Revue des sciences de l'eau
Journal of Water Science

Justification des formules de probabilité empirique basées sur la médiane de la statistique d'ordre
A defense of the median plotting position

D. Rosbjerc, J. Corréa and P. F. Rasmussen

Volume 5, Number 4, 1992

URI: https://id.erudit.org/iderudit/705145ar
DOI: https://doi.org/10.7202/705145ar

See table of contents

Publisher(s)

Université du Québec - INRS-Eau, Terre et Environnement (INRS-ETE)

ISSN

0992-7158 (print)
1718-8598 (digital)

Explore this journal

Cite this article

Rosbjerc, D., Corréa, J. & Rasmussen, P. F. (1992). Justification des formules de probabilité empirique basées sur la médiane de la statistique d'ordre. *Revue des sciences de l'eau / Journal of Water Science*, *5*(4), 529–540. https://doi.org/10.7202/705145ar

Article abstract

Plotting position formulae (PPFs) maintain an important role in engineering practice. In the area of flood frequency analysis they are used to assign exceedance probabilities to observed floods. In the present study we review various principles for the choice of PPFs. These can be divided into three main categories : (1) formulae based on the observed sample frequencies, (2) formulae based on the distribution of sample frequencies, and (3) formulae based on the distribution of order statistics. PPFs in the first two categories are distribution-free, meaning that no assumption needs to be made regarding the form of the parent distribution of events. The Hazen PPF is an example of a formula belonging to the first category.

The Weibull PPF, which is probably the most used formula in practice, belongs to the second category. It can be shown that the frequency corresponding to a particular order statistic is beta distributed regardless of the form of the parent distribution. Being equivalent to the expected value in the beta distribution the Weibull plotting position, $pm = m/(n + 1)$, therefore corresponds to the mean value of sample frequencies. In his book on statistical extremes GUMBEL (1958) recommended the Weibull formula, because it fulfils a set of criteria which he found important. Some of these criteria have later been questioned, for instance by CUNNANE (1978). Various studies have demonstrated that the Weibull PPF is significantly biased in the event domain for most common distributions (an exception is the uniform distribution, where the Weibull PPF is the exact unbiased plotting position).

In recent years most attention has been paid to PPFs of the third category. CUNNANE (1978) strongly recommended the use of unbiased PPFs, i.e. formulae for the exceedance probabilities of the expected values of order statistics. In the last decade much effort has been devoted to the development of unbiased plotting positions. As unbiased PPs in virtue of their definition are related to the parent distribution, they will differ for each individual distribution. In general, it is not possible to derive exact unbiased PPFs, but good approximations have been developed for most common distributions such as the normal, the Gumbel, the generalized extreme value, and the Pearson type III distributions (see *table 1*). If a distribution contains a shape parameter, this must be reflected in its unbiased PPF. Approximate formulae for such distributions are therefore in general of a more complex form. This fact along with the need for distribution-dependent formulae is probably the reason why the Weibull formula is still the most used PPF in practice.

In the present study we emphasize the practical convenience of having a distribution-free PPF which at the same time has a statistical interpretation and is related to the distribution of order statistics. The median PP fulfils these points. It is easily seen to be distribution-free by observing that the median in the distribution of order statistics corresponds to the median in the distribution of frequencies (beta distribution). In general, one of two different principles is commonly adopted when developing estimators : either the choice of the modal value (maximum likelihood principle) or the choice of the mean value (principle of unbiasedness). Unfortunately these distinct principles usually lead to different estimators. For continuous distributions used in flood frequency, the median of order statistics is located in between the modal value and the mean value. In general, it is much less biased thon the Weibull PPF, which for the EV1 distribution is a reasonable approximation to the modal PP. Although no analytical expression exists for the median plotting position, an approximation has been deduced by BENARO and BOS-LEVENBACH (1953), namely $pm = (m - 0.3)/(n + 0.4)$. This formula is also known as the Chegodayev PPF.

Various PPFs are exemplified in the case of three different parent distributions, namely the normal, the log-normal, and the Gumbel distributions. The sample size n = 10 is considered. The results of applying different PPFs are presented in *table 2* (for the largest order statistic in the sample). Several observations can be made : 1) The Benard and Bos-Levenbach-formula is a good approximation to the median PP; 2) The Weibull-formula is close to the modal value PP when the parent is EV1, but differs significantly in the case of other parents; 3) The unbiased PPs depend strongly on the underlying distribution, and an unbiased PPF suitable for all distributions can therefore not be found; 4) The median PP is a fair compromise between the mean and the modal values of the order statistics in all three cases, and it is therefore recommended as a good choice of a standard PPF.

REVUE DES SCIENCES DE L'EAU, 5(1992) 529-540

Justification des formules de probabilité empirique basées sur la médiane de la statistique d'ordre

A defense of the median plotting position

D. ROSBJERG[1], J. CORRÉA[2], P.F. RASMUSSEN[3]

Reçu le 21 septembre 1990, accepté pour publication le 1er juin 1992*.

