}{\partial \mu} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i - \mu), \qquad \frac{\partial \ln L(\mu, \sigma^2)}{\partial \sigma^2} = \frac{1}{2\sigma^4}\sum_{i=1}^{n}(x_i - \mu)^2 - \frac{n}{2\sigma^2} \tag{3.14}$$

La résolution du système obtenu en égalant ces deux équations à zéro donne, pour μ et σ^2 respectivement, la solution du maximum de vraisemblance suivante :

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i, \qquad \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{\mu})^2 \tag{3.15}$$

On peut montrer que les valeurs $\hat{\mu}$ et $\hat{\sigma}^2$ maximisent la fonction $L(\mu, \sigma^2)$. Ainsi, $\hat{\mu}$ et $\hat{\sigma}^2$ sont les estimateurs du maximum de vraisemblance des paramètres μ et σ^2 de la loi normale.

Pour plusieurs lois de probabilité incluses dans le logiciel ***AJUSTE-II***, le système d'équations (3.9) n'admet pas de solution explicite et il faut le résoudre numériquement. On utilise alors une méthode de type Newton-Raphson qui permet de résoudre de façon itérative de tels systèmes d'équations non-linéaires.

De manière générale, si les estimateurs issus de la méthode du maximum de vraisemblance existent, ils possèdent, tout au moins pour de grands échantillons ($n \to +\infty$), les trois propriétés d'optimalité présentées à la Section 3.1. Ils sont convergents, asymptotiquement non-biaisés et asymptotiquement efficaces. Cependant, dans le cas de lois de probabilité possédant un paramètre d'origine, c'est-à-dire lorsque le domaine de la variable X dépend d'un des paramètres (lois Pearson Type 3, log-normale à 3 paramètres, par exemple), la méthode du maximum de vraisemblance peut conduire à des résultats erratiques et certaines des propriétés d'optimalité ne tiennent plus.

3.3.2 Méthode des moments (MOM)

Lorsqu'on considère une variable aléatoire X distribuée selon une loi de probabilité $F(x; \underline{\theta})$ donnée qui dépend de p paramètres $\underline{\theta} = (\theta_1, \theta_2, \ldots, \theta_p)$, il est possible de déterminer les moments non-centrés $\mu'_r(X)$, les moments centrés $\mu_r(X)$ et les coefficients (*Cv*, *Cs*, *Ck*) théoriques de la variable aléatoire à partir des expressions (2.17), (2.19), (2.25), (2.26) et (2.27). Ces caractéristiques de la loi de probabilité sont des fonctions des p paramètres comme on a pu le montrer au chapitre 2.

Pour un échantillon aléatoire $X_1, X_2, X_3, \ldots, X_n$ de taille n provenant d'une loi quelconque, on peut, à partir d'un échantillon observé $x_1, x_2, x_3, \ldots, x_n$, déterminer les moments m'_r et m_r, ainsi que les divers coefficients de l'échantillon (Section 3.2). Si l'échantillon considéré est représentatif d'une loi de probabilité particulière, les moments de l'échantillon sont de bonnes approximations des grandeurs correspondantes de la population (moments théoriques). La méthode des moments, qui permet d'estimer les paramètres de la loi, s'appuie sur cette correspondance.

Généralement, la méthode des moments consiste à solutionner le système d'équations formé en égalant les moments non-centrés théoriques $\mu'_r(X)$ à ceux de l'échantillon m'_r. Pour obtenir une solution unique, on doit disposer d'autant d'équations indépendantes que de paramètres à estimer. Si théoriquement on peut choisir n'importe quel ensemble de moments différents pour construire le système d'équations à résoudre, en pratique, et surtout lorsque la taille de l'échantillon est petite, on vise à utiliser les moments d'ordre r petit. En effet, tel que cela a été mentionné à la section 3.2, les estimateurs des moments d'ordres élevés sont généralement biaisés et ont une variance plus grande.

Ainsi, si on considère n variables aléatoires $X_1, X_2, X_3, \ldots, X_n$ indépendantes et identiquement distribuées selon une loi $F(x;\underline{\theta})$ **donnée** dépendant de p paramètres $\underline{\theta} = (\theta_1, \theta_2, \ldots, \theta_p)$, et $x_1, x_2, x_3, \ldots, x_n$ les réalisations correspondantes, les solutions $\tilde{\theta}_1, \tilde{\theta}_2, \ldots, \tilde{\theta}_p$ du système à p inconnus et p équations suivant :

$$\mu'_j(X) = m'_j, \qquad j = 1, 2, \ldots, p \tag{3.16}$$

sont des estimateurs des paramètres $\theta_1, \theta_2, \ldots, \theta_p$ obtenus par la méthode des moments.

Exemple 3.2. *Soit un échantillon aléatoire $X_1, X_2, X_3, \ldots, X_n$ provenant d'une loi gamma de paramètres α et λ dont la f.d.p. est donnée par*

$$f(x) = \frac{\alpha^\lambda}{\Gamma(\lambda)} x^{\lambda-1} e^{-\alpha x}, \quad 0 < x < +\infty \tag{3.17}$$

Les moments non-centrés théoriques sont donnés par (voir Tableau 2.1) :

$$\mu'_r(X) = \frac{1}{\alpha^r} \frac{\Gamma(\lambda + r)}{\Gamma(\lambda)} \tag{3.18}$$

Puisque cette loi possède deux paramètres, les deux premiers moments non-centrés suffisent pour déterminer les estimateurs. On déduit de l'équation (3.18) et du fait que $\Gamma(x) = (x-1)!$, que ces moments s'expriment respectivement comme suit :

$$\mu'_1(X) = \frac{\lambda}{\alpha}, \qquad \mu'_2(X) = \frac{\lambda(\lambda+1)}{\alpha^2} \tag{3.19}$$

Si on égale chacun des 2 moments théoriques précédents à son équivalent échantillonal m'_r, et qu'on résout le système d'équations, on obtient comme solutions pour α et λ respectivement :

$$\tilde{\alpha} = \frac{m_1'}{m_2' - m_1'^2} = \frac{\bar{x}}{\hat{\sigma}^2}, \qquad \tilde{\lambda} = \frac{m_1'^2}{m_2' - m_1'^2} = \frac{\bar{x}^2}{\hat{\sigma}^2} \tag{3.20}$$

Ainsi, $\tilde{\alpha}$ et $\tilde{\lambda}$ sont les estimateurs des paramètres α et λ de la loi gamma obtenus par la méthode des moments.

De la même manière que pour les moments non-centrés, des estimateurs équivalents de la méthode des moments peuvent être obtenus en utilisant les moments centrés ou les coefficients *Cv*, *Cs* et *Ck*, à condition que les équations demeurent indépendantes. En particulier, dans le logiciel ***AJUSTE-II***, on considère le système d'équations formé à partir de la moyenne $\mu_1'(X)$, de la variance $\mu_2(X)$ et du coefficient d'asymétrie *Cs* pour les lois à 3 paramètres. Dans ce cas, on utilise les estimations corrigées S^2 et $\hat{C}_{S1}$, $\hat{C}_{S2}$ ou $\hat{C}_{S3}$ présentées à la Section 3.2 pour diminuer le biais.

La méthode des moments a l'avantage d'être très facile d'utilisation et permet d'obtenir plus souvent une solution explicite comparativement à la méthode du maximum de vraisemblance. La loi gamma est un exemple où les estimateurs du maximum de vraisemblance doivent être déterminés par une méthode numérique, alors que la méthode des moments donne une solution explicite (Exemple 3.2). Toutefois, même si les estimateurs issus de la méthode des moments sont convergents, ils sont biaisés et généralement non-efficaces.

3.3.3 Autres méthodes d'estimation (MM1, WRC, SAM et MMP)

Les méthodes du maximum de vraisemblance et des moments ont été étudiées par nombre de statisticiens à la fois théoriquement et empiriquement. Leurs propriétés sont très bien connues. En particulier, lorsque la taille d'échantillon est grande, la méthode du maximum de vraisemblance est optimale pour les lois ne possédant pas de paramètre d'origine. La méthode des moments, pour sa part, est simple et fournit presque toujours une solution.

En hydrologie, les séries de données disponibles sont souvent de taille réduite. Ainsi, dans certains cas extrêmes, c'est-à-dire lorsque la taille d'échantillon est très petite, les propriétés enviables des méthodes du maximum de vraisemblance et des moments ne peuvent s'appliquer. C'est pourquoi, certains hydrologues ont proposé de nouvelles méthodes d'estimation, et ont montré empiriquement que dans certains cas particuliers celles-ci peuvent fournir plus souvent une solution, et des estimateurs plus efficaces que ceux des deux méthodes classiques. Il est important de souligner que ces méthodes ont été très peu

étudiées théoriquement. Ainsi, on ne peut généraliser leurs propriétés comme il est possible de le faire avec la méthode du maximum de vraisemblance : **ces approches donnent de meilleurs résultats pour certaines lois, certaines tailles d'échantillon et certains ensembles de paramètres**.

Quatre méthodes d'estimation des paramètres, outre les deux méthodes classiques, ont été incorporées dans *AJUSTE-II* pour la loi log-Pearson Type 3, la loi gamma généralisée et la loi généralisée des valeurs extrêmes. Les méthodes MM1 et SAM ont été ajoutées au logiciel à la suite d'une étude comparative effectuée pour déterminer les méthodes d'estimation des paramètres les plus adéquates pour les lois de la famille gamma (Messaoudi, 1994). La méthode WRC a surtout été incorporée parce qu'elle est devenue une norme aux États-Unis lors de l'utilisation de la loi log-Pearson Type 3. Enfin, la méthode MMP a été ajoutée pour la loi généralisée des valeurs extrême puisqu'elle est souvent utilisée en hydrologie. Pour développer les méthodes MM1, SAM et WRC, les hydrologues se sont inspirés de la méthode des moments et du fait que les moments d'ordres inférieurs de l'échantillon sont moins biaisés et moins variables. La méthode MMP est quelque peu différente puisqu'elle considère un autre type de moments, les moments pondérés. Nous les présentons brièvement dans ce qui suit.

1. Méthode indirecte des moments (WRC)

Cette méthode a été proposée par le "United States Water Resources Council" (WRC, 1967) qui à cette époque a recommandé l'emploi systématique de la loi log-Pearson Type 3 pour modéliser les débits maximums annuels. Adaptée à la loi log-Pearson Type 3, cette méthode s'appuie sur la relation qui existe entre cette loi et la loi Pearson Type 3. En effet, si une variable aléatoire X suit une loi log-Pearson Type 3, alors la variable transformée $\ln X$ est distribuée selon une loi Pearson Type 3.

La méthode WRC consiste tout simplement à résoudre le système d'équations de la méthode classique des moments (Section 3.3.2) appliquée à la série des logarithmes des valeurs observées qui suivent une loi Pearson Type 3. Ainsi, on égale les moments théoriques $\mu'_r(X)$ de la loi Pearson Type 3 et les moments empiriques m'_r calculés à partir des données transformées. Cette méthode est expliquée plus en détail dans Bobée et Ashkar (1991, chapitre 7).

Cette approche a été critiquée (Bobée, 1975) parce qu'elle donne un poids égal aux logarithmes des observations et non aux observations elles-mêmes. Ceci peut réduire considérablement l'importance relative des observations extrêmes d'un échantillon donné.

2. Méthode mixte des moments (MM1)

L'emploi de la méthode classique des moments pour une loi à 3 paramètres nécessite l'utilisation du moment d'ordre 3 ou, de façon équivalente, du coefficient d'asymétrie. Plusieurs études ont montré que les estimations du moment d'ordre 3 et du coefficient d'asymétrie sont biaisées, très variables et grandement influencées par les valeurs extrêmes de l'échantillon. Une façon de corriger ce problème est d'utiliser la méthode mixte des moments MM1. Cette méthode modifie la méthode classique des moments en remplaçant dans le système d'équation le moment d'ordre 3 par un moment d'ordre inférieur, en l'occurrence le moment d'ordre 1 du logarithme de la variable aléatoire. La méthode MM1 combine alors les deux premières équations de la méthode classique des moments (moyenne et variance de la série observée) et la première équation de la méthode WRC (moyenne des valeurs transformées en logarithme).

Il est intéressant de noter que le moment d'ordre 1 du logarithme de la variable X est en fait le logarithme de la moyenne géométrique de la variable aléatoire X, une mesure de tendance centrale au même titre que la moyenne arithmétique. Il est appelé le moment non-centré d'ordre "quasi-zéro" puisqu'il correspond au moment non-centré d'ordre r, lorsque r tend vers zéro. La méthode MM1 ainsi que la notion de moment d'ordre "quasi-zéro" sont expliquées en détail dans Bobée et Ashkar (1991, chapitre 7).

3. Méthode des moyennes (SAM)

La méthode SAM, proposée par Bobée (1988), consiste à remplacer par un moment d'ordre inférieur dans le système d'équations de la méthode classique des moments, non seulement le moment d'ordre 3, mais aussi le moment d'ordre 2. En effet, les estimateurs issus de cette méthode sont les solutions du système formé des équations qui font correspondre les moments non-centrés théoriques et échantillonnaux d'ordre -1, "quasi-zéro" et 1. Le moment d'ordre -1 de la variable X aléatoire est appelée la moyenne harmonique et est aussi une mesure de tendance centrale. On retrouve une description de la méthode dans Bobée (1988).

4. Méthode des moments pondérés (MMP)

La méthode MMP a été introduite sous sa forme la plus générale par Greenwood et al. (1979). Cette méthode a été développée principalement pour les lois dont la fonction de répartition F peut s'exprimée explicitement (par exemple, les lois Gumbel, Weibull, exponentielle et généralisée des valeurs extrêmes, cf. Tableau 2.1). Elle a été, entre autres, fortement recommandée par Hosking *et al.* (1985) pour l'ajustement de la loi généralisée des valeurs extrêmes (GEV).

Considérons n variables aléatoires $X_1, X_2, X_3, \ldots, X_n$ indépendantes et identiquement distribuées selon une loi $F(x;\underline{\theta})$ donnée dépendant de p paramètres $\underline{\theta} = (\theta_1, \theta_2, \ldots, \theta_p)$, et $x_1, x_2, x_3, \ldots, x_n$ les réalisations correspondantes. Les moments pondérés théoriques d'ordre r utilisés en pratique sont définis de la façon suivante :

$$\beta_r(X) = E\{X F^r\}, \qquad r = 0, 1, 2, \ldots \tag{3.21}$$

où E désigne l'espérance mathématique (cf. Chapitre 2). Les moments pondérés d'ordre r de l'échantillon sont donnés par :

$$b_r = \frac{1}{n}\sum_{i=1}^{n} \frac{(i-1)(i-2)\cdots(i-r)}{(n-1)(n-2)\cdots(n-r)} x_{(i)} \qquad r = 0, 1, 2, \ldots \tag{3.22}$$

où $x_{(i)}$ est la $i$ème statistique d'ordre de l'échantillon classé en ordre croissant $x_{(1)} \le x_{(2)} \le x_{(3)} \le \ldots \le x_{(n)}$. La statistique b_r est un estimateur non-biaisé du moment pondéré théorique $\beta_r(X)$.

De façon analogue à la méthode classique des moments (cf. Section 3.3.2), les estimateurs déduits de la méthode des moments pondérés $\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_p$ sont obtenus en résolvant le système à p inconnus et p équations suivant :

$$\beta_j(X) = b_j, \qquad j = 1, 2, \ldots, p \tag{3.23}$$

Le Tableau 3.1 donne les méthodes d'estimation retenues pour chaque loi de probabilité incluse dans le logiciel AJUSTE-II. On y retrouve aussi les références des principaux documents où sont présentés les détails théoriques correspondant à l'estimation des paramètres de ces lois.

Tableau 3.1. Lois et méthodes d'estimation du logiciel ***AJUSTE-II***.

Lois	Méthodes	Références
gamma	**MXV, MOM**	Bobée et Ashkar (1991), Perreault *et al.* (1992c)
Pearson Type 3	**MXV, MOM**	Bobée et Ashkar (1991), Perreault *et al.* (1992c)
log-Pearson Type 3	**MM, SAM, WRC**	Bobée et Ashkar (1991), Perreault *et al.* (1992c)
gamma généralisée	**MOM, MM1**	Bobée et Ashkar (1991), Perreault *et al.* (1992c)
gamma inverse	**MXV**	Kotz et Johnson (1983)
GEV	**MXV, MOM, MMP**	Perreault *et al.* (1992a)
Gumbel	**MXV, MOM**	Perreault *et al.* (1992a)
Weibull	**MXV, MOM**	Perreault *et al.* (1992b)
normale	**MXV**	Perreault *et al.* (1992d)
log-normale à 2 paramètres	**MXV**	Aitchison et Brown (1957)
log-normale à 3 paramètres	**MXV, MOM**	Aitchison et Brown (1957)
exponentielle	**MXV**	Lehmann (1983)
Halphen Type A	**MXV**	Perreault *et al.* (1994)
Halphen Type B	**MXV**	Perreault *et al.* (1994)
Halphen Type B^{-1}	**MXV**	Perreault *et al.* (1994)

3.4 Estimation du quantile x_T de période de retour T

On a vu, à la Section 2.3, que le quantile théorique x_T de période de retour T d'une loi $F(x;\underline{\theta})$ est déterminé à partir de la fonction de répartition et qu'il peut s'exprimer comme une fonction de T et des paramètres $\underline{\theta}=(\theta_1, \theta_2, \ldots, \theta_p)$ de la loi. Si on s'intéresse à l'événement extrême $\{X > x_T\}$ (par exemple, au risque associé aux débits de crue), on a :

$$x_T = F^{-1}\left(\theta_1, \theta_2, \ldots, \theta_p, 1-\frac{1}{T}\right) \tag{3.24}$$

Considérons maintenant n variables aléatoires $X_1, X_2, X_3, \ldots, X_n$ indépendantes que l'on suppose identiquement distribuées selon une loi $F(x;\underline{\theta})$ **donnée** de f.d.p. $f(x;\underline{\theta})$. Soient $\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_p$ les estimateurs des paramètres $\theta_1, \theta_2, \ldots, \theta_p$ de la loi obtenus par une méthode quelconque. Un estimateur $\hat{x}_T$ du quantile théorique x_T de période de retour T de la loi $F(x;\underline{\theta})$ est alors déduit en remplaçant, dans l'équation (3.24), les paramètres $\theta_1, \theta_2, \ldots, \theta_p$ par leurs estimations correspondantes $\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_p$:

$$\hat{x}_T = F^{-1}\left(\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_p, 1-\frac{1}{T}\right) \tag{3.25}$$

Selon la méthode utilisée pour estimer les paramètres de la loi $\theta_1, \theta_2, \ldots, \theta_p$, $\hat{x}_T$ sera soit un estimateur du maximum de vraisemblance, soit un estimateur issu de la méthode des moments, etc..

