Partie II : CORRELATIONS ET ANALYSE DES DONNEES

1. Corrélation

Introduction

L'étude des corrélations entre deux variables est un domaine qui peut parfois révéler beaucoup sur les mécanismes sous-jacents. Par exemple, chez les conducteurs automobiles, il existe une très forte corrélation entre le fait de posséder un téléphone cellulaire et le nombre d'accident automobile. Évidemment, la cause de cette corrélation est très simple: les conducteurs qui parlent dans leur cellulaire sont beaucoup moins attentifs à la route et ont donc des réactions plus lentes en cas de danger, ce qui augmente la probabilité d'accidents.

On peut presque dire que la possession d'un cellulaire cause un accroissement des accidents.

Cependant, toutes les corrélations ne sont pas aussi faciles à comprendre.

Définition

Étudier la corrélation entre deux ou plusieurs variables, c'est étudier l'intensité de la liaison qui peut exister entre ces variables. La liaison recherchée est une relation dont la représentation graphique est une droite. Une mesure de cette corrélation est obtenue par le calcul du coefficient de corrélation linéaire.

2. Les types de relations entre deux caractères quantitatifs

En amont de toute mesure de corrélation à l'aide de coefficients appropriés, il est nécessaire de définir la forme d'une éventuelle relation entre deux caractères à l'aide d'une représentation graphique appropriée. En effet, selon la forme de la relation observée, on ne fera pas les mêmes hypothèses et on n'utilisera pas les mêmes outils de mesure.

2.1 Le diagramme de corrélation

Pour savoir s'il existe une relation entre deux caractères, on établit un diagramme de corrélation, c'est à dire un diagramme croisant les modalités de X et de Y. Chaque élément i est représenté par le point de coordonnées (Xi,Yi). L'ensemble des points forme un nuage de points dont la forme permet de caractériser la relation à l'aide de trois critères :

- intensité de la relation
- forme de la relation
- sens de la relation

2.1.2L'intensité de la relation

Une relation est forte si les unités ayant des valeurs voisines sur X ont également des valeurs voisines sur Y, c'est à dire si l'on a la relation suivante

Xi proche de Xj => Yi proche de Yj

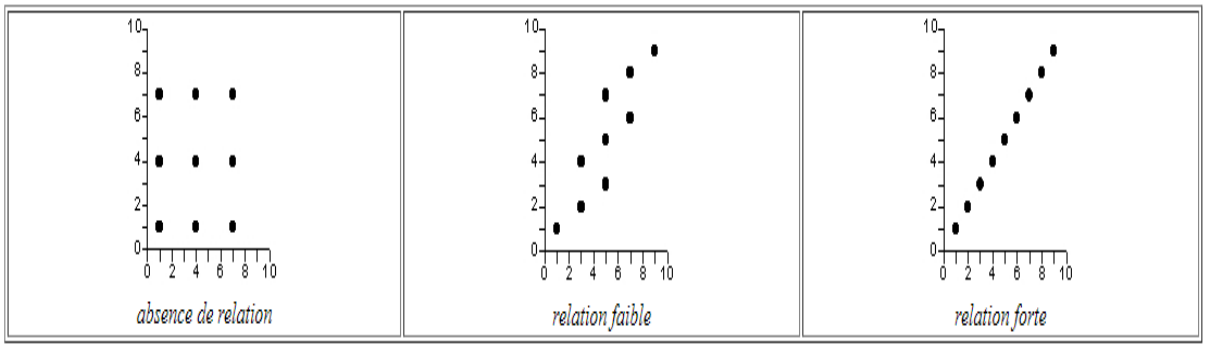
=> le nuage de point prend alors la forme d'une ligne ou d'une courbe dont les points s'écartent peu.

Une relation est faible si les unités ayant des valeurs voisines sur X peuvent avoir des valeurs éloignées sur Y, c'est à dire si deux valeurs proches de X peuvent correspondre à deux valeurs très différentes de Y

=> le nuage de point n'a pas la forme d'une ligne ou d'une courbe, ou seulement de façon très grossière.

Une relation est nulle si les valeurs de X ne permettent aucunement de prédire les valeurs de Y

=> le nuage de point a la forme d'un carré, d'un cercle, d'une "patate" sans véritables lignes directrices.