SUMMARY

Plotting position formulae (PPFs) maintain an important role in engineering practice. In the area of flood frequency analysis they are used to assign exceedance probabilities to observed floods. In the present study we review various principles for the choice of PPFs. These can be divided into three main categories : (1) formulae based on the observed sample frequencies, (2) formulae based on the distribution of sample frequencies, and (3) formulae based on the distribution of order statistics. PPFs in the first two categories are distribution-free, meaning that no assumption needs to be made regarding the form of the parent distribution of events. The Hazen PPF is an example of a formula belonging to the first category.

The Weibull PPF, which is probably the most used formula in practice, belongs to the second category. It can be shown that the frequency corresponding to a particular order statistic is beta distributed regardless of the form of the parent distribution. Being equivalent to the expected value in the beta distribution the Weibull plotting position, $p_m = m/(n+1)$, therefore corresponds to the mean value of sample frequencies. In his book on statistical extremes GUMBEL (1958) recommended the Weibull formula, because it fulfils a set of criteria which he found important. Some of these criteria have later been questioned, for instance by CUNNANE (1978). Various studies have demonstrated that the Weibull PPF is significantly biased in the event domain for most common distributions (an exception is the uniform distribution, where the Weibull PPF is the exact unbiased plotting position).

1. Institute of Hydrodynamics and Hydraulic Engineering, Technical University of Denmark, DK-2800 Lyngby, Danemark.
2. Département de Génie Civil, Ecole Polytechnique de Montréal, Montréal, Québec, H3C 3A7, Canada.
3. Institut National de la Recherche Scientifique, Université du Québec, 2800 rue Einstein suite 105, c.p. 7500 Sainte-Foy, Québec G1V 4C7, Canada.

* Les commentaires seront reçus jusqu'au 15 juin 1993.

In recent years most attention has been paid to PPFs of the third category. CUNNANE (1978) strongly recommended the use of unbiased PPFs, i.e. formulae for the exceedance probabilities of the expected values of order statistics. In the last decade much effort has been devoted to the development of unbiased plotting positions. As unbiased PPs in virtue of their definition are related to the parent distribution, they will differ for each individual distribution. In general, it is not possible to derive exact unbiased PPFs, but good approximations have been developed for most common distributions such as the normal, the Gumbel, the generalized extreme value, and the Pearson type III distributions (see *table 1)*. If a distribution contains a shape parameter, this must be reflected in its unbiased PPF. Approximate formulae for such distributions are therefore in general of a more complex form. This fact along with the need for distribution-dependent formulae is probably the reason why the Weibull formula is still the most used PPF in practice.

In the present study we emphasize the practical convenience of having a distribution-free PPF which at the same time has a statistical interpretation and is related to the distribution of order statistics. The median PP fulfils these points. It is easily seen to be distribution-free by observing that the median in the distribution of order statistics corresponds to the median in the distribution of frequencies (beta distribution). In general, one of two different principles is commonly adopted when developing estimators : either the choice of the modal value (maximum likelihood principle) or the choice of the mean value (principle of unbiasedness). Unfortunately these distinct principles usually lead to different estimators. For continuous distributions used in flood frequency, the median of order statistics is located in between the modal value and the mean value. In general, it is much less biased than the Weibull PPF, which for the EV1 distribution is a reasonable approximation to the modal PP. Although no analytical expression exists for the median plotting position, an approximation has been deduced by BENARD and BOS-LEVENBACH (1953), namely $p_m = (m - 0.3)/(n + 0.4)$. This formula is also known as the Chegodayev PPF.

Various PPFs are exemplified in the case of three different parent distributions, namely the normal, the log-normal, and the Gumbel distributions. The sample size n = 10 is considered. The results of applying different PPFs are presented in *table 2* (for the largest order statistic in the sample). Several observations can be made : 1) The Benard and Bos-Levenbach-formula is a good approximation to the median PP ; 2) The Weibull-formula is close to the modal value PP when the parent is EV1, but differs significantly in the case of other parents ; 3) The unbiased PPs depend strongly on the underlying distribution, and an unbiased PPF suitable for all distributions can therefore not be found ; 4) The median PP is a fair compromise between the mean and the modal values of the order statistics in all three cases, and it is therefore recommended as a good choice of a standard PPF.

Key words : *plotting position formulae, unbiasedness, median, Beta distribution, Benard and Bos-Levenbach formula.*

RÉSUMÉ

Au cours de ces dernières années, beaucoup d'efforts ont été consacrés au développement de formules de probabilité empirique (FPE) non biaisées. En raison même de leur définition, les FPE non biaisées sont dépendantes de la distribution parente des échantillons considérés, et une formule doit donc être établie pour chaque distribution. Dans cette étude on passe en revue les différentes approches pour le développement des FPE, et on montre que la FPE basée sur la médiane des statistiques d'ordre peut constituer un compromis acceptable entre les FPE non biaisées (i.e. correspondant à la moyenne des statistiques d'ordre), et les FPE basées sur le mode des statistiques d'ordre.