Évidement, cette technique de substitution s'applique aussi lorsqu'on est intéressé à l'événement extrême $\{Y \le y_T\}$, par exemple au risque associé aux périodes sèches. Ainsi, en utilisant l'équation (2.42), on a que :

$$\hat{y}_T = G^{-1}\left(\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_p, \frac{1}{T}\right) \tag{3.26}$$

Exemple 3.3. *Considérons le problème de l'estimation des débits minimums annuels y_T de période de retour T par la méthode du maximum de vraisemblance. On s'intéresse donc à l'événement extrême $\{Y \le y_T\}$. Soit un échantillon aléatoire $Y_1, Y_2, Y_3, \ldots, Y_n$ de débits minimums annuels que l'on suppose indépendants et identiquement distribués selon une loi exponentielle de paramètres α et m dont la f.d.p. est donnée par (Exemple 2.5)* :

$$g(x) = \frac{1}{\alpha}\exp\left\{-\frac{y-m}{\alpha}\right\}, \quad m < y < +\infty \tag{3.27}$$

Si on applique la méthodologie présentée à la Section 3.3.1, on obtient les estimateurs du maximum de vraisemblance des paramètres α et m (Lehmann, 1983 et NERC, 1975) :

$$\hat{m} = y_{(1)} - \frac{1}{n(n-1)}\sum_{i=1}^{n}\left(y_i - y_{(1)}\right), \qquad \hat{\alpha} = \frac{1}{(n-1)}\sum_{i=1}^{n}\left(y_i - y_{(1)}\right) \tag{3.28}$$

où $y_{(1)}$ désigne la plus petite valeur observée de l'échantillon. On a vu à l'Exemple 2.7 que le quantile y_T de période de retour T de la loi exponentielle, lorsqu'on s'intéresse à l'événement extrême $\{Y \le y_T\}$, s'exprime théoriquement de la façon suivante :

$$y_T = m - \alpha \ln\left(1 - \frac{1}{T}\right) \tag{3.29}$$

On déduit alors, à l'aide des équations (3.28) et (3.29) et après quelques manipulations algébriques, l'estimateur $\hat{y}_T$ du maximum de vraisemblance correspondant :

$$\hat{y}_T = y_{(1)} - \left[\frac{1}{n(n-1)} + \frac{1}{(n-1)} \ln\left(1 - \frac{1}{T}\right)\right] \sum_{i=1}^{n} \left(y_i - y_{(1)}\right) \tag{3.30}$$

4 PRÉCISION DES ESTIMATEURS

Nous donnons dans ce chapitre les outils permettant de quantifier la précision d'un estimateur du quantile x_T de période de retour T. Nous considérons le cas général d'une variable aléatoire X distribuée selon une loi de probabilité à 3 paramètres, et l'événement extrême $\{X > x_T\}$.

4.1 Variance asymptotique du quantile x_T de période de retour T

On a présenté au chapitre 3 des techniques nous permettant d'obtenir des estimations des moments et des paramètres d'une loi de probabilité desquelles on en a déduit ensuite des estimateurs du quantile x_T de période de retour T. L'estimation du quantile est cependant inexacte puisqu'elle est basée sur un échantillon de taille finie. Cette incertitude d'échantillonnage qui se traduit par des écarts entre l'estimation et la quantité théorique doit être quantifiée. Pour ce faire, on utilise traditionnellement la variance de l'estimateur. Cette mesure de précision traduit la dispersion de celui-ci autour de sa moyenne. On a vu, à la Section 3.1, que la variance d'un estimateur détermine sont efficacité (propriété 3).

Toutefois, il est souvent impossible de calculer la variance exacte d'un estimateur du quantile x_T de période de retour T car sa loi de probabilité est difficile à identifier. Aussi, considérons-nous plutôt la variance asymptotique, qui repose sur les hypothèses que la taille de l'échantillon est grande, et que la loi asymptotique de l'estimateur est la loi normale. Cette variance est donc une approximation de la variance exacte. Plus la taille d'échantillon est grande, plus l'approximation est bonne. L'hypothèse concernant la loi normale est raisonnable puisque les estimateurs issus des méthodes présentées au chapitre 3, tout au moins des méthodes classiques du maximum de vraisemblance et des moments, sont asymptotiquement distribués selon une loi normale.

Considérons n variables aléatoires $X_1, X_2, X_3, \ldots, X_n$ indépendantes que l'on suppose indépendantes et identiquement distribuées selon une loi $F(x;\underline{\theta})$ à 3 paramètres. Le quantile théorique x_T de période de retour T s'exprime comme une fonction des paramètres et de T (équation 2.37), que l'on écrit ici de la façon suivante :

$$x_T = F^{-1}\left(\theta_1, \theta_2, \theta_3; 1-\frac{1}{T}\right) = h(\theta_1, \theta_2, \theta_3, T) \tag{4.1}$$

Soient $\hat{\theta}_1, \hat{\theta}_2$ et $\hat{\theta}_3$ les estimateurs des paramètres θ_1, θ_2 et θ_3 de la loi obtenus par une méthode quelconque. Un estimateur $\hat{x}_T$ de cette quantité est alors obtenu, comme nous l'avons vu au Chapitre 3, en remplaçant dans l'équation (4.1) les paramètres par leurs estimations correspondantes :

$$x_T = F^{-1}\left(\hat{\theta}_1, \hat{\theta}_2, \hat{\theta}_3; 1-\frac{1}{T}\right) = h\left(\hat{\theta}_1, \hat{\theta}_2, \hat{\theta}_3, T\right) \tag{4.2}$$

Selon la théorie des grands nombres (théorème de la limite centrale, Bickel et Docksum, 1977), l'estimateur $\hat{x}_T$ est alors distribué asymptotiquement selon une loi normale de moyenne x_T et de variance donnée par :

$$Var\{\hat{x}_T\} = \sum_{i=1}^{3}\left(\frac{\partial g}{\partial \theta_i}\right)^2 Var\{\hat{\theta}_i\} + \sum_{i=1}^{3}\sum_{j\neq i}^{3}\left(\frac{\partial g}{\partial \theta_i}\right)\left(\frac{\partial g}{\partial \theta_j}\right) Cov\{\hat{\theta}_i, \hat{\theta}_j\} \tag{4.3}$$

où $Cov\{X, Y\}$ désigne la covariance entre les variables aléatoires X et Y, et mesure la relation linéaire entre deux variables. Si X et Y sont des variables aléatoires indépendantes, alors la covariance est nulle.

La variance asymptotique de l'estimateur $\hat{x}_T$ (Équation 4.3) dépend :

- des dérivées partielles de la fonction g par rapport à chacun des paramètres de la loi. Elles sont aisément obtenues lorsqu'on peut déterminer explicitement x_T (par exemple, pour la loi Gumbel). Si aucune forme explicite n'est disponible, on utilise une dérivée numérique.
- des variances et covariances asymptotiques des estimateurs des paramètres de la loi de probabilité.

Il faut donc aussi déterminer ces caractéristiques. Les sections qui suivent présentent les variances et covariances asymptotiques des estimateurs des paramètres issus des méthodes d'estimations présentées au chapitre 3.

4.2 Variance et covariances asymptotiques des estimateurs du maximum de vraisemblance

Comme on l'a vu précédemment (Section 3.3.1), les propriétés des estimateurs du maximum de vraisemblance, lorsque la taille d'échantillon est grande, sont bien connues. Ils sont convergents, asymptotiquement non-biaisés et asymptotiquement efficaces. De plus, toujours en considérant une loi à 3 paramètres, si la fonction de vraisemblance $L(\theta_1, \theta_2, \theta_3)$

admet qu'un seul maximum, les variables aléatoires $\sqrt{n}\left(\hat{\theta}_1 - \theta_1\right), \sqrt{n}\left(\hat{\theta}_2 - \theta_2\right), \sqrt{n}\left(\hat{\theta}_2 - \theta_2\right)$ sont distribuées selon une loi normale multidimensionnelle de moyenne nulle et de matrice des variances et covariances $\boldsymbol{\Sigma}$ dont les éléments correspondent à ceux de l'inverse de la matrice d'information de Fisher $\mathbf{I}_f$. Les éléments $\left(\mathbf{I}_f\right)_{ij}$ de la matrice d'information de Fisher sont donnés par :

$$\left(\mathbf{I}_f\right)_{ij} = -E\left\{\frac{\partial^2 \ln L(\theta_1, \theta_2, \theta_3)}{\partial\theta_i \partial\theta_j}\right\}, \qquad \text{pour } i \text{ et } j \in \{1, 2, 3\} \tag{4.4}$$

En notation matricielle, on a :

$$\boldsymbol{\Sigma} = \mathbf{I}_f^{-1} = \begin{pmatrix} Var\left\{\hat{\theta}_1\right\} & Cov\left\{\hat{\theta}_1, \hat{\theta}_2\right\} & Cov\left\{\hat{\theta}_1, \hat{\theta}_3\right\} \\ Cov\left\{\hat{\theta}_2, \hat{\theta}_1\right\} & Var\left\{\hat{\theta}_2\right\} & Cov\left\{\hat{\theta}_2, \hat{\theta}_3\right\} \\ Cov\left\{\hat{\theta}_3, \hat{\theta}_1\right\} & Cov\left\{\hat{\theta}_3, \hat{\theta}_2\right\} & Var\left\{\hat{\theta}_3\right\} \end{pmatrix} \tag{4.5}$$

Pour déterminer les variances et les covariances des estimateurs du maximum de vraisemblance des paramètres d'une loi $F(x; \underline{\theta})$, il faut donc :

1. obtenir les dérivées secondes de la fonction de vraisemblance logarithmique par rapport à chacun des paramètres. Ces dérivées sont fonction des observations et des paramètres;

2. en déduire la matrice d'information de Fisher $\mathbf{I}_f$ en calculant l'espérance mathématique de ces dérivées secondes ;

3. inverser la matrice $\mathbf{I}_f$ ainsi obtenue.

4.3 Variances et covariances asymptotiques des estimateurs de la méthode des moments

Pour décrire l'approche utilisée afin de déterminer les variances et covariances asymptotiques des estimateurs issus de la méthode des moments, nous traitons ici, à titre d'exemple, le cas particulier où on utilise les trois premiers moments non-centrés. La technique est analogue lorsqu'on utilise tout autre ensemble de moments. Elle peut donc être appliquée aux méthodes WRC, MM1 et SAM. Le cas général est présenté dans Bobée et Ashkar (1991).

Les variances et les covariances asymptotiques des estimateurs de la méthode des moments sont reliées à celles des moments de l'échantillon décrits à la Section 3.1. Plus précisément, on a :

$$Cov\{m'_r, m'_q\} = \sum_{i=1}^{3}\left(\frac{\partial \mu'_r}{\partial \theta_i}\right)\left(\frac{\partial \mu'_q}{\partial \theta_i}\right) Var\{\hat{\theta}_i\} + \sum_{i=1}^{3}\sum_{j\neq i}^{3}\left(\frac{\partial \mu'_r}{\partial \theta_i}\right)\left(\frac{\partial \mu'_q}{\partial \theta_j}\right) Cov\{\hat{\theta}_i, \hat{\theta}_j\} \qquad (4.6)$$

pour r et $q \in \{1, 2, 3\}$, où $Cov\{m'_r, m'_r\} = Var\{m'_r\}$. La détermination des variances et covariances asymptotiques des estimateurs issus de la méthode des moments repose sur cette relation. En effet, l'application de l'équation (4.6) conduit au système suivant :

$$\begin{pmatrix} Var\{m'_1\} \\ Var\{m'_2\} \\ Var\{m'_3\} \\ Cov\{m'_1, m'_2\} \\ Cov\{m'_1, m'_3\} \\ Cov\{m'_2, m'_3\} \end{pmatrix} = \begin{pmatrix} v_{11} & \cdots & v_{16} \\ & & \\ \vdots & & \vdots \\ \vdots & & \vdots \\ & & \\ v_{61} & \cdots & v_{66} \end{pmatrix} \bullet \begin{pmatrix} Var\{\hat{\theta}_1\} \\ Var\{\hat{\theta}_2\} \\ Var\{\hat{\theta}_3\} \\ Cov\{\hat{\theta}_1, \hat{\theta}_2\} \\ Cov\{\hat{\theta}_1, \hat{\theta}_3\} \\ Cov\{\hat{\theta}_2, \hat{\theta}_3\} \end{pmatrix} \qquad (4.7)$$

qui peut s'exprimer, en notation matricielle :

$$\mathbf{V}_m = \mathbf{V}\,\mathbf{V}_p \qquad (4.8)$$

où $\mathbf{V}_m$ est le vecteur des variances et covariances asymptotiques des moments de l'échantillon, $\mathbf{V}_p$ celui des variances et covariances asymptotiques des estimateurs des paramètres (ce que l'on veut déterminer) et $\mathbf{V}$ la matrice définie comme suit :

$$\mathbf{V} = \begin{pmatrix} A_{11}^2 & A_{12}^2 & A_{13}^2 & 2A_{11}A_{12} & 2A_{11}A_{13} & 2A_{12}A_{13} \\ A_{21}^2 & A_{22}^2 & A_{23}^2 & 2A_{21}A_{22} & 2A_{21}A_{23} & 2A_{22}A_{23} \\ A_{31}^2 & A_{32}^2 & A_{33}^2 & 2A_{31}A_{32} & 2A_{31}A_{33} & 2A_{32}A_{33} \\ A_{11}A_{21} & A_{12}A_{22} & A_{13}A_{23} & (A_{11}A_{22}+A_{12}A_{21}) & (A_{11}A_{23}+A_{13}A_{21}) & (A_{12}A_{23}+A_{13}A_{22}) \\ A_{11}A_{31} & A_{12}A_{32} & A_{13}A_{33} & (A_{11}A_{32}+A_{12}A_{31}) & (A_{11}A_{33}+A_{13}A_{31}) & (A_{12}A_{33}+A_{13}A_{32}) \\ A_{21}A_{31} & A_{22}A_{32} & A_{23}A_{33} & (A_{21}A_{32}+A_{22}A_{31}) & (A_{21}A_{33}+A_{23}A_{31}) & (A_{22}A_{33}+A_{23}A_{32}) \end{pmatrix}$$

Le terme A_{rj} de cette matrice est la dérivée partielle du moment théorique μ'_r par rapport au paramètre θ_j :

$$A_{rj} = \frac{\partial \mu'_r}{\partial \theta_j} \tag{4.9}$$

Les éléments de $\mathbf{V}_m$ peuvent être obtenus pour tout moment de l'échantillon à partir des expressions données dans Kendall et Stuart (1987, chapitre 10). Une fois $\mathbf{V}_m$ et $\mathbf{V}$ déterminés, on peut en déduire les variances et covariances asymptotiques des estimateurs obtenus par la méthode des moments en appliquant la relation :

$$\mathbf{V}_p = \mathbf{V}^{-1}\,\mathbf{V}_m \tag{4.10}$$

En résumé, pour déterminer les variances et les covariances des estimateurs de la méthode des moments des paramètres d'une loi $F(x;\underline{\theta})$, on doit :

1. calculer les variances et covariances asymptotiques des moments de l'échantillon utilisés pour déterminer les estimateurs. On forme ainsi le vecteur $\mathbf{V}_m$;
2. calculer les dérivées partielles des moments théoriques utilisés par rapport à chacun des paramètres et construire alors la matrice $\mathbf{V}$;
3. inverser la matrice $\mathbf{V}$ ainsi obtenue;
4. déduire le vecteur $\mathbf{V}_p$ à l'aide de l'expression (4.10).

Cette technique peut aussi être appliquée pour déduire les variances et covariances asymptotiques des estimateurs déduits de la méthode des moments pondérés MMP. Toutefois, les variances et covariances asymptotiques des moments pondérés de l'échantillon b_r (équation 3.22) utilisés pour déterminer les estimateurs (éléments du vecteur $\mathbf{V}_m$) sont plus difficiles à obtenir. Les détails peuvent être obtenus dans Perreault *et al.* (1992a).

4.4 Intervalle de confiance asymptotique

Nous sommes maintenant en mesure de déterminer un estimateur ponctuel du quantile x_T de période de retour T (chapitre 3) et de calculer sa variance asymptotique (Sections 4.1, 4.2 et 4.3). Cette variance, telle que présentée aux sections précédentes, est théorique puisqu'elle est fonction des paramètres inconnus de la loi de probabilité. Pour en déduire une valeur numérique à partir d'un échantillon donné, il suffit de remplacer les paramètres par leur estimation obtenue à partir d'une méthode quelconque. Pour distinguer la variance théorique et sa valeur numérique (l'estimation de la variance), nous notons cette dernière $S^2_{\hat{x}_T}$.

La variance asymptotique est une mesure de la précision de l'estimateur. Elle est particulièrement intéressante pour comparer deux estimations issues de méthodes différentes. Toutefois, la variance demeure en pratique difficile à interpréter. Une approche permettant d'avoir une meilleure idée de l'exactitude de l'estimation est la construction d'un intervalle de confiance. Un **intervalle de confiance** de niveau 100(1 - α)% pour une quantité théorique θ inconnue, est un intervalle $[L_1, L_2]$ tel que :

$$Prob\{L_1 \leq \theta \leq L_2\} = 1 - \alpha \tag{4.11}$$

où L_1 et L_2 sont des statistiques indépendantes de θ. Cette approche nous permet donc d'obtenir un intervalle contenant la valeur théorique inconnue avec une probabilité $(1 - \alpha)$.

Pour construire un tel intervalle, on doit trouver une statistique qui fait intervenir la quantité θ, et dont la loi probabilité théorique ou approchée est connue. Dans le cas de la construction d'un intervalle de confiance pour le quantile x_T de période de retour T, cette statistique est simple à obtenir. En effet, on a vu (Section 4.1) que l'estimateur $\hat{x}_T$ est distribué asymptotiquement selon une loi normale de moyenne x_T et de variance donnée par $Var\{\hat{x}_T\}$. La statistique :

$$S = \frac{\hat{x}_T - x_T}{\sqrt{Var\{\hat{x}_T\}}} \tag{4.12}$$

est alors distribuée asymptotiquement selon une loi normale centrée-réduite. Remarquons que S fait intervenir la quantité théorique inconnue x_T. On déduit donc que :

$$\text{Prob}\left\{ z_{\alpha/2} \leq \frac{\hat{x}_T - x_T}{\sqrt{Var\{\hat{x}_T\}}} \leq z_{1-\alpha/2} \right\} = 1 - \alpha \tag{4.13}$$

où $z_{\alpha/2}$ et $z_{1-\alpha/2}$ sont respectivement les quantiles de probabilité au non-dépassement $\alpha/2$ et $1-\alpha/2$ de la loi normale centrée-réduite. En isolant x_T dans l'expression (4.13) et en remarquant que pour la loi normale $z_{\alpha/2} = -z_{1-\alpha/2}$ (loi symétrique par rapport à zéro), on obtient :

$$\text{Prob}\left\{ \hat{x}_T - z_{1-\alpha/2}\sqrt{Var\{\hat{x}_T\}} \leq x_T \leq \hat{x}_T + z_{1-\alpha/2}\sqrt{Var\{\hat{x}_T\}} \right\} = 1 - \alpha \tag{4.14}$$

et l'intervalle

$$\left[\hat{x}_T - z_{1-\alpha/2}\sqrt{Var\{\hat{x}_T\}} \;;\; \hat{x}_T + z_{1-\alpha/2}\sqrt{Var\{\hat{x}_T\}} \right] \tag{4.15}$$

est un intervalle de confiance de niveau 100(1 - α)% pour le quantile théorique x_T de période de retour T. Cet intervalle est dit **intervalle de confiance asymptotique** puisqu'il est construit l'aide d'une approximation de la loi exacte de $\hat{x}_T$ (loi asymptotique). Pour obtenir un intervalle de confiance pour x_T à partir de l'échantillon, on remplace $Var\{\hat{x}_T\}$ par $S^2_{\hat{x}_T}$ dans l'intervalle (4.15).

