

Chapitre 1 : Analyse fréquentielle et statistique appliquée à l'hydrologie

1.1 Rappels de notions de probabilité et d'analyse statistique

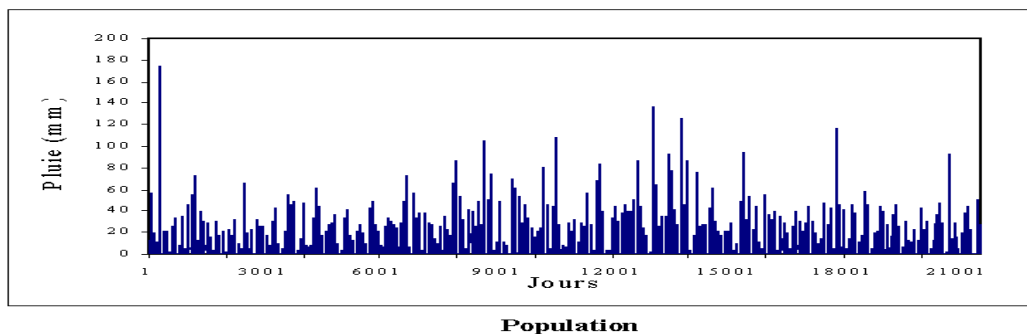
1.1.1 Variable aléatoire, population, échantillon et lois de probabilité

L'analyse statistique est opérée sur un échantillon ou une série d'observations (données).

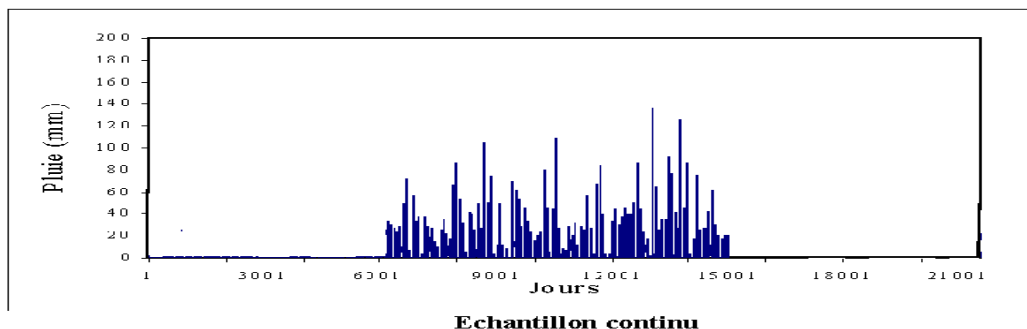
Ces données (variables) peuvent être continues ou discrètes.

- Variables continues : ce sont généralement des valeurs mesurées (précipitation, débit, ...).
- Variables discrètes : résultent habituellement d'un comptage (le fait de pleuvoir ou non est une variable discrète), (nombre de jours de pluie, de neige ou de gel...).

Elles proviennent d'une population, qui est un phénomène hydrologique dont une description est requise.



On appelle population un ensemble fini ou infini d'éléments.



Un échantillon est un sous-ensemble de la population.

On dira qu'un échantillon est exhaustif lorsque sa taille est celle de la population.

En hydrologie de surface, les populations sont généralement infinies, et par conséquent, les échantillons sont non exhaustifs.

Ces données doivent répondre à certains critères statistiques pour que ce genre d'analyse soit valable :

Elles doivent être indépendantes, *homogènes*, *stationnaires* et *aléatoires*.

Les variables que l'on manipule en hydrologie (précipitations, débits, mais aussi niveau de nappe phréatique, durée d'insolation, etc.) sont considérées comme des variables aléatoires.

La Variable Aléatoire, parfois notée V.A., est une variable formelle, notée en majuscule, par exemple X :

Cette variable prendra une valeur x_k à chaque « tirage aléatoire », à chaque réalisation k .

Par exemple, en 1980, la variable X a pris la valeur $x_{80} = 734 \text{ mm}$.

Chapitre 1 : Analyse fréquentielle et statistique appliquée à l'hydrologie

Il n'en reste pas moins que, si l'on veut dimensionner un barrage pour compenser le manque d'eau nécessaire à l'AEP, à Alger, il faudra s'intéresser aux années futures (par exemple de 1987, fin de la construction de l'ouvrage (barrage de Keddara), à 2037, fin de la période d'amortissement). Or on ne savait pas en 1987 ce que seraient les réalisations de la variable aléatoire X en 1988, 1989, etc., c'est-à-dire x_{88} , x_{89} , x_{90} , etc.

On se trouve alors en avenir incertain : aucune approche déterministe, aucune mesure ou méthode déductive, ne peut nous dire exactement, en 1987, ce que sera la réalisation X_{92} de X en 1992.

Tout au plus pourra-t-on supposer que les phénomènes générateurs de la pluie seront les mêmes que dans le passé récent, et on fera l'hypothèse que les réalisations futures de la variable aléatoire X auront les mêmes caractéristiques, la même distribution statistique que par le passé.

La proximité temporelle ne garantit pas la stationnarité. Toute extrapolation, même sur quelques décennies, doit discuter les évolutions du climat, du bassin et des usages ; une durée de vie d'ouvrage ne constitue pas un critère statistique de validité.

Remarque — stationnarité et changement climatique. Une série est stationnaire si ses propriétés statistiques (moyenne, variance, loi de probabilité) ne varient pas dans le temps. Cette hypothèse, fondement de l'analyse fréquentielle classique, peut être mise en défaut par :

- le changement climatique, qui modifie la fréquence et l'intensité des pluies extrêmes et des sécheresses ;
- les modifications du bassin versant : urbanisation, déforestation, incendies, changements de pratiques agricoles ;
- les aménagements et usages : barrages, prélèvements, transferts d'eau, rejets.

Avant l'analyse, examiner les métadonnées, les tendances et les ruptures, puis choisir des tests adaptés (Mann-Kendall, Pettitt, etc.). L'autocorrélation peut modifier leurs seuils. Un test non significatif ne prouve pas l'absence de changement.

- travailler sur une sous-période homogène ;
- utiliser des modèles non stationnaires, par exemple une loi GEV dont les paramètres dépendent du temps ou d'une covariable climatique ;
- raisonner en termes de risque sur la durée de vie de l'ouvrage plutôt qu'en période de retour fixe, et tester la sensibilité des valeurs de projet.

Milly et al. (2008) ont résumé ce débat par la formule « Stationarity is dead ». Le choix de l'hypothèse doit toujours être justifié et discuté dans une étude.

On va donc chercher à décrire et à résumer un échantillon, considéré comme un sous-ensemble d'une population qui sera souvent infinie.

Sur cette population, on peut définir une loi de probabilité : $F(x)$ où x correspond à une valeur numérique.

Cette loi de probabilité, ou fonction de répartition, exprime la : « Probabilité que la Variable Aléatoire X reste inférieure ou égale à la valeur Numérique x . »

$$F(x) = P(X \leq x)$$

La fréquence (f) d'un événement est le nombre de fois que cet événement se répète divisé par le nombre total d'événements observés.

$$f = \frac{n}{N}$$

Chapitre 1 : Analyse fréquentielle et statistique appliquée à l'hydrologie

Lorsque le nombre total d'événements ou d'observations tend vers l'infini, cette mesure tend vers une valeur constante qui s'appelle la probabilité (p).

En hydrologie, il convient de distinguer la fréquence empirique, calculée à partir des observations, de la probabilité théorique, définie par un modèle probabiliste.

Le terme fréquence s'applique aussi à des intervalles fixes comprenant chacun un certain nombre d'observations.

Les couples (x_i, n_i) décrivent une distribution empirique discrète ou groupée. La fréquence relative de classe n_i/N est une masse de probabilité empirique ; pour une variable continue, la hauteur d'un histogramme de densité est $n_i/(N \Delta x_i)$.

$$f_e(x_i) = \frac{n_i}{N}$$

Et que la fonction de répartition est :

$$FND = F_e(x_i) = \sum_{j=1}^i f_e(x_j) \quad \text{et} \quad FD = F_{1,e}(x_i) = 1 - F_e(x_i) = \sum_{j=i+1}^k f_e(x_j)$$

Ces fonctions f_e et F_e (e pour échantillon) sont définies pour un échantillon.

Pour une population ces fonctions sont $f(x)$ et $F(x)$ et sont définies comme suit :

$$f = \lim_{\substack{N \rightarrow \infty \\ \Delta x \rightarrow 0}} \left\{ \frac{f_e(x)}{\Delta x} \right\} \quad \text{et} \quad F(x) = \lim_{\substack{N \rightarrow \infty \\ \Delta x \rightarrow 0}} F_e(x)$$

$f(x)$ = fonction de densité de probabilité (f.d.p.)

$F(x)$ = fonction de répartition (f.r.)

Pour une variable continue, un histogramme normalisé en densité estime $f(x)$: la somme des aires des rectangles vaut 1. Réduire les classes sans augmenter suffisamment la taille de l'échantillon ne garantit pas une estimation régulière.

Variable aléatoire discrète (VAD)

Définition d'une VAD :

- 1) X peut prendre les valeurs $x_1, x_2, \dots, x_n$
- 2) Les probabilités respectives sont $f(x_1), f(x_2), \dots, f(x_n)$
- 3) Caractéristiques des probabilités :

$$\sum_{i=1}^n f(x_i) = 1 \text{ avec } 0 \leq f(x_i) \leq 1$$

- 4) Fonction cumulative F :

$$F(x_i) = \sum_{j=1}^i f(x_j)$$

$$f(x_i) = F(x_i) - F(x_{i-1})$$

Probabilité que X se trouve entre les valeurs a et b :

$$P(X \leq a = x_i) = F(x_i)$$

$$P(X \leq b = x_j) = F(x_j) \quad x_j > x_i$$

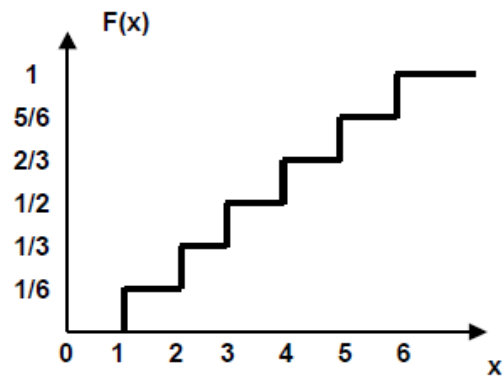
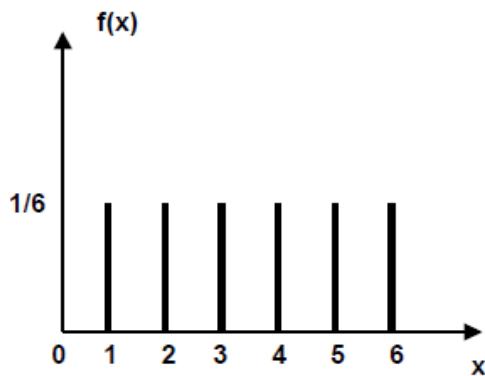
Chapitre 1 : Analyse fréquentielle et statistique appliquée à l'hydrologie

$$P(a < X \leq b) = F(x_j) - F(x_i) = \sum_{x_k > a}^{x_k \leq b} f(x_k)$$

Exemple 1 :

Dé à 6 faces classique (dé non lancé) :

Résultat n°	i	1	2	3	4	5	6
Valeurs possibles	x_i	1	2	3	4	5	6
Probabilité	$f(x_i)$	1/6	1/6	1/6	1/6	1/6	1/6



$$P(2 \leq X \leq 5) = \sum_{2 \leq x_i \leq 5} f(x_i) = f(2) + f(3) + f(4) + f(5) = 2/3$$

Variable aléatoire continue (VAC)

C'est une variable qui peut prendre toutes les valeurs d'un intervalle fini ou infini, et pour laquelle il existe une fonction $f(x)$, appelée fonction de densité de probabilité, telle que :

$$(1) f(x) \geq 0 \forall x$$

$$(2) \int_{-\infty}^{\infty} f(x) dx = 1$$

$$(3) P(a \leq X \leq b) = \int_a^b f(x) dx$$

Chapitre 1 : Analyse fréquentielle et statistique appliquée à l'hydrologie

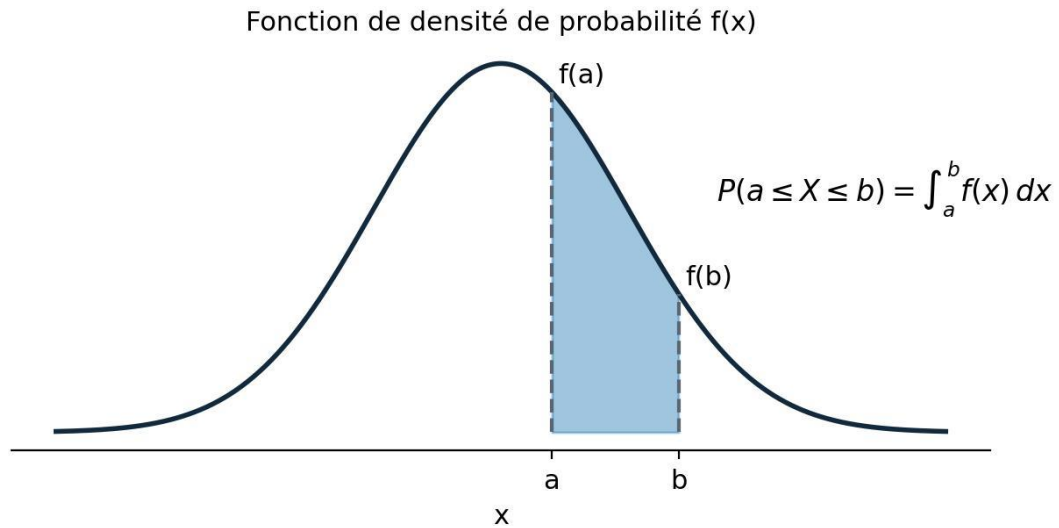


Figure : illustration de la fonction de densité de probabilité.

Fonction de répartition :

$$F(x) = P(X \leq x) = \int_{-\infty}^x f(x) dx$$

Relation entre $f(x)$ et $F(x)$

$$f(x) = \frac{dF(x)}{dx}$$

Propriétés de $F(x)$:

$$0 \leq F(x) \leq 1$$

$$F(-\infty) = 0 \text{ et } F(\infty) = 1$$

$$P(a \leq X \leq b) = F(b) - F(a)$$

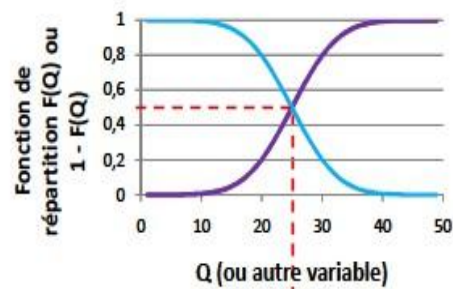
Remarque

En génie, beaucoup de variables sont bornées à gauche, c'est-à-dire ≥ 0 (débit, vitesse du vent, etc.).

Dès lors :

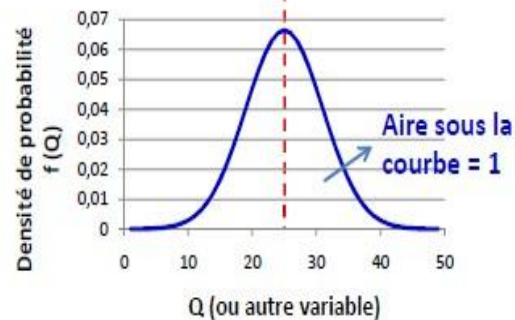
$$F(x) = \int_0^x f(x) dx \text{ et } \int_0^{\infty} f(x) dx = 1$$

- Fonction de répartition (probabilités au non-dépassement théoriques)
 $F(x)$ et $1 - F(x)$ (dépassement)



- Fonction de densité de probabilité

$$f(x) = \frac{d}{dx} F(x) \text{ et donc } F(x) = \int_{-\infty}^x f(x) dx$$



Les données hydrologiques doivent être traitées statistiquement selon trois étapes principales :

a) Description

Il est d'usage de condenser l'information hydrologique s'étendant sur de nombreuses années à une station fixe et de la remplacer par quelques caractéristiques bien choisies, à condition, toutefois, que ces dernières représentent la série chronologique de manière quasi exhaustive.

La statistique descriptive s'adapte parfaitement à ce type de problème. Elle définit certains paramètres types analysant fidèlement le phénomène à étudier.

La série d'observations définie sera classée et décrite par trois types de paramètres :

- Valeurs centrales
- Paramètres de dispersion
- Caractéristiques de forme des courbes de fréquence : histogrammes

b) L'analyse

Cette étape consiste à formaliser les données expérimentales par une expression mathématique tenant compte des valeurs types calculées à la première étape.

Le problème à résoudre sera de choisir le modèle probabiliste adéquat qui représente au mieux la série expérimentale. On appelle cette phase la recherche d'un **ajustement théorique** adéquat.

c) La prévision

Dans cette étape, l'ingénieur aura à projeter dans l'avenir le modèle choisi pour pouvoir organiser l'avenir de la façon la plus avantageuse et pour pouvoir prendre des décisions optimales et sécuritaires.

Remarque :

Chapitre 1 : Analyse fréquentielle et statistique appliquée à l'hydrologie

Ces 3 phases de l'analyse statistique supposent que les données utilisées sont homogènes, indépendantes et qu'elles ont subi la phase de **contrôle et de critique**, phase ultime pour la fiabilité des résultats qui dépend de la qualité des données.

Dans les exposés des méthodes statistiques qui suivront, nous supposerons que l'indépendance et l'homogénéité existent soit d'emblée, soit après correction des données.

1.1.2 Probabilité d'apparition d'une valeur – Intervalle moyen de récurrence

Pour le dimensionnement des structures conditionnées par un phénomène naturel, il est important que l'ingénieur connaisse la probabilité d'apparition de ce phénomène et son importance afin d'établir des critères de dimensionnement adéquats. Des considérations socioéconomiques ont établi certaines règles (règles de l'Art) qui font que l'ingénieur assurera des capacités hydrauliques variables, selon le dommage qui peut résulter d'un événement dépassant cette capacité.

Exemples illustratifs : choix de la fréquence de dimensionnement des ouvrages

Les deux exemples suivants sont donnés à titre illustratif ; les valeurs réellement retenues dépendent de la réglementation, de la classe de l'ouvrage et du niveau de risque accepté (voir les ordres de grandeur ci-après). Un déversoir de barrage pourra par exemple être conçu pour évacuer une crue d'une **fréquence probable de 0,001 (soit T = 1 000 ans)** à cause des dommages à la propriété et des pertes de vie qui résulteraient d'une crue supérieure à la capacité du déversoir (danger pour la stabilité du barrage lui-même).

Par contre un égout pluvial peut être insuffisant de temps à autre, créant certains dommages à la propriété, dommages d'un montant inférieur toutefois au capital qu'il aurait fallu investir pour construire l'égout plus gros et éviter l'inondation.

La capacité de l'égout pluvial sera basée sur un débit de fréquence probable plus forte que le déversoir, par exemple de l'ordre de 0,2 (soit T = 5 ans).

On définit l'intervalle moyen de récurrence, dit période de retour, par l'équation :

$$T = 1/P$$

P : probabilité qu'une valeur au moins égale à une valeur donnée se produise. La probabilité est exprimée en fraction.

Un débit d'inondation dont la probabilité **d'apparition ou de dépassement** est 0.033 est appelé crue de 30 ans (**$T=1/0.033=30$ ans**) car la probabilité est établie à l'aide des crues annuelles. L'unité de temps de T est la même que celle de la variable qui a servi à déterminer la probabilité.

Il ne faut pas conclure qu'un débit de 30 ans se produira à intervalles fixes de 30 ans ou que, s'étant produit une fois, il ne se produira plus pendant 30 ans. On doit comprendre plutôt que sur une longue période, 300 ans par exemple, il se produira en moyenne 10 crues au moins égales à ce débit.

