Institut National Polytechnique
de Grenoble

E.N.S. d'Hydraulique et Mécanique de Grenoble
ENSHMG

INTRODUCTION au TRAITEMENT de DONNEES en HYDROLOGIE

par **Ph. Bois**, **Ch. Obled** *et* **I. Zin**

Professeurs et Maître de Conférences à l'ENSHMG

7ème édition revue et complétée – Janvier 2007

L'Edition du Millénaire

COURS POLYCOPIE

Transmettre vos remarques à Isabella Zin, Maître de Conférences à l'ENSHMG, responsable de ce cours depuis 2004: Isabella.Zin@hmg.inpg.fr

"TRAITEMENT de DONNEES en HYDROLOGIE"

AVERTISSEMENT AU LECTEUR

Cet ensemble d'opuscules rassemblés en un document polycopié n'est pas un Traité de Statistiques!... Ce n'est qu'une introduction, destinée plus précisément aux ***Applications*** de la Statistique ***en Hydrologie.*** Ceci s'adresse principalement à des étudiants de 2ème cycle, du niveau 2ème année d'Ecole d'Ingénieurs, ainsi qu'à des formations professionnalisées du type DESS ou formation continue.

Mais d'abord: *Pourquoi utilise-t-on, (- et de manière assez intensive..! -), les statistiques en Hydrologie?*

Réponse: Parce que l'hydrologie doit apporter des éléments de décision (dimensionnement d'ouvrages par exemple) qui concernent le futur, et donc un avenir incertain. Que ce soit pour anticiper les apports qui viendront remplir un réservoir, pour choisir le débit à évacuer par un ouvrage de sécurité en cas de crue "extrême", ou pour décider de ce que peut être une sècheresse sévère et s'en prémunir, les démarches employées s'appuieront toujours sur les données observées dans le passé.., et en tireront des conclusions pour le futur...

L'objectif de ces documents est donc de présenter, parfois succinctement, les concepts élémentaires de quelques méthodes statistiques les plus couramment utilisées en Hydrologie. Ce cours est conçu pour venir après des cours d'initiation aux Probabilités et aux Statistiques, souvent placés en première année de second cycle. Mais à l'issue de ce premier contact, il apparaît que les étudiants ont encore peu de pratique ou d'expérience, (- par exemple sur ce que recouvre la notion de fluctuations d'échantillonnage...-), et même parfois un début d'allergie vis à vis de ces matières..!

Par ailleurs, un petit nombre d'entre eux aborde en fait la statistique directement par le biais de l'hydrologie. On trouvera donc aussi quelques rappels de notions théoriques, présentées parfois d'une manière "intuitive" qui doit parfois faire frémir certains de nos collègues mathématiciens...

Les méthodes décrites ici seront utilisées par les élèves sur des exemples concrets, traités essentiellement à la main, afin que l'outil informatique n'occulte pas le concept à acquérir. Cependant, on utilisera parfois aussi des ***logiciels*** adaptés, ou on signalera leur existence. Outre ceux développés en interne à l'Ecole d'Hydraulique, on citera notamment ***SAFARHY*** *(Logiciel de calculs statistiques et d'analyse fréquentielle adapté à l'évaluation du risque en Hydrologie)* distribué par les Editions de l'IRD (Institut de Recherche en Développement, ex-ORSTOM), ainsi que des logiciels commerciaux comme ***STATISTICA***® (marque protégée), avec lequel la plupart des graphiques de ce document ont été tracés. Néanmoins, leur évolution est tellement rapide qu'il faudra toujours refaire une petite étude de marché au moment d'en choisir un..

Sur un plan plus méthodologique, voire pédagogique, on fera assez souvent appel à la "simulation stochastique"; c'est à dire qu'un certain nombre d'exemples s'appuieront sur des échantillons synthétiques, générés aléatoirement, mais provenant de lois de probabilités bien définies, choisies et imposées a priori , donc connues. Ceci permettra par exemple d'initier le lecteur aux problèmes de tests (une loi donnée est-elle acceptable pour représenter cet échantillon ?) et d'échantillonnage.

Notre objectif est qu'à la fin de cette courte formation, l'élève ait acquis une autonomie suffisante pour comprendre et acquérir par lui-même d'autres méthodes ou approfondir celles qu'il aura apprises.

Ce document d'" Introduction au Traitement de Données en Hydrologie" est donc loin d'être exhaustif, et on y trouvera surtout les quelques méthodes statistiques les plus utilisées, notamment pour l'Hydrologie de Projet. Il a été écrit conçu initialement pour les élèves des filières "Ressources en Eau et Aménagements" et "Génie Hydraulique et Ouvrages" du Département GENIE de l'ENVIRONNEMENT à l'Ecole d'Hydraulique de Grenoble (INPG-ENSHMG), et pour le DESS "Eaux Souterraines" de l'Université Joseph FOURIER. Il a été utilisé aussi en Maîtrise de Géologie, de Mécanique, ainsi que pour la filière Hydraulique de l'ENTPE.

Il est en voie d'être complété par un autre fascicule, intitulé *"Hydrologie Opérationnelle"*, dans lequel ces notions élémentaires de traitement de données sont largement utilisées pour les problèmes notamment de " crues de projet ".

Cependant, les hydrologues confirmés utilisent aussi d'autres techniques d'analyse statistique, encore plus élaborées. Certaines sont présentées à l'ENSHMG au cours de la 3ème année, dans la filière "Ressources en Eau", et dans le DEA "Géophysique et Environnement". Ce sont par exemple l'analyse des séries temporelles, l'analyse de données multidimensionnelles, ou la géostatistique des processus spatiaux. On se réfèrera aux documents correspondants de MM. Duband, Bois et Obled.

Enfin, ce document est le résultat d'un travail collectif. De nombreux emprunts ont été faits, soit à des ouvrages cités en référence, soit à des documents de travail ou des rapports d'études faits par des collègues que nous tenons à remercier ici et que nous citerons au fil du texte. En dépit des efforts d'homogénéisation faits par les rédacteurs, nul doute qu'il reste quelques différences de style ou incohérences de notations, sans compter quelques erreurs sur lesquelles pourra s'exercer la sagacité du lecteur... Merci de nous les signaler.

Donc à tous, bon courage, et bonne lecture..!

Les auteurs-compositeurs

Ph. BOIS et Ch. OBLED

Note : Par rapport aux éditions antérieures, on a ajouté le chapitre sur la corrélation multiple et le chapitre sur la critique des données

PLAN GENERAL

"TRAITEMENT de DONNEES en HYDROLOGIE"

1ère Partie: ***MODELES PROBABILISTES*** 7

Chap. I: **DESCRIPTION D'UN ECHANTILLON** 7

Chap. II: **MODELES PROBABILISTES LES PLUS COURANTS** 35

Chap. III: **ESTIMATION ET TECHNIQUES D'AJUSTEMENT A UN ECHANTILLON** 85

2ème Partie: ***LIAISONS STOCHASTIQUES ENTRE VARIABLES*** 129

Chap. IV: **CORRELATION LINEAIRE SIMPLE** 131

Chap. V: **CORRELATION LINEAIRE MULTIPLE** 173

3ème Partie: ***CRITIQUE DE DONNEES*** 203

Chap. VI: **QUELQUES METHODES SIMPLES** 205

Chap. VII: **LA METHODE DU CUMUL DES RESIDUS** 237

4ème Partie: ***Annexes Tables de Student et du Chi2*** 263

Note Importante (*):

Dans les chapitres qui suivent, certains paragraphes sont marqués d'un astérisque (*****). Cela signifie qu'ils comportent des **développements ou des démonstrations qui peuvent être ignorés en première lecture.**

1ére Partie: MODELES PROBABILISTES

CHAPITRE I :

DESCRIPTION D'UN ECHANTILLON

I) Rappel sur les Variables Aléatoires: 9

I-1) Exemples et Définitions: 9
I-2) Rappels sur les **Lois de Probabilité**: 9
I-3) **Moments** d'une Loi de Probabilité: 12
I-4) Analyse d'un **échantillon**: 13

II) Description numérique d'un échantillon : 14

II-1) Paramètres de **Position:** 14
II-2) Paramètres de **Dispersion** : 15
II-3) Paramètres d' **Asymétrie** : 18
II-4) Paramètres d' **Aplatissement** : 18

III) Description graphique : 23

III-1) Histogramme des fréquences empiriques : 23
III-2) Courbe des **fréquences cumulées.** Fonction de répartition **empirique**: 26

IV) Compléments théoriques : 29

IV-1) Notion de **Période de retour** 29
IV-2) Changements de variables 32

1ère Partie - **CHAPITRE I :**

DESCRIPTION D'UN ECHANTILLON

I) RAPPEL sur les VARIABLES ALEATOIRES:

I-1) EXEMPLES et DEFINITIONS:

Les variables que l'on manipule en hydrologie (précipitations, débits, températures, mais aussi niveau de nappe phréatique, hauteur d'enneigement, durée d'insolation, etc...), vont être considérées comme des **Variables Aléatoires**.

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

X = "Précipitation annuelle à la station de Grenoble"

Cette variable prendra une valeur x_k à chaque "tirage aléatoire", à chaque réalisation k. Cela peut choquer certains de considérer comme aléatoire quelque chose que l'on peut (avec les moyens adéquats), mesurer exactement.
Par exemple, en 1988, la variable X a pris la valeur x_{88} = 734 mm.

Il n'en reste pas moins que, si l'on veut dimensionner un barrage pour compenser le manque d'eau nécessaire à certaines cultures, il faudra s'intéresser aux années futures (- par exemple de 1994, fin de la construction de l'ouvrage, à 2044, fin de la période d'amortissement -). Or on ne savait pas en 1994 ce que seraient les réalisations de la variable aléatoire X en 1995, 96 etc.., c'est à dire x_{95}, x_{96}, x_{97} etc...

On se trouve alors en *avenir incertain*: aucune approche déterministe, aucune mesure ou méthode déductive ne peut nous dire exactement, en 1994, ce que sera la réalisation x_{98} de X en 1998...

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.... (Autrement dit, on suppose que le Dieu de la Pluie tirera toujours dans la *même* urne pour décider de la pluie de l'année suivante...). Naturellement, cette hypothèse ne s'appliquera qu'à un futur relativement *" proche "* : sur une durée un peu supérieure à la durée d'amortissement de l'ouvrage, ou encore de l'ordre de grandeur de sa durée de vie utile, c'est à dire sur quelques dizaines d'années…

Un autre type de problème courant en hydrologie conduit à utiliser les mêmes outils : il ne concerne plus le futur, mais concerne l'*échantillonnage dans l'espace*. Par exemple, si on considère la conductivité hydraulique à saturation d'un sol, on conçoit qu'il s'agit d'un paramètre déterministe, que l'on peut mesurer en tout point avec un infiltromètre.

Mais on conçoit aussi que pour un bassin versant, ou une parcelle agricole de taille importante, il soit économiquement impossible de faire ces essais partout. On les réalisera donc en quelques points seulement, supposés représentatifs du domaine. On constatera que les valeurs mesurées varient, de manière difficile à prévoir, mais dans une gamme de valeurs assez stables (même si on augmente l'échantillon).

On fera alors l'hypothèse que, en un ou des points non mesurés, la variable aléatoire **X** = Conductivité hydraulique à saturation, prend des valeurs inconnues, difficiles voire impossibles à prédire exactement, mais qui auront les mêmes caractéristiques, la même **distribution statistique** que l'échantillon des valeurs effectivement mesurées en quelques points.

I-2) RAPPELS sur les **LOIS de PROBABILITE**:

On va donc chercher bientôt 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."

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

Exemple:
Probabilité que la Variable Aléatoire "Pluie Journalière à Grenoble" reste inférieure à la valeur numérique x =10 mm: $\Rightarrow$ c'est bien une fonction de x, car si au lieu de x =10 mm, on met x = 15 mm, la probabilité change. Cette probabilité est même plus grande, car on a plus de chance d'être en dessous de 15 que de 10 mm.

Evidemment, cette loi dépend aussi de la population considérée, par une forme analytique et des valeurs de coefficients particuliers, propres à cette population.
Rappelons cependant quelques propriétés générales d'une loi de probabilité:

+ la fonction de répartition de la variable aléatoire X est une **fonction monotone non décroissante** de la variable réelle x *(cf. exemple cf. figure 1)*.

En effet, si x = 15 et x + dx = 17, il est évident … (*mais il faut s'en convaincre!*) que:

$F(x) = F(15) = \Pr(X \leq 15)$ est plus petit que $F(x+dx) = F(17) = \Pr(X \leq 17)$

Mais par contre, on ne peut avoir: *(cf. contre-exemple figure 1-b ci-contre)*

$F(15) = \Pr(X \leq 15)$ plus grand que $F(17) = \Pr(X \leq 17)$

On donne ci après quelques exemples de formes possibles pour la fonction de répartition :

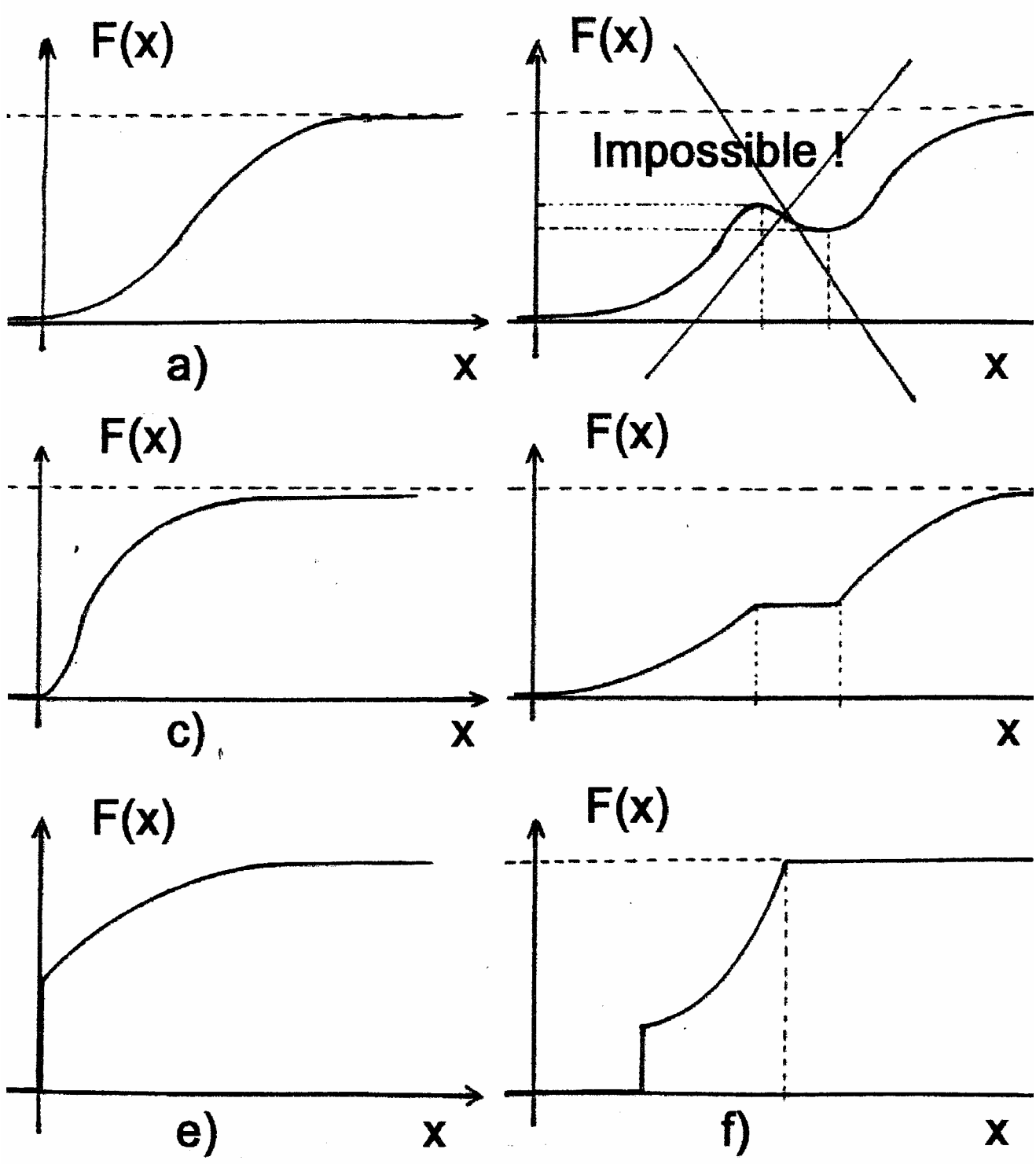


Figure 1

+ La probabilité que X tombe dans l'intervalle:

$$x < X \leq x + dx$$

est évidemment (- *... mais là aussi il faut s'en convaincre!* -):

$$\Pr(x < X \leq x + dx) = \Pr(X \leq x + dx) - \Pr(X \leq x) = F(x + dx) - F(x)$$

+ On voudrait d'ailleurs connaître aussi la probabilité que X soit strictement égal à x... Mais parmi l'infinité des valeurs possibles, cette probabilité Pr (X = x) est quasi nulle si la variable x est continue (on verra plus loin le cas des variables discrètes).

Par contre, si on se donne un peu plus de latitude, par exemple si on se donne un intervalle dx et que l'on veut :

$$Pr\ (x < X \leq x+dx)$$

alors cette probabilité dépend :

- de la longueur de ***dx*** :
  (plus dx ↗ augmente, plus on a de chance de tomber dans l'intervalle [x, x + dx])

- mais aussi de la ***position de x*** :
  Il y a des valeurs de x autour desquelles la ***densité*** d'individus, (- ou encore : de réalisations de la V.A. X) est plus grande qu'ailleurs.

⇒ On exprime cela en écrivant que:

$$Pr\ (x \langle X \leq x+dx) = \mathbf{f(\ x\ )}.dx$$

et on appelle la fonction f(x) la **densité de probabilité** de X

+ *Mais alors, qu'est-ce que f(x)?*

On a défini f(x), pour dx " petit " comme:

$$\Pr\ (x \langle X \leq x+dx) = \mathbf{f(\ x\ )}\ .dx$$

Mais on peut vérifier (- …bien réfléchir à nouveau ...-) que:

$$\Pr(\ x \langle\ X\ \leq\ x+dx\ )\ =\ \mathrm{F}(x+dx)\ -\ F(x)$$

on obtient donc : $\mathrm{f}(x).dx\ =\ \mathrm{F}(x+dx)\ -\ F(x)$ ou encore $\mathrm{f}(x)\ =\ \dfrac{\mathrm{F}(x+dx)\ -\ F(x)}{dx}$

et si on réduit l'intervalle considéré (dx → 0) alors:

$$\mathrm{f}\ (\mathrm{x}\ ) = \mathrm{F}'(\mathrm{x})$$

⇒ et la *densité de probabilité* est la ***dérivée première*** *de la fonction de répartition.*

I-3) MOMENTS d'une LOI de PROBABILITE:

On considèrera aussi que certaines caractéristiques de cette loi, et donc de cette population, sont contenues dans les **moments** de la loi F(x).

On démontre même que si tous les moments de la loi sont connus, la loi est connue complètement. (cf. par exemple VIALAR 1986).

Mais définissons d'abord les moments, par exemple la moyenne $\mu_{\mathbf{X}}$ et l'écart-type $\sigma_{\mathbf{X}}$ de la population.
On appelle **moment d'ordre 1** l'intégrale:

$$\mu_{1_x} = \int_{-\infty}^{+\infty} x.f(x).dx \quad \text{que l'on appellera encore simplement } \mu_{\mathrm{x}}$$

C'est l'**espérance mathématique** ou encore la moyenne de la population, que l'on peut voir de deux manières équivalentes comme :

- la somme de toutes les *tirages* possibles, même si certaines valeurs sortent plusieurs fois , (divisée par le nombre de *tirages* possible)
- ou la somme de toutes les *valeurs* possibles, mais chacune étant pondérée par son nombre d'apparition (divisé par le nombre de *tirages* possible), donc pondérée par… sa probabilité d'apparaître… !

Le **moment d'ordre 2** s'écrit: $\mu_{2_x} = \int_{-\infty}^{+\infty} x^2 . f(x).dx$

mais à partir de l'ordre 2, on préfère utiliser les **moments centrés**, c'est à dire :

$$\mu_{2_x} = \int_{-\infty}^{+\infty} (x - \mu_x)^2 . f(x).dx$$ encore appelé Variance et noté σ_X^2

de même on calculerait le **moment d'ordre 3** :

$$\mu_{3_x} = \int_{-\infty}^{+\infty} (x - \mu_x)^3 . f(x).dx \quad , \ \mu_{4_x}, ..., \mu_{p_x}, etc...$$

Et on verra plus loin que l'on dépasse rarement l'ordre 4..!

*** **Notations** Dans tout ce document, nous noterons :

*en **lettres grecques*** les caractéristiques de la **population**, par exemple μ_X et σ_X

et

*en **lettres latines*** les caractéristiques d'un **échantillon,** par exemple m_X et s_X

I-4) ANALYSE d'un **ECHANTILLON**:

A défaut de pouvoir appréhender toute la population qui nous intéresse, on dispose souvent d'un échantillon de n valeurs d'une variable X.

Exemple:

les températures moyennes mensuelles de Février à Grenoble de 1900 à 1990,
soit 91 valeurs…

Mais dés que n est grand ($\geq$ quelques dizaines), la lecture du tableau n'est pas aisée, et il n'est pas utile de le transmettre entièrement pour permettre à un interlocuteur de s'en faire une idée.

C'est pourquoi il est intéressant d'effectuer une ***synthèse*** de ce tableau:

+ synthèse numérique (on le résume en quelques chiffres)

+ synthèse graphique (on le résume en une courbe)

+ synthèse analytique (on le résume par une fonction analytique, un modèle… cf. chapitre II)

Certes on perdra de l'information mais on y gagnera en clarté. C'est ce que nous allons voir dans le paragraphe et les chapitres suivants.

II- DESCRIPTION NUMERIQUE D'UN ECHANTILLON:

Soit **x_i, (i de 1 à n),** les n valeurs de l'échantillon.

On va chercher à tirer de ce tableau quelques repères numériques, ***représentatifs*** non seulement de l'échantillon, mais si possible aussi de la population dont il est extrait.

Pour éclairer ces notions simples, on utilisera la *"simulation stochastique*", c'est à dire un moyen "simple" pour "fabriquer" des échantillons issus d'une même population (i.e. tirés de la même urne de caractéristiques imposées). Ainsi l'on pourra travailler sur un grand nombre d'échantillons tous différents mais dont on sait, pour les avoir fabriqués, qu'ils proviennent de la même population de caractéristiques connues.

II-1) Paramètres de **POSITION:**

Ce sont des paramètres qui précisent à peu près l'ordre de grandeur le plus courant de X. On utilise couramment:

a) **Moyenne arithmétique** :

On la définit (- en *lettres latines* car elle est estimée sur un échantillon -) par:

$$\bar{x} \text{ ou } m_x = \frac{1}{n}\sum_{i=1}^{n} x_i$$

C'est un descripteur simple, qui a les **avantages** d'être :

+ ***Robuste*** : ne varie pas trop d'un échantillon à l'autre (on aura des précisions dans la suite du cours pour certaines populations).

+ ***Convergent*** : si n tend vers l'infini, la moyenne ainsi définie tend vers la moyenne de la population (ce qui aurait été également le cas si on avait divisé par n-1 au lieu de n).

+ ***Non biaisé*** : si on fait le calcul pour un grand nombre d'échantillons différents de taille n, la moyenne de ces moyennes est une bonne estimation, ni plutôt par excès ni plutôt par défaut de la moyenne de la population (ce qui n'aurait pas été le cas si on avait divisé par n-1 au lieu de n).

mais qui présente des **défauts**:

- Ne donne aucune idée des variations de x_i autour de cette valeur.

- Pour certaines distributions (notamment asymétriques ou multimodales), la moyenne n'est pas toujours une valeur très probable.

Exemple:

A Grenoble, la moyenne de l'insolation journalière en Février est de 4 heures; mais en fait, peu de journées ont autour de 4 heures d'insolation: schématiquement, ou bien il fait beau, et il y a 8 heures d'insolation, ou bien il fait mauvais, et il n'y pas d'insolation du tout. (Pour l'anecdote, la méconnaissance de cette observation élémentaire a amené certains constructeurs d'installations solaires à mal dimensionner ces installations).

Mais on peut penser à d'autres paramètres de position:

b) **la Médiane** :

C'est la valeur x_{Med} ou $x_{50\%}$ telle que :

X a 50% de chance d'être supérieure à x_{Med}
mais aussi 50% de chance de lui être inférieure.

c) **le Mode** :

C'est la valeur $\mathbf{x_{Mod}}$ autour de laquelle on trouve le plus de valeurs , celle qui est la plus fréquente, ou la plus probable.

Exemple:
Si on considère une variable aléatoire comme " le salaire des salariés déclaré lors du recensement de la population de 1992 ", on constate que le salaire moyen est voisin de 9000 FF (car il inclue notamment quelques "gros salaires", qui apparaissent épisodiquement dans un journal satirique paraissant le Mercredi....).
Par contre le salaire Médian est plutôt voisin de 8500 FF (la moitié des français gagnent moins et l'autre gagne plus).
Enfin le salaire le plus fréquent est encore le SMIC, voisin de 5000 FF...
(Pour le lecteur soucieux d'être à jour, on rappelle qu'un Euro = 6,55957 FF …)

Complément:
On notera aussi que d'un point de vue analytique, le mode correspond au maximum de la densité de probabilité f(x) et donc vérifie que sa dérivée $f'(\mathbf{x_{Mod}}) = 0$

II-2) Paramètres de DISPERSION:

Après avoir "positionné" la gamme de valeurs de X, on cherche à donner une idée de la fluctuation des x_i dans l'échantillon.

a) ***Extrêmes*** (étendue)

Une façon simple consiste à préciser minimum et maximum de l'échantillon.

Simples à déterminer sur un échantillon, ils ont le défaut d'être peu robustes, c'est à dire de varier considérablement d'un échantillon à l'autre d'une même population (sauf évidemment pour des populations bornées comme l'insolation).

Il en est de même de l'étendue = Max – Min

b) ***Variance et écart type*** :

On le définit sur l'échantillon par : $\frac{1}{n}\sum_{i=1}^{n}(x_i - m_x)^2 = V = s_x^2$

Soit $\mu_2 = \sigma^2$ la valeur de ce terme dans la population infinie.

On conçoit que si n tend vers l'infini, V tend vers σ^2, c'est à dire que s_X est un estimateur **convergent** de σ^2.

Mais si pour n donné, on effectue ce calcul pour un grand nombre d'échantillons (en utilisant pour *centrer* chaque échantillon la moyenne *empirique* m_X de cet échantillon), on va trouver que la moyenne des V est en général inférieure à $\sigma 2 \Rightarrow$ V est donc un estimateur convergent mais **biaisé** de σ.

Il est alors intéressant de le débiaiser, d'où les définitions:

Variance = Carré de l'écart type = σ_X^2, sera *estimée* par:

$$s_x^2 = \frac{1}{\underbrace{n-1}_{\Uparrow}} \sum_{i=1}^{n} (x_i - m_x)^2$$

si $\overline{x}$ ou m_X est calculé sur l'échantillon.
Par contre:

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

si on connaît μ_X la vraie moyenne de la population. (Ce deuxième cas est pratiquement inconnu en Hydrologie..!).

c) ***Coefficient de variation CV*** *:*

On définit aussi:

$$CV = \frac{\sigma_x}{\mu_x} \text{ estimé par } \frac{s_x}{m_x} \text{ ou } \frac{s_x}{\overline{x}}$$

qui compare donc la fluctuation à la valeur moyenne.

C'est une grandeur **adimensionnelle,** qui ne dépend pas des unités, si x est une mesure, mais qui dépend de l'origine choisie pour la variable X

(-*Attention* par exemple aux températures exprimées en unités ordinaires Celsius ou Fahrenheit!. $\Rightarrow$ le coefficient de variation de la température exprimée en degré Kelvin est beaucoup plus faible que celui de la température en Celsius...!
Et ce coefficient de variation est même absurde, en ° Celsius, pour une station de montagne comme le grand Saint Bernard où la moyenne est proche de 0°C, car il devient quasi infini.. !)

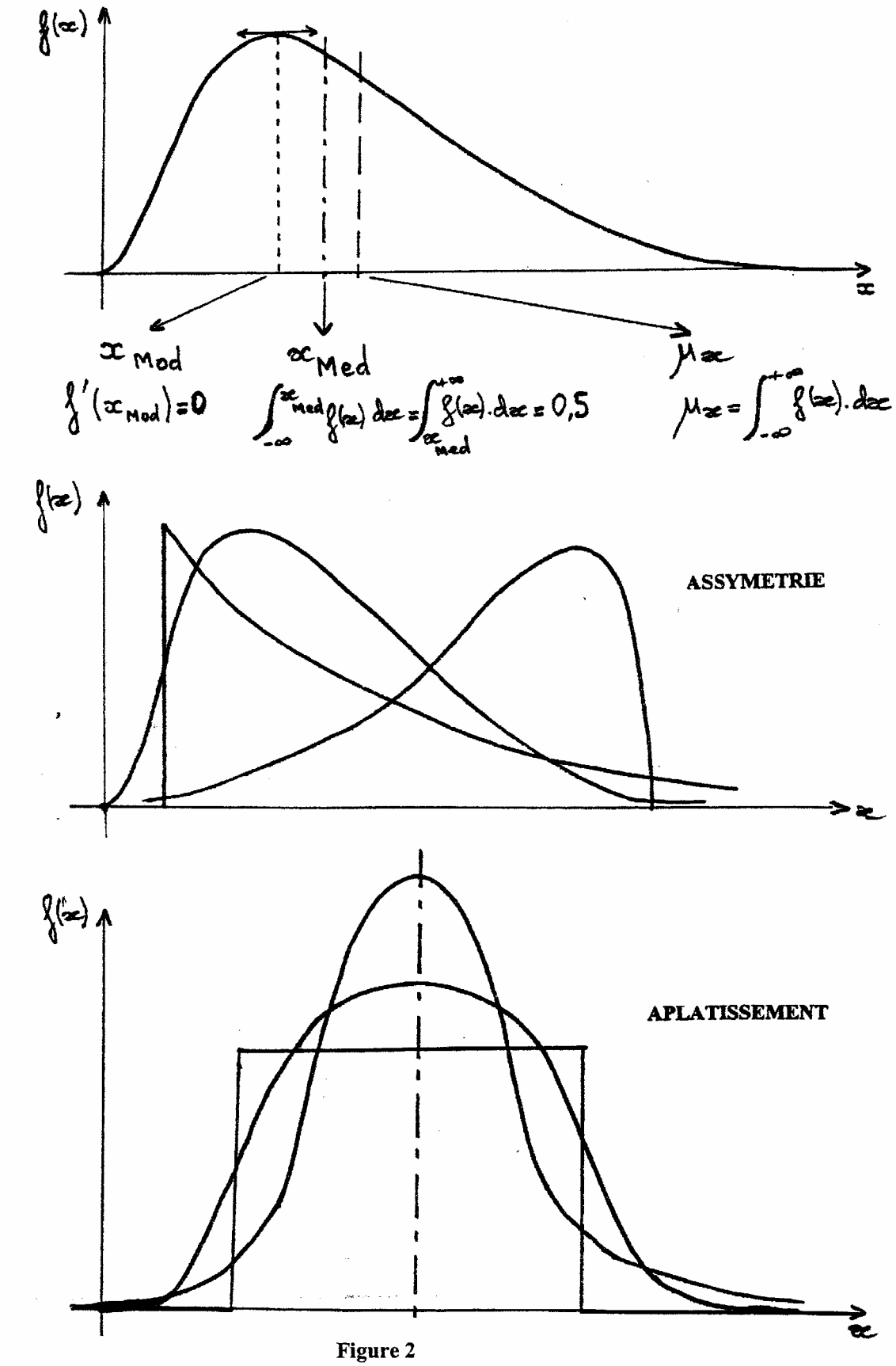

f(x)
x
x Mod
x Med
μx
f'(x Mod) = 0
∫ f(x) dx = ∫ f(x) dx = 0,5
μx = ∫ f(x) dx
ASSYMETRIE
APLATISSEMENT

Figure 2

d) *Paramètres de **distribution : médiane et quantiles***

On a déjà vu la médiane, qui est le quantile 50%.

Plus généralement, on dira que **$Q_{k\%}$ est le "quantile k %"** de l'échantillon si k% des valeurs observées x_i sont inférieures ou égales à $Q_{k\%}$.
Les plus utilisés sont :

Premier décile : Valeur non dépassée dans 10 % des cas
Dernier décile : Valeur non dépassée dans 90 % des cas
(ou non atteinte dans 10 % des cas)

Médiane : Valeur non dépassée dans 50 % des cas.

Ces paramètres sont relativement robustes (plus que les extrêmes !).

On parlera parfois, pour caractériser la dispersion, d'intervalles **interquantiles** :

$X_{90} - X_{10} \rightarrow$ inter**décile** $X_{75} - X_{25} \rightarrow$ inter**quartile**

II-3) Paramètres d'ASYMETRIE :

On définit le coefficient d'asymétrie CS (*Coefficient of Skewness*) sur la population par:

$$CS = \frac{\mu_{3_x}}{\mu_{2_x}^{\frac{3}{2}}} \quad \text{estimé sur l'échantillon par} \quad CS = \frac{m_{3_x}}{s_x^3}$$

où $\mu_{2X} = \sigma_X^2$ et μ_{3X} sont respectivement les Moments centrés d'ordre 2 et 3.

On a vu que le premier était estimé par:

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

quand à μ_{3X}, moment d'ordre 3, on l'estime par:

$$m_{3_x} = \frac{1}{(n-1).(n-2)}\left[n.\sum_{i=1}^{n} x_i^3 - 3.\sum_{i=1}^{n} x_i . \sum_{i=1}^{n} x_i^2 + \frac{2}{n}.\left(\sum_{i=1}^{n} x_i\right)^3 \right]$$

CS est un paramètre peu robuste si n est petit (i.e. limité à quelques dizaines).

II-4) Paramètres d'APLATISSEMENT :

Déjà moins utilisés, ils caractérisent si, pour une même valeur des paramètres précédents, la distribution est plus ou moins aplatie ou au contraire concentrée en pic autour de l'axe.

Ce paramètre (appelé *kurtosis* en anglais) dépend du moment d'ordre 4 de la population; il s'écrit:

$$\frac{\mu_{4x}}{\sigma_x^4}$$

Là encore, si l'échantillon est petit, il est peu robuste et surtout très sensible aux valeurs extrêmes.(On l'utilise peu en hydrologie)

Exemples sur données simulées:

On a tiré d'une loi de Gauss, (loi de probabilité simple et assez répandue - cf. Chap. II), de moyenne *théorique* 1000 et d'écart type *théorique* 200, 20 échantillons différents (10 de taille 10 et 10 de taille 100). On verra dans un chapitre ultérieur comment on génère des données simulées.

Pour chaque échantillon, on a calculé la moyenne, les extrêmes, l'écart type, la médiane (que l'on a pris comme moyenne des 5° et 6° valeurs dans l'ordre croissant pour les échantillons de taille 10, et moyenne des 50° et 51° valeurs dans l'ordre croissant pour les échantillons de taille 100. Le tableau I décrit ces valeurs.

Echantillon N° ↓	Moy	Min	Max	1° déc.	Médiane	9° déc.	s	N	
n=10									
1	1067	746	1408	-	1070	-	211	10	
2	1036	827	1284	-	1040	-	130	10	
3	1002	868	1149	-	982	-	87	10	
4	983	584	1457	-	860	-	329	10	
5	974	644	1250	-	1014	-	164	10	
6	893	757	1203	-	860	-	129	10	
7	973	764	1253	-	950	-	179	10	
8	1006	655	1368	-	990	-	241	10	
9	1046	699	1345	-	1050	-	191	10	
10	977	700	1295	-	980	-	175	10	
n=100									
11	1016	552	1487	720	1005	1200	200	100	
12	1003	559	1589	760	992	1200	184	100	
14	975	459	1481	737	988	1218	191	100	
15	992	463	1506	715	990	1315	212	100	
16	995	634	1409	752	990	1222	181	100	
17	1001	611	1529	750	992	1240	191	100	
18	1025	562	1518	749	1017	1311	209	100	
19	979	550	1577	766	963	1182	176	100	
20	1031	560	1474	784	1020	1290	185	100	

TABLEAU I : Echantillons générés aléatoirement.