Le logiciel ***AJUSTE-II*** fournit pour chaque estimation de quantile, l'intervalle de confiance qui y est associé. Le niveau de confiance $(1-\alpha)$ peut être sélectionné par l'utilisateur.

5 TESTS STATISTIQUES

Pour que les résultats d'une analyse hydrologique de fréquence soient théoriquement valides, les observations utilisées doivent satisfaire tout d'abord certaines caractéristiques statistiques liées à la façon dont elles ont été acquises et mesurées. En effet, les observations doivent être indépendantes et identiquement distribuées selon une loi de probabilité bien spécifiée. Si ces hypothèses sont vérifiées, les observations sont alors homogènes et stationnaires, c'est-à-dire que les caractéristiques statistiques de l'échantillon (en particulier, la moyenne) sont invariantes dans le temps.

Le logiciel *AJUSTE-II* permet de vérifier, d'une part, si l'échantillon satisfait aux caractéristiques statistiques d'indépendance, d'homogénéité et de stationnarité, et d'autre part, s'il provient bel et bien de la loi de probabilité envisagée. Ces vérifications sont effectuées à l'aide de tests d'hypothèses qui sont décrits dans ce chapitre.

Dans un premier temps, nous présentons brièvement les principales notions de base concernant la théorie des tests d'hypothèses.

5.1 Notions de base

Pour introduire le plus simplement possible les notions de base de la théorie des tests d'hypothèses, nous considérons dans cette section, sans perte de généralité, les tests concernant la valeur théorique d'un paramètre plutôt que ceux touchant les caractéristiques de l'échantillon, comme l'indépendance, la stationnarité ou l'homogénéité des observations.

Nous avons présenté dans les chapitres précédents l'estimation ponctuelle d'une quantité théorique. Une estimation de ce paramètre inconnu (paramètre d'une loi de probabilité, moment, quantile, etc.), qu'on note ici θ, peut être obtenu à partir d'un échantillon. En général, lorsqu'on fait une estimation, nous n'avons aucune idée a priori de la valeur que peut prendre le paramètre. Nous tentons de lui attribuer la meilleure valeur possible à partir d'une méthode d'estimation. Toutefois, lorsqu'on effectue un test d'hypothèses concernant le paramètre θ, on a une idée préconçue de la ou des valeurs que peut prendre ce paramètre. Ceci implique donc que deux hypothèses sont considérées lorsqu'on effectue un test statistique: l'hypothèse suggérée par le spécialiste (H_0) qui est appelée l'**hypothèse nulle**, et sa négation (H_1) appelée l'**hypothèse alternative**. L'objectif d'un test statistique est de

vérifier, à l'aide de l'information contenue dans l'échantillon, s'il y a assez d'évidences statistiques pour rejeter l'hypothèse nulle.

Une fois les hypothèses H_0 et H_1 bien formulées, et étant donné un échantillon $x_1, x_2, x_3, \ldots, x_n$, la décision de rejeter ou non l'hypothèse nulle est prise en se basant sur une fonction des observations $T(x_1, x_2, \ldots, x_n)$ dont la loi exacte, ou tout au moins une approximation de cette loi, est connue sous l'hypothèse H_0. Cette fonction traduit l'écart entre les hypothèses du test et est appelée la **statistique du test**. Considérons le test qui vérifie les hypothèses suivantes :

$$H_0 : \theta = \theta_0 \quad \text{contre} \quad H_1 : \theta > \theta_0$$

Si la statistique du test prend une valeur rarement rencontrée lorsque $\theta = \theta_0$ et qui tend à favoriser l'hypothèse alternative, on rejette alors H_0. Si plutôt la valeur de la statistique se réalise souvent sous l'hypothèse nulle, alors on ne rejette pas H_0 en faveur de H_1. L'ensemble des valeurs de la statistique T pour lesquelles on rejette et on ne rejette pas l'hypothèse nulle H_0 sont appelés respectivement la **région critique** et la **région d'acceptation** du test. Ces deux ensembles sont notés RC et RA.

Lorsqu'on effectue un test d'hypothèses deux types d'erreurs peuvent survenir :

- L'hypothèse H_0 est rejetée alors qu'elle est effectivement vraie. Une telle erreur est appelée l'**erreur de type I**, et sa probabilité, calculée sous l'hypothèse nulle, est notée α. Cette probabilité est le **niveau de signification** du test et s'exprime de la façon suivante :

$$\alpha = \text{Prob}\{T(x_1, x_2, \ldots, x_n) \in RC \mid H_0 \text{ est vraie}\} \tag{5.1}$$

- L'hypothèse H_0 n'est pas rejetée alors qu'elle est fausse. Cette erreur est appelée l'**erreur de type II**, et sa probabilité, calculée sous l'hypothèse alternative, est notée β :

$$\beta = \text{Prob}\{T(x_1, x_2, \ldots, x_n) \in RA \mid H_0 \text{ est fausse}\} \tag{5.2}$$

On déduit que la probabilité que H_0 soit rejetée, alors que cette hypothèse est fausse, est égale à $1\text{-}\beta$. Cette probabilité correspond à la **puissance du test**, qui est un indicateur de son efficacité. Évidemment, un test idéal aurait $\alpha = 0$ et $1 - \beta = 1$. On ne peut, en pratique, construire un tel test étant donnée l'information limitée contenue dans un échantillon de taille finie. En général, lorsqu'on développe un test d'hypothèses, on fixe d'abord le niveau de

signification α à une petite valeur, et on tente de construire une statistique permettant d'obtenir une puissance $1-\beta$ élevée. Cette opération peut être difficile puisque plus le niveau signification α est faible plus la puissance diminue.

En résumé, les principales étapes pour effectuer un test statistique sont les suivantes :

1. Formuler les hypothèses;
2. Choisir la statistique du test;
3. Choisir le niveau de signification;
4. Définir la région critique qui constitue la règle de décision. La région critique dépend du niveau de signification et de la loi de la statistique lorsque l'hypothèse nulle est vraie.
5. Calculer la statistique du test à partir des données de l'échantillon.
6. Appliquer la règle de décision définie en 4, c'est-à-dire vérifier si la valeur calculée de la statistique appartient à la région critique.

Pour illustrer ces 6 étapes, nous considérons l'exemple suivant qui consiste à tester une hypothèse concernant la moyenne théorique.

Exemple 5.1. *Supposons qu'il existe une norme stipulant que le débit moyen d'une rivière donnée doit être maintenu à 100 m³/s, et surtout ne pas dépasser ce seuil. On est donc intéressé à vérifier les hypothèses suivantes :*

Étape 1

$$H_0 : \mu = 100 \quad \textit{contre} \quad H_1 : \mu > 100$$

où le paramètre μ désigne la moyenne théorique du débit moyen. De plus, supposons que la variable aléatoire X représentant le débit est distribuée selon une loi normale de moyenne μ inconnue et de variance $\sigma^2 = 400$ connue.

Pour vérifier les hypothèses sur la moyenne théorique d'une loi normale, il est naturel d'utiliser une statistique basée sur la moyenne arithmétique $\overline{X}$. En effet, on a vu à l'Exemple 3.1 que cette statistique est l'estimateur du maximum de vraisemblance du paramètre μ. Une statistique traduisant l'écart entre les hypothèses peut s'écrire comme suit :

Étape 2

$$T(x_1, x_2, \ldots, x_n) = \sqrt{n}\,\frac{\overline{X} - 100}{20}$$

et est distribuée selon une loi normale centrée et réduite.

Étape 3

Le choix du niveau de signification α est conditionné par les conséquences que peut entraîner une erreur de type I. Nous choisissons ici $\alpha = 5\%$. Ainsi, nous sommes prêts à accepter que l'on rejette à tort l'hypothèse nulle 5 fois sur 100.

Étape 4

Il ne reste plus qu'à définir la région critique et ensuite appliquer le test. Selon les hypothèses formulées et la statistique T, il est naturel de rejeter l'hypothèse nulle si cette statistique prend de grandes valeurs positives. Ainsi, la région critique est de la forme $RC = \{T > c\}$, où c est appelée ***valeur critique*** *du test et est telle que :*

$$5\% = \text{Prob}\{T(x_1, x_2, \ldots, x_n) \in RC \mid H_0 \text{ est vraie}\}$$
$$= \text{Prob}\{T(x_1, x_2, \ldots, x_n) > c\}$$

Puisque T est distribuée selon une loi normale centrée-réduite, c est le quantile de probabilité au dépassement égal à 5% de cette loi, soit c = 1,645 (valeur obtenue de la table standard de la loi normale).

Étapes 5 et 6

Supposons maintenant que l'on dispose d'un échantillon de taille n=25 dont la moyenne arithmétique est $\overline{x} = 105$. Alors la statistique calculée, notée t, est égale à 1,25 et est inférieure à la valeur critique 1,645. Ainsi, si on accepte une erreur de type I de 5%, il n'y a pas assez d'évidence statistique pour rejeter l'hypothèse que la moyenne théorique est de 100 m^3/s.

Avant de présenter les tests disponibles dans le logiciel ***AJUSTE-II***, nous décrivons une dernière notion de base des tests d'hypothèse. La simple règle de décision classique qui consiste à rejeter ou non l'hypothèse nulle peut être critiquée puisqu'elle ne fournit aucun indicateur nous permettant de quantifier notre décision suite à un test statistique. Un indice utilisé à cette fin est le **niveau de signification observé** $\hat{\alpha}$ qui est défini comme étant la probabilité, sous l'hypothèse nulle, d'obtenir une valeur plus extrême de la statistique que celle calculée à partir de l'échantillon (plus extrême au sens de l'hypothèse alternative). En

particulier, pour l'Exemple 5.1, le niveau de signification observé correspond à la probabilité au dépassement de la valeur t calculée, c'est-à-dire :

$$\begin{aligned}\hat{\alpha} &= \text{Prob}\{T(x_1, x_2, \ldots, x_n) > t \mid H_0 \text{ est vraie}\} \\ &= \text{Prob}\{T(x_1, x_2, \ldots, x_n) > 1,25\} = 0,1056\end{aligned} \tag{5.3}$$

Une grande valeur du niveau de signification observé $\hat{\alpha}$ supporte l'hypothèse nulle alors qu'une faible valeur favorise le rejet de cette hypothèse. Plus cette valeur diffère du niveau de signification fixé a priori α, plus on aura confiance en notre décision. On remarque que pour effectuer le test, comparer t au quantile c de la loi normale centrée-réduite est équivalent à comparer $\hat{\alpha}$ au niveau de signification fixé a priori $\alpha = 5\%$. Ainsi, puisque $0,1056 > 0,05$, on ne peut rejeter l'hypothèse nulle dans le cas de l'Exemple 5.1.

5.2 Vérification des hypothèses

L'analyse hydrologique de fréquence repose sur des hypothèses statistiques. En effet, pour que les résultats d'une telle analyse soient théoriquement valides, les observations utilisées doivent être indépendantes et identiquement distribuées; ce qui implique qu'elles sont homogènes et stationnaires. Nous présentons, dans ce qui suit, ces caractéristiques ainsi que les tests qui permettent de vérifier ces hypothèses. Notons que les tests présentés dans cette section sont non paramétriques, c'est-à-dire que la loi de leur statistique est indépendante de celle des observations. Il n'est donc pas nécessaire pour appliquer ces tests de connaître la loi d'où proviennent les données comme c'est le cas de plusieurs tests usuels connus (test de Student, test de Fisher, etc.), qui s'appuient généralement sur l'hypothèse de normalité des observations.

5.2.1 Indépendance : test de Wald-Wolfowitz

Des observations sont indépendantes si la probabilité d'occurrence de chacune d'entre elles n'est pas influencée par les autres observations. Par exemple, on considère les débits maximums annuels indépendants si l'intensité d'une crue n'est pas influencée par celle observée l'année précédente. En d'autres mots, on ne peut tirer aucune information d'un débit maximum annuel pour prédire celui de l'année suivante. Une dépendance peut généralement être observée lorsque l'intervalle de temps entre les observations est réduit. En effet, il est clair que les débits journaliers ne sont pas indépendants puisqu'il y a forte chance qu'un débit observé soit élevé si celui du jour précédent est élevé, et faible lorsque

l'observation du jour précédent est faible. On dira alors que les observations sont autocorrélées, et dans ce cas on ne peut pas utiliser l'analyse hydrologique de fréquence.

Le test utilisé dans ***AJUSTE-II*** pour vérifier d'indépendance des observations est le test de Wald-Wolfowitz (1943) qui compare les hypothèses suivantes :

$$H_0 : X_1, X_2, \ldots, X_n \text{ sont indépendantes}$$
$$\text{contre}$$
$$H_0 : X_1, X_2, \ldots, X_n \text{ ne sont pas indépendantes}$$

Considérons n variables aléatoires $X_1, X_2, X_3, \ldots, X_n$ (par exemple, le débit de crue des n dernières années) et les réalisations correspondantes $x_1, x_2, \ldots, x_n$ (valeurs numériques correspondantes). La statistique de Wald-Wolfowitz R s'exprime de la façon suivante :

$$R = \sum_{i=1}^{n-1} X_i X_{i+1} + X_1 X_n \tag{5.4}$$

Sous l'hypothèse nulle, c'est-à-dire lorsque les n variables aléatoires sont indépendantes, la statistique R est distribuée asymptotiquement (lorsque $n \to +\infty$) selon une loi normale de moyenne et de variance données respectivement par :

$$E\{R\} = \frac{s_1^2 - s_2}{n-1} \quad \text{et} \quad Var\{R\} = \frac{s_2^2 - s_4}{n-1} - E^2\{R\} + \frac{s_1^4 - 4s_1^2 s_2 + 4s_1 s_3 + s_2^2 - 2s_4}{(n-1)(n-2)} \tag{5.5}$$

où $s_r = n m_r'$, m_r' étant le moment non-centré d'ordre r de l'échantillon (Section 3.1). La statistique standardisée :

$$U = \frac{R - E\{R\}}{\sqrt{Var\{R\}}} \tag{5.6}$$

est distribuée asymptotiquement selon une loi normale centrée-réduite. Les observations seront indépendantes si la valeur de R est proche de sa moyenne. Ainsi, on rejette l'hypothèse nulle pour de grandes valeurs de la statistique U en valeur absolue, calculée à partir des valeurs numériques $x_1, x_2, \ldots, x_n$. La région critique du test de Wald-Wolfowitz au niveau de signification α est alors de la forme $\{|U| > z_{\alpha/2}\}$, où $z_{\alpha/2}$ est le quantile de probabilité au **dépassement** égale à $\alpha/2$ de la loi normale centrée réduite.

La règle de décision pour effectuer ce test à un niveau de signification donné α est donc la suivante :

- Si $|U| > z_{\alpha/2}$, on rejette H_0, les observations ne peuvent être considérées comme indépendantes;
- Sinon, on ne rejette pas H_0.

5.2.2 Homogénéité : test de Wilcoxon

On entend par échantillon aléatoire homogène, un échantillon dont toutes les observations proviennent de la même population statistique. En pratique, une série d'observations sera homogène si les données sont toujours acquises dans les mêmes conditions. Par exemple, les débits peuvent être hétérogènes s'il y eu déplacement de la station de mesure durant la période d'acquisition des données. Une façon de vérifier dans ce cas l'hypothèse d'homogénéité est de comparer la moyenne des débits obtenus avant la date de modification de la procédure d'acquisition avec celle des débits obtenus après cette date. C'est ce que le logiciel ***AJUSTE-II*** permet d'effectuer à l'aide du test de rang de Wilcoxon.

Considérons n variables aléatoires $X_1, X_2, X_3, \ldots, X_n$ et la série de réalisations correspondantes $x_1, x_2, \ldots, x_n$ que l'on divise en deux sous-échantillons, l'un formé des observations mesurées $x_1, x_2, \ldots, x_{n_1}$ avant la date de modification de la procédure d'acquisition et l'autre des données mesurées suite au changement $y_1, y_2, \ldots, y_{n_2}$. Supposons de plus qu'ils proviennent respectivement de deux populations de moyenne et de variance (μ_1, σ_1^2) et (μ_2, σ_2^2). Le test de Wilcoxon vérifie les hypothèses :

$$H_0 : \mu_1 = \mu_2 \quad \text{contre} \quad H_1 : \mu_1 \neq \mu_2$$

La statistique utilisée, qui est asymptotiquement distribuée selon une loi normale centrée-réduite, est donnée par :

$$W = \frac{V - \frac{n_1(n+1)}{2} + \frac{1}{2}}{\sqrt{Var\{V\}}} \tag{5.7}$$

où :

- $V = \sum_{i=1}^{n_1+n_2} R_i \, s(R_i)$
- R_i est le rang correspondant à l'observation i de l'échantillon combiné de taille $n_1 + n_2 = n$ classé en ordre croissant
- $s(R_i) = \begin{cases} 0 & si\ R_i \text{ correspond à une donnée du sous-échantillon des } y_i \\ 1 & si\ R_i \text{ correspond à une donnée du sous-échantillon des } x_i \end{cases}$

- $Var\{V\} = \frac{n_1 n_2 (n+1)}{12} - \frac{n_1 n_2 \sum_{k=1}^{h} d_k^3 - d_k}{12n(n-1)}$, en supposant que nous avons dans l'échantillon combiné h groupes distincts contenant des observations égales, et que le nombre d'observations égales dans chacun de ces groupes soient respectivement $d_1, d_2, \ldots, d_h$. Si toutes les observations sont distinctes, on a $h = n$, $d_k^3 - d_k = 0$ et donc :

$$Var\{V\} = \frac{n_1 n_2 (n+1)}{12}$$

La règle de décision pour effectuer ce test à un niveau de signification donné α est la suivante :

- Si $|W| > z_{\alpha/2}$, on rejette H_0
- Sinon, on ne rejette pas H_0

où $z_{\alpha/2}$ est le quantile de probabilité au **dépassement** $\alpha/2$ de la loi normale centrée-réduite. Un test tout à fait équivalent au test de Wilcoxon est celui de Mann-Withney. Lehmann (1975, Chap. 1) donne les détails théoriques concernant ces deux tests.

5.2.3 Stationnarité : test de Kendall

On dit que les observations sont stationnaires si, outre les fluctuations aléatoires de la série, les caractéristiques statistiques (moyenne, variance, etc.) de la série ne varient pas dans le temps. La non-stationnarité se traduit généralement par des changements brusques ou graduels de la moyenne des observations. Cette hypothèse peut être vérifiée dans le cas de changements brusques en comparant les moyennes de deux sous-échantillons à l'aide du test de Wilcoxon (Section 5.2.2), si on connaît a priori la date du saut. Toutefois, lorsqu'on n'a pas cette information, ou si on soupçonne un changement graduel de la moyenne, il est préférable d'utiliser un autre test, en l'occurrence le test de Kendall.