Autrement dit, chaque année, il y a 3,3 % de chances *qu'un tel débit soit atteint*.

Les périodes de retour ci-dessous sont des ordres de grandeur indicatifs, couramment cités dans la littérature technique. Elles ne constituent pas une norme universelle : les valeurs à retenir dépendent de la réglementation et des guides techniques du pays, de la classe de l'ouvrage et du niveau de risque accepté (enjeux humains et économiques à l'aval). À titre indicatif :

- a) Déversoirs de barrages où une crue dépassant la capacité peut mettre le barrage en danger et créer des dommages considérables et des pertes de vie : 500 à 1 000 ans, voire 10 000 ans.
- b) Ponts sur routes importantes où l'exhaussement de l'eau créé par le pont peut entraîner des dommages importants ou la perte du pont : 50 à 100 ans.
- c) Ponts sur routes secondaires ou ponceaux sur routes importantes : 25 ans.

Chapitre 1 : Analyse fréquentielle et statistique appliquée à l'hydrologie

d) Ponceaux sur routes secondaires, égouts pluviaux, fossés de drainage : 5 à 10 ans.

e) Égouts pluviaux de moindre importance : 1 à 2 ans.

Le calcul des probabilités permet de calculer quelles sont les chances de non-apparition d'une valeur égale ou supérieure à X connaissant sa probabilité d'apparition au cours de l'année calculée par : $p = 1/T$.

La probabilité de non-apparition se calcule par : $q = 1 - P = 1 - 1/T$

1.1.3 Distribution d'une série statistique

Un échantillon hydrologique formé de n valeurs constitue une série continue dont l'effectif ou la fréquence d'apparition varie avec la valeur de la variable. On peut construire avec cet échantillon un histogramme ou polygone des fréquences d'apparition, une courbe des valeurs classées, une courbe de distribution des fréquences.

a. Polygone des fréquences d'apparition

Soit un échantillon de N observations décrivant une variable aléatoire X ($x_1, x_2, \dots, x_N$).

La construction d'un histogramme des fréquences de la variable X consiste à graduer l'axe des abscisses en valeur croissante de la variable étudiée et découpé en intervalles de classes. On porte alors en ordonnée le nombre d'apparitions constatées dans chaque intervalle. On obtient ainsi un graphique en « escalier ».

b. Courbe des valeurs classées

Ces courbes sont obtenues en portant :

En ordonnée : les valeurs observées, classées en ordre décroissant

En abscisse : la fréquence d'apparition de l'ensemble des valeurs supérieures à la valeur portée en ordonnée.

Ce sont des courbes qui permettent de donner la probabilité (en %) qu'une valeur observée soit égale ou dépassée. Elles ont généralement l'allure d'un S couché.

c. La courbe de distribution des fréquences

On peut construire soit la distribution des fréquences cumulées au non-dépassement soit la distribution des fréquences cumulées au dépassement.

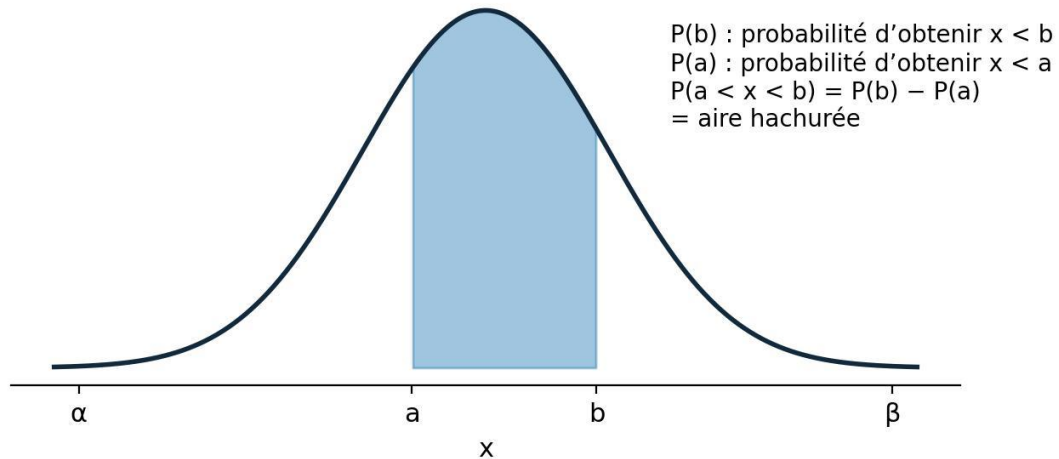
Si on calcule les fréquences cumulatives de toutes les valeurs inférieures ou égales à une valeur donnée x_i , on obtient la fréquence cumulée au non-dépassement de cette valeur. Le classement, dans ce cas, des valeurs observées, doit être fait en ordre croissant.

Inversement, le classement des données de l'échantillon par ordre décroissant et le calcul de la fréquence cumulée de toutes les valeurs supérieures à une valeur donnée x_i , permet de tracer le graphique des fréquences cumulées au dépassement.

d. Fonction de densité de probabilité - Fonction de répartition

Lorsque la taille de l'échantillon augmente, un histogramme correctement normalisé peut approcher la densité $f(x)$. Pour une variable continue, $P(X = x) = 0$; la densité $f(x)$ n'est pas une probabilité ponctuelle et peut dépasser 1.

Chapitre 1 : Analyse fréquentielle et statistique appliquée à l'hydrologie



Soit la courbe décrivant les fréquences d'apparition des événements d'une variable X .

$$Prob(a \leq x \leq b) = P(b) - P(a)$$

On peut démontrer que la probabilité d'avoir $a < x < b$

est égale à la surface hachurée sous la courbe.

$$Prob(a \leq x \leq b) = P(b) - P(a) = \int_a^b f(x)dx$$

Si α est la plus petite valeur possible de x et β la plus grande, on aura par définition : $\int_{\alpha}^x f(x)dx = P(x)$ et $\int_{\alpha}^{\beta} f(x)dx = 1$

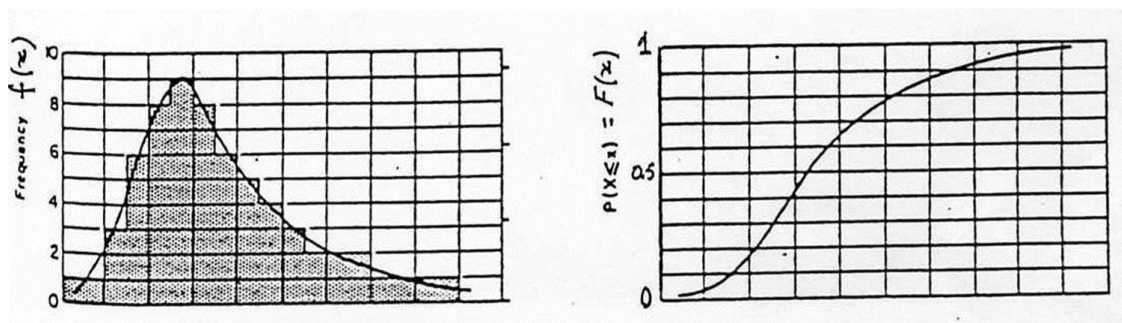
La dernière expression est évidente, puisqu'elle exprime que la probabilité pour que x soit situé entre les deux extrêmes possibles, est égale à 1 (ou 100 % de chances).

Elle exprime aussi que la totalité de l'aire sous la courbe de la fonction de distribution est égale à l'unité.

La moyenne et l'écart-type sont deux caractéristiques importantes d'une distribution, mais ne suffisent pas, en général, à la définir entièrement (asymétrie, aplatissement et paramètres de forme peuvent également intervenir).

De même, la limite du polygone des fréquences cumulées définit une fonction de répartition appelée loi de répartition de la population, notée $F(x)$ telle que :

$$\frac{dF(x)}{dx} = f(x)$$



Fonction de répartition

Soit un échantillon X de valeurs observées $(x_1, \dots, x_N)$, classées en ordre croissant.

Chapitre 1 : Analyse fréquentielle et statistique appliquée à l'hydrologie

Fonction de répartition empirique : $F_n(x)$ est la proportion d'observations inférieures ou égales à x .

Pour une population continue, on utilise l'intégrale ci-dessous ; pour une population discrète, la somme des masses de probabilité.

C'est la probabilité qu'une valeur de la variable X soit inférieure ou égale à la valeur x_i .

$$F(x_i) = \int_{-\infty}^{x_i} f(x) dx$$

Cette fonction est connue aussi sous le nom de la fonction de répartition au non-dépassement. Elle prend des valeurs de 0 à 1.

Par opposition, le complément de cette fonction s'appelle la fonction de répartition au dépassement et se définit par :

$$F_1(x_i) = P(X > x_i) = 1 - F(x_i)$$

Le dépassement vaut $1 - F(x)$ pour l'événement $X > x$. Un classement décroissant facilite son tracé mais n'est pas nécessaire à sa définition.

Une loi non rejetée par un test demeure un modèle plausible de l'échantillon. Ce résultat ne démontre pas qu'elle est la loi réelle du phénomène, ni qu'elle décrit avec certitude des événements bien au-delà des observations.

Les échantillons étant petits en hydrologie, il sera parfois difficile de déclarer si une loi de probabilité s'applique d'une façon absolue à un phénomène. On devra souvent se contenter d'utiliser la loi qui semble le mieux s'appliquer au phénomène étudié, représentée d'une façon plus ou moins parfaite par l'échantillon disponible.

1.1.4 Utilisation de modèles fréquents : les principales lois de distribution utilisées en hydrologie

Compte tenu de la taille parfois limitée des échantillons hydrologiques et des incertitudes associées aux observations, il est difficile de discriminer finement entre de nombreuses lois théoriques ; l'hydrologue a donc recours, en pratique, à un nombre limité de lois :

- La loi normale peut convenir à certaines sommes ou moyennes hydrologiques, si leurs conditions statistiques le permettent ; elle doit être contrôlée et sa probabilité de valeurs négatives examinée pour les grandeurs positives.
- La loi de Gumbel est un modèle de maxima à asymétrie fixe. Son adéquation ne peut être présumée pour tous les extrêmes ; la GEV offre un paramètre de forme supplémentaire.
- La loi de Pearson III est une loi gamma éventuellement translatée. Sa forme est contrôlée par son paramètre de forme ou son asymétrie ; le coefficient de variation seul ne détermine pas une équivalence avec Gumbel.
- On utilise également **les lois de Galton**, où le logarithme de la variable suit une loi de Gauss et **la loi de Fréchet**, où le logarithme de la variable suit une loi de Gumbel.

Ces lois présentent l'avantage de pouvoir être utilisées comme des lois à trois paramètres, ce qui leur confère une souplesse presque aussi grande que celle de la loi de Pearson III.

D'autres familles peuvent être pertinentes selon la variable, la stratégie d'échantillonnage et le comportement des queues. Le choix repose sur les diagnostics, les hypothèses et l'incertitude des quantiles.

Remarque : le théorème central limite (TCL) et ses conditions. Soit $X_1, \dots, X_n$ des variables aléatoires de moyenne μ et d'écart-type σ . Le TCL affirme que la moyenne $\bar{X}_n$ tend vers la loi normale $N(\mu, \sigma^2/n)$ lorsque n augmente, à condition que :

Chapitre 1 : Analyse fréquentielle et statistique appliquée à l'hydrologie

- les variables soient indépendantes (ou faiblement dépendantes) ;
- elles aient la même loi, ou au moins qu'aucune d'elles ne domine la somme (condition de Lindeberg) ;
- leur variance soit finie ;
- n soit suffisamment grand : la convergence est d'autant plus lente que la loi de départ est dissymétrique.

Conséquences en hydrologie : Les modules et pluies annuels, qui cumulent de nombreux apports, sont souvent proches de la normale en climat humide. Ce n'est pas le cas en zone aride ou semi-aride : les apports sont fortement dissymétriques et les apports journaliers sont autocorrélés, ce qui viole l'indépendance. Les maxima annuels (pluies journalières maximales, débits de pointe) ne relèvent pas du TCL mais de la théorie des valeurs extrêmes (lois de Gumbel et GEV, §1.1.4.3 et §1.1.4.4). Dans tous les cas, la normalité doit être vérifiée par un test d'adéquation (§1.3).

Les lois normale, log-normale, Gumbel et GEV sont développées ci-dessous.

1.1.4.1 La loi Normale (loi de Gauss)

Distribution normale : loi de Gauss, notée $N(\mu, \sigma^2)$

Convention : dans ce polycopié, $N(\mu, \sigma^2)$ désigne la loi normale de moyenne μ et de variance σ^2 (la loi normale standard est donc $N(0, 1)$). Certains ouvrages écrivent $N(\mu, \sigma)$ en donnant l'écart-type comme second paramètre ; il faut toujours vérifier la convention utilisée avant de lire une table ou un logiciel.

Lorsqu'une variable x subit l'influence de causes nombreuses très petites et indépendantes les unes des autres, les valeurs de cette variable se distribuent suivant une fonction de distribution, dite « Normale ». C'est la loi la plus commune et la mieux étudiée. Sa fonction de densité de probabilité est définie par :

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

dont les deux seuls paramètres sont la moyenne μ et l'écart-type σ de la population étudiée.

Dans cette distribution, x peut varier de manière continue de $-\infty$ à $+\infty$ et

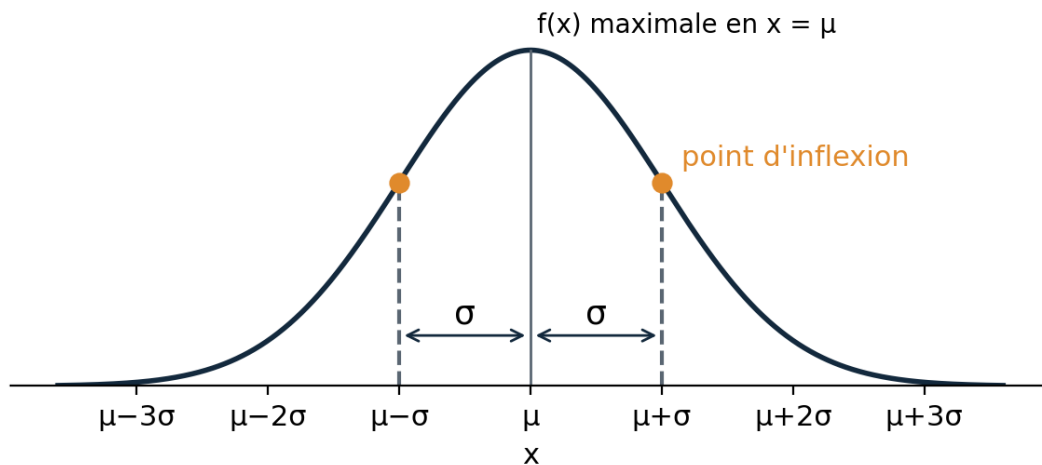
$$\int_{-\infty}^{+\infty} f(x)dx = 1$$

La fonction de distribution ou de répartition est définie par :

$$F(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^x e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx$$

La fonction $f(x)$ se représente par une courbe en forme de cloche, symétrique par rapport à $x=\mu$, admettant un maximum pour $x=\mu$ et tendant asymptotiquement vers 0 pour $x \rightarrow -\infty$ et $x \rightarrow +\infty$.

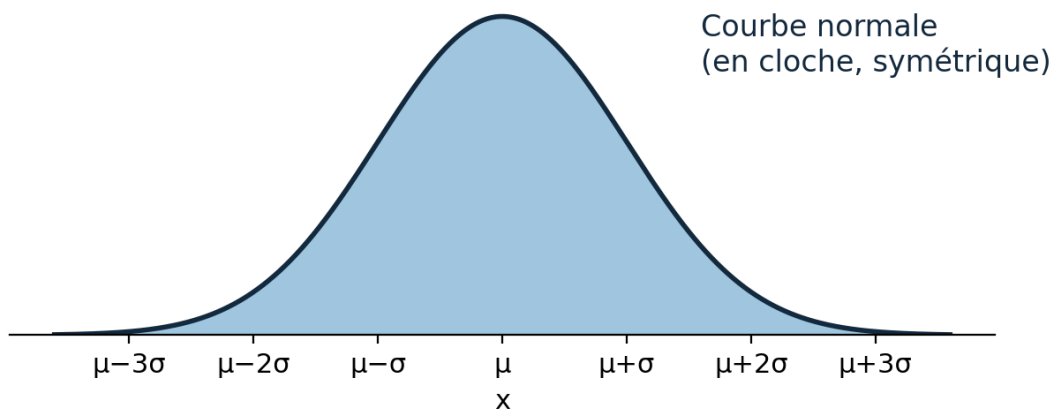
Chapitre 1 : Analyse fréquentielle et statistique appliquée à l'hydrologie



Les points d'inflexion se situent à $(\mu-\sigma)$ et $(\mu+\sigma)$.

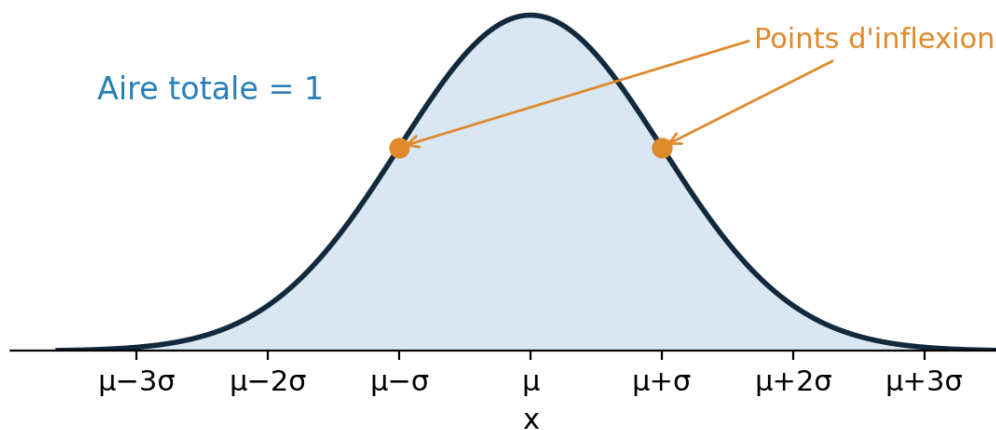
Propriétés de la distribution normale

La plus importante distribution des probabilités en statistiques est la distribution normale.



Une distribution normale est une distribution de probabilité continue d'une variable aléatoire, X . Le graphique d'une distribution normale est appelé courbe normale.

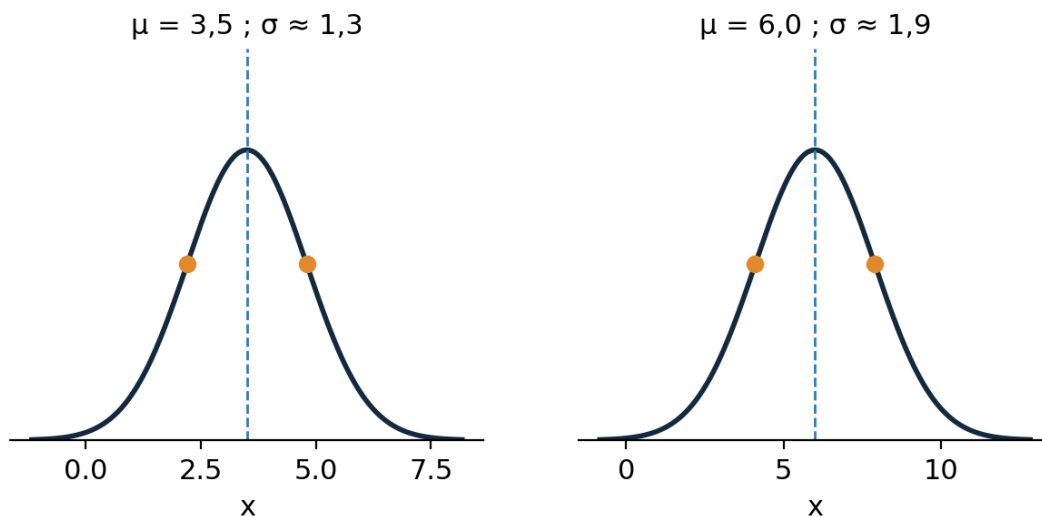
Chapitre 1 : Analyse fréquentielle et statistique appliquée à l'hydrologie



Il existe une famille entière de lois normales. Elles se différencient par leur moyenne et leur variance.