Contrairement à ces dernières, la FPE basée sur la médiane des statistiques d'ordre est indépendante de la distribution parente des échantillons, et peut donc être utilisée de façon standard. Par ailleurs, la FPE basée sur la médiane des statistiques d'ordre est moins biaisée que la FPE de Weibull, qui est également indépendante de la distribution parente des échantillons. Bien qu'il n'existe pas d'expression analytique exacte pour la FPE basée sur la médiane des statistiques d'ordre, BENARD et BOS-LEVENBACH en ont proposé une très bonne approximation, $p_m = (m - 0.3)/(n + 0.4)$. Cette formule est aussi connue sous le nom de formule de Chegodayev.

Mots-clés : *formules de probabilité empirique, distribution des fréquences échantillonnales, médiane des statistiques d'ordre, formule de Benard et Bos-Levenbach.*

1 - INTRODUCTION

La détermination de la meilleure formule de probabilité empirique (FPE) a toujours été un défi pour les hydrologues depuis l'adoption des méthodes probabilistes. De nombreuses formules ont été proposées par différents auteurs et de bonnes études de synthèse existent dans la littérature, par exemple CUNNANE (1978), JI, *et al.* (1984), et HARTER (1984). Ces formules sont cependant assez semblables d'un point de vue pratique, ce qui laisse supposer que des efforts supplémentaires sur ce sujet seraient peu justifiés. Toutefois, dans le cas particulier important en pratique en hydrologie où l'on traite de petits échantillons, le choix de la FPE à utiliser peut être crucial pour la détermination de la période de retour des valeurs extrêmes. Par ailleurs, d'un point de vue strictement théorique, ce problème présente un intérêt toujours actuel. Il s'agit, en bref, d'estimer la probabilité de dépassement (ou de non-dépassement) pour chacune des valeurs d'un échantillon ordonné. Trois approches différentes peuvent être adoptées et elles sont présentées dans ce qui suit.

1.1 FPE basées sur les fréquences échantillonnales

La première approche prend en compte uniquement les fréquences cumulées des observations. Lorsque les n valeurs d'un échantillon sont classées en ordre décroissant, $x_1 \geq x_2 \geq \ldots \geq x_m \geq \ldots \geq x_n$, la valeur de rang m ($x_m$) est dépassée par m – 1 autres valeurs, ce qui permet d'estimer la probabilité de dépassement de x_m par :

$$p_m = \frac{m-1}{n} \tag{1}$$

Cependant, si m = 1, on obtient $p_1 = 0$, ce qui n'a en pratique pas de sens. Pour pallier cet inconvénient la méthode dite de Californie compte le nombre d'observations égales ou supérieures à la valeur de rang m, d'où la formule :

$$p_m = \frac{m}{n} \tag{2}$$

La probabilité de dépassement de la plus grande valeur observée est alors évaluée à 1/n, ce qui est plus acceptable. Cependant cette formule présente l'inconvénient d'obtenir pour probabilité de dépassement de la plus petite observation (x_n) la valeur $p_n = 1$. Ainsi HAZEN (1914) a suggéré d'utiliser la moyenne arithmétique des estimations données par les relations (1) et (2) soit :

$$p_m = \frac{m - 0{,}5}{n} \tag{3}$$

On trouve ainsi $p_1 = 1/2n$ et $p_n = 1 - 1/2n$, ce qui paraît plus acceptable.

1.2 FPE basées sur la distribution des fréquences échantillonnales

La deuxième approche part également du principe que les FPE devraient être indépendantes de la distribution parente mais est basée non pas sur les fréquences échantillonnales elles-mêmes, mais sur leur distribution. La figure 1 montre la fonction de répartition (FR) F(x) et la fonction de densité de probabilité (FDP) de la plus grande valeur observée X_1, soit $f_1(x)$. Pour chaque valeur possible de X_1 on définit une valeur p_1 sur l'axe des probabilités par la transformation $p_1 = 1 - F(x_1)$. En considérant que les éléments de l'échantillon sont indépendants, la FR de X_1 est donnée par

$$F_1(x) = [F(x)]^n \tag{4}$$

La FR et la FDP de $P_1 = 1 - F(X_1)$, notées respectivement $G_1(p)$ et $g_1(p)$, peuvent être déduites de la manière suivante :

$$\begin{aligned} G_1(p) &= P\{1 - F(X_1) \le p\} \\ &= P\{X_1 \ge F^{-1}(1-p)\} \\ &= 1 - F_1(F^{-1}(1-p)) \\ &= 1 - [F(F^{-1}(1-p))]^n \\ &= 1 - (1-p)^n \qquad 0 \le p \le 1 \end{aligned} \tag{5}$$

et

$$g_1(p) = G_1'(p) = n(1-p)^{n-1} \qquad 0 \le p \le 1 \tag{6}$$

Notons que cette dérivation est permise par le fait que la FR est strictement croissante. Les relations (5) et (6) montrent que la distribution de P_1 est indépendante de F(x) et dépend uniquement de la taille de l'échantillon. La moyenne $\bar{p}_1$ est déterminée par :

$$\bar{p}_1 = E[P_1] = \int_0^1 pn(1-p)^{n-1}\,dp = 1/(n+1) \tag{7}$$

et sa valeur médiane $\hat{p}_1$ peut être obtenue par :