Le test de Kendall compare les hypothèses suivantes :

H_0 : La moyenne des variables aléatoires est constante dans le temps

contre

H_1 : La moyenne des variables aléatoires n'est pas constante dans le temps

Considérons n variables aléatoires $X_1, X_2, X_3, \ldots, X_n$ classées par ordre chronologique. La statistique S du test de Kendall (1975) s'exprime de la façon suivante :

$$S = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} sgn(X_j - X_i) \tag{5.8}$$

où la fonction $sgn(.)$ est donnée par :

$$sgn(x) = \begin{cases} 1 & \text{si } x > 0 \\ 0 & \text{si } x = 0 \\ -1 & \text{si } x < 0 \end{cases} \tag{5.9}$$

Sous l'hypothèse nulle, c'est-à-dire lorsque les variables aléatoires sont stationnaires, la statistique S est distribuée asymptotiquement selon une loi normale de moyenne nulle et de variance donnée par :

$$Var\{S\} = \frac{1}{18}\left[n(n-1)(2n+5) - \sum_t t(t-1)(2t+5) \right] \tag{5.10}$$

où t désigne les nombres de valeurs égales dans un ensemble donné d'observations identiques et $\sum_t$ est la sommation sur tous les ensembles d'observations identiques rencontrés dans l'échantillon. En ajoutant une correction de continuité, la statistique standardisée :

$$K = \begin{cases} \dfrac{S-1}{\sqrt{Var\{S\}}} & \text{si S} > 0 \\ 0 & \text{si S} = 0 \\ \dfrac{S+1}{\sqrt{Var\{S\}}} & \text{si S} < 0 \end{cases} \tag{5.11}$$

est distribuée asymptotiquement selon une loi normale centrée-réduite. Les observations seront stationnaires si la valeur de S est proche de sa moyenne, c'est-à-dire nulle. Ainsi, on rejette l'hypothèse nulle pour de grandes valeurs de la statistique $|K|$ en valeur absolue, calculée à partir des observations $x_1, x_2, \ldots, x_n$. La région critique du test de Kendall au niveau de signification α est alors de la forme $\{|K| > z_{\alpha/2}\}$ où $z_{\alpha/2}$ est le quantile de probabilité au **dépassement** égale à $\alpha/2$ de la loi normale centrée-réduite.

La règle de décision pour effectuer ce test à un niveau de signification donné α est donc la suivante :

- Si $|K| > z_{\alpha/2}$, on rejette H_0, les observations ne peuvent être considérées stationnaires;
- Sinon, on ne rejette pas H_0.

5.3 Tests d'adéquation du modèle

Nous avons présenté au chapitre 4 une manière de quantifier l'incertitude d'échantillonnage en introduisant la notion de variance des estimateurs. Un autre type d'incertitude doit être aussi considéré lors de l'estimation des quantiles de période de retour. C'est l'incertitude associée au choix de la loi de probabilité. En effet, les probabilités d'occurrence théoriques des événements étudiés sont inconnues et on cherche une loi de probabilité qui en donne une bonne approximation. Ainsi, l'approche présentée dans ce rapport repose sur l'hypothèse que les observations proviennent d'une loi de probabilité bien identifiée. Cette méthode est souvent peu robuste et peut, si la loi choisie n'est pas adéquate pour le phénomène étudié, conduire à des résultats erronés. Il est alors important de pouvoir s'assurer que la distribution envisagée est compatible avec les observations. Un certain nombre de techniques, dont les tests d'adéquation, ont été développées pour vérifier cette compatibilité.

Il s'agit ici de savoir si l'on peut considérer que l'échantillon $x_1, x_2, x_3, \ldots, x_n$ que l'on possède provient ou non d'une loi $F(x; \underline{\theta})$ bien déterminée de paramètres $\underline{\theta} = (\theta_1, \theta_2, \ldots, \theta_p)$ inconnus. Plus précisément, nous désirons tester les hypothèses :

$$H_0 : x_1, x_2, \ldots, x_n \in F(x; \underline{\theta}) \quad \text{contre} \quad H_1 : x_1, x_2, \ldots, x_n \notin F(x; \underline{\theta})$$

Cette section est consacrée à la présentation des techniques et des tests statistiques disponibles dans le logiciel ***AJUSTE-II*** pour vérifier ces hypothèses.

5.3.1 Test du khi-deux

Le test d'adéquation du khi-deux, certainement le plus ancien et le plus connu, a été introduit au début du siècle par Karl Pearson. Tel que développé à l'origine, ce test suppose que la valeur des paramètres $\underline{\theta} = (\theta_1, \theta_2, \ldots, \theta_p)$ est connue a priori. Pour définir le test du khi-deux, on commence par faire une partition en M classes de l'ensemble des valeurs possibles d'une observation provenant de la loi envisagée (domaine $\mathcal{D}$ de la variable aléatoire X distribuée selon la loi $F(x; \underline{\theta})$). Désignons par $C_1, C_2, \ldots, C_M$ les M classes construites et, pour tout i appartenant à $\{1, 2, \ldots, M\}$, notons N_i le nombre d'observations de l'échantillon $x_1, x_2, x_3, \ldots, x_n$ qui appartiennent à la classe C_i; on a $\sum_i N_i = n$. Le test du khi-deux est

alors une fonction de l'écart entre les fréquences N_i/n et les probabilités correspondantes des classes C_i sous l'hypothèse H_0, c'est-à-dire celles calculées à l'aide de la loi $F(x;\underline{\theta})$. Ces probabilités, comme on l'a vu au chapitre 2, dépendent bien entendu du vecteur de paramètres $\underline{\theta} = (\theta_1, \theta_2, \ldots, \theta_p)$ et sont données, pour tout i, par

$$p_i(\underline{\theta}) = \int_{C_i} f(x;\underline{\theta})dx \tag{5.12}$$

où $f(x;\underline{\theta})$ est la f.d.p.. Par exemple, si la classe C_i est un intervalle $[z_{i-1}, z_i]$, alors

$$p_i(\underline{\theta}) = \int_{z_{i-1}}^{z_i} f(x;\underline{\theta})dx \tag{5.13}$$

Les écarts entre les fréquences observées N_i/n et les probabilités $p_i(\underline{\theta})$ reflètent en effet l'erreur d'ajustement des observations à la loi $F(x;\underline{\theta})$. Plus précisément, considérons la statistique suivante :

$$X^2(\underline{\theta}) = \sum_{i=1}^{M} \frac{(N_i - np_i(\underline{\theta}))^2}{np_i(\underline{\theta})} \tag{5.14}$$

La statistique $X^2(\underline{\theta})$ permet de tester l'hypothèse H_0 que la distribution considérée représente adéquatement les observations contre l'hypothèse alternative H_1. La région critique de ce test est de la forme $\{X^2(\underline{\theta}) \geq c_\alpha\}$. Comme on l'a vu à la Section 5.1, la valeur critique c_α dépend de la loi de la statistique $X^2(\underline{\theta})$ et du niveau de signification α que l'on se fixe a priori.

La loi exacte de la statistique $X^2(\underline{\theta})$ est difficile à déterminer mais nous pouvons remarquer, à partir de l'expression (5.14), que cette loi dépend, sous l'hypothèse nulle H_0, du nombre M de classes. Il en est aussi de même pour la loi asymptotique ($n \to +\infty$) de cette statistique, qui est plus facilement identifiable. On peut en effet montrer que, lorsque les paramètres de la loi sont connus, la statistique $X^2(\underline{\theta})$ a pour loi asymptotique la loi de khi-deux à M - 1 degrés de liberté.

En pratique, particulièrement en hydrologie, il est très rare de pouvoir spécifier entièrement a priori la loi envisagée dans l'hypothèse nulle. En effet, tel que mentionné au chapitre 3, on n'a aucune information permettant d'attribuer une valeur précise aux paramètres de cette loi. C'est pourquoi, en pratique, on utilise plutôt une modification du test du khi-deux proposée aussi par Pearson. Il suffit d'estimer les paramètres de la loi $F(x;\underline{\theta})$, et de substituer dans les équations (5.12) et (5.14) le vecteur de paramètres $\underline{\theta} = (\theta_1, \theta_2, \ldots, \theta_p)$ par l'estimation obtenue $\underline{\hat{\theta}} = (\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_p)$. La statistique du test devient alors :

$$X^2\left(\underline{\hat{\theta}}\right) = \sum_{i=1}^{M} \frac{\left(N_i - np_i(\underline{\hat{\theta}})\right)^2}{np_i(\underline{\hat{\theta}})} \tag{5.15}$$

Ici, la loi asymptotique de la statistique $X^2\left(\underline{\hat{\theta}}\right)$ dépend non seulement du nombre M de classes mais aussi de la méthode d'estimation choisie pour déterminer $\underline{\hat{\theta}} = \left(\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_p\right)$. La méthode qui doit être utilisée est la méthode du maximum de vraisemblance. On peut montrer, dans ce cas, qu'une bonne approximation de la loi asymptotique de la statistique est donnée par la loi de khi-deux à M-p-1 degrés de liberté, p étant le nombre de paramètres estimés. Les détails concernant la détermination de la loi asymptotique de cette statistique sont donnés dans D'Agostino et Stephens (1986).

Un point faible des tests du khi-deux est la subjectivité introduite par la nécessité de diviser le domaine de la variable aléatoire. En effet, pour appliquer ce test, on doit choisir un nombre de classes et ensuite les construire. Le choix de ces classes est guidé par deux considérations : la puissance du test (probabilité de rejeter l'hypothèse nulle alors que cette hypothèse est fausse) et l'utilisation de la loi asymptotique pour approximer la loi exacte de la statistique pour une taille d'échantillon finie.

Le choix du nombre de classes et du type de classes à utiliser pour le calcul de la statistique du khi-deux est un sujet qui a été beaucoup étudié, surtout lorsque la loi des observations est entièrement spécifiée (paramètres connus). Ainsi, Mann et Wald (1942) ont montré qu'il était préférable, lorsque les paramètres de la loi sont connus, de prendre des classes équiprobables pour cette loi. Cohen et Sackrovitz (1975) ont montré que, dans ce cas, le test du khi-deux est non-biaisé, c'est-à-dire que le niveau de signification réel est égal au niveau de signification fixé a priori. Or, comme nous l'avons vu, les paramètres ne sont généralement pas connus pour le phénomène étudié et on doit plutôt utiliser le test du khi-deux modifié (équation 5.15). Si l'on veut étendre à ce cas la règle des classes équiprobables, on peut les construire de la façon suivante. On estime le vecteur de paramètres $\underline{\theta} = \left(\theta_1, \theta_2, \ldots, \theta_p\right)$ par la méthode du maximum de vraisemblance et on choisit M classes C_i de telle sorte qu'elles aient des probabilités égales lorsque celles-ci sont calculées avec la loi spécifiée dans l'hypothèse nulle, les paramètres étant remplacés par $\underline{\hat{\theta}} = \left(\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_p\right)$. Nous remarquons alors que les classes C_i ne sont plus fixées, mais aléatoires, puisqu'elles dépendent des estimateurs $\underline{\hat{\theta}} = \left(\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_p\right)$. Une des questions soulevées par ce choix de classes (non fixées) est de savoir si la loi asymptotique de la statistique (khi-deux à M-p-1 degrés de liberté) est toujours valable. Roy (1956) et Watson

(1957, 1958 et 1959) ont étudié ce problème empiriquement et selon leurs résultats, il est admis en pratique que cette loi est adéquate pour effectuer le test.

Le choix du nombre M de classes est dicté principalement par des considérations de puissance du test. De nombreuses études (entre autres, Mann et Wald, 1942) ont montré, dans le cas de classes aléatoires équiprobables, qu'une règle empirique généralement admise consiste à prendre M au plus égal à :

$$M = 4\left\{\frac{2n^2}{u(\alpha)}\right\} \tag{5.15}$$

où $u(\alpha)$ est le quantile de probabilité au non-dépassement égale à $1-\alpha$ de la loi normale centrée-réduite. En particulier, Schorr (1974) montre par simulation, que pour un niveau de signification $\alpha = 5\%$, le nombre de classes M optimal est donné par $\left[2n^{2/5}\right]$, où [.] désigne "la partie entière de". Plus de détails sur ce sujet, sont donnés par D'Agostino et Stephens (1986, Chapitre 3).

Le test du khi-deux est disponible dans le logiciel pour toutes les lois à l'exception des lois normale et log-normale, pour lesquelles des tests d'adéquation plus efficaces ont été développés. Nous décrivons, dans ce qui suit, le test de Shapiro-Wilk et le test des moments empiriques adaptés à la loi normale. Pour la loi log-normale, il suffit d'appliquer ces tests aux observations transformées en logarithme naturel.

5.3.2 Test de Shapiro-Wilk ($n \leq 50$)

Pour la loi normale, le test de Shapiro-Wilk est plus puissant que tout autre test connu lorsque la taille de l'échantillon est inférieure à 50 observations. Considérons l'échantillon aléatoire ordonné $X_{(1)} \leq X_{(2)} \leq \ldots \leq X_{(n)}$ de taille n que l'on suppose distribué selon une loi normale de paramètres μ et σ^2. Soient $Z_{(1)} \leq Z_{(2)} \leq \ldots \leq Z_{(n)}$ les variables standardisées correspondantes. On a donc que:

$$X_{(i)} = \mu + \sigma Z_{(i)}, \quad i = 1, \ldots, n \tag{5.16}$$

et que :

$$E\{X_{(i)}\} = \mu + \sigma E\{Z_{(i)}\}, \quad i = 1, \ldots, n \tag{5.17}$$

Les valeurs $E\{Z_{(i)}\}$ étant des constantes que l'on peut calculer ou estimer (elles sont tabulées dans Harter, 1961), un graphique des observations ordonnées $x_{(i)}$ en fonction des espérances mathématiques $E\{Z_{(i)}\}$ devrait être approximativement linéaire avec une

ordonnée à l'origine μ et une pente σ si elles proviennent bel et bien d'une loi normale. C'est l'idée permettant de construire le papier de probabilité normal.

D'un tel modèle linéaire ($x_{(i)} = \mu + \sigma E\{Z_{(i)}\}$), on peut déterminer un estimateur $\tilde{\sigma}$ de σ par la méthode des moindres carrés comme en régression linéaire simple. Cet estimateur est appelé l'estimateur BLUE (Best Linear Unbiased Estimator) et est déduit du théorème de Gauss-Markov (Bickel et Docksum, 1977). Ce modèle linéaire et l'estimateur $\tilde{\sigma}$ sont à la base du test de Shapiro-Wilk.

Shapiro et Wilk (1965) ont proposé, pour tester la normalité d'un échantillon, de comparer l'estimation $\tilde{\sigma}$ de σ obtenue par le modèle linéaire à l'estimateur du maximum de vraisemblance $\hat{\sigma}$ donné à l'Exemple 3.1. La statistique SW proposée par ces auteurs est donnée par :

$$SW = K\frac{\tilde{\sigma}^2}{\hat{\sigma}^2} \tag{5.18}$$

où K est une constante qui dépend des espérances $E\{Z_{(i)}\}$ et des covariances $Cov\{Z_{(i)}, Z_{(j)}\}$ des variables $Z_{(i)}$. Cette statistique peut aussi s'écrire sous la forme :

$$SW = \frac{1}{n\hat{\sigma}^2}\left[\sum_{i=1}^{n} a_i\, x_{(i)}\right]^2 \tag{5.19}$$

Sous l'hypothèse "H_0 : les observations proviennent d'une loi normale", $\tilde{\sigma}^2$ et $\hat{\sigma}^2$ sont deux estimateurs du même paramètre σ^2. Ainsi, à une constante de normalisation près K (cette constante a été choisie de sorte que $0 < SW < 1$), la statistique SW est le rapport de deux estimateurs de σ^2 lorsque la loi des observations est la loi normale. On peut donc s'attendre à ce que cette statistique prenne des valeurs proches de 1 lorsque l'échantillon provient d'une loi normale, et des petites valeurs autrement. C'est ce que Shapiro et Wilk ont constaté à partir de nombreuses simulations. La statistique SW peut aussi être interprétée comme un coefficient de détermination R^2 mesurant la qualité de l'ajustement linéaire des données sur un papier de probabilité normal.

Pour tester l'hypothèse "H_0 : les observations proviennent d'une loi normale" contre l'hypothèse alternative H_1, on calcule donc SW à partir de l'échantillon et on rejette l'hypothèse nulle à un seuil de signification α si $SW \le c_\alpha$. Les valeurs optimales des coefficients a_i ont été calculées par Shapiro et Wilk (1965) pour $2 \le n \le 50$ et sont données

dans la Table A.1 de l'annexe A. La Table A.2 de cette annexe donne les valeurs critiques c_α correspondantes pour $\alpha = 1\%, 5\%$.

5.3.3 Tests des moments empiriques

La forme de la loi normale de paramètres μ et σ^2 est entièrement caractérisée par les moments centrés d'ordre 3 et 4. Plus précisément, on a pour cette distribution $Cs = 0$ et $Ck = 3$. Le coefficient *Cs*, comme on l'a vu à la Section 2.2, caractérise l'asymétrie d'une loi. Si une distribution est symétrique par rapport à la moyenne, comme la loi normale, $Cs = 0$. Des valeurs de *Cs* différentes de zéro indiquent l'asymétrie et donc la non-normalité. Le coefficient *Ck* caractérise le poids aux extrémités des lois. Des valeurs de *Ck* différentes de 3 indiquent la non-normalité. Les valeurs de *Ck* supérieures à 3 sont obtenues pour des lois aux extrémités plus lourdes (heavy tails) que la loi normale, alors que des valeurs de *Ck* inférieures à 3 sont associées à des lois aux extrémités plus légères (thin tails).

Il est donc naturel de vouloir détecter une déviation par rapport à la normalité en testant si la loi des observations possède les caractéristiques $Cs = 0$ et $Ck = 3$. Nous décrivons brièvement dans ce qui suit deux tests, l'un pour la symétrie, l'autre pour l'aplatissement, qui sont effectués simultanément pour tester l'adéquation de la loi normale lorsque la taille d'échantillon est supérieure à 50 observations.