Moyennes et écarts-types

Une distribution normale peut avoir n'importe quelle moyenne et n'importe quel écart-type positif.



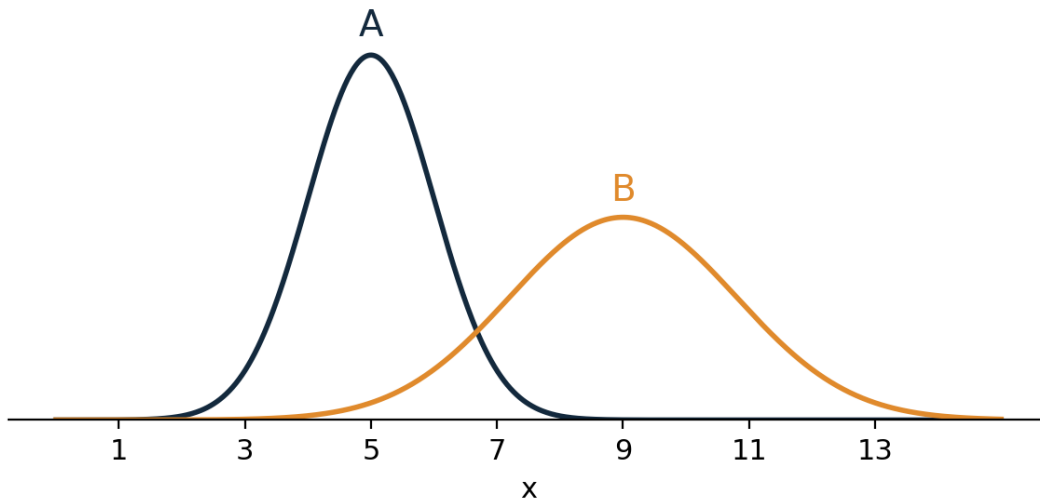
Points orange : points d'inflexion ($\mu \pm \sigma$) — tirets : axe de symétrie (moyenne)

L'écart-type décrit la dispersion des données.

Exemple :

- Quelle est la courbe qui a la plus grande moyenne ?
- Quelle est la courbe qui a le plus grand écart-type ?

Chapitre 1 : Analyse fréquentielle et statistique appliquée à l'hydrologie

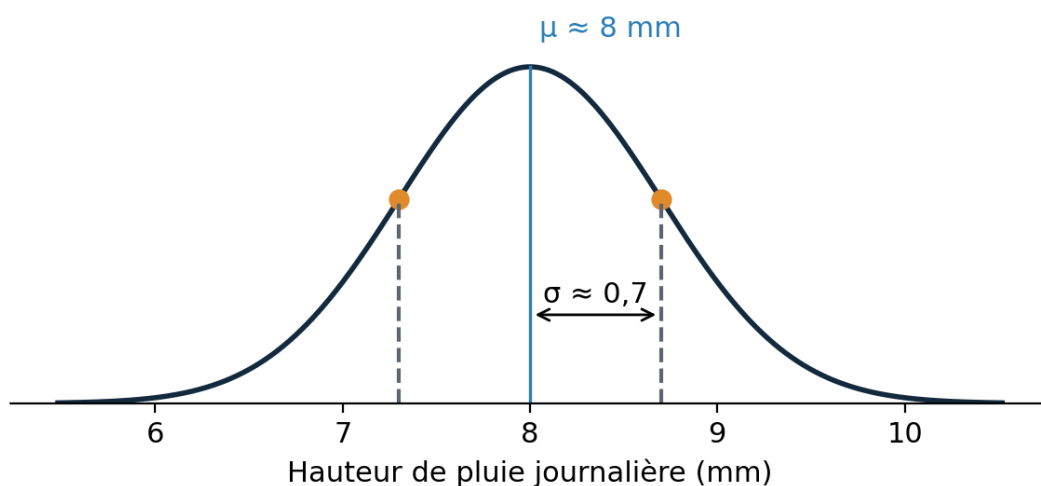


- L'axe de symétrie de la courbe A se produit à $x = 5$.
- L'axe de symétrie de la courbe B se produit à $x = 9$.
- La courbe B a la plus grande moyenne.
- La courbe B est plus étendue que la courbe A, de sorte que la courbe B a le plus grand écart-type.

Interprétation de graphes

Exemple :

Exemple purement pédagogique : on suppose ici une variable de hauteur de pluie de loi normale. Lire sa moyenne et son écart-type sur la courbe. Cette hypothèse ne décrit pas, en général, les pluies journalières réelles, qui comportent des jours secs et des valeurs positives dissymétriques.



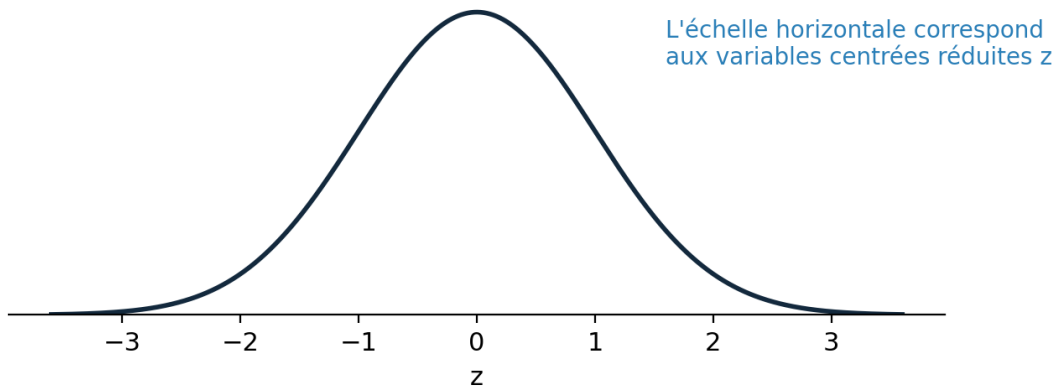
Les points d'inflexion sont à une distance d'un écart-type de la moyenne.

Dans cet exemple hypothétique, $\mu \approx 8 \text{ mm}$ et $\sigma \approx 0,7 \text{ mm}$.

La distribution normale standard

La distribution normale standard des probabilités est une distribution normale avec une moyenne nulle et un écart-type égal à 1, notée $N(0, 1)$.

Chapitre 1 : Analyse fréquentielle et statistique appliquée à l'hydrologie

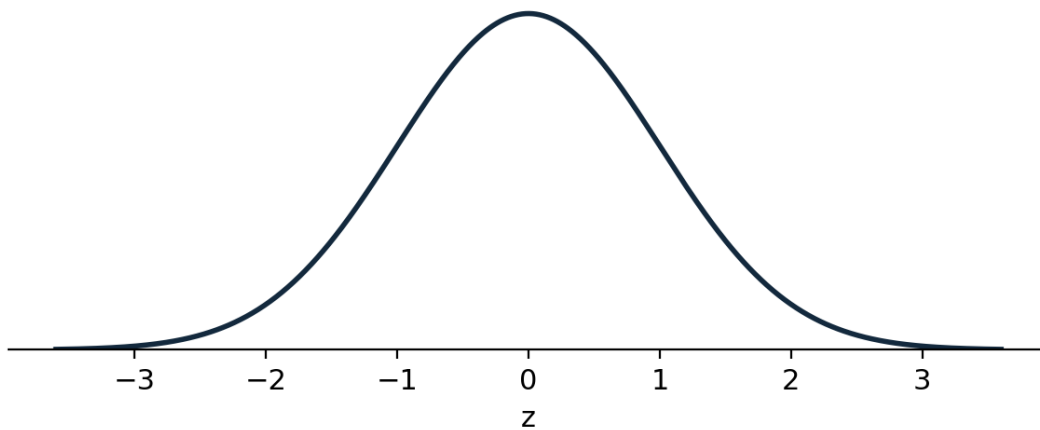


N'importe quelle valeur observée peut être transformée en une Variable Centrée Réduite (VCR) z en utilisant la formule :

$$Z = \frac{X - \mu}{\sigma} \sim \mathcal{N}(0,1)$$

Si chaque valeur d'une variable aléatoire x , distribuée normalement, est transformée en une variable centrée réduite z , le résultat sera la distribution normale standard.

Loi normale standard $\mathcal{N}(0, 1)$: $\mu = 0$, $\sigma = 1$



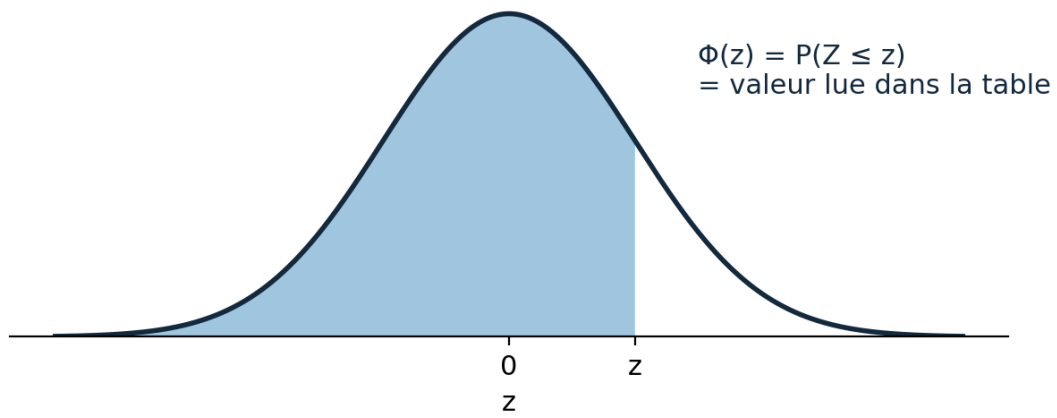
L'aire qui se situe dans l'intervalle sous la courbe normale non standard (pour les valeurs de x) est la même que celle de la courbe normale standard (dans les limites correspondantes à z).

La formule est utilisée pour transformer une valeur de x en une VCR z , alors que la table de la loi normale standard est utilisée pour calculer l'aire cumulée sous la courbe.

Chapitre 1 : Analyse fréquentielle et statistique appliquée à l'hydrologie

La table de la distribution normale standard

$$\mu = 0 ; \sigma = 1$$



Chapitre 1 : Analyse fréquentielle et statistique appliquée à l'hydrologie

Table de la loi normale standard : $\Phi(z) = P(Z \leq z)$, $z \geq 0$
(pour $z < 0$: $\Phi(-z) = 1 - \Phi(z)$)

z	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,5000	0,5040	0,5080	0,5120	0,5160	0,5199	0,5239	0,5279	0,5319	0,5359
0,1	0,5398	0,5438	0,5478	0,5517	0,5557	0,5596	0,5636	0,5675	0,5714	0,5753
0,2	0,5793	0,5832	0,5871	0,5910	0,5948	0,5987	0,6026	0,6064	0,6103	0,6141
0,3	0,6179	0,6217	0,6255	0,6293	0,6331	0,6368	0,6406	0,6443	0,6480	0,6517
0,4	0,6554	0,6591	0,6628	0,6664	0,6700	0,6736	0,6772	0,6808	0,6844	0,6879
0,5	0,6915	0,6950	0,6985	0,7019	0,7054	0,7088	0,7123	0,7157	0,7190	0,7224
0,6	0,7257	0,7291	0,7324	0,7357	0,7389	0,7422	0,7454	0,7486	0,7517	0,7549
0,7	0,7580	0,7611	0,7642	0,7673	0,7704	0,7734	0,7764	0,7794	0,7823	0,7852
0,8	0,7881	0,7910	0,7939	0,7967	0,7995	0,8023	0,8051	0,8078	0,8106	0,8133
0,9	0,8159	0,8186	0,8212	0,8238	0,8264	0,8289	0,8315	0,8340	0,8365	0,8389
1,0	0,8413	0,8438	0,8461	0,8485	0,8508	0,8531	0,8554	0,8577	0,8599	0,8621
1,1	0,8643	0,8665	0,8686	0,8708	0,8729	0,8749	0,8770	0,8790	0,8810	0,8830
1,2	0,8849	0,8869	0,8888	0,8907	0,8925	0,8944	0,8962	0,8980	0,8997	0,9015
1,3	0,9032	0,9049	0,9066	0,9082	0,9099	0,9115	0,9131	0,9147	0,9162	0,9177
1,4	0,9192	0,9207	0,9222	0,9236	0,9251	0,9265	0,9279	0,9292	0,9306	0,9319
1,5	0,9332	0,9345	0,9357	0,9370	0,9382	0,9394	0,9406	0,9418	0,9429	0,9441
1,6	0,9452	0,9463	0,9474	0,9484	0,9495	0,9505	0,9515	0,9525	0,9535	0,9545
1,7	0,9554	0,9564	0,9573	0,9582	0,9591	0,9599	0,9608	0,9616	0,9625	0,9633
1,8	0,9641	0,9649	0,9656	0,9664	0,9671	0,9678	0,9686	0,9693	0,9699	0,9706
1,9	0,9713	0,9719	0,9726	0,9732	0,9738	0,9744	0,9750	0,9756	0,9761	0,9767
2,0	0,9772	0,9778	0,9783	0,9788	0,9793	0,9798	0,9803	0,9808	0,9812	0,9817
2,1	0,9821	0,9826	0,9830	0,9834	0,9838	0,9842	0,9846	0,9850	0,9854	0,9857
2,2	0,9861	0,9864	0,9868	0,9871	0,9875	0,9878	0,9881	0,9884	0,9887	0,9890
2,3	0,9893	0,9896	0,9898	0,9901	0,9904	0,9906	0,9909	0,9911	0,9913	0,9916
2,4	0,9918	0,9920	0,9922	0,9925	0,9927	0,9929	0,9931	0,9932	0,9934	0,9936
2,5	0,9938	0,9940	0,9941	0,9943	0,9945	0,9946	0,9948	0,9949	0,9951	0,9952
2,6	0,9953	0,9955	0,9956	0,9957	0,9959	0,9960	0,9961	0,9962	0,9963	0,9964
2,7	0,9965	0,9966	0,9967	0,9968	0,9969	0,9970	0,9971	0,9972	0,9973	0,9974
2,8	0,9974	0,9975	0,9976	0,9977	0,9977	0,9978	0,9979	0,9979	0,9980	0,9981
2,9	0,9981	0,9982	0,9982	0,9983	0,9984	0,9984	0,9985	0,9985	0,9986	0,9986
3,0	0,9987	0,9987	0,9987	0,9988	0,9988	0,9989	0,9989	0,9989	0,9990	0,9990
3,1	0,9990	0,9991	0,9991	0,9991	0,9992	0,9992	0,9992	0,9992	0,9993	0,9993
3,2	0,9993	0,9993	0,9994	0,9994	0,9994	0,9994	0,9994	0,9995	0,9995	0,9995
3,3	0,9995	0,9995	0,9995	0,9996	0,9996	0,9996	0,9996	0,9996	0,9996	0,9997
3,4	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9998

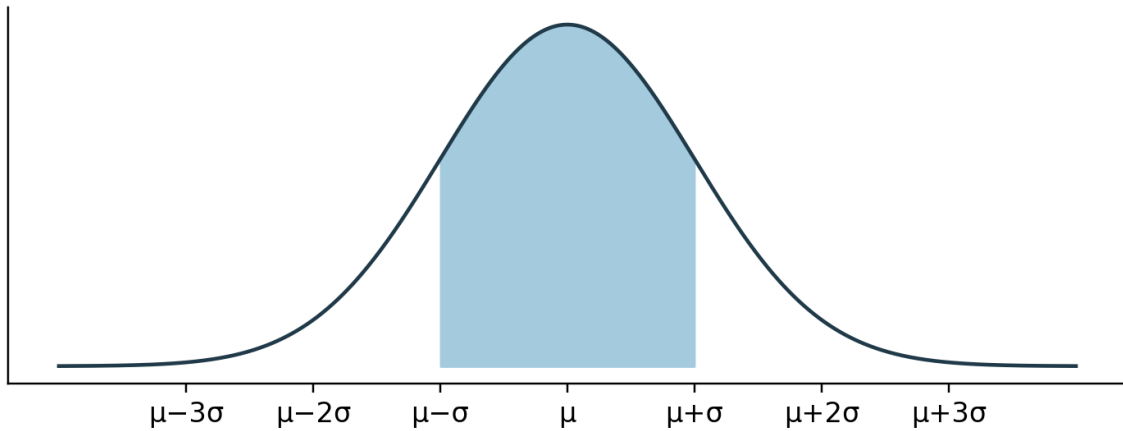
La table donne l'aire cumulée (probabilité) située à gauche de la VCR z , c'est-à-dire $P(Z \leq z)$.

Chapitre 1 : Analyse fréquentielle et statistique appliquée à l'hydrologie

Propriétés de la distribution normale

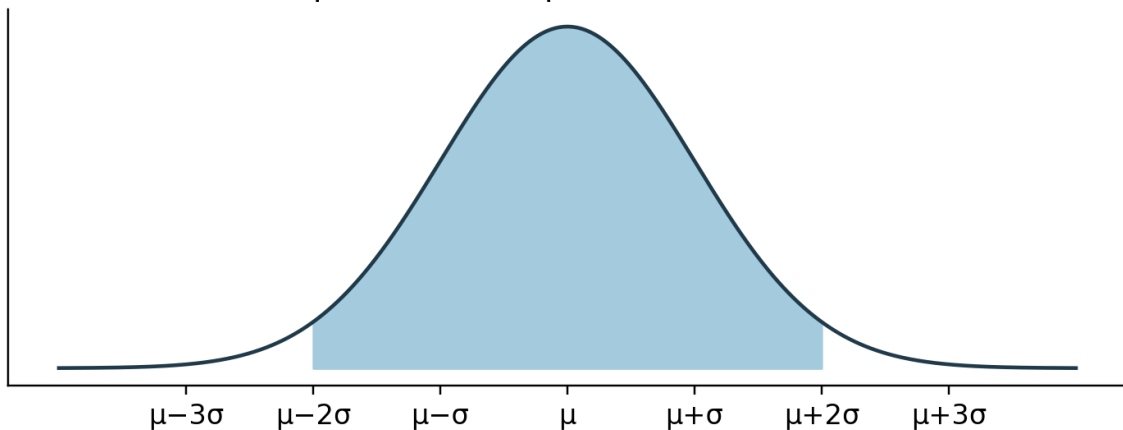
Règle empirique

$$P(\mu - 1\sigma \leq X \leq \mu + 1\sigma) = 68.27 \%$$

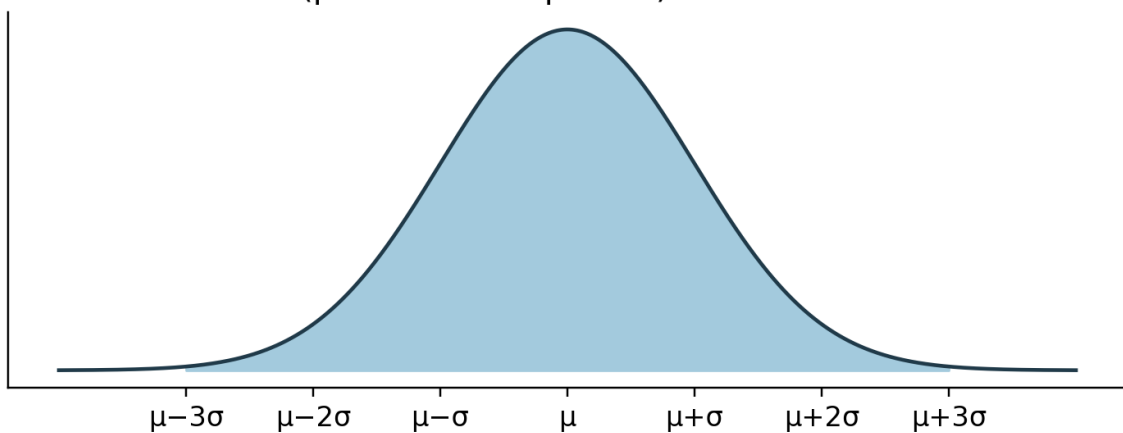


Pour une population normale, environ 68,27 % de la probabilité se situe entre $\mu - \sigma$ et $\mu + \sigma$. Les intervalles construits avec la moyenne et l'écart-type estimés de l'échantillon sont des approximations.

$$P(\mu - 2\sigma \leq X \leq \mu + 2\sigma) = 95.45 \%$$



$$P(\mu - 3\sigma \leq X \leq \mu + 3\sigma) = 99.73 \%$$



La distribution normale standard

Propriétés de la distribution normale standard :

1. La surface cumulée est proche de 0 pour des VCR z proches de $z = -3,49$.

Chapitre 1 : Analyse fréquentielle et statistique appliquée à l'hydrologie

2. La surface cumulée augmente à mesure que les VCR z augmentent.
3. La surface cumulée pour $z = 0$ est 0,5000.
4. La surface cumulée est proche de 1 pour des VCR z proches de $z = 3,49$.