$$1 - (1 - \hat{p}_1)^n = 0{,}5 \tag{8}$$

d ' où

$$\hat{p}_1 = 1 - 0{,}5^{1/n} \tag{9}$$

La valeur de p correspondant au maximum de $g_1(p)$, $\tilde{p}_1$, est :

$$1 - \tilde{p}_1 = 1 \tag{10}$$

d ' où

$$\tilde{p}_1 = 0 \tag{11}$$

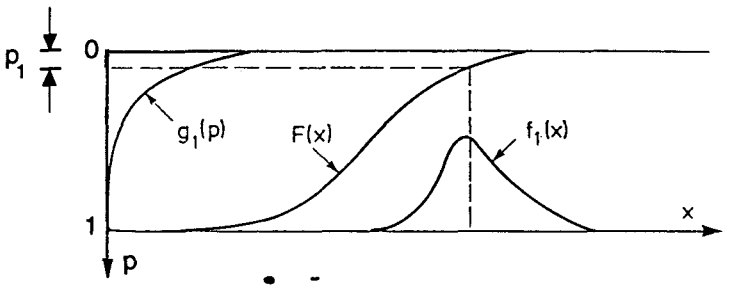

Figure 1 Schéma de la FDP de la plus grande valeur observée x_1 et de la FR F (x)

Probability density functions of the largest value in the sample, X_1 and its corresponding exceedance probability $P_1 = 1 - F(X_1)$. F (x) is the cumulative distribution function to be estimated.

De façon générale, on trouve que $P_m = 1 - F(X_m)$, où X_m est la statistique d'ordre m, est indépendante de F(x) et suit une distribution bêta (KENDALL et STUART, 1977) de FDP :

$$g_m(p) = \frac{1}{B(m, n-m+1)} [1-p]^{n-m} [p]^{m-1} \qquad 0 \le p \le 1 \tag{12}$$

WEIBULL (1939) a utilisé la valeur moyenne de P_m, $\bar{p}_m$, comme probabilité empirique :

$$p_m = \bar{p}_m = \frac{m}{n+1} \tag{13}$$

Ce choix a été confirmé par GUMBEL (1958) qui a indiqué des conditions que devrait respecter une FPE :

– La FPE doit être telle que toutes les observations peuvent être représentées.

– La probabilité empirique de la valeur de rang m doit se situer entre les fréquences observées (m – 1)/n et m/n et être indépendante de la distribution parente.

– La période de retour d'une valeur supérieure ou égale à la plus grande valeur observée doit être proche de n.

– Les observations doivent être également espacées sur l'échelle des fréquences, i. e. l'écart entre la valeur de rang (m + 1) et celle de rang m doit être une fonction de n seulement et être indépendant de m.

– La FPE doit avoir une signification intuitive et être analytiquement simple.

BEARD (1943) a quant à lui choisi la médiane $\hat{p}_1$ comme FPE de la plus grande valeur observée. Malheureusement il n'existe pas de formule explicite donnant la médiane d'une distribution bêta en général. BEARD (1943) a proposé la FPE approximative suivante :

$$p_m = \hat{p}_m = \frac{m - 0{,}31}{n + 0{,}38} \tag{14}$$

Plus récemment BENARD et BOS-LEVENBACH (1953) ont proposé cette autre FPE qui est un peu plus simple :

$$p_m = \hat{p}_m = \frac{m - 0{,}3}{n + 0{,}4} \tag{15}$$

Cette relation est également connue sous le nom de formule de Chegodayev (CHOW, 1964).

Correspondant au mode de P_m, $\tilde{p}_m$, on a la FPE suivante :

$$p_m = \tilde{p}_m = \frac{m - 1}{n - 1} \tag{16}$$

Cette formule n'a cependant pas été beaucoup utilisée pour le calcul des probabilités empiriques en raison des mauvais résultats auxquels elle conduit pour les valeurs extrêmes.

1.3 FPE basées sur la distribution des statistiques d'ordre

La troisième approche possible est basée sur la distribution de la statistique d'ordre X_m elle même au lieu de celle de sa fréquence. Des probabilités empiriques non biaisées sont obtenues par :

$$p_m = 1 - F(\bar{x}_m) \tag{17}$$

où $\bar{x}_m = E[X_m]$ est la valeur moyenne de X_m. L'utilisation de FPE non biaisées pour l'estimation de quantiles et l'estimation des paramètres d'une distribution a été encouragée par CUNNANE (1978) qui a constaté que la formule de Weibull occasionne des biais importants dans le cas de distributions à asymétrie positive. D'autres auteurs ont également recommandé l'utilisation de FPE non biaisées (IN-NA et NGUYEN, 1989 ; ARNELL *et al.*, 1986). La distribution de la statistique d'ordre X_m peut être déduite en notant que dans un échantillon de taille n, extrait d'une distribution continue F(x), la probabilité d'avoir n – m valeurs inférieures à x, m – 1 valeurs supérieures à x, et une dans l'intervalle x ± 1/2 dx est :

$$dF_m(x) = m\binom{n}{m} [F(x)]^{n-m} [1 - F(x)]^{m-1} f(x)\, dx \tag{18}$$

On peut en déduire (KENDALL et STUART, 1977) :

$$f_m(x) = \frac{1}{B(m, n - m + 1)} [F(x)]^{n-m} [1 - F(x)]^{m-1} f(x) \tag{19}$$