Note : on a noté 1°déc. et 9° déc. = premier et dernier décile;
ceux ci n'ont pas été déterminés pour les échantillons 1 à 10 de taille 10.

On constate :

- la robustesse des moyennes et des médianes.

- la grande variabilité des extrêmes d'un échantillon à l'autre.

En outre, on pourrait retrouver que la précision d'estimation (écart entre la valeur dans la population et dans l'échantillon) est fonction de la racine carrée de la taille; c'est à dire que les paramètres calculés sur les échantillons de taille 100 ne sont pas 10 fois plus précis que ceux calculés sur les échantillons de taille 10 mais plutôt 3 fois plus précis.

Résumé:

Pour décrire numériquement et simplement un échantillon, on donnera en général:

- la moyenne arithmétique
- l'écart type
- la médiane
- les déciles inférieurs et supérieurs

Exemple sur données réelles *:*

On donne, ci-contre, un tableau de valeurs de débits de la Romanche à Rioupéroux. Il est difficile en la scrutant de s'en faire une idée rapide.
Mais comme on peut le voir ci dessous, le petit résumé des valeurs précédemment définies renseigne rapidement sur les valeurs de la fluctuation des débits:

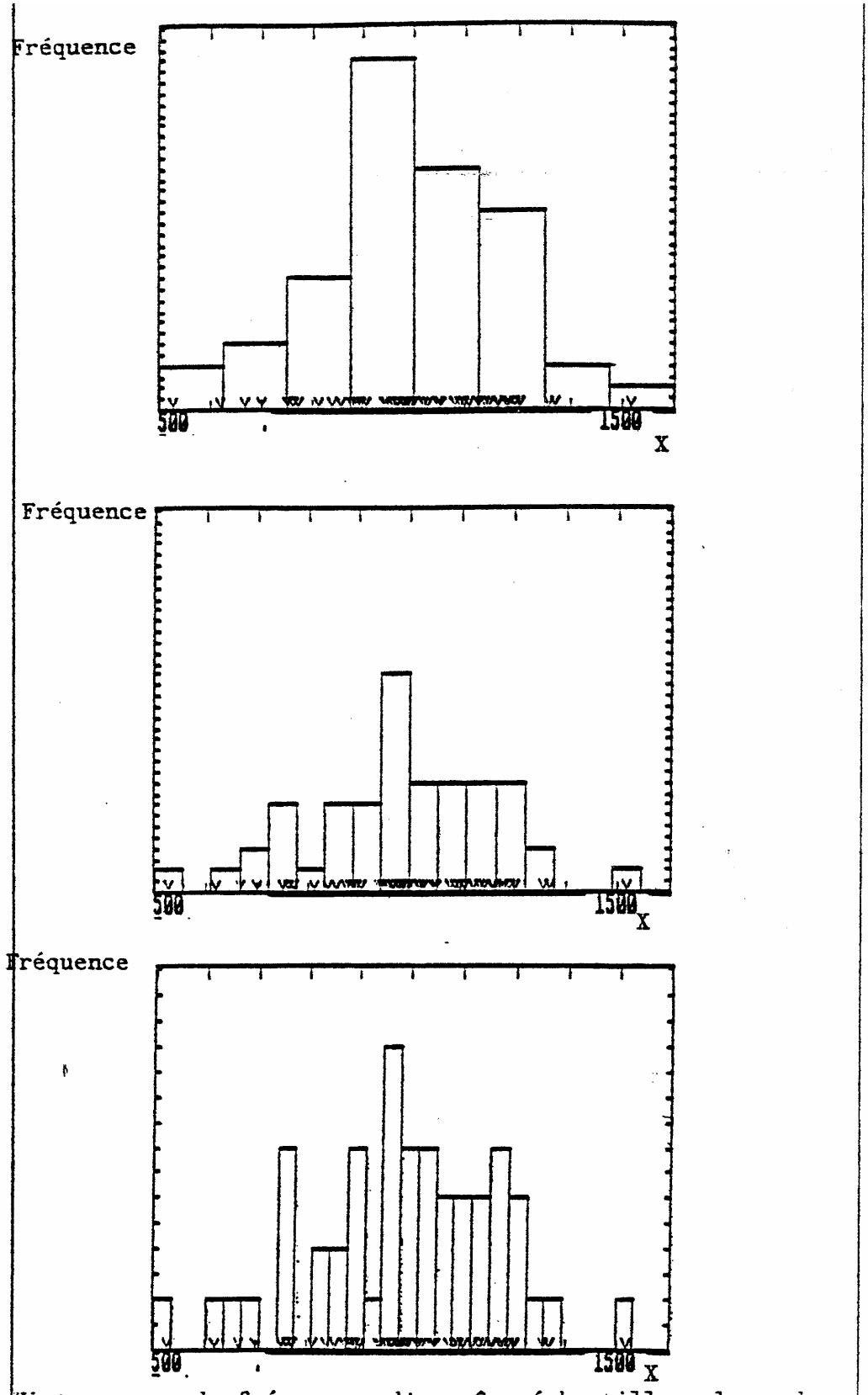
Moy:	12.7	13.	17.2	29.2	65.	87.3	75.6	55.2	37.9	27.6	23.6	16.3	38.4
s :	5.5	6.5	7.1	11.4	28.6	26.1	19.9	14.0	13.8	14.7	15.3	6.8	8.2
1°d.	7.2	6.5	10.	14.	37	53	56	41	22	16	11	9.2	29
Méd:	12	12.5	16.	28	61	85	72	52	35	22	19	15.2	37
9°d.	17	18	26	45	94	120	100	76	58	52	38	24.2	52
Min	2.2	5.1	7.2	8.1	22.9	35.9	42	22.6	17.8	12.3	5.6	2.9	16.9
Max	37.5	40	41.2	57.1	182	140	143	86.6	71.3	86.8	89.1	38.2	58

Tableau récapitulatif des valeurs numériques les plus significatives des débits de la Romanche à Rioupéroux

1°d : = 1° décile (valeur non atteinte dans 10% des cas)
Médiane : Valeur non atteinte dans 50% des cas
9°d : = 9° décile (valeur non atteinte dans 90% des cas.

Débits mensuels (en m^3/s) de la Romanche à Rioupéroux de 1907 à 1948 .

AN	J	F	M	A	M	J	J	A	S	O	N	D	Ann.
1907	7.0	6.7	13.3	22.2	67.8	111	65	63.8	26.2	24.5	16.2	23.8	37.3
1908	9.0	9.2	9.7	14.4	83.2	82.4	63.3	43.2	28.1	17.0	11.1	8.7	31.6
1909	7.6	6.4	7.3	33.0	43.5	46.9	54.7	44.0	20.9	22.2	14.6	14.7	26.3
1910	16.7	13.1	13.4	21.3	47.9	117.	92.9	58.6	21.1	24.9	25.5	24.1	39.8
1911	15.2	13.0	14.3	18.9	40.3	95.9	83.8	51.0	27.6	22.9	15.5	12.0	34.2
1912	13.1	12.9	18.5	31.1	81.7	92.8	78.0	64.6	23.2	28.1	16.5	10.3	39.2
1913	9.1	8.9	16.0	25.9	62.8	104.	50.2	40.6	36.0	27.5	23.1	14.0	34.9
1914	8.2	10.1	22.9	57.1	61.6	60.4	86.5	86.6	33.0	17.8	16.3	13.2	39.5
1915	11.5	13.2	14.0	21.6	107.	111.	93.5	48.7	21.0	17.0	15.9	23.4	41.5
1916	12.9	16.6	14.1	30.0	89.2	86.6	75.1	46.2	26.0	18.1	26.3	18.9	38.3
1917	13.3	9.0	10.1	12.1	96.9	101.	58.6	48.3	29.0	29.5	15.3	9.9	36.1
1918	8.6	7.9	7.8	8.1	45.1	55.5	65.6	44.0	58.1	15.6	9.8	20.3	28.9
1919	14.4	13.0	14.9	26.3	81.6	130.	64.6	61.5	24.5	20.6	22.2	17.6	41.0
1920	20.5	16.0	25.8	30.3	93.2	77.7	77.6	22.6	41.2	18.1	11.0	6.7	36.0
1921	7.2	6.4	7.2	10.0	22.9	35.9	42.0	32.1	17.8	12.3	5.6	2.9	16.9
1922	2.2	5.1	11.4	14.0	47.0	54.8	67.9	57.8	28.6	20.3	30.1	15.2	29.5
1923	11.8	13.5	17.0	25.0	77.6	68.6	107.	50.2	27.4	42.4	23.8	22.9	40.7
1924	16.7	10.7	19.0	43.8	115.	99.1	76.6	31.4	50.5	21.5	22.1	15.6	43.5
1925	9.2	11.3	10.6	19.2	52.9	93.9	80.8	84.4	24.7	16.5	14.4	14.0	36.0
1926	18.4	26.1	24.3	43.0	67.0	132.	95.5	58.0	39.0	57.2	53.2	18.3	52.7
1927	14.0	13.7	21.2	41.8	182.	118.	88.2	76.0	58.4	21.1	34.1	16.0	57.1
1928	14.3	20.1	17.2	24.0	41.1	99.0	52.1	43.2	56.0	86.8	61.7	22.2	44.8
1929	14.4	13.1	17.4	21.5	61.6	99.7	60.7	47.8	30.0	24.2	15.2	17.0	35.2
1930	15.8	13.0	22.4	28.6	73.4	122.	82.0	58.1	40.4	53.3	33.1	22.0	47.0
1931	16.3	12.5	41.2	30.0	47.5	140.	65.0	82.8	46.7	32.9	21.6	13.1	45.8
1932	12.3	9.0	10.2	14.9	42.7	55.9	66.8	50.4	42.3	37.5	14.6	10.2	30.6
1933	10.7	11.1	11.8	24.1	39.1	52.4	81.0	52.7	48.0	53.7	29.1	16.2	35.8
1934	11.2	11.9	14.8	35.5	75.0	72.5	58.3	51.2	33.8	21.2	18.3	24.2	35.7
1935	11.3	12.6	19.9	28.5	62.5	110.	106.	74.7	45.0	51.1	89.1	26.6	53.2
1936	37.5	40.0	28.8	44.5	83.4	93.2	143.	73.8	68.2	35.1	26.4	21.7	57.9
1937	17.0	32.2	33.5	45.6	84.1	108.	87.6	64.1	60.3	28.6	19.7	38.2	51.6
1938	15.8	14.3	17.2	22.4	23.6	59.1	61.1	45.7	23.5	19.6	14.3	13.0	27.5
1939	9.8	9.8	11.8	26.3	30.6	83.7	78.5	52.4	33.8	34.3	37.6	16.1	35.4
1940	11.2	13.2	21.5	31.0	51.6	80.4	92.8	52.4	56.2	31.0	33.2	17.4	41.0
1941	14.9	16.0	20.1	33.4	51.8	123.	122.	77.3	34.5	16.9	17.4	9.8	44.7
1942	7.2	6.4	18.9	28.8	49.4	66.6	57.8	48.2	44.7	22.9	26.4	10.2	32.3
1943	9.0	9.5	15.6	39.9	58.7	72.6	59.2	59.9	71.3	17.6	10.3	10.8	36.2
1944	7.8	8.8	10.6	30.5	41.3	45.6	59.5	51.6	47.2	32.9	41.4	29.2	33.9
1945	10.8	12.7	17.5	47.8	76.4	83.7	70.9	51.8	25.9	16.0	14.0	10.5	36.5
1946	9.0	11.5	15.7	38.4	36.5	72.2	78.4	52.5	34.9	12.3	10.6	9.2	31.8
1947	12.5	11.4	29.4	51.6	79.2	73.2	65.1	50.3	34.6	16.3	16.1	13.0	37.7
1948	18.7	16.0	26.4	31.6	54.9	77.3	61.4	65.5	51.3	20.0	19.2	10.1	37.7



Histogrammes de fréquences d'un même échantillon; le nombre de classes est différent et fait croire à 3 échantillons différents

Figure 3

III) DESCRIPTION GRAPHIQUE :

Objectifs *:*

Présenter sur un graphique les caractéristiques essentielles de l'échantillon.

III-1) HISTOGRAMME des FREQUENCES EMPIRIQUES:

C'est une idée simple:
on se fixe des classes $\mathbf{C_k}$ définies par leurs bornes $[a_k , a_{k+1}]$ on compte le nombre de valeurs de l'échantillon dans chaque classe.

Avantages : Facile à comprendre

Défaut : Le nombre de classes et le choix des classes est laissé à l'initiative de l'individu. Si bien que pour un même échantillon, surtout s'il est de taille assez réduite (qq. dizaines d'éléments), les aspects de ces histogrammes peuvent être assez différents selon le choix effectué.

Les figures 3 ci-contre illustrent cette variabilité de tracés d'un choix à l'autre. C'est pourquoi ce mode de description n'est pas très utilisé surtout si l'échantillon est de taille assez réduite.

Une *règle empirique* consiste à prendre:

Nc = nombre de classes = $1 + 4/3 \log(N)$

(avec N = taille de l'échantillon et le log est Népérien)

Exemple : Pour N = 30 , on fera environ 5 classes, pour N = 50 , 6 classes
et pour N = 100 , 7 classes…

Le tracé de **l'histogramme**, surtout avec un échantillon bien fourni, *permet de supputer la forme de la* **densité de probabilité** f(x) (symétrique ou non, uni- ou multimodale etc...) et de choisir un ou des modèles possibles.

Ceux ci seront ensuite testés et validés, mais *plutôt sur la fonction de répartition*.

**** Complément d'interprétation (sur l'histogramme):*

Pour aider à la compréhension, on peut donner une petite analogie "mécanique" à la moyenne: quand on construit l'histogramme, on donne un **poids** de 1 à chaque individu.
Si on considère l'axe des x comme le bras d'une balance, on peut alors chercher le point pivot de cet axe tel que le **moment** des forces qui s'exercent à droite et à gauche se compensent. C'est le barycentre, ou encore la *moyenne*.
On comprend alors que, si on ajoute ne serait-ce qu'un seul point mais très écarté de la distribution, son bras de levier est tel qu'il faut sensiblement déplacer le pivot pour compenser son effet et rétablir l'équilibre.

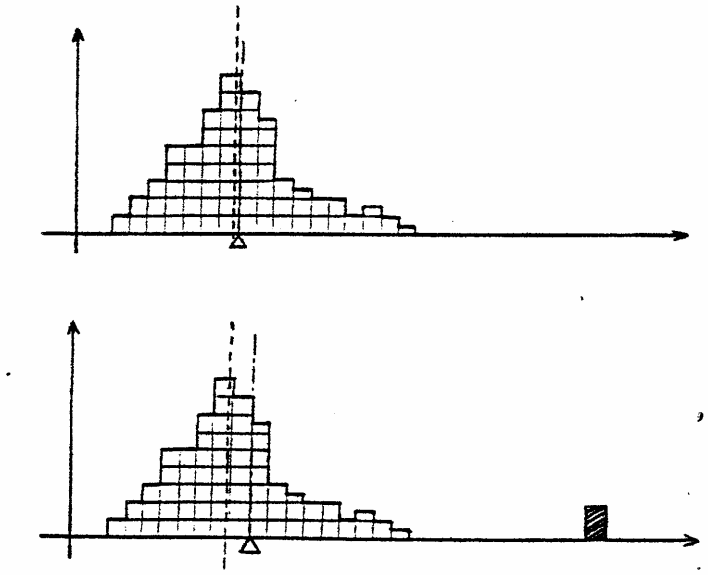


Figure 4

Par contre, ce point ne modifiera pas beaucoup la médiane, telle que 50% des points sont à gauche et 50% à droite, (mais peu importe leur éloignement sur l'axe...!):

⇒ **La médiane est donc plus robuste que la moyenne.**

De même on peut penser décrire la dispersion autour de la moyenne comme le font les mécaniciens pour décrire l'**inertie à la rotation** d'un corps autour d'un axe. Si on prend un axe vertical passant par la moyenne m_X, et que l'on fait tourner l'histogramme autour de cet axe, le moment d'inertie des points d'abscisse x_i et de masse 1 sur une droite serait:

$\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2$ et on pourrait en prendre la moyenne par individu:

(cf. Théorème de Huyghens: le moment d'inertie d'ordre 2 par rapport à un axe est minimum si cet axe est situé au centre de gravité).
Ici encore, l'adjonction d'un individu éloigné de l'axe augmente sensiblement l'inertie de rotation, et donc la variance empirique (qui sera moins robuste qu'un intervalle interdécile).

Enfin, plus on considère des moments d'ordre élevé, plus un individu "extrême", un **horsain**, aura de poids dans le calcul de ce moment (d'où une sensibilité croissante des moments à l'échantillonnage quand leur ordre augmente)

On remarquera aussi que des échantillons (ou des populations) plus "étalés" ou dispersés ont évidemment une variance plus grande, et donc qu'il faut "mécaniquement" plus d'énergie pour les mettre en rotation autour de leur axe.

Note:
Ces considérations "mécanistes" n'ont pas pour seul but d'aider les personnes de formation mécanicienne à se raccrocher à des notions connues. Elles seront souvent à la base des raisonnements utilisés en statistique multidimensionnelle (analyse en composantes principales, analyse discriminante, etc...)

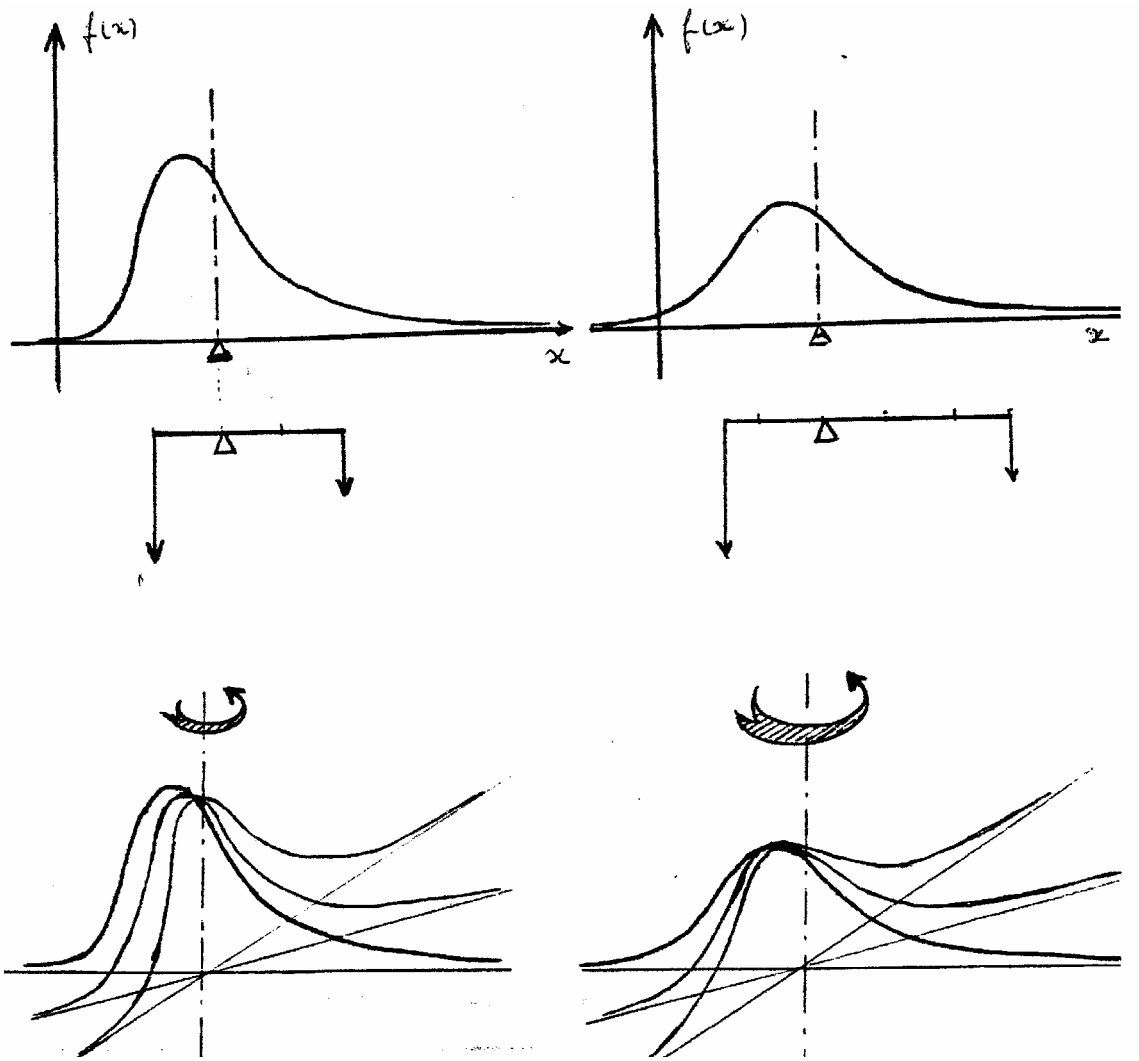


Figure 5

III-2) COURBE des **FREQUENCES CUMULEES**
FONCTION DE REPARTITION EMPIRIQUE

Objectifs *:*

Trouver une représentation graphique assez complète pour décrire l'échantillon.

Cette fois on va chercher :

- à utiliser toute l'information donnée par l'ensemble des valeurs (ce que l'on ne faisait pas quand on regroupait en classes avec l'histogramme des fréquences relatives).
- à anticiper sur les méthodes d'ajustements probabilistes (cf. . Chap. II)

La première idée est de tracer la *courbe en escalier :*

$F^*(x_i)$ = Proportion des valeurs de l'échantillon inférieures ou égales à x_i

= Fréquence empirique, observée, des valeurs x_i inférieures ou égales à x_i. = $\frac{i}{N}$ (où N est la taille de l'échantillon).

Le défaut est que l'on ne donne pas la même importance au minimum qu'au maximum,

puisque: $F^*(Min) = \frac{1}{N}$ et $F^*(Max) = 1$.

$\Rightarrow$ D'où l'idée des statisticiens :

- ***si*** l'échantillon est tiré d'une loi de probabilité définie par sa fonction de répartition F(x) = Probabilité qu'une valeur X tirée au hasard de la population soit inférieure ou égale à x,
- ***essayons*** de tracer à partir de l'échantillon une courbe la plus voisine de F(x) (en général inconnue).

Ceci permettra non seulement une description de l'échantillon mais peut être une aide à la recherche de F(x).

Pour cela ***classons les n valeurs*** x_i *dans l'ordre croissant*

$\Rightarrow$ d'où un échantillon de N valeurs x_i classées.

On montre qu'une bonne estimation assez simple de $F(x_j) = Pr(X \leq x_j)$ est fournie par :

$$F^*(x_i) = \frac{i-a}{N+b}$$

où a et b ont un optimum *qui dépendent de la loi dont sont issus les échantillons…*
Il faudrait donc la connaître **a priori** pour bien choisir la façon de pointer les valeurs observées , alors que l'on fait ce pointé justement pour essayer de déterminer la loi la plus plausible… On fera donc des paris et des compromis…

Exemples: Loi Normale (Gauss) a = 0.375 b = 0.25 (cf. définitions de ces lois dans le chapitre II)

Loi de Gumbel a = 0 b = 1

Nous prendrons souvent: a = 0.5 et b = 0.5 ou **a = 0.5 et b = 0**

d'où les formules d'estimation de la probabilité empirique

$$Pr(X \leq x_{\mathbf{i}}) = \frac{2.i-1}{2.N+1} \text{ ou } \frac{2.i-1}{2.N}$$

avec **i** le **rang** de la valeur $\mathbf{x_i}$

Attention:

Le choix de cette façon d'estimer la probabilité et de la pointer sur un diagramme ("plotting position" en anglais) n'est pas tout à fait neutre et a reçu une grande attention
de la part de certains auteurs (cf. Yevjevitch V. 1972 ou Haan Ch.T. 1977, p. 135 ou, plus récemment, et pour une loi particulière, l'article de Nophadol et Nguyen 1989).
On verra dans l'analyse des valeurs extrêmes que cela a une certaine importance.

_____

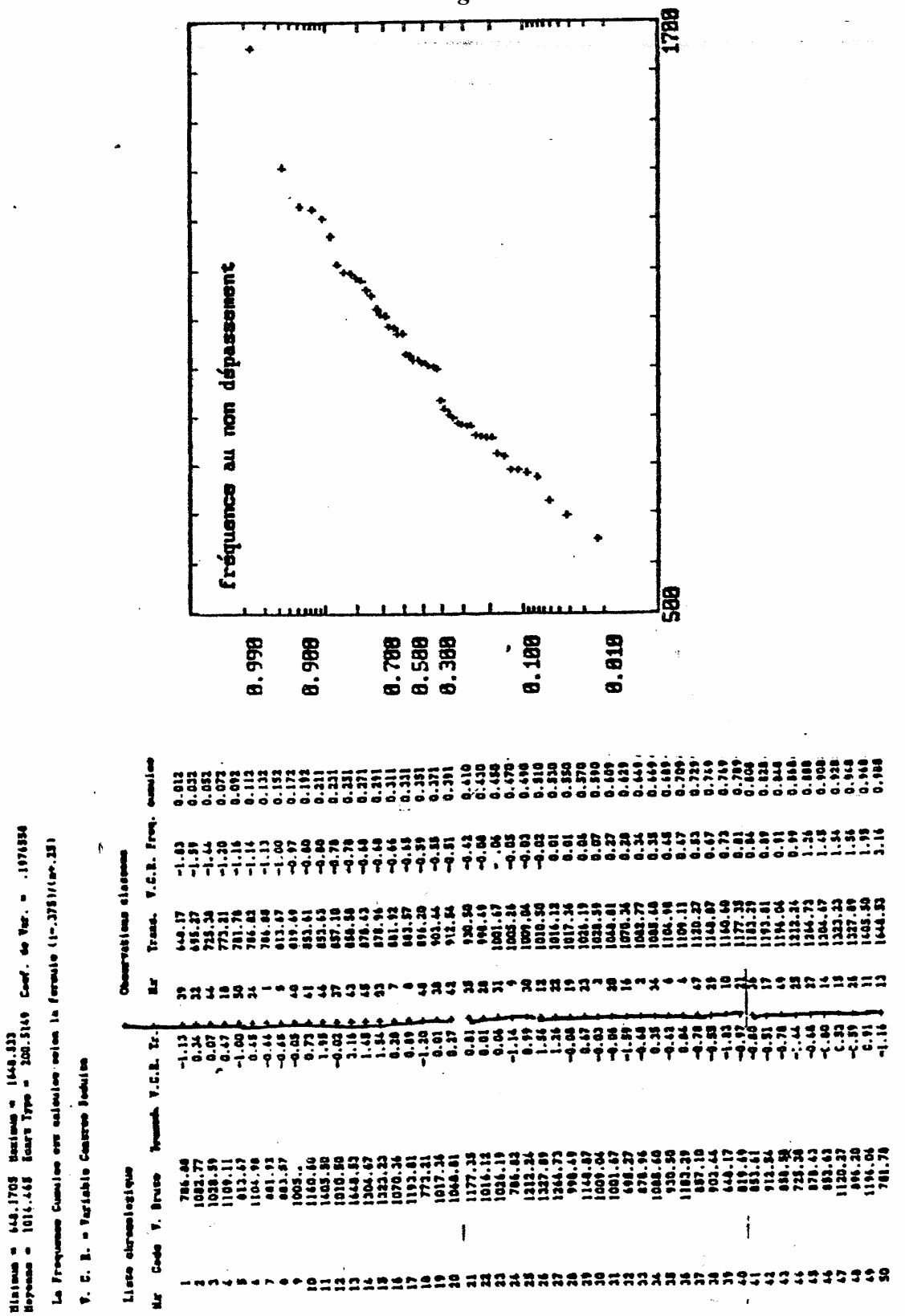
On trace ensuite les points sur un diagramme.
Mais en diagramme *arithmétique*, où les axes Ox et Oy sont gradués linéairement, les courbes obtenues ont souvent la forme d'un S (sigmoïdes) et il est difficile d'en déduire une forme de loi et de les distinguer. C'est pourquoi on utilise souvent des papiers où l'échelle des F* est distordue (papier de Gauss, papier de Gumbel).

L'intérêt de ces **diagrammes fonctionnels,** dits de probabilité ,

- meilleure lecture pour certaines probabilités
  (les extrêmes par exemple pour Gumbel)
- tracé plus aisé de certaines lois
  (droite pour une loi de Gauss sur papier de Gauss).

Les tableaux et la figure 5 de la page suivante illustrent cette description. Le papier utilisé est un papier de Gauss dont on verra la construction par la suite.

Figure 5


fréquence au non dépassement
1700
500
0.990
0.900
0.700
0.500
0.300
0.100
0.010

IV) COMPLEMENTS THEORIQUES

IV-1) NOTION de **PERIODE DE RETOUR:**

a) *Variables aléatoires en Hydrologie*
Période de retour, Durée de retour.

Quand on définit une variable aléatoire, il est fréquent qu'on lui associe un intervalle de temps:

X1 = Total de la pluie du mois d'Octobre.
X2 = débit moyen annuel
X3 = durée d'insolation des mois d'été, etc...
X4 = Pluie maximale journalière de chaque année

On définit donc implicitement:

- Une notion d'évènement, ou de "**tirage**" aléatoire dans l'espace des évènements
- souvent associée, dans le cas où les variables sont en fait des processus temporels se déroulant dans le temps, à un **intervalle de temps.**

Exemples:

Pour X1, c'est le "mois d'Octobre" (Il n'y en a qu'un par an et on considère que les autres mois, le total pluviométrique a un comportement différent).

Pour X2, c'est l'année. (On considère que deux années successives, bien qu'aboutées, correspondent à 2 tirages "indépendants" de la variable).

Pour X3, c'est la "saison d'été". (Il n'y en a qu'une par an, car on considère là aussi que l'insolation a un comportement différent sur les autres saisons).

Pour X4, c'est l'année, dans laquelle on va chercher quel est le total pluviométrique journalier le plus fort.

….etc…

_____

Quand ensuite, on dit que : $Pr(X \leq \alpha) = 90\%$,

cela signifie que: - si on fait un tirage indépendant de la variable X
- il y a 9 chances sur 10 d'être inférieur ou égal à α

Statistiquement, si on faisait plusieurs fois (par exemple K fois) des paquets de N tirages indépendants, on trouverait que, **en moyenne** sur les K fois**,** sur les N tirages d'un paquet, 0.10.N dépassent α. (même si pour un paquet donné de N tirages, on peut avoir un résultat différent de 0.10.N)
On dira alors que la valeur α est dépassée **en moyenne** 1 fois tous les 10 tirages.

Par abus de langage, on dit que la valeur α "revient" en moyenne tous les 10 tirages et donc qu'elle a une **"période" de retour moyenne** de T= 10, en fait de 1 fois tous les 10 tirages.

Quand en plus, chaque "tirage" est associé lui-même à un intervalle de temps, par exemple si on ne fait que un tirage par an, on dira que la valeur α, qui "revient" en moyenne tous les 10 tirages, a une **durée de retour** moyenne de **T** =10 ans (exprimée dans la même unité que l'intervalle inter-tirages), et que la valeur α est **décennale**.

Si, au lieu de prendre un seuil particulier $F(\alpha) = 0.9$, on prend un seuil quelconque $F(x_F) = F$ fixée , avec F prise de manière quelconque $\in [0,1]$, alors la **période de retour** est:

$$T = \frac{1}{1-F}$$

et ainsi:

F = 0.9 $\quad T = \frac{1}{1-0.9}$ $\quad$ T = 10

F = 0.95 $\quad T = \frac{1}{1-0.95}$ $\quad$ T = 20 etc...

Exemples:

Si Pr (X2 >250 m3/s) = 0.1, on dira que le débit moyen annuel de 250 m3/s est dépassé en moyenne 1 tirage sur 10, donc 1 année sur 10 en moyenne, donc a une "période de retour" *décennale*.

De même si Pr (X1 < 100 mm) = 0.9, on dira que la valeur 100 mm est dépassée en moyenne 1 tirage sur 10, donc 1 mois d'Octobre sur 10, et donc a une période de retour *décennale* (car il n'y a qu'un mois d'Octobre et donc qu'un tirage possible par an)

De même pour X3.

b) *Complément sur les probabilités empiriques*
(et les ajustements graphiques)

On a vu dans l'analyse des échantillons qu'il fallait associer à chaque valeur x_i de rang i une probabilité empirique au non dépassement.

La plus simple consiste à prendre: $F^*(x_i) = Pr(X \leq x_i) = \frac{i}{N}$

Si pour illustrer, on prend N= 100, on voit que:

$F(x_1) = 0.01$ mais que $F(x_N) = 1$ ...!

Ceci est gênant puisqu'alors Pr ($X > x_N$) = 0 ...!
or on a toute raison de penser que si on augmente l'échantillon on trouvera des valeurs supérieures à x_N….

C'est pourquoi on a "bricolé" des formules de la forme: $P_i = \frac{i-a}{N+b}$

¤ Dans le cas de $P_i = \frac{i-0.5}{N}$, on voit que (avec N = 100):

$P(X \leq x_1) = 0.005$ et $P(X < x_N) = 0.995$ ou $P(X > x_N) = 0.005$

Soit encore, *en terme de période de retour*:
on considère, **et on impose**, par cette formule que les valeurs x_1 et x_{100}, min. et max. d'un échantillon de 100 valeurs, reviennent en moyenne 1 fois tous les **200** tirages.

¤ Dans le cas où on choisit une formule, tout aussi symétrique entre mini et maxi:

$$P_i = \frac{i}{N+1}$$

on voit que (avec N =100) cela revient à considérer que:

$P(X \leq x_1) = 0.01$ et $P(X > x_N) = 0.01$

soit encore que, en terme de période de retour, il reviennent tous les **100 tirages**,
soit **deux fois plus souvent** qu'avec la formule précédente...

Par contre la probabilité de l'évènement médian, x_{50} , reste dans les deux formules très proche de 50% et la période de retour correspondante proche de 2.

Conclusions:

Il faut donc considérer que :la probabilité *empirique* est proche de la probabilité réelle, (- ou au moins est estimée de façon stable -), **dans la partie *centrale* de l'échantillon, mais certainement pas dans les *queues de la distribution*** à gauche et à droite.

En conséquence, dans les ajustements graphiques, il faudrait pondérer plus faiblement les points extrêmes, car on leur a attribué une probabilité *empirique* parfois éloignée de la réalité, et surtout trop dépendante de la formule d'estimation retenue.

Notons cependant que la formule $P_i = \frac{i}{N+1}$ tend à considérer les évènements extrêmes comme plus fréquents, et donc *va dans le sens d'une certaine sécurité*.

Ces notions de durée de retour seront largement utilisées en Hydrologie de Projet.

IV-2) CHANGEMENTS de VARIABLES:(*)

Soit une variable aléatoire **X** dont la densité de probabilité est **f**(x).
On va souvent chercher à savoir quelle est la forme de la distribution de la variable aléatoire **U**, obtenue par une transformation U = **g**(X). On appellera cette nouvelle distribution , i.e. la densité de probabilité de U, **h**(u).

On montre alors, (cf. Benjamin and Cornell 1970) que:

$$h(u) = f(x).\frac{dx}{du} \quad \text{avec} \quad x = g^{-1}(u) \ \text{ et } \ \frac{du}{dx} = g'(g^{-1}(u))$$

soit encore:

$$h(u) \ = \ \frac{f\left[g^{-1}(u)\right]}{g'\left[g^{-1}(u)\right]}$$

Exemple: (tiré de T. Haan)

Soit une variable X variant entre 0 et 5 et de densité de probabilité $f(x) = \frac{3.x^2}{125}$

On vérifie que: $F(x) = \int_0^{x \le 5} f(t).dt = \int_0^x \frac{3.t^2}{125} dt = \left[\frac{t^3}{125}\right]_0^x = \frac{x^3}{125}$

et donc que F(0) = 0 et F(5) = 1

On considère maintenant la variable U = X² avec cette fois $0 \le U \le 25$.

alors: $U = g(X) = X^2 \ \Rightarrow \ X = g^{-1}(U) = \sqrt{U}$

de plus: $\frac{dU}{dx} = g'(X) = 2.X \ = g'\left[g^{-1}(U)\right] \ \Rightarrow \ = 2.\sqrt{U}$

Si on reporte dans :

$$h(u) \ = \ \frac{f\left[g^{-1}(u)\right]}{g'\left[g^{-1}(u)\right]} \quad \text{alors} \quad h(u) \ = \ \frac{f\left[g^{-1}(u)\right]}{g'\left[g^{-1}(u)\right]} \ = \ \frac{\frac{3.(\sqrt{u})^2}{125}}{2.\sqrt{u}} \ = \ \frac{3}{2}.\frac{\sqrt{u}}{125}$$

On peut même vérifier que h(u) est bien une densité de probabilité. Par exemple:

$$\int_{u=0}^{u=25} h(u).du = \frac{1}{125}\int_{u=0}^{u=25} \frac{3}{2}.\sqrt{\mathrm{u}}.du = \frac{1}{125}.\left[u^{\frac{3}{2}}\right]_0^{25} = 1$$

Utilisation:
Il arrivera fréquemment que, après transformation de la variable d'intérêt, la variable transformée suive une loi "simple" et pratique à manipuler.
On fera donc référence en quelques occasions à ce paragraphe.