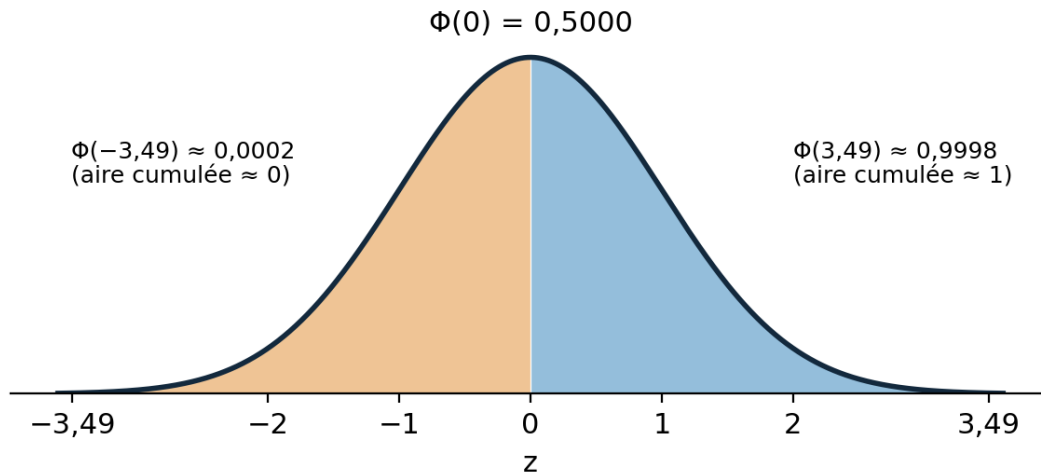


Table de la distribution normale standard

Exemple:

Trouvez l'aire cumulée qui correspond à une VCR $z = 2,71$.

Table de la distribution normale standard

2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981

Se déplacer dans la colonne de gauche jusqu'à trouver la valeur de $z = 2,7$, puis se déplacer en horizontale jusqu'à la colonne 0.01.

L'aire située à gauche de $z = 2,71$ est 0,9966.

Pour lire l'aire (probabilité) correspondant à une VCR z donnée :

- repérer la partie entière et la première décimale de z dans la colonne de gauche de la table de la loi normale standard, et la deuxième décimale dans la ligne du haut ;
- l'aire se lit à l'intersection de cette ligne et de cette colonne ;
- inversement, pour trouver la VCR z correspondant à une aire donnée, on cherche cette aire à l'intérieur de la table et on lit z dans les marges.

Il est important de distinguer la recherche d'une probabilité (aire) associée à une VCR z donnée de la recherche de la VCR z associée à une probabilité donnée.

Savoir reconnaître la bonne méthode pour chacun de ces deux cas est essentiel.

Utiliser la symétrie pour trouver l'aire à gauche de la moyenne

En raison de la symétrie, ces deux aires sont égales.

Chapitre 1 : Analyse fréquentielle et statistique appliquée à l'hydrologie

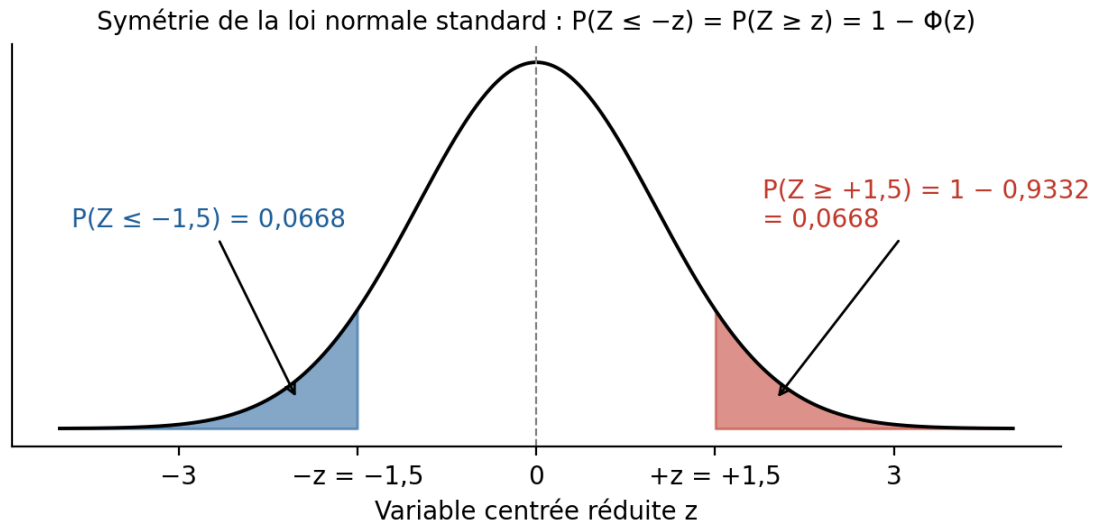


Figure 1.1 : Symétrie de la loi normale standard : $P(Z \leq -z) = P(Z \geq z) = 1 - \Phi(z)$.

NOTE : Même si la VCR z peut être négative, l'aire sous la courbe (ou la probabilité correspondante) ne peut jamais l'être.

Probabilité de la moitié d'une distribution

La courbe normale standard étant symétrique par rapport à $z = 0$, chaque moitié de la distribution contient exactement la moitié de la probabilité totale : $P(Z \leq 0) = P(Z \geq 0) = 0.5000$. Ce résultat permet de calculer une probabilité d'un seul côté de la moyenne, par exemple $P(0 \leq Z \leq z) = \Phi(z) - 0.5$.

Chaque moitié de la courbe normale représente une probabilité de 0,5

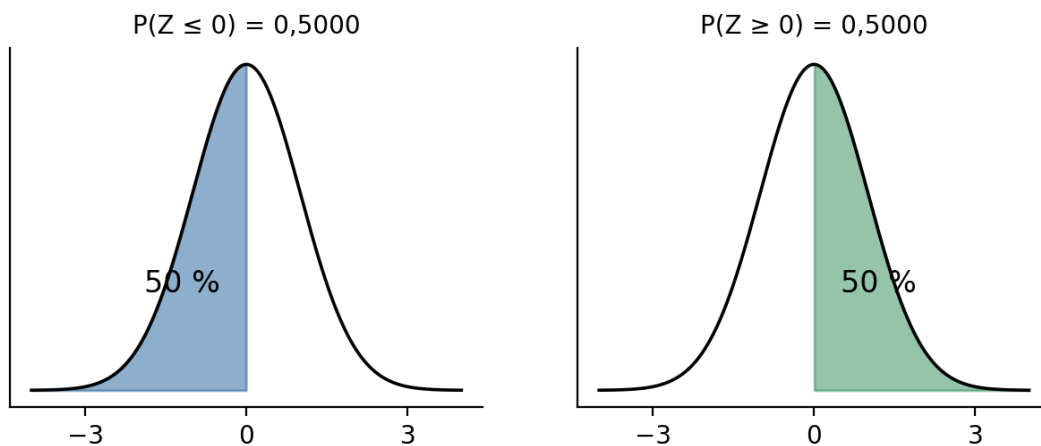


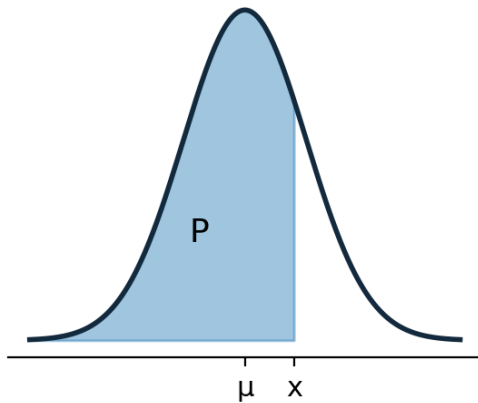
Figure 1.2 : Chaque moitié de la loi normale standard représente une probabilité de 0,5.

Conversion de distribution normale standard

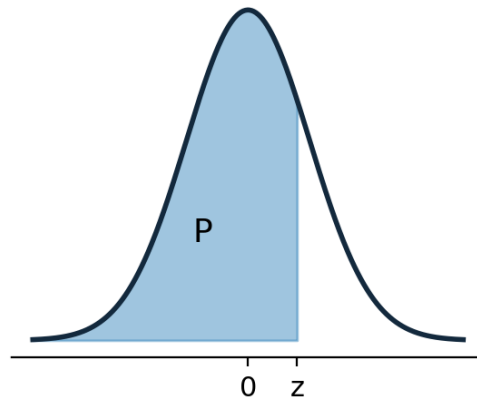
Chapitre 1 : Analyse fréquentielle et statistique appliquée à l'hydrologie

$$z = (x - \mu) / \sigma \rightarrow \text{même aire } P$$

(a) loi normale $N(\mu, \sigma^2)$



(b) loi normale standard $N(0, 1)$

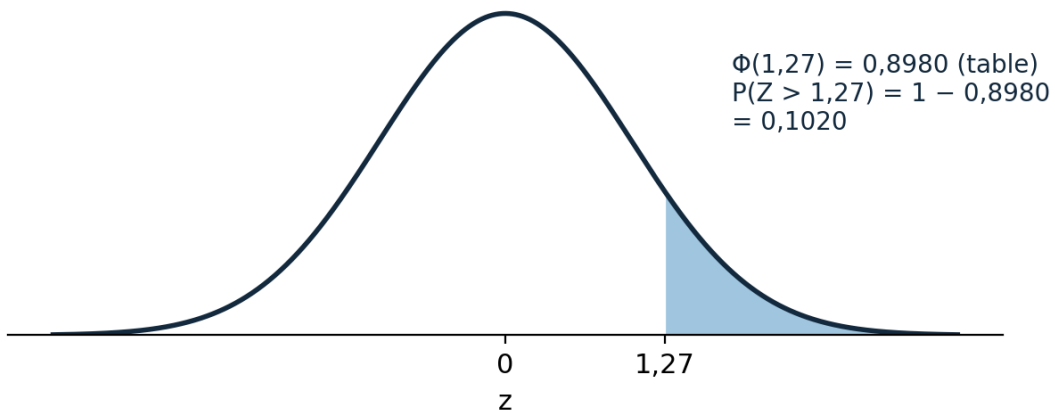


Note :

$$P(x < a) = \text{lecture directe}$$

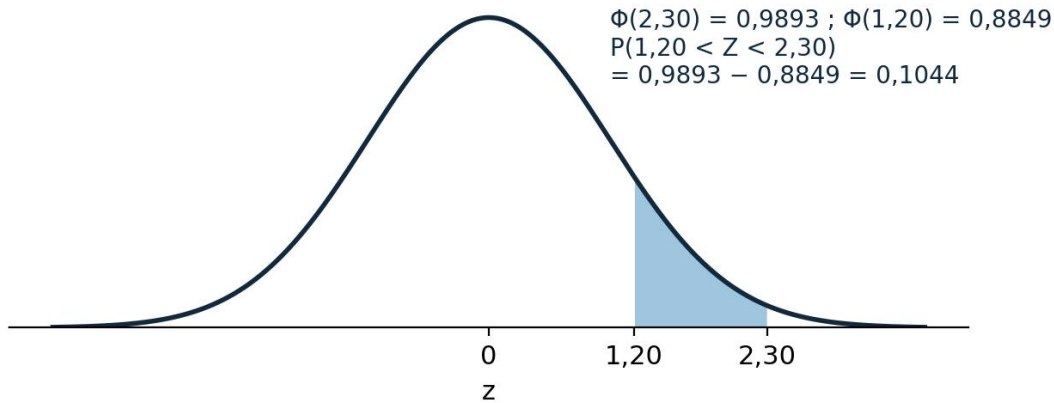
$$P(Z > a) = 1 - \Phi(a) ; \quad P(Z < -a) = 1 - \Phi(a) ; \quad P(a < Z < b) = \Phi(b) - \Phi(a)$$

Déterminer l'aire à droite de $z = 1.27$



Déterminer l'aire comprise entre $z = 1.20$ et $z = 2.30$

Chapitre 1 : Analyse fréquentielle et statistique appliquée à l'hydrologie

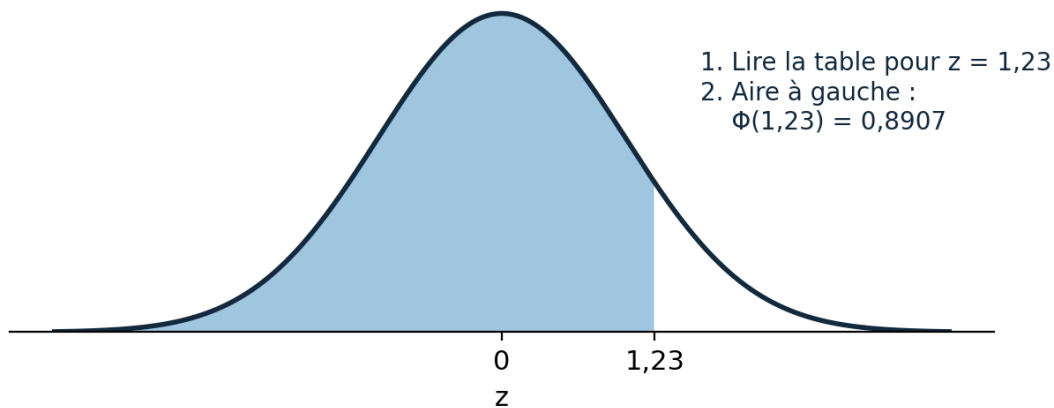


$$P(a < x < b) = P(x < b) - P(x < a)$$

Détermination des aires

Détermination des aires sous la courbe normale standard

1. Dessinez la courbe normale standard et ombragez la zone appropriée sous la courbe.
2. Trouvez l'aire en suivant les indications suivantes :
 - a. Pour déterminer l'aire située à gauche de z, trouver l'aire qui correspond à z dans la table de la loi normale standard.

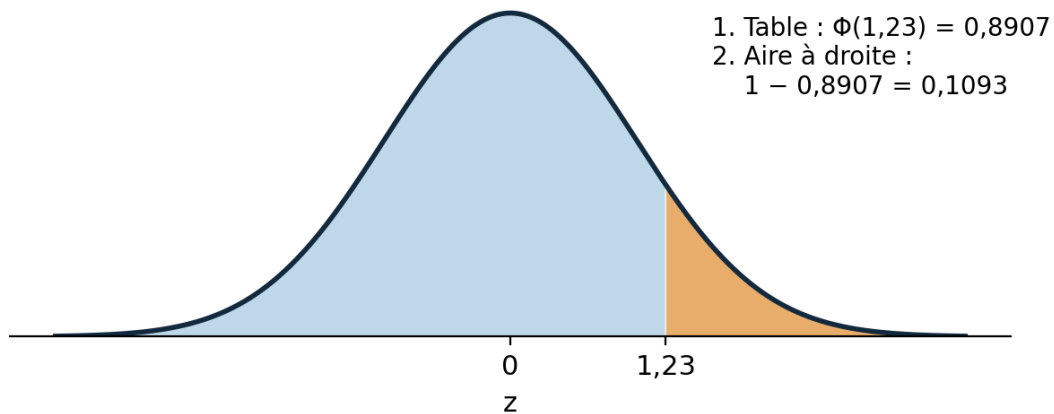


Détermination des aires sous la courbe normale standard

- b. Pour déterminer l'aire située à droite de z, utilisez la table de la loi normale standard pour trouver l'aire qui correspond à z.

Puis soustraire cette aire ou ce résultat de 1.

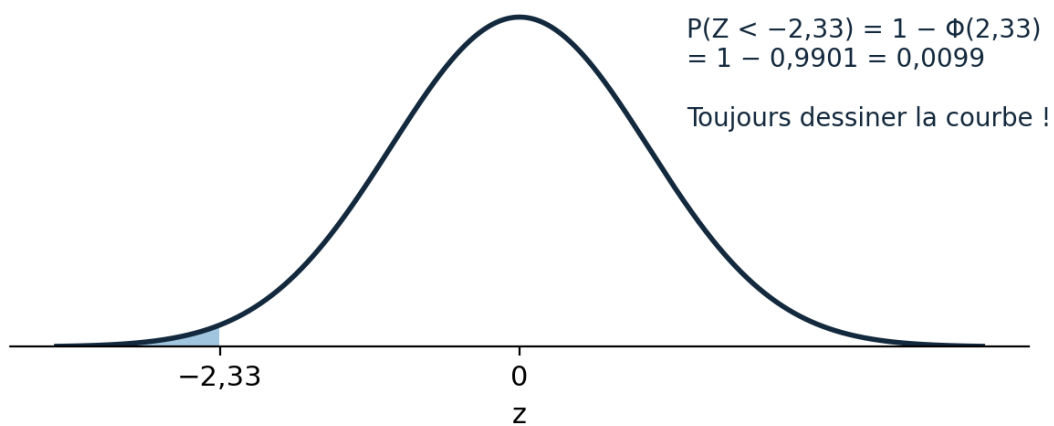
Chapitre 1 : Analyse fréquentielle et statistique appliquée à l'hydrologie



Détermination des aires

Exemple:

Trouvez l'aire sous la courbe normale standard à gauche de $z = -2,33$

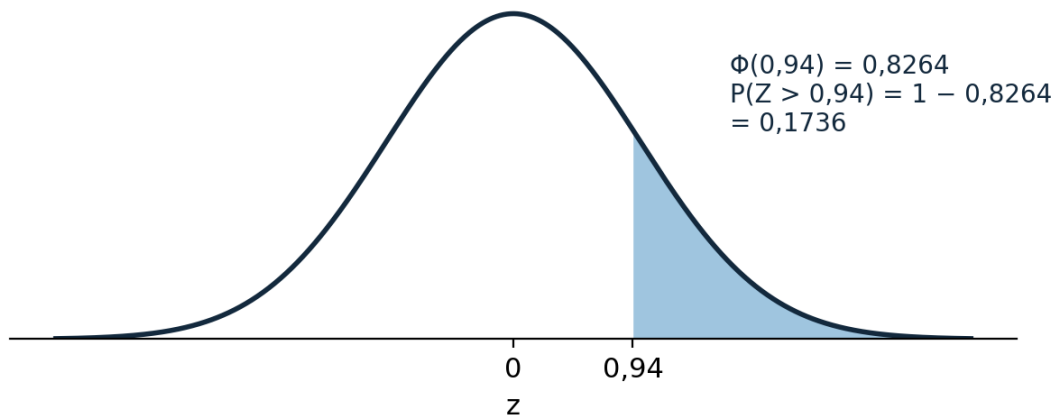


De la table de la loi normale standard, l'aire est égale à 0,0099.