A partir de l'équation précédente on voit que $E[X_m]$ s'écrit :

$$E[X_m] = \bar{x}_m = \int_{-\infty}^{\infty} x f_m(x)\, dx$$

$$= \frac{1}{B(m, n - m + 1)} \int_{-\infty}^{\infty} x [F(x)]^{n-m} [1 - F(x)]^{m-1} f(x)\, dx \tag{20}$$

Des fréquences exactes peuvent être obtenues en utilisant (17) et (20) mais cela requiert de procéder généralement par intégration numérique (dans certains cas, on peut aussi utiliser la méthode des moments pondérés comme proposé par ARNELL *et al.*, 1986). Excepté le cas de la distribution uniforme, il n'existe pas d'expression exacte simple du point de vue analytique. Les meilleures approximations qui ont été proposées sont données sur le tableau 1. Notons que chaque distribution requiert sa propre FPE. Toutefois CUNNANE (1978) a trouvé que la relation :

$$p_m = \frac{m - 2/5}{n + 1/5} \tag{21}$$

est un compromis satisfaisant dans la plupart des cas.

Tableau 1 Formules approximatives de probabilités empiriques non biaisées.

Table 1 *Approximate unbiased plotting positions.*

Distribution	FPE	Référence
Normale	$\frac{m-3/8}{n+1/4}$	BLOM (1958)
EV1	$\frac{m-0,44}{n+0,12}$	GRINGORTEN (1963)
GEV[1]	$\frac{m-0,13\,\tau-0,27}{n-0,08\,\tau+0,38}$	IN-NA et NGUYEN (1989)[2]
Pearson III[1]	$\frac{m-0,42}{n+0,3\,\tau+0,05}$	NGUYEN *et al.* (1989)
Exponentielle[3]	$\exp\left(-\sum_{i=1}^{n+1-m}\frac{1}{n+1-i}\right)$	SUKHAME (1938)
Uniforme[3]	$\frac{m}{n+1}$	WEIBULL (1939)

1 τ représente le coefficient d'asymétrie.
2 La formule pour la loi GEV a été discutée récemment par GUO (1990).
3 Exacte.

La distribution des statistiques d'ordre étant en général asymétrique, il en résulte que la moyenne $\bar{x}_m$, le mode $\tilde{x}_m$, et la médiane $\hat{x}_m$ ne sont pas confondus. Pour le mode on obtient la formule :

$$p_m = 1 - F(\tilde{x}_m) \tag{22}$$

Le mode de la statistique d'ordre dépend de la distribution parente F(x) et il n'existe pas de formule explicite générale. La solution de l'équation (22) peut cependant être tabulée par intégration numérique. Dans le cas de la loi des valeurs extrêmes type I la probabilité empirique basée sur le mode de X_1, $\tilde{x}_1$, peut être aisément déduite :

$$p_1 = 1 - e^{-1/n} \approx \frac{1}{n + 0,5} \tag{23}$$

ce qui est un résultat proche de celui obtenu par la formule de Weibull, $p_1 = 1/(n + 1)$, [eq. (13)].

Contrairement aux deux cas précédents, si la probabilité empirique est basée sur la médiane de X_m, $\hat{x}_m$, on obtient une formule indépendante de la distribution parente. En effet, alors que $F(\bar{x}_m) \neq 1 - \bar{p}_m$, excepté pour la distribution uniforme, on a $F(\hat{x}_m) = 1 - \hat{p}_m$ *(figure 2)*. Ainsi la FPE $p_m = 1 - F(\hat{x}_m)$ basée sur la médiane de la statistique d'ordre X_m est équivalente à la FPE $p_m = \hat{p}_m$ basée sur la médiane des fréquences échantillonales. Ceci constitue une particularité appréciable pour les FPE basées sur la médiane. On peut retrouver la probabilité empirique en faisant :

$$F_m(\hat{x}_m) = \frac{1}{B(m, n-m+1)} \int_{-\infty}^{\hat{x}_m} [F(x)]^{n-m} [1 - F(x)]^{m-1} f(x)\, dx \tag{24}$$

$$= \frac{1}{B(m, n-m+1)} \int_{\hat{p}_m}^{1} [1-p]^{n-m} p^{m-1}\, dp = \frac{1}{2}$$

On voit que p_m est évaluée indépendamment de F(x). La FPE $p_m = 1 - F(\hat{x}_m) = G_m^{-1}(1/2)$ est donc indépendante de la distribution parente.