BIBLIOGRAPHIE:

BENJAMIN J.R and CORNELL C.A. (1970).
Probability, Statistics and Decision for Civil Engineers
Mac Graw Hill Pub. Comp. 684 p.

Groupe CHADULE (1974)
Initiation aux méthodes statistiques en Géographie.
(Ouvrage collectif) Masson et Cie ed. 192 p.
(Ouvrage probablement épuisé mais disponible en bibliothèque)

HAAN Ch. T. (1977)
Statistical Methods in Hydrology.
Iowa state University Press 2ème ed. 1979, 378 p.

KOTTEGODA N.T. and R. ROSSO (1997)
Probability, Statistics and Reliability for Civil Engineers and Environmental Engineers
The Mac Graw Hill Pub. Comp. Inc. 735 p.

MORLAT G. (1954)
Les méthodes statistiques
Conférences faites par G. Morlat du 21 Avril au 9 Juin 1952. rassemblées dans un ouvrage. Direction des Etudes et Recherches d'EDF -(Pour les bibliophiles : disponible en photocopie auprès du service de documentation d'EDF).

NOPHADOL IN-NA and VAN-THANH- VAN NGUYEN (1989)
An unbiaised plotting position formula for the general extreme value distribution
Journal of Hydrology, vol. 106, p. 193-209

VIALAR 1986
Probabilités et Statistiques (5 fascicules)
Cours de l'Ecole Nationale de la Météorologie

YEVJEVICH V. (1972)
Probability and Statistics in Hydrology
Water Ressources Publications Ed Fort Collins Co USA. 302 p.
(Ouvrage très complet sur les modèles probabilistes- le Pr Yevjevich est sorti de l'ENS d' Hydraulique de Grenoble en 1939)

1ère Partie : MODELES PROBABILISTES

CHAPITRE II

MODELES PROBABILISTES LES PLUS COURANTS 35

I-) GENERALITES sur les LOIS de PROBABILITE 37

I-1) Objectifs du chapitre 37
I-2) Lois de probabilité paramétrées 37
I-3) Aperçu sur le calage des paramètres 39

II- FAMILLE DES LOIS NORMALES et DERIVEES: 41

II-1) Loi de Gauss (dite également Loi Normale): 41
II-2) Loi Lognormale (dite également Loi de GALTON) 52
II-3) Aperçu sur d'autres lois dérivées 56

III- FAMILLE DES LOIS GAMMA et DERIVEES: 59

III-1) Loi Gamma à 2 paramètres (ou loi de Pearson) 59
III-2) Calcul des Moments (en fonction des paramètres) 62
III-3) Tables de la loi Gamma 63
III-4) Aperçu sur les lois Bêta 65

IV- FAMILLE DES LOIS EXPONENTIELLES ET VALEURS EXTRÊMES 67

IV-1) Loi exponentielle 67
IV-2) Extension de la loi Exponentielle (Somme d'exponentielles) 69
IV-3) Loi de Gumbel 71
IV-4) Aperçu sur d'autres lois de valeurs extrêmes (Weibull et GEV) 74

V-) QUELQUES LOIS de VARIABLES DISCRETES: 77

V-1) Loi de Poisson 77
V-2) Loi Binomiale 79

VI-) LOIS UTILISEES DANS LES TESTS d'HYPOTHESES:
81

VI-1) Loi du Chi 2 81
VI-2) Loi de Student 81
VI-3) Loi de Fisher-Snedecor 83

1ère Partie - CHAPITRE II :

MODELES PROBABILISTES LES PLUS COURANTS

I-) GENERALITES sur les LOIS de PROBABILITE

I-1) OBJECTIFS de ce CHAPITRE:

Dans le chapitre I, nous avons montré quelques présentations numériques ou graphiques de séries de données, sans faire aucune hypothèse probabiliste sur la population d'origine.

Dans certains cas, on peut penser que ces données peuvent être décrites par une ou plusieurs lois de probabilité courantes et simples d'emploi, au moins dans une certaine gamme de probabilité.

Il est alors intéressant de chercher à ***ajuster*** sur ces données une, ou des lois pour faciliter l'utilisation numérique et parfois, sous certaines réserves, pour en tirer des informations de type probabiliste.

Exemple 1 :

Pour dimensionner une protection contre les crues à Grenoble, on envisage de construire des digues. Plus les digues sont hautes, plus on est protégé, mais plus leur coût est élevé.

Il est donc important de savoir calculer la probabilité d'être inondé pour une hauteur de digues donnée, afin de résoudre ensuite le problème du choix de leur hauteur en termes économiques.

Exemple 2 :

On sait, par expérience, que les pluies annuelles en France sont bien décrites par des lois de Gauss (appelée loi Normale par la suite) dont les moyennes et écarts types varient considérablement d'un endroit à l'autre.

La simple information qu'à Grenoble la moyenne est de 1100 mm et l'écart type de 300 mm permet, après consultation d'une table de Gauss (ou utilisation d'une calculette comportant les fonctions statistiques), de calculer qu'il y a une chance sur dix pour que l'an prochain, il tombe moins de 616 mm.

Le même type de calcul sur les pluies mensuelles ou saisonnières intéressera évidemment les agriculteurs pendant la période de croissance ou de récolte...!

Après les analyses exploratoires du Chapitre I, notamment la forme de l'histogramme, on peut déjà se faire une idée de la *forme* de loi de probabilité adaptée à la représentation de l'échantillon dont on dispose. On va ensuite chercher, parmi les lois que l'on connaît, si une (ou plusieurs) présente une forme analogue, susceptible d'être ajustée à l'échantillon.

Le but de ce chapitre II va donc être de décrire les lois les plus couramment utilisées, avec pour objectif de disposer d'une **boîte à outils**, plus ou moins riche et complète, plus ou moins adaptée à une grande variété de situations.

Exemple:

Un mécanicien peut souhaiter disposer de toute la gamme des clés plates, des clés à anneaux , etc..., mais reconnaître aussi qu'une bonne clé à mollette répond déjà "parcimonieusement" à beaucoup de situations...!).

Ensuite, ayant décrit les outils disponibles et découvert leur propriétés, il va falloir s'en servir et les **ajuster au mieux** sur les données disponibles: $\Rightarrow$ ce sera l'objet du Chapitre III. Ces deux chapitres, indissociables en pratique, ne l'ont été que pour la clarté de l'exposé.

I-2) FONCTIONS **PARAMETREES**

Nous ne décrirons que quelques lois: les plus couramment utilisées en Hydrologie, ainsi que quelques autres d'intérêt général (utilisées par exemple dans les tests d'hypothèses).

Une fonction ***paramétrée*** est en fait *une famille* de courbes qui se résume par une équation *unique* de la variable x, mais comportant des coefficients, des paramètres, qui peuvent prendre une infinité de valeurs. Par exemple les paraboles se résument en un polynôme du second degré en x : $y(x,a,b,c) = a.x^2 + b.x + c$

mais selon les valeurs que l'on donnera aux paramètres a,b,c, on aura une infinité de courbes possibles...

De même la plupart des lois de probabilité s'exprimeront sous la forme:

$f(x, \alpha_1, \alpha_2, ..., \alpha_p)$: Densité de probabilité

c'est à dire que la probabilité de tirer au hasard une valeur de la variable aléatoire X entre x-dx/2 et x+dx/2 est égale à $f(x, \alpha_1, \alpha_2, ..., \alpha_p)dx$

De même on utilisera aussi:

$F(x, \alpha_1, \alpha_2, ..., \alpha_p)$: Fonction de répartition

c'est à dire que la probabilité de tirer au hasard $X<x$ est $F(x, \alpha_1, \alpha_2, .., \alpha_p)$.

Plutôt qu'une fonction particulière, ce seront donc des familles, ou des ***classes de fonctions*** de la variable x **et** d'un certain nombre de ***paramètres*** α_k .

Ces fonctions théoriques correspondront en quelque sorte aux fonctions empiriques que sont l'histogramme de fréquences relatives (Densité de probabilité) et le diagramme des fréquences cumulées (Fonction de répartition) vues au chapitre I.

I-3) APERCU sur le **CALAGE des PARAMETRES**:

Pour déterminer les paramètres α_k, plusieurs méthodes seront utilisées; nous décrirons les plus classiques dans le chapitre III, en détaillant le calcul pour certaines lois.
Dans ce chapitre, nous insisterons donc parfois sur certaines propriétés mathématiques des lois: c'est parce qu'elles sont utiles ensuite dans la mise en oeuvre des techniques d'ajustement.
Signalons donc simplement , parmi ces techniques:

a)-Méthode des Moments :

Soit $f(x, \alpha_1, \alpha_2,...\alpha_p)$ la famille de lois (-une expression théorique paramétrée-),
et soit un échantillon de n valeurs x_i de la variable X.

Dans cette famille de lois, on choisira **la** loi spécifique (-donc on choisira les valeurs spécifiques des paramètres $\alpha_1, \alpha_2,...\alpha_p$ -) telle que:
p Moments *théoriques* de cette loi f(x,...)
soient égaux aux :
p Moments *empiriques* correspondants, calculés sur les x_i.

D'où un système plus ou moins compliqué de p équations à p inconnues (- les α_k -), qui nécessite d'expliciter les relations entre les paramètres et l'expression théorique de ces moments.
Cette méthode donne pour de nombreuses lois des résultats simples, aussi est-elle couramment utilisée. Mais elle donne beaucoup de poids aux valeurs extrêmes, ce qui peut être problématique.

b)- Méthode du Maximum de Vraisemblance :

La probabilité d'avoir eu dans l'échantillon une valeur comprise entre $x_i + dx/2$ et $x_i - dx/2$ est, selon la loi définie par sa fonction densité :

$$f(x_i, \alpha_1,...., \alpha_p)dx = Pr\ (x_i - dx/2 < X < x_i + dx/2)$$

Si les valeurs x_i sont indépendantes, $\Rightarrow$ la probabilité d'avoir tiré (dans n'importe quel ordre) les n valeurs x1 , x2,..., xn (à plus ou moins dx/2) est le produit de ces n probabilités;
$\Rightarrow$ c'est donc une fonction des p paramètres pour les n valeurs x_i données.

La méthode du maximum de vraisemblance consiste à maximiser cette probabilité, c'est à dire choisir les valeurs des p paramètres qui rendent cet échantillon *le plus probable possible*, au vu d'une loi choisie préalablement.
La résolution analytique de cette maximisation est plus ou moins simple selon les lois...

c) Méthode graphique

Elle consiste à trouver un **diagramme fonctionnel** tel que:

- si l'échantillon suit raisonnablement la loi pour laquelle ce diagramme a été conçu,
- alors cela se traduira par un alignement, selon une droite, facile à apprécier à l'oeil.

Si la pratique en est aisée, la conception du diagramme doit être bien comprise et repose sur une bonne compréhension des propriétés de la loi choisie.

On voit donc que ces méthodes nécessitent aussi une bonne connaissance analytique des différentes lois et de leurs moments, ce que nous allons étudier ci-après.

Nous présenterons d'abord quelques familles de lois couramment utilisées en Hydrologie pour des variables réelles, puis quelques lois appropriées à des variables discrètes (- prenant seulement des valeurs entières-).

II- FAMILLE DES LOIS NORMALES et DERIVEES:

II-1-) LOI de GAUSS (dite également **Loi Normale**):

a) Forme analytique:

C'est une loi à **2 paramètres α et β.** La densité de probabilité s'écrit:

$$f(x,\alpha,\beta)=\frac{1}{\alpha\sqrt{2\pi}}.e^{-\frac{1}{2}\left(\frac{x-\beta}{\alpha}\right)^2}$$

et la Fonction de répartition , que l'on écrira souvent $N(\alpha,\beta)$ pour loi Normale de paramètres α, β :

$$F(x,\alpha,\beta)=\int_{-\infty}^{x}\frac{1}{\alpha\sqrt{2\pi}}.e^{-\frac{1}{2}\left(\frac{t-\beta}{\alpha}\right)^2}.dt = \mathrm{N}(\beta,\alpha)$$

et Prob(X≤x) = F(x,α, β)

Si on effectue sur x la transformation linéaire : $x \rightarrow u=\frac{x-\beta}{\alpha}$, on peut montrer que la nouvelle variable *u* suit encore une loi de Gauss (- on le démontrera et on l'utilisera plusieurs fois ci-après -).

⇒ Donc toutes les lois de Gauss peuvent se ramener à la même loi normale centrée réduite ***N(0,1)*** dite *loi standard*, calculée il y a un siècle!
De même on peut revenir de ***N(0,1)*** à $N(\alpha,\beta)$. En effet, nous allons voir que les paramètres sont tels que β est la moyenne et α l'écart type .

Caractéristiques essentielles de cette loi :

- symétrique (d'où **Moyenne ≡ Médiane) ,** et la moyenne correspond aussi à la probabilité de 50% au non dépassement)

- unimodale (la fonction densité n'a qu'un maximum: **Mode** = Moyenne = β= μ_X)

- non bornée à droite comme à gauche

Intérêt de cette loi :

On démontre que, sous certaines restrictions:
- *si* X est la ***somme*** de k variables aléatoires **indépendantes**, tirées dans des lois quelconques
- *mais d'ordres de grandeur voisins en moyenne et écart-type,*
- *alors*, si le nombre k tend vers l'infini, X suit une loi de Gauss.

(En fait il suffit que k dépasse une dizaine pour que cela constitue déjà une bonne approximation).

Or dans la nature, de nombreux phénomènes sont le résultat d'addition de variables aléatoires indépendantes (par exemple les pluies annuelles en France, ou en zone tempérée, là où il pleut souvent), d'où le choix fréquent de cette loi dans ce cas.

Mais attention: d'autres phénomènes aléatoires ne sont pas du tout décrits par des lois de Gauss (par exemple les pluies journalières *maximales* en France.., ou les pluies annuelles au Sahara car ce n'est alors que la somme d'une ou deux pluies journalières!).

b) Calcul des moments **(*)**

Soit : $$f(x,\alpha,\beta)=\frac{1}{\alpha\sqrt{2\pi}}.e^{-\frac{1}{2}.\left(\frac{x-\beta}{\alpha}\right)^2}$$

Par définition, le moments d'ordre 1 va s'écrire:

$$\mu_1 = E[X] = \int_{-\infty}^{+\infty} x.f(x,\alpha,\beta).dx = \int_{-\infty}^{+\infty} x.\frac{1}{\alpha\sqrt{2\pi}}.e^{-\frac{1}{2}.\left(\frac{x-\beta}{\alpha}\right)^2}.dx$$

Si on pose : $v = \frac{x-\beta}{\alpha}$ *donc* $dv = \frac{dx}{\alpha}$ et $x = \alpha.v + \beta$ *d'*où $dx = \alpha.dv$

alors :

$$\mu_1 = \int_{-\infty}^{+\infty} (\alpha.v+\beta).\frac{1}{\alpha\sqrt{2\pi}}.e^{-\frac{1}{2}.v^2}.\alpha.dv = \int_{-\infty}^{+\infty} v.\frac{1}{\sqrt{2\pi}}.e^{-\frac{1}{2}.v^2}.\alpha.dv + \int_{-\infty}^{+\infty} \beta.\frac{1}{\sqrt{2\pi}}.e^{-\frac{1}{2}.v^2}.dv$$

soit encore: $$\mu_1 = \frac{\alpha}{\sqrt{2\pi}}\int_{-\infty}^{+\infty} e^{-\frac{1}{2}.v^2} v.dv + \beta.\int_{-\infty}^{+\infty}\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}.v^2}.dv$$

La seconde intégrale : $\int_{-\infty}^{+\infty}\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}.v^2}.dv$ est justement l'intégrale d'une densité de probabilité (ou encore de la fonction mathématique Fonction Erreur Erf(v)) et donc vaut 1.

La première intégrale s'intègre en : $$\int_{-\infty}^{+\infty} e^{-\frac{1}{2}.v^2} v.dv = \left[-e^{-\frac{1}{2}.v^2}\right]_{-\infty}^{+\infty} = 0$$

D'où il reste que : $\mu_1 = \beta$

$\Rightarrow$ La **moyenne (ou espérance de x)** est égale au **paramètre** β de la loi f(x,α,β).

Les moments suivants seront en général calculés de manière **centrée,** $\Rightarrow$ en écart à la moyenne, i.e. le moment d'ordre 1 μ_1 , que l'on note plus couramment μ_X.

Par exemple, le **moment centré** d'ordre 2 ou **variance** s'écrit:

$$\mu_2 = E\left[(X-\mu_1)^2\right] = \int_{-\infty}^{+\infty}(x-\mu_1)^2 f(x,\alpha,\beta).dx = \int_{-\infty}^{+\infty}(x-\beta)^2 . \frac{1}{\alpha\sqrt{2\pi}} . e^{-\frac{1}{2}\left(\frac{x-\beta}{\alpha}\right)^2} .dx$$

On démontre alors (le faire en exercice, pour le plaisir..!) que ce moment devient:

$$\mu_2 \quad \text{encore noté} \quad V_x \quad \text{ou} \quad \sigma_x^{\,2} \;=\; \alpha^2$$

et donc que, pour la loi normale:

⇒ **l'écart-type σ coïncide avec le paramètre α**
dans l'expression analytique de la loi...!

Plus généralement, on démontre que :

- tous les moments **centrés** d'**ordre impair** (au delà de l'ordre 1) sont nuls:

$$\mu_{2p+1} \;=\; 0 \quad \forall\ p$$

- les moments **centrés** d'ordre **pair** ont pour expression:

$$\mu_{2p} \;=\; \frac{(2\mathrm{p})!}{2^{\mathrm{p}}.p!}\ .\sigma^{2p} \quad \forall\ p$$

On retrouve évidemment pour: p = 1 , $\mu_2 = \sigma^2$.

On pourra s'en convaincre aisément en faisant le calcul et en intégrant par parties….Sinon, on en trouvera le détail par exemple dans l'ouvrage de référence de Benjamin et Cornell (1970, p.258).

On voit aussi que l'on obtient: $\mu_4 = 3\,\sigma^4$

et comme on avait vu que le coefficient d'aplatissement (*kurtosis en anglais*) s'écrit en général : $\dfrac{\mu_4}{\sigma^4}$

⇒ cela donne, pour la loi normale, un **coefficient d'aplatissement égal à 3**

c) Table de la Loi Normale

Par ailleurs, on montre (*cf. compléments ci après*) que :

- si une variable X suit une loi normale,
- toute transformation linéaire de X, soit Y= a.x+b, suit encore une loi normale.

Cela permet notamment le changement de variable linéaire :

$$X \to U = \frac{X - \mu_x}{\sigma_x}$$

qui ramène à la ***Loi Normale Standard***, où la variable U est **centrée réduite** de moyenne 0, (puisque la moyenne des u_i est nulle) et d'écart-type 1 qui est l'écart-type des u_i, loi encore notée $N(0,1)$.

Cette Loi Normale centrée réduite s'écrit:

$$F(u) = \int_{-\infty}^{u} \frac{1}{\sqrt{2\pi}} . e^{-\frac{1}{2}.t^2} .dt$$

Nous l'avons tracée graphiquement sur la page suivante. On reconnaît évidemment la fameuse allure de "courbe en cloche"!

En général, cette loi se trouve tabulée dans tous les ouvrages: nous en donnons un exemple dans la page qui suit le graphique. Cette table permet notamment de vérifier ou de retrouver des **intervalles interquantiles** remarquables, propres à la loi normale:

- Intervalle **interdécile** [10% - 90%] =
  $\pm$ **1,28** écart-type de part et d'autre de la moyenne,
  contient **80 %** des valeurs de la population

- l'Intervalle de $\pm$**1** écart-type de part et d'autre de la moyenne,
  contient **68 %** des valeurs

- l'Intervalle de $\pm$**2** écart-types de part et d'autre de la moyenne,
  contient **95 %** des valeurs

(Il est vivement recommandé d'en retenir quelques-uns, et d'apprendre à se servir de la table...)

Loi Normale Standard (moyenne μ=0 , écart-type σ=1)
Fonction de répartition et densité de probabilité

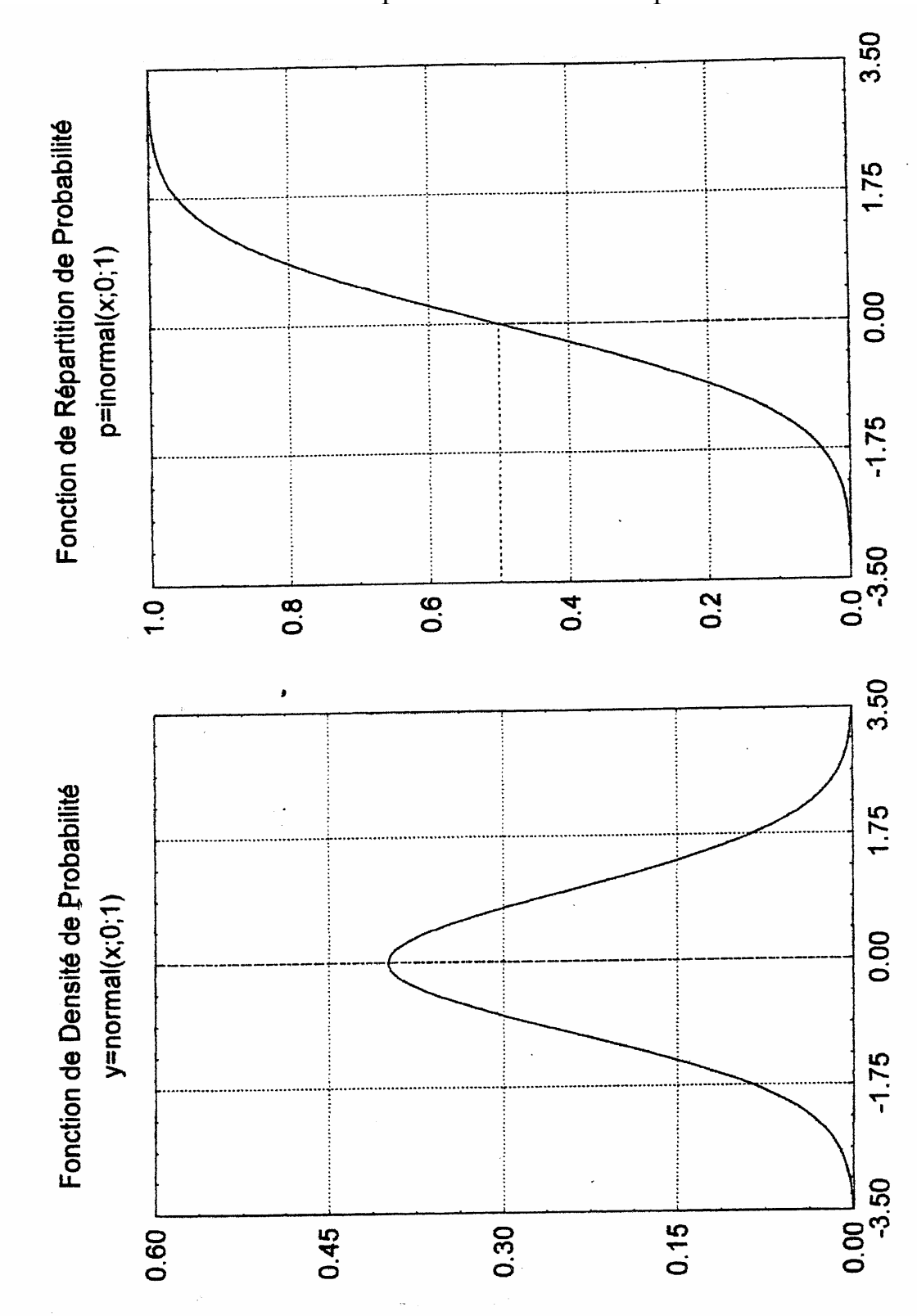
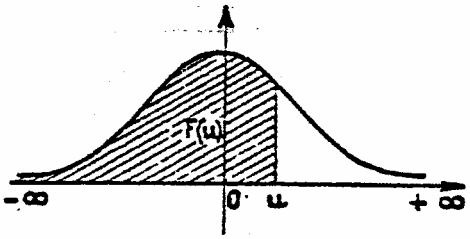


TABLE 2-1

FONCTION DE REPARTITION DE LA LOI NORMALE REDUITE

(Probabilité de trouver une valeur inférieure à u)



u	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	0,7290	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	0,9779	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

Table pour les grandes valeurs de u

u	3,0	3,1	3,2	3,3	3,4	3,5	3,6	3,8	4,0	4,5
F(u)	0,99865	0,99904	0,99931	0,99952	0,99966	0,99976	0,999841	0,999928	0,999968	0,999997

Nota – La table donne les valeurs de F(u) pour u positif. Lorsque u est négatif il faut prendre le complément à l'unité de la valeur lue dans la table.

Exemple . pour u = 1,37 F(u) = 0,9147
pour u = - 1,37 F(u) = 0,0853

Complément de démonstration(*)

Comme on l'a signalé dans le paragraphe IV du Chapitre I, il est intéressant de regarder ce que produit un changement de variable et ce que devient la loi de probabilité de la variable transformée.

Si l'on prend ici une loi normale classique que l'on écrit:

$$f(x) = \frac{1}{\sigma_x \sqrt{2\pi}} . e^{-\frac{1}{2}\left(\frac{x-\mu_x}{\sigma_x}\right)^2}$$

et si l'on applique une transformation linéaire sur x, soit : y = a .x + b = g(x)

Alors on a : $x = g^{-1}(y) = \frac{y-b}{a}$ et $g'(x) = a$

d'où en reportant dans:

$$h(u) = \frac{f\left[g^{-1}(u)\right]}{g'\left[g^{-1}(u)\right]} \quad \Rightarrow \quad h(y) = \frac{1}{\sigma_x \sqrt{2\pi}} . e^{-\frac{1}{2}\left[\frac{\frac{y-b}{a}-\mu_x}{\sigma_x}\right]^2} . \frac{1}{a} = \frac{1}{a.\sigma_x \sqrt{2\pi}} . e^{-\frac{1}{2}\left[\frac{y-b-a.\mu_x}{a.\sigma_x}\right]^2}$$

Or :

$$a.\sigma_x = \sigma_y \quad \text{et} \quad \mu_y = a.\mu_x + b \quad \Rightarrow \quad h(y) = \frac{1}{\sigma_y \sqrt{2\pi}} . e^{-\frac{1}{2}\left[\frac{y-\mu_y}{\sigma_y}\right]^2}$$

et donc la densité de probabilité de y a bien encore la forme analytique d'une loi normale.

Donc:

"si une variable x suit une loi normale,
toute transformation linéaire de x en y fournit une variable y
qui suit aussi une loi normale".

En particulier, si on fait la transformation:

$$x \quad \rightarrow \quad u = \frac{x - \mu_x}{\sigma_x} \quad \textbf{(standardisation)}$$

alors la loi de u devient, avec:

$$\sigma_u = \frac{1}{\sigma_x} . \sigma_x = 1 \quad \text{et} \quad \mu_x = \frac{1}{\sigma_x} . \mu_u + \mu_x \quad \Rightarrow \mu_u = 0 \ , \quad h(u) = \frac{1}{\sqrt{2\pi}} . e^{-\frac{1}{2}.u^2}$$

c'est à dire la loi standard de moyenne 0 et d'écart-type 1 encore notée $N(0,1)$.

Ce résultat va être utilisé pour construire un ***papier fonctionnel***.

d) Diagramme Gausso-arithmétique ou **"Papier de Gauss" :**

L'utilisation d'un papier "dit de Gauss" est très simple et nous la verrons plus loin. Mais nous donnons d'abord une idée de la:

Construction du Papier de Gauss(*):

Elle va comporter 3 étapes (cf. figure page ci-contre)

1) nous traçons d'abord sur un premier diagramme, à échelles arithmétiques, la fonction normale standard en fonction de u.
$\Rightarrow$ A chaque valeur de u_i correspond une probabilité au non dépassement P_i que nous trouvons dans la table, ou inversement à chaque valeur Pj correspond une valeur u_j.

2) si nous considérons une autre variable **normale** X. Comme elle est normale, elle peut s'obtenir par transformation linéaire de U, donc elle est en relation linéaire avec U.

$\Rightarrow$ Donc dans un second diagramme à échelles arithmétiques en U et X, les valeurs u_j et x_j **correspondant à la même probabilité Pj** sont en relation linéaire, et donc alignées selon une droite. La position de cette droite dépendra évidemment des coefficients de la transformation linéaire.

Donc pour une valeur x_j dont on connaît la probabilité $Pr(X \leq x_j) = P_j$:

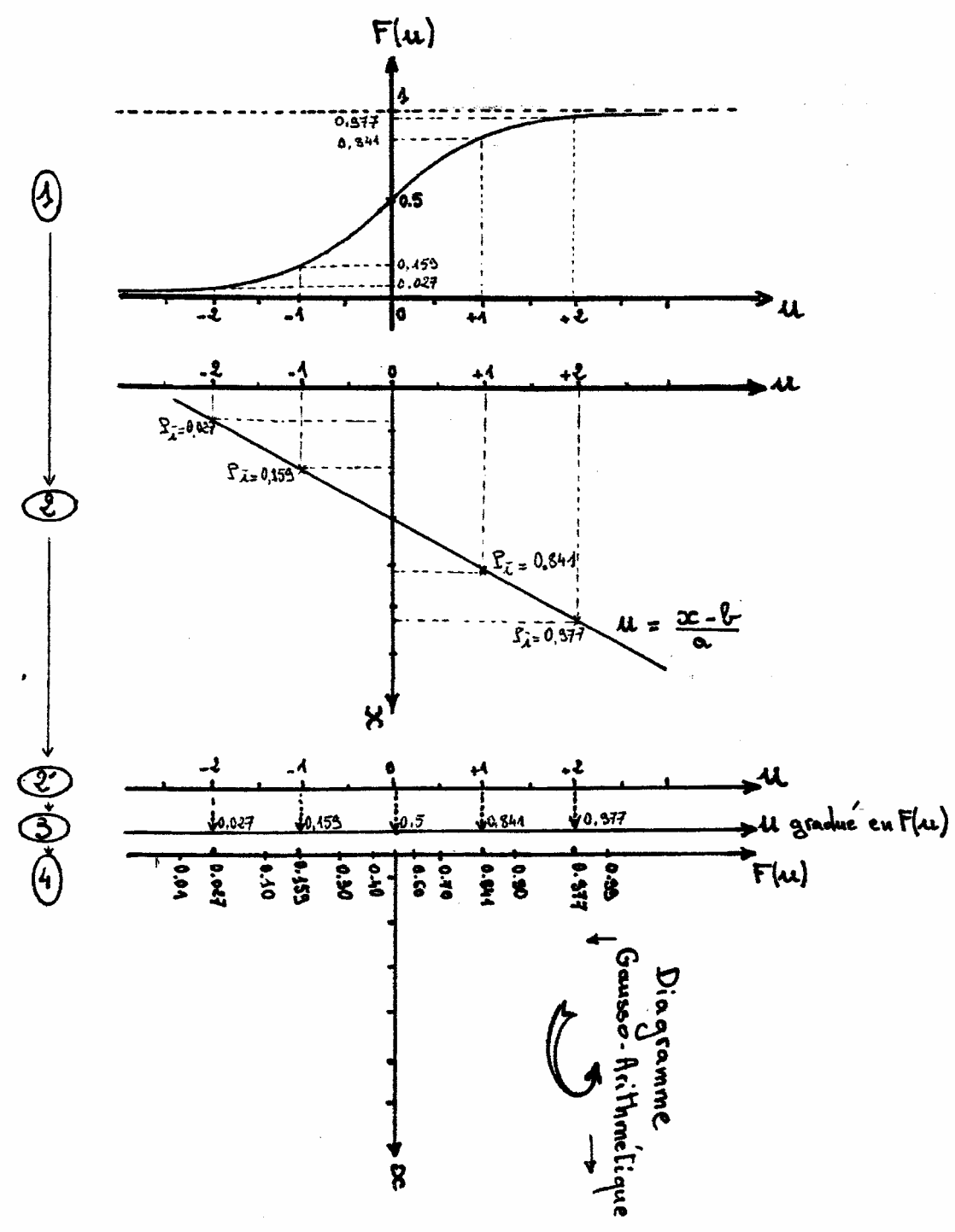
- on porte d'abord x_j sur l'axe des x,
- puis on regarde sur le 1er diagramme la valeur de u_j telle que $F(u_j) = Pj$,
- et on porte cette valeur u_j en ordonnée.

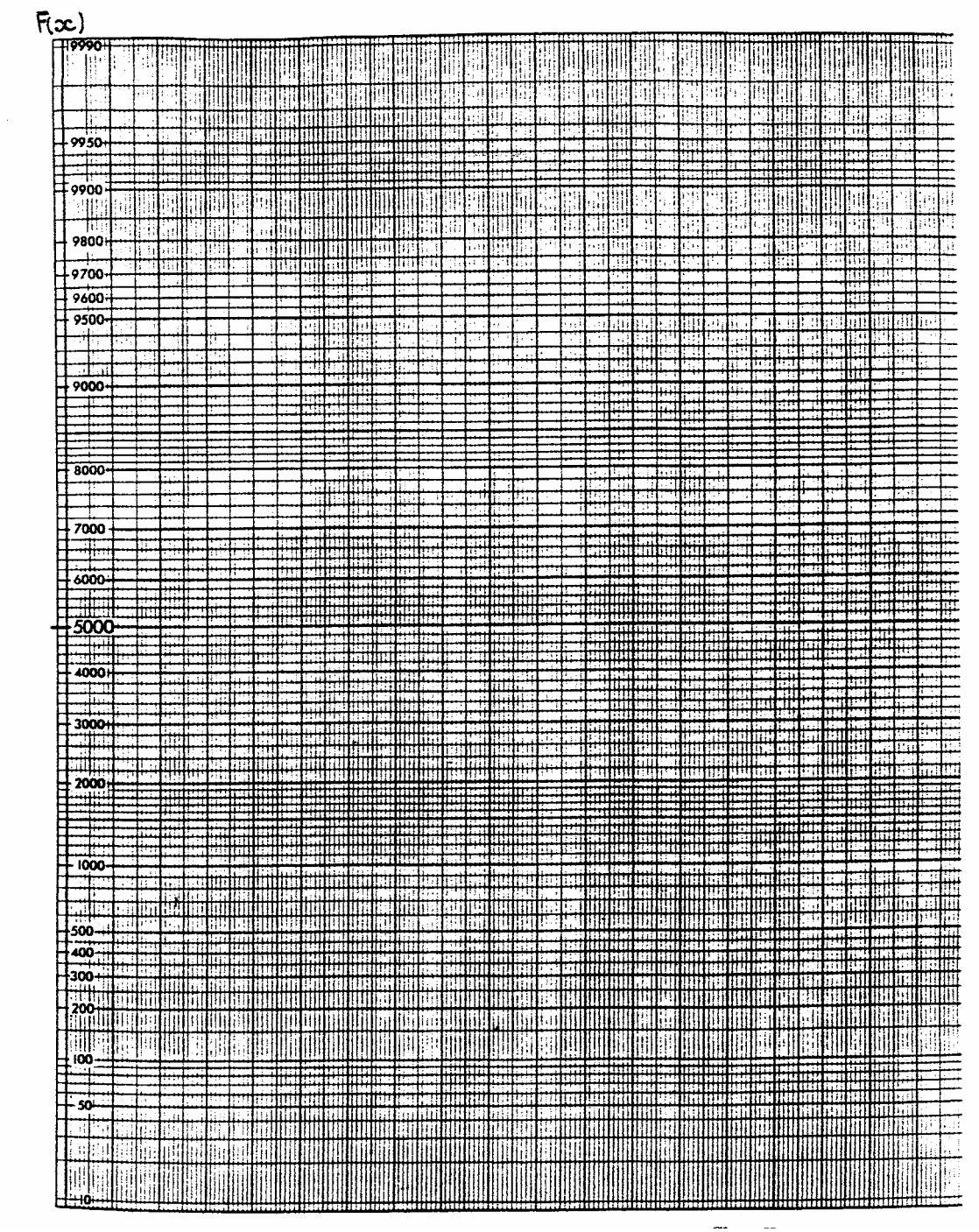
$\Rightarrow$ et les points (x_j, u_j)doivent être alignés.

