

QUELQUES NOTIONS DE STATISTIQUES

A. INTRODUCTION

Les statistiques sont une science qui utilise des méthodes scientifiques pour collecter, organiser, synthétiser, présenter et analyser les données de tel ou tel phénomène. Elles permettent aussi de tirer des conclusions valables et de prendre des décisions raisonnables sur la base de ces analyses.

Les statistiques permettent d'exploiter les informations recueillies pour établir toute relation de causalité par l'interprétation et l'analyse.

Un phénomène aléatoire est un phénomène comportant des variables aléatoires, c'est-à-dire des variables liées au hasard et dont les valeurs ne peuvent, en conséquence, être connues à l'avance.

Les statistiques sont appliquées dans presque tous les domaines de l'activité scientifique. Lorsqu'on analyse des données relatives à un groupe d'individus ou d'objets, par exemple les tailles des étudiants, les hauteurs et/ou les diamètres des troncs d'arbres dans une forêt, les débits d'un cours d'eau, il est souvent impossible ou pas pratique d'examiner tous les éléments du groupe appelé population ; on examine alors une petite partie du groupe appelée échantillon.

Si l'échantillon est représentatif de la population, des conclusions importantes peuvent être tirées à partir de l'analyse de l'échantillon.

B. ANALYSE STATISTIQUE

Une série statistique est constituée par l'ensemble des valeurs du caractère étudié.

Par exemple, nous disposons de la série suivante de débits maximum annuels d'un oued, en m³/s :

c) Rapport de confluence R_c :

$$R_c = \frac{N_i}{N_{i+1}}$$

N_i = nombre de thalwegs d'ordre i.

d) Coefficient de torrentialité C_t :

$$C_t = D_d \times F_1.$$

$$C_t = 0,755 \times 0,220 = 0,166$$

D. BIBLIOGRAPHIE

- Roche M. (1963) : *Hydrologie de Surface*, Gauthier-Villars éd. Paris.
- Grisoni, M., Decroux, J. (1972) : *Cours d'Hydrologie Superficielle, Initiation à l'Hydrologie*, S.E.S., Secrétariat d'Etat à l'Hydraulique, Alger.
- Dubreuil, P. (1974) : *Initiation à l'Analyse Hydrologique*, Masson et Cie éd. Paris.
- Linsley, R.K., Kohler, Paulhus (1982) : *Hydrology for Engineers*, Mc Graw Hill Company, New York.
- Wilson, E. M. (1983) : *Engineering Hydrology*, MacMillan Publishers Ltd, London.
- Réménieras, G. (1986) : *L'Hydrologie de l'Ingénieur*, éd. Eyrolles, Paris.
- Chow, V.T., Maidment, D.R., Mays, L.W. (1988) : *Applied Hydrology*, Mc Graw Hill Book Company, New York.
- Linsley, R.K., Franzini, J.B., Freyberg, D.L., Tchobanoglous, G. (1992) : *Water Resources Engineering*, Mc Graw Hill Inc., New York.

1	2	1	2	1	2	1	2	1	2
An	Q	An	Q	An	Q	An	Q	An	Q
1965	28	1970	36	1975	44	1980	77	1985	58
1966	37	1971	69	1976	49	1981	77	1986	39
1967	52	1972	99	1977	53	1982	59	1987	38
1968	34	1973	77	1978	58	1983	54	1988	103
1969	44	1974	62	1979	64	1984	49	1989	47

Tableau III-1 Débits maximum d'un oued

En général, ces données brutes ne sont pas organisées. Pour pouvoir analyser une telle série et mettre en relief ses caractéristiques essentielles, l'on procède comme suit :

1 - Ordre de la série :

On peut ranger les valeurs étudiées soit dans l'ordre croissant soit dans l'ordre décroissant. La différence entre la plus grande valeur et la plus petite est appelée l'amplitude de la série.

Une valeur n'est inscrite qu'une seule fois et, en face, on indique le nombre de fois où l'on a observé cette valeur. Ce nombre est l'effectif de la valeur ou sa fréquence absolue (n_i) ; ainsi, la fréquence absolue du débit 38 m³/s est 1.

On peut également indiquer pour cette valeur la fréquence relative (f_i) qui est le rapport entre la fréquence absolue de la valeur et le total des fréquences absolues $N = \sum n_i = 25$ de la série ; ainsi, la fréquence relative du débit 38 m³/s est $1/25 = 0,04$.

Toutes ces opérations sont indiquées dans le tableau II - 2.

2 - Groupement des valeurs:

Pour mettre en relief les caractéristiques de la série étudiée, on opère des groupements en classes de valeurs.

Un groupement consiste à remplacer toutes les valeurs situées dans un intervalle donné par une valeur unique, appelée "centre de classe", à laquelle on attribue une fréquence égale à la somme des fréquences des valeurs appartenant à cet intervalle. Dans la plupart des cas, l'on recherche un découpage en intervalles égaux.