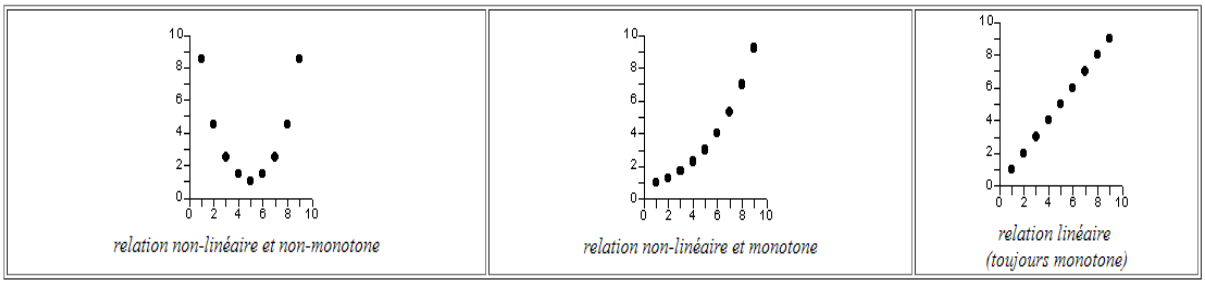


2.1.3La forme de la relation

a) Une relation est linéaire : si l'on peut trouver une relation entre X et Y de la forme Y=ax+b, c'est à dire si le nuage de point peut s'ajuster correctement à une droite.

b) Une relation est non-linéaire : si la relation entre X et Y n'est pas de la forme Y=ax+b, mais de type différent (parabole, hyperbole, sinusoïde, etc). Le nuage de point présente alors une forme complexe avec des courbures.

c) Une relation non-linéaire est monotone si elle est strictement croissante ou strictement décroissante, c'est-à-dire si elle ne comporte pas de minima ou de maxima. Toutes les relations linéaires sont monotones.



2.1.4 Le sens de la relation

Une relation monotone (linéaire ou non) est positive si les deux caractères varient dans le même sens, c'est à dire si l'on observe en général que :

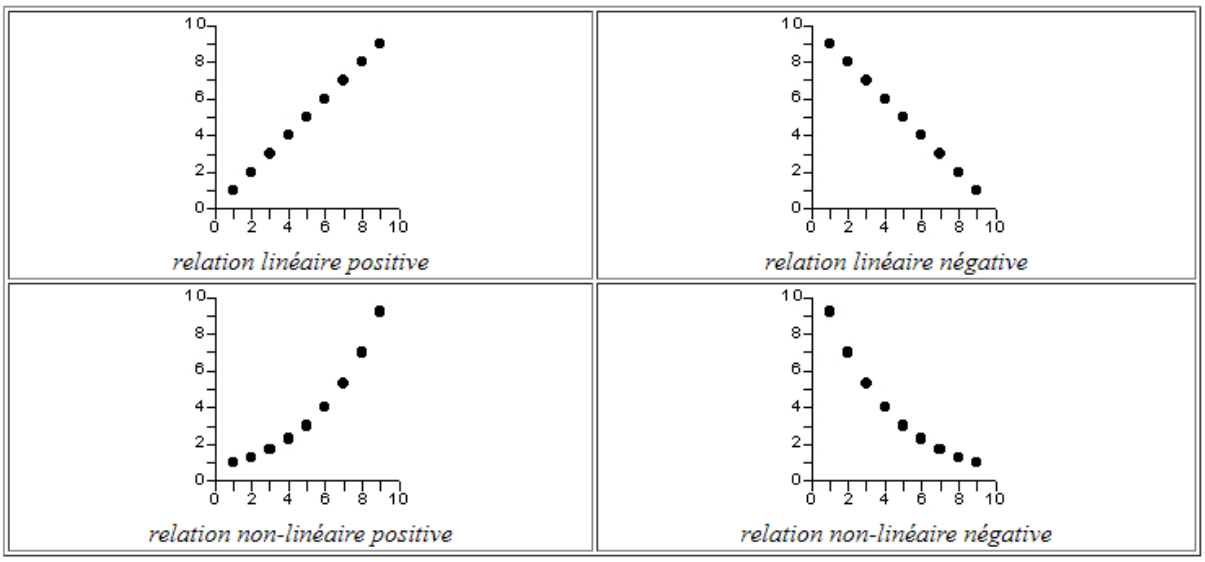
Xi > Xj => Yi > Yj

- les valeurs fortes de X correspondent généralement aux valeurs fortes de Y
- les valeurs moyennes de X correspondent généralement aux valeurs moyennes de Y
- les valeurs faibles de X correspondent généralement aux valeurs faibles de Y

Une relation monotone est négative si les deux caractères varient en sens inverse, c'est à dire si l'on observe en général que

Xi > Xj => Yi < Yj

- les valeurs fortes de X correspondent généralement aux valeurs faibles de Y
- les valeurs moyennes de X correspondent généralement aux valeurs moyennes de Y
- les valeurs faibles de X correspondent généralement aux valeurs fortes de Y



3. Calcul du coefficient de corrélation

3.1 Le coefficient de corrélation de Pearson

Par exemple, pour calculer le coefficient de corrélation entre deux séries de même longueur (cas typique : une régression), on suppose qu'on a les valeurs suivants : $X(x_1,\ldots x_n)$ et $Y(y_1,\ldots,y_n)$ pour chacune des deux séries. Alors, pour connaître le coefficient de corrélation liant ces deux séries, on applique la formule suivante :