Tout d'abord, rappelons que les coefficients d'asymétrie *Cs* et d'aplatissement *Ck*, peuvent être estimés à partir des moments de l'échantillon $x_1, x_2, x_3, \ldots, x_n$ par $\hat{C}s$ et $\hat{C}k$ (équation 3.3). De plus, lorsque les observations proviennent d'une loi normale, on peut montrer (Kendall et Stuart, 1987) que :

$$E\left\{\hat{C}s\right\} = 0 \quad \text{et} \quad Var\left\{\hat{C}s\right\} = \frac{6(n-2)}{(n+1)(n+3)} \tag{5.20}$$

et que

$$E\left\{\hat{C}k\right\} = \frac{3(n-1)}{(n+1)} \quad \text{et} \quad Var\left\{\hat{C}k\right\} = \frac{24n(n-2)(n-3)}{(n+1)^2(n+3)(n+5)} \tag{5.21}$$

Le premier test est construit à partir de la statistique $\hat{C}s$. Pour tester l'hypothèse "H_0 : les observations proviennent d'une loi normale" contre sa négation, on utilise une région critique de la forme $\left\{|\hat{C}s| \geq c_\alpha\right\}$. Ceci revient en fait à tester si la loi des observations est symétrique. La loi exacte de $\hat{C}s$ est difficile à obtenir. D'Agostino et Tietjen (1973) ont étudié ce problème et ont suggéré d'approcher la loi exacte de cette statistique sous l'hypothèse nulle

par une loi de Student. Pour ce faire, ils utilisent une statistique T_1 modifiée. Plus précisément, si nous posons

$$\beta = \frac{3(n^2+27n-70)(n+1)(n+3)}{(n-2)(n+5)(n+7)(n+9)} \quad \text{et} \quad \nu = \frac{4\beta-6}{\beta-3} \tag{5.22}$$

alors la statistique

$$T_1 = \frac{\hat{C}s}{\sqrt{Var\{\hat{C}s\}}}\left(\frac{\nu}{\nu-2}\right)^{1/2} \tag{5.23}$$

suit approximativement une loi de Student à $\|[\nu]\|$ (valeur absolue de la valeur entière de ν) degrés de liberté sous l'hypothèse nulle. Cette approximation est applicable pour des tailles d'échantillon au moins supérieures à huit observations. Elle est plus efficace que l'approximation normale classique qui consiste tout simplement à diviser le coefficient d'asymétrie par son écart-type et à comparer la valeur obtenue au quantile de la loi normale standardisée (celle-ci nécessite des échantillons de tailles supérieures à 150 observations, D'Agostino et Stephens, 1986).

Le deuxième test est construit à partir de la statistique $\hat{C}k$. Pour tester l'hypothèse "H_0 : les observations proviennent d'une loi normale" contre sa négation, on utilise une région critique de la forme $\{\hat{C}k \le c'_\alpha\} \cup \{\hat{C}k \ge c''_\alpha\}$ car la loi exacte de $\hat{C}k$ sous l'hypothèse nulle n'est pas symétrique (D'Agostino et Stephens, 1986). Ceci revient en fait à tester si l'aplatissement de la loi des observations est différent de celui de la loi normale. Anscombe et Glynn (1983) ont proposé d'effectuer le test en utilisant une fonction T_2 de $\hat{C}k$ dont la loi exacte peut être approchée adéquatement par la loi normale centrée-réduite. Cette fonction est définie de la façon suivante. Considérons le coefficient d'aplatissement standardisé :

$$Z = \frac{\hat{C}k - E\{\hat{C}k\}}{\sqrt{Var\{\hat{C}k\}}} \tag{5.24}$$

et posons :

$$A = \frac{6(n^2-5n+2)}{(n+7)(n+9)}\left[\frac{6(n+3)(n+5)}{n(n-2)(n-3)}\right]^{1/2}$$
$$B = 6 + \frac{8}{A}\left[\frac{2}{A} + \left(1+\frac{4}{A^2}\right)^{1/2}\right] \tag{5.25}$$

Nous pouvons alors admettre que la loi de la statistique T_2 définie comme suit :

$$T_2 = \sqrt{\frac{2}{9B}}\left\{\left(1-\frac{2}{9B}\right)-\left[\frac{1-\frac{2}{B}}{1+Z\sqrt{\frac{2}{B-4}}}\right]^{1/3}\right\} \qquad (5.26)$$

peut être approchée par la loi normale centrée-réduite. Ainsi, une fois la statistique T_2 calculée, on rejette ou l'on accepte l'hypothèse H_0 en se référant à une table de la loi normale. Par exemple, pour un seuil de signification de 5%, on rejette la normalité si $|T_2| \geq 1,96$. Cette approximation peut être utilisée efficacement pour $n > 20$ et est de loin plus efficace que l'approximation normale classique (équation 5.24) qui nécessite des tailles supérieures à 1000 observations pour donner des résultats valides (D'Agostino et Stephens, 1986).

5.4 Tests de discordance

Les résultats d'une analyse hydrologique de fréquence sont grandement influencés par les observations extrêmes d'un échantillon (plus grande ou plus petite observation). Ces données mesurées, que l'on appelle **données singulières** à cause de leur caractère distinct, peuvent être réelles et il est alors important de les conserver dans l'analyse. Toutefois, celles-ci peuvent très bien être le résultat d'erreurs lors de l'acquisition des données. Elles sont alors **aberrantes,** et il faut les rejeter de l'échantillon puisqu'elles ne sont pas représentatives du phénomène étudié et peuvent conduire à des résultats inadéquats. Il est donc important, lors d'une analyse hydrologique de fréquence, de détecter les données singulières. On doit ensuite examiner, à partir de considérations hydrologiques, si les données singulières identifiées sont aberrantes.

Une donnée singulière, dans le cadre d'une analyse hydrologique de fréquence, est une observation extrême de l'échantillon qui ne semble pas provenir de la loi envisagée. Les tests permettant de détecter une telle observation sont les **tests de discordance**. Pour chacune des lois de ***AJUSTE-II***, à l'exception des lois de Halphen, deux tests de discordance sont disponibles, l'un concernant la plus petite observation de l'échantillon, l'autre concernant la plus grande donnée. Le principe général d'un test de discordance est le suivant.

Considérons un échantillon aléatoire de taille n $X_1, X_2, X_3, \ldots, X_n$ distribué selon une loi de probabilité $F(x; \underline{\theta})$, et les variables aléatoires correspondantes $X_{(1)} \leq X_{(2)} \leq X_{(3)} \leq \ldots \leq X_{(n)}$

rangé par ordre croissant. Les valeurs extrêmes, et potentiellement singulières, sont donc $X_{(1)}$ et $X_{(n)}$. Un test de discordance est défini par une statistique qui vérifie, à un niveau de signification donné, si l'une des deux variables aléatoires $X_{(1)}$ et $X_{(n)}$ est incompatible avec le modèle de loi de probabilité $F(x; \underline{\theta})$. Si c'est le cas, on dit respectivement que $X_{(1)}$ est une **donnée singulière inférieure** et $X_{(n)}$ est une **donnée singulière supérieure**. Les hypothèses testées sont alors, pour $X_{(1)}$:

$$H_0 : X_1, X_2, \ldots, X_n \in F(x; \underline{\theta}) \quad \text{contre} \quad H_1 : X_{(1)} \notin F(x; \underline{\theta})$$

et pour $X_{(n)}$:

$$H_0 : X_1, X_2, \ldots, X_n \in F(x; \underline{\theta}) \quad \text{contre} \quad H_1 : X_{(n)} \notin F(x; \underline{\theta})$$

Il est important de noter que la loi asymptotique des statistiques des tests de discordance est très difficile à obtenir. En général, nous ne disposons que d'une loi approximative. Ainsi, les niveaux de signification observés $\hat{\alpha}$ donnés dans le logiciel *AJUSTE-II* sont des valeurs approchées. Souvent seule une borne supérieure ou inférieure du niveau de signification peut être calculée. Dans certains cas, aucune conclusion n'est possible lorsqu'on effectue un test de discordance.

Enfin, la détection d'une donnée singulière à l'aide d'un test de discordance ne justifie pas nécessairement le rejet de celle-ci. On doit aussi condidérer la physique de l'événement singulier. Le test de discordance est un outil statistique que l'hydrologue doit utiliser comme indicateur. Pour plus de détails concernant les tests de discordance, on peut se référer à Huitorel *et al.* (1992) et Barnett et Lewis (1984).

6 APPLICATION

Ce chapitre présente une application complète de l'analyse hydrologique de fréquence effectuée à l'aide du logiciel ***AJUSTE-II***. Cet exemple peut servir de guide général pour un utilisateur de ***AJUSTE-II*** quoique que toute analyse statistique comme l'AHF comporte des problèmes particuliers reliés à l'ensemble de données traité.

On considère, dans la présente application de l'AHF, le dimensionnement d'un barrage sur la rivière Harricana. L'événement extrême qui nous intéresse est dans ce cas le débit de crue puisque celui-ci pourrait engendrer une défaillance de cet ouvrage hydroélectrique (débordement, etc.). Pour évaluer le risque de défaillance, on estime les débits maximums annuels de période de retour T tels que définis au chapitre 2 (équation 2.37). L'échantillon est constitué des débits maximums annuels historiques mesurés à la rivière Harricana de 1915 à 1983. La Figure 4.1 donne le graphique des débits maximums annuels observés à chacune des années.

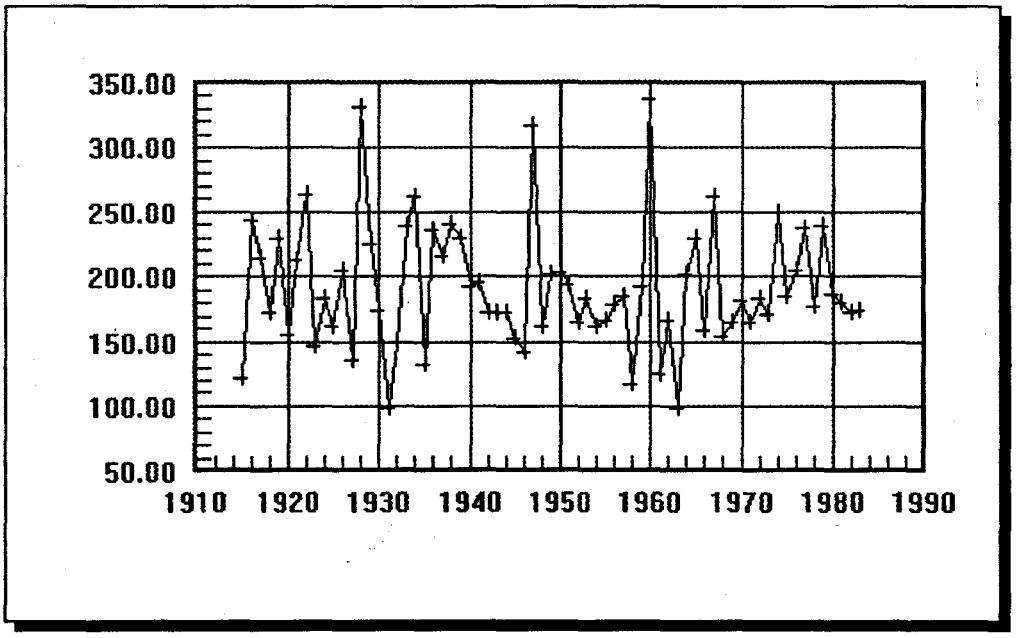


Figure 4.1. Série chronologique des débits maximums annuels de la rivière Harricana

Avant d'effectuer le choix du modèle de loi de probabilité et l'estimation des quantiles, on doit vérifier les hypothèses de base que sont l'indépendance, l'homogénéité et la stationnarité des observations de l'échantillon (cf. chapitre 5). Nous avons choisi d'effectuer ici tous les

tests d'hypothèse à un niveau de signification de 5% puisque nous jugeons qu'une telle probabilité de rejeter à tort l'hypothèse nulle est acceptable dans cette étude.

En général, on peut considérer que les débits maximums annuels sont indépendants entre eux. En effet, aucune raison évidente nous permet d'affirmer que la probabilité d'occurrence d'un débit de crue est influencée par l'intensité de la crue de l'année précédente. Toutefois, étant donné que nous ne pouvons connaître exactement tous les facteurs naturels régissant le phénomène étudié, on effectue quand même un test statistique. La Figure 4.2 donne le résultat du test de Wald-Wolfowitz qui vérifie si les données sont indépendantes.

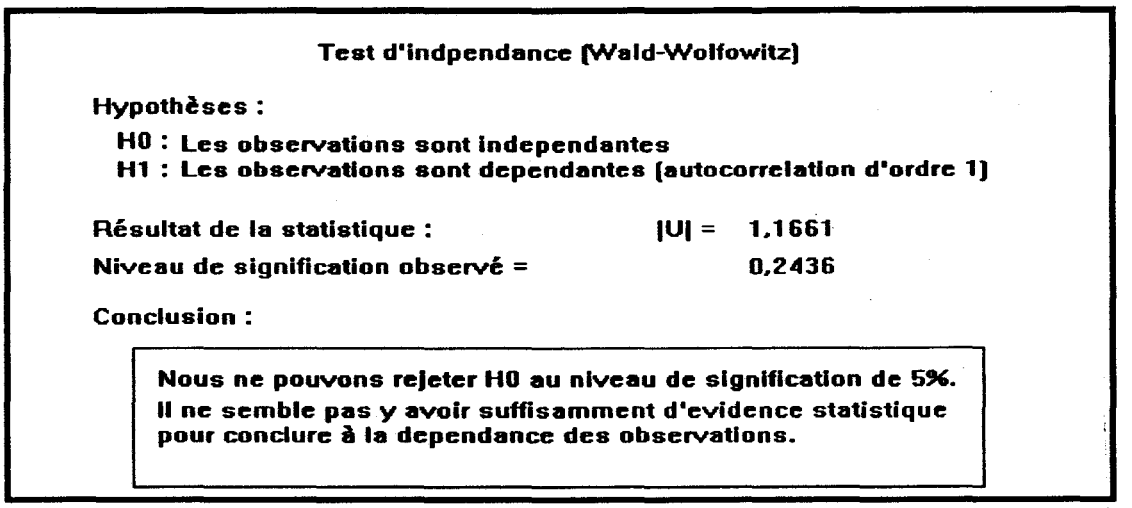


Figure 4.2. Résultat du test d'indépendance de Wald-Wolfowitz

Le niveau de signification observé 0,2436 étant supérieur à 5%, on ne peut rejeter l'hypothèse nulle et on considère que les débits maximums annuels de la rivière Harricana sont indépendants.

Des documents concernant les appareils de mesure indiquent que la jauge de la rivière Harricana a été déplacée à la fin de l'année 1972. Ce déplacement peut avoir modifié les caractéristiques statistiques de la série, rendant ainsi les observations hétérogènes. Il est donc important de tester l'homogénéité de la série en comparant les débits maximums annuels moyens avant et après le déplacement de l'appareil. Le test utilisé est le test de Wilcoxon présenté à la Section 5.2.2. La Figure 4.3 présente les résultats obtenus qui n'indiquent aucune différence significative entre les moyennes des deux sous-échantillons. Les observations sont alors statistiquement homogènes.

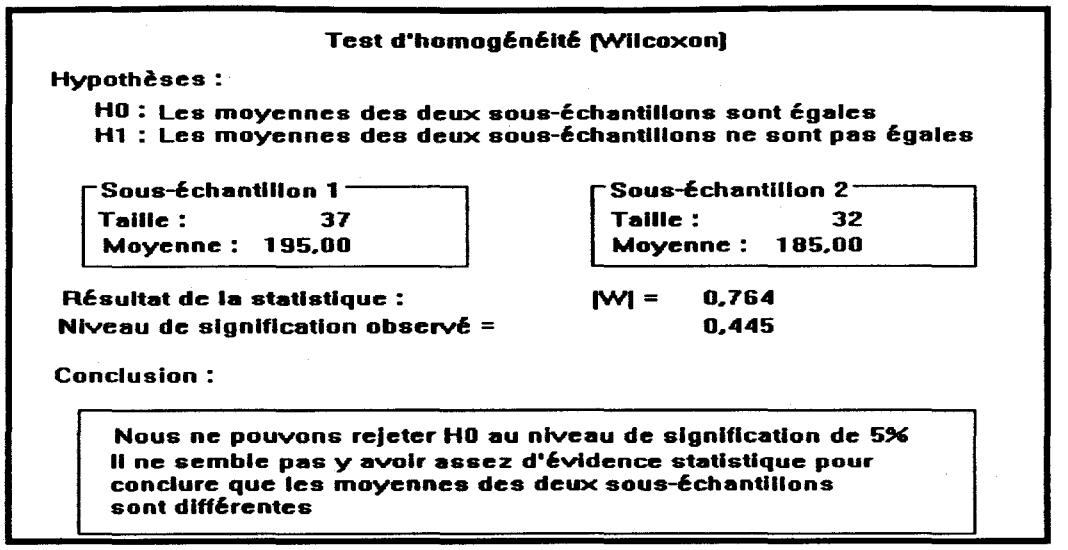


Figure 4.3. Résultats du test d'homogénéité de Wilcoxon.

L'examen de la Figure 4.1 ne nous permet pas d'identifier clairement un problème de non-stationnarité. En effet, il ne semble pas y avoir de tendances graduelles ou de sauts brusques dans la série d'observations. Il est toutefois préférable d'effectuer le test de Kendall pour s'en assurer. Les résultats de ce test sont présentés à la Figure 4.4. Les observations sont stationnaires au niveau de signification considéré puisque le niveau observé (0,7757) est de loin supérieur à 5%.

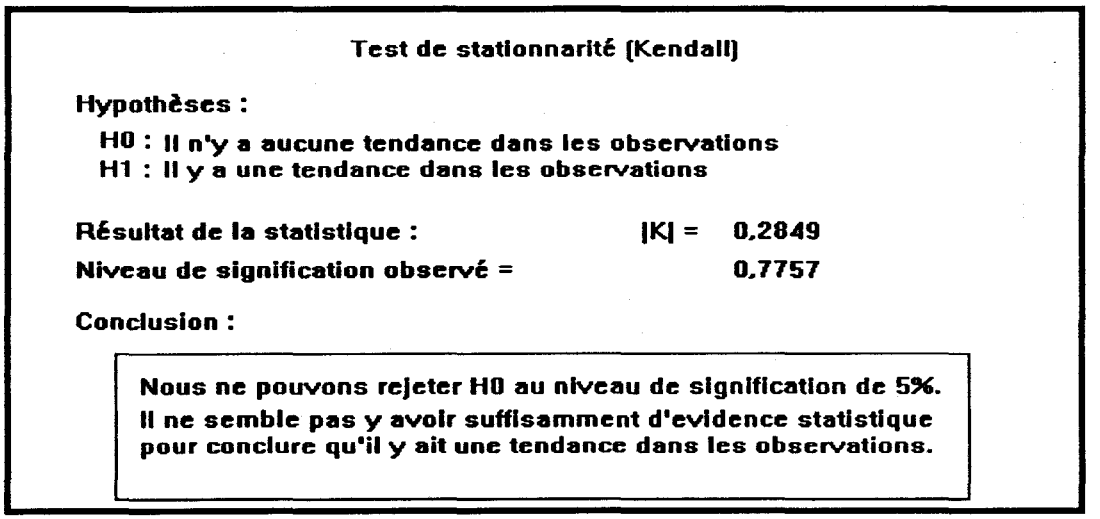


Figure 4.4. Résultat du test de stationnarité de Kendall.

Les débits maximums annuels de la rivière Harricana répondent donc, pour un niveau de signification de 5%, aux hypothèses de base de l'analyse hydrologique de fréquence. On

peut maintenant procéder à l'estimation des quantiles x_T de période de retour T à l'aide d'une loi de probabilité.

Pour aider l'usager à effectuer un choix a priori de la distribution, ou tout au moins à éliminer quelques lois qui seraient inadéquates pour représenter l'échantillon, le logiciel *AJUSTE-II* permet d'examiner l'histogramme des données. Ce graphique donne une indication de la forme de la loi de probabilité d'où proviennent les observations. La Figure 4.5 présente l'histogramme des débits maximums annuels de la rivière Harricana.