Pratiquement, pour trouver les distributions fréquentielles, l'on procède comme suit :

1.- On détermine la donnée la plus grande et la donnée la plus petite, et on calcule l'amplitude de notre échantillon, qui est égale à la différence entre ces deux valeurs.

1	2	3	1	2	3
Variable (x_i)	Effectif ou Fréquence Absolue (n_i)	Fréquence Relative (f_i)	(x_i)	(n_i)	(f_i)
28	1	0,04	53	1	0,04
34	1	0,04	54	1	0,04
36	1	0,04	58	2	0,08
37	1	0,04	59	1	0,04
38	1	0,04	62	1	0,04
39	1	0,04	64	1	0,04
44	2	0,08	69	1	0,04
47	1	0,04	77	3	0,12
49	2	0,04	99	1	0,04
52	1	0,04	103	1	0,04

Tableau III-2 Rangement des valeurs des débits

2.- On divise cette amplitude en un nombre convenable d'intervalles de même grandeur. Le nombre d'intervalles se situe entre 5 et 20 selon les cas. On utilise souvent la formule suivante:

$$k = 1 + \frac{10 \times \log N}{3}$$

où k = nombre d'intervalles et N = grandeur de l'échantillon.

Les limites des classes ne doivent pas coïncider avec les données de l'échantillon.

Numéro de classe (i)	Bornes des classes	Centre de classe (x_i)	Effectif ou fréquence absolue (n_i)	Fréquence relative (f_i)
1	$25,5 \leq x < 37,5$	31,5	4	0,16
2	$37,5 \leq x < 49,5$	43,5	7	0,28
3	$49,5 \leq x < 61,5$	55,5	6	0,24
4	$61,5 \leq x < 73,5$	67,5	3	0,12
5	$73,5 \leq x < 85,5$	79,5	3	0,12
6	$85,5 \leq x < 97,5$	91,5	0	0
7	$97,5 \leq x < 109,5$	103,5	2	0,08

Tableau III-3 Groupement des valeurs des débits

3.- On détermine le nombre d'observations (ou de données) dans chaque intervalle, c'est-à-dire la fréquence absolue n_i de chaque classe.

4.- On détermine la fréquence relative $f_i = (n_i / N)$, (N = nombre total d'observations contenues dans l'échantillon).

L'ensemble des couples (x_i, n_i) ou (x_i, f_i) définit ce qu'on appelle la fonction de distribution de la variable x .

On remarque ce qui suit :

1° - une classe contient la limite inférieure, mais pas la limite supérieure, ceci afin d'éviter qu'une valeur chevauche entre deux classes ou soit comptée deux fois ;

2° - le groupement dénature la série initiale : perte d'information et altération de son contenu ;

3° - le groupement, étant dépendant du statisticien, est arbitraire.

Le groupement des données fait perdre beaucoup d'informations. Ainsi, pour la classe $[40,5 - 55,5]$, l'on ne sait pas comment sont distribuées les données à l'intérieur de l'intervalle. Cependant le groupement présente un avantage majeur, qui est celui d'avoir une vision globale de l'échantillon et les caractéristiques principales de l'échantillon deviennent plus apparentes.

Le groupement devra donc être choisi de manière à concilier les avantages de la synthèse et les inconvénients d'une trop grande altération.

3 - Histogramme et polygone de fréquences:

C'est la représentation graphique de la fonction de distribution.

Un histogramme est une série de rectangles ayant:

a- leurs bases sur l'axe des x centrées au milieu des intervalles et dont la longueur est égale à la grandeur de l'intervalle.

b- leurs hauteurs sont égales aux fréquences. De ce fait la surface d'un rectangle est proportionnelle à la fréquence de l'intervalle qu'il représente.

Le polygone des fréquences est obtenu en joignant les milieux des sommets des rectangles de l'histogramme. On complète la ligne polygonale au moyen des segments AB et HI, de façon telle que l'aire du polygone soit égale à l'aire de l'histogramme.

La fréquence relative d'une classe est obtenue en divisant la fréquence absolue de la classe par le nombre total de données dans l'échantillon. Par exemple, la fréquence relative de la 2ème classe est :

$$8 / 25 = 0,32 \text{ ou } 32 \%$$

Si on remplace, dans le tableau, la colonne des fréquences absolues par celle des fréquences relatives, on obtient la distribution des fréquences relatives.

La représentation graphique de la distribution des fréquences relatives s'obtient à partir de l'histogramme ou du polygone des fréquences en changeant l'échelle verticale des fréquences par les fréquences relatives.