Exemple:

Trouvez l'aire sous la courbe normale standard à droite de $z = 0.94$.

Chapitre 1 : Analyse fréquentielle et statistique appliquée à l'hydrologie

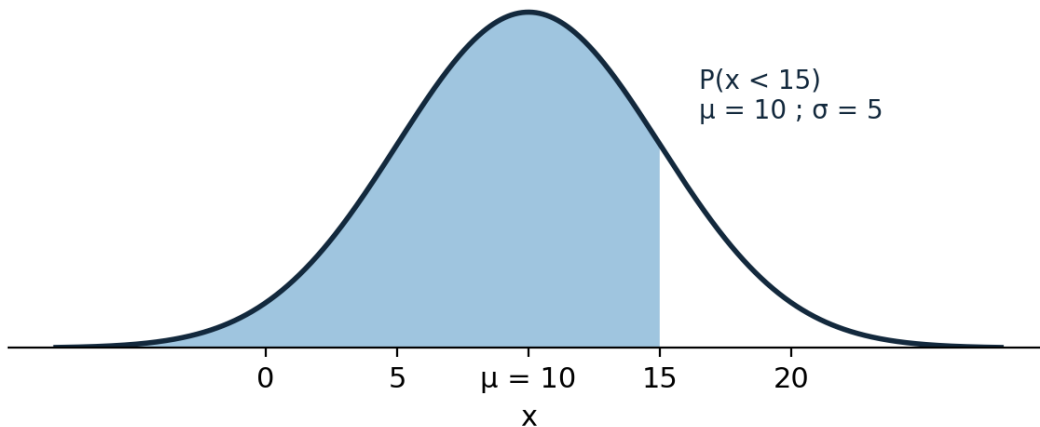


À partir de la table de la loi normale standard, l'aire est égale à 0,1736.

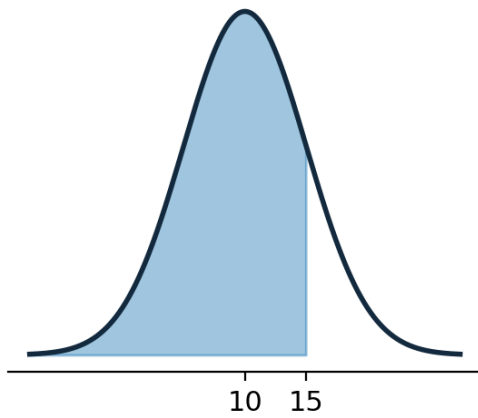
Probabilité et distributions normales

Si une variable aléatoire, x , est normalement distribuée, on peut déterminer la probabilité que x se trouve dans un intervalle donné par le calcul de l'aire sous la courbe normale pour cet intervalle.

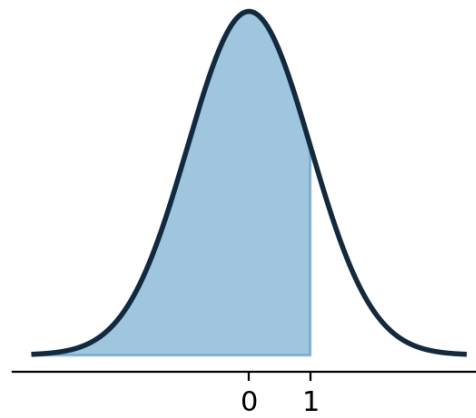
Chapitre 1 : Analyse fréquentielle et statistique appliquée à l'hydrologie



Loi normale $N(10 ; 5^2)$



Loi normale standard $N(0, 1)$

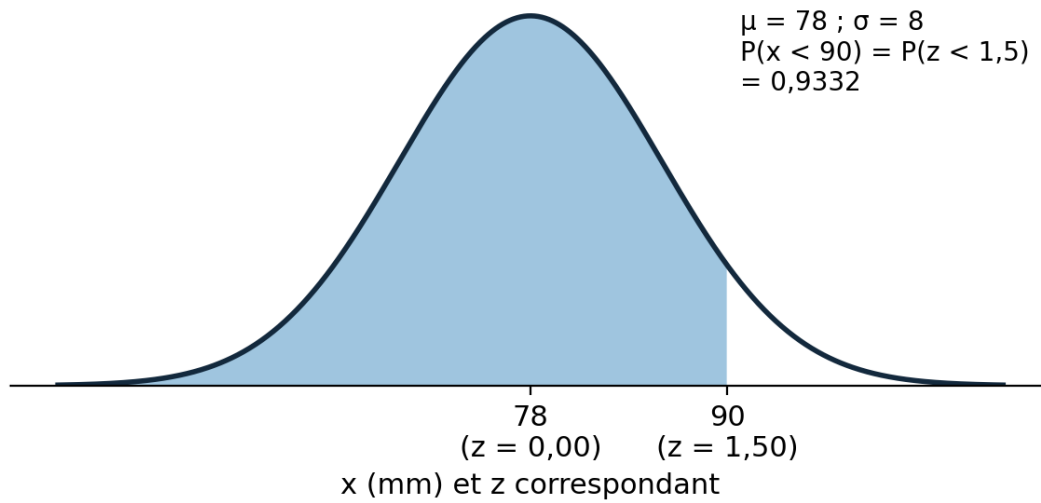


$$P(x < 15) = P(z < 1) = 0,8413 : \text{même aire}$$

Exemple:

La moyenne d'une série de pluies était de 78 mm avec un écart-type de 8 mm. Si ces pluies sont normalement distribuées, trouver la probabilité que la pluie soit inférieure à 90 mm.

Chapitre 1 : Analyse fréquentielle et statistique appliquée à l'hydrologie



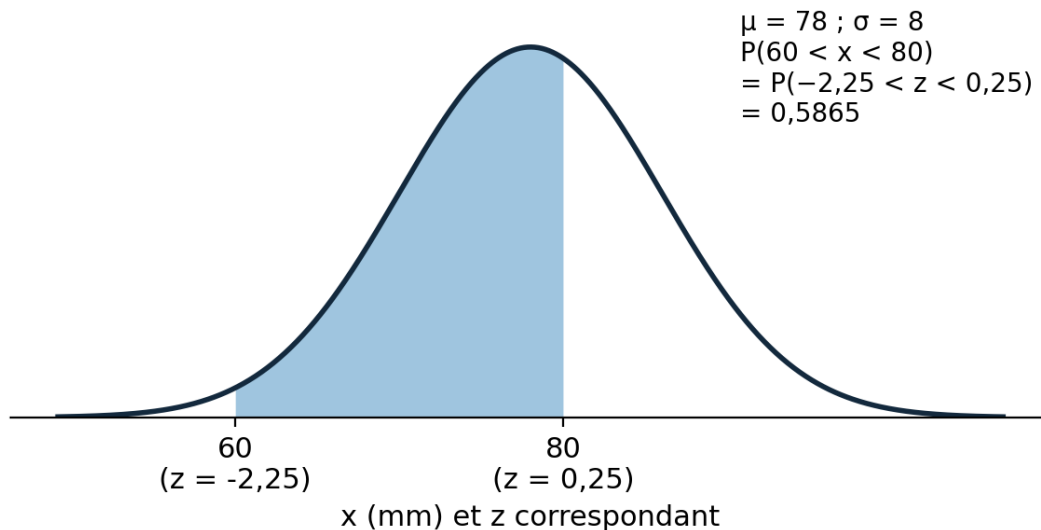
$$Z = \frac{X - \mu}{\sigma} = \frac{90 - 78}{8} = 1.5$$

La probabilité qu'une pluie soit inférieure à 90 mm est 0,9332.

$$P(x < 90) = P(z < 1.5) = 0.9332$$

Exemple:

La moyenne d'une série de pluies était de 78 mm avec un écart-type de 8 mm. Si ces pluies sont normalement distribuées, trouver la probabilité que la pluie soit comprise entre 60 et 80 mm.



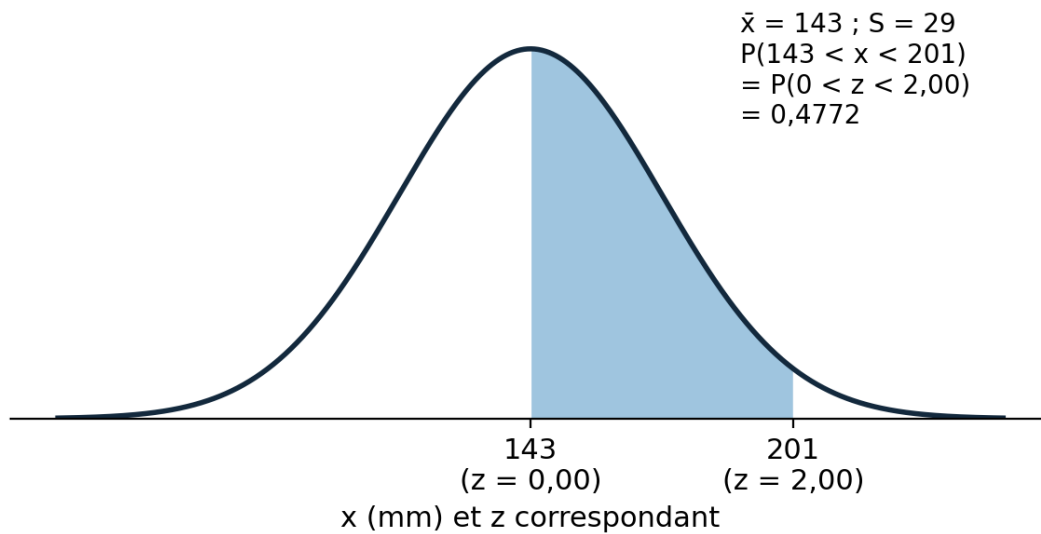
$$Z_1 = \frac{x - \mu}{\sigma} = \frac{60 - 78}{8} = -2.25 \quad Z_2 = \frac{x - \mu}{\sigma} = \frac{80 - 78}{8} = 0.25$$

La probabilité qu'une pluie soit comprise entre 60 et 80 mm est 0,5865.

$$P(60 < x < 80) = P(-2.25 < z < 0.25) = P(z < 0.25) - 1 + P(z < 2.25) = 0.5987 - 1 + 0.9878 = 0.5865$$

Probabilité d'une pluie annuelle comprise entre 143 et 201 mm

Chapitre 1 : Analyse fréquentielle et statistique appliquée à l'hydrologie



Solution : avec $\bar{x} = 143$ mm et $S = 29$ mm, les bornes réduites sont $z_1 = (143 - 143)/29 = 0$ et $z_2 = (201 - 143)/29 = 2,00$.

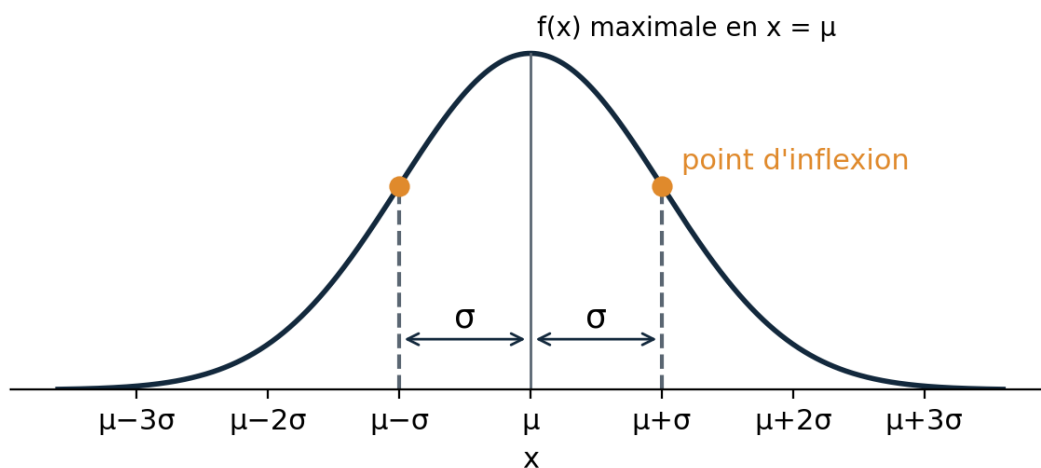
$$P(143 < x < 201) = P(0 < z < 2,00) = \Phi(2,00) - \Phi(0) = 0,9772 - 0,5000 = 0,4772$$

Il y a donc environ 47,7 % de chances que la pluie annuelle soit comprise entre 143 et 201 mm.

Loi Normale ou loi de Gauss

Fonction de densité : $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

Fonction de distribution : $F(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^x e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx$



La moyenne : $\bar{X} = \frac{\sum x}{n}$

L'écart-type : $S = \sqrt{\frac{\sum (x - \bar{x})^2}{N-1}}$

La droite de Henry : $x = \bar{x} + z \times s$

Chapitre 1 : Analyse fréquentielle et statistique appliquée à l'hydrologie

Intervalles de confiance sous hypothèse de normalité

On suppose des observations indépendantes issues d'une même loi $N(\mu, \sigma^2)$. Un intervalle à $100(1 - \alpha)$ % est construit par une procédure qui contient le paramètre fixe dans cette proportion de répétitions.

Après observation, on ne lui attribue pas une probabilité fréquentiste d'être dans l'intervalle calculé.

Moyenne si σ est connu : intervalle exact normal. Si σ est inconnu, utiliser l'écart-type empirique S et Student à $N - 1$ degrés de liberté.

$$\bar{x} \pm z_{1-\alpha/2} \frac{\sigma}{\sqrt{N}} \quad (\sigma \text{ connu})$$

$$\bar{x} \pm t_{N-1;1-\alpha/2} \frac{S}{\sqrt{N}} \quad (\sigma \text{ inconnu})$$

Écart-type : l'intervalle exact n'est généralement pas symétrique autour de S . On note $\chi^2_{v;p}$ le quantile de probabilité cumulée p de la loi du khi-deux à $v = N - 1$ degrés de liberté.

$$\sqrt{\frac{(N-1)S^2}{\chi^2_{N-1;1-\alpha/2}}} \leq \sigma \leq \sqrt{\frac{(N-1)S^2}{\chi^2_{N-1;\alpha/2}}}$$

Quantile : $q_p = \mu + z_p \sigma$ est inconnu ; son estimateur est $\hat{q}_p = \bar{x} + z_p S$. Pour un échantillon assez grand, une approximation par la méthode delta est :

$$\hat{q}_p = \bar{x} + z_p S ; \quad SE(\hat{q}_p) \simeq S \sqrt{\frac{1}{N} + \frac{z_p^2}{2(N-1)}}$$

$$IC_{1-\alpha}(q_p) \simeq \hat{q}_p \pm z_{1-\alpha/2} SE(\hat{q}_p)$$

L'approximation des quantiles doit être utilisée avec prudence pour des échantillons courts ou des probabilités très proches de 1 ; un bootstrap paramétrique avec réestimation des paramètres est alors utile. Les intervalles ci-dessus ne couvrent ni les erreurs de mesure ni l'incertitude de choix de loi.

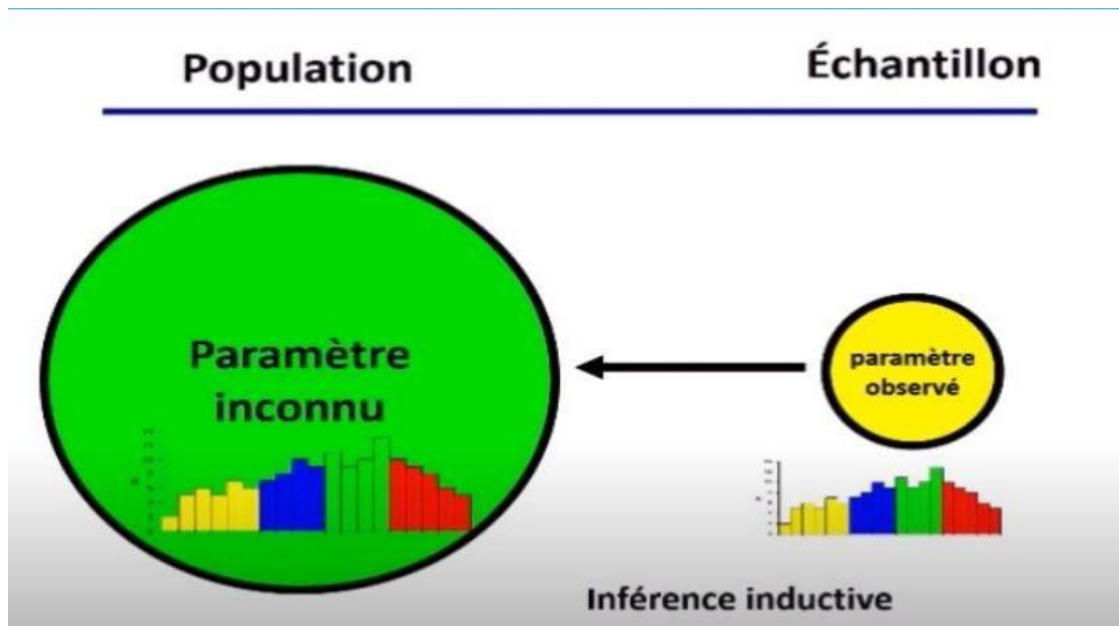
Le risque α

Le seuil α est la probabilité de rejeter H_0 lorsqu'elle est vraie, sous les hypothèses du test.

La p-value est la probabilité, sous H_0 , d'obtenir une statistique au moins aussi extrême que celle observée.

Ni α ni la p-value ne représentent la probabilité que H_0 ou une affirmation comme $A > B$ soit fausse.

Au seuil de 3 %, un test bilatéral répartit le risque en deux queues de 1,5 % ; un test unilatéral utilise une seule queue de 3 %.



Domaine du risque α

- Estimation d'un paramètre par intervalle de confiance
- Résultat d'un test statistique
- Adéquation d'un modèle avec situation observée

Estimation d'un paramètre par intervalle de confiance

Exemple : étude de l'écart-type des pluies

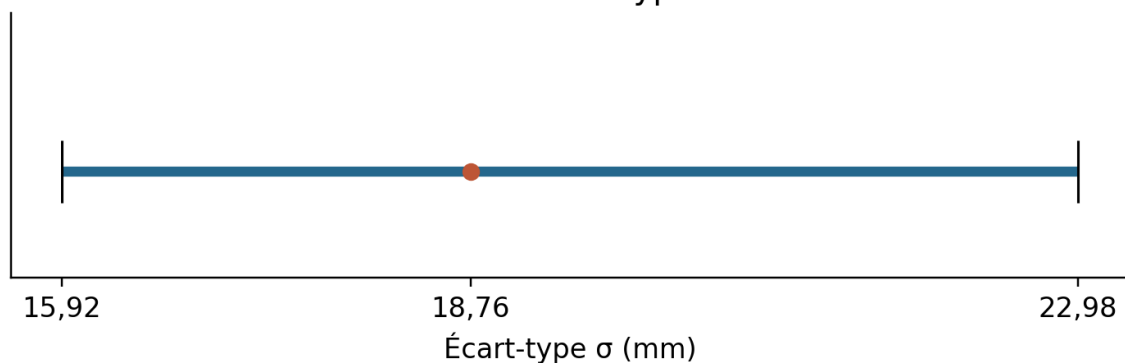
L'écart-type de la pluie :

L'écart-type de l'échantillon = 18,76 mm ; N = 42

L'écart-type de la population : σ ?

Avec N = 42 et S = 18,76 mm, l'intervalle exact à 90 % est [15,92 ; 22,98] mm ($\nu = 41$). Il suppose une population normale et des observations indépendantes.

Intervalle exact à 90 % sous hypothèse de normalité



1.1.4.2 La loi Log-Normale (loi de Galton)

Si le logarithme d'une variable quelconque a une distribution normale, alors cette variable possède une distribution log-normale.

Si, par exemple, $y = \ln x$ suit une loi normale, alors x suit une loi log-normale.