3) Mais en fait, il est inutile d'aller à chaque fois recalculer les u_j à partir de F(u)! Il suffit de porter directement sur l'axe, en même temps, et même carrément *à la place* des u_j la valeur $F(u_j)$.

4) En pratique on utilisera donc le seul diagramme inférieur (cf. page suivante), avec la seule graduation F(u) en ordonnées .

Schéma de construction du diagramme Gausso-arithmétique.



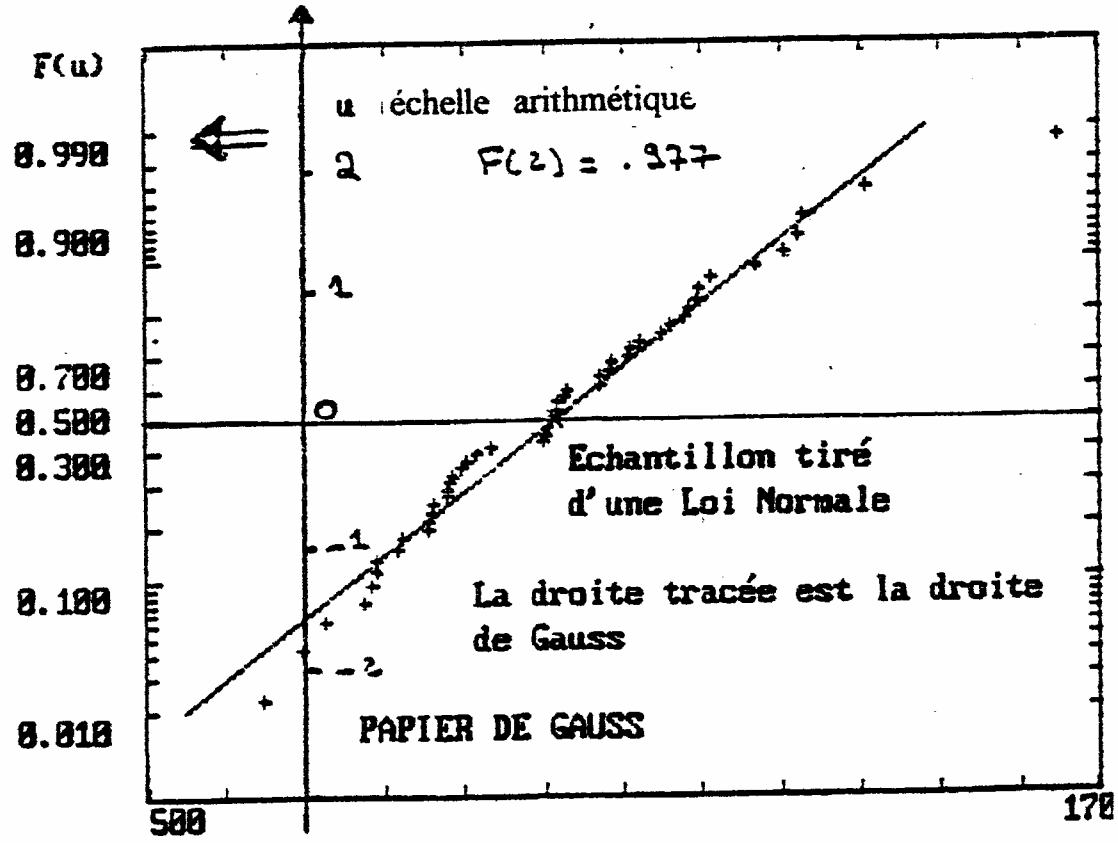


Papier de Gauss

Utilisation du Papier de Gauss:

On classe donc l'échantillon par valeurs croissantes x_j de rang i.
A chaque valeur on associe une probabilité au non-dépassement P^*_j
(estimée empiriquement)
On porte sur le diagramme les points $[x_j , F(u_j) = P^*_j]$

Si cette fonction de répartition empirique est proche d'une droite sur de diagramme, alors on peut considérer que l'échantillon est tiré d'une loi Normale.



Si la fonction de répartition empirique de l'échantillon est représentée par une courbe assez voisine d'une droite sur ce papier, cela signifie aussi qu'une loi de Gauss le décrit assez bien en termes de probabilités.

En outre, ce papier dilate les probabilités vers les extrêmes ce qui peut être parfois intéressant. On l'utilisera donc comme support de tracé *même* dans des cas où l'on ne s'attend pas à un comportement gaussien.

Ces diagrammes sont en vente dans (presque toutes…) les bonnes papeteries.

II-2) LOI **LOGNORMALE** (dite également LOI de GALTON):

Une façon courante d'enrichir la boîte à outils consiste:
à faire une *transformation simple* sur la variable aléatoire X,
$\Rightarrow$ soit Y = g(X),
et à voir si la nouvelle variable Y ne serait pas normale..?

On tente couramment la racine carrée $Y = \sqrt{X}$ (- dans ce cas on construit l'échantillon des valeurs en racine carrée-), ou dans ce qui suit, le logarithme.

On distinguera :

a) loi lognormale à **2** paramètres:

où si X> 0 Y = Log X et $$h(y,\alpha,\beta) = \frac{1}{\alpha.\sqrt{2\pi}}.e^{-\frac{1}{2}\left(\frac{y-\beta}{\alpha}\right)^2} \qquad (1)$$

Intérêt de cette loi :

Comme pour la loi de Gauss, on démontre que, sous certaines restrictions:
- *si* le phénomène X est le ***produit*** de k variables aléatoires **indépendantes**,
- *alors*, si k tend vers l'infini, X suit une loi Lognormale.

Dans la nature, on peut citer le cas:
- de la granulométrie des sédiments, qui résultent de chocs indépendants qui enlèvent chacun un pourcentage ($\Rightarrow$ multiplicatif) aléatoire du grain,
- de phénomènes de fatigue où l'effet est *proportionnel* à l'état déjà atteint (cf. Benjamin et Cornell 1970)
- de certains débits (par exemple mensuels) qui sont en première approche le produit de la pluie par des coefficients d'écoulement aléatoires…, etc...

b) loi lognormale à **3** paramètres:

où Y = Log (X - $\mathbf{x_0}$) et $$h(y,a,b) = \frac{1}{a\sqrt{2\pi}}.e^{-\frac{1}{2}\left(\frac{y-b}{a}\right)^2} \qquad (2)$$

$\Rightarrow$ incluant un **troisième paramètre $\mathbf{x_0}$** ,
qui sera optimisé pour rendre la variable transformée la plus gaussienne possible.

On montre ainsi (cf. M. Roche, 1963) que l'on peut choisir x_0 de manière à ce que le coefficient d'asymétrie Cs de Y soit nul (condition nécessaire pour que la loi de Y soit normale), ce qui entraîne que x_0 devient solution de:

$$\frac{(\mu_x - x_0)^3}{\sigma_x{}^2 + 3.(\mu_x - x_0)^2} = \frac{\sigma_x{}^4}{\mu_{3x}} \qquad (3)$$

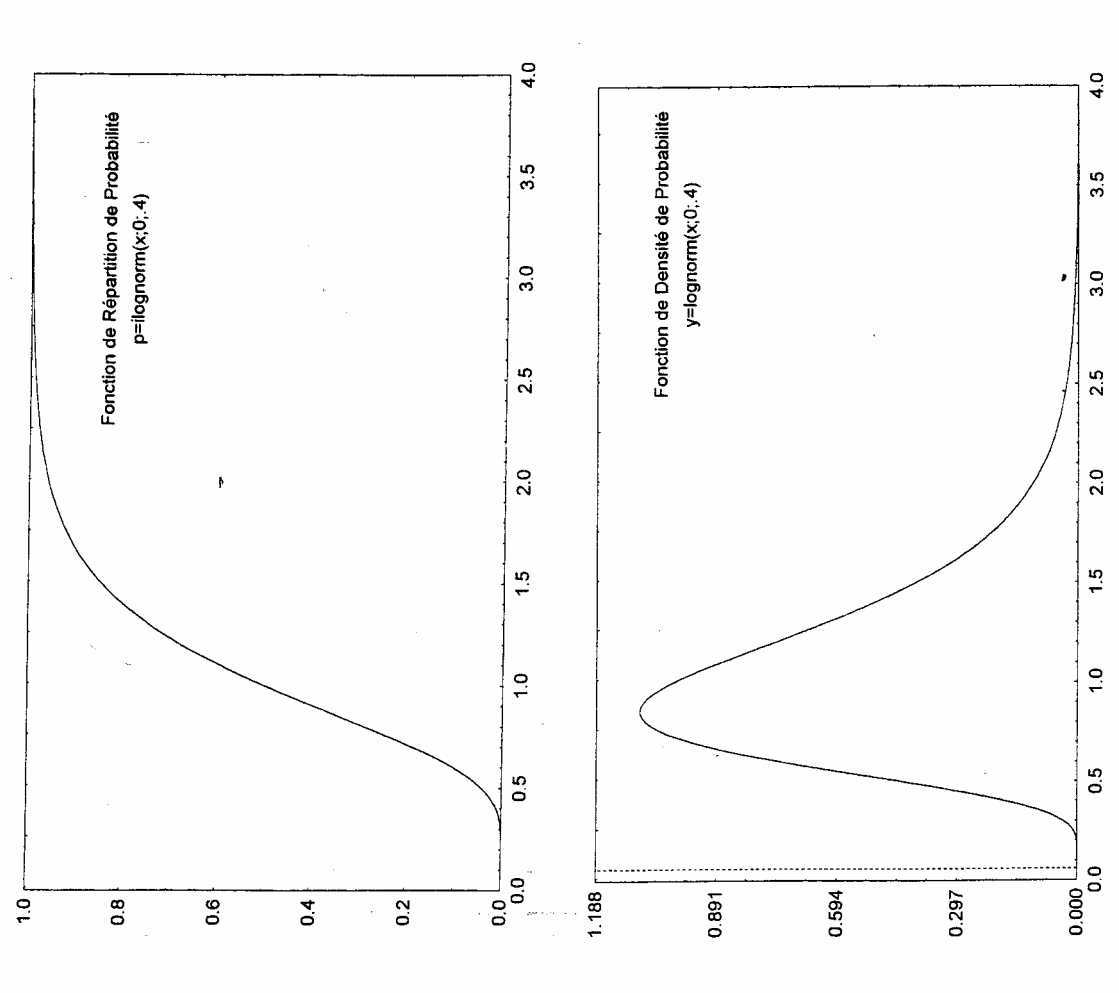
avec μ_{3x} = moment centré d'ordre 3 = $E[x^3] - 3\mu_x.\sigma_x{}^2 - \mu_x{}^3$

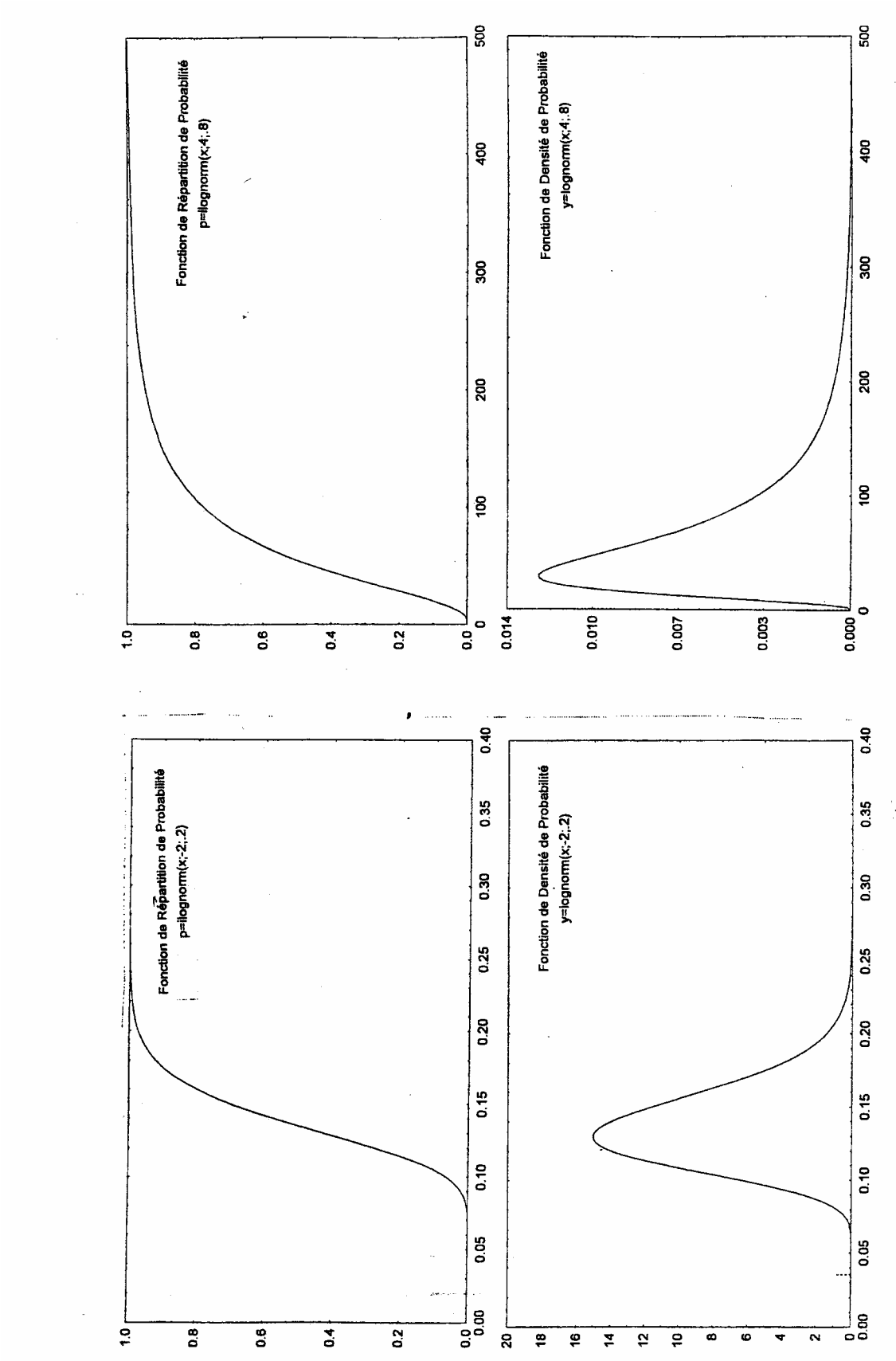
L'estimation de μ_{3X} se fait sur l'échantillon par m_{3X} (vue au chap. I)

$$m_{3x} = \frac{1}{(n-1).(n-2)}.\left[n.\sum_{i=1}^{n} x_i^3 - 3.\sum_{i=1}^{n} x_i.\sum_{i=1}^{n} x_i^{\,2} + \frac{2}{n}\left(\sum_{i=1}^{n} x_i\right)^3 \right] \quad (4)$$

La forme de la densité Lognormale est intéressante, puisqu'elle démarre à l'origine, (resp. en x_0) avec une tangente variable selon les paramètres et qu'elle est dissymétrique (la moyenne est plus grande que la médiane).

On donne ci-après quelques exemples pour différentes valeurs des paramètres α et β, qui accentuent plus ou moins la dissymétrie.




Fonction de Répartition de Probabilité
p=ilognorm(x;4;.8)
Fonction de Densité de Probabilité
y=ilognorm(x;4;.8)
Fonction de Répartition de Probabilité
p=ilognorm(x;-2;.2)
Fonction de Densité de Probabilité
y=ilognorm(x;-2;.2)

Compléments théoriques (*)**:** sur la loi Lognormale

Un premier résultat consiste à noter que:

- quand X suit une loi Lognormale, on peut montrer alors que
- toute transformation puissance, donc de la forme $Z = a X^b$ suit aussi une loi Lognormale.

Un autre ensemble de résultats provient de ce que l'on sait exprimer:

- la forme analytique de la loi de probabilité de la variable transformée
- $y = Log(x-x_0)$, puisque c'est la loi normale classique,
- mais aussi celle de x, qui est plus compliquée.

Pour l'obtenir, on peut appliquer ici aussi les résultats sur le changement de variables (*cf. chap. I , p. 24-25*):

Soit $x = g(y) = e^y + x_0$ et $y = g^{-1}(x) = Log(x - x_0)$

de même:

$$g'(y) = e^y \quad \text{et donc} \quad g'\left[g^{-1}(x)\right] = e^{Log(x-x_0)} = x - x_0$$

d'où:

$$h(x) = \frac{f\left[g^{-1}(x)\right]}{g'\left[g^{-1}(x)\right]} = \frac{1}{x - x_0} \cdot \frac{1}{\sigma_y \sqrt{2\pi}} e^{-\frac{1}{2}\left\{\frac{Log(x-x_0)-\mu_y}{\sigma_y}\right\}^2}$$

On comprend donc que l'on n'utilise pas couramment cette expression...!

Elle permet pourtant des développements intéressants (cf. Yevjevich 1972 ou Benjamin & Cornell 1970 pour plus de détails) , comme les relations entre les moments de la variable brute X et ceux de la variable transformée $Y = Log(X-x_0)$.

On montre notamment que:

- si μ_X^k est le moment d'ordre k (*non centré*) de la variable brute X,

- alors tous ces moments s'expriment en fonction des seuls deux premiers moments de la variable transformée Y, soit μ_Y et σ_Y par :

$$\mu_X^k = e^{k.\mu_Y + k^2.\frac{\sigma_Y^2}{2}} \qquad (5)$$

On en déduit d'ailleurs que la moyenne vérifie: $$\mu_X = e^{\mu_Y + \frac{\sigma_Y^2}{2}} \qquad (6)$$

ou respectivement : $\mu_X - x_0 = e^{\mu_Y + \frac{\sigma_Y^2}{2}}$ quand Y = Log (X-x_0)

et l'écart-type (en combinant les formules (5) et (6) ci-dessus):

$$\sigma_x^2 = E[X^2] - (E[X])^2 = \mu_x^{k=2} - (\mu_x^{k=1})^2$$

$$\sigma_X^2 = \mu_X^2 . (e^{\sigma_Y^2} - 1) \qquad (7)$$

En fait, on utilise plutôt ces formules dans l'ordre inverse, pour exprimer les paramètres de la loi f(y) en fonction des moments de X, soit:

$$\mu_Y = Log\left(\frac{\mu_X^2}{\sqrt{\mu_X^2 + \sigma_X^2}}\right) \quad (8) \qquad \text{et} \qquad \sigma_Y^2 = Log\left(1 + \frac{\sigma_X^2}{\mu_X^2}\right) \quad (9)$$

Ceci pourra être utilisé, mais avec circonspection...!, dans la méthode d'ajustement par les moments (cf. Chap. III, p. 6-9)

c) <u>diagramme lognormal:</u>

Par contre, on comprend facilement comment adapter le diagramme gausso-arithmétique à cette nouvelle variable :

- pour tester si le Log de X est gaussien,
- il suffit de remplacer l'échelle arithmétique des abscisses par une **échelle logarithmique**,
- et de porter les valeurs naturelles de x sur cette échelle.

Au besoin, si les points ne sont pas alignés, on retranchera par tâtonnement une quantité x_0 pour tenter d'améliorer l'alignement. On pourra aussi utiliser les relations (3) et (4) de ce chapitre sur la loi Lognormale pour estimer x_0.

On en verra plus en détail l'utilisation au chap. III.

<u>II-3)</u> APERCU SUR D'AUTRES LOIS DERIVEES: (de la Loi Normale)

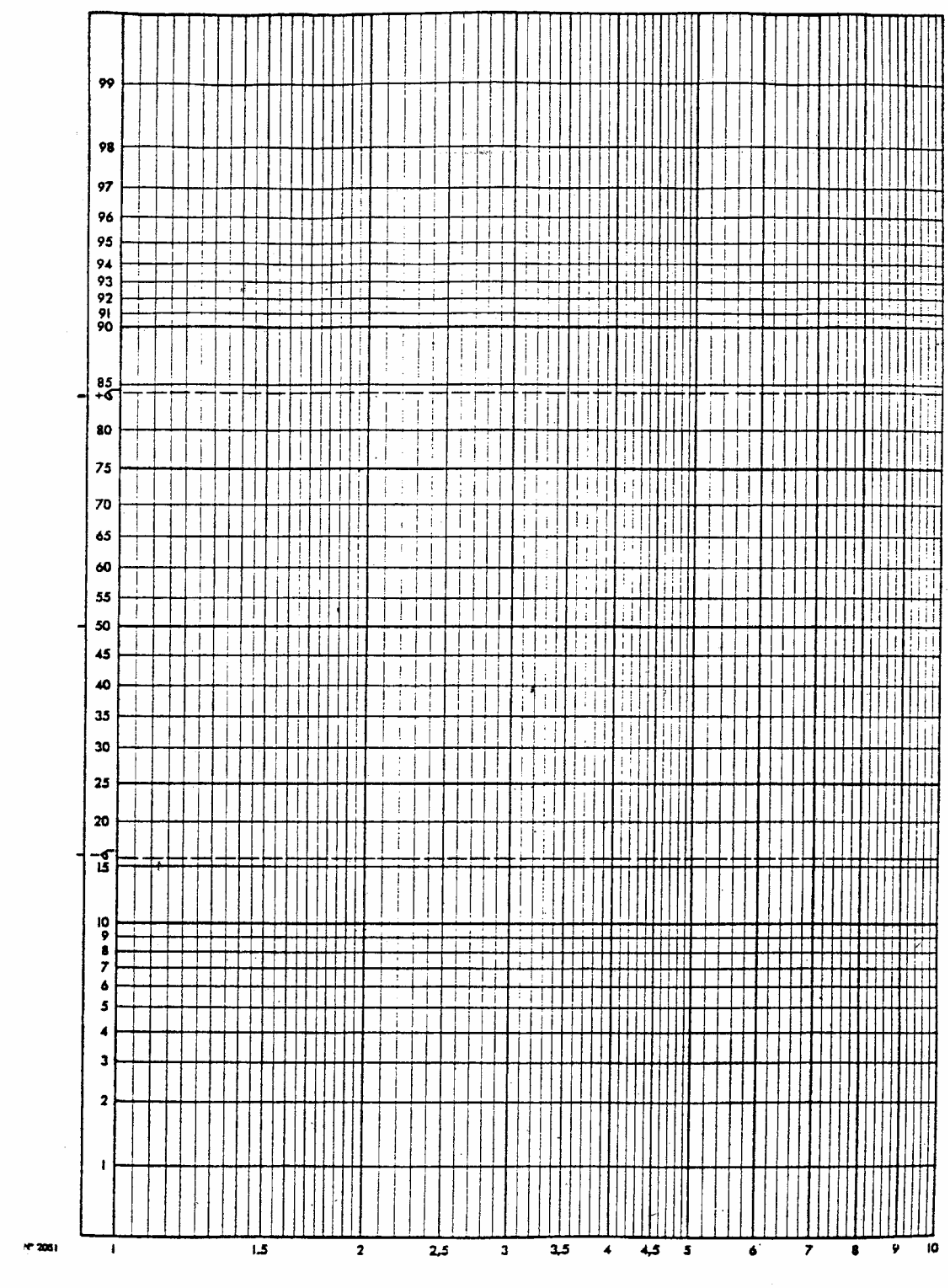
Un autre exemple d'extension de la loi normale que nous nous contenterons d'évoquer est celui ou la ***Racine Carrée*** de X suit une loi normale:

Cet exemple est intéressant car la contrainte X > 0 entraîne aussi Y > 0 et donc on ne doit considérer que la partie de la loi où les valeurs de Y sont > 0.

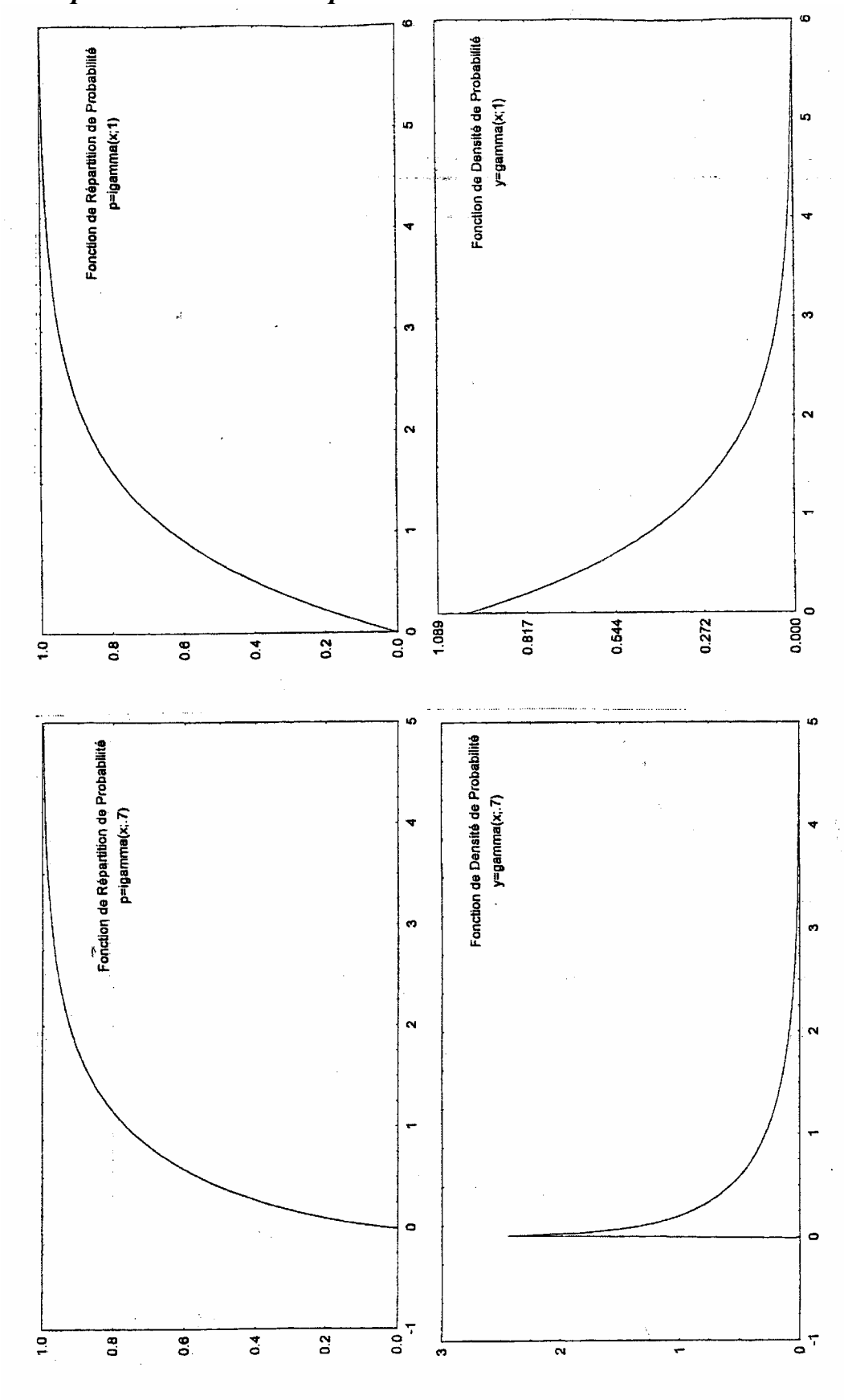
Il s'agit alors d'une loi normale **tronquée**, comme on en verra une plus loin pour la loi exponentielle. Cette loi est parfois préconisée pour les valeurs de pluies mensuelles non nulles.

On en trouvera les propriétés dans Lubès et al. (1994).

(diagramme lognormal)



Exemples de Lois Gamma pour $\lambda<1$ ou $=1$:



III) LOIS GAMMA et DERIVEES :

III-1) Loi GAMMA à 2 paramètres:

Cette loi est à 2 paramètres, ρ et λ.

Elle est définie pour une variable continue $\mathbf{x \geq 0}$ positive ou nulle.

Son intérêt majeur est une *grande flexibilité de forme*, qui en fait un outil susceptible de s'adapter à des histogrammes très variés. On verra qu'elle peut même entrer en compétition avec les lois normales et lognormales.

L'un des paramètres (ρ) a la dimension de x (paramètre d'**échelle**),
l'autre est adimensionnel (λ = paramètre de **forme**).

Sa densité est définie par:

$$f(x,\lambda,\rho) = \frac{1}{\Gamma(\lambda)}.e^{-\frac{x}{\rho}}.(\frac{x}{\rho})^{\lambda-1}.\frac{1}{\rho}$$

avec Γ(λ) la fonction spéciale dite fonction Gamma qui:

- pour λ entier vaut Γ(λ) = (λ-1)! (c'est à dire le factoriel de λ)
  Par convention , pour λ =1 Γ(1) = 0 et Γ(0) = 0 , Γ(1/2) = $\sqrt{\pi}$
- et, pour λ non entier, elle est définie par : $\Gamma(\lambda) = \int_0^{+\infty} z^{\lambda-1}.e^{-z}.dz$

Exemples de formes:

Paramètre de forme **λ<1** : (*cf. page ci-contre*)

C'est le cas éventuel des pluies journalières. La forme est quasiment hyperbolique, mais part d'une ordonnée finie fonction de λ (et de ρ, mais on prendra ce paramètre d'échelle égal à 1 ici). On peut par exemple faire l'étude pour λ= 0.5

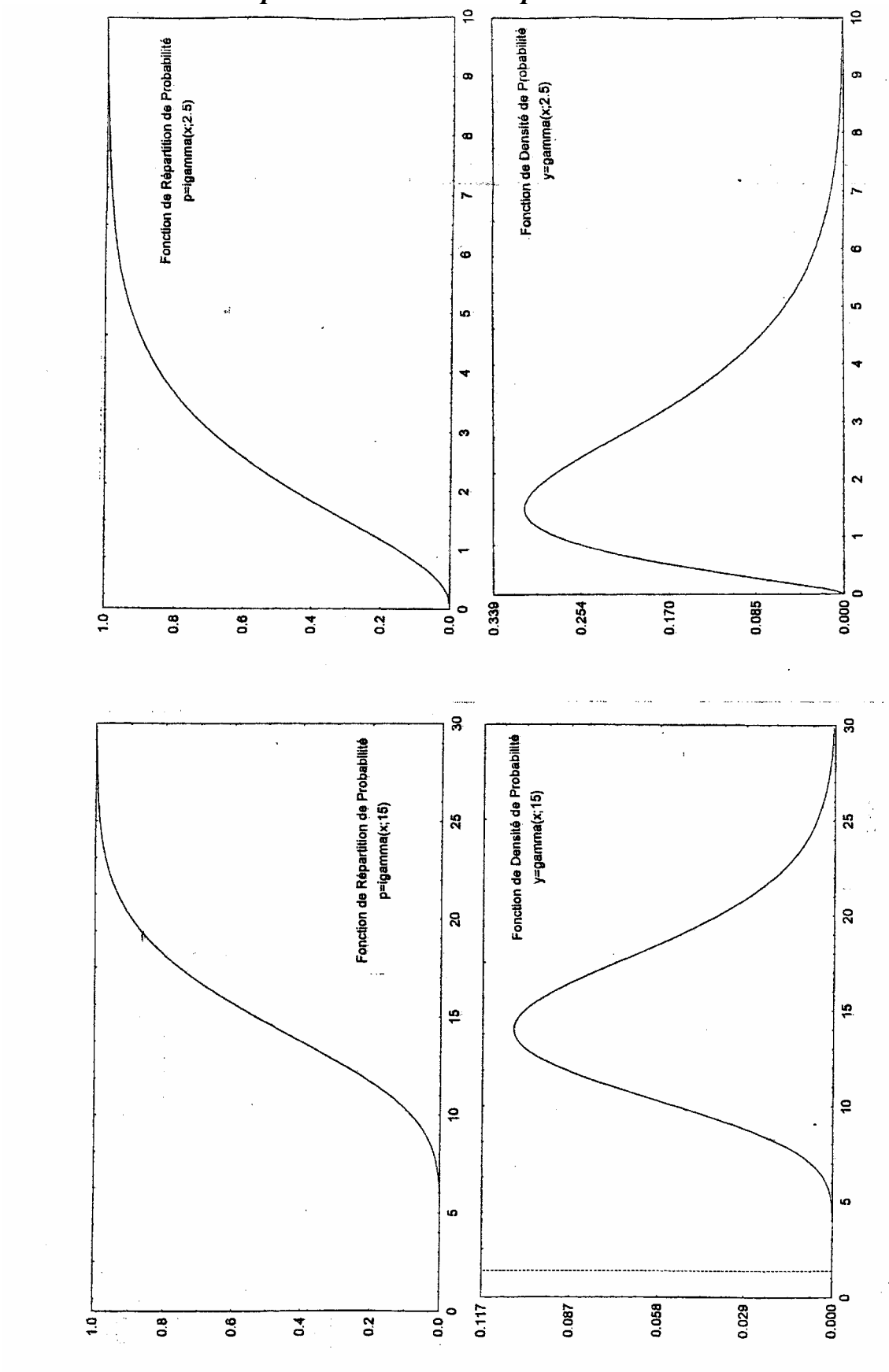
Paramètre de forme **λ=1** : (cf. page ci-contre)

On trouve alors comme cas particulier la loi **exponentielle:**

$$f(x,1,\rho) = f(x,\rho) = \frac{1}{\rho}.e^{-\frac{x}{\rho}}$$

(cas souvent utilisé aussi pour les pluies à court pas de temps - jusqu'à 24 heures)
Cette dernière loi sera étudiée plus en détail au paragraphe suivant.

Exemples de Lois Gamma pour λ>1 et λ>20:



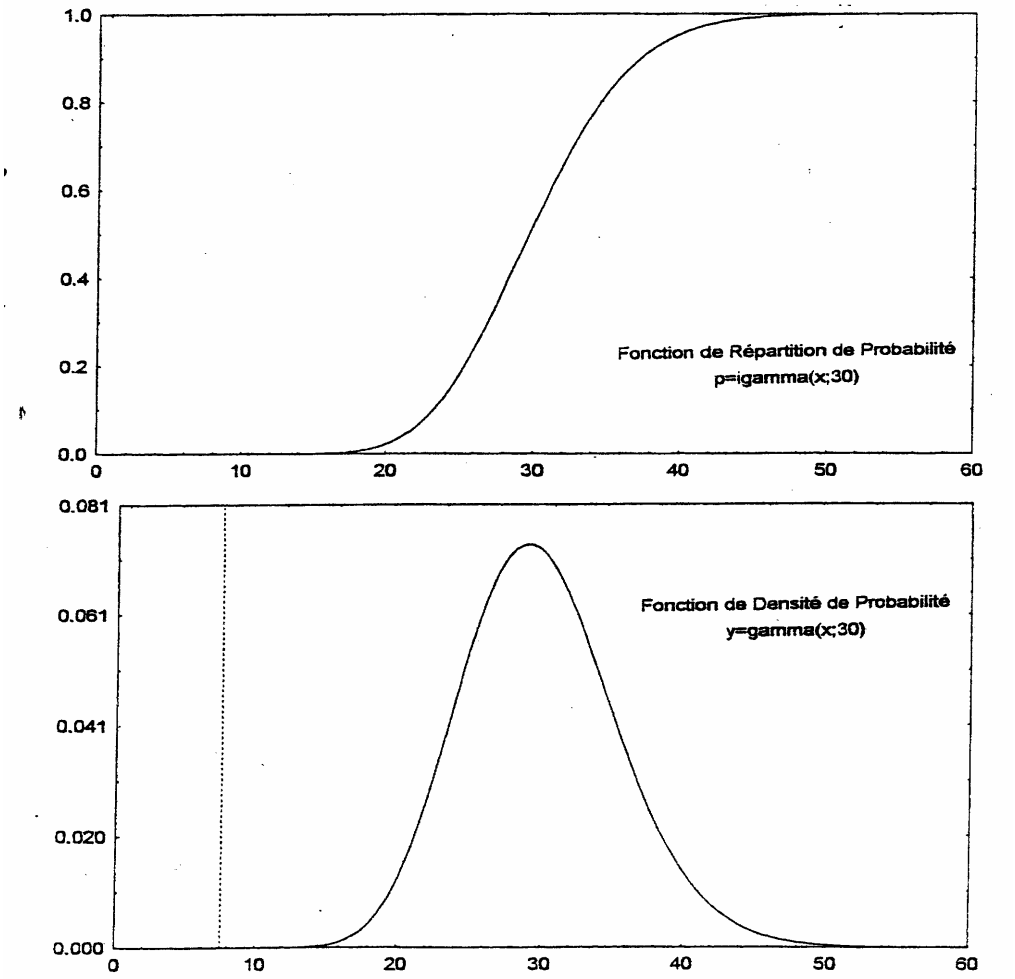
Paramètre de forme λ**>1** : (cf. page ci-contre)

La forme devient en cloche dissymétrique, proche d'une loi lognormale (l'oeil aura du mal à les distinguer)
C'est souvent le cas de pluies mensuelles (non nulles), dont on a vu aussi que la racine carrée pouvait être normale, ce qui montre la difficulté à choisir entre plusieurs représentations..!