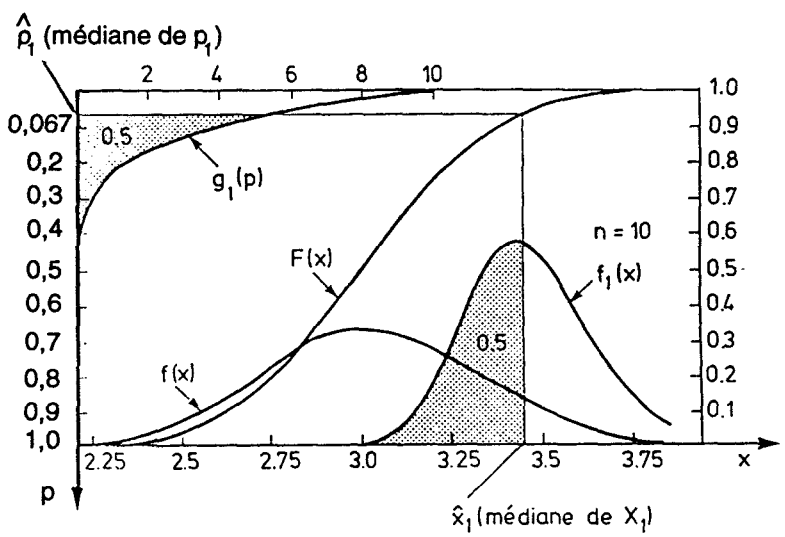

Figure 2 Correspondance entre la médiane de la probabilité de dépassement $\hat{p}_1$ et la médiane de la plus grande valeur observée $\hat{x}_1$ ($\hat{p}_1 = 1 - F(\hat{x}_1)$).

Relationship between the median of the largest value in the sample, $\hat{x}_1$, and its corresponding exceedance probability $\hat{p}_1 = 1 - F(\hat{x}_1)$. F (x) and f (x) denote the parent cumulative distribution function and probability density function, respectively.

2 - DISCUSSION

Deux principes souvent utilisés pour le développement d'estimateurs sont le choix de la valeur la plus probable (estimateur de maximum de vraisemblance), et le choix de la valeur moyenne (estimateur non biaisé). Toutefois, ces principes conduisent en général à des estimateurs différents.

GUMBEL (1958) a recommandé les probabilités empiriques basées sur la valeur la plus probable (mode) et a trouvé que la formule de Weibull, tout au moins pour la distribution EVI, est satisfaisante de ce point de vue. Par ailleurs, du fait qu'elle respecte l'ensemble des cinq conditions indiquées précédemment, il a recommandé son utilisation systématique (GUMBEL, 1958). Il s'avère cependant que la formule de Weibull conduit à des biais importants. Ainsi CUNNANE (1978) a remarqué que la formule de Weibull donne des biais significatifs pour les valeurs extrêmes lorsque la distribution parente a une asymétrie positive, ce qui est souvent le cas en hydrologie. Dans sa rétrospective de la problématique des FPE, CUNNANE a remis en question deux des conditions proposées par GUMBEL, et l'usage systématique de la formule de Weibull (recommandée par GUMBEL compte tenu des cinq conditions postulées). Il a recommandé l'usage de probabilités empiriques non biaisées et liées à la distribution échantillonnale, même si celles-ci peuvent dépendre d'un paramètre de forme pour certaines distributions à trois paramètres. Reconnaissant cependant que cette approche ne convient pas lorsque les distributions sont inconnues *a priori*, il a présenté une formule générale de compromis [eq. (21)].

La préférence pour la valeur la plus probable (ou valeur modale) a aussi été critiquée par BEARD (1943), qui soutient que même si le mode est plus probable que la moyenne ou la médiane, la probabilité de l'obtenir demeure infinitésimale. Il a préféré la médiane à la moyenne et au mode parce qu'elle conduit à des probabilités empiriques basées sur les statistiques d'ordre et indépendantes de la distribution parente. Étrangement, ce fait était ignoré par GUMBEL (1958) qui, parallèlement à sa deuxième condition a affirmé que les fréquences empiriques basées sur les probabilités de la moyenne, du mode ou de la médiane de la statistique d'ordre devaient être exclues parce qu'étant différentes pour différentes distributions. Ceci est tout à fait vrai en ce qui concerne le mode ou la moyenne mais ne l'est pas en ce qui concerne la médiane. Par ailleurs JI *et al.* (1984) ont déterminé par la méthode de Monte-Carlo, des probabilités empiriques non biaisées pour la loi Pearson type III, avec différentes valeurs de coefficient d'asymétrie. Ils ont comparé leurs fréquences empiriques avec celles données par la formule de HAZEN [Eq. (3)], BENARD et BOS-LEVENBACH (eq. [15]) et Weibull (eq. [13]) et ont trouvé que les formules de HAZEN et de BENARD et BOS-LEVENBACH (basée sur la médiane) n'ont donné que des biais légers alors que la formule de Weibull a donné des biais significatifs.