Si les intervalles des classes sont égaux entre eux, on a:

$$\text{aire de l'histogramme} = \text{aire du polygone des fréquences} = \sum f_i = 1$$

et

$$\text{aire de l'histogramme} = \text{aire du polygone des fréquences} = \sum n_i = N; \quad (N = 25 \text{ dans notre cas}).$$

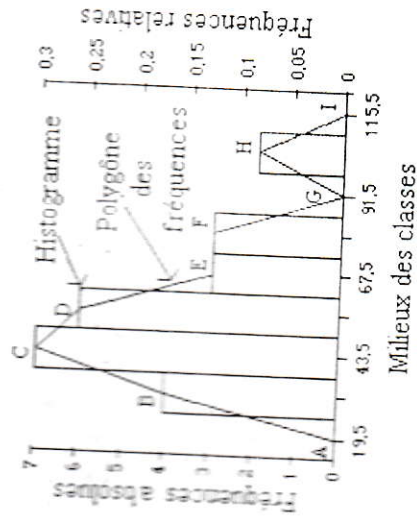


Figure III-1 Histogramme et polygone des fréquences

4 - Courbes de fréquences cumulées ou fonction de répartition:

La fonction de distribution d'une variable est constituée par l'ensemble des couples (x_i, n_i) ou (x_i, f_i) .

Débits	Effectif cumulé	Fréquence cumulée au non-dépassement (FND)	Débits	Effectif cumulé	Fréquence cumulée au dépassement (FD)
< 25.5	0	0	> 25.5	25	1.00
< 37.5	4	0.16	> 37.5	21	0.84
< 49.5	11	0.44	> 49.5	14	0.56
< 61.5	17	0.68	> 61.5	8	0.32
< 73.5	20	0.80	> 73.5	5	0.20
< 85.5	23	0.92	> 85.5	2	0.08
< 97.5	23	0.92	> 97.5	2	0.08
< 109.5	25	1.00	> 109.5	0	0.00

Tableau III-4 Calcul des fréquences cumulées

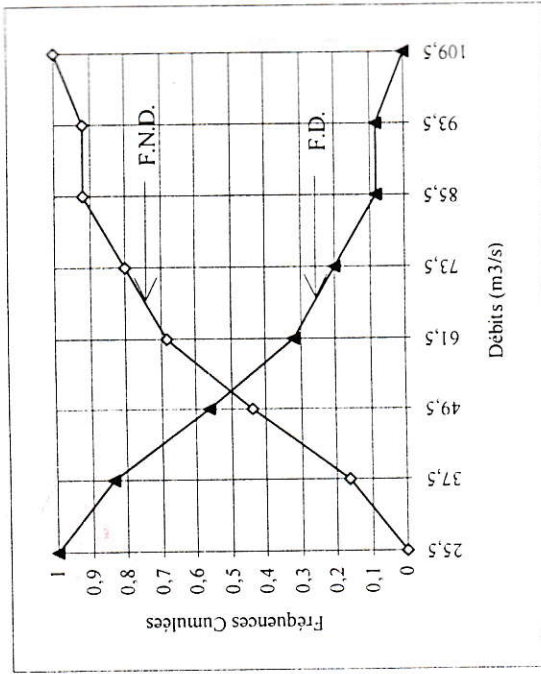


Figure III-2 Courbes des fréquences cumulées

La fonction de répartition, dite aussi fonction intégrale, est constituée par l'ensemble des couples suivants :

- $(x_i, \text{cumul des fréquences de la plus petite valeur jusqu'à celle de } x_i)$ pour les cumuls ascendants (c'est à dire de la plus petite à la plus grande valeur). C'est la fréquence cumulée au non-dépassement (FND)
- ou $(x_i, \text{cumul des fréquences depuis celle de } x_i \text{ jusqu'à celle de la plus grande valeur de l'échantillon})$ pour les cumuls descendants. C'est la fréquence cumulée au dépassement (FD).

Dans le tableau III - 4 on a calculé les fréquences cumulées jusqu'aux bornes des intervalles. La somme des fréquences de toutes les

valeurs plus petites que la limite supérieure d'un intervalle est appelée Fréquence cumulée au non - dépassement (FND) : ainsi 80 % des débits maximum annuels considérés sont inférieurs à 73,5 m³/s. En outre, la somme des fréquences de toutes les valeurs plus grandes que la limite inférieure d'un intervalle est appelée fréquence cumulée au dépassement (FD). Ainsi, 20% des débits maximums annuels de notre échantillon sont supérieurs à 73,5 m³/s.