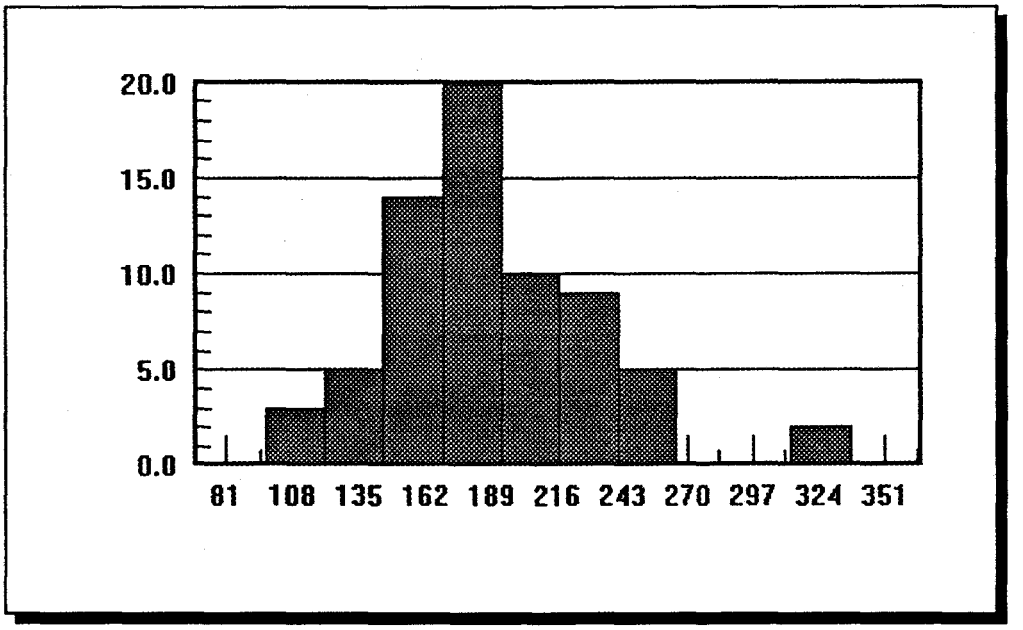


Figure 4.5. Histogramme des débits maximums annuels de la rivière Harricana

Cet histogramme permet d'observer que la distribution des données a la forme d'une loi à asymétrie positive puisqu'elle décroît plus lentement vers zéro pour des débits élevés que pour des petits débits (cf. Figure 2.6). Cette observation nous indique a priori qu'une distribution symétrique comme la loi normale n'est probablement pas un bon modèle pour l'échantillon Harricana. On remarque, de plus, que le mode de la distribution (classe de l'histogramme où la fréquence est la plus grande) se situe relativement loin de la plus petite valeur de l'échantillon. Ainsi, une loi de probabilité telle que la distribution exponentielle (cf. Figure 2.5) n'est guère adaptée aux débits maximums annuels de la rivière Harricana.

Étant donné que la loi normale est une distribution fort connue, simple à appliquer et que l'asymétrie de l'échantillon ne semble pas très importante, nous avons utilisé cette loi comme premier modèle. Les estimations sont obtenues à l'aide de la méthode du maximum de

vraisemblance (Section 3.3.1). La Figure 4.6 donnent les résultats tels présentés par le logiciel *AJUSTE-II*. On y retrouve :

- l'estimation des paramètres μ et σ de la loi normale;
- l'estimation des quantiles x_T correspondant à 21 périodes de retour T;
- l'écart-type asymptotique de chacune des estimations de x_T;
- l'intervalle de confiance asymptotique de niveau de confiance 95% pour chaque x_T.

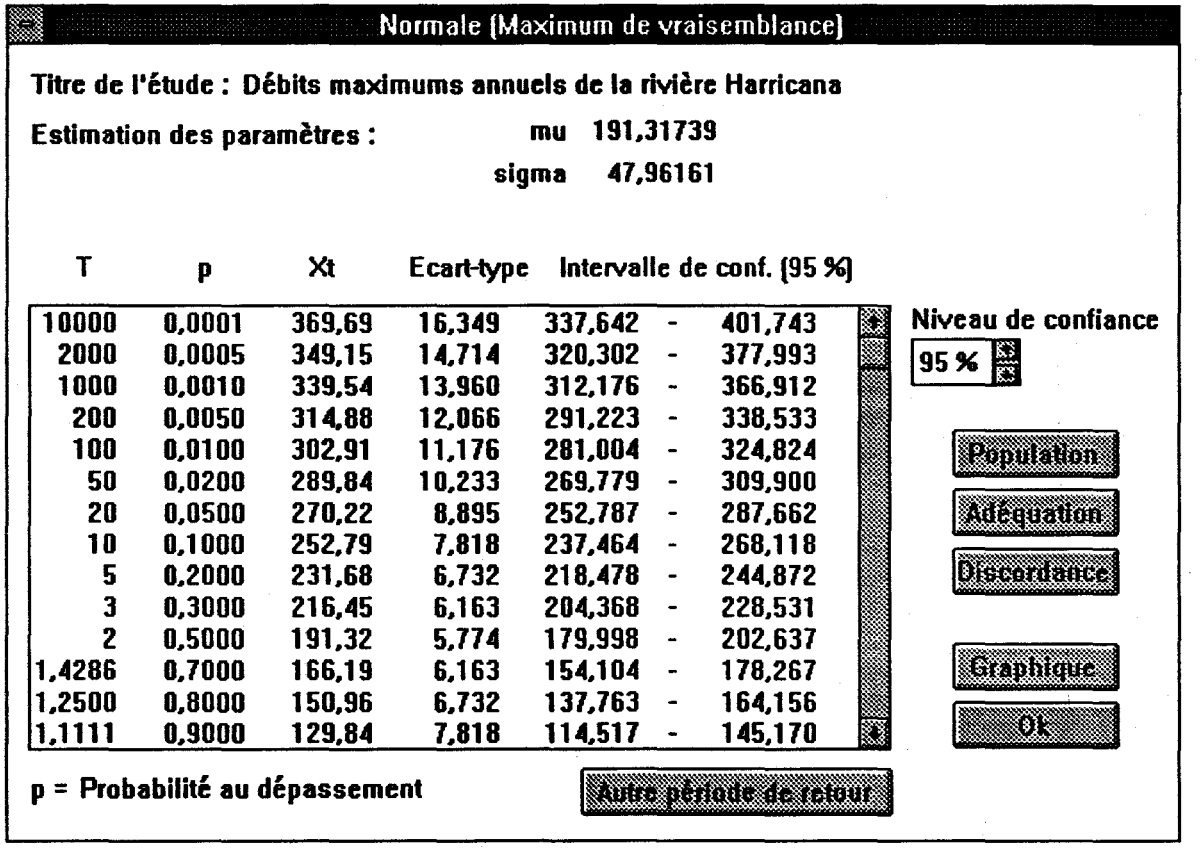


Figure 4.6. Résultats de l'ajustement de la loi normale aux données de la rivière Harricana.

Ainsi, selon le modèle de la loi normale, le débit maximum annuel centennal (de probabilité au dépassement $p = 0{,}01$) est estimé à 302,91 m^3s. De plus, ce débit théorique inconnu est compris dans l'intervalle [281,004 ; 324,824] avec une probabilité de 95%. Pour vérifier si la loi normale est une bonne approximation de la vraie loi des observations et que les résultats obtenus sont adéquats, nous effectuons un test d'adéquation (cf. Section 5.3). Les résultats sont présentés à la Figure 4.7.

Test d'adequation (moments empiriques)

Hypothèses :

H0 : L'echantillon provient d'une loi Normale
H1 : L'echantillon ne provient pas d'une loi Normale

	Pour Cs		Pour Ck	
Statistiques :	T1 =	**3,1714**	T2 =	**0,0173**
Probabilités au dép. :		**0,0015**		**0,9862**

Conclusion :

Nous DEVONS REJETER H0 au niveau de signification de 5%.
Il y a suffisamment d'evidence statistique pour conclure que les observations NE SUIVENT PAS une loi Normale.

Figure 4.7. Résultats du test d'adéquation pour la loi normale

Le test effectué est celui des moments empiriques puisque la taille de l'échantillon est supérieure à 50 observations. Comme on a vu à la Section 5.3.3, deux tests sont réalisés : l'un sur le coefficient d'asymétrie et l'autre sur le coefficient d'aplatissement. Selon le premier test (statistique T_1), on doit rejeter l'hypothèse de normalité des observations car le niveau de signification observé (0,0015) est inférieur à 5%. La loi normale n'est donc pas compatible avec les observations de la rivière Harricana et doit être rejetée pour représenter les données étudiées.

Nous devons maintenant envisager d'autres lois de probabilité afin de modéliser les débits maximums annuels de la rivière Harricana. Deux distributions ayant une asymétrie positive ont été considérées, les lois Gumbel et log-normale à deux paramètres. La méthode du maximum de vraisemblance a été employée pour estimer les paramètres de ces deux lois. Les résultats de l'ajustement et du test d'adéquation sont présentés aux Figures 4.8 et 4.9 respectivement pour les lois Gumbel et log-normale.

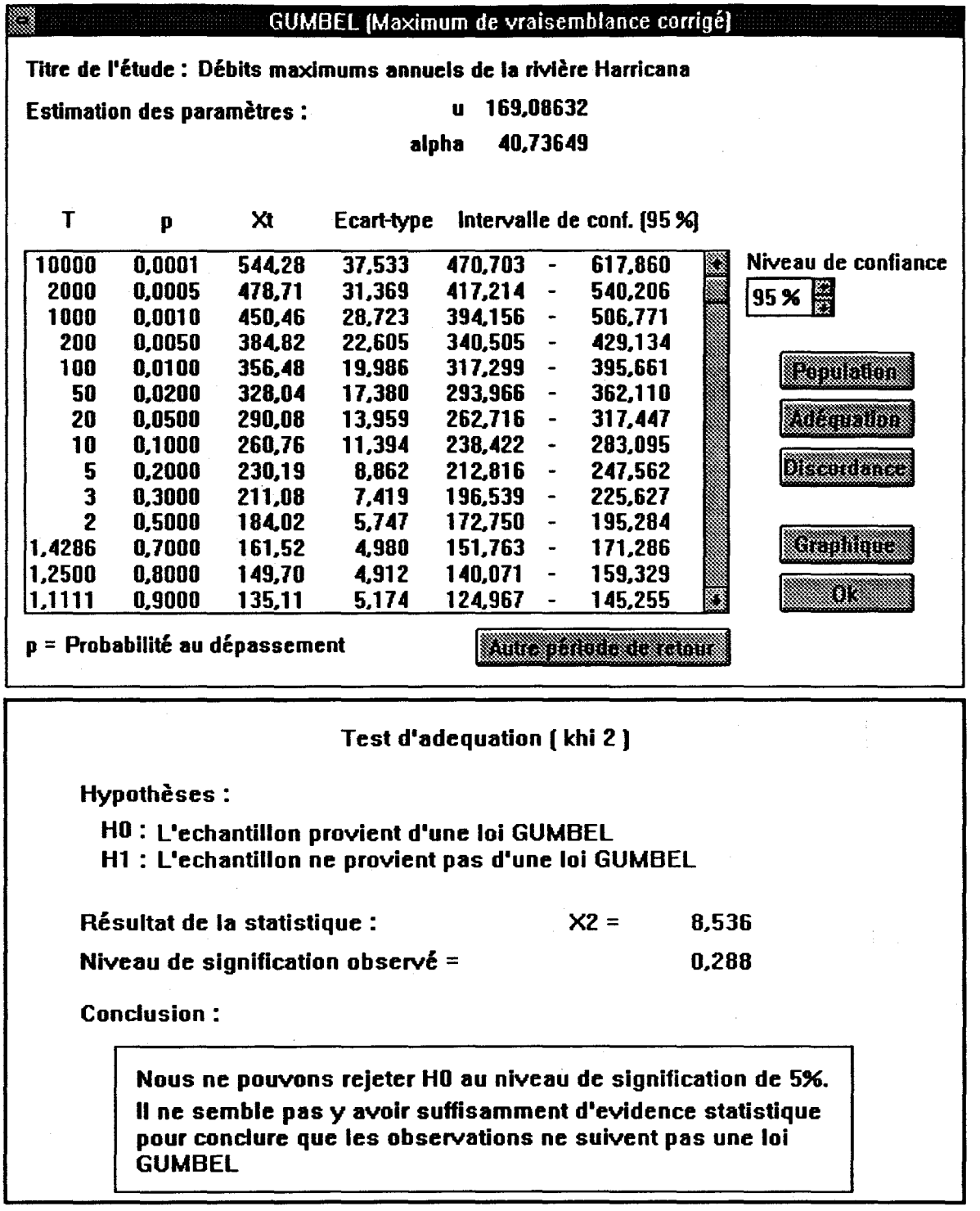
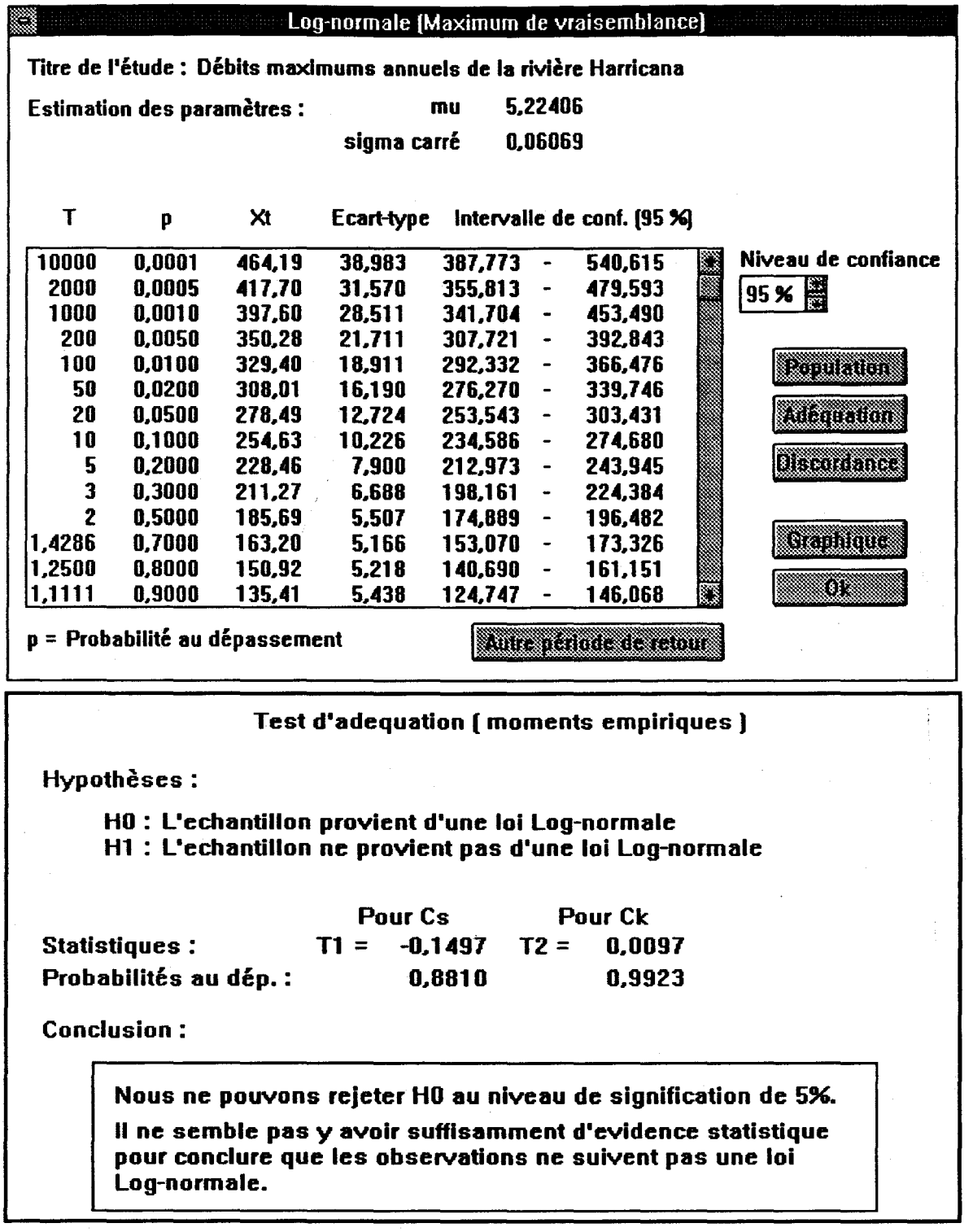


Figure 4.8. Résultats de l'ajustement et du test d'adéquation de la loi Gumbel.



Log-normale (Maximum de vraisemblance)

Titre de l'étude : Débits maximums annuels de la rivière Harricana

Estimation des paramètres : mu 5,22406
sigma carré 0,06069

T	p	Xt	Ecart-type	Intervalle de conf. (95 %)
10000	0,0001	464,19	38,983	387,773 - 540,615
2000	0,0005	417,70	31,570	355,813 - 479,593
1000	0,0010	397,60	28,511	341,704 - 453,490
200	0,0050	350,28	21,711	307,721 - 392,843
100	0,0100	329,40	18,911	292,332 - 366,476
50	0,0200	308,01	16,190	276,270 - 339,746
20	0,0500	278,49	12,724	253,543 - 303,431
10	0,1000	254,63	10,226	234,586 - 274,680
5	0,2000	228,46	7,900	212,973 - 243,945
3	0,3000	211,27	6,688	198,161 - 224,384
2	0,5000	185,69	5,507	174,889 - 196,482
1,4286	0,7000	163,20	5,166	153,070 - 173,326
1,2500	0,8000	150,92	5,218	140,690 - 161,151
1,1111	0,9000	135,41	5,438	124,747 - 146,068

Niveau de confiance 95 %

Population
Adéquation
Discordance
Graphique
Ok

p = Probabilité au dépassement

Autre période de retour

Test d'adequation (moments empiriques)

Hypothèses :

H0 : L'echantillon provient d'une loi Log-normale
H1 : L'echantillon ne provient pas d'une loi Log-normale

	Pour Cs	Pour Ck
Statistiques :	T1 = -0,1497	T2 = 0,0097
Probabilités au dép. :	0,8810	0,9923

Conclusion :

Nous ne pouvons rejeter H0 au niveau de signification de 5%.
Il ne semble pas y avoir suffisamment d'evidence statistique pour conclure que les observations ne suivent pas une loi Log-normale.

Figure 4.9. Résultats de l'ajustement et du test d'adéquation de la loi log-normale.

Si on se réfère aux tests d'adéquation, il n'y a pas assez d'évidence statistique pour rejeter l'une de ces deux lois de probabilité. En effet, les niveaux de significations observés sont dans les deux cas supérieurs à 5%.

On remarque, lorsqu'on examine de près les quantiles estimés $\hat{x}_T$, que la loi Gumbel fournit des estimations plus grandes que celle obtenues avec la loi log-normale, en particulier pour les grandes périodes de retour T. On peut mieux constater cette observation sur le graphique comparatif des deux ajustements présenté à la Figure 4.10. L'axe des abscisses correspond aux probabilités au non-dépassement ($1-p=1-1/T$) et l'axe des ordonnés aux débits maximums annuels. On retrouve sur cette figure les valeurs estimées des quantiles des deux lois reliées par un trait continu ainsi que les observations de l'échantillon représentées par une croix. Ce graphique nous permet d'apprécier l'adéquation de chacune des lois et aussi de comparer visuellement les deux ajustements.