$r_p = \frac{\sigma_{xy}}{\sigma_x * \sigma_y}$ $r_p = \frac{\sum_{i=1}^{N}(x_i - \bar{x}) * (y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N}(x_i - \bar{x})^2} * \sqrt{\sum_{i=1}^{N}(y_i - \bar{y})^2}}$

$\sigma_{xy} = \frac{1}{N} \sum_{i=1}^{N}(x_i - \bar{x}) * (y_i - \bar{y})$ Est la covariance entre X et Y et

$\sigma_x = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2}$ est l'écart-type de X.

$\sigma_{xy} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \bar{y})^2}$ est l'écart-type de Y.

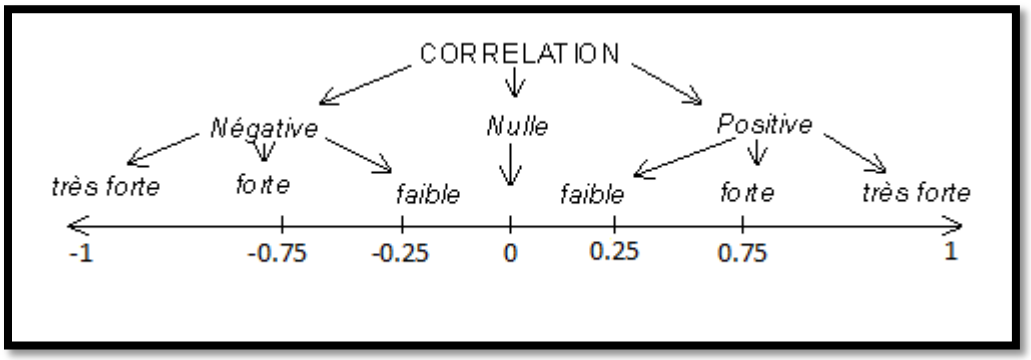
Remarque : lorsque deux caractères sont standardisés, leur coefficient de corrélation est égal à leur covariance puisque leurs écarts-types sont égaux à 1.

3.1.2 Propriétés et interprétation de r (XY)

On peut démontrer que ce coefficient varie entre -1 et +1. Son interprétation est la suivante :

- si r est proche de 0, il n'y a pas de relation linéaire entre X et Y
- si r est proche de -1, il existe une forte relation linéaire négative entre X et Y
- si r est proche de 1, il existe une forte relation linéaire positive entre X et Y

Le signe de r indique donc le sens de la relation tandis que la valeur absolue de r indique l'intensité de la relation c'est-à-dire la capacité à prédire les valeurs de Y en fonctions de celles de X.



3.2 Le coefficient de corrélation de rang de Spearman

Le coefficient de corrélation de rang (appelé coefficient de Spearman) examine s'il existe une relation entre le rang des observations pour deux caractères X et Y, ce qui permet de détecter l'existence de relations monotones (croissante ou décroissante), quelle que soit leur forme précise (linéaire, exponentiel, puissance, ...). Ce coefficient est donc très utile lorsque

l'analyse du nuage de point révèle une forme curviligne dans une relation qui semble mal s'ajuster à une droite. On notera également qu'il est préférable au coefficient de Pearson lorsque les distributions X et Y sont dissymétriques et/ou comportent des valeurs exceptionnelles.

Le coefficient de Spearman est fondé sur l'étude de la différence des rangs entre les attributs des individus pour les deux caractères X et Y :

$$r_s = 1 - \frac{6 * \sum_{i=1}^{N} [r(X_i) - r(y_i)]^2}{N^3 - N}$$

$r(X_i)$: Rang de X_i dans la distribution $X_1 \dots \dots X_N$

$r(Y_i)$: Rang de Y_i dans la distribution $X_1 \dots \dots X_N$

Ce coefficient varie entre -1 et +1. Son interprétation est la même que celui de Pearson, mais il permet de mettre en évidence des relations non-linéaires lorsqu'elles sont positives ou négatives.