La fonction de densité de probabilité s'écrit alors :

Chapitre 1 : Analyse fréquentielle et statistique appliquée à l'hydrologie

$$f(x) = \frac{1}{x\sigma_n\sqrt{2\pi}} e^{-\frac{(\ln x - \mu_n)^2}{2\sigma_n^2}}$$

avec $x > 0$, où μ_n et σ_n sont respectivement la moyenne et l'écart-type de $\ln x$.

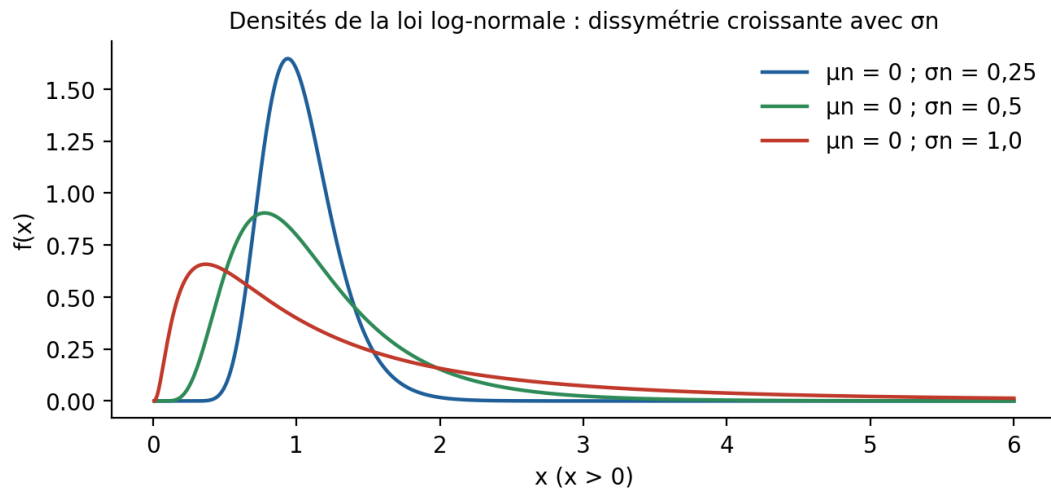


Figure 1.3 : Densités de probabilité de la loi log-normale pour $\mu_n = 0$ et différentes valeurs de σ_n .

La moyenne et la variance de la distribution log-normale peuvent aussi s'écrire :

$$\begin{cases} \mu_n = \frac{1}{2} \ln \left(\frac{\mu_x^2}{C_v^2 + 1} \right) \\ \sigma_n^2 = \ln(C_v^2 + 1) \end{cases} \rightarrow \begin{cases} \mu_x = e^{\mu_n + \frac{\sigma_n^2}{2}} \\ \sigma_x^2 = \mu_x^2 (e^{\sigma_n^2} - 1) \end{cases}$$

où μ_x est la moyenne de la variable étudiée (sans logarithme) et C_v est le coefficient de variation :

$$C_v = \frac{\sigma}{\mu}$$

La moyenne : $\overline{\ln p} = \frac{\sum \ln p_i}{N}$

L'écart-type : $S_{\ln p} = \sqrt{\frac{\sum (\ln p_i)^2 - N(\overline{\ln p})^2}{N-1}}$

La droite de Henry : $\ln p_i = \overline{\ln p} + Z_i \times S_{\ln p}$

Intervalles de confiance des paramètres logarithmiques

L'intervalle de confiance de la loi Log-Normale (GALTON)

La moyenne ci-dessous est celle de $Y = \ln(X)$, et non $E(X)$. Employer Student lorsque son écart-type est inconnu.

$$\bar{y} \pm t_{N-1; 1-\alpha/2} \frac{S_y}{\sqrt{N}}$$

où μ_n désigne la moyenne (inconnue) de $\ln x$ dans la population et σ_n son écart-type.

L'écart-type :

$$\sqrt{\frac{(N-1)S_y^2}{\chi_{N-1; 1-\alpha/2}^2}} \leq \sigma_y \leq \sqrt{\frac{(N-1)S_y^2}{\chi_{N-1; \alpha/2}^2}}$$

Un quantile x_p de probabilité p :

$$\hat{y}_p = \bar{y} + z_p S_y ; \quad q_p \in [\exp(\hat{y}_p - z_{1-\alpha/2} SE), \exp(\hat{y}_p + z_{1-\alpha/2} SE)]$$

Chapitre 1 : Analyse fréquentielle et statistique appliquée à l'hydrologie

Ici $SE \approx S_y \sqrt{[1/N + z_p^2/(2(N-1))]}$. Il s'agit d'une approximation pour le quantile ; l'exponentielle de l'intervalle de μ_y fournit un intervalle de la médiane de X, pas de sa moyenne. Pour la moyenne $E(X) = \exp(\mu_y + \sigma_y^2/2)$, utiliser une méthode tenant compte des deux paramètres, par exemple un bootstrap paramétrique.

1.1.4.3 La loi de Gumbel

C'est une loi très importante, qui sert dans l'analyse fréquentielle des valeurs extrêmes, et sera notamment l'ingrédient essentiel, en hydrologie opérationnelle, de la méthode du Gradex pour le calcul des crues de projet.

Pour l'étude des pluies extrêmes (ou n'importe quel autre événement d'une rare fréquence), on constitue un échantillon de N valeurs, chacune d'elles représentant la précipitation journalière la plus forte d'une des N années.

On parvient généralement à ajuster à cet échantillon la loi de Gumbel ou la loi Log-Normale.

La fonction de répartition de la loi de Gumbel est :

$$F(x) = e^{-e^{-\alpha(x-\beta)}}$$

où $F(x)$ = fréquence au non-dépassement = FND = F.

et α, β = coefficients d'ajustement.

Par un changement de variable $y = \alpha (x-\beta)$, la loi de Gumbel est :

$$F(x) = F(y) = e^{-e^{-y}}$$

où y est la variable réduite de Gumbel, liée à la probabilité attachée à la valeur x, et $F(y)$ = fréquence au non-dépassement de la variable réduite y.

L'équation $y = \alpha (x-\beta)$ présentée sous forme :

$$x = \frac{1}{\alpha} y + \beta$$

est l'équation d'une droite qui représente la loi de Gumbel sur du papier à l'échelle de probabilité Gumbel.

Le papier de probabilité Gumbel porte en graduation d'abscisse deux (2) échelles :

- Une échelle de fréquence au non-dépassement FND ;
- Une échelle arithmétique de la variable réduite y.

À chaque valeur de y de la seconde échelle correspond, sur le papier de probabilité, la valeur de la fréquence au non-dépassement calculée par l'expression :

$$F(x) = e^{-e^{-y}}$$

L'ordonnée, sur le papier de probabilité Gumbel, représente, sur une échelle arithmétique, la variable étudiée x.

La représentation graphique de l'échantillon des valeurs extrêmes sur papier Gumbel est obtenue en portant en ordonnées les valeurs de x et en abscisses les fréquences expérimentales au non-dépassement :

$$F_i = \frac{i - 0.5}{N} \quad (\text{Hazen, rang croissant})$$

Les valeurs de $1/\alpha$ et β sont déterminées en utilisant :

- Soit la méthode des moindres carrés qui minimise la somme des carrés des écarts entre les valeurs observées et les valeurs estimées par le modèle considéré.

Chapitre 1 : Analyse fréquentielle et statistique appliquée à l'hydrologie

- Soit la résolution d'un système d'équations formé avec les moments des deux premiers ordres.

Ces méthodes ne sont pas détaillées ici.

L'estimation des paramètres α et β par la méthode des moments donne :

$$\begin{cases} \alpha = \frac{\pi}{\sigma\sqrt{6}} \\ \text{et} \\ \beta = \mu - \frac{0.5772}{\alpha} \end{cases}$$

où μ et σ sont respectivement la moyenne et l'écart-type de la série étudiée.

Une fois les paramètres de la droite de Gumbel déterminés, on trace la droite en calculant trois valeurs de x à partir de la valeur de y en utilisant l'équation de la variable réduite de Gumbel.

L'estimation de la valeur que prendrait la variable étudiée pour une probabilité donnée peut se faire soit par la lecture directe du graphe, soit en la calculant grâce à la formule :

$$x_T = \frac{1}{\alpha} y + \beta \text{ avec } y = -\ln(-\ln(F(x)))$$

En remplaçant α et β par leurs expressions respectives on obtient :

$$\begin{aligned} x_T &= \frac{\sigma\sqrt{6}}{\pi} y + \mu - \frac{0.5772}{\alpha} = \frac{\sigma\sqrt{6}}{\pi} y + \mu - 0.5772 \times \frac{\sigma\sqrt{6}}{\pi} \\ x_T &= \mu - \frac{\sigma\sqrt{6}}{\pi} (0.5772 - y) \end{aligned}$$

Or :

$$y = -\ln(-\ln(F_{ND}(x))) \text{ et } F_{ND} = \frac{T-1}{T}$$

$$y = -\ln\left(-\ln\left(\frac{T-1}{T}\right)\right) = -\ln\left(\ln\frac{T}{T-1}\right)$$

D'où :

$$\begin{aligned} x_T &= \mu - \frac{\sqrt{6}}{\pi} \left\{ 0.5772 + \ln\left[\ln\frac{T}{T-1}\right] \right\} \times \sigma \\ x_T &= \mu + K_T \times \sigma \end{aligned}$$

$$\text{Avec } K_T = -\frac{\sqrt{6}}{\pi} \left\{ 0.5772 + \ln\left[\ln\frac{T}{T-1}\right] \right\}$$

$$K_T = -0.7797 \left[0.5772 + \ln\left\{\ln\left(\frac{T}{T-1}\right)\right\} \right]$$

K_T est appelé **facteur de fréquence**.

$$\text{La moyenne : } \bar{X} = \frac{\sum x}{n}$$

$$\text{L'écart-type : } S = \sqrt{\frac{\sum (x - \bar{x})^2}{N-1}}$$

$$\text{La droite de Henry : } x = \frac{1}{\alpha} y + X_0$$

$$\frac{1}{\alpha} = \frac{\sqrt{6}}{\pi} S \approx 0.78 \times S \quad ; \quad X_0 = \bar{X} - \frac{0.577}{\alpha}$$

Intervalle de confiance des quantiles de Gumbel

Chapitre 1 : Analyse fréquentielle et statistique appliquée à l'hydrologie

Les incertitudes dépendent de la méthode d'estimation. Procédure bootstrap paramétrique : ajuster Gumbel, simuler B échantillons de taille N, réestimer les paramètres par la même méthode, calculer chaque qT , puis prendre les quantiles $\alpha/2$ et $1 - \alpha/2$ de ces B estimations. Employer un B suffisant (par exemple 2 000) et vérifier la stabilité. Cet intervalle reste conditionnel à la loi de Gumbel choisie.

Application : pluies journalières maximales (tableau 1.1)

Pour la série des pluies journalières maximales du tableau 1.1 (§1.3.1 ; $N = 53$, $\bar{x} = 37,35$ mm, $S = 11,14$ mm), la méthode des moments donne :

$$\frac{1}{\alpha} = 0,78 \times 11,14 = 8,69 \text{ mm} \quad x_0 = 37,35 - 0,577 \times 8,69 = 32,33 \text{ mm}$$

La droite de Gumbel s'écrit donc $x = 32,33 + 8,69 y$. Pour $T = 10$ ans, $y = -\ln(-\ln 0,9) = 2,250$, d'où $x_{10} \approx 51,9$ mm ; pour $T = 100$ ans, $y = 4,600$, d'où $x_{100} \approx 72,3$ mm. La figure 1.4 montre l'ajustement graphique sur papier de Gumbel, les points expérimentaux étant placés avec la fréquence de Hazen $F_i = (i - 0,5)/N$.

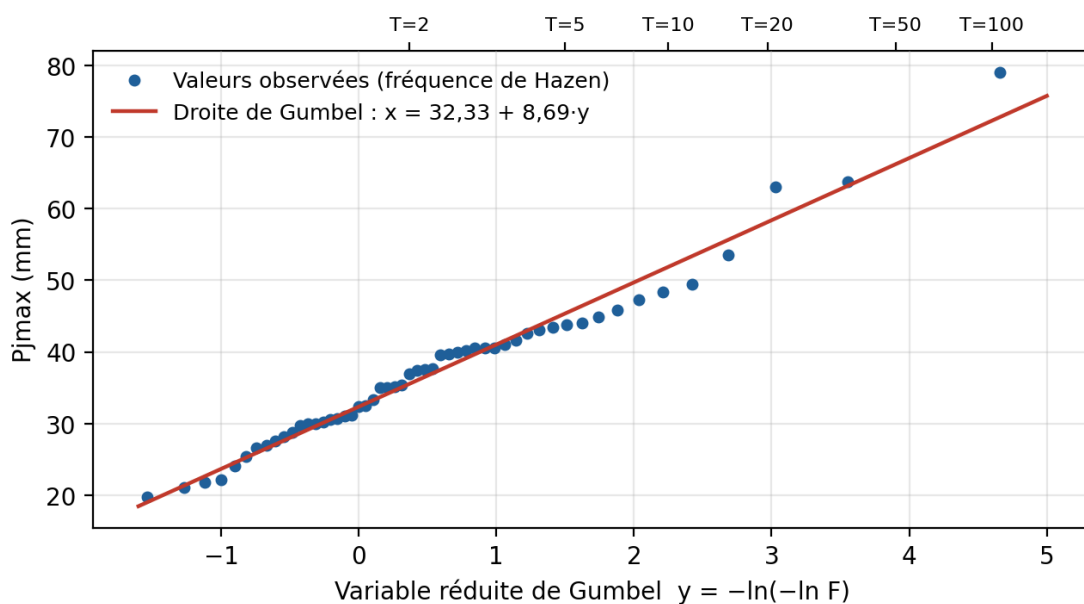


Figure 1.4 : Ajustement de la loi de Gumbel (méthode des moments) aux pluies journalières maximales du tableau 1.1.

1.1.4.4 Loi généralisée des valeurs extrêmes (GEV, ou loi de Jenkinson)

La loi de Gumbel est très largement utilisée pour caractériser la distribution de valeurs extrêmes. En fait, elle n'est qu'un des types de lois utilisables.

Jenkinson (1955) a montré que les lois de distribution des extrêmes pouvaient se mettre sous une forme unique :

$$F(x) = e^{-\left[1 - \frac{k(x-x_0)}{s}\right]^{1/k}}$$

Domaine : $s > 0$ et $1 - k(x - x_0)/s > 0$. La convention de ce cours est $k = -\xi$ par rapport à la notation internationale usuelle ξ . Aux bornes, prolonger F par 0 ou 1 suivant le support. Les moments pondérés ci-dessous exigent $k > -1$; pour $k = 0$, employer les limites de Gumbel.

Trois paramètres interviennent :

- x_0 : le paramètre de position,
- s : le paramètre d'échelle,

Chapitre 1 : Analyse fréquentielle et statistique appliquée à l'hydrologie

- k : le paramètre de forme.

Sous cette formulation unique, on retrouve les trois types de distribution selon les valeurs prises par k :

- Si $k > 0$, on retrouve une loi proposée par Weibull (GEV de type III) où x est défini sur l'intervalle $-\infty, x_0'$.
- La borne supérieure de l'intervalle de définition se déduit des paramètres s, k, x_0 par la relation :

$$x_0' = \frac{s}{k} + x_0$$

- Une borne supérieure peut être plausible ou incompatible avec la physique et les données selon le bassin ; elle ne doit pas être exclue a priori, ni acceptée sans examen des quantiles extrêmes.

Si $k = 0$, on constate que :

$$\left[1 - \frac{k(x - x_0)}{s}\right]^{1/k} \rightarrow e^{\frac{-(x-x_0)}{s}}$$

- On retrouve donc la loi de Gumbel (GEV de type I), définie sur l'intervalle $-\infty, +\infty$.

Si $k < 0$, on trouve une loi proposée par Fréchet (GEV de type II), où x est défini sur l'intervalle $x_0', +\infty$.

La borne inférieure de l'intervalle de définition se déduit des paramètres s, k, x_0 par la relation :

$$x_0' = \frac{s}{k} + x_0$$

- Ce type de loi peut se rencontrer en hydrologie, mais la difficulté réside dans l'estimation du paramètre de forme, très sensible aux aléas de l'échantillonnage.

Remarque : suivant les auteurs, le signe de k est parfois inversé et k est donc remplacé par $-k$.

- **Fonction de répartition et densité de probabilité**

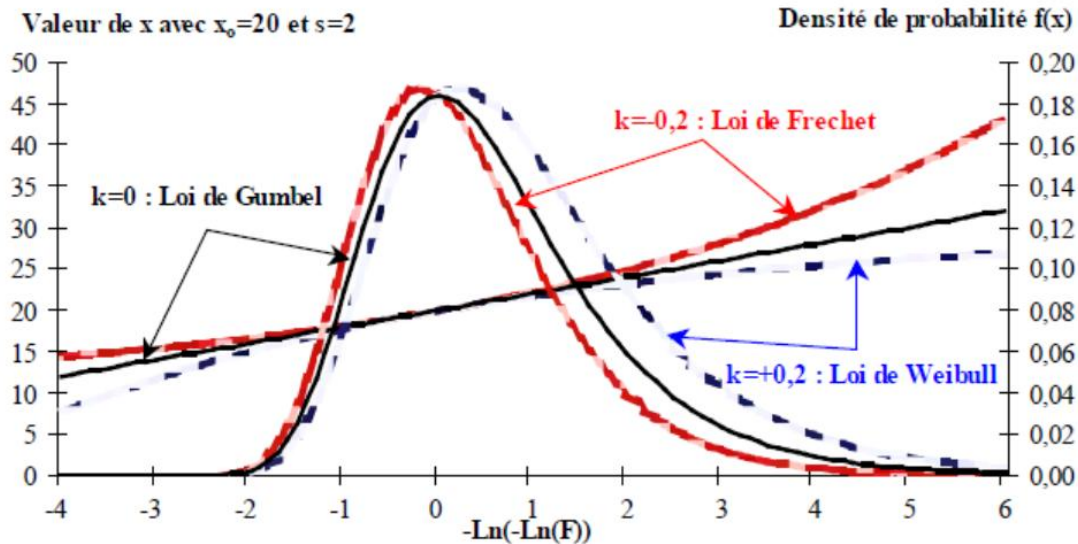
La fonction de distribution étant :

$$F(x) = e^{-\left[1 - \frac{k(x-x_0)}{s}\right]^{1/k}}$$

On en déduit la densité de probabilité :

$$f(x) = \frac{1}{s} \left[1 - \frac{k(x-x_0)}{s}\right]^{1/k-1} e^{-\left[1 - \frac{k(x-x_0)}{s}\right]^{1/k}}$$

Chapitre 1 : Analyse fréquentielle et statistique appliquée à l'hydrologie



Quantiles $x(y)$ et densité $fX(x(y))$ de la GEV en fonction de la variable réduite de Gumbel y . La courbe de densité utilise l'axe droit ; ce n'est pas la densité de y .

Comme l'illustre la figure ci-dessus, les points expérimentaux peuvent être reportés dans un graphique de **Gumbel**

$(x, -\ln(-\ln(F)))$, mais les points ne seront alignés que si $k = 0$.

La courbe rouge indique bien que si k est négatif (loi de Fréchet), les valeurs les plus fortes sont plus fréquentes que dans une loi de Gumbel.