Les FPE basées sur la médiane de la statistique d'ordre se particularisent ainsi par le fait que d'une part, elles donnent généralement des estimations moins biaisées que les FPE basées sur le mode de la statistique d'ordre, et que d'autre part, elles sont indépendantes de la distribution parente, contrairement aux FPE basées sur la moyenne de la statistique d'ordre. De plus, les

estimations obtenues sont généralement comprises entre celles données par les FPE basées sur le mode et celles données par les FPE basées sur la moyenne. Les FPE basées sur la médiane de la statistique d'ordre constituent donc un compromis intéressant entre ces deux alternatives. Ceci est illustré par quelques exemples ci-après.

3 - APPLICATION

Dans ce qui suit nous considérons trois échantillons de taille n = 10 extraits respectivement de populations lognormale, normale et EVI.

La discussion est détaillée dans le cas de l'échantillon lognormal et les résultats sont présentés pour l'ensemble des trois distributions.

Considérons un échantillon de taille 10 extrait d'une distribution lognormale dont les valeurs transformées ont une moyenne de 3,0 et un écart type de 0,3. La figure 3 montre la FDP et la FR de la population ainsi que la FDP de la plus grande statistique d'ordre. Les probabilités empiriques correspondant au mode, à la médiane et à la moyenne de f_1 (x) ont été calculées avec exactitude et leur inverse, i.e les périodes de retour correspondant à la plus grande valeur observée suivant les différentes approches, sont également données sur la figure. On constate qu'il y a une grande variation dans les périodes de retour calculées. Le tableau 2 présente les résultats pour ce cas de distribution lognormale et ceux obtenus pour des échantillons extraits de distributions normale et EVI. On vérifie que la probabilité empirique basée sur la médiane, du fait qu'elle n'est pas liée au type de distribution, est la même dans les trois cas. Les résultats exacts obtenus peuvent être comparés à ceux obtenus par les différentes formules approximatives proposées. Par la formule de Blom (établie pour la loi normale) on obtient 16,4 ans, et par la formule de Gringorten (établie pour la loi EVI) on obtient 18,1 ans. La formule générale de Cunnane [eq. (21)] donne 17 ans, ce qui est un compromis raisonnable entre le cas de la loi normale et celui de l'EVI, mais reste très différent de la fréquence empirique non biaisée obtenue pour le cas de la loi lognormale. La formule approximative de Benard et Bos-Levenbach, eq. (15) basée sur la médiane donne comme résultat 14,9 ans, ce qui a également été obtenu par un calcul exact.

Le tableau 2 présente également les périodes de retour obtenues par des FPE basées sur les fréquences échantillonnales, ainsi que celles données par des FPE basées sur la distribution des fréquences. Une comparaison des différents résultats montre que :

– La formule de Weibull donne des résultats relativement voisins de ceux obtenus à partir du mode de X_1 dans le cas de la distribution de Gumbel (EVI), mais ceci n'est pas le cas pour les deux autres distributions.

– Les probabilités empiriques calculées sur la base de la moyenne de X_1 (fréquences non biaisées) donnent des résultats très différents suivant le type de distribution. Une formule de compromis qui serait non biaisée dans tous les cas est par conséquent impossible à trouver.

– Parmi les FPE basées sur la distribution des statistiques d'ordre considérées dans cette étude, celle correspondant à la médiane est la seule qui puisse être utilisée de façon standard lorsque la distribution n'est pas connue. En effet, la période de retour calculée est indépendante de la distribution. (Ceci est une particularité des FPE basées sur un quantile des statistiques d'ordre).

– Pour les distributions examinées, les probabilités empiriques basées sur la médiane semblent être un bon compromis entre les probabilités empiriques non biaisées et celles basées sur le mode de la statistique d'ordre car elles donnent des résultats intermédiaires.

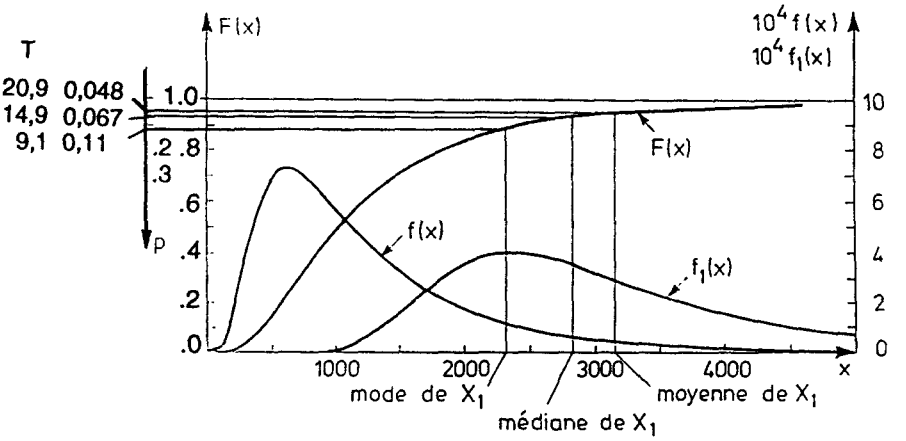

Figure 3 Probabilités empiriques basées sur la distribution de la plus grande valeur observée dans le cas d'une distribution lognormale ($y = \log_{10} x$; $\mu_y = 3$; $\sigma_y = 0.3$).
Plotting positions corresponding to the mean, the mode, and the median of the largest value in a log-normal sample of size 10 ($Y = \log_{10} X$; $\mu_Y = 3$; $\sigma_Y = 0.3$).