4. Régression

Introduction

L´étude de certains phénomènes hydrologiques s´avère complexe, parfois du fait de la nature même du phénomène et parfois du manque de données fiables sur eux. Certains phénomènes sont liés entre eux par des relations de cause à effet. Il arrive donc qu´on connaisse relativement assez bien la cause (phénomène (s) Xi) et relativement mal le(s) phénomène(s) Yi) (Par exemple relation Pluie – débit). En hydrologie, il arrive rarement qu´on étudie les phénomènes comme des variables aléatoires isolées, c´est-à-dire sans prise en compte de leur dépendance vis à vis d´autres phénomènes ou facteurs. D´habitude, nous mesurons les grandeurs hydrologiques et autres afin de déterminer leurs rapports et leur dépendance mutuelle. L´hydrologie et surtout sa branche hydrologie statistique a développé des modèles mathématiques qui décrivent les liaisons (si elles existent !) entre ces phénomènes stochastiques (variables aléatoires). Ces modèles sont appelés modèles de régression. On les divise :

Selon le nombre de phénomènes (variables) mis en liaison en :

- Modèles de régression simple
- Modèles de régression multiple

Chacun de ces types de modèles peut être encore subdivisé en :

- Modèles de régression linéaire
- Modèles de régression non linéaire

En général, on peut rechercher une liaison mutuelle entre m variables aléatoires. On parle alors de corrélation à m dimensions. Nous considérons une de ces m variables aléatoires comme dépendante (syn. : expliquée, endogène) et les autres variables comme indépendantes (syn. : explicatives, exogènes). Dans la pratique hydrologique, il n´arrive que rarement où on

est amené à étudier la corrélation multiple (corrélation et régression sont utilisées ici comme synonymes). Très souvent, le nombre de variables aléatoires étudiées est $m \leq 3$. Le cas le plus courant est la recherche d´une corrélation entre 2 variables aléatoires X et Y.

Nous nous intéresserons ici qu´aux régressions linéaires car certaines régressions non linéaires peuvent être ramenées aux régressions linéaires par linéarisation et changement de variables (log,..etc.).