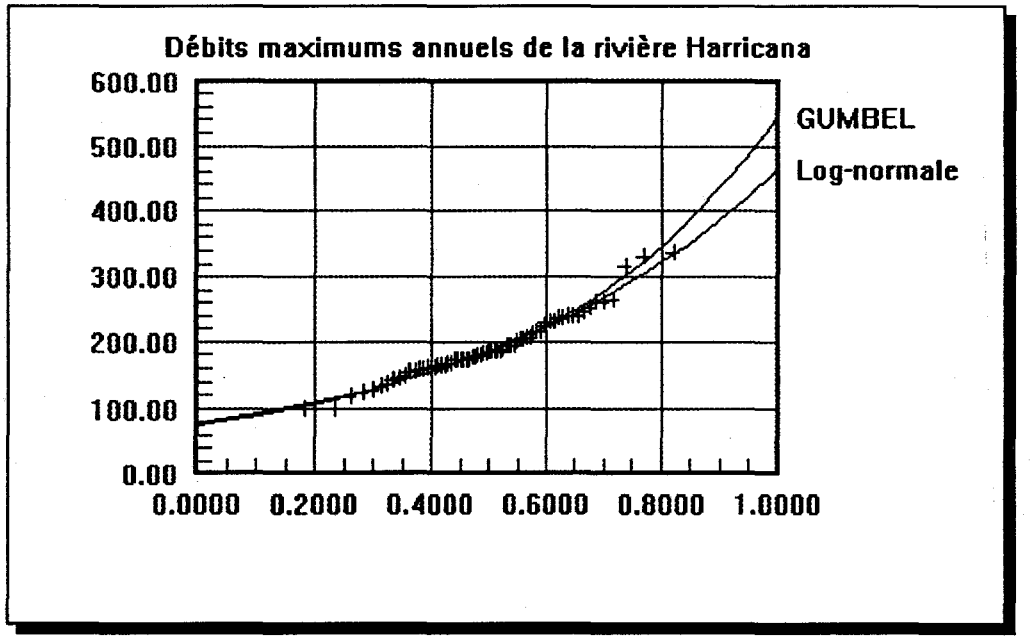


Figure 4.10. Comparaison des ajustements des lois Gumbel et log-normal.

La Figure 4.10 montre que ces lois de probabilité à deux paramètres représentent convenablement les observations de la rivière Harricana. Cette constatation nous amène à exclure les lois à trois paramètres (Pearson Type 3, log-Pearson Type 3, Halphen, GEV, etc.). En effet, le gain en adéquation obtenu par l'ajout d'un paramètre n'est probablement pas important par rapport à l'erreur d'échantillonnage supplémentaire introduite en estimant ce paramètre additionnel.

Il est difficile d'arrêter notre choix à l'une des deux distributions. Étant données la variance des quantiles estimés et les erreurs de mesure toujours présentes dans ce type de données hydrologiques, les deux ajustements ne sont pas significativement différents. En effet, on réalise en examinant de près les intervalles de confiance des quantiles x_T des deux lois de probabilité (Figures 4.8, 4.9, 4.11 et 4.12) que ceux-ci se recoupent systématiquement quelle que soit la période de retour T (ceci ne constitue pas un test statistique mais seulement une indication). De plus, on remarque que les écarts-types asymptotiques sont pratiquement identiques, ce qui indique que les estimations obtenues sont aussi précises pour les deux distributions considérées.

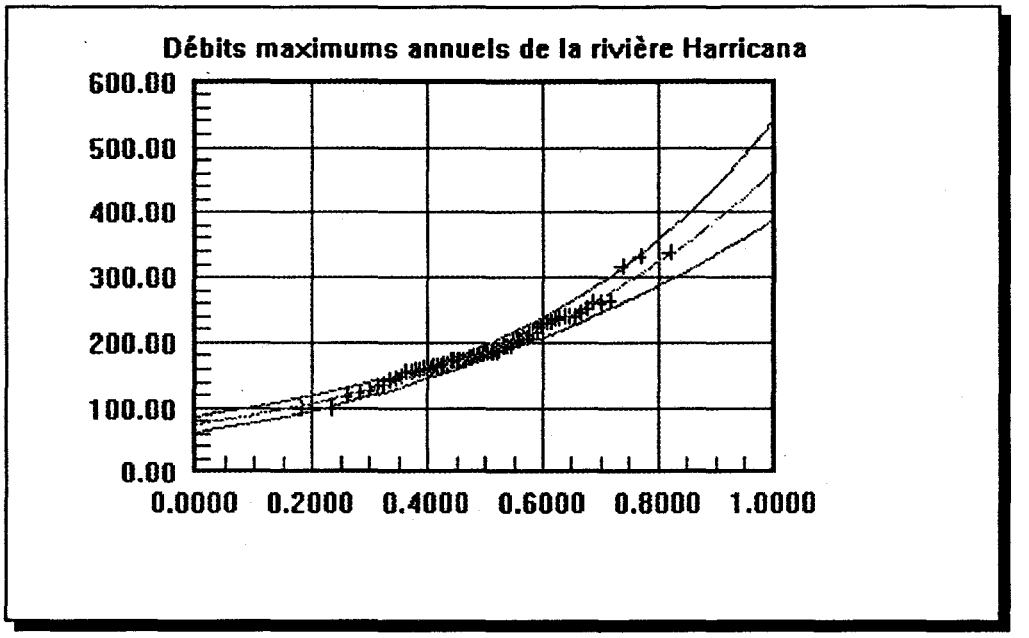


Figure 4.11. Ajustements de la loi log-normal accompagné des intervalles de confiance.

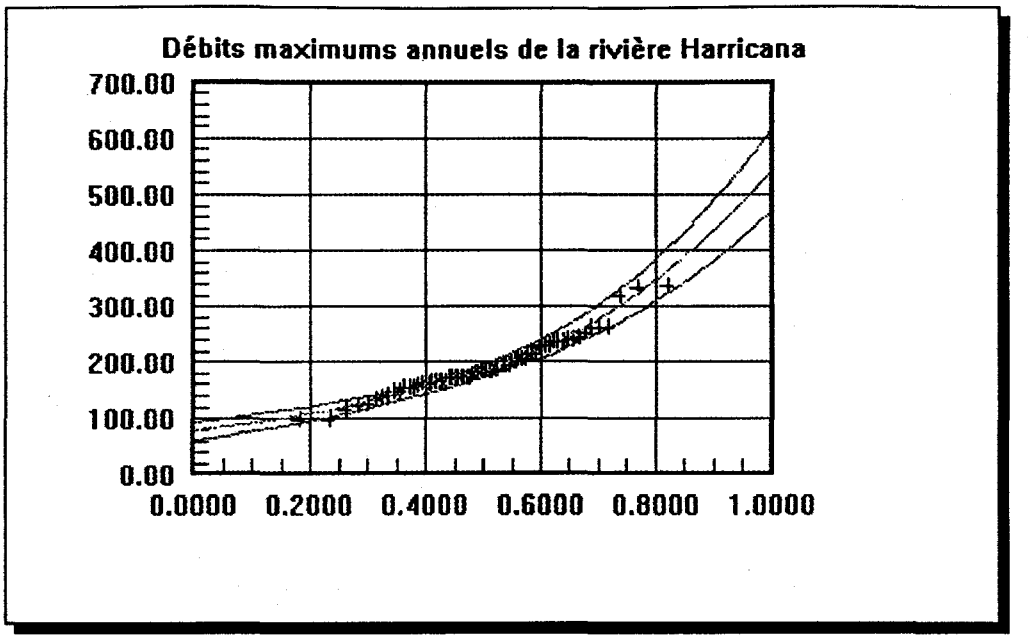


Figure 4.12. Ajustement de la loi Gumbel accompagné des intervalles de confiance.

Le choix final de la loi de probabilité, et donc des débits de période de retour que l'on utilisera pour aider à dimensionner le barrage, doit être, à ce stade-ci de l'étude, motivé par des considérations autres que statistiques. Par exemple, on pourrait choisir la loi Gumbel car elle fournit des quantiles estimés supérieurs à ceux de la loi log-normale. On augmente ainsi les coûts de construction de l'ouvrage, mais au profit d'une diminution du risque de défaillance (débordement, etc.) par rapport à la loi log-normale.

7 CONCLUSION ET RECOMMANDATIONS

7.1 Mises en garde

Ce manuel général présente les principales techniques statistiques qu'un spécialiste doit connaître pour utiliser adéquatement le logiciel ***AJUSTE-II*** et effectuer une analyse hydrologique de fréquence. Il est évidemment possible d'aller plus en profondeur sur le sujet et pour ce faire le lecteur peut consulter les ouvrages cités dans le texte. Dans ce manuel, nous avons mis l'accent sur le traitement des débits maximums ou minimums annuels. Toutefois, les procédures présentées peuvent très bien être appliquées à d'autres ensembles de données hydrologiques qui satisfont aux hypothèses de base (chapitre 5) notamment, les précipitations, les volumes, les températures, les mesures de qualité de l'eau, etc.

L'AHF, comme toute méthode statistique, a ses limites et ne peut être utilisée telle qu'elle seulement si toutes les hypothèses de base (indépendance, homogénéité et stationnarité) sont satisfaites. Lorsque les données ne sont pas indépendantes, le spécialiste doit éviter d'effectuer une AHF. Nous proposons, dans ce cas, d'utiliser des modèles de prévision qui prennent en compte cette caractéristique des données (modèle de série de temps, Box et Jenkins, 1976). Si les données sont indépendantes mais qu'une tendance brusque (changement de moyenne) est détectée à l'aide des tests de Wilcoxon ou de Kendall, nous suggérons d'effectuer deux AHF en traitant indépendamment les deux sous-échantillons de moyennes différentes. Outre les hypothèses de base, d'autres facteurs aussi importants doivent être pris en compte lors d'une analyse hydrologique de fréquence. Nous donnons dans ce qui suit quelques indications qui peuvent aider l'utilisateur de ***AJUSTE-II***.

7.1.1 Taille de l'échantillon

Pour la majorité des lois incluses dans le logiciel, les propriétés des estimateurs ne sont connues qu'asymptotiquement. Ainsi, le calcul des variances et des intervalles de confiance est valide seulement si la taille de l'échantillon est grande. Nous considérons qu'en général un échantillon d'au moins 30 observations est nécessaire pour interpréter tous les résultats d'une AHF. Cette taille d'échantillon minimale est recommandée par la plupart des statisticiens lorsqu'une méthode statistique comme l'AHF repose sur la théorie des grands

nombres. Évidemment, on peut tout de même traiter des échantillons de tailles inférieures à titre **indicatif**, mais il faut être conscient que la validité des résultats peut en être considérablement affectée. Nous suggérons, pour de faibles tailles d'échantillon ($n < 30$) :

- d'éviter d'utiliser des lois de probabilité possédant plus de deux paramètres de façon à minimiser l'erreur d'échantillonnage;
- de ne pas interpréter les quantiles estimés de grandes périodes de retour ($T > 50$ ou 100).

7.1.2 Méthode d'estimation

L'efficacité d'une méthode d'estimation dépend de la loi de probabilité, de la taille d'échantillon et de certaines caractéristiques de l'échantillon. Le choix d'une méthode est donc un exercice difficile à effectuer. Toutefois, de nombreuses études comparatives, autant empiriques que théoriques, ont été menées afin de déterminer dans quelles circonstances une méthode d'estimation est la plus efficace pour une loi donnée. Le choix des méthodes d'estimation incorporées dans le logiciel *AJUSTE-II* s'appuie en grande partie sur ces études. Nous vous invitons à consulter les ouvrages cités dans le Tableau 3.1. Il est important de noter que si dans le logiciel une seule méthode est disponible pour une loi donnée (par exemple, lois normale et exponentielle), c'est que celle-ci est optimale.

7.1.3 Tests d'adéquation

Les tests d'adéquation (chapitre 5), particulièrement le test du khi-deux, sont généralement **peu puissants**. En effet, ils possèdent une probabilité assez faible de rejeter l'hypothèse nulle lorsqu'il le faut. Aussi peut-il arriver qu'une loi donnée ne soit pas rejetée par le test d'hypothèses alors que ce modèle n'est pas tout à fait adéquat pour représenter les observations. Toutefois, si le test d'adéquation conclut à la non compatibilité du modèle, la loi considérée pourra alors être rejetée avec confiance en faveur d'une autre distribution. L'examen visuel des graphiques d'ajustement, comme ceux présentés au chapitre 6, est un bon complément aux tests d'adéquation.

7.1.4 Tests de discordance

Les tests de discordance permettent uniquement de vérifier si les observations extrêmes de l'échantillon (la plus grande ou la plus petite données) sont compatibles avec la **loi choisie**. Ce test ne peut être employé pour déterminer si une observation est aberrante

indépendamment du modèle choisi. En effet, l'analyse hydrologique de fréquence repose sur l'hypothèse de la connaissance a priori de la loi des observations et ainsi la notion de donnée singulière ou aberrante dans le cadre d'une telle méthode statistique est directement reliée au modèle. C'est pourquoi, un test distinct pour chaque loi a été incorporé dans *AJUSTE-II*.

7.2 Recherche future

De plus en plus d'hydrologues-statisticiens sont maintenant soucieux du fait que les hypothèses de base des modèles mathématiques sont susceptibles d'influencer considérablement les résultats d'une analyse statistique. En particulier, pour estimer les débits extrêmes de période de retour (quantiles), l'acceptation, sans trop d'investigation, de l'hypothèse usuelle d'une loi de probabilité bien fixée inquiète un nombre grandissant d'analystes dans le domaine. En effet, faire l'hypothèse que les observations suivent une loi stricte alors qu'en réalité elles peuvent provenir d'une autre loi, peut engendrer des conclusions erronées lors de l'estimation des quantiles. Pour pallier ce problème et tenir compte des incertitudes inhérentes au choix du modèle, on peut envisager l'emploi de méthodes s'appuyant sur des modèles moins restrictifs.

En statistique classique, ce problème d'incertitude de modèle a été étudié par plusieurs auteurs particulièrement pour l'estimation d'un paramètre de localisation (paramètre de tendance centrale). Ainsi, différentes méthodes d'estimation du paramètre de localisation reposant sur des modèles statistiques plus ou moins larges ont été proposés, notamment les estimateurs robustes.

Les estimateurs robustes pour un paramètre de localisation s'appuient sur un modèle plus large (Hampel, 1986, et Huber, 1981). On considère tout d'abord un modèle paramétrique et on souhaite tenir compte du fait que les observations peuvent très bien, pour diverses raisons, venir, non pas d'une loi précisée par ce modèle, mais plutôt d'une loi "assez proche" de celles du modèle paramétrique considéré. Les raisons de cette déviation possible du modèle peuvent être dues à des erreurs d'expérimentation, des erreurs de mesures, des perturbations aléatoires non prises en compte dans le modèle, etc.. Pour représenter cette déviation, on introduit un ensemble de lois "voisines" de celles spécifiées par le modèle paramétrique de la façon suivante.

Notons F une loi de l'ensemble correspondant au modèle paramétrique et Φ l'ensemble de toutes les lois continues. Alors, une loi "voisine" de F est une loi de l'ensemble

$$\{G\ ;\ \ G=(1-\varepsilon)F+\varepsilon H,\ \ H\in\Phi,\ \ \varepsilon\in[0,1[\}$$

L'approche robuste consiste, intuitivement, à chercher des solutions qui

- sont valides et ont une bonne efficacité pour F;
- restent relativement valides et efficaces si on s'éloigne un peu de F vers une loi voisine (petit ε);
- n'ont pas une validité et une efficacité nulles si on est très éloigné de F (ε voisin de 1).

Cette idée de robustesse pourrait être mise à profit lors de l'estimation des débits extrêmes de période de retour. D'ailleurs, Bernier (1991) propose une approche dont les fondements de base s'apparentent à ceux du modèle de robustesse. En effet, il utilise un modèle paramétrique, la loi Weibull, qu'il élargit afin de tenir compte des incertitudes liées au choix du modèle. La loi Weibull n'est pas considérée comme loi stricte et bien identifiée mais comme une approximation dans une famille de distributions plus large : la famille générée par une transformation de Box-Cox sur une variable de loi gamma généralisée logarithmique.

Ce sujet est une avenue de recherche que nous croyons intéressante à approfondir. Ce travail de recherche permettrait d'une part, de clarifier la notion de robustesse souvent mal utilisée dans la littérature en hydrologie, et d'autre part, de fournir une approche originale considérant non seulement les incertitudes d'échantillonnages (incertitudes liées à la taille d'échantillon et aux méthodes d'estimations), mais aussi les incertitudes inhérentes au choix de la loi de probabilité.

8 RÉFÉRENCES BIBLIOGRAPHIQUES

Aitchison, J. et J.A.C. Brown (1957). *The Lognormal Distribution.* Cambridge University Press, London : 176 p.

Anscombe, F.J. et W.J. Glynn (1983). Distribution of the kurtosis statistic b_2 for normal statistics. *Biometrika*, **70**: 227-234.

Barnett, V. et T. Lewis (1984). *Outliers in Statistical Data.* Wiley, New York : 463 p.

Benjamin, J.R. et C.A. Cornell (1970). *Probability, Statistics and Decision for Civil Engineers.* McGraw Hill Company.

Bernier, J. (1991). Bayesian analysis of robustness of models in water and environmental sciences. *OTAN/ASI on Risk and Reliability in Water Resources and Environmental Engineering*, Porto Carras, Greece, 18-28 May.

Bickel, P.J. et K.A. Docksum (1977). *Mathematical Statistics. Holden-Day*, Oakland : 492 p.

Bobée B. et F. Ashkar (1991). *The Gamma and Derived Distributions Applied in Hydrology.* Water Resources Publications, Littleton, Co., Yevjevitch (Ed.) : 202 p.

Bobée, B. (1975). The Log Pearson type 3 distribution and its application in hydrology. *Water Resources Research*, **11**(5): 681-689.

Bobée, B. (1988). The generalized method of moments applied to the log-Pearson 3 distribution. *Journal of Hydraulic Engineering, ASCE*, **114**(8): 899-909.

Bobée, B. et Robitaille (1975). Correction of bias in the estimation of the coefficient of skewness. *Water Resources Research*, **11**(6): 851-854.

Box, G.E.P. et G.M. Jenkins (1976). *Time Series Analysis: Forcasting and Control.* Holden-Day, San Francisco: 575 p.

Burden, R.L. and J.D. Faires (1981). *Numerical Analysis.* PWS-Kent, Boston: 729 p.

Cohen, A. et H.B. Sackrovitz (1975). Unbiasedness of the chi-square, likelihood ratio, and other goodness of fit tests for equal cell case. *Annals of Statistics*, **4**: 959-964.

D'Agostino, R.B. et M.A. Stephens (1986). *Goodness-of-fit Techniques.* Marcel-Dekker, New York.

D'Agostino, R.B. et G.L. Tietjen (1973). Approaches to the null distribution of $\sqrt{b_1}$. *Biometrika*, 60: 169-173.