Paramètre de forme λ**> 20** : (cf. page ci-dessous)

Enfin si le paramètre de forme λ est grand (supérieur à 20), on retrouve quasiment la forme d'une *loi Normale* dans les valeurs centrales (intervalle interdécile). Par exemple, pour $\lambda = 15$, on constate encore une légère dissymétrie, que l'on ne détecte même plus pour $\lambda=30$. Pour un tel histogramme, on ne peut donc pas dire à l'oeil si sa forme analytique est la fonction erreur (Gauss) ou la loi Gamma... ! bien que les expressions analytiques soient très différentes….

Exemple de Loi Gamma pour $\lambda=30$



III-2) CALCUL des MOMENTS (en fonction des paramètres)

Par intégration (- les amateurs éclairés peuvent le faire à titre d'exercice, c'est relativement aisé ..-), on trouve les relations suivantes:

$$E[X] = \mu_x = \lambda.\rho \qquad V[X] = \mu_{2x} = \sigma_x^2 = \lambda.\rho^2$$

qui permettent immédiatement d'en déduire les paramètres en fonction des deux premiers moments:

$$\rho = \frac{\sigma^2}{\mu} \quad \text{et} \quad \lambda = \frac{\mu^2}{\sigma_x^2} = \frac{1}{CV^2}$$

où CV est le coefficient de variation défini au chapitre I

Le paramètre de forme **λ est l'inverse du carré du *coefficient de variation*:** Il est donc d'autant plus grand que la fluctuation de X est petite par rapport à sa moyenne

Le paramètre d'échelle ρ est d'autant plus grand que la fluctuation est grande par rapport à la moyenne; ce paramètre a la dimension de la variable.

On montrerait de même que les moments suivants: $\mu_{3x} = 2.\lambda.\rho^3$

et : $\mu_{4x} = 3.\lambda.(\lambda+2)\rho^4$

La dissymétrie, qui s'exprime par le **coefficient d'asymétrie**:

$$G = \frac{\mu_{3x}}{\sigma^3} = \frac{2.\lambda.\rho^3}{(\sqrt{\lambda}.\rho)^3} = \frac{2}{\sqrt{\lambda}}$$

est donc d'autant plus faible que λ est grand (d'où le fait que pour λ >30 on retombe sur une loi symétrique quasi normale)

Bien entendu, il est possible d'utiliser un **3ème paramètre x_0**

$$f(x,\lambda,\rho,x_0) = \frac{1}{\Gamma(\lambda)}.e^{-\frac{x-x_0}{\rho}}.(\frac{x-x_0}{\rho})^{\lambda-1}.\frac{1}{\rho}$$

permettant la prise en compte d'une origine non nulle ou l'optimisation du choix de l'origine pour maximiser l'adéquation à une fonction Gamma.

III-3) TABLES de la LOI GAMMA (en fonction des paramètres)

La loi Gamma incomplète a 2 paramètres, et il n'est pas possible de trouver une expression plus simple (par exemple par un changement de variable); d'où des tables donnant pour diverses valeurs du paramètre de forme les valeurs de la fonction de répartition en fonction bien souvent de la variable réduite (mais non centrée), c'est à dire de la variable divisée par son écart type, ce qui permet d'éliminer le problème de dimension.

On donne en annexe ci-contre un exemple de table ; il faut bien évidemment *interpoler* pour les utiliser quand la valeur estimée de λ ne figure pas dans la table.

(Table de la loi gamma)

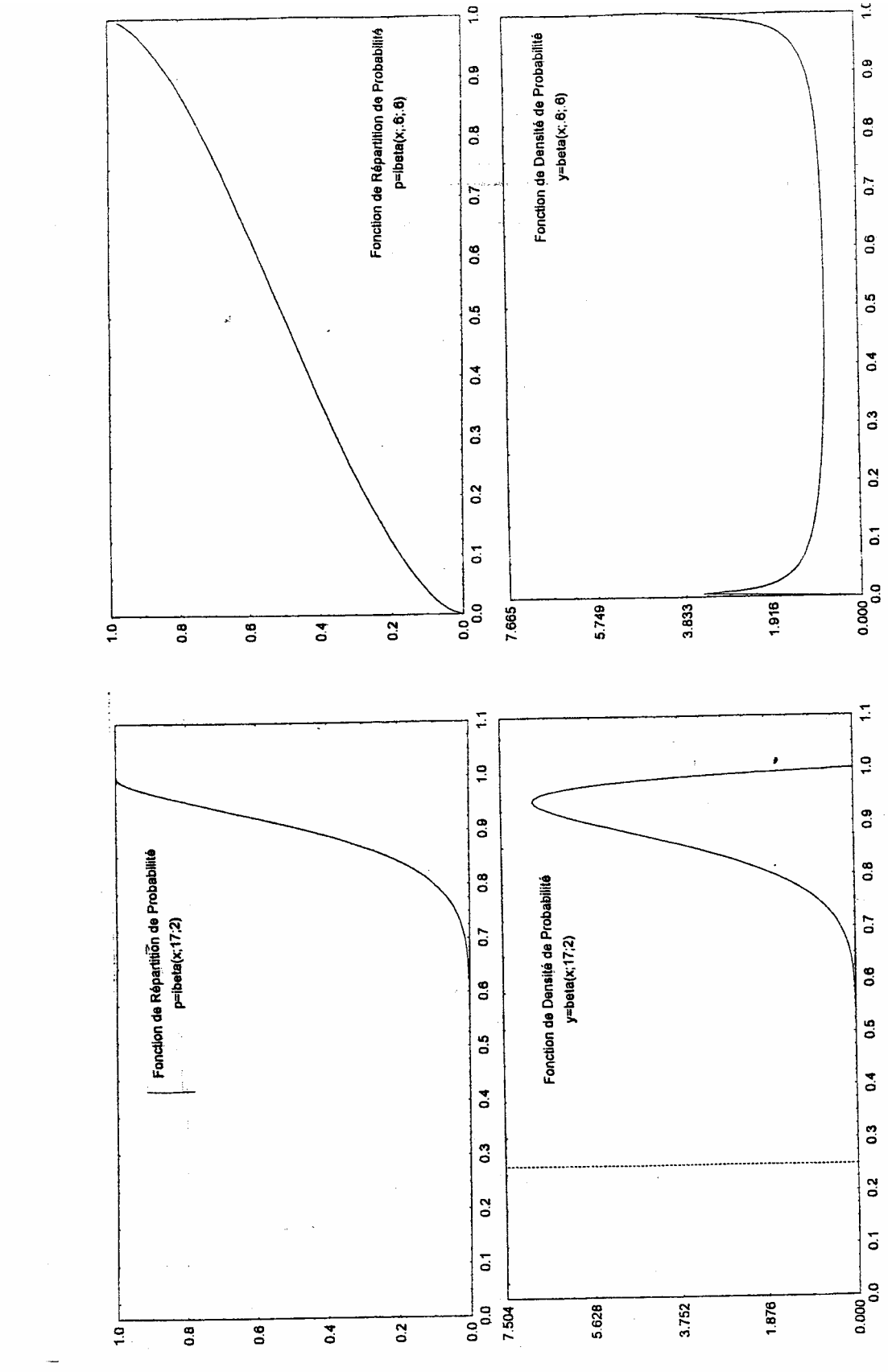
Lamda→	0.50	0.60	0.70	0.80	0.90	1.00	1.50	2.00	2.50	3.00	3.50	4.00	4.50	5.00	6.00	7.00	8.00	9.00	10.00	11.00	12.00	13.00	14.00	15.00
F(u) : ▼					Valeurs de u :																			
0.001	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.03	0.07	0.11	0.16	0.21	0.27	0.33	0.45	0.57	0.70	0.82	0.94	1.05	1.17	1.28	1.39	1.50
0.005	0.00	0.00	0.00	0.00	0.00	0.01	0.03	0.07	0.13	0.20	0.26	0.34	0.41	0.48	0.63	0.77	0.91	1.04	1.18	1.30	1.43	1.55	1.67	1.78
0.010	0.00	0.00	0.00	0.00	0.01	0.01	0.05	0.11	0.18	0.25	0.33	0.41	0.49	0.57	0.73	0.88	1.03	1.17	1.31	1.44	1.57	1.69	1.81	1.93
0.025	0.00	0.00	0.01	0.01	0.02	0.03	0.09	0.17	0.26	0.36	0.45	0.54	0.64	0.73	0.90	1.06	1.22	1.37	1.52	1.66	1.79	1.92	2.05	2.17
0.050	0.00	0.01	0.01	0.02	0.04	0.05	0.14	0.25	0.36	0.47	0.58	0.68	0.78	0.88	1.07	1.24	1.41	1.57	1.72	1.86	2.00	2.13	2.26	2.39
0.075	0.01	0.01	0.03	0.04	0.06	0.08	0.19	0.32	0.44	0.56	0.68	0.79	0.89	0.99	1.19	1.37	1.54	1.70	1.85	2.00	2.14	2.28	2.41	2.54
0.100	0.01	0.02	0.04	0.06	0.08	0.11	0.24	0.38	0.51	0.64	0.76	0.87	0.98	1.09	1.29	1.47	1.65	1.81	1.97	2.12	2.26	2.40	2.53	2.66
0.150	0.03	0.05	0.07	0.10	0.13	0.16	0.33	0.48	0.63	0.77	0.90	1.02	1.14	1.25	1.45	1.64	1.82	1.99	2.15	2.30	2.45	2.59	2.72	2.85
0.200	0.05	0.08	0.11	0.15	0.18	0.22	0.41	0.58	0.74	0.89	1.02	1.15	1.27	1.38	1.59	1.79	1.97	2.14	2.31	2.46	2.61	2.75	2.88	3.02
0.250	0.07	0.11	0.16	0.20	0.24	0.29	0.50	0.68	0.85	1.00	1.14	1.27	1.39	1.51	1.72	1.92	2.11	2.28	2.44	2.60	2.75	2.89	3.03	3.16
0.300	0.10	0.15	0.21	0.26	0.31	0.36	0.58	0.78	0.95	1.10	1.25	1.38	1.51	1.62	1.84	2.05	2.23	2.41	2.57	2.73	2.88	3.02	3.16	3.29
0.350	0.15	0.21	0.26	0.32	0.38	0.43	0.67	0.87	1.05	1.21	1.36	1.49	1.62	1.74	1.96	2.16	2.35	2.53	2.70	2.85	3.00	3.15	3.29	3.42
0.400	0.19	0.26	0.33	0.39	0.45	0.51	0.76	0.97	1.16	1.32	1.47	1.61	1.73	1.85	2.08	2.28	2.47	2.65	2.82	2.97	3.13	3.27	3.41	3.54
0.450	0.25	0.33	0.40	0.47	0.54	0.60	0.86	1.08	1.26	1.43	1.58	1.72	1.85	1.97	2.20	2.40	2.59	2.77	2.94	3.09	3.25	3.39	3.53	3.66
0.500	0.32	0.41	0.49	0.56	0.63	0.69	0.97	1.19	1.38	1.54	1.70	1.84	1.97	2.09	2.31	2.52	2.71	2.89	3.06	3.22	3.37	3.51	3.65	3.79
0.550	0.40	0.50	0.58	0.66	0.73	0.80	1.08	1.30	1.50	1.66	1.82	1.96	2.09	2.21	2.44	2.65	2.84	3.01	3.18	3.34	3.49	3.64	3.78	3.91
0.600	0.50	0.60	0.69	0.77	0.85	0.92	1.20	1.43	1.62	1.79	1.95	2.09	2.22	2.34	2.57	2.78	2.97	3.14	3.31	3.47	3.62	3.77	3.91	4.04
0.650	0.62	0.72	0.82	0.90	0.98	1.05	1.34	1.57	1.76	1.93	2.09	2.23	2.36	2.48	2.71	2.91	3.11	3.28	3.45	3.61	3.76	3.91	4.05	4.18
0.700	0.76	0.87	0.97	1.05	1.13	1.20	1.50	1.72	1.92	2.09	2.24	2.38	2.51	2.63	2.86	3.07	3.26	3.43	3.60	3.76	3.91	4.06	4.19	4.33
0.750	0.94	1.05	1.15	1.24	1.31	1.39	1.68	1.90	2.10	2.26	2.42	2.55	2.68	2.81	3.03	3.23	3.42	3.60	3.77	3.93	4.08	4.22	4.36	4.49
0.800	1.16	1.28	1.38	1.46	1.54	1.61	1.89	2.12	2.31	2.47	2.62	2.76	2.89	3.01	3.23	3.43	3.62	3.79	3.96	4.12	4.27	4.41	4.55	4.68
0.850	1.47	1.58	1.67	1.76	1.83	1.90	2.17	2.38	2.57	2.73	2.87	3.01	3.13	3.25	3.47	3.67	3.85	4.03	4.19	4.35	4.49	4.64	4.77	4.90
0.900	1.91	2.01	2.10	2.17	2.24	2.30	2.55	2.75	2.92	3.07	3.21	3.34	3.46	3.57	3.79	3.98	4.16	4.33	4.49	4.65	4.79	4.93	5.07	5.20
0.925	2.24	2.33	2.41	2.48	2.54	2.59	2.82	3.00	3.16	3.31	3.44	3.57	3.68	3.79	4.00	4.19	4.37	4.54	4.69	4.85	4.99	5.13	5.26	5.39
0.950	2.72	2.79	2.85	2.90	2.95	3.00	3.19	3.35	3.50	3.63	3.76	3.88	3.99	4.09	4.29	4.48	4.65	4.81	4.97	5.11	5.26	5.39	5.52	5.65
0.975	3.55	3.58	3.61	3.64	3.66	3.69	3.82	3.94	4.06	4.17	4.28	4.38	4.48	4.58	4.76	4.94	5.10	5.25	5.40	5.54	5.68	5.81	5.94	6.06
0.990	4.69	4.66	4.63	4.62	4.61	4.61	4.63	4.69	4.77	4.85	4.94	5.02	5.11	5.19	5.35	5.51	5.66	5.80	5.94	6.07	6.20	6.33	6.45	6.57
0.995	5.57	5.48	5.41	5.36	5.33	5.30	5.24	5.25	5.30	5.35	5.42	5.49	5.56	5.63	5.78	5.92	6.06	6.19	6.32	6.45	6.58	6.70	6.81	6.93
0.999	7.66	7.42	7.25	7.11	7.00	6.91	6.64	6.53	6.49	6.48	6.50	6.53	6.57	6.62	6.72	6.83	6.94	7.05	7.16	7.28	7.39	7.50	7.60	7.71

Table de la loi Gamma Incomplète en valeurs reduites (non centrées); u est une variable réduite (non centrée). x=u*écart type de x

Lamda = (moyenne des x)^2 / (écart type des x)^2

Exemple : si Lamda= 1.5 ; F(u)=.1 donne u= 0.24. Soit x=.24*écart type de x, pour avoir F(x)=0.1

(Exemples de graphes de la loi bêta)



III-4) APERCU SUR LES LOIS **BETA (*)**

Cette famille de lois est reliée à celle des lois Gamma d'abord par les ingrédients analytiques qu'elle utilise. Elle s'exprime par :

a) Cas de deux paramètres: la loi ***B1:***

Cette loi B1 a pour expression, pour $x \in [0,1]$:

$$f(x,\alpha,\beta) = \frac{1}{B(\alpha+1,\beta+1)}.x^{\alpha}.(1-x)^{\beta} \quad \text{avec} \quad B(\alpha+1,\beta+1) = \frac{\Gamma(\alpha+1).\Gamma(\beta+1)}{\Gamma(\alpha+\beta+2)}$$

qui varie entre les *bornes* 0 et 1
(c'est par exemple le cas de la durée d'insolation, qui varie entre 0 et 100% du potentiel astronomique).
C'est une des utilisations les plus originales, car les autres lois vues précédemment ne sont en général bornées que d'un côté. On en donne une illustration ci-contre.

b) Cas de deux paramètres: la loi ***B2:***

L'expression de la loi B2 est, pour x >0:

$$f(x,\alpha,\beta) = \frac{1}{B(\alpha+1,\beta+1)}.x^{\alpha}.(1+x)^{\beta}$$

qui varie entre 0 et l'∝.

On en donne quelques illustrations (page ci-contre et page suivante) qui montre qu'elle peut ressembler à la loi Gamma, mais qu'elle permet aussi de représenter des dissymétries inversées.

c) Cas de quatre paramètres:

Quand les bornes ne sont pas 0 et 1, mais a et b (comme par exemple la direction du vent entre 0 et 360°) on a alors une loi à 4 paramètres de la forme :

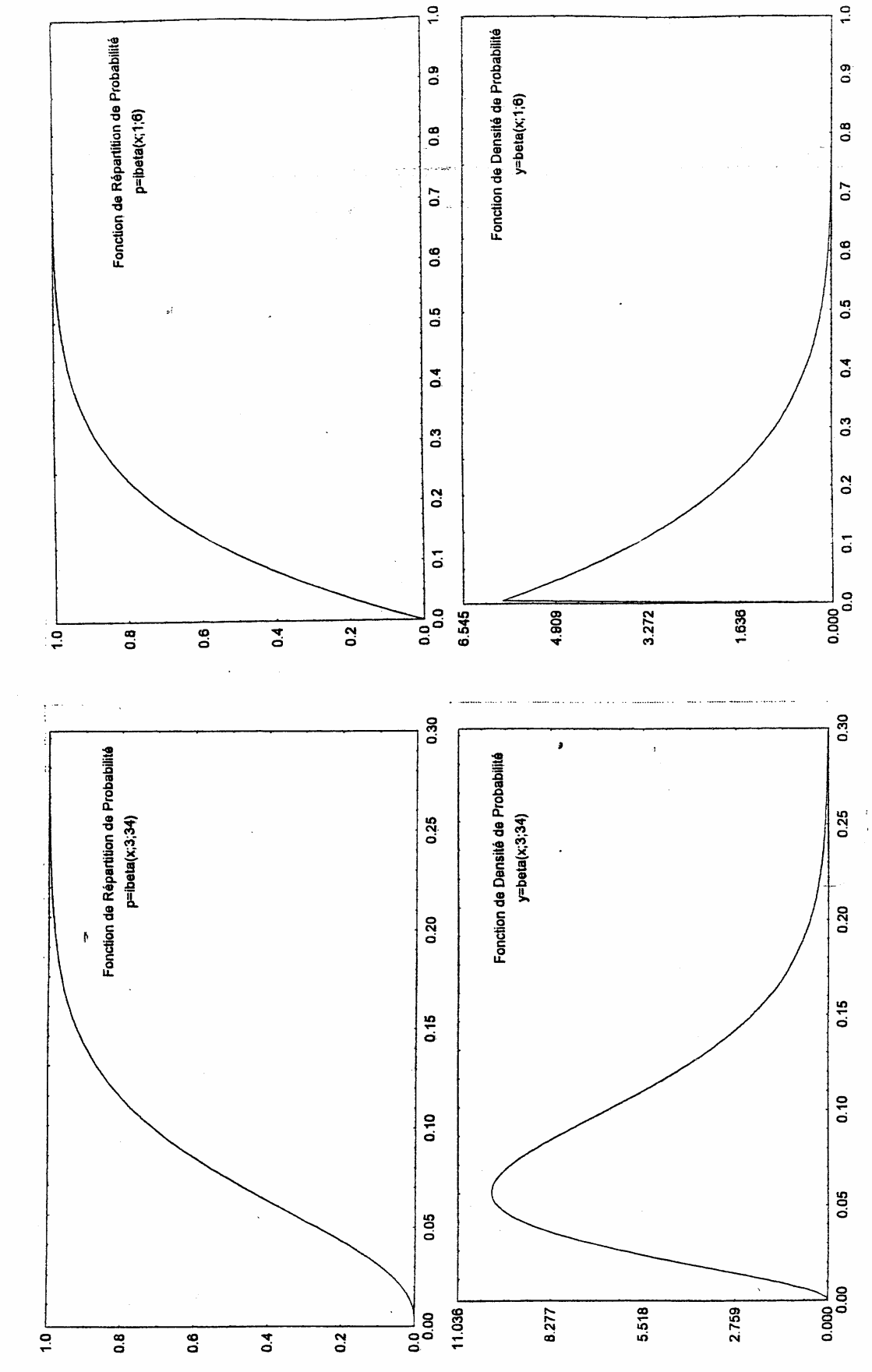
$$f(x,\alpha,\beta) = \frac{1}{(b-a).B(\alpha+1,\beta+1)}.\left(\frac{x-a}{b-a}\right)^{\alpha}.\left(\frac{b-x}{b-a}\right)^{\beta}$$

On renvoie aux ouvrages spécialisés pour leurs propriétés.

Utilisation:

En hydroclimatologie, nous avons été amenés à utiliser ces lois pour représenter la fréquence des durées d'insolation ou des quantités de rayonnement (bornées entre 0 et le maximum astronomique qui lui même dépend de la date dans l'année. Pour éviter cet aspect saisonnier, nous utilisons plutôt le rapport d'insolation, exprimé en pourcentage du maximum astronomique possible, qui varie donc entre 0 et 100). On se reportera par exemple à M. MARQUES (1982).

(Exemples de graphiques de loi bêta)



IV- FAMILLE DES LOIS EXPONENTIELLES ET LOIS DES VALEURS EXTRÊMES

IV-1) LOI EXPONENTIELLE

On a vu que cette loi fait partie de la famille des lois Gamma. C'est le cas particulier où $\lambda = 1$.
On comprend donc que la forme est fixée (c'est une exponentielle), et qu'elle a **un seul paramètre d'échelle** ρ.

Elle s'écrit, respectivement en fonction de répartition ou en densité de probabilité:

$$F(x,\rho) = 1 - e^{-\frac{x}{\rho}} \qquad f(x,\rho) = \frac{1}{\rho}.e^{-\frac{x}{\rho}}$$

mais on la trouve aussi écrite avec $\alpha = \frac{1}{\rho}$, soit alors:

$$F(x,\alpha) = 1 - e^{-\alpha.x} \qquad f(x,\alpha) = \alpha.e^{-\alpha.x}$$

a) Calcul des Moments **(*)**

On rappelle que:

$$\mu_1 = E[X] = \int_0^{+\infty} x.f(x,\rho).dx = \int_0^{+\infty} x.\frac{1}{\rho}.e^{-\frac{x}{\rho}}.dx$$

Si on pose:

$$u = \frac{x}{\rho} \qquad \text{d'où} \quad du = \frac{dx}{\rho}$$

$$dv = e^{-\frac{x}{\rho}}.\frac{dx}{\rho} \qquad \text{et} \quad v = e^{-\frac{x}{\rho}}$$

et que l'on intègre par parties:

$$\mu_1 = \int_0^{+\infty} x.\frac{1}{\rho}.e^{-\frac{x}{\rho}}.dx = \rho.\int_0^{+\infty} \frac{x}{\rho}.e^{-\frac{x}{\rho}}.\frac{dx}{\rho} = \rho.\int_0^{+\infty} u.dv = \rho.[u.v]_0^{+\infty} - \rho.\int_0^{+\infty} v.du = \rho.\left[\frac{x}{\rho}.e^{-\frac{x}{\rho}}\right]_0^{+\infty} + \rho.\int_0^{+\infty} e^{-\frac{x}{\rho}}.d(\frac{x}{\rho})$$

On vérifie que le premier terme est nul et que le second devient:

$$\mu_1 = \rho.\int_0^{+\infty} e^{-\frac{x}{\rho}}.d(\frac{x}{\rho}) = \rho.\left[-e^{-\frac{x}{\rho}}\right]_0^{+\infty} = \rho.1$$

D'où, pour une loi exponentielle:

⇒ **la moyenne est égale au paramètre d'échelle ρ:** $\mu_1 = \rho$

On calculerait de même le moment centré d 'ordre 2:

$$\mu_2 = E\left[(X-\mu_1)^2\right] = \text{Variance } V_x = \sigma_x^2 = \int_0^{+\infty} (x-\mu_1)^2 . f(x,\rho).dx = \int_0^{+\infty} (x-\rho)^2 . \frac{1}{\rho} . e^{-\frac{x}{\rho}} . dx$$

En faisant 2 intégrations par parties successives (*le faire en exercice*), on trouvera...:

$$\text{Variance } V_x = \sigma_x^2 = \rho^2 \quad \text{et} \quad \sigma_x = \rho$$

Donc:

$\Rightarrow$ **l'écart-type d'une loi exponentielle** σ est aussi **égal à la moyenne** et au paramètre d'échelle ρ.

On ajoutera dans ces propriétés mathématiques que ρ est aussi, pour la densité de probabilité $f(x,\rho) = \frac{1}{\rho} . e^{-\frac{x}{\rho}}$, l'inverse de l'ordonnée à l'origine.

Enfin, on pourra calculer X_{med}, la **médiane** de la distribution, donc telle que:

$$\int_0^{Xmed} f(x,\rho).dx = \int_0^{Xmed} \frac{1}{\rho} . e^{-\frac{x}{\rho}} . dx = \frac{1}{2}, \quad \text{et vérifier que :}$$

$$Xmed = \rho.\text{Log}2 \text{ et donc } \text{Xmed} < \mu = \rho$$

b) Diagramme Fonctionnel:

Si la fonction de répartition est : $F(x,\rho) = 1 - e^{-\frac{x}{\rho}}$

alors $1 - F(x,\rho) = e^{-\frac{x}{\rho}}$ et $Log\left[1 - F(x,\rho)\right] = -\frac{x}{\rho}$

D'où une relation linéaire (décroissante) entre :

$$Log\left[1 - F(x,\rho)\right] = Log\left[\Pr(X \geq x)\right] \quad \text{et} \quad x.$$

Il suffit donc de calculer la probabilité empirique au *dépassement*, soit 1-Pi dans les notations précédentes, et de le porter dans un **diagramme log-arithmétique**.

On prendra autant de modules logarithmiques qu'il le faut (en général 3 suffisent).

On démarrera le premier module *en haut* par la valeur 1.0, donc le précédent par 0.1, le troisième par 0.001 etc…. (cf. exercice à faire en T.D. et exemple du chapitre III)

IV-2) EXTENSION de la loi Exponentielle (Somme d'exponentielles)

a) Cas d'une discontinuité à l'origine:

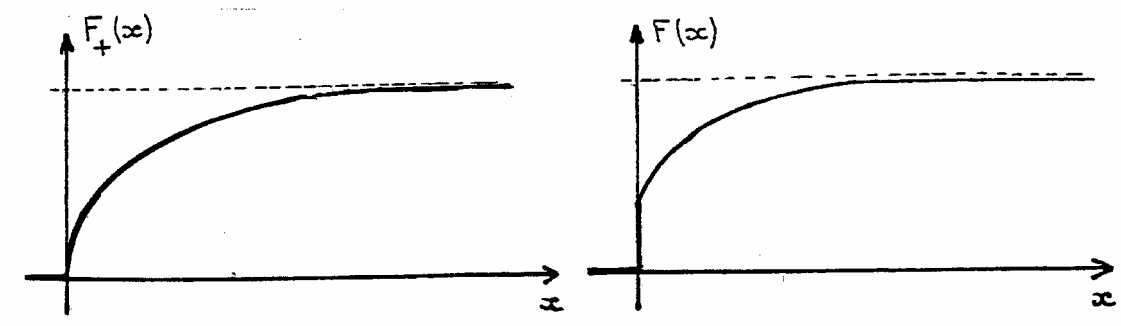
La loi exponentielle est très utilisée pour la distribution des pluies à courts pas de temps (ex. :l'épisode, la journée ou quelques heures). Dans ce cas, une fraction importante des valeurs est *nulle*:

$$\Pr(X = 0) = F(0) = 30 \text{ à } 70\ \%$$

(selon la région et le pas de temps considéré)

Dans un premier temps donc, on doit le faire apparaître dans l'expression de la fonction de répartition:

- Si l'on appelle $\mathbf{F_+(x)}$ la distribution des valeurs positives non nulles de x
- et F(0) la proportion de valeurs strictement nulles



alors on montre que la distribution de toutes les valeurs **≥ 0** devient:

$$F_+(x) = \frac{F(x)-F(0)}{F(\infty)-F(0)} \quad \text{ou encore} \quad F(x) = F(0) + [1-F(0)].F_+(x)$$

On va donc introduire un **nouveau paramètre θ** = fréquence des valeurs strictement nulles, et proposer comme distribution de toutes les valeurs:

$$F(x) = \theta + [1-\theta]\left[1 - e^{-\frac{x}{\rho}}\right] = 1-(1-\theta).e^{-\frac{x}{\rho}}$$

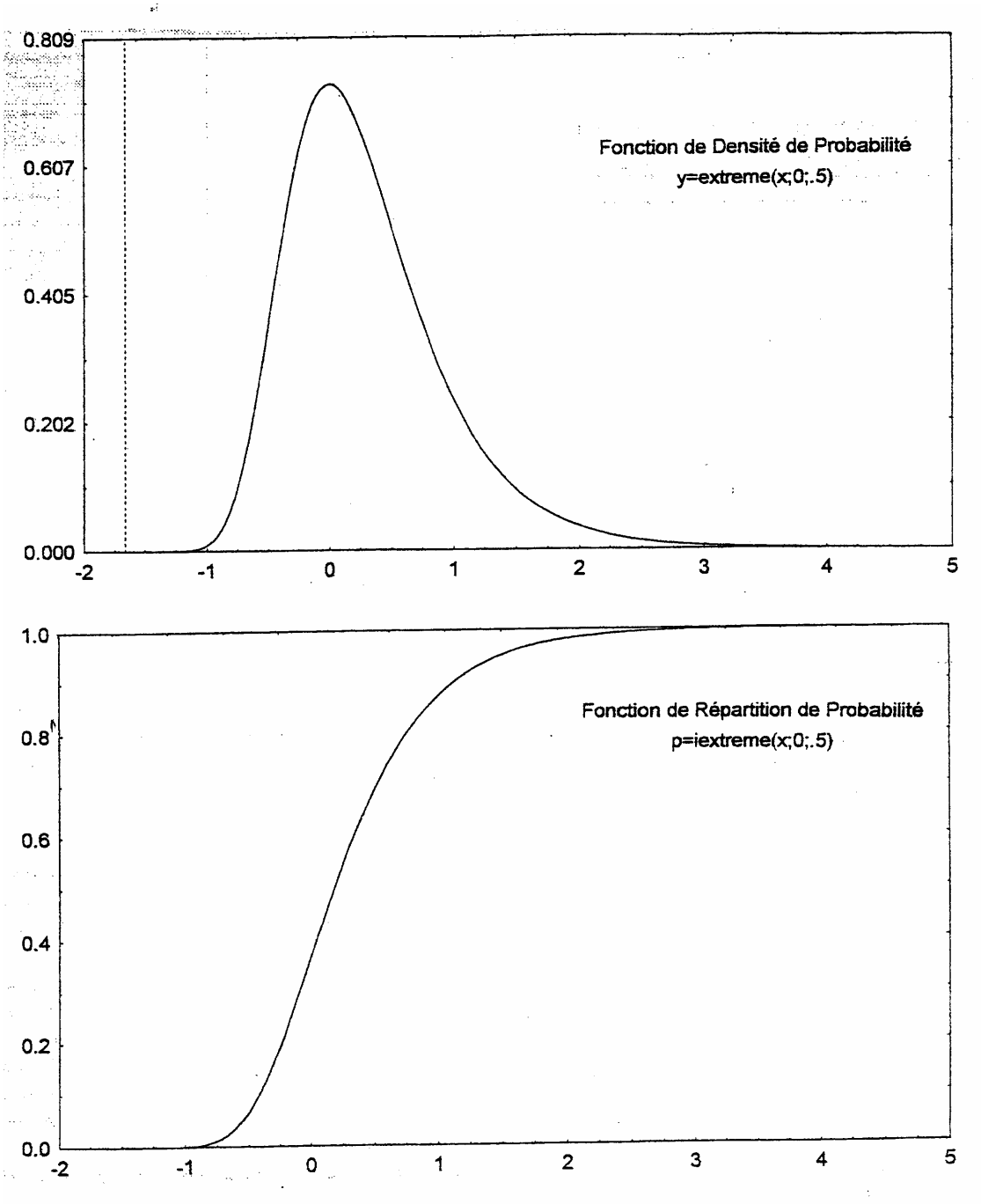
b) Somme de deux exponentielles **(*)**

Sur papier log-arithmétique, il est fréquent que les séries de pluies s'ajustent non pas à une droite mais à deux ou plusieurs tronçons de droite ⇒ d'où l'idée qu'il n'y a pas une mais plusieurs exponentielles qui se superposent, et que la fonction de répartition s'écrit plutôt:

$$F(x) = 1 - A.e^{-\frac{x}{\alpha}} - B.e^{-\frac{x}{\beta}}$$

Cette fonction se révèle meilleure que la fonction Gamma en particulier au voisinage de 0 et dans les valeurs extrêmes (-où l'exponentielle la moins décroissante devient dominante et où la distribution se ramène à cette seule distribution exponentielle-)

Fonction de répartition et densité de probabilité de la loi de GUMBEL



IV-3) LOI de **GUMBEL** (ou Loi des Valeurs Extrêmes de type I)

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.

On la rattache ici à la famille des lois exponentielles, mais les spécialistes la rattachent aussi à la loi Généralisée des Valeurs Extrêmes ***(G.E.V.)*** ou loi de Jenkinson, dont elle est un cas particulier (cf. parag. IV-4 ci après)

a) Forme analytique:

C'est une loi à *2 paramètres* α *et* β*,* tous les deux de même dimension que x. Elle est définie pour toute valeur de x par sa **fonction de répartition** F(x, α , β) :

$$F(x,\alpha,\beta) = e^{-e^{-\left(\frac{x-\beta}{\alpha}\right)}}$$

Sa densité s'écrit: $$f(x,\alpha,\beta) = \frac{1}{\alpha}.e^{-\left(\frac{x-\beta}{\alpha}\right)}.e^{-e^{-\left(\frac{x-\beta}{\alpha}\right)}}$$

et on vérifiera que le maximum de cette densité ou *Mode* est obtenu pour x = β.

Elle est souvent utilisée pour l'étude des valeurs extrêmes (crues, pluies extrêmes, hauteur de vagues), car elle repose sur une théorie qui se résume ainsi :

- pour une variable respectant certaines conditions
- *si* on prend k échantillons de taille N
- et *si* sur chaque échantillon de N individus on sélectionne le max, ou le min, *alors*
- les k maxima ou minima observés suivent une loi de Gumbel.

Il arrive que ces conditions soient assez bien remplies dans la nature (cas des pluies extrêmes à pas de temps assez fin), mais cela n'est quand même pas général.
Il est pourtant fréquent de la voir appelée "loi des valeurs extrêmes", comme si elle s'appliquait à tous les cas ...