Type des relations

$Y = a(x)+b$

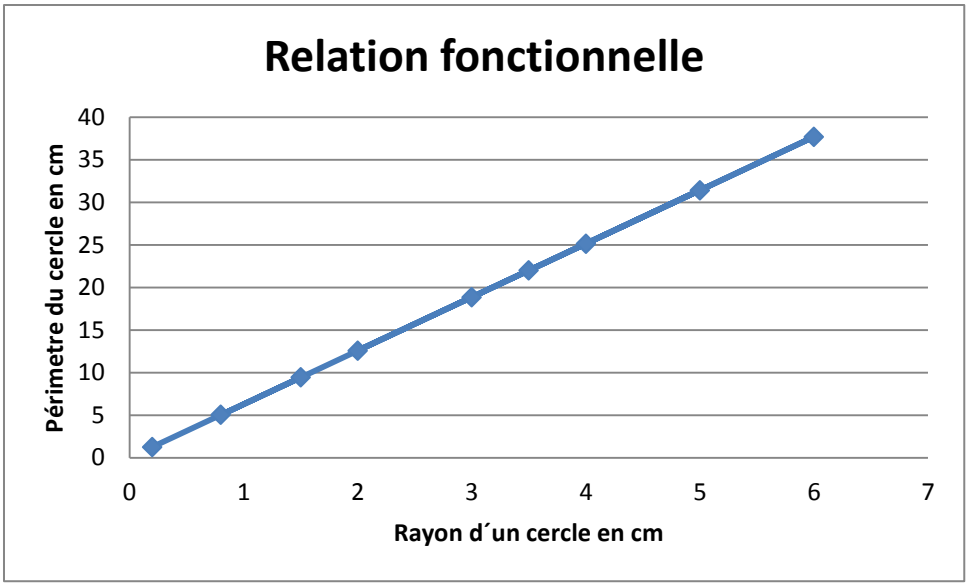
$Y = a \log(x) +b$

$\text{Log } (y) = a\ (x) +b.$

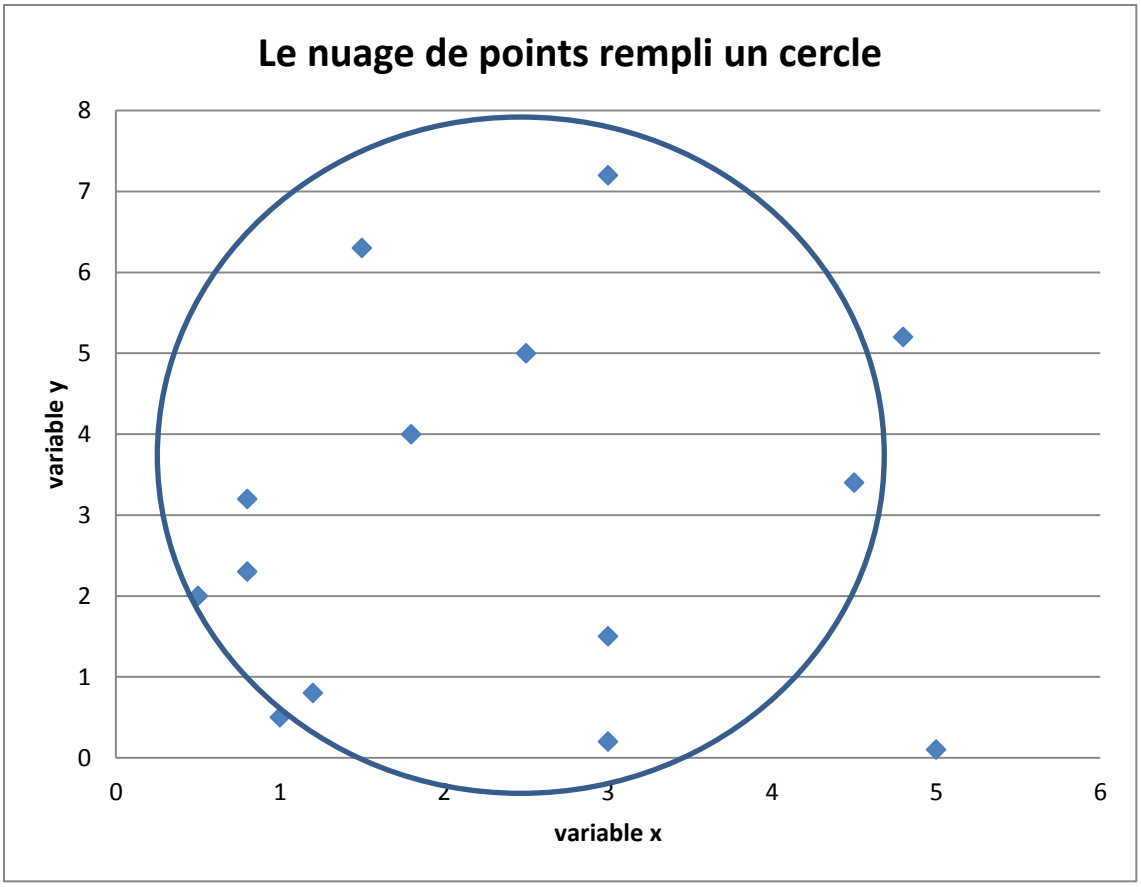
4.1 Régression linéaire simple

Considérons un échantillon de n éléments sur lesquels on mesure les valeurs de 2 variables aléatoires. Nous obtenons alors n couples de résultats de mesure (x_i , y_i). Si nous portons les points (x_i , y_i) sur un système orthogonal X, Y, nous obtenons un nuage de points qui représente une des 3 variantes suivantes :

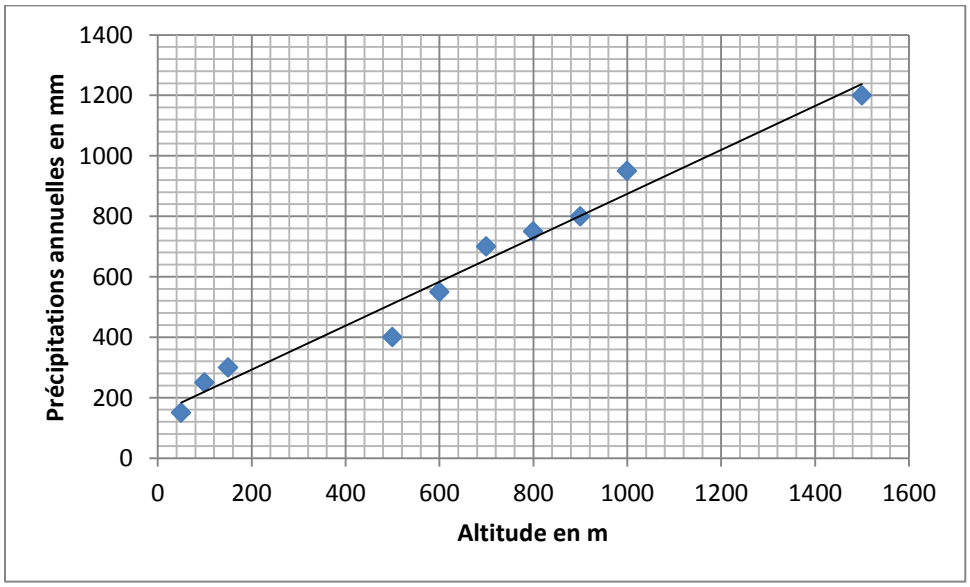
a) tous les points sont alignés sur une ligne continue (droite) qu´on peut exprimer par l´équation y= f(x). Cette relation où à une valeur fixe de la variable indépendante X correspond une et une seule valeur de la variable aléatoire dépendante Y est appelée relation fonctionnelle. En hydrologie, ce type de relation pratiquement n´apparait pas ;