- **Méthode d'ajustement**

Dans le cas général, il est préconisé d'utiliser la **méthode des moments pondérés**. On estime d'abord les trois premiers moments pondérés b_0, b_1, b_2 :

$$b_0 = \frac{1}{n} \sum_{i=1}^n x_i$$

$$b_1 = \frac{1}{n} \sum_{i=1}^n \frac{i-1}{n-1} x_i$$

$$b_2 = \frac{1}{n} \sum_{i=1}^n \frac{(i-1)(i-2)}{(n-1)(n-2)} x_i$$

(n étant la taille de l'échantillon et i le rang dans l'échantillon classé par ordre croissant)

Ces moments sont liés aux trois paramètres x_0, s, k par les relations :

$$\frac{3b_2 - b_0}{2b_1 - b_0} = \frac{1 - 3^{-k}}{1 - 2^{-k}}$$

$$b_0 = x_0 + \frac{s}{k} [1 - \Gamma(1 + k)]$$

$$2b_1 - b_0 = \frac{s}{k} \Gamma(1 + k) (1 - 2^{-k})$$

La fonction gamma est définie par :

$$\Gamma(1 + k) = \int_0^{\infty} u^k e^{-u} du$$

Chapitre 1 : Analyse fréquentielle et statistique appliquée à l'hydrologie

et l'on montre que :

$$\Gamma(1+k) = k\Gamma(k)$$

Pour les valeurs entières de k :

$$\Gamma(1+k) = k!$$

Une formule approchée pour $0 < k < 1$ est :

$$\Gamma(1+k) \approx 1 - 0.5748646k + 0.9512363k^2 - 0.6998588k^3 + 0.4245549k^4 - 0.1010678k^5$$

(Approximation d'Abramowitz et Stegun, erreur inférieure à 5×10^{-5} pour $0 \leq k \leq 1$. Pour $-1 < k < 0$, on utilise la relation $\Gamma(1+k) = \frac{\Gamma(2+k)}{1+k}$.)

La première équation ne contenant qu'une seule inconnue (k), on peut la résoudre par itérations successives.

Dans le cas général $-0.5 < k < 0.5$, on peut approcher par :

$$k = 7.8590c + 2.9554c^2$$

avec

$$c = \frac{2b_1 - b_0}{3b_2 - b_0} - \frac{\ln(2)}{\ln(3)}$$

On en déduit immédiatement les valeurs de s et x_0 :

$$s = \frac{(2b_1 - b_0)k}{(1 - 2^{-k})\Gamma(1+k)}$$
$$x_0 = b_0 + \frac{s}{k} [\Gamma(1+k) - 1]$$

- **Représentation graphique**

En reprenant la fonction de distribution :

$$F(x) = e^{-\left[1 - \frac{k(x-x_0)}{s}\right]^{1/k}}$$

On en déduit une représentation linéaire en fonction des points expérimentaux X_i et de leur fréquence expérimentale $F(x_i)$:

$$-\ln[F(x)] = \left[1 - \frac{k(x-x_0)}{s}\right]^{1/k}$$

En posant $y = -\ln(-\ln F)$ (variable réduite de Gumbel), il vient $(-\ln F)^k = e^{-ky} = 1 - \frac{k(x-x_0)}{s}$, d'où l'expression générale du quantile :

$$x = x_0 + \frac{s}{k} [1 - e^{-ky}]$$

On en déduit une représentation linéaire dans chacun des deux cas suivants (avec la borne $x_0' = x_0 + s/k$) :

Cas 1 : $k > 0$ (loi de Weibull)

$$x_0' - x = \frac{s}{k} e^{-ky} \quad \text{avec} \quad x_0' = x_0 + \frac{s}{k}$$

$$\ln(x_0' - x) = \ln\left(\frac{s}{k}\right) - ky$$

$\Rightarrow$ droite de pente $-k$ dans le repère $(y ; \ln(x_0' - x))$

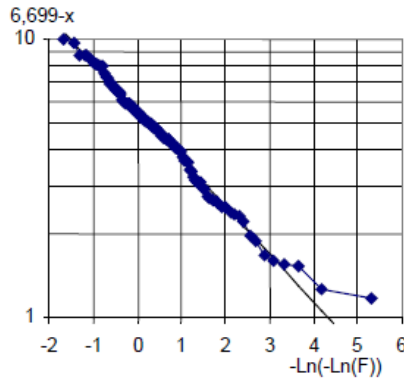
Chapitre 1 : Analyse fréquentielle et statistique appliquée à l'hydrologie

Cas 2 : $k < 0$ (loi de Fréchet)

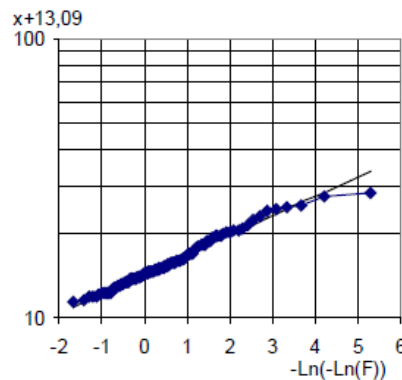
$$x - x_0' = \frac{s}{-k} e^{-ky} \quad \text{avec} \quad x_0' = x_0 + \frac{s}{k}$$

$$\ln(x - x_0') = \ln\left(-\frac{s}{k}\right) - ky$$

$\Rightarrow$ droite de pente $-k (> 0)$ dans le repère $(y ; \ln(x - x_0'))$



Exemple d'ajustement à une loi de Weibull avec :
 $k=0.39$, $s=2.165$, $x_0=1.177$ et $x_0'=6.699$



Exemple d'ajustement à une loi de Fréchet avec :
 $k=-0.162$, $s=2.319$, $x_0=1.257$ et $x_0'=-13.09$

1.2 Estimation des paramètres du modèle fréquentiel

Méthodes d'ajustement

Le problème consiste à déterminer les paramètres d'une loi (dont la formulation a déjà été choisie) en fonction de l'échantillon observé.

1.2.1 La méthode du maximum de vraisemblance

Dans le cas d'une loi discrète où on a observé un échantillon $(x_1, x_2, \dots, x_n)$, chaque valeur ayant une probabilité $(P_1, P_2, \dots, P_n)$, la probabilité P de voir apparaître cet échantillon est :

$$P = P_1 * P_2 * \dots * P_n$$

Pour une variable continue, $P(X = x_i) = 0$. La vraisemblance est un produit de densités, et non une probabilité ponctuelle de l'échantillon. Elle peut dépasser 1 et dépend des unités de mesure.

$$L(\theta) = \prod_{i=1}^N f(x_i; \theta)$$

$$L = f(x_1, a, b, \dots) f(x_2, a, b, \dots) \dots f(x_n, a, b, \dots)$$

où $a, b, \dots, k$ sont les paramètres de la loi.

Maximiser L ou sa logarithmique sous les contraintes des paramètres. Les dérivées nulles ne donnent que des candidats intérieurs ; vérifier qu'il s'agit d'un maximum et examiner les frontières du domaine.

$$\frac{\partial L}{\partial l} = 0 \text{ avec } l = a, b, \dots, k$$

On obtient ainsi un système de k équations à k inconnues.

Il est bien souvent plus simple d'effectuer les transformations suivantes :

$$\sum_{i=1}^n \frac{1}{f(x)} \frac{\partial f(x)}{\partial a} = 0 \text{ et } \sum_{i=1}^n \frac{1}{f(x)} \frac{\partial f(x)}{\partial k} = 0$$

Chapitre 1 : Analyse fréquentielle et statistique appliquée à l'hydrologie

$$\text{Log } L = \sum_{i=1}^n \log f(x_i)$$

Sous des conditions de régularité et d'identifiabilité, les estimateurs du maximum de vraisemblance sont convergents et asymptotiquement efficaces ; ils peuvent être biaisés à taille finie. Lorsque le support dépend des paramètres (GEV notamment), les conditions usuelles peuvent échouer : utiliser un algorithme contraint et vérifier la validité des résultats.

1.2.2 La méthode des moments

Égaler autant de moments empiriques que de paramètres aux moments théoriques, lorsqu'ils existent, puis résoudre le système. La convergence et l'efficacité dépendent de la loi et des conditions ; cette méthode n'est pas systématiquement inefficace.

Ajustement graphique

Cette méthode consiste à porter dans un graphique les valeurs x de l'échantillon en fonction de leur fréquence expérimentale, les axes de ce graphique étant gradués de façon à ce que les points soient alignés si le choix de la loi est correct.

Graphiquement, on fait passer une droite dans le nuage de points et on détermine ainsi deux paramètres. Des lois à trois paramètres pourront aussi être ajustées, le troisième paramètre servant à provoquer l'alignement par tâtonnements.

1.2.3 Intervalles de confiance rangs et risque de dépassement

Les intervalles de confiance sont développés au §1.1.4.1 pour la normale, au §1.1.4.2 sur l'échelle logarithmique et au §1.1.4.3 pour Gumbel. Pour la GEV, un bootstrap paramétrique avec réajustement permet d'estimer l'incertitude des quantiles ; surveiller particulièrement le paramètre de forme et la convergence.

1.2.3.1 Probabilité des valeurs d'un échantillon observé

1. La loi de distribution de la population mère est connue (par son expression mathématique et les valeurs numériques des paramètres). On démontre alors que la loi de distribution des fréquences au non-dépassement F_i , associées aux valeurs de rang i , est une loi Bêta incomplète de moyenne :

$$F_i = \frac{i}{n+1}$$

Cette formule est parfois employée pour calculer une fréquence expérimentale mais il serait alors plus judicieux d'utiliser non pas la valeur moyenne de F_i mais sa médiane. Une valeur bien approchée de la médiane est donnée par la formule :

$$F_i = \frac{i - 0,3}{n + 0,4}$$

2. Lorsque la loi est ajustée à la série, Hazen est une convention de positionnement fréquemment employée :

$$F_i = \frac{i - \frac{1}{2}}{n}$$

La formule de Hazen n'est pas universellement optimale. Préciser le classement et la convention ; les positions de Weibull ou de Gringorten sont d'autres choix possibles.

3. Cas mixte :

On a observé des crues durant 20 ans et on connaît les trois plus fortes crues enregistrées

Chapitre 1 : Analyse fréquentielle et statistique appliquée à l'hydrologie

durant les 50 ans précédents. Comment voir si ces trois valeurs sont compatibles avec ce que l'on a pu constater sur les 20 dernières années ?

On choisit le type de loi et on ajuste les paramètres à partir de l'échantillon de 20 ans. On portera donc les valeurs des crues en fonction des fréquences calculées par :

$$F_i = \frac{i - 0,5}{n}$$

Pour les trois crues historiques, utiliser leur rang dans la période historique totale et une méthode adaptée aux observations censurées. Les autres crues sont connues seulement comme inférieures au seuil historique : elles ne doivent pas être traitées comme des données manquantes ordinaires.

1.2.3.2 Danger d'apparition et période de retour

Pour des maxima annuels stationnaires, le seuil haut a $T = 1/[1 - F(x)]$ ans. Pour un seuil bas atteint avec probabilité $F(x)$ chaque année, $T_{\text{bas}} = 1/F(x)$ ans. Ce choix dépend de l'événement défini, et non d'une règle imposant $F > 0,5$.

$$T_{\text{haut}} = \frac{\Delta t}{1 - F(x)}; \quad T_{\text{bas}} = \frac{\Delta t}{F(x)}$$
$$F_1 = 1 - F$$

F et $1 - F$ sont des probabilités sans dimension. Avec un pas d'observation Δt , $T = \Delta t/p$; seul le taux d'occurrence $p/\Delta t$ a la dimension inverse d'un temps.

Exemple :

Si on a établi la statistique des crues maximales annuelles (un débit par an), la crue de fréquence au non-dépassement 0,9 a une période de retour de :

$$T = \frac{1}{1 - 0,9} = 10 \text{ ans}$$

En moyenne, cette crue est dépassée une fois tous les dix ans.

Cependant, la crue de fréquence au non-dépassement 0,1, a elle aussi une période de retour de 10 ans, mais on dira que cette crue est non dépassée en moyenne une fois tous les dix ans.

1.2.3.3 Fréquence d'apparition d'une valeur connue « a priori »

Quelle est la probabilité P de rencontrer au moins une fois, une valeur supérieure au quantile x_F sur n observations ?

Soit P_1 la probabilité de ne jamais rencontrer $x > x_F$ sur ces n observations ; on a alors :

$$P_1 + P = 1.$$

Pour ne jamais rencontrer $x > x_F$, on doit avoir $x < x_F$ à la première observation, $x < x_F$ à la seconde, etc. Or,

chaque hypothèse est vérifiée avec une probabilité F (F : fréquence au non-dépassement) ; donc :

$$P_1 = (F)^n \Rightarrow P = 1 - (F)^n$$

1.3 Tests d'adéquation

Supposons qu'une répartition empirique soit approchée par une courbe théorique $F(x)$.

Même si la courbe théorique est bien choisie, certains écarts entre celle-ci et les points expérimentaux sont inévitables.

Une question se pose alors :

Ces écarts sont-ils dus au hasard, vu le nombre limité d'observations ?

Ou bien sont-ils structurels et proviennent-ils du fait que la courbe théorique a été mal choisie ?

Chapitre 1 : Analyse fréquentielle et statistique appliquée à l'hydrologie

Pour y répondre, l'on fait appel aux tests d'adéquation ou de conformité.

Dans ce cours, nous décrirons trois tests d'adéquation :

- Le test du Khi-deux χ^2 ,
- Le test de Kolmogorov-Smirnov,
- Le test d'Anderson-Darling.

L'application de ces derniers consiste à vérifier l'hypothèse H_0 selon laquelle une certaine loi de répartition $F(x)$ s'ajuste à notre échantillon.

C'est-à-dire que H_0 est considérée comme hypothèse vraie et est appelée hypothèse nulle.

Toute autre hypothèse est appelée hypothèse alternative H_1 .

Le risque consenti, noté α , de rejeter H_0 alors qu'elle est vraie est choisi à l'avance ; on l'appelle seuil de signification. On a :

α = probabilité de rejeter H_0 alors qu'elle est vraie, ou bien,

α ne représente pas la probabilité globale d'une décision erronée : le risque de type II dépend notamment de l'alternative et de la taille de l'échantillon.

1.3.1 Le test du Khi-deux (χ^2)

H_0 affirme que les observations suivent la loi spécifiée. Ses paramètres peuvent être fixés a priori ou estimés ; dans ce dernier cas, les degrés de liberté et la calibration du test doivent en tenir compte.

Pour confirmer ou infirmer cette hypothèse, on utilise le test de Pearson, encore appelé le test du khi-deux (χ^2).

Il permet de juger de la qualité de l'ajustement d'une distribution théorique à une distribution expérimentale.

La procédure d'utilisation de ce test est la suivante :

1. Choisir des classes couvrant tout le support et ayant des effectifs théoriques suffisamment élevés (repère usuel : $E_i \geq 5$). Regrouper les classes trop peu attendues, même si leurs effectifs observés sont supérieurs à 5.
 - La classe i est bornée par les valeurs x_{i-1} et x_i choisies arbitrairement.

On détermine la fréquence absolue observée ou l'effectif de chaque classe : $f_{01}, f_{02}, \dots, f_{0k}$, avec $\sum f_{0i} = N$; N = taille de l'échantillon.

2. On peut obtenir les fréquences théoriques :

$f_{t1} = N \cdot p_1, f_{t2} = N \cdot p_2, f_{tk} = N \cdot p_k$ où p_i représente la probabilité que la variable étudiée prenne une valeur appartenant à la classe i , c'est-à-dire que : $p_i = \int_{x_{i-1}}^{x_i} f(x) dx = FND_i - FND_{i-1}$

où $f(x)$ = la fonction de densité de probabilité choisie, les limites de l'intégrale étant x_{i-1} et x_i .

On a aussi $\sum f_{ti} = N$.

Pour évaluer l'ampleur de l'écart entre les fréquences absolues observées et les fréquences théoriques f_{ti} obtenues à partir de la loi théorique que l'on suppose adéquate, on utilise la quantité :

$$\chi^2 = \frac{(f_{01} - f_{t1})^2}{f_{t1}} + \frac{(f_{02} - f_{t2})^2}{f_{t2}} + \dots + \frac{(f_{0k} - f_{tk})^2}{f_{tk}} = \sum_{i=1}^k \frac{(f_{0i} - f_{ti})^2}{f_{ti}}$$

Pearson a démontré que la distribution de cette quantité est approximativement celle du khi-deux avec v degrés de liberté, avec $v = k-1-r$,

Chapitre 1 : Analyse fréquentielle et statistique appliquée à l'hydrologie

r = nombre de paramètres qui caractérisent complètement la distribution théorique (dans le cas de la loi Normale $r = 2$).

Les conditions d'utilisation du test du χ^2 sont :

- l'échantillon est prélevé au hasard dans la population ;
- la taille de l'échantillon est suffisamment importante.

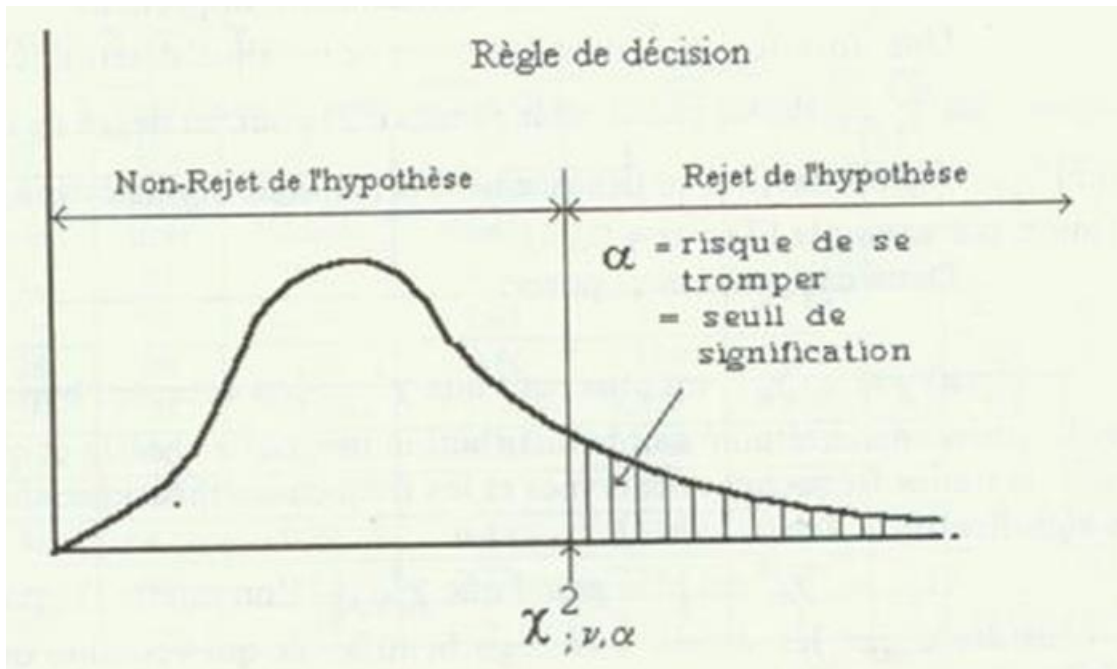


Figure 1.5 : Règle de décision du test du khi-deux

Une fois le χ^2_e relatif à notre échantillon déterminé, on le compare au $\chi^2_{v,\alpha}$ donné par la table pour un degré de liberté connu v et une probabilité au dépassement α (seuil de signification fixé à l'avance, par exemple, $FD = \alpha = 0.05$).

Deux cas peuvent se poser :

- Si $\chi^2_{\text{calculé}} \leq \chi^2_{v,1-\alpha}$, H_0 n'est pas rejetée au seuil choisi. Ce résultat ne prouve pas l'exactitude de la loi.
- Si $\chi^2_{\text{calculé}} > \chi^2_{v,1-\alpha}$, rejeter H_0 ; examiner le choix de loi, les données et les hypothèses d'indépendance, sans conclure automatiquement qu'une autre loi précise est vraie.