On constate que: F.N.D. + F.D. = 80 % + 20 % = 100 %

5 - Les paramètres de position:

a) Le mode (ou dominante)

C'est la valeur dont la fréquence est la plus grande (qui se répète le plus souvent). Dans notre cas, cette valeur est 77 m³ /s.

b) La moyenne arithmétique x_a :

$$\bar{x}_a = \frac{(\sum x_i)}{N} = \frac{\text{somme des valeurs des variables}}{\text{nombre total des variables}} = 56,28 \text{ m}^3/\text{s}$$

c) La moyenne géométrique x_g :

$$\bar{x}_g = (x_1 \times x_2 \times \dots \times x_N)^{1/N} = \left(\prod x_i \right)^{1/N} ; (1/N) \sum \ln x_i = 99,45 / 25 = 3,98; \bar{x}_g = e^{3,98} \Rightarrow \bar{x}_g = 53,40 \text{ m}^3 / \text{s}$$

d) La moyenne harmonique x_h :

$$\bar{x}_h = N / \left(\sum 1 / x_i \right) = 25 / 0,4925 = 50,76 \text{ m}^3 / \text{s}$$

e) La moyenne quadratique x_q :

$$\bar{x}_q = \sqrt{\frac{\sum x_i^2}{N}} = \sqrt{87,989 / 25} = 59,33 \text{ m}^3 / \text{s}$$

f) Hiérarchie des moyennes:

On appelle moyenne d'ordre α la valeur x_α telle que

$$(\bar{x}_\alpha)^\alpha = \frac{x_1^\alpha + x_2^\alpha + \dots + x_N^\alpha}{N}$$

et l'on voit que si:

- $\alpha = 2$, on retrouve la moyenne quadratique,
- $\alpha = 1$, on retrouve la moyenne arithmétique,
- $\alpha = -1$, on retrouve la moyenne harmonique,

si α tend vers 0, on montre que la limite est la moyenne géométrique.

Nous noterons que la fonction $\bar{x}_\alpha$ est toujours croissante. La hiérarchie des moyennes est donc celle des valeurs de α c'est-à-dire:

$$\bar{x}_h < \bar{x}_g < \bar{x}_a < \bar{x}_q$$

Dans notre exemple on a bien:

$$\begin{aligned} \bar{x}_h &= 50,76 \text{ m}^3 / \text{s} < \bar{x}_g = 53,40 \text{ m}^3 / \text{s} \\ &< \bar{x}_a = 56,28 \text{ m}^3 / \text{s} < \bar{x}_q = 59,33 \text{ m}^3 / \text{s} \end{aligned}$$

g) La médiane:

La médiane d'une série de chiffres ordonnés est la valeur du milieu, c'est à dire celle qui partage l'échantillon en deux parties égales. Géométriquement, la médiane est la valeur de l'abscisse x qui correspond à la verticale qui divise l'histogramme en deux parties d'égale surface.

On calcule la médiane par la formule suivante:

$$\text{Médiane} = L_1 + \left(\frac{\frac{N}{2} - (\sum f_i)}{f_{\text{médiane}}} \right) \times c$$

où:

L_1 = limite inférieure de la classe médiane,

N = nombre de valeurs dans l'échantillon,

$\sum f_i$ = somme des fréquences absolues de toutes les classes inférieures à la classe médiane.

$f_{\text{médiane}}$ = fréquence de la classe médiane

c = grandeur de la classe médiane.

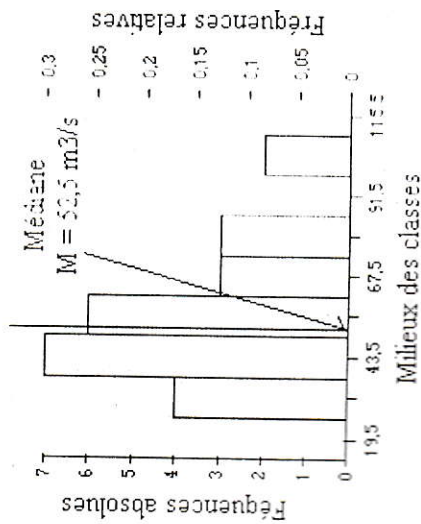


Figure III-3 Calcul de la médiane

Dans notre cas on a:

$$\text{Médiane} = 49,5 + \left(\frac{(25/2) - (4+7)}{6} \right) \times 12 = 52,5 \text{ m}^3 / \text{s}$$

On vérifie que:

$$\begin{aligned} 4 \times 12 + 7 \times 12 + 6 \times (52,5 - 49,5) &= 6 \times (61,5 - 52,5) + 6 \times 12 + 2 \times 12 \\ 48 + 84 + 18 = 54 + 72 + 24 \\ 150 &= 150 \end{aligned}$$

L'égalité est vérifiée.

6 - Les paramètres de dispersion :

a) La variance s^2

La variance d'une série de valeurs est la moyenne arithmétique des carrés des écarts entre ces valeurs et leur moyenne:

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{N - 1} = 366,79$$

b) L'écart-type s

L'écart-type s est égal à la racine carrée de la variance. Il mesure la dispersion des valeurs étudiées autour de la moyenne. Dans notre cas: $s = 19,15 \text{ m}^3 / \text{s}$.