En dépit de ses propriétés particulières, surtout intéressantes "dans la queue de la distribution" (pour les probabilités proches de 1), l'allure de la courbe est assez banale, proche d'une loi Gamma ou Lognormale dans sa partie médiane (*cf. page ci- contre*).

b) Calcul des Moments

On trouve, en effectuant le calcul analytique:

$$\mu_x = \beta + 0.577.\alpha \qquad \text{et} \qquad V_x = 1.645.\alpha^2$$

et inversement, en exprimant les paramètres en fonction des moments:

$$\alpha = \frac{\sqrt{6}}{\pi} = 0.7797.\sigma_x \qquad \text{et} \qquad \beta = \mu_x - 0.577.\alpha = \mu_x - 0.444.\sigma_x$$

Notons que ceci sera utilisé dans la méthode d'ajustement dite méthode des Moments pour le calage des paramètres d'une loi de Gumbel. Mais il existe d'autres méthodes que nous verrons au chapitre suivant.

Enfin, on peut vérifier que son coefficient d'asymétrie: $CS \text{ ou } \gamma = \frac{\mu_3}{\sigma_x^{\,3}} = 1.14$

est constant, de même que son coefficient d'aplatissement $\frac{\mu_4}{\sigma_x^{\,4}} = 5.4$

c) Papier de Gumbel :

Comme, d'après l'expression de la loi:

$$-Log\left(-Log\left[F(x)\right]\right) = \frac{x-\beta}{\alpha}$$

si on porte sur un papier à échelles arithmétiques:

-Log(-Log[F*(x)]) en fonction de x,

(où F*(x) est la probabilité empirique estimée sur l'échantillon)

alors les points (si n est grand) seront à peu près alignés (puisque l'on a alors l'équation d'une droite).

Un papier de Gumbel est donc constitué d'une échelle arithmétique pour la variable x et d'une échelle doublement logarithmique en probabilité, mais:

- arithmétique en **u = -Log(-Log[F(x)])** $= \frac{x-\beta}{\alpha}$, dite **variable de Gumbel**,

- mais *graduée* en fait en valeurs de F(x)

En pratique, cette échelle est complètement dilatée vers les valeurs de fortes probabilités au non dépassement.

On donne en annexe un papier de Gumbel, qui est très utilisé pour décrire un échantillon de valeurs extrêmes.

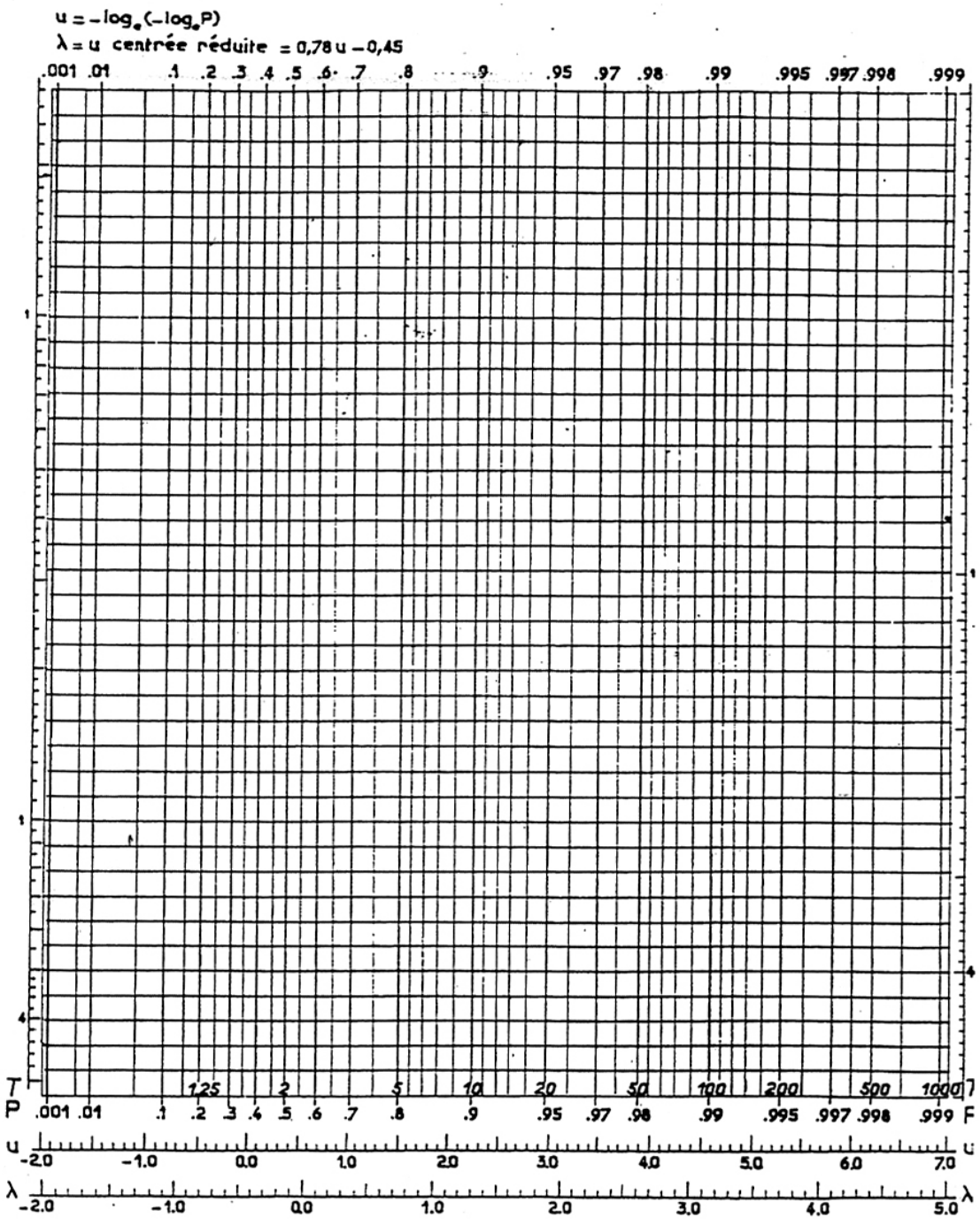
d) Extension de la loi de Gumbel :

Comme précédemment, on peut être tenté d'enrichir la boîte à outils en se demandant si ce n'est pas une transformée de la variable X qui suit une loi de Gumbel.

Par exemple, si Y = Log $(x-x_0)$ suit une loi de Gumbel, alors X suit une **loi de Fréchet**.

C'est la raison pour laquelle le papier de Gumbel est souvent proposé avec un axe supplémentaire à échelle logarithmique.

papier de Gumbel



IV-4) APERCU SUR D'AUTRES LOIS DE VALEURS EXTRÊMES: (*)
Loi de WEIBULL ET G.E.V.

Quand on maîtrisera à peu près la Loi de Gumbel, on pourra s'interroger sur l'utilisation d'autres lois pour représenter la distribution des valeurs extrêmes (minima ou maxima).
On en trouvera une description assez complète, mais didactique, dans le remarquable ouvrage de Kottegoda et Rosso (1997)

La loi de Gumbel est souvent appelée loi des valeurs extrêmes de **Type I.**

La loi de **Type II** s'écrira $$F(x,u,k) = e^{-\left(\frac{u}{x}\right)^k}$$

C'est la loi de Fréchet, qui est à la loi de type I ce que la loi lognormale est à la loi normale.

La loi de **Type III** s'écrira: $$F(x,u,k) = 1 - e^{-\left(\frac{x}{u}\right)^k}$$

ou Loi de Weibull et elle est utilisée pour les valeurs minimales.

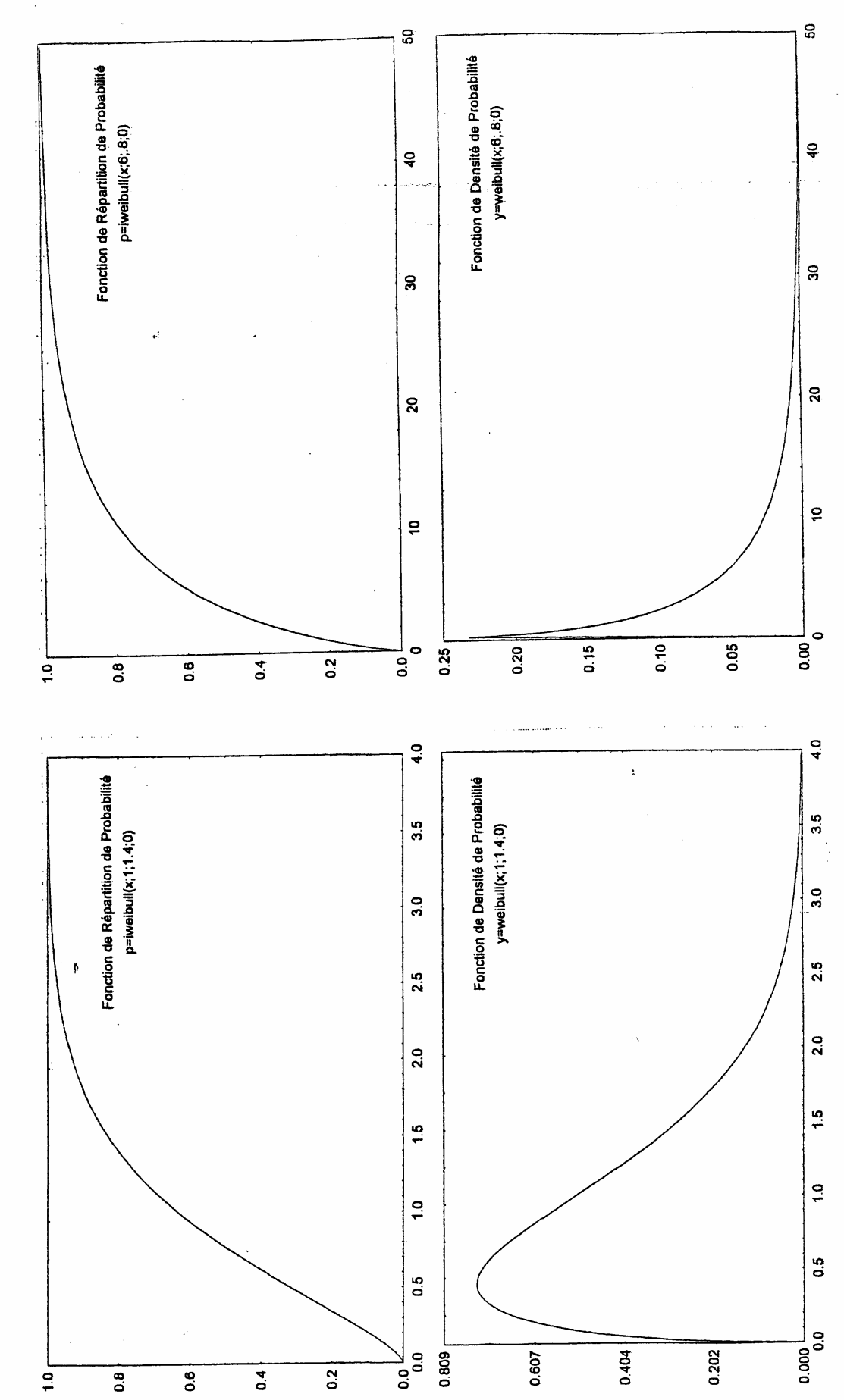
A. F. Jenkinson (1955) a trouvé une formulation générale de ces trois lois sous la forme:

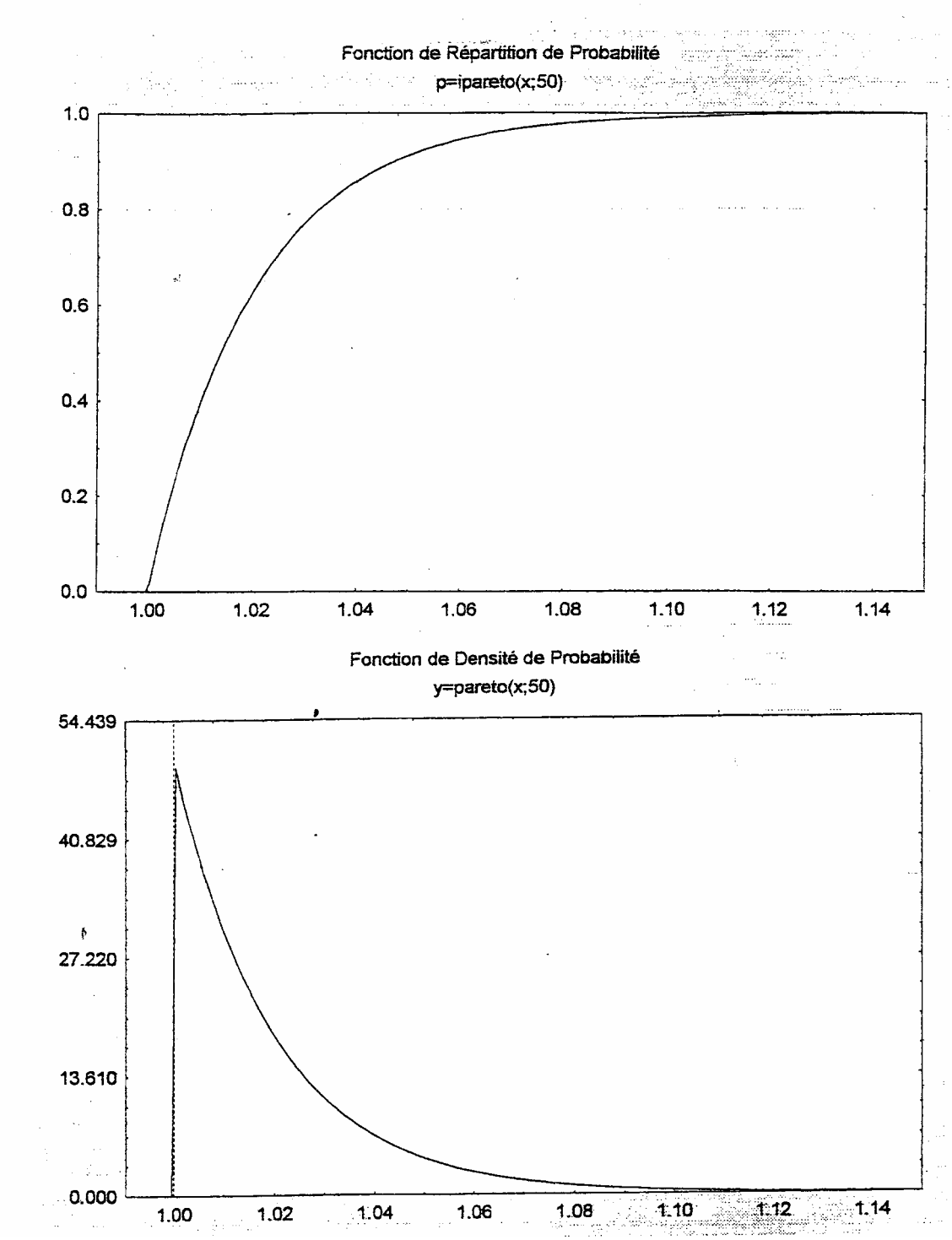
$$F\left(x,x_0,\alpha,k\right) = e^{\left[-\left(1-\frac{k(x-x_0)^{\frac{1}{k}}}{\alpha}\right)\right]} \qquad k \neq 0 \quad \alpha > 0$$

qui dégénère en Loi de Gumbel pour k = 0.

Ces lois sont de plus en plus utilisées et on complètera ces aspects dans le cours de 3ème année sur le calcul des valeurs extrêmes pour les crues de projet.

On donne ci-dessous et page suivante quelques illustrations de la loi de Weibull et de PARETO, non présentée en détail. Juste pour information, la loi de Pareto est de la forme (cf. Kottedoga et Rosso 1997 : $$\boldsymbol{F(x, x_0, \theta)} = 1 - \left(\frac{\boldsymbol{x_0}}{\boldsymbol{x}}\right)^{\boldsymbol{\theta}}$$


Fonction de Répartition de Probabilité
p=iweibull(x;6;.8;0)
Fonction de Densité de Probabilité
y=weibull(x;6;.8;0)
Fonction de Répartition de Probabilité
p=iweibull(x;1;1.4;0)
Fonction de Densité de Probabilité
y=weibull(x;1;1.4;0)


Fonction de Répartition de Probabilité
p=ipareto(x;50)
1.0
0.8
0.6
0.4
0.2
0.0
1.00
1.02
1.04
1.06
1.08
1.10
1.12
1.14
Fonction de Densité de Probabilité
y=pareto(x;50)
54.439
40.829
27.220
13.610
0.000
1.00
1.02
1.04
1.06
1.08
1.10
1.12
1.14

V-) QUELQUES LOIS DE VARIABLES DISCRETES:

Il s'agit de lois destinées à traiter des variables *discrètes* (qui ne peuvent prendre que certaines valeurs préfixées)

Exemples:

- la TVA n'a que 4 taux possibles: 5%, 18.6%, 20.6% ou 33% (si tant est que ce soit une variable aléatoire..!)
- le résultat du jet d'un dé (1, 2, 3, 4, 5, 6) ou de plusieurs dés (pour 3 dés, les valeurs vont de 3 à 18, etc...)
- le nombre de véhicules passant dans un temps donné à un péage autoroutier...

En Hydrologie, on utilisera des variables comme:

- nombre de jours pluvieux d'un mois donné (de 0 à 31)
- nombre de crues (-dans l'année-), supérieures à un certain seuil de débit, etc...

V-1) LOI de POISSON:

C'est une loi définie pour x entier positif ou nul, elle n'a qu'un seul paramètre **a**

$$\Pr[X = x] = \frac{a^x . e^{-a}}{x!}$$

On peut calculer son moment d'ordre 1:

$$\mu_X = \sum_{x=0}^{\infty} x.\Pr[X = x] = \sum_{x=0}^{\infty} x.\frac{a^x . e^{-a}}{x!} = 0.\Pr[X = 0] + \sum_{x=1}^{\infty} x.\frac{a^x . e^{-a}}{x!}$$

$$= \sum_{x=1}^{\infty} x.\frac{a^x . e^{-a}}{x!} = e^{-a}.\sum_{x=1}^{\infty} x.\frac{a^x}{x!} = e^{-a}.a.\sum_{x=1}^{\infty} \frac{a^{x-1}}{(x-1)!} = e^{a}.a\ .e^{-a} = a$$

(car le Σ est en fait le développement en série de e^{-a}.)

Dans le chapitre III, on verra donc que pour la méthode des moments ou du maximum de vraisemblance, on prendra tout simplement le paramètre **a** égal à la moyenne empirique de l'échantillon.

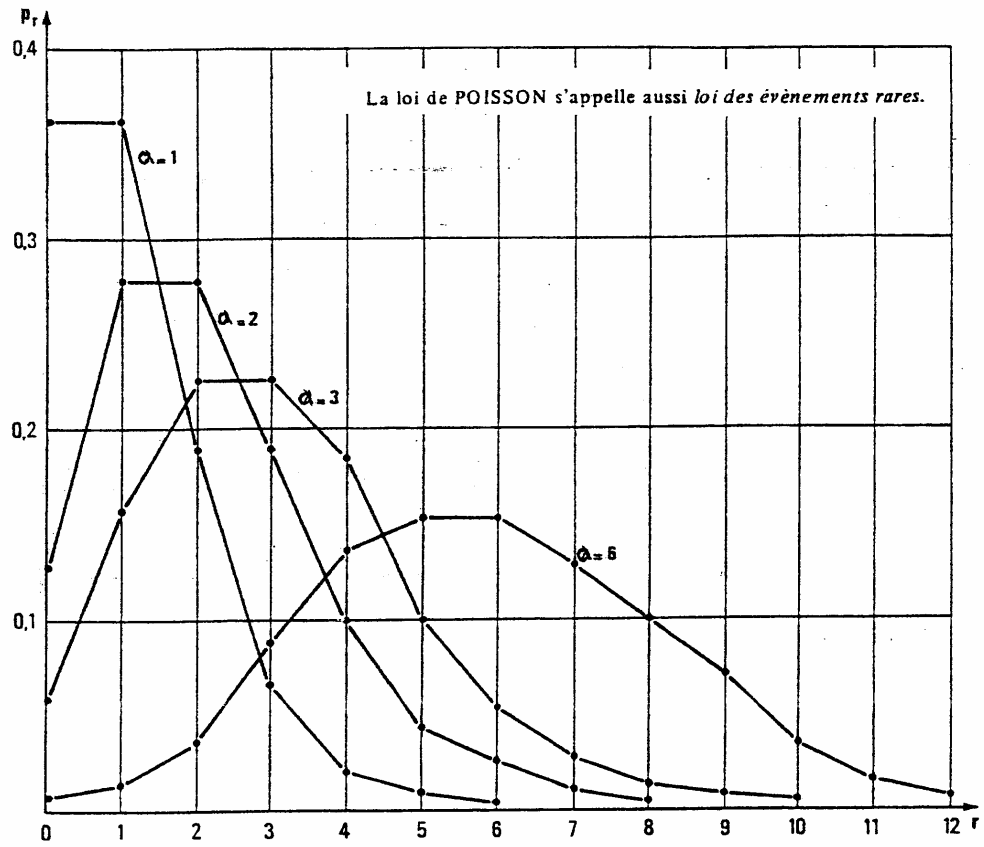
$$a = m_x = \bar{x}$$

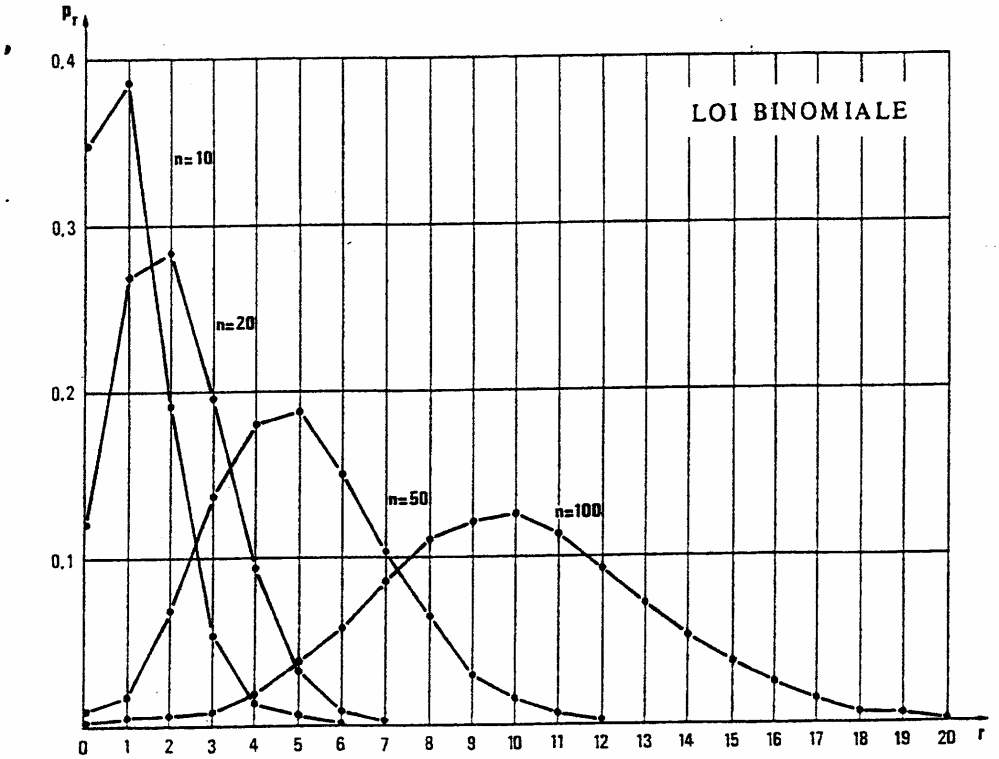
Exemple :

Nombre de crues par an de l'Isère dépassant 500 m3/s à Grenoble,
(si on fait l'hypothèse (à vérifier) que cette variable est bien décrite par la loi de Poisson de moyenne "3 crues par an") :

	Nb de crues par an :							
(dépassant 500 m3/s):	0	1	2	3	4	5	6	7
Probabilité (en %):	5%	15%	22%	22%	17%	10%	5%	2%

Notons au passage le caractère non symétrique de cette loi.





D'après VIALAR 1986

V-2) LOI **BINOMIALE**

C'est une loi définie pour les n+1 valeurs entières 0, 1, 2, 3,... n;
la valeur k tirée au hasard d'une loi binomiale ayant pour probabilité :

$$\Pr[X = k] = C_n^k p^k . q^{n-k} = \frac{n!}{k!(n-k)!} p^k . q^{n-k} \qquad \text{avec p + q = 1}$$

Calcul des moments :
On montre que : $E[X] = \mu_X = n.p$
et que

$$Var[X] = E\left[(X - n.p)^2\right] = n.p.q \qquad et \qquad \sigma_X = \sqrt{n.p.q}$$

toujours avec p+q =1

On remarquera que cette loi n'a qu'un paramètre (puisque p et q sont liés) et que l'on vient de donner 2 relations pour calculer ce paramètre!.

Exemple :

"Probabilité pour avoir en 50 ans 2 crues maximales annuelles (et seulement 2) supérieures à la crue centennale "(cette dernière est la crue qui a une chance sur 100 d'être dépassée chaque année).

L'évènement de base est :
" la crue maximale annuelle est supérieure à la crue centennale".
et comme on considère n = 50 ans, l'évènement peut apparaître:
0 , ou 1 , ou 2 , ou ... k , ou n = 50 fois

La variable k est le nombre de fois où, en 50 ans, k maximas annuels dépassent sur un échantillon infini la valeur de la crue centennale.
$\Rightarrow$ k a donc une moyenne de .5 (50 ans/ 100 ans).

L'évènement de base: "le débit max annuel dépasse la crue centennale"
a une probabilité élémentaire de : p = .01
et son complément q = .99.

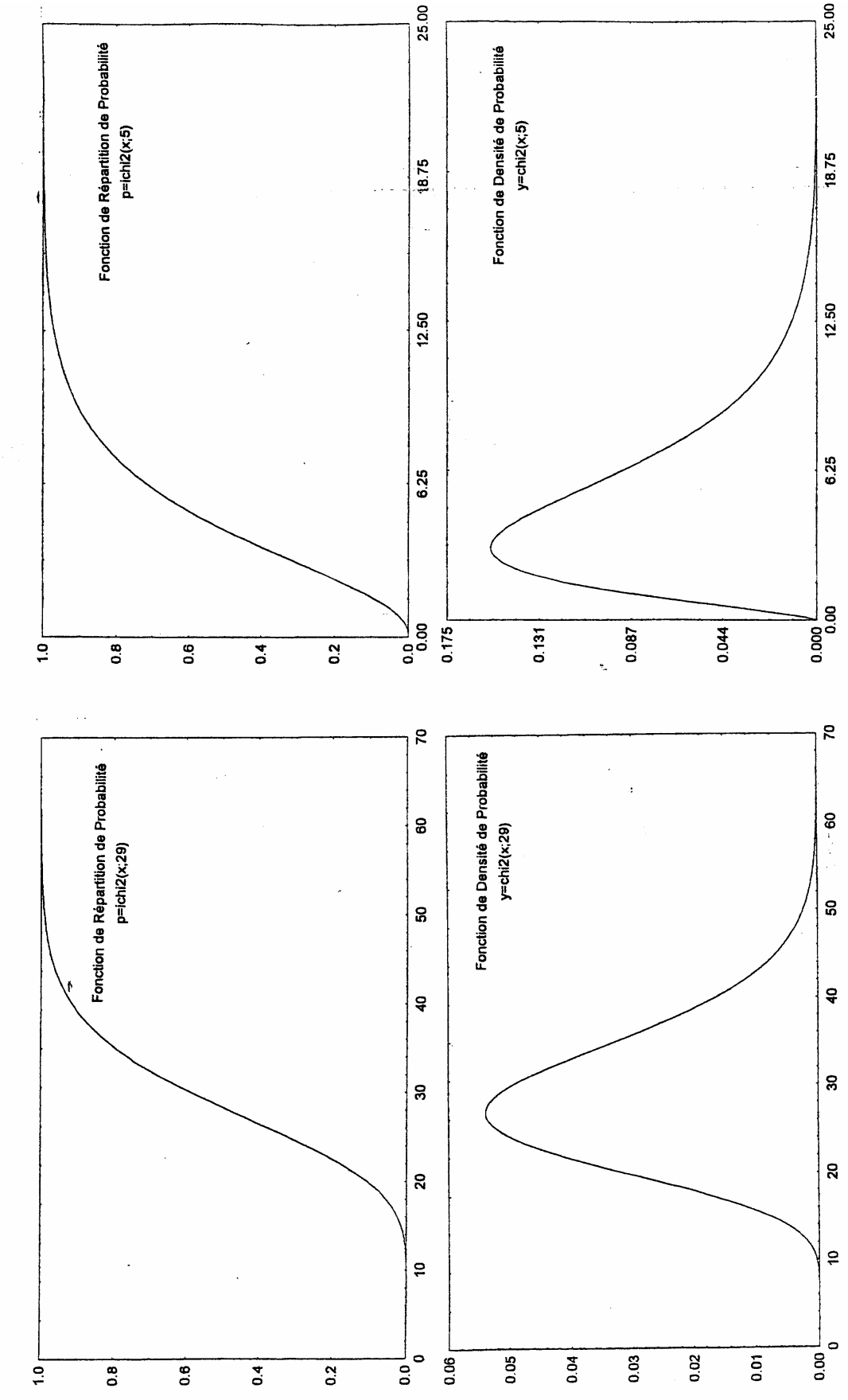
Donc la probabilité que le nombre soit strictement égal à k est :

$$\Pr[X = k] = \frac{n!}{k!(n-k)!} p^k . q^{n-k}$$

soit, pour k = 2 $\Pr[X = 2] = \frac{50!}{2!48!}(0.01)^2 . (0.99)^{48} = 0.08$

$\Rightarrow$ On a donc 8% de probabilité d'observer en 50 ans
2 et *seulement* 2 crues dépassant la crue centennale.

graphes du Chi 2 et de Student



VI-) LOIS UTILISEES DANS LES TESTS d'HYPOTHESES(*):

Les lois qui suivent sont rarement utilisées comme modèle que l'on cherchera à ajuster à un jeu de données. Par contre, on y fait souvent référence dans les *tests d'hypothèses*. Ceux-ci sont utilisés pour décider de l'adéquation d'un modèle, ou dans les distributions des effets de l'échantillonnage (cf. Chap. III de cette partie ou IIème partie sur la corrélation).

VI-1) LOI du **CHI 2 :**

C'est une loi à *un paramètre n* définie pour x > 0 ; l'expression de sa densité est la suivante :

$$f(x,n) = \frac{1}{2^{\frac{n}{2}}.G\left(\frac{n}{2}\right)}.x^{\frac{n}{2}-1}.e^{-\frac{x}{2}}$$

où G est la fonction Gamma classique, et *n est le nombre de* ***degrés de liberté***.
On en donne deux exemples sur les figures ci-contre.

Son origine : c'est la loi de la somme des carrés de n variables normales centrées réduites. Nous l'utiliserons surtout dans les tests d'ajustement.

Tables : On donne en général pour diverses valeurs de n la probabilité au dépassement ou au non dépassement (cf. annexe).

VI-2) LOI de **STUDENT:**

C'est une loi à un paramètre n : $$f(x,n) = \frac{1}{\sqrt{n.p}}.\frac{G\left(\frac{n+1}{2}\right)}{G\left(\frac{n}{2}\right)}.\frac{1}{\left(1+\frac{x^2}{n}\right)^{\frac{n+1}{2}}}$$

où G est la fonction Gamma classique. Là encore on voit un exemple de graphe ci-contre.

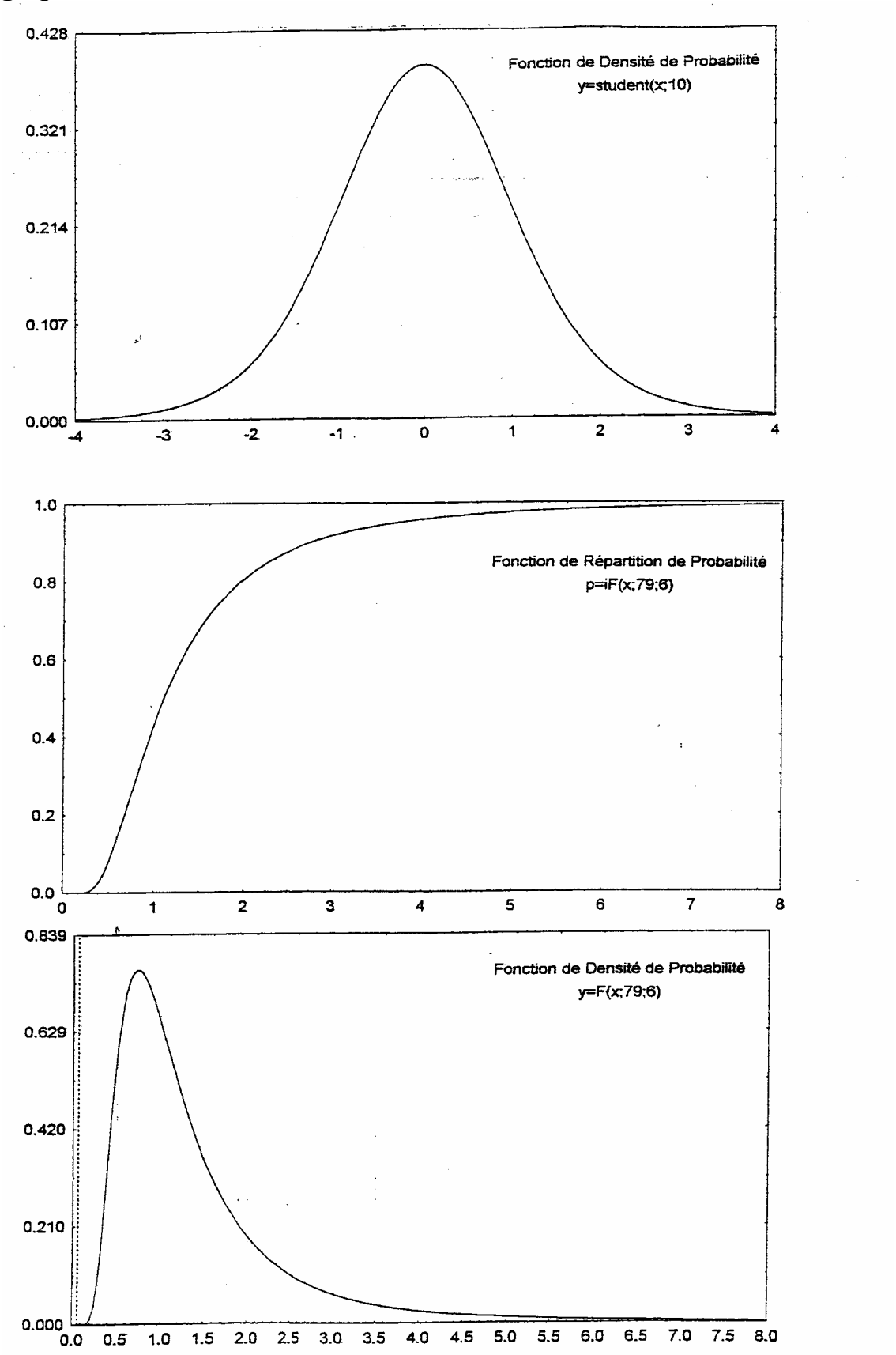
Son origine : Si on prend n variables centrées réduites *gaussiennes* $X_1, X_2, ..X_n$, et alors la variable t , définie ainsi:

$$t = \frac{X}{\sqrt{\frac{1}{n}\sum_{i=1}^{n} X_i^2}}$$

suit une loi de Student ; le paramètre n est appelé nombre de degrés de liberté.

Tables : (cf. annexe)
On l'utilise surtout dans des tests de comparaison de moyennes d'échantillons gaussiens ou pour tester des coefficients de corrélation partielle en corrélation multiple.

graphes Fisher Snedecor



VI-3) LOI de **FISHER SNEDECOR :**

C'est la loi du rapport de deux variables indépendantes du Chi 2, U et V respectivement à m et n degrés de liberté.

$$X = \frac{\frac{U}{m}}{\frac{V}{n}}$$

Cette distribution est définie par sa densité g(x, m,n):

$$g(x,m,n) = \frac{G\left(\frac{m+n}{2}\right)}{G\left(\frac{m}{2}\right).G\left(\frac{n}{2}\right)}.m^{\frac{m}{2}}.n^{\frac{n}{2}}.x^{\frac{m-2}{2}}.(m+n.x)^{-\frac{m+n}{2}}$$

En pratique, on préfère appeler la variable aléatoire F plutôt que X, et la densité devient:

$$g(f,m,n) = \frac{G\left(\frac{m+n}{2}\right)}{G\left(\frac{m}{2}\right).G\left(\frac{n}{2}\right)}.m^{\frac{m}{2}}.n^{\frac{n}{2}}.f^{\frac{m-2}{2}}.(m+n.f)^{-\frac{m+n}{2}}$$

Selon "Statistical Methods in Hydrology" (Ch. Hahn 1977), sa moyenne vaut:

$$E[F] = \frac{m}{n-2}$$

et sa variance vaut: $Var[F] = n^2.\frac{m+2}{m.(n-2).(n-4)}$

Cette loi est surtout utilisée dans les tests d'homogénéité pour comparer des variances d'échantillons. C'est une loi à 2 paramètres m et n. On trouvera une table de cette loi en annexe.