Exemple : soit la série des pluies journalières maximales ($P_{j\max}$)

Année	Pluie	Année	Pluie	Année	Pluie	Année	Pluie	Année	Pluie
1922	44	1933	35.4	1944	44.9	1955	27	1965	36.9
1923	29.7	1934	40.6	1945	21	1956	43.4	1966	24
1924	30.2	1935	26.6	1946	47.3	1957	42.6	1967	35
1925	40.5	1936	30	1947	39.7	1958	63.8	1968	27.5
1926	63	1937	40.5	1948	39.6	1959	32.4	1969	43.1
1927	35.1	1938	32.5	1949	29.9	1960	37.7	1970	48.4
1928	41.6	1939	31.2	1950	79.1	1961	35	1971	19.7
1929	49.5	1940	40.2	1951	28.7	1962	30.7	1972	37.5
1930	43.8	1941	45.8	1952	31	1963	28.2	1973	33.3
1931	53.5	1942	25.4	1953	30.6	1964	37.4	1974	41

Chapitre 1 : Analyse fréquentielle et statistique appliquée à l'hydrologie

Année	Pluie	Année	Pluie	Année	Pluie	Année	Pluie	Année	Pluie
1932	22.1	1943	40	1954	21.8				

Tableau 1.1 — Série des pluies journalières maximales en mm

On essaye de savoir si une loi normale, avec une moyenne $\bar{P} = 37.35mm$, et un écart-type $S = 11.14mm$, s'ajuste à notre échantillon de pluie journalière maximale (tableau 1.1) à un seuil de signification de 0.05.

Pour cela on applique le test du khi-deux à notre échantillon.

Pour rendre les effectifs attendus homogènes, choisir ici huit classes équiprobables de la normale ajustée. Chaque classe a $E_i = 53/8 = 6,625 \geq 5$; les bornes sont ses quantiles $i/8$.

$$b_i = \hat{\mu} + \hat{\sigma}\Phi^{-1}(i/8), \quad i = 1, \dots, 7$$

Calculer les bornes sans arrondi pour affecter les observations ; les valeurs affichées dans le tableau sont arrondies.

Tableau 1.2 — Test du khi-deux avec huit classes équiprobables

On cherche maintenant le $\chi^2_{v,\alpha}$ avec $\alpha = 0.05$

Et v = nombre de degrés de liberté = $K - 1 - r$, où

* K = nombre de classes = 8,

* $r = 2$ = nombre de paramètres qui déterminent la loi normale.

* $V = K - 1 - r \rightarrow V = 8 - 1 - 2 \rightarrow V = 5$

* La table du χ^2 donne $\chi^2_{5,0.05} = 11.07$

Avec huit classes et deux paramètres ajustés, $v = 5$. On obtient $\chi^2 \approx 5,717 < 11,070$: H_0 n'est pas rejetée à 5 %. Le χ^2 est un diagnostic approché ; l'approximation des degrés de liberté est plus délicate lorsque les paramètres et les bornes ont été estimés sur les mêmes données.

Le tableau ci-dessous est reproductible à partir de la série du tableau 1.1.

Classe	Borne inf. mm	Borne sup. mm	O_i	E_i	Contribution χ^2
1	$-\infty$	24,532	5	6,625	0,399
2	24,532	29,833	7	6,625	0,021
3	29,833	33,797	10	6,625	1,719
4	33,797	37,347	5	6,625	0,399
5	37,347	40,897	10	6,625	1,719
6	40,897	44,861	7	6,625	0,021
7	44,861	50,162	5	6,625	0,399
8	50,162	$+\infty$	4	6,625	1,040

La décision dépend des classes ; compléter ce diagnostic par KS avec calibration adaptée, Anderson-Darling et examen des queues. Pour des échantillons courts, un bootstrap paramétrique réestimant les paramètres et reconstruisant les classes fournit une calibration cohérente.

1.3.2 Le test de Kolmogorov-Smirnov

C'est un test d'ajustement qui permet de comparer une distribution de valeurs observées à une distribution théorique. Ce test joue le même rôle que celui du khi-deux.

Pour accepter ou rejeter l'hypothèse H_0 que la loi choisie s'ajuste bien à notre échantillon, on considère la grandeur D qui caractérise la différence entre la répartition empirique et la répartition théorique.

La grandeur D peut être choisie de plusieurs façons.

Chapitre 1 : Analyse fréquentielle et statistique appliquée à l'hydrologie

Cette grandeur D est elle-même une variable aléatoire dont la loi de répartition, dans certains cas, pour N suffisamment grand, ne dépend pratiquement pas de la fonction $F(x)$.

Avec le test de Kolmogorov-Smirnov, on cherche la valeur maximale de la valeur absolue de la différence entre la fonction de répartition empirique $F_N(x)$ d'un échantillon de N valeurs et la fonction de répartition théorique $F(x)$ correspondante soit :

$$D_N = \max_{1 \leq i \leq N} \left\{ \frac{i}{N} - F(x_{(i)}), F(x_{(i)}) - \frac{i-1}{N} \right\}$$

Si les observations sont indépendantes et la fonction de répartition continue entièrement spécifiée, la limite de la loi de $\sqrt{N} D_N$ est la loi de Kolmogorov.

$$P(\sqrt{N} D_N \leq y) \rightarrow 1 - 2 \sum_{j=1}^{\infty} (-1)^{j-1} \exp(-2j^2 y^2), \quad y > 0$$

Les valeurs de cette probabilité ont été tabulées.

On rejettera l'hypothèse H_0 que la loi choisie représente notre échantillon, au niveau de signification choisi, lorsque D_N est supérieur ou égal à d_n , la valeur critique théorique.

N = 20;
 $\alpha = 0.05$

ANNEXE 3

Table du test de Kolmogorov-Smirnov

$D_n = \sup |F_n^*(x) - f(x)|$

Valeurs de d_n telles que $P = P(D_n < d_n)$

n	P=.80	P=.90	P=.95	P=.98	P=.99
1	.90000	.95000	.97500	.99000	.99500
2	.68377	.77639	.84189	.90000	.92929
3	.56481	.63604	.70760	.78456	.82900
4	.49265	.56522	.62394	.68887	.73424
5	.44698	.50945	.56328	.62718	.66853
6	.41037	.46799	.51926	.57741	.61661
7	.38148	.43607	.48342	.53844	.57581
8	.35381	.40962	.45427	.51654	.54179
9	.33910	.38746	.43001	.47960	.51332
10	.32260	.36866	.40925	.45662	.48893
11	.30829	.35242	.39122	.43670	.46770
12	.29577	.33815	.37543	.41918	.44905
13	.28470	.32549	.36143	.40362	.43247
14	.27481	.31417	.34890	.38970	.41762
15	.26588	.30397	.33760	.37713	.40420
16	.25778	.29472	.32733	.36571	.39201
17	.25039	.28627	.31796	.35528	.38086
18	.24360	.27851	.30936	.34569	.37062
19	.23735	.27136	.30143	.33685	.36117
20	.23156	.26473	.29408	.32866	.35241
21	.22617	.25858	.28724	.32104	.34427
22	.22115	.25283	.28087	.31394	.33666
23	.21645	.24746	.27490	.30728	.32954
24	.21205	.24242	.26931	.30104	.32286
25	.20790	.23768	.26404	.29516	.31657
26	.20399	.23320	.25907	.28962	.31064
27	.20030	.22898	.25438	.28438	.30502
28	.19680	.22497	.24993	.27942	.29971
29	.19348	.22117	.24571	.27471	.29466
30	.19032	.21756	.24170	.27023	.28987
31	.18732	.21412	.23788	.26596	.28530
32	.18445	.21085	.23424	.26189	.28094
33	.18171	.20771	.23076	.25801	.27677
34	.17909	.20472	.22743	.25429	.27279
35	.17659	.20185	.22425	.25073	.26897

La table KS ci-dessus concerne une loi entièrement spécifiée. Ses annotations $N = 20$ sont un exemple de lecture, pas l'effectif de l'application ($N = 53$).

Application : pour les 53 valeurs du tableau 1.1, $\hat{\mu} = 37,347$ mm, $S = 11,140$ mm et $D = 0,10538$. Le seuil $1,36/\sqrt{53} \approx 0,187$ suppose des paramètres connus et n'est pas valide ici. Une approximation de Lilliefors à 5 % donne environ $0,886/\sqrt{53} \approx 0,122$; H_0 n'est pas rejetée. Pour une décision précise, employer la calibration de Lilliefors ou un bootstrap paramétrique réestimant μ et σ dans chaque échantillon.

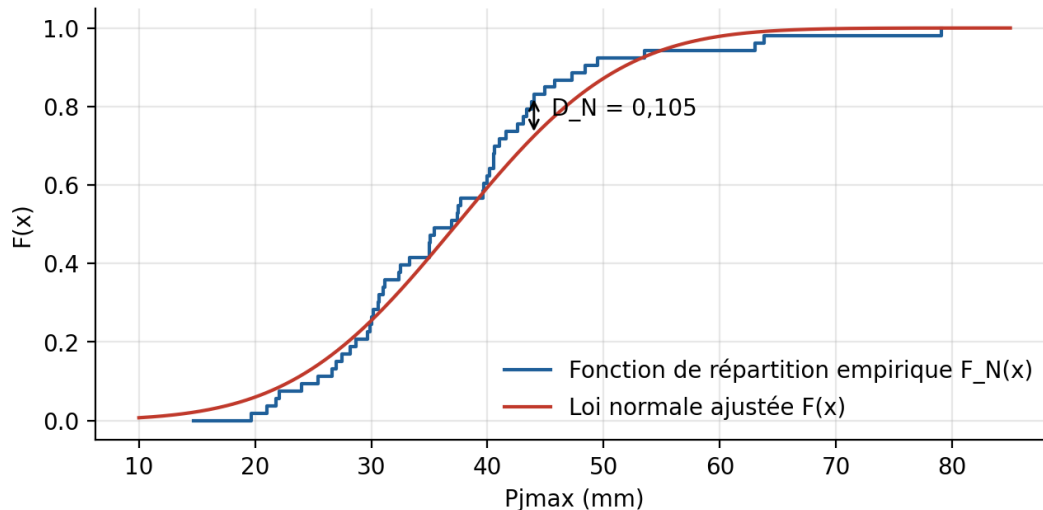


Figure 1.6 : Statistique de Kolmogorov-Smirnov : écart maximal D entre les fonctions de répartition empirique et théorique.

1.3.3 Le test d'Anderson-Darling

Le test d'Anderson-Darling est, comme celui de Kolmogorov-Smirnov, un test d'ajustement basé sur la comparaison entre la fonction de répartition empirique $F_N(x)$ et la fonction de répartition théorique $F(x)$. Il présente l'avantage, par rapport au test de Kolmogorov-Smirnov, d'accorder plus de poids aux écarts observés dans les queues de distribution — ce qui le rend particulièrement adapté à l'hydrologie statistique, où l'intérêt porte précisément sur les événements extrêmes (crues rares, étiages sévères).

Pour un échantillon de N valeurs classées par ordre croissant $x_1 \leq x_2 \leq \dots \leq x_N$, la statistique du test s'écrit :

$$A^2 = -N - \frac{1}{N} \sum_{i=1}^N (2i - 1) [\ln F(x_i) + \ln(1 - F(x_{N+1-i}))]$$

où $F(x_i)$ est la valeur de la fonction de répartition théorique ajustée, évaluée en x_i .

Procédure de décision :

1. On calcule la statistique A^2 à partir de l'échantillon et de la loi ajustée ;
2. On compare A^2 à une valeur critique A_{crit}^2 , tabulée en fonction du seuil de signification α et, selon les cas, corrigée pour tenir compte de la taille de l'échantillon et du nombre de paramètres estimés ;
3. Si $A^2 < A_{crit}^2$, l'hypothèse H_0 (la loi théorique ajuste convenablement l'échantillon) n'est pas rejetée ; dans le cas contraire, elle est rejetée.

Remarque : le test d'Anderson-Darling, le test de Kolmogorov-Smirnov et le test du Khi-deux répondent à la même question (l'adéquation d'une loi théorique à un échantillon) mais ne sont pas équivalents : il est fréquent, en pratique, qu'ils ne mènent pas exactement à la même conclusion pour un même échantillon, notamment lorsque le désaccord entre modèle et données se situe surtout dans les valeurs extrêmes. Il est donc recommandé, chaque fois que possible, de croiser plusieurs tests plutôt que de se fier à un seul.

1.3.4 Comparaison des modèles : critères d'information d'Akaike (AIC) et bayésien (BIC)

Les tests d'adéquation précédents (Khi-deux, Kolmogorov-Smirnov, Anderson-Darling) permettent de juger si *une* loi donnée s'ajuste convenablement à un échantillon, mais ne permettent pas de comparer directement l'ajustement de *plusieurs* lois candidates entre elles, notamment lorsque celles-ci n'ont pas le même nombre de paramètres. Un modèle à trois ou quatre paramètres (Pearson III, GEV) ajustera presque toujours un échantillon avec un peu moins d'écart qu'un modèle à deux paramètres (Gauss, Gumbel), sans que cela signifie qu'il soit réellement le mieux adapté au phénomène étudié : il faut pénaliser la complexité du modèle. C'est l'objet des critères d'information.

Critère d'information d'Akaike (AIC)

Proposé par Akaike (1974), il s'écrit :

$$AIC = -2\ln(L) + 2k$$

où L est la valeur maximale de la fonction de vraisemblance du modèle ajusté (obtenue par la méthode du maximum de vraisemblance, voir §1.2.1) et k le nombre de paramètres de la loi ($k = 2$ pour la loi Normale ou la loi de Gumbel, $k = 3$ pour la loi GEV ou Pearson III, etc.).

Pour des échantillons de petite taille ($n/k < 40$, situation fréquente en hydrologie), on utilise de préférence la version corrigée :

$$AIC_c = AIC + \frac{2k(k+1)}{n-k-1}$$

Critère d'information bayésien (BIC)

Proposé par Schwarz (1978), il pénalise davantage la complexité du modèle lorsque la taille de l'échantillon est grande :

$$BIC = -2\ln(L) + k\ln(n)$$

Choisir le plus petit AIC ou BIC parmi les modèles comparables ajustés par maximum de vraisemblance aux mêmes observations. Une différence d'AIC inférieure à 2 fournit peu d'éléments pour départager ; supérieure à 10, elle favorise nettement le plus petit AIC. Ces écarts ne sont pas des tests de significativité.

Les critères d'information évaluent un compromis ajustement-complexité, sans garantir l'adéquation absolue. Pour la log-normale, calculer la vraisemblance sur X , avec son facteur $1/x$: une vraisemblance calculée sur $\ln(X)$ seule ne se compare pas directement à celle des lois ajustées sur X . AICc exige $n > k + 1$.

1.4 Utilisation du modèle fréquentiel pour la construction des courbes IDF

Les courbes Intensité-Durée-Fréquence (courbes IDF) constituent l'un des principaux produits opérationnels de l'analyse fréquentielle en hydrologie urbaine et en assainissement pluvial. Elles relient, pour une station donnée, trois grandeurs :

- L'intensité moyenne de la pluie i (mm/h) ;
- La **durée** d de l'épisode pluvieux considéré (généralement comprise entre quelques minutes et 24 heures) ;
- La **fréquence**, exprimée sous forme de période de retour T (années).

Principe de construction

1. Pour chaque durée d retenue (5 min, 15 min, 30 min, 1 h, 2 h, ..., 24 h), on constitue, à partir des données pluviographiques disponibles, la série des hauteurs (ou intensités) maximales annuelles associées à cette durée ;

Chapitre 1 : Analyse fréquentielle et statistique appliquée à l'hydrologie

2. On ajuste à chacune de ces séries annuelles une loi de probabilité (le plus souvent la loi de Gumbel, mieux adaptée aux valeurs extrêmes — voir §1.1.4.3 — mais la loi GEV ou la loi log-Normale peuvent également être utilisées), selon la méthode exposée aux paragraphes 1.1 à 1.3 ;
3. Pour chaque durée d et chaque période de retour T souhaitée, on calcule le quantile correspondant à l'aide de la relation $x_T = \mu + K_T \sigma$ (facteur de fréquence, voir §1.1.4.3), ce qui fournit la hauteur de pluie $h(d, T)$, puis l'intensité moyenne $i(d, T) = \frac{h(d, T)}{d}$;
4. On ajuste enfin, pour chaque période de retour T , une relation mathématique liant i à d . La forme la plus courante, dite de **Montana**, est :

$$i(d, T) = \frac{a(T)}{d^{b(T)}}$$

où $a(T)$ et $b(T)$ sont des coefficients ajustés (par régression linéaire sur $\ln(i)$ en fonction de $\ln(d)$, voir §2.2) pour chaque période de retour, $b(T)$ étant compris typiquement entre 0,3 et 0,8.

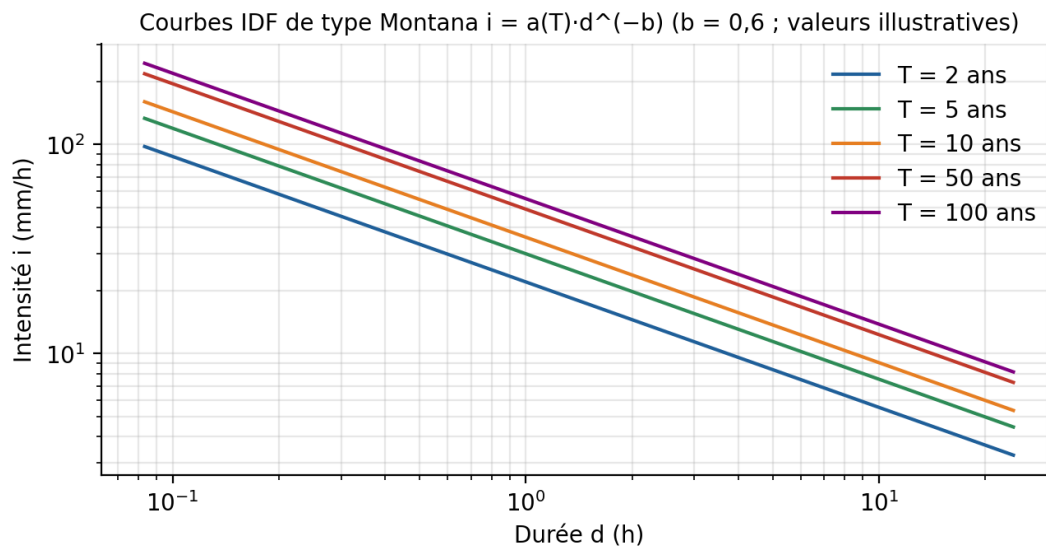


Figure 1.7 : Allure de courbes IDF de type Montana pour différentes périodes de retour (valeurs illustratives).

Utilisation

Les courbes IDF (ou les coefficients de Montana correspondants) constituent une donnée d'entrée essentielle pour le dimensionnement des ouvrages d'assainissement pluvial (réseaux, bassins de rétention, exutoires), pour la méthode rationnelle et pour la construction des hyétogrammes de projet (par exemple par la méthode des blocs alternés), en fournissant l'intensité de pluie de projet associée à la durée caractéristique du bassin versant étudié (temps de concentration) et à la période de retour retenue pour l'ouvrage.

1.4.1 Exemple de lecture et limites des courbes IDF

Exemple fictif : $i = 35 d^{-0,6}$ pour $T = 10$ ans, avec i en mm/h et d en heures. À $d = 0,5$ h, $i \approx 53,05$ mm/h et $h = i d \approx 26,53$ mm. Si d est exprimée en minutes, le coefficient devient $35 \times 60^{0,6} \approx 408,28$; son unité dépend de la convention. Ne pas appliquer les coefficients hors de l'intervalle de durées ajusté.

Construire chaque série à partir de fenêtres glissantes de durée d , vérifier les lacunes et les années incomplètes, puis contrôler la cohérence : à durée fixe, le quantile augmente avec T ; à T fixe, l'intensité diminue généralement et la hauteur cumulée ne doit pas diminuer avec la durée. Une pluie

Chapitre 1 : Analyse fréquentielle et statistique appliquée à l'hydrologie

de période de retour T ne produit pas automatiquement un débit de même période de retour : les conditions initiales et la structure temporelle interviennent.