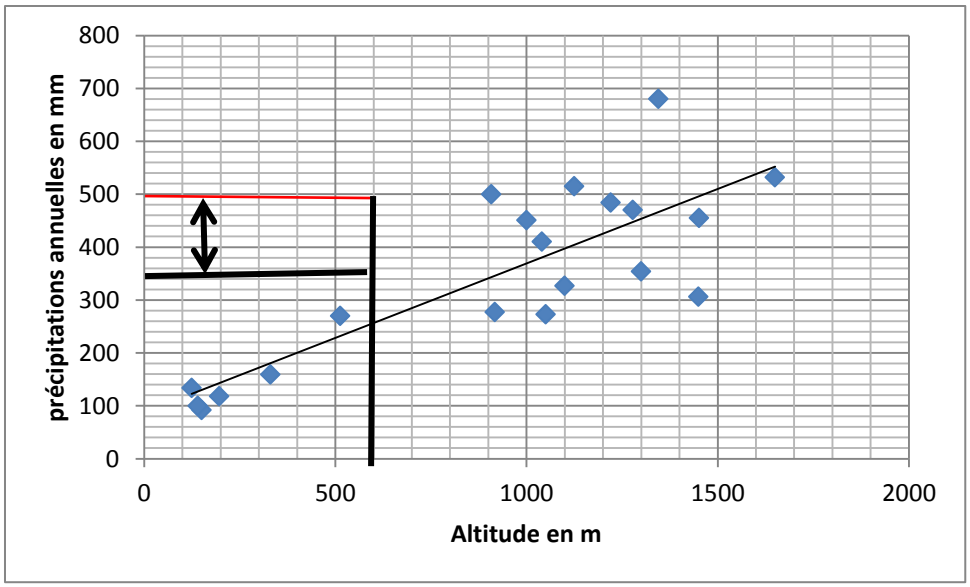
b) Les points (x_i, y_i) sont dispersés à l´intérieur d´un cercle. Aux différentes valeurs d´une variable correspond différentes valeurs d´une autre variable. Dans ce cas, nous parlons d´indépendance statistique des variables X et Y ;



c) Les points (x_i, y_i) s´alignent de part et d´autre d´une droite



- Le premier cas a lieu lorsque la liaison est parfaite ; le coefficient de corrélation que nous décrirons ci-après est égal à r=1 (resp. r= -1),
- Le deuxième cas décrit un manque de liaison (indépendance des v.a. X et Y), donc r=0 ;
- Le troisième cas concerne une liaison plus ou moins marquée, 0 < r < 1 pour une liaison directe et -1 < r < 0 pour une liaison indirecte (c´est à dire lorsque x augmente, y diminue ou le contraire). C´est ce cas qui arrive le plus couramment en hydrologie et que nous étudierons en détail dans ce polycopié.



Lorsqu´on observe dans un diagramme de dispersion de deux variables aléatoires X et Y une certaine dépendance entre les deux variables, Il est possible d´estimer au mieux la valeur prise par l´une des variables en fonction d´une valeur donnée de l´autre variable. Cette estimation s´appelle **régression.** Lorsque la dépendance entre les deux variables X et Y est exprimée par une fonction linéaire, il s´agit alors de régression linéaire simple.

Nous pouvons donc exprimer les droites de régression des deux variables par :

$y_x = a\,x + b$ la droite de y en x (f1)

4.2.1Coefficient de détermination

$$r_p = \left(\frac{\sum_{i=1}^{N}(x_i - \bar{x}) * (y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N}(x_i - \bar{x})^2} * \sqrt{\sum_{i=1}^{N}(y_i - \bar{y})^2}}\right)^2$$

4.2 Estimation des paramètres

L´estimation des paramètres a et b de l´équation (y = ax+b) se fait par la méthode des moindres carrés. Le principe consiste à minimiser la somme des carrés des résidus e_i (ou écarts) $(y_{x(i)} - y_i)$, pour i=1 à n, où n est le nombre de couples $(y_{x(i)}, y_i)$

$D_{y/x}$ => y = ax +b

a = $\frac{\sum(x_i - \overline{x})\,(y_i - \overline{y})}{\sum(x_i - \overline{x})^2}$

b = $\bar{y}$- a $\bar{x}$

5. Homogénéisation des donnés

Les erreurs dans les séries de mesures pluviométriques modifient le caractère aléatoire des phénomènes et les conditions de leur avènement.

Si ces conditions changent cela veut dire que les données mesurées ne proviennent pas de la même population et que la série de mesures n'est pas homogène.

Les données à traiter dans cette partie sont les séries des pluies annuelles des vingt-quatre stations, de la période d'observation déférent. Plusieurs tests statistiques sont utilisés pour s'assurer de l'homogénéité d'une série statistique nous étudierons l'un des tests

Suivants :

* le test de Wilcoxon.

* le test de Mann-Whitney.

* la méthode des doubles cumuls.

5.1 Le test de Wilcoxon

Principe de test

C'est un test non paramétrique qui utilise la série des rangs des observations, au lieu de la série de leurs valeurs.