Résumé et conclusions

Nous venons de décrire quelques lois en donnant leurs propriétés essentielles: c'est la ***" boîte à outils***" de base pour l'ingénieur hydrologue.

Mais on imagine facilement qu'il est impossible d'être exhaustif, et qu'*il faut être prêt*, face à une variable nouvelle, à rechercher éventuellement dans les ouvrages spécialisés une nouvelle loi, correspondant à un histogramme particulier.

Nous verrons ci-après comment ajuster leurs paramètres à partir de données observées.

BIBLIOGRAPHIE:

BENJAMIN J.R and CORNELL C.A. (1970).
Probability, Statistics and Decision for Civil Engineers
Mac Graw Hill Pub. Comp. 684 p.

B. BOBEE et F. ASHKAR (1990)
The Gamma family and derived distributions applied in hydrology
Water Ressources Publications PO Box 26 0026 Highlands Ranch Co 80 126 0026 USA Ed Fort Collins 218 p. (+ 10 disquettes - optionnel)

HAAN Ch. T. (1977)
Statistical Methods in Hydrology.
Iowa state University Press 2ème ed. 1979, 378 p.

JENKINSON A.F. (1955)
The frequency distribution of the annual maximum or minimum values of meteorological elements.
Quarterly Journal of the Royal Meteorological Society, vol. 81, pp. 158-171

KOTTEGODA N.T. and R. ROSSO (1997)
Probability, Statistics and Reliability for Civil Engineers and Environmental Engineers
The Mac Graw Hill Pub. Comp. Inc. 735 p.
Ouvrage très complet, très didactique et illustré d'exemples

LUBES H., MASSON J.M., RAOUS P. , TAPIAU M. (1994)
SAFARHY, Logiciel de calculs statistiques et d'analyse fréquentielle adapté à l'évaluation du risque en Hydrologie. Manuel de référence. Editions ORSTOM

M. MARQUES (1982)
Conception d'un modèle stochastique de simulation des rayonnement solaires direct et global à pas de temps fin. Application aux données de Grenoble.
Thèse Université Scientifique et Médicale de Grenoble

ROCHE M. (1963)
Hydrologie de l'Ingénieur
Ed. Gauthier Villard 1963

VIALAR 1986
Probabilités et Statistiques (5 fascicules)
Cours de l'Ecole Nationale de la Météorologie
(disponible auprès du Service des publications de Météo France)

YEVJEVICH V. (1972)
Probability and Statistics in Hydrology
Water Ressources Publications Ed Fort Collins Co USA. 302 p.
(Ouvrage très complet sur les modèles probabilistes –
Le Pr Yevjevich est d'ailleurs sorti de l'ENS d' Hydraulique de Grenoble en 1939)4.

1ére Partie: MODELES PROBABILISTES

CHAPITRE III :

ESTIMATION ET TECHNIQUES D'AJUSTEMENT D'UNE LOI DE PROBABILITE A UN ECHANTILLON

I) Principes de l'ajustement: 87

II) Méthode des Moments 88

II-1) Principe et problème d'estimation associés: 88
II-2) Applications à la loi normale et à ses dérivées 90
II-3) Applications à la loi Gamma 95
II-4) Applications à la loi exponentielle et à ses dérivées 96
II-5) Applications à la loi de Gumbel 97

III) Méthodes Graphiques (d'ajustement d'un échantillon): 98

III-1) Principe et problèmes associés: 98
III-2) Le diagramme Gausso-arithmétique (ou papier "normal") 99
III-3) Le diagramme Log-normal 101
III-4) Loi exponentielle et diagramme Log-arithmétique 102
III-5) Le diagramme de Gumbel 106
III-6) Extensions 106

IV) Méthode du Maximum de Vraisemblance 108

IV-1) Principe et problème d'estimation associés 108
IV-2) Applications à la loi de Poisson 109
IV-3) Application à la loi normale
110
IV-4) Applications à la loi exponentielle et à ses dérivées 110
IV-5) Applications à la loi de Gumbel 111

V) Tests d'hypothèses 114

V-1) Objectif 114
V-2) Test du Chi 2 (χ^2) 115
V-3) Test de Kolmogorov Smirnoff 120

VI) Compléments théoriques : 122

VI-1) La méthode des moments pondérés(*) 122
VI-2) Incertitudes sur les estimateurs - Effets de l'échantillonnage 126

1ère Partie - **CHAPITRE III :**

ESTIMATION ET TECHNIQUES D'AJUSTEMENT D'UNE LOI DE PROBABILITE A UN ECHANTILLON

I) PRINCIPES DE L'AJUSTEMENT:

Après l'analyse descriptive et exploratoire d'un échantillon, on peut avoir une certaine intuition du "Modèle Probabiliste" le plus adéquat pour résumer/représenter cet échantillon.

On choisit en général ce modèle dans une famille de lois bien connues, et comportant certains paramètres:

F(x, ***a1, a2,...ap***) ou f(x, ***a1, a2,...ap***)

⇒ le problème consiste alors à:

trouver, dans cette famille de lois,
celle qui représentera au mieux l'échantillon considéré:.

Cela revient donc à:

fixer, de manière unique et reproductible,
les paramètres du modèle concerné.

Pour caler ces paramètres: α_k, **plusieurs méthodes sont couramment utilisées**. Nous ne décrirons que les plus classiques en détaillant le calcul pour certaines lois.

Signalons tout de suite que, selon les lois et les échantillons, ces méthodes donnent des résultats plus ou moins différents et satisfaisants.

Enfin, nous terminerons, au paragraphe VI, par la présentation d'une méthode dont l'utilisation, récente, tend à se répandre, notamment pour l'étude des valeurs extrêmes, et qui illustre les tendances en cours en recherche.
Nous évoquerons aussi les quelques aspects des effets d'échantillonnage.

Note: Ce chapitre pourra parfois sembler fastidieux. En effet, les démonstrations des méthodes sont souvent laborieuses.

*On se rappellera que l'on est d'abord **utilisateu**r de ces méthodes, en vue de leur application à des problèmes d'ingénierie hydrologique.*

On ne donnera donc que quelques exemples de démonstrations, pour bien assimiler le principe de chaque méthode. Les justifications exhaustives seront à rechercher dans les ouvrages spécialisés.

Cependant, certaines méthodes, bien qu'admissibles, sont connues pour donner avec certaines lois ou dans certaines conditions de piètres résultats: il faudra aussi intégrer cette information et l'utiliser (par exemple le logiciel SAFARHY inclue ce type de "conseils à l'utilisateur " pour un large éventail de lois).

II) METHODE des MOMENTS:

II-1) Principe :

Soit f (x, α_1, α_2,...α_p) la loi retenue, avec les paramètres α_1, α_2,...etc,
et soit un échantillon observé de n valeurs xi de la variable X.

La méthode des moments s'appuie sur les propriétés suivantes:

a) Un théorème nous dit qu'une loi de probabilité est connue:
- soit par l'expression de sa forme analytique
- soit, **de manière équivalente**, par la connaissance de *tous ses moments* (qui sont une infinité)

(Note : Le lecteur intéressé pourra se reporter à un cours de Probabilités, où il verra à ce propos la notion de "Fonction Caractéristique" - cf. par exemple Vialar 1986).

b) D'autre part, la définition et le **calcul théorique des moments** montrent qu'ils sont évidemment *en relation avec les paramètres* de la loi considérée:

$$\mu_k = \int_{-\infty}^{+\infty} \left[x - \mu_1(\alpha_1, \alpha_2, ..\alpha_p) \right]^k . f(x, \alpha_1, \alpha_2, ..\alpha_p).dx = \mu_k(\alpha_1, \alpha_2, ..\alpha_p)$$

où chaque moment, même d'ordre k supérieur à p, ne dépend plus que des p paramètres α_j .

Donc **inversement**, on peut écrire que ces paramètres sont en relation avec les moments par:

$$\alpha_k = G_k(\mu_1, \mu_2, ..\mu_m, ...) \quad \forall\, k$$

et **même**, puisqu'il n'y a que p paramètres :

$\Rightarrow$ il suffit d'écrire les *p premiers moments* pour obtenir *p relations à p inconnues*:

$$\begin{array}{lll}
\mu_1 = \mu_1(\alpha_1, \alpha_2, ..\alpha_p) & & \alpha_1 = G_1(\mu_1, \mu_2, ..\mu_p) \\
\mu_2 = \mu_2(\alpha_1, \alpha_2, ..\alpha_p) & & \alpha_2 = G_2(\mu_1, \mu_2, ..\mu_p) \\
\quad ..... & & \quad .... \\
\mu_k = \mu_k(\alpha_1, \alpha_2, ..\alpha_p) & \Rightarrow & \alpha_k = G_p(\mu_1, \mu_2, ..\mu_p) \\
\quad ..... & & \quad .... \\
\mu_p = \mu_p(\alpha_1, \alpha_2, ..\alpha_p) & & \alpha_p = G_p(\mu_1, \mu_2, ..\mu_p)
\end{array}$$

(et cela même s'il n'est pas toujours évident d'inverser les relations pour trouver explicitement les fonctions G_k , et donc les paramètres...)

c) On sait aussi **estimer les moments *théoriques*** d'ordre k μ_k de la population à partir des ***moments empiriques*** m_k calculés sur l'échantillon.

d) Partant des relations de b), et considérant que l'on a p paramètres à caler, on va écrire que les p paramètres de la loi satisfont les équations:

$$\alpha_1 = G_1(\mu_1, \mu_2, ..\mu_p)$$
$$\alpha_2 = G_2(\mu_1, \mu_2, ..\mu_p)$$
$$....$$
$$\alpha_k = G_p(\mu_1, \mu_2, ..\mu_p)$$
$$....$$
$$\alpha_p = G_p(\mu_1, \mu_2, ..\mu_p)$$

puis : - en **remplaçant** les p **Moments *théoriques*** μ_k de la loi f(x,...)
- par les p **Moments *empiriques*** $\mathbf{m_k}$ calculés sur l'échantillon à partir des x_i .

$\Rightarrow$ on obtient alors:

- les **estimations** $\mathbf{a_k}$ *(en lettres latines car ce sont des estimations)*
- des "vrais" paramètres α_k (inaccessibles! car il faudrait toute la population)

et cela par:

- un système plus ou moins compliqué
- de p équations , (-souvent non-linéaires-) à p inconnues (- les a_k -) :

$$a_1 = G_1(m_1, m_2, ..m_p)$$
$$a_2 = G_2(m_1, m_2, ..m_p)$$
$$....$$
$$a_k = G_p(m_1, m_2, ..m_p)$$
$$....$$
$$a_p = G_p(m_1, m_2, ..m_p)$$

Comme en général, on ne manipule guère de lois à plus de 3 paramètres, on a au plus un système de 3 équations à 3 inconnues, certes non linéaires, mais soluble par les méthodes classiques (Newton-Raphson, etc...)

Remarque:

On peut théoriquement prendre n'importe quelle relation entre paramètres et moments. En pratique, on utilise toujours les relations entre les p **premiers** moments et les paramètres, car la théorie de l'échantillonnage nous montre que ce sont les premiers moments (ceux d'ordre le plus bas) qui sont le mieux estimés.

En effet, pour estimer un moment d'ordre k, on utilise les estimations des moments d'ordre inférieur k-1, k-2, etc... et donc on propage les erreurs faites sur ceux qui précèdent… !

On va maintenant voir des exemples de cette méthode sur les lois étudiées au Chapitre II.

II-2) APPLICATIONS à la LOI NORMALE et à ses DERIVEES

II-2-a) Cas de la loi normale

La loi Normale a donc 2 paramètres α et β de même dimension que la variable X.
Soit un échantillon de n valeurs de la variable X : cherchons quelles sont les valeurs de a et de b permettant à une loi Normale de s'ajuster au mieux avec les n valeurs de l'échantillon. Pour cela, nous allons utiliser la méthode des Moments .

La loi Normale ayant *deux* paramètres α et β, on va donc :

- chercher les valeurs estimées a et b de α et β,
- telles que les *deux* premiers moments (moyenne et variance),
  (dont l'expression est fournie au chapitre II p. II-7 et II-8)
- soient égales à la moyenne et à la variance des valeurs x_i *de l'échantillon*.

Pour la loi théorique paramétrée: $f(x,\alpha,\beta) = \frac{1}{\alpha.\sqrt{2\pi}}.e^{-\frac{1}{2}.\left(\frac{x-\beta}{\alpha}\right)^2}$,

on a trouvé : Moment théorique d'ordre 1 : $\mu_1 = \mu_X = \beta$
Moment théorique d'ordre 2 : $\mu_2 = \sigma_X^2 = \alpha^2$

Et ces deux relations suffisent théoriquement pour déterminer les paramètres α et β à partir des moments *théoriques* μ_1 et μ_2.

Comme ceux-ci μ_X et σ_X sont inconnus, ***on les remplace*** par les moments *empiriques* m_X et s_X , ou plus simplement m et s, calculés sur l'échantillon.

D'où finalement les *valeurs estimées des paramètres:* (estimées par la méthode des moments)

b = m_X = moyenne de l'échantillon
a = s_X = écart type de l'échantillon

et la loi normale, *ajustée* à cet échantillon particulier *par la méthode des moments*, aura pour expression :

$$f(x,a,b) = \frac{1}{s_x.\sqrt{2\pi}}.e^{-\frac{1}{2}.\left(\frac{x-m_x}{s_x}\right)^2}$$

On aura donc *forcé l'égalité stricte* entre les deux premiers Moments *théoriques* de cette loi et les deux premiers Moments *empiriques* de l'échantillon.

On pourra aussi comparer sur un graphique la droite correspondant à cette loi ajustée avec une autre droite obtenue par ajustement graphique direct (à l'oeil), (cf. paragraphe suivant III-2).

II-2-b) Cas de la loi Log-normale à **deux** paramètres:

On a vu au chapitre II (parag. II-2) les expressions *théoriques* des moments
- de la loi normale sur la variable Y = Log X ,
- en fonction de ceux de la variable X
- et réciproquement.

Si dans ces expressions:
- on remplace les moments *théoriques*
- par les moyenne et écart-type *empiriques* m_X et s_X calculés sur l'échantillon,

on obtient (cf. par ex. Chadule p. 71):

$$m_Y = Log\left(\frac{m_X^2}{\sqrt{m_X^2 + s_X^2}}\right)$$

et pour variance:

$$s^2{}_Y = Log\left(1 + \frac{s_X^2}{m_X^2}\right)$$

Et on les utilisera ensuite sur le graphique lognormal (cf. paragraphe III-3 suivant) pour tracer une droite que l'on comparera éventuellement à un tracé direct à l'oeil (méthode graphique).

ATTENTION ! ! ! Mise en garde :

Quand X suit ***exactement*** une loi Log-normale, ces estimateurs semblent déjà assez sensibles à l'échantillonnage.
Pour le vérifier :

1)- on calculera le logarithme de chaque valeur brute, et ensuite la moyenne et l'écart-type de ces logarithmes, soit m_Y et s_Y .

2)- parallèlement, on appliquera les formules ci-dessus, qui, à partir du calcul de m_X et s_X proposent une autre estimation m^*_Y et s^*_Y .

3)- on comparera alors les estimations
directes par le calcul de m_y et s_y
ou indirectes à partir des formules et du calcul de m_x et s_x:
$\Rightarrow$ elles diffèrent souvent de 10 ou 20% , surtout pour l'écart-type s_y

Cependant, quand la distribution est dissymétrique ***mais n'est pas strictement lognormale*****,** on peut toujours, comme en 1)-, calculer directement m_y et s_y à partir du logarithme des valeurs brutes et travailler dessus..

Par contre, dans ce dernier cas, l'application des formules théoriques qui les relient aux valeurs initiale en x, formules strictement *valables seulement et seulement pour une distribution lognormale*, n'est plus adaptée et ne se justifie plus.
$\Rightarrow$ Les écarts entre s^*_y et s_y sont souvent d'un ***facteur*** 1,5 ou 2 ..!

Ces formules étaient attractives avant l'avènement des calculateurs, quand il fallait consulter une table pour trouver le log de chaque valeur, puisqu'elles évitaient justement de calculer le logarithme.
Aujourd'hui il vaut mieux calculer les logarithmes et travailler directement dessus.

C'est pourquoi *nous* ***déconseillons plutôt d'utiliser ces formules*** (et la méthode des moments) pour la loi Lognormale.

II-2-c) Cas de la loi Log-normale à **trois** paramètres:

Dans ce cas, ce n'est plus la variable:

Y = Log X qui suit une loi normale *mais* la variable Y = Log (X - $\mathbf{x_0}$)

Et on cherche à déterminer x_0 de sorte que Y soit le plus "normal" possible.

On utilise pour cela une propriété de la loi normale, à savoir que son asymétrie est nulle, donc que le moment d'ordre 3 *de la variable Y* devrait être nul.

Si on l'exprime en fonction des moments de la variable initiale X, et qu'on l'annule, on voit que x_0 doit satisfaire la relation:

$$\frac{(\mu_x - x_0)^3}{\sigma_x^2 + 3.(\mu_x - x_0)^2} = \frac{\sigma_x^4}{\mu_{3x}}$$

avec $\mu 3_X$ le moment centré d'ordre 3 de X : $\mu_{3x} = E\left[X^3\right] - 3.\mu_X.\sigma_X^2 - \mu_X^3$

On connaît déjà des estimateurs empiriques de μ_X et σ_X . $\Rightarrow$ Pour $\mu 3_X$ on prendra:

$$m_{3x} = \frac{1}{(n-1).(n-2)}\left[n\sum_{i=1}^{n} x_i^3 - 3\sum_{i=1}^{n} x_i^2 . \sum_{i=1}^{n} x_i + \frac{2}{n}\left(\sum_{i=1}^{n} x_i \right)^3 \right]$$

On en trouvera la démonstration dans l'ouvrage de M. Roche (1965), et une application un peu voisine dans SAFARHY.

On traitera en exercice une application à une variable bien adaptée à la loi LogNormale (les débits mensuels de la Romanche), mais on donne ici quelques exemples un peu plus problématiques.

EXEMPLES

a) *Exemples simulés:* Soit un échantillon de 20 valeurs ***"vraiment"*** tirées d'une loi Log normale

N° de -l'observation	Val. Brutes Xi	Transform. Yi = LnXi
1	245.74	5.504273
2	218.52	5.386859
3	227.06	5.425231
4	52.57	3.962241
5	169.52	5.132996
6	133.74	4.895869
7	96.10	4.565418
8	48.66	3.884836
9	556.26	6.321229
10	282.82	5.644801
11	252.27	5.530508
12	199.04	5.293523
13	598.22	6.393953
14	176.79	5.174989
15	468.64	6.149843
16	106.43	4.667445
17	39.39	3.673489
18	1853.22	7.524682
19	16.86	2.824712
20	858.81	6.755545
moyenne	330.0331 m_x	5.2356221 m_y
écart-type	418.5698 s_x	1.1214482 s_y

Et les formules théoriques proposent :

$$m^*_Y = Log\left(\frac{m_X^2}{\sqrt{m_X^2 + s_X^2}}\right) = 5.3198$$ à comparer à 5.2356

et pour l'écart-type:

$$s^*_Y = \sqrt{Log\left(1 + \frac{s_X^2}{m_X^2}\right)} = 0.979$$ à comparer à 1.121

De même pour un échantillon de 500 valeurs simulées:

moyenne	225.857985 m_x	4.9244716 m_y
écart-type	306.663117 s_x	0.9942243 s_y

à comparer avec des valeurs "théoriques" de : $m^*_y = 4.897$ et $s^*_y = 1.045$

Donc "en gros", les deux estimateurs (empiriques et théoriques) sont voisins pour la moyenne, un peu plus sensibles à l'échantillonnage pour l'écart-type.

b) Cas réels: *(où l'on n'a pas de certitude sur l'appartenance à une loi Lognormale)*

Exemple des pluies annuelles à BILMA (Niger) (données tirées de CHADULE, p. 60)

Année	Pluie annuelle X	Y = Ln X
1941	6	
1942	2	
1943	20	
1944	3	
1945	28	
1946	57	
1947	34	
1948	6	*à compléter*
1949	47	*comme exercice*
1950	40	
1951	9	
1952	17	
1953	54	
1954	21	
1955	13	
1956	9	
1957	9	
1958	10	
1959	15	
1960	4	
1961	16	
1962	25	
1963	17	
1964	39	
1965	9	
1966	15	
1967	8	
1968	27	
1969	14	
1970	7	
Moyenne	19.37	2.6524
Ecart-type	15.16	0.8447

Et les formules théoriques proposent: $m^*_Y = Log\left(\frac{m_X^2}{\sqrt{m_X^2 + s_X^2}}\right) = 2.7248$ à comparer à 2.6524

mais surtout, pour l'écart-type: $s^*_Y = \sqrt{Log\left(1 + \frac{s_X^2}{m_X^2}\right)} = 0.6912$ à comparer à 0.8447

soit une différence supérieure à 25 %...!

II-3) APPLICATIONS à la LOI GAMMA et à ses DERIVEES

II-3-a) Cas de la loi Gamma à ***deux*** paramètres:

On rappelle l'expression de la loi:

$$f(x,\lambda,\rho)=\frac{1}{\Gamma(\lambda)}.e^{-\frac{x}{\rho}}.\left(\frac{x}{\rho}\right)^{\lambda-1}.\frac{1}{\rho}$$

Le calcul des moments, que nous ne détaillerons pas ici, fournit:

$$E[X]=\mu_x=\lambda.\rho \qquad \text{et} \qquad V[X]=\sigma_x^2=\lambda.\rho^2$$

D'où en résolvant les 2 équations:

$$\rho=\frac{\sigma_x^2}{\mu_x} \text{ et} \qquad \lambda=\frac{\mu_x^2}{\sigma_x^2}=\frac{1}{CV^2}$$

avec CV coefficient de variation.

Remarque :
A noter que l'on pourrait calculer des moments d'ordre plus élevé, par exemple:

$$\mu_{3x}=2.\lambda.\rho^3 \quad \text{et} \qquad \mu_{4x}=3.\lambda.(\lambda+2)\rho^4$$

entre lesquels on pourrait facilement aussi calculer λ et ρ.

Mais on sait que plus le moment est *d'ordre élevé*, plus il perd de degrés de liberté, (à cause de la nécessité d'utiliser , pour calculer ces moments d'ordre élevé, des moments d'ordre inférieur !) et donc moins il a de robustesse :

$\Rightarrow$ *on utilisera toujours les moments d'ordre le plus faible possible*.

II-3-b) Cas de la loi Gamma à ***trois*** paramètres:
Dans ce cas, l'expression de la loi devient :

$$f(x,\lambda,\rho,x_0)=\frac{1}{\Gamma(\lambda)}.e^{-\frac{(x-x_0)}{\rho}}.\left(\frac{x-x_0}{\rho}\right)^{\lambda-1}.\frac{1}{\rho}$$

c'est à dire que l'on va chercher à déterminer en plus l'origine x_0 (-s'il n'y a pas de valeur imposée par la "physique" du phénomène-) et il va falloir utiliser un 3éme moment pour trouver aussi x_0.

II-3-c) Utilisation de la ***table de la loi Gamma*** à deux paramètres:

La méthode des moments est la plus couramment utilisée pour la loi gamma, vu sa facilité.

$\Rightarrow$ On calcule la valeur de $\lambda=\frac{1}{CV^2}$.

$\Rightarrow$ On trouve dans la table la colonne correspondant à cette valeur de λ.

Au besoin, si la valeur du λ obtenu ne figure pas dans la table, on interpolera entre deux colonnes.

Pour une valeur de F(u) (par exemple F(u) =.80), et λ = 10 :

- on lit dans la table la valeur u correspondante, soit u = 3,96
- et on repasse à x par x = u. σ_X , associé à F(x) = .80
- d'où un point de la courbe de la loi Gamma théorique qui a les mêmes (deux premiers) moments que l'échantillon, et de même pour les autres points F = 0.1, 0.2, etc...

On porte ensuite les points obtenus sur un diagramme quelconque (il n'existe pas de diagramme "gamma"), souvent un diagramme gausso-arithmétique , comme on le verra en III-6)

On traitera par exemple en exercice une application à des pluies mensuelles qui, pour les mois secs notamment, ont une dissymétrie qui les éloigne de la loi normale.

II-4) APPLICATIONS à la LOI EXPONENTIELLE et à ses DERIVEES

La loi exponentielle est un cas très particulier de la loi Gamma et peut s'écrire indifféremment:

$$F(x) = 1 - e^{-\lambda.x} = 1 - e^{-\frac{x}{\rho}} \qquad \text{ou, en densité de probabilité} \quad f(x) = \lambda.e^{-\lambda.x} = \frac{1}{\rho} e^{-\frac{x}{\rho}}$$

Il y a donc un seul paramètre; on prendra indifféremment λ ou ρ, ce dernier ayant l'avantage d'avoir la même dimension que X.

On montre aisément, par un petit calcul d'intégration (à la portée d'un étudiant de DEUG B…), que le premier moment:

$$\mu_X = E[X] = \int_0^{+\infty} x.\frac{1}{\rho}.e^{-\frac{x}{\rho}}.dx = \rho$$

Un autre petit calcul d'intégration, (à la portée d'un étudiant de classes préparatoires…), montre que le moment centré d'ordre 2 s'écrit:

$$\sigma_X^2 = E\left[(X - \mu_X)^2\right] = \int_0^{+\infty} (x - \mu_X)^2.\frac{1}{\rho}.e^{-\frac{x}{\rho}}.dx = \rho^2$$

Soit encore, pour la loi exponentielle: $\boldsymbol{\sigma_X = \rho}$

$\Rightarrow$ C'est d'ailleurs un moyen de vérifier que la loi est exponentielle:

sa moyenne est égale à l'écart-type.

Naturellement, la méthode des moments utilisera le moment d'ordre le moins élevé et prendra :

$$\rho = m_X$$

car σ_X est plus sensible aux valeurs fortes de l'échantillon, via l'élévation au carré.

On traitera en détail de cette loi sur un exercice (pluies journalières à Seyssel – cf. paragraphe III)

II-5) APPLICATIONS à la LOI de GUMBEL

On rappelle la forme de cette loi:

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

Le calcul des deux premiers moments *théoriques* n'est pas tout à fait évident. (Les amateurs éclairés pourront chercher la démonstration dans la thèse de M. Slimani (1985), vol. d'annexe IA2). Cela fournit:

$$\mu_x = \beta + \alpha.0.577$$

et

$$\sigma_x = \alpha.1.2826$$

d'où les relations entre paramètres et moments:

$$\alpha = 0.78\ .\sigma_x \qquad \text{et} \qquad \beta = \mu_x - 0.577.\alpha$$

Pour les estimer, il suffit de calculer les moments empiriques m_X et σ_X sur un échantillon et d'en tirer a et b.

Rappel: α, ou son estimation a , est appelé **"gradex"** (pour gradient de l'exponentielle), car c'est la quantité (en unité de l'utilisateur) dont augmente x si on augmente d'une unité de Gumbel sur l'axe des probabilités

Un autre ingrédient couramment utilisé est le quantile de probabilité fixée:

- soit une probabilité fixée q , donc telle que $F(x_q) = q$
- alors ce ***quantile*** $\mathbf{x_q}$ est tel que :

$$x_q = \beta + \alpha.\, u_q$$

où u_q est la valeur :

$$u_q = -Ln\left[-Ln(q)\right]$$

ou encore, sur le graphique (cf. parag. III-5), celle qui correspond sur l'axe de Gumbel à la probabilité q.

En fait plutôt que de raisonner en probabilité, on raisonne souvent en *Période de Retour* :

$$T = \frac{1}{1-q} \qquad \text{d'où} \qquad x_q = x_T$$

Exemple: si q = 0.999 , T = 1000, et $x_{.999} = x_{1000}$

III) METHODES GRAPHIQUES:

III-1) Principe :

On part en général des distributions empiriques présentées au chapitre I-2, et plus particulièrement de la *courbe des fréquences cumulées.*

Comme on l'a vu, celle-ci nécessite:

- le classement de l'échantillon

mais surtout ***un choix*** :

- l'*affectation* à chaque individu, à partir de son rang de classement i,
- *d'une* ***probabilité empirique*** *estimée* … $\text{Prob}[X<x_i] \sim f(i, n)$

Or ce choix pose un délicat problème, et impose certaines hypothèses (- notamment sur la fréquence des valeurs les plus fortes et les plus faibles de l'échantillon- cf. biblio du chapitre I , notamment des articles comme celui de NOPHADOL IN-NA and VAN-THANH- VAN NGUYEN (1989))

Cette probabilité empirique s'écrit, pour la valeur de rang i parmi n:

$$\text{Probabilité associée à la valeur de rang i} = P_i = \frac{i-a}{n+b}$$

qui, dans le cas courant de la formule de Hazen devient: $P_i = \frac{i-0.5}{n}$

$\Rightarrow$ On peut alors porter sur un diagramme P_i en fonction de x_i.

On peut constater que l'allure de cette courbe, en diagramme arithmétique (sur du papier millimétré classique), est souvent chaotique à cause du faible échantillonnage $\Rightarrow$ Elle ne permet guère de reconnaître et de distinguer les modèles probabilistes les plus courants, ni de juger de la symétrie. $\Rightarrow$ Pour ce diagnostic, (-c'est à dire pour identifier le type de loi-), on lui préfèrera souvent l'histogramme.

Le principe des méthodes graphiques va donc reposer plutôt sur la Fonction de répartition et sur la construction d'un **diagramme fonctionnel** associé, (-comme on a vu qu'il en existe pour certaines lois au chapitre II).

Cela consiste à **réaliser une anamorphose** (une transformation analytique) telle que:

- dans le nouveau diagramme transformé,
- le modèle considéré, (i.e. la Fonction de Répartition, approchée par la courbe empirique des Fréquences Cumulées) ,
- prendra une allure que l'oeil humain sait reconnaître aisément :

….. **une ligne droite.**

Les axes seront gradués en valeurs arithmétiques usuelles (- entre 0 et 1 pour les probabilités et entre les valeurs admissibles pour la variable X -) *mais* les graduations verront leur écartement et leur progression évoluer selon une fonction particulière.

Si les points empiriques s'alignent *(à peu près...)* correctement sur un diagramme fonctionnel donné,

⇒ c'est que l'échantillon suit *(à peu près...)* le modèle utilisé pour construire ce diagramme,

⇒ et donc que le modèle constitue un compromis acceptable pour représenter / résumer l'échantillon..

La *qualité de l'alignement* constitue de plus *un test* de l'adéquation du modèle à l'échantillon utilisé.

Le mieux est de donner quelques exemples d'usage de ces papiers fonctionnels. On a vu au chapitre II la façon de les construire et un peu de les utiliser. On va rappeler ces possibilités, en *insistant ici sur les côtés pratiques.*

III-2) Le diagramme Gausso-arithmétique (ou papier" normal")

On a vu au chapitre II comment construire un diagramme gaussien, en s'appuyant sur le fait que toute transformation linéaire d'un variable normale reste une variable normale.

On utilisera ce diagramme en exercice dans une application à des pluies annuelles.

Compléments: Comparaison avec l'ajustement obtenu par la méthode des moments

La méthode des moments nous permet de déterminer une loi particulière dans la famille des lois normales, celle qui a même moments m_X et s_X que l'échantillon.
On peut donc la représenter sur papier de Gauss: c'est une droite qui passe:

- par le point d'ordonnée $P = 0,5$ et d'abscisse $x = m_X$

- par les points d'ordonnées $P = 0,1$ et d'abscisse $x = m - 1,28.s_X$
  et $P = 0,9$ et d'abscisse $x = m + 1,28.s_X$

On utilisera d'ailleurs ces propriétés pour déduire *directement* du graphique :

- la moyenne estimée de la population :
  en lisant la valeur de x correspondant à la probabilité 50%
  puisque pour une loi normale $m_x = X_{med} = x_{50\%}$
- et l'écart-type :
  en lisant l'amplitude qu'il y a entre les valeurs de x correspondant aux probabilité 10 et 90%. Or cet intervalle , qui contient 80% des individus, correspond pour la loi normale à 2,56 écart-types.

Exemple :

On montre un petit exemple (en fait litigieux !), tiré de HUBERT P. et H. BENDJOUDI (1998) A propos de la distribution statistique des cumuls pluviométriques annuels : Faut-il en finir avec la normalité ? (*Revue des Sciences de l'Eau)*

Dans cet article un peu provocateur, les auteurs remettent en cause une démarche couramment acceptée (et présentée comme telle au chapitre I) , à savoir que les pluies annuelles suivent une loi normale…

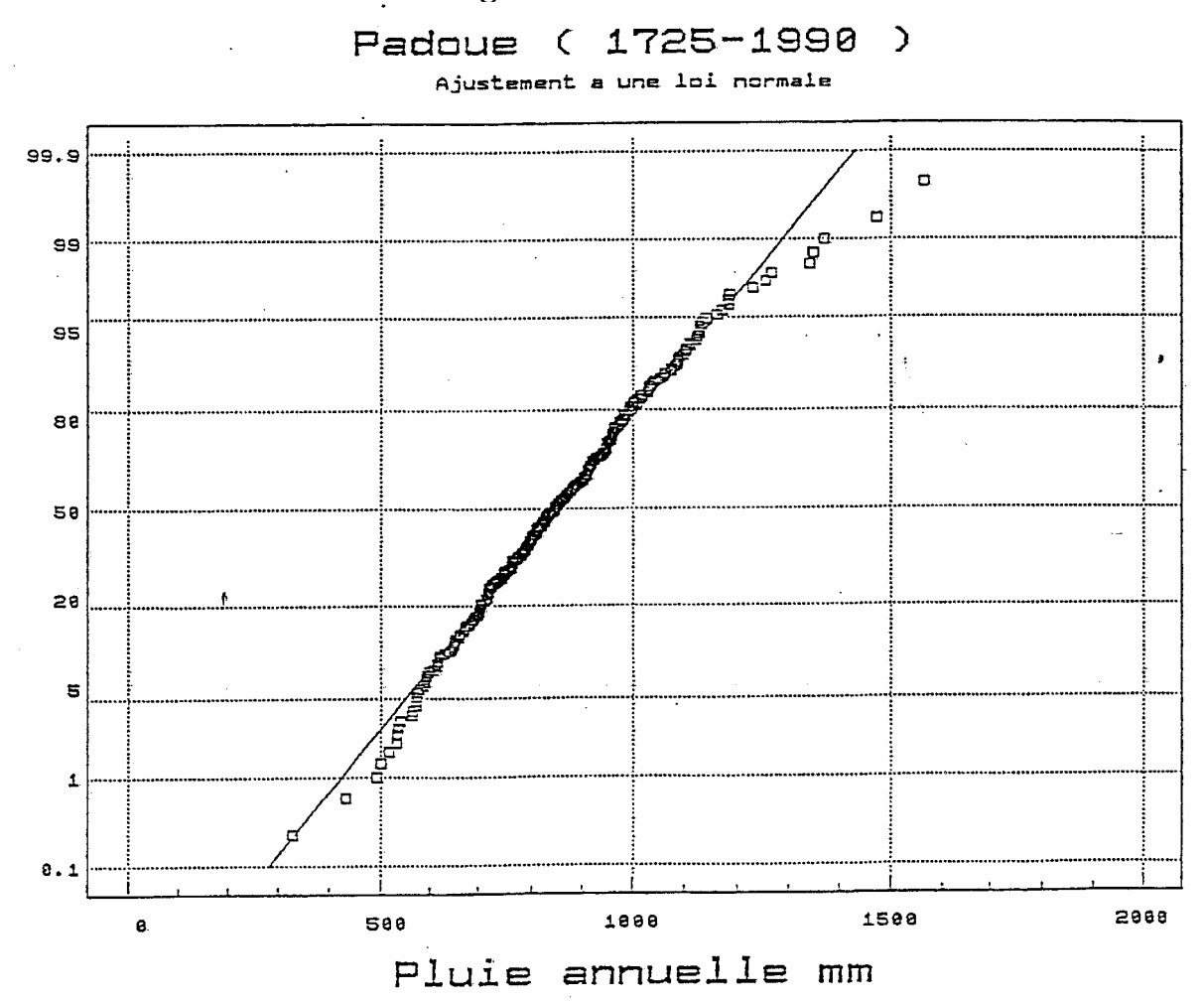
Les arguments sont empiriques (les histogrammes sont à peu près gaussiens, les fréquences cumulée aussi), et théoriques : la pluie annuelle est une variable somme de nombreuses variables (les évènements pluvieux) indépendantes et de même ordre de grandeur…

Or pour les valeurs extrêmes cet argument ne tient plus :

- on a des volumes (-qui ne suivent pas une loi normale-) apportés par certains épisodes pluvieux , qui peuvent à eux seuls représenter plus que la valeurs annuelles courantes
- certaines valeurs annuelles sont donc plus le reflet d'un épisode exceptionnel que d'une moyenne de nombreux épisodes indépendants. Elles suivent plutôt la loi de ces épisodes qu'un loi normale.