Greenwood, J.A., Landwher, J.M., Matalas, N.C. and J.R. Wallis (1979). Probability weighted moments : Definition and relation to parameters of several distributions expressable in inverse form. *Water Resources Research*, **15**(5): 1049-1054.

Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J. et A. Stahel (1986). *Robust Statistics : The Approach Based on Influence Functions*. Wiley, New York.

Harter, H.L. (1961). Expected values of normal order statistics. *Biometrika*, **48**: 151-165.

Hazen, A. (1924). Discussion on "Theoritical frequency curves" by Foster. *Transactions ASCE*, **87**: 174-179.

Hosking, J.R.M., Wallis, J.R. and E.F. Wood (1985). Estimation of the generalized extreme-value distribution by the method of probability-weighted moments. *Technometrics*, **27**(3): 251-261.

Huber, P.J. (1981). *Robust Statistics*. Wiley, New York.

Huitorel, N., Perreault, L. et B. Bobée (1992). Tests de détection de données singulières pour quelques lois du logiciel AJUSTE-2. *INRS-Eau, rapport de recherche* n^{o} 360, 65 p.

Kendall, M.G. (1975). *Rank Correlation Methods*. Charles Griffin, London.

Kendall M.G. et A. Stuart (1987). *Advanced Theory of Statistics. Volume 1: Distribution Theory*. Oxford University Press, New York.

Kirby, W. (1974). Algebraic boundedness of sample statistics. *Water Resources Reshearch*, **10**(2): 220-222.

Kotz, S. et L.M. Johnson (1983). *Encyclopedia of Statistical Sciences, Vol. 4*. Wiley, New York.

Lehmann, E.L. (1975). *Nonparametrics: Statistical Methods Based of Ranks*. Holden-Day, Oakland.

Lehmann, E.L. (1983). *Theory of Point Estimation*. Wiley, New York.

Mann, H.B. et A. Wald (1942). On the choice of the number of class intervals in the application of the chi-square test. *Annals of Mathematical Statistics*, **13**: 306-317.

Mathier L. , Perreault, L., Bobée, B. et F. Ashkar (1992). The use of geometric and gamma-related distributions for frequency analysis of water deficit. *Stochastic Hydrology and Hydraulics*, **6**(4): 239-254.

Messaouidi, H. (1994). Thèse de doctorat, INRS-Eau. En préparation.

NERC (1975). *Flood Studies Report, Vol. 1, Hydrological Studies*. Natural Environment Research Council, London.

Perreault L. et B. Bobée (1992a). Loi généralisée des valeurs extrêmes. Propriétés mathématiques et statistiques. Estimation des paramètres et des quantiles X_T de période de retour T. *INRS-Eau, rapport de recherche* n° 350, 56 p.

Perreault L. et B. Bobée (1992b). Loi Weibull à deux paramètres. Propriétés mathématiques et statistiques. Estimation des paramètres et des quantiles X_T de période de retour T. *INRS-Eau, rapport de recherche* n° 351, 29 p.

Perreault, L., Bobée, B. et V. Fortin (1992c). Approximation des quantiles de la loi Pearson Type 3 standardisée par les polynômes de Tchebichef. *INRS-Eau, rapport de recherche* n° 346, 36 p.

Perreault, L. et B. Bobée (1992d). Loi normale: propriétés mathématiques et statistiques. Estimation des paramètres et des quantiles X_T de période de retour T. *INRS-Eau, rapport de recherche* n° 352, 18 p.

Perreault L. et B. Bobée (1994). Les lois de Halphen. Propriétés mathématiques et statistiques. Estimation des paramètres et des quantiles X_T de période de retour T. En préparation.

Perron H. (1994). Logiciel AJUSTE-II: manuel du programmeur. En préparation.

Perron H., Perreault L. et B. Bobée (1994) . Logiciel AJUSTE-II: guide de l'utilisateur. En préparation.

Roy, A.R. (1956). On chi-square statistics with variable intervals. *Technical Report* n° 1, Department of Statistics. Stanford University.

Shapiro, S.S. et M.B. Wilk (1965). An analysis of variance test for normality (complete samples). *Biometrika*, **52**: 591-611.

Wald, A. et J. Wolfowitz (1943). An exact test for randomness in the nonparametric case based on serial correlation. *Annals of Mathematical Statistics*, **14**: 378-388.

Wallis, J.R., Matalas, N.C. et J.R. Slack (1985). Just a moment. Water Resources Research, 10(2): 211-219.

Watson, G.S. (1957). The chi-square goodness-of-fit test for normal distributions. *Biometrika*, **44**: 336-348.

Watson, G.S. (1958). On chi-square goodness-of-fit tests for continuous distributions. *Journal of the Royal Statistical Association, B*, **20**: 44-61.

Watson, G.S. (1959). Some recent results in chi-square goodness-of-fit tests. *Biometrics*, **15**: 440-468.

WRC (1967). *Guidelines for Determining Flood Flow Frequency*. US Water Resources Council Hydrology Committee, Washingtion.

ANNEXE A

Coefficients et valeurs critiques du test de Shapiro-Wilk

COEFFICIENTS a_j

j \ n	2	3	4	5	6	7	8	9	10	n \ j
1	0,7071	0,7071	0,6872	0,6646	0,6431	0,6233	0,6052	0,5888	0,5739	1
2		0,0000	0,1677	0,2413	0,2806	0,3031	0,3164	0,3244	0,3291	2
3				0,0000	0,0875	0,1401	0,1743	0,1976	0,2141	3
4						0,0000	0,0561	0,0947	0,1224	4
5								0,0000	0,0399	5

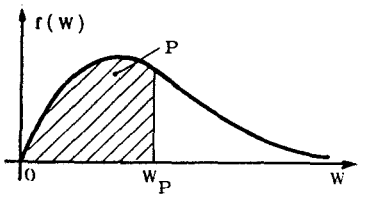
j \ n	11	12	13	14	15	16	17	18	19	20	n \ j
1	0,5601	0,5475	0,5359	0,5251	0,5150	0,5056	0,4963	0,4886	0,4808	0,4734	1
2	0,3315	0,3325	0,3325	0,3318	0,3306	0,3290	0,3273	0,3253	0,3232	0,3211	2
3	0,2260	0,2347	0,2412	0,2460	0,2495	0,2521	0,2540	0,2553	0,2561	0,2565	3
4	0,1429	0,1586	0,1707	0,1802	0,1878	0,1939	0,1988	0,2027	0,2059	0,2085	4
5	0,0695	0,0922	0,1099	0,1240	0,1353	0,1447	0,1524	0,1587	0,1641	0,1686	5
6	0,0000	0,0303	0,0539	0,0727	0,0880	0,1005	0,1109	0,1197	0,1271	0,1334	6
7			0,0000	0,0240	0,0433	0,0593	0,0725	0,0837	0,0932	0,1013	7
8					0,0000	0,0196	0,0359	0,0496	0,0612	0,0711	8
9							0,0000	0,0163	0,0303	0,0422	9
10									0,0000	0,0140	10

j \ n	21	22	23	24	25	26	27	28	29	30	n \ j
1	0,4643	0,4590	0,4542	0,4493	0,4450	0,4407	0,4366	0,4328	0,4291	0,4254	1
2	0,3185	0,3156	0,3126	0,3098	0,3069	0,3043	0,3018	0,2992	0,2968	0,2944	2
3	0,2578	0,2571	0,2563	0,2554	0,2543	0,2533	0,2522	0,2510	0,2499	0,2487	3
4	0,2119	0,2131	0,2139	0,2145	0,2148	0,2151	0,2152	0,2151	0,2150	0,2148	4
5	0,1736	0,1764	0,1787	0,1807	0,1822	0,1836	0,1848	0,1857	0,1064	0,1870	5
6	0,1399	0,1443	0,1480	0,1512	0,1539	0,1563	0,1584	0,1601	0,1616	0,1630	6
7	0,1092	0,1150	0,1201	0,1245	0,1283	0,1316	0,1346	0,1372	0,1395	0,1415	7
8	0,0804	0,0878	0,0941	0,0997	0,1046	0,1089	0,1128	0,1162	0,1192	0,1219	8
9	0,0530	0,0618	0,0696	0,0764	0,0823	0,0876	0,0923	0,0965	0,1002	0,1036	9
10	0,0263	0,0368	0,0459	0,0539	0,0610	0,0672	0,0728	0,0778	0,0822	0,0862	10
11	0,0000	0,0122	0,0228	0,0321	0,0403	0,0476	0,0540	0,0598	0,0650	0,0697	11
12			0,0000	0,0107	0,0200	0,0284	0,0358	0,0424	0,0483	0,0537	12
13					0,0000	0,0094	0,0178	0,0253	0,0320	0,0381	13
14							0,0000	0,0084	0,0159	0,0227	14
15									0,0000	0,0076	15

COEFFICIENTS a_j

j \ n	31	32	33	34	35	36	37	38	39	40	n \ j
1	0,4220	0,4188	0,4156	0,4127	0,4096	0,4068	0,4040	0,4015	0,3989	0,3964	1
2	0,2921	0,2898	0,2876	0,2854	0,2834	0,2813	0,2794	0,2774	0,2755	0,2737	2
3	0,2475	0,2463	0,2451	0,2439	0,2427	0,2415	0,2403	0,2391	0,2380	0,2368	3
4	0,2145	0,2141	0,2137	0,2132	0,2127	0,2121	0,2116	0,2110	0,2104	0,2098	4
5	0,1874	0,1878	0,1880	0,1882	0,1883	0,1883	0,1883	0,1881	0,1880	0,1878	5
6	0,1641	0,1651	0,1660	0,1667	0,1673	0,1678	0,1683	0,1686	0,1689	0,1691	6
7	0,1433	0,1449	0,1463	0,1475	0,1487	0,1496	0,1505	0,1513	0,1520	0,1526	7
8	0,1243	0,1265	0,1284	0,1301	0,1317	0,1331	0,1344	0,1356	0,1366	0,1376	8
9	0,1066	0,1093	0,1118	0,1140	0,1160	0,1179	0,1196	0,1211	0,1225	0,1237	9
10	0,0899	0,0931	0,0961	0,0988	0,1013	0,1036	0,1056	0,1075	0,1092	0,1108	10
11	0,0739	0,0777	0,0812	0,0844	0,0873	0,0900	0,0924	0,0947	0,0967	0,0986	11
12	0,0585	0,0629	0,0669	0,0706	0,0739	0,0770	0,0798	0,0824	0,0848	0,0870	12
13	0,0435	0,0485	0,0530	0,0572	0,0610	0,0645	0,0677	0,0706	0,0733	0,0759	13
14	0,0289	0,0344	0,0395	0,0441	0,0484	0,0523	0,0559	0,0592	0,0622	0,0651	14
15	0,0144	0,0206	0,0262	0,0314	0,0361	0,0404	0,0444	0,0481	0,0515	0,0546	15
16	0,0000	0,0068	0,0131	0,0187	0,0239	0,0287	0,0331	0,0372	0,0409	0,0444	16
17			0,0000	0,0062	0,0119	0,0172	0,0220	0,0264	0,0305	0,0343	17
18					0,0000	0,0057	0,0110	0,0158	0,0203	0,0244	18
19							0,0000	0,0053	0,0101	0,0146	19
20									0,0000	0,0049	20

j \ n	41	42	43	44	45	46	47	48	49	50	n \ j
1	0,3940	0,3917	0,3894	0,3872	0,3850	0,3830	0,3808	0,3789	0,3770	0,3751	1
2	0,2719	0,2701	0,2684	0,2667	0,2651	0,2635	0,2620	0,2804	0,2589	0,2574	2
3	0,2357	0,2345	0,2334	0,2323	0,2313	0,2302	0,2291	0,2281	0,2271	0,2260	3
4	0,2091	0,2085	0,2078	0,2072	0,2065	0,2058	0,2052	0,2045	0,2038	0,2032	4
5	0,1876	0,1874	0,1871	0,1868	0,1865	0,1862	0,1859	0,1855	0,1851	0,1847	5
6	0,1693	0,1694	0,1695	0,1695	0,1695	0,1695	0,1695	0,1693	0,1692	0,1691	6
7	0,1531	0,1535	0,1539	0,1542	0,1545	0,1548	0,1550	0,1551	0,1553	0,1554	7
8	0,1384	0,1392	0,1398	0,1405	0,1410	0,1415	0,1420	0,1423	0,1427	0,1430	8
9	0,1249	0,1259	0,1269	0,1278	0,1286	0,1293	0,1300	0,1306	0,1312	0,1317	9
10	0,1123	0,1136	0,1149	0,1160	0,1170	0,1180	0,1189	0,1197	0,1205	0,1212	10
11	0,1004	0,1020	0,1035	0,1049	0,1052	0,1073	0,1085	0,1095	0,1105	0,1113	11
12	0,0891	0,0909	0,0927	0,0943	0,0959	0,0972	0,0986	0,0998	0,1010	0,1020	12
13	0,0782	0,0804	0,0824	0,0842	0,0860	0,0876	0,0892	0,0906	0,0919	0,0932	13
14	0,0677	0,0701	0,0724	0,0745	0,0765	0,0783	0,0801	0,0817	0,0832	0,0846	14
15	0,0575	0,0602	0,0628	0,0651	0,0673	0,0694	0,0713	0,0731	0,0748	0,0764	15
16	0,0476	0,0506	0,0534	0,0560	0,0584	0,0607	0,0628	0,0648	0,0667	0,0685	16
17	0,0379	0,0411	0,0442	0,0471	0,0497	0,0522	0,0546	0,0568	0,0588	0,0608	17
18	0,0283	0,0318	0,0352	0,0383	0,0412	0,0439	0,0465	0,0489	0,0511	0,0532	18
19	0,0188	0,0227	0,0263	0,0296	0,0328	0,0357	0,0385	0,0411	0,0436	0,0459	19
20	0,0094	0,0136	0,0175	0,0211	0,0245	0,0277	0,0307	0,0335	0,0361	0,0386	20
21	0,0000	0,0045	0,0087	0,0126	0,0163	0,0197	0,0229	0,0259	0,0288	0,0314	21
22			0,0000	0,0042	0,0081	0,0118	0,0153	0,0185	0,0215	0,0244	22
23					0,0000	0,0039	0,0076	0,0111	0,0143	0,0174	23
24							0,0000	0,0037	0,0071	0,0104	24
25									0,0000	0,0035	25



VALEURS DE W_P

n \ P	0,01	0,02	0,05	0,10	0,50	0,90	0,95	0,98	0,99	P \ n
3	0,753	0,756	0,767	0,789	0,959	0,998	0,999	1,000	1,000	3
4	0,687	0,707	0,748	0,792	0,935	0,987	0,992	0,996	0,997	4
5	0,686	0,715	0,762	0,806	0,927	0,979	0,986	0,991	0,993	5
6	0,713	0,743	0,788	0,826	0,927	0,974	0,981	0,986	0,989	6
7	0,730	0,760	0,803	0,838	0,928	0,972	0,979	0,985	0,988	7
8	0,749	0,778	0,818	0,851	0,932	0,972	0,978	0,984	0,987	8
9	0,764	0,791	0,829	0,859	0,935	0,972	0,978	0,984	0,986	9
10	0,781	0,806	0,842	0,869	0,938	0,972	0,978	0,983	0,986	10
11	0,792	0,817	0,850	0,876	0,940	0,973	0,979	0,984	0,986	11
12	0,803	0,828	0,859	0,883	0,943	0,973	0,979	0,984	0,986	12
13	0,814	0,837	0,866	0,889	0,945	0,974	0,979	0,984	0,986	13
14	0,825	0,846	0,874	0,895	0,947	0,975	0,980	0,984	0,986	14
15	0,835	0,855	0,881	0,901	0,950	0,975	0,980	0,984	0,987	15
16	0,844	0,863	0,887	0,906	0,952	0,976	0,981	0,985	0,987	16
17	0,851	0,869	0,892	0,910	0,954	0,977	0,981	0,985	0,987	17
18	0,858	0,874	0,897	0,914	0,956	0,978	0,982	0,986	0,988	18
19	0,863	0,879	0,901	0,917	0,957	0,978	0,982	0,986	0,988	19
20	0,868	0,884	0,905	0,920	0,959	0,979	0,983	0,986	0,988	20
21	0,873	0,888	0,908	0,923	0,960	0,980	0,983	0,987	0,989	21
22	0,878	0,892	0,911	0,926	0,961	0,980	0,984	0,987	0,989	22
23	0,881	0,895	0,914	0,928	0,962	0,981	0,984	0,987	0,989	23
24	0,884	0,898	0,916	0,930	0,963	0,981	0,984	0,987	0,989	24
25	0,888	0,901	0,918	0,931	0,964	0,981	0,985	0,988	0,989	25
26	0,891	0,904	0,920	0,933	0,965	0,982	0,985	0,988	0,989	26
27	0,894	0,906	0,923	0,935	0,965	0,982	0,985	0,988	0,990	27
28	0,896	0,908	0,924	0,936	0,966	0,982	0,985	0,988	0,990	28
29	0,898	0,910	0,926	0,937	0,966	0,982	0,985	0,988	0,990	29
30	0,900	0,912	0,927	0,939	0,967	0,983	0,985	0,988	0,990	30
31	0,902	0,914	0,929	0,940	0,967	0,983	0,986	0,988	0,990	31
32	0,904	0,915	0,930	0,941	0,968	0,983	0,986	0,988	0,990	32
33	0,906	0,917	0,931	0,942	0,968	0,983	0,986	0,989	0,990	33
34	0,908	0,919	0,933	0,943	0,969	0,983	0,986	0,989	0,990	34
35	0,910	0,920	0,934	0,944	0,969	0,984	0,986	0,989	0,990	35
36	0,912	0,922	0,935	0,945	0,970	0,984	0,986	0,989	0,990	36
37	0,914	0,924	0,936	0,946	0,970	0,984	0,987	0,989	0,990	37
38	0,916	0,925	0,938	0,947	0,971	0,984	0,987	0,989	0,990	38
39	0,917	0,927	0,939	0,948	0,971	0,984	0,987	0,989	0,991	39
40	0,919	0,928	0,940	0,949	0,972	0,985	0,987	0,989	0,991	40
41	0,920	0,929	0,941	0,950	0,972	0,985	0,987	0,989	0,991	41
42	0,922	0,930	0,942	0,951	0,972	0,985	0,987	0,989	0,991	42
43	0,923	0,932	0,943	0,951	0,973	0,985	0,987	0,990	0,991	43
44	0,924	0,933	0,944	0,952	0,973	0,985	0,987	0,990	0,991	44
45	0,926	0,934	0,945	0,953	0,973	0,985	0,988	0,990	0,991	45
46	0,927	0,935	0,945	0,953	0,974	0,985	0,988	0,990	0,991	46
47	0,928	0,936	0,946	0,954	0,974	0,985	0,988	0,990	0,991	47
48	0,929	0,937	0,947	0,954	0,974	0,985	0,988	0,990	0,991	48
49	0,929	0,937	0,947	0,955	0,974	0,985	0,988	0,990	0,991	49
50	0,930	0,938	0,947	0,955	0,974	0,985	0,988	0,990	0,991	50
n \ P	0,01	0,02	0,05	0,10	0,50	0,90	0,95	0,98	0,99	n \ P