Si l'échantillon (de pluie par exemple) X est issue d'une même population Y, l'échantillon X U Y (union de X et de Y) en est également issu.

On procède ainsi:

Soit une série d'observation de longueur N à partir de laquelle on tire deux échantillons X et Y : N1 et N2 sont respectivement les tailles de ces échantillons, avec N = N1 + N2 et N1 ≤ N2.

En classe ensuite les valeurs de notre série par ordre croissant. Par la suite, nous ne nous intéresserons qu'au rang de chacun des éléments des deux échantillons dans cette série. Si une valeur se répète plusieurs fois, on lui associe le rang moyen correspondant.

On calcule ensuite la somme W x des rangs des éléments du premier échantillon dans la série commune: W x = Σ Rang x.

Wilcoxon a constitué une série homogène, la quantité Wx est comprise entre deux bornes

W max et W min donnée par les formules suivantes:

$$Wmin = \frac{(N1 + N2 + 1)N1 - 1}{2} - Z_{1-\frac{\propto}{2}} \sqrt{\frac{N1\ N2(N1 + N2 + 1)}{12}}$$

Et

$$Wmax = (N1 + N2 + 1)N1 - Wmin$$

$Z_{1-\frac{\propto}{2}}$ Représente la valeur de la variable centrée réduite de la loi normale correspondant à $1 - \frac{\propto}{2}$ [au seuil de confiance de 95% nous avons]

$Z_{1-\frac{\propto}{2}}$ = 1,96

Nous allons utiliser le test de Wilcoxon pour vérifier l'homogénéité des données pluviométriques de toutes les vingt-quatre stations au niveau de signification de 5%.
Dans la suite du travail, à titre explicative, nous allons détailler les calculs pour la station d'AIN BERDA.

5.1.2 Cas des pluies annuelles de la station Ain Barda

Les données sont reportées sur le tableau (1).

Tableau (1) : série des pluies annuelles de la station d'ain berda

Année	Pluie Annuelle	Année	Pluie Annuelle	Année	Pluie Annuelle	Année	Pluie Annuelle
1970	588,9	1981	585,1	1991	623,7	2001	368,2
1971	706,3	1982	567,1	1992	680,6	2002	973
1972	791,2	1983	715,9	1993	500	2003	858,9
1973	391,7	1984	833,4	1994	585,5	2004	842,9
1974	418,1	1985	448,4	1995	734,6	2005	574,2
1975	597,6	1986	813,6	1996	391,8	2006	594
1976	705,1	1987	382,1	1997	863	2007	528,4
1977	555,5	1988	480,1	1998	735	2008	888,5
1979	464,1	1989	530,9	1999	562,1	-	-
1980	577,6	1990	638,6	2000	547,1	-	-

Nous formons ensuite le Tableau (2) pour faciliter les calculs. On commence par diviser notre série pluviométrique en deux échantillons de longueurs respectives N1= 19 valeurs et N2 =20 valeurs. Dans la première colonne, on porte le premier échantillon X; dans la deuxième colonne, on porte le deuxième échantillon Y; dans la troisième et la quatrième colonne, on porte respectivement les rangs et les valeurs classées de la séries originale et, dans la cinquième colonne, l'origine de la valeur de la série, c'est-à-dire on note si elle provient de l'échantillon X ou de Y.

Tableau (2) : application de la méthode de Wilcoxon pour vérifier l'homogénéité de la série des pluies maximales journalières de la **station d'Ain Barda.**

Données	X	Y	Rang	XUY	ORIGINE	Σ Rang x
588,9	588,9	638,6	1	368,2	Y	
706,3	706,3	623,7	2	382,1	X	2
791,2	791,2	680,6	3	391,7	X	3
391,7	391,7	500	4	391,8	Y	
418,1	418,1	585,5	5	418,1	X	5
597,6	597,6	734,6	6	448,4	X	6
705,1	705,1	391,8	7	464,1	X	7
555,5	555,5	863	8	480,1	X	8
654,5	654,5	735	9	500	Y	
464,1	464,1	562,1	10	528,4	Y	
577,6	577,6	547,1	11	530,9	Y	
585,1	585,1	368,2	12	547,1	Y	
567,1	567,1	973	13	555,5	X	13
715,9	715,9	858,9	14	562,1	Y	
833,4	833,4	842,9	15	567,1	X	15
448,4	448,4	574,2	16	574,2	Y	
813,6	813,6	594	17	577,6	X	17
382,1	382,1	528,4	18	585,1	X	18
480,1	480,1	888,5	19	585,5	Y	
530,9		530,9	20	588,9	X	20
638,6			21	594	Y	
623,7			22	597,6	X	22
680,6			23	623,7	Y	
500			24	638,6	Y	
585,5			25	654,5	X	25
734,6			26	680,6	Y	
391,8			27	705,1	X	27
863			28	706,3	X	28
735			29	715,9	X	29
562,1			30	734,6	Y	
547,1			31	735	Y	
368,2			32	791,2	X	32
973			33	813,6	X	33
858,9			34	833,4	X	34
842,9			35	842,9	Y	
574,2			36	858,9	Y	
594			37	863	Y	
528,4			38	888,5	Y	
888,5			39	973	Y	