Tableau 2 Période de retour calculée pour la plus grande valeur observée dans un échantillon de taille 10.

Table 2 *Return period for the largest value in samples of size 10 computed according to various plotting position formulae.*

Formule de probabilité empirique	Distribution		
	Normale	Lognormale	EVI
$1 - F(x_1)$ (mode)	12,8	9,1	10,5
$1 - F(x_1)$ (médiane)	14,9	14,9	14,9
$1 - F(x_1)$ (moyenne)	16,2	20,9	18,3
BLOM : (m – 3/8)/(n + 1/4)	16,4	16,4	16,4
GRINGORTEN : (m – 0,44)/(n + 0,12)	18,1	18,1	18,1
CUNNANE : (m – 2/5)/(n + 1/5)	17	17	17
Empirique : (m – 1)/n	∞	∞	∞
Californie : m/n	10	10	10
HAZEN : (m – 0,5)/n	20	20	20
WEILBULL : m/(n + 1)	11	11	11
BENARD et BOS-LEVENBACH : (m – 0,3)/(n + 0,4)	14,9	14,9	14,9
Mode : (m – 1)/(n – 1)	∞	∞	∞

CONCLUSION

Le calcul des probabilités empiriques devrait être basé sur la distribution des statistiques d'ordre. Le support théorique pour des probabilités empiriques basées sur le mode ou la moyenne existe, mais parce qu'elles dépendent du type de distribution, ces méthodes ne sont pas applicables pour des distributions inconnues *a priori.* Par ailleurs elles sont difficiles à mettre en œuvre dans le cas où intervient un paramètre de forme et des formules approximatives ne sont disponibles que pour un petit nombre de distributions particulières.

Par contre les probabilités empiriques basées sur la médiane des statistiques d'ordre sont indépendantes du type de distribution. Elles sont donc utiles en hydrologie car la distribution parente pour un échantillon donné est rarement connue. Le biais occasionné est généralement faible et le seul inconvénient est l'inexistence d'une formule exacte. Toutefois, une très bonne et simple formule approximative existe, celle de BENARD et BOS-LEVENBACH [eq. (15)]. Ainsi, malgré des propriétés plausibles, la formule de Weibull a été remise en question au bénéfice de FPE non biaisées. Cependant, tel qu'illustré dans cette étude, il semble plus raisonnable de réhabiliter l'approche des FPE basées sur la médiane et d'utiliser la formule approximative de BENARD et BOS-LEVENBACH [eq. (15)] comme standard.

RÉFÉRENCES BIBLIOGRAPHIQUES

ARNELL N.W., BERAN M., HOSKING J.R.M., 1986. Unbiased plotting positions for the General Extreme Value distribution, *J. Hydrol., 86,* 59-69.

BEARD L.R., 1943. Statistical analysis in hydrology, Trans., *ASCE,* 108, 1110-1160.

BENARD A., BOS-LEVENBACH E.C., 1953. The plotting of observations on probability paper, *Statistica Neerlandica, 7,* 163-173.

BLOM G., 1958. Statistical estimates and transformed beta variables, Wiley, New York.

CHOW V.T. (ed.), 1964. Handbook of applied hydrology, McGraw Hill, New York, N.Y.

CUNNANE, C., 1978. Unbiased plotting positions – a review, *J. Hydrol., 37,* 205-222.

GRINGORTEN I.I., 1963. A plotting rule for extreme probability paper, *J. Geophys. Res., 68,* 813-814.

GUMBEL E.J., 1958. Statistics of extremes, Columbia University Press, New York.

GUO S.L., 1990. A discussion on unbiased plotting position for the General Extreme Value distribution, *J. Hydrol., 121,* 33-44.

HARTER H.L., 1984. Another look at plotting positions, *Commun. Stat. Theor. Meth., 13,* 1613-1633.

HAZEN A., 1914. Storage to be provided in impounding reservoirs for municipal water supply, Trans., *ASCE, 77,* 1547-1550.

IN-NA N., NGUYEN V.T.V., 1989. An unbiased plotting position formula for the General Extreme Value distribution, *J. Hydrol., 106,* 193-209.

JI X., JING D., SHEN H.W., SALAS J.D., 1984. Plotting positions for Pearson type-III distribution, *J. Hydrol., 74,* 1-29.

KENDALL S.M., STUART A., 1977. The advanced theory of statistics, vol.1, Distribution theory, 4^{e} ed., MacMillan Pub. Co., New York.

NGUYEN V.T.V., IN-NA N., BOBEE B., 1989. New plotting position formula for Pearson type-III distribution, *J. Hydraul. Eng., 115,* 709-730.

SUKHAME P.V., 1938. Test of significance for samples of the chisquare population with two degrees of freedom, *Ann. Engen.,* London, *8,* 52-56.

WEIBULL W., 1939. A statistical theory of strength of materials, Ing. Vetenskaps Akad., Handl, Stockolm.