Somme Rang x = 344

X = 19

Y = 20

Wmin = 290,34

Wmax = 469,76

Sachant que $Z_{1-\frac{\propto}{2}}$ = 1,96 pour un niveau significatif **α** =5 %.
On vérifier l'égalité Wmin < Σ Rang x < Wmax

C'est-à-dire que 290,34 < 344 < 469,76

L'inégalité est donc vérifiée, et notre série elle est homogène.

5.2. Test de Mann-Whitney

Le test non paramétrique de Mann-Whitney a pour comparer deux échantillons indépendants de petite taille. Il est valide sur des données cardinales ou ordinales, voire des variables différentes observées sur deux populations. Toutefois, en pratique, il permet surtout d'estimer si les variables de deux échantillons suivent la même loi de probabilité. Ce qui revient souvent à se demander si ces échantillons proviennent de la même population.

Pour appliquer le test de Mann-Whitney, on possède comme suit :

On divise notre échantillon original en deux sous-ensembles de tailles respectives

N1 et N2 avec N1>N2.

x1 , x2, ……xi……xN1

y1 , y2, ……yi……yN2

La taille de l'échantillon original est N = N1+N2.

On classe ensuite nos valeurs par ordre croissant de 1 à N et l'on note les rangs R(xi) des éléments du premier sous-ensemble et ceux R(yi) des éléments du second sous-ensemble dans l'échantillon original.

On définit K et S comme suit :

$K = L - \frac{N_1 (N_1 + 1)}{2}$ et S = N1N2 – k avec $L = \sum_{i=1}^{N1} R(x_i)$; c'est-à-dire la somme des rangs des élements de l'échantillon 1 dans l'échantillon original.

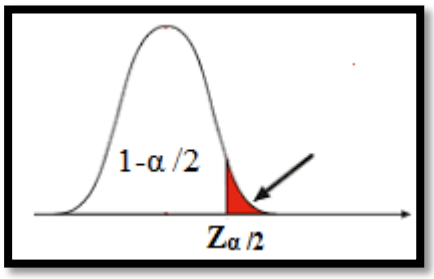
K est la somme des nombres de dépassements de chaque élément du second échantillon par ceux du premier échantillon.

S et la somme des nombres de dépassements des éléments du premier sous-ensemble (ou échantillon) par ceux du second.

On montre que lorsque N >20, N1 >3 et N2 >3 ; K et S sont distribués selon une loi normale ayant :

- Une moyenne égale à $\overline{K} = \bar{S} = \frac{N1\ N2}{2}$
- Un écart-type à$s_k = s_s = \frac{N1\ N2}{12}$ (N1+N2+1)

On peut alors tester l'hypothèse H_0 selon laquelle les deux sous-ensembles proviennent de la même population{XE " $population$"}, au niveau de signification α, en comparant la grandeur : $T = \left|\frac{K-\bar{K}}{s_k}\right|$ avec la variable normale centrée réduite ayant une probabilité de dépassement 1-α /2.



Si $T<Z_{\alpha/2}$ on accepte H_0

5.3. Analyse des doubles accumulations

La méthode des doubles accumulations 'double-mass analysis' (KohleL 1949) a été développée pour détecter et corriger un saut dans une série d'observations sur une longue période du temps, Elle est particulièrement applicable aux températures moyennes et aux précipitations totales mensuelles, saisonnières ou annuelles. Elle consiste à faire une régression linéaire sur les valeurs cumulées de la série de base (v) en fonction des valeurs accumulées de la série de référence (x). Puisque les précipitations observées à deux sites voisins sont la plupart du temps proportionnelles, il n'y a pas d'ordonnée à l'origine dans le modèle de régression. Le seul paramètre à estimer est la pente. Lorsque la pente est estimée, il faut faire un graphique des couples de points sur lesquels on superpose la droite de régression. Lorsque les séries sont homogènes, les points sont disposés aléatoirement autour de la droite de régression. Par contre, un changement à l'une ou l'autre des deux stations se remarque par une cassure de la pente. Dans ce cas, il faut ajuster deux modèles de régression avant et après cette date et la correction de la série se fait en multipliant le dernier segment par le rapport des deux pentes. Il existe une variante de cette méthode dans laquelle la régression est calculée sur la série de base en fonction de la série de référence. Par la suite, les résidus sont cumulés et l'analyse graphique se fait sur le cumul des résidus

Références bibliographiques

- Site internet