Cela peut se voir sur de (vraiment … !) longues séries , comme ici Padoue (Italie) où l'on dispose de 266 ans et où l'on peut admettre que la loi normale convient bien pour des probabilités inférieures à .95, mais s'écarte de la loi empirique au delà.

diagramme PADOUE



III-3) Le diagramme Log-normal (ou papier "Gausso-logarithmique")

On a vu aussi au chapitre II comment construire un diagramme lognormal, en s'appuyant sur le fait que l'on peut revenir à une variable normale en utilisant une échelle logarithmique.

On utilisera ce diagramme en exercice dans une application à des pluies ou à des débits mensuels.

Compléments: *Comparaison avec l'ajustement obtenu par la méthode des moments*

La méthode des moments nous a permis de déterminer la loi normale particulière qui a mêmes moments m_Y et s_Y que **l'échantillon transformé en Logarithme**.

Rappelons que l'on a obtenu:

$$m_Y = Log\left(\frac{m_X^2}{\sqrt{m_X^2 + s_X^2}}\right)$$

et pour l'écart-type

$$s_Y = Log\left(1 + \frac{s_X^2}{m_X^2}\right)$$

On peut donc la représenter sur papier de Gauss (gausso-logarithmique) , **mais avec quelques précautions** :

1) Puisque le diagramme fait lui-même la transformation logarithmique, pour représenter le point moyen : $m_Y = Log\left(\frac{m_X^2}{\sqrt{m_X^2 + s_X^2}}\right)$ qui va correspondre au point d'ordonnée P = 0.5 ,

on portera l'abscisse: $X_{50} = X_{m_Y} = \frac{m_X^2}{\sqrt{m_X^2 + s_X^2}}$

2) C'est un peu plus complexe pour les quantiles, par exemple les déciles correspondant aux points d'ordonnées P = 0,1 et P = 0,9.
En effet, ils ont pour abscisses **en Y** : $Y_{10} = m_Y - 1{,}28.s_Y$ et $Y_{90} = s_Y + 1{,}28.\, s_Y$
On calcule donc ces valeurs:

$$Y_{10} = Log\left(\frac{m_X^2}{\sqrt{m_X^2 + s_X^2}}\right) - 1{,}28.\sqrt{Log\left(1 + \frac{s_X^2}{m_X^2}\right)}$$

et

$$Y_{90} = Log\left(\frac{m_X^2}{\sqrt{m_X^2 + s_X^2}}\right) + 1{,}28.\sqrt{Log\left(1 + \frac{s_X^2}{m_X^2}\right)}$$

et on porte sur le diagramme, pour les points d'ordonnées: P = 0,1 et P = 0,9

les abscisses correspondant à : $X_{10} = e^{Y_{10}}$ et $X_{90} = e^{Y_{90}}$

III-4) Loi exponentielle et diagramme Log-arithmétique

On rappelle l'expression de cette loi

$$F(x) = 1 - e^{-\frac{x}{\rho}} \qquad \text{ou, en densité de probabilité} \quad f(x) = \frac{1}{\rho} e^{-\frac{x}{\rho}}$$

et on cherche une relation entre :

- la probabilité au dépassement (ou au non-dépassement F(x))
- et la valeur x,
- de sorte que cette relation devienne linéaire grâce à une transformation simple.

Ici, on a: $\Pr(X \le x) = F(x, \rho) = 1 - e^{-\frac{x}{\rho}}$

ou $\Pr(X \ge x) = 1 - F(x, \rho) = e^{-\frac{x}{\rho}}$

et en prenant le logarithme des 2 membres et en changeant de signe:

$$\frac{x}{\rho} = -Log[1 - F(x, \rho)]$$

D'où une relation linéaire entre x et le logarithme de la probabilité de *dépassement.*

Or il existe déjà un **diagramme log-arithmétique** : le classique papier " log "
$\Rightarrow$ il suffira de l'adapter à cet usage *probabiliste!*

Utilisation pratique du diagramme:

On prend un diagramme log-arithmétique à deux ou 3 modules.

Pour x = 0 , on a $\Pr[X \le 0] = 0$ car pour cette loi $X \ge 0$
donc : $1 - \Pr[X \le 0] = 1$

et on gradue à la probabilité 1 (*au dépassement*) , le sommet de l'échelle logarithmique. Ensuite on descend et on gradue 0,1 le sommet du module suivant, puis 0.01, et etc... de module en module.

Mais il s'agit de probabilité au dépassement. Donc pour revenir à F(x), la probabilité au *non* dépassement, il suffit d'afficher en face le complément à 1, c'est à dire 0, 0.9, 0.99, etc...
On le voit sur le diagramme associé à l'exemple donné ci après.

On constate aussi que le diagramme dilate les échelles dans les grandes valeurs et écrase dans le premier module 90% des valeurs courantes :
$\Rightarrow$ le tracé va tendre à s'appuyer surtout sur les grandes valeurs...
(d'où un risque de biais …comme dans la méthode des moments)

Exemple: *Traitement graphique d'une loi exponentielle:*

On considère ici (*cf. exercice distribué en cours*), les pluies journalières non nulles à Seyssel (74) sur la période Mai à Septembre inclus de 1919 à 1967, soit 49 ans.

Il y a N = 2268 jours pluvieux sur 7497.

Borne sup. x_i de la classe i (en mm)	Effectif n_i cumulé jusqu'à x_i	$Pr(X \leq x_{i.}) = n_i/N$	$Pr(X \geq x_{i.}) = 1 - n_i/N$
1	181	0.080	0.920
2	445	.196	.804
3	673	.297	.703
4	839	.370	.630
5	1003	.449	.558
6	1148	.506	.494
8	1357	.598	.402
10	1524	.672	.328
15	1805	.796	.204
20	1975	.871	.129
25	2067	.911	.059
30	2143	.945	.055
35	2188	.965	.035
40	2214	.976	.024
45	2235	.985	.015
50	2246	.990	.010
55	2255	.994	.006
60	2260	.996	.004
65	2262	.997	.003
75	2266	.999	.0009
85	2268	1.000	0.0

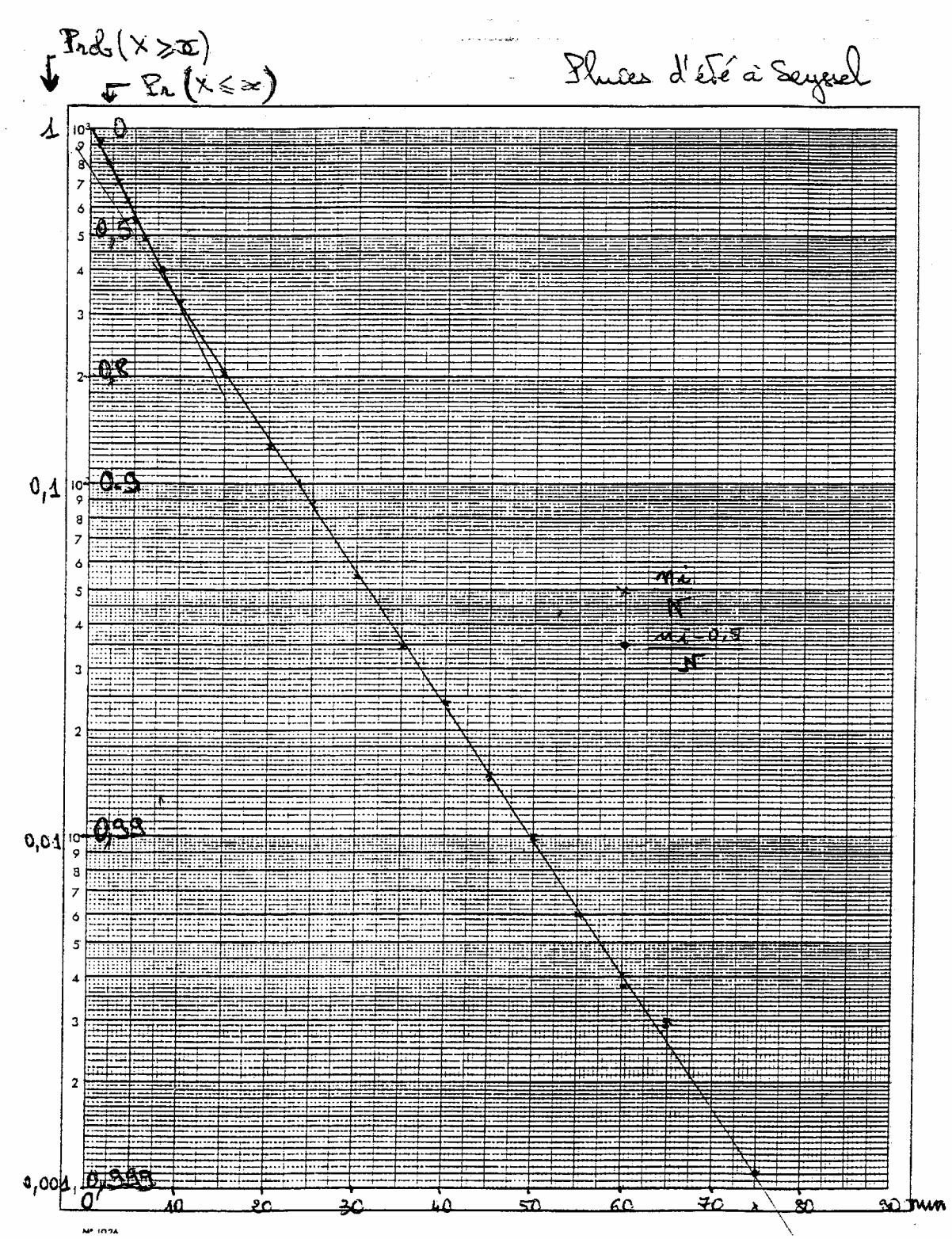
Ici, on a classé toutes les valeurs , mais on ne leur a pas affecté individuellement de probabilité empirique. On est dans un contexte " riche " en données, donc on peut se contenter de faire un découpage en classes assez nombreuses.
Pour chaque borne supérieure de classe b_k, on sait le nombre total d'individus $n_k$contenu dans la classe k et les k-1 qui la précèdent, et on a donc la probabilité estimée :

$$Pr[X < b_k] = (n_1+n_2+\ldots..+n_k) / n$$

Ici, il est inutile de " pinailler " la formule vu le nombre d'individus disponibles.

⇒ On trace le diagramme
et on vérifie, à l'oeil, que les points sont alignés,
ce qui est vrai sauf pour les pluies faibles (< 10 mm).

Distribution des pluies journalières d'été à SEYSSEL(74)



Il faut alors déterminer d'après le diagramme le paramètre ρ.

On peut prendre pour cela 2 points quelconques, i.e. deux valeurs x_1 et x_2 et écrire que :

$$\Pr(X \geq x_1) = . = e^{-\frac{x_1}{\rho}} \qquad \Rightarrow \qquad \frac{x_1}{\rho} = -Log[\Pr(X \geq x_1)]$$

$$\Pr(X \geq x_2) = . = e^{-\frac{x_2}{\rho}} \qquad \Rightarrow \qquad \frac{x_2}{\rho} = -Log[\Pr(X \geq x_2)]$$

D'où par différence:

$$\frac{x_2}{\rho} - \frac{x_1}{\rho} = Log[\Pr(X \geq x_1)] - Log[\Pr(X \geq x_2)] = Log\frac{\Pr(X \geq x_1)}{\Pr(X \geq x_2)}$$

On peut même prendre comme point de départ $x_1 = 0 \Rightarrow \Pr[X \geq 0] = 1$ et il reste alors:

$$\frac{x_2}{\rho} = Log\frac{1}{\Pr(X \geq x_2)} = -Log[\Pr(X \geq x_2)]$$

Si dans cet exemple on prend:

$x_2 = 50$ mm on lit $\Pr[X \geq x_2] \approx 0.01 = 10^{-2}$

$$Log[\Pr(X \geq x_2)] = -2.Log10 \approx -2x2.30 = -4.6 \qquad \text{d'où} \qquad \rho = \frac{50}{4.6} = 10.87mm$$

On comparera cette valeur avec celle trouvée par la méthode des moments: $\rho = 9.59\ mm$ en prenant la moyenne des x_i....

III-5) Le diagramme de Gumbel

(Note: ce paragraphe est emprunté pour partie à M. Paul. Meylan , de l'EPFL Lausanne, que nous remercions vivement)

On a vu au chapitre II pourquoi, en prenant deux fois le Log de F(x), on avait une relation linéaire entre x et cette valeur transformée :

$$u[F(x)] = -Log[-Log\{F(x)\}]$$

On peut se contenter de tracer "à l'oeil" la droite qui passe par les points, mais une méthode astucieuse a été proposée par Gumbel, qui est *recommandée par l'OMM.*

Compléments: *la droite des moindres rectangles:*

L'idée est de trouver une droite (dite des *moindres rectangles*), **bissectrice** des deux droites classiques de la régression de x en y et de y en x (Cf. II^{ème} Partie - Chapitre IV). Celle-ci a l'avantage (contrairement à la régression) de ne pas faire intervenir de produits croisés entre les deux variables considérées.

Dans le cas général de 2 variables y et x , cette droite s'écrit:

$$y = a.x + b \quad \text{avec} \quad a = \frac{s_y}{s_x} \quad \text{et} \quad b = m_y - a.m_x$$

Dans notre cas particulier, les 2 variables sont x et u , d'où:

$$u = a.x + b \quad \text{avec} \quad a = \frac{s_x}{s_u} \quad \text{et} \quad b = m_x - a.m_u$$

L'astuce consiste à remarquer que:

- pour un échantillon de taille n fixé
- les probabilités empiriques P_i sont fixées
- donc aussi les valeurs $u_i = -Ln[-Ln\ P_i]$

et que donc ⇒ - les valeurs de $\mathbf{m_u(n)}$ **et** $\mathbf{s_u(n)}$ *ne changent pas* (une fois la taille de l'échantillon n fixée)

Alors ⇒ - on peut les tabuler une fois pour toutes (cf. Table ci-contre proposée par P. Meylan 1992),

- et calculer aisément le tracé de la droite .

L'utilisation de ce graphique sera vue (et pratiquée !) assez intensivement dans le cours d'Hydrologie Opérationnelle sur l'étude des crues extrêmes et la méthode du Gradex .

III-6) Extensions de l'utilisation des graphiques:

On utilisera parfois certains graphiques, issus d'une loi particulière, pour projeter les observations ou la courbe d'une autre loi.

On utilisera alors seulement le fait qu'il propose une distorsion appréciée dans une partie de l'échelle des fréquences…

Exemple : utilisation du papier de Gauss pour projeter des lois comme les lois Gamma...

Tableau

Caractéristiques de l'échantillon des variables u de Gumbel (Hazen)

n	$\bar{u}(n)$	$s_u(n)$	n	$\bar{u}(n)$	$s_u(n)$	n	$\bar{u}(n)$	$s_u(n)$
6	.5336	1.0909	71	.5731	1.2585	136	.5750	1.2688
7	.5395	1.1141	72	.5732	1.2588	137	.5750	1.2689
8	.5439	1.1318	73	.5732	1.2591	138	.5751	1.2690
9	.5474	1.1459	74	.5733	1.2593	139	.5751	1.2691
10	.5502	1.1574	75	.5733	1.2596	140	.5751	1.2691
11	.5525	1.1670	76	.5734	1.2599	141	.5751	1.2692
12	.5545	1.1750	77	.5734	1.2601	142	.5751	1.2693
13	.5561	1.1820	78	.5735	1.2604	143	.5751	1.2694
14	.5576	1.1880	79	.5735	1.2606	144	.5751	1.2695
15	.5588	1.1933	80	.5735	1.2608	145	.5752	1.2695
16	.5599	1.1980	81	.5736	1.2611	146	.5752	1.2696
17	.5609	1.2022	82	.5736	1.2613	147	.5752	1.2697
18	.5617	1.2059	83	.5737	1.2615	148	.5752	1.2698
19	.5625	1.2093	84	.5737	1.2617	149	.5752	1.2698
20	.5632	1.2124	85	.5738	1.2619	150	.5752	1.2699
21	.5639	1.2152	86	.5738	1.2622	151	.5752	1.2700
22	.5644	1.2178	87	.5738	1.2624	152	.5753	1.2701
23	.5650	1.2201	88	.5739	1.2625	153	.5753	1.2701
24	.5655	1.2223	89	.5739	1.2627	154	.5753	1.2702
25	.5659	1.2244	90	.5739	1.2629	155	.5753	1.2703
26	.5663	1.2262	91	.5740	1.2631	156	.5753	1.2703
27	.5667	1.2280	92	.5740	1.2633	157	.5753	1.2704
28	.5671	1.2296	93	.5740	1.2635	158	.5753	1.2705
29	.5674	1.2312	94	.5741	1.2637	159	.5753	1.2705
30	.5677	1.2326	95	.5741	1.2638	160	.5753	1.2706
31	.5680	1.2340	96	.5741	1.2640	161	.5754	1.2707
32	.5683	1.2353	97	.5742	1.2642	162	.5754	1.2707
33	.5686	1.2365	98	.5742	1.2643	163	.5754	1.2708
34	.5688	1.2376	99	.5742	1.2645	164	.5754	1.2709
35	.5690	1.2387	100	.5743	1.2646	165	.5754	1.2709
36	.5693	1.2397	101	.5743	1.2648	166	.5754	1.2710
37	.5695	1.2407	102	.5743	1.2649	167	.5754	1.2710
38	.5697	1.2417	103	.5743	1.2651	168	.5754	1.2711
39	.5699	1.2425	104	.5744	1.2652	169	.5754	1.2712
40	.5700	1.2434	105	.5744	1.2654	170	.5755	1.2712
41	.5702	1.2442	106	.5744	1.2655	171	.5755	1.2713
42	.5704	1.2450	107	.5745	1.2656	172	.5755	1.2713
43	.5705	1.2457	108	.5745	1.2658	173	.5755	1.2714
44	.5707	1.2464	109	.5745	1.2659	174	.5755	1.2714
45	.5708	1.2471	110	.5745	1.2660	175	.5755	1.2715
46	.5709	1.2478	111	.5745	1.2662	176	.5755	1.2716
47	.5711	1.2484	112	.5746	1.2663	177	.5755	1.2716
48	.5712	1.2490	113	.5746	1.2664	178	.5755	1.2717
49	.5713	1.2496	114	.5746	1.2665	179	.5755	1.2717
50	.5714	1.2501	115	.5746	1.2667	180	.5756	1.2718
51	.5715	1.2507	116	.5747	1.2668	181	.5756	1.2718
52	.5716	1.2512	117	.5747	1.2669	182	.5756	1.2719
53	.5717	1.2517	118	.5747	1.2670	183	.5756	1.2719
54	.5718	1.2522	119	.5747	1.2671	184	.5756	1.2720
55	.5719	1.2527	120	.5747	1.2672	185	.5756	1.2720
56	.5720	1.2531	121	.5748	1.2673	186	.5756	1.2721
57	.5721	1.2536	122	.5748	1.2674	187	.5756	1.2721
58	.5722	1.2540	123	.5748	1.2676	188	.5756	1.2722
59	.5723	1.2544	124	.5748	1.2677	189	.5756	1.2722
60	.5724	1.2548	125	.5748	1.2678	190	.5756	1.2723
61	.5724	1.2552	126	.5749	1.2679	191	.5756	1.2723
62	.5725	1.2556	127	.5749	1.2680	192	.5757	1.2724
63	.5726	1.2559	128	.5749	1.2681	193	.5757	1.2724
64	.5727	1.2563	129	.5749	1.2682	194	.5757	1.2724
65	.5727	1.2566	130	.5749	1.2683	195	.5757	1.2725
66	.5728	1.2570	131	.5749	1.2683	196	.5757	1.2725
67	.5729	1.2573	132	.5750	1.2684	197	.5757	1.2726
68	.5729	1.2576	133	.5750	1.2685	198	.5757	1.2726
69	.5730	1.2579	134	.5750	1.2686	199	.5757	1.2727
70	.5730	1.2582	135	.5750	1.2687	200	.5757	1.2727

IV) METHODE DU MAXIMUM DE VRAISEMBLANCE

IV-1) Principe

On rappelle que si l'on considère la probabilité, en effectuant un tirage au hasard, d'obtenir exactement la valeur x_i :

$\Rightarrow$ cette probabilité est infinitésimale, quasiment nulle..!

Mais si on se fixe un intervalle, une tolérance, de $\pm$ dx/2 , la probabilité d'avoir eu dans l'échantillon une valeur x_i comprise entre $x_i + dx/2$ et $x_i - dx/2$ est, selon la loi définie par sa fonction densité :

$$Pr[x_i - dx/2 < X < x_i + dx/2] = f(x_i, \alpha_1,...., \alpha_p).dx$$

Et si les tirages sont indépendants, donc si les valeurs x_i sont indépendantes:

- la probabilité d'avoir tiré (dans n'importe quel ordre)
- *les n valeurs $x_1 , x_2,..., x_n$* (- à plus ou moins dx/2 -)
- est le *produit* de ces n probabilités, soit:

$$Pr[\{x_1-dx/2 < X < x_1+dx/2 \}\cap\{x_2-dx/2 < X < x_2+dx/2 \}\cap...\cap\{x_n-dx/2 < X < x_n+dx/2 \}]$$

$$= f(x_1, \alpha_1,...., \alpha p).dx \; . \; f(x_2, \alpha_1,...., \alpha p).dx.....f(x_n, \alpha_1,...., \alpha p).dx$$

$\Rightarrow$ c'est donc une fonction des p paramètres $\alpha_1,...., \alpha p$

(- car les n valeurs observées x_i sont alors des données).

La méthode du *Maximum de Vraisemblance* fait alors une hypothèse quasi philosophique:

- si cet échantillon , (- le seul même dont on dispose...)
- est celui qui est apparu,

alors

- c'est qu'il avait une probabilité "forte" d'apparaître: $\Rightarrow$ il était très probable!
- *et même* sans doute parmi les échantillons les plus probables...

Il est donc cohérent que le choix des valeurs des paramètres traduise cette "forte" probabilité..

$\Rightarrow$ La méthode du Maximum de Vraisemblance consiste à choisir les valeurs *estimées* des paramètres $a_1, ….a_p$, de manière à *maximiser cette probabilité*, c'est à dire à rendre cet échantillon observé le plus probable, le plus " vraisemblable " possible, dans le contexte d'une loi choisie au préalable.

La maximisation se fait :

- grâce aux paramètres,
- en annulant les dérivées partielles de la probabilité de l'échantillon.
- ou d'une transformation monotone de cette probabilité.

La résolution de cette maximisation sera d'ailleurs plus ou moins simple selon les lois utilisées... On va en voir quelques exemples sur des lois classiques.

IV-2) Application à la loi de Poisson

Nous donnons un premier exemple avec la loi de Poisson .
La loi de Poisson (cf. Chap. II) est définie pour des valeurs entières positives de X.
Par exemple, en Hydrologie, cette loi donne la probabilité que lors d'une année prise au hasard, il y ait x crues supérieures à une valeur donnée Q0.

Cette loi n'a qu'un paramètre a.

Sa densité (qui est ici une probabilité qu'une valeur, -un nombre entier-, tirée au hasard soit égale à x) est : $\Pr[X = x] = f(x,a) = \frac{a^x}{x!}.e^{-a}$

Soit un échantillon de n valeurs de X: $x_1, x_2, ..., x_n$

La probabilité de l'échantillon, c'est à dire la probabilité que n valeurs de x soient égales à celles de l'échantillon (dans n'importe quel ordre) vaut :

$$P = \Pr\left[1^{er}\ tirage\, de\, X = x_1 \cap 2^{ème}\ tirage\, de\, X = x_2 \cap .... \cap n^{ème}\ tirage\, de\, X = x_n\right]$$

et pour cette loi de Poisson: $P = \frac{a^{\left(x_1 + x_2 + ... + x_n\right)}}{x_1!x_2!...x_n!}.e^{-n.a}$

Maximisons cette probabilité P par rapport au paramètre a :

⇒ pour cela cherchons la valeur de a qui annule la dérivée de P par rapport à a.

En fait, la fonction P(a) aura son maximum pour la même valeur de a que toute transformation de P(a) par une fonction monotone, par exemple Log[P(a)].

⇒ Donc on peut chercher la valeur de a qui annule la dérivée Logarithmique:

$$\frac{d\,LogP}{d\,a} = 0$$

or ici :

$$Log\ P(a) = -n.a + \left(x_1 + x_2 + ... + x_n\right).Log\ a - \mathrm{Log}\left(x_1!x_2!...x_n!\right)$$

et $\frac{d\,LogP}{d\,a} = -n + \frac{1}{a}\sum_{i=1}^{n} x_i$

Donc la condition: $\frac{d\,LogP}{d\,a} = 0$ fournit: $a = \frac{1}{n}\sum_{i=1}^{n} x_i = m_x$

⇒ la méthode du Maximum de Vraisemblance propose:

- d'ajuster le paramètre a d'une loi de Poisson
- en le prenant égal à la moyenne de l'échantillon.

Note :
Dans ce cas, pour cette loi particulière, le résultat est le même que par la méthode des moments…

IV-3) Application à la loi Normale:

Celle-ci a pour expression $f(x,\alpha,\beta) = \frac{1}{\alpha.\sqrt{2\pi}}.e^{-\frac{1}{2}.\left(\frac{x-\beta}{\alpha}\right)^2}$

et la méthode du Maximum de Vraisemblance va consister à trouver les valeurs a et b de α et β qui permettent de maximiser la probabilité de l'échantillon, c'est à dire de maximiser le produit :

$$P = f(x1, a\ b)*f(x2, a, b)*...*f(xn, a, b)$$

Avec l'expression précédente de la densité de la loi Normale, celle ci s'exprime:

$$P = \frac{1}{\alpha^n}.\frac{1}{2\pi^{\frac{n}{2}}}.e^{-\frac{1}{2}.\sum_{i=1}^{n}\left(\frac{x_i-\beta}{\alpha}\right)^2}$$

Minimiser P revient au même que minimiser Log P, et la dérivation par rapport aux deux paramètres: $\frac{\partial LogP}{\partial a} = 0$ et $\frac{\partial LogP}{\partial \beta} = 0$ fournit les deux équations:

$$\frac{\partial\ LogP}{\partial\ \beta} = -\frac{1}{2}\sum_{i=1}^{n} 2.(x_i - \beta).(-1) = 0 \qquad \text{soit} \qquad b = \frac{1}{n}.\sum_{i=1}^{n} x_i = m_x$$

et $$\frac{\partial\ LogP}{\partial\ a} = -n.\frac{1}{\alpha} - \frac{1}{2}\sum_{i=1}^{n} 2.\left(\frac{x_i - m_x}{\alpha}\right).\left(-\frac{1}{\alpha^2}\right) = 0$$

d'où à l'optimum la valeur: $a^2 = \frac{1}{n}.\sum_{i=1}^{n}(x_i - m_x)^2$ ou $a = s_x$

Note :

$\Rightarrow$ on retrouve ici aussi les résultats obtenus précédemment par la méthode des moments...!

IV-4) Application à la loi exponentielle

Ici aussi le résultat va être trivial.

Avec la densité de probabilité : $f(x) = \lambda.e^{-\lambda.x} = \frac{1}{\rho}e^{-\frac{x}{\rho}}$

$$Log\ P(\rho) = n.Log\frac{1}{\rho} - \sum_{i=1}^{n}\frac{x_i}{\rho} = -n.Log\rho - \frac{1}{\rho}\left(\sum_{i=1}^{n} x_i\right)$$

et

$\frac{\partial LogP}{\partial \rho} = 0$ fournit l'estimation : $\rho = \frac{1}{n}\sum_{i=1}^{n} x_i$

Note :

$\Rightarrow$ et on retrouve ici encore les résultats obtenus précédemment par la méthode des moments...!

Par contre pour des lois assez proches comme les lois Gamma, ce sera beaucoup plus compliqué (cf. Haan 1977 par exemple). Nous en donnerons un seul exemple : celui de la loi de Gumbel.

IV-5) Application à la loi de Gumbel

Ici, le résultat ne sera pas trivial... Et comme il est assez couramment utilisé, on en propose une démonstration un peu détaillée(*).

Cette loi a pour expression $F(x,\alpha,\beta) = e^{-e^{-\left(\frac{x-\beta}{\alpha}\right)}}$

et pour densité de probabilité $f(x,\alpha,\beta) = \frac{1}{\alpha}.e^{-\left(\frac{x-\beta}{\alpha}\right)}.e^{-e^{-\left(\frac{x-\beta}{\alpha}\right)}}$

et la méthode du Maximum de Vraisemblance va consister à trouver les valeurs a et b de α et β qui permettent de maximiser la probabilité de l'échantillon, c'est à dire de maximiser le produit :

$$P = f(x_1, a, b)*f(x_2, a, b)*...*f(xn, a, b)$$

Avec l'expression précédente de la densité, celle ci s'exprime:

$$P = \frac{1}{\alpha^n}.\prod_{i=1}^{n} e^{-\left(\frac{x_i-\beta}{\alpha}\right)}.e^{-e^{-\left(\frac{x_i-\beta}{\alpha}\right)}}$$

Minimiser P revient au même que minimiser Log P:

$$Log\ P = -n.Log\alpha - \sum_{i=1}^{n}\left(\frac{x_i-\beta}{\alpha}\right) - \sum_{i=1}^{n} e^{-\left(\frac{x_i-\beta}{\alpha}\right)}$$

et la dérivation par rapport au deux paramètres : $\frac{\partial LogP}{\partial a} = 0$ et $\frac{\partial LogP}{\partial \beta} = 0$ fournit les deux équations:

$$\frac{\partial\ LogP}{\partial\ \beta} = 0 - \sum_{i+1}^{n}\left(-\frac{1}{\alpha}\right) - \sum_{i=1}^{n} e^{-\left(\frac{x_i-\beta}{\alpha}\right)}.\frac{1}{\alpha} = 0$$

soit $\frac{n}{\alpha} - \frac{1}{\alpha}.\sum_{i=1}^{n} e^{-\left(\frac{x_i-\beta}{\alpha}\right)} = 0$ ou encore: $n = \sum_{i=1}^{n} e^{-\left(\frac{x_i-\beta}{\alpha}\right)}$

De même:

$$\frac{\partial\, LogP}{\partial\, a} = -n.\frac{1}{\alpha} + 0 \quad \text{d'où} \quad \frac{1}{\alpha^2}.\sum_{i=1}^{n}(x_i-\beta) + \sum_{i=1}^{n} e^{-\left(\frac{x_i-\beta}{\alpha}\right)}.\left(-\frac{x_i-\beta}{\alpha^2}\right) = 0$$

Et on obtient finalement l'ensemble de deux équations, donnant les estimations a et b de α et β:

$$a = m_x - \frac{\sum_{i=1}^{n} x_i.e^{-\frac{x_i}{a}}}{\sum_{i=1}^{n} e^{-\frac{x_i}{a}}}.$$

$$b = a\,.\,Log\left(\frac{n}{\sum_{i=1}^{n} e^{-\frac{x_i}{a}}}\right)$$

La résolution de ce système d'équations dépend surtout de la première, qu'il faut résoudre de manière *itérative*:

- on prend comme valeur initiale de a , soit $a_{(0)}$, la valeur proposée par la méthode des moments
- puis on itère, passant de $a_{(k)}$ à $a_{(k+1)}$.
- Kimball, cité par Gumbel (1958), propose pour accélérer la convergence, d'utiliser à l'issue de l'itération la formule suivante:

$$a^*_{(k+1)} = \frac{1}{\frac{1}{a_{(k+1)}} + \frac{\frac{1}{a_{(k)}} - \frac{1}{a_{(k+1)}}}{3}}$$

Enfin, lorsque la taille n de l'échantillon est faible, on démontre que ces valeurs sont *biaisées* d'où une correction proposée par Fiorentino et Gabriele (1984):

$$\hat{a} = \frac{a}{1. - 0.8 \,.\, n} \qquad \text{et} \qquad \hat{b} = \hat{a} \,.\, Log\left(\frac{n}{\sum_{i=1}^{n} e^{-\frac{x_i}{\hat{a}}}} \right) - 0.7.\frac{\hat{a}}{n}$$

Cette correction tend à limiter la sous-estimation systématique du *gradex* a par cette méthode du maximum de vraisemblance.

On verra au paragraphe VI-1 un aperçu de méthodes plus avancées encore pour estimer ces paramètres...

V- TESTS D'HYPOTHESE

Le seul problème que nous avons résolu jusqu'à présent, en donnant parfois plusieurs solutions, est de:

- choisir arbitrairement (-au vu de l'histogramme par exemple-),
  une *famille* de lois (- par exemple la famille des lois Gamma incomplètes -)
- puis de trouver parmi cette famille,
  l'individu , i.e. la ou les lois s'ajustant au mieux à l'échantillon présenté
- au vu d'un ou plusieurs *critères* (égalité des Moments, max. de vraisemblance,…).

Mais il se peut que cet échantillon puisse être relativement bien décrit par cette loi (- sans qu'il en soit issu...! -), et peut-être encore mieux par une autre...! A ce stade de l'analyse, nous avons donc besoin d'un *outil d'évaluation et de comparaison*; c'est pourquoi nous allons tenter de répondre à cette question dans la suite du chapitre.

V-1) Objectif :

On possède une série de n valeurs $\{x_i , i = 1 \text{ à } n \}$ et on veut infirmer ou confirmer *l'hypothèse* suivante : cet échantillon peut raisonnablement être considéré comme tiré d'une population ayant une certaine distribution de probabilité que l'on précise a priori.

Si c'est le cas, (si l'hypothèse est vérifiée) cette distribution de probabilité pourra être facilement utilisée pour calculer des valeurs x de probabilité donnée ou les probabilités associées à des valeurs de x fixées.

Mais attention :

- on ne pourra jamais prouver que cette hypothèse est exacte.. ! (- au mieux on pourra donner une idée de la vraisemblance de l'hypothèse….-)

- il faudra se méfier des extrapolations … ! (- valeurs de probabilité au non dépassement très faibles ou très fortes -) si l'hypothèse est retenue.

- il sera bien souvent possible de trouver plusieurs lois de probabilité assez classiques pour lesquelles l'hypothèse " que l'échantillon pourrait en être issu " soit raisonnablement acceptable pour ces différentes lois … (- sans pouvoir en choisir une plutôt qu'une autre…-)

Exemple : A partir d'un échantillon de *40 valeurs des pluies mensuelles* de Novembre à Gap, on a accepté l'hypothèse d'appartenance à la loi Gamma incomplète de moyenne 142 mm et d'écart type 105 mm. A partir d'une table de la loi Gamma incomplète, on en tire que la pluie d'un mois de Novembre d'une année tirée au hasard (par exemple, l'an prochain) a 90% de chances de dépasser 30 mm et 95 % de chance d'être inférieure à 340 mm.

Mais il serait aventureux de calculer à partir d'une table de la loi Gamma incomplète (et d'un échantillon de 40 valeurs…) la pluie qui n'a qu'une chance sur 1000 d'être dépassée en Novembre...

Méthodologie pour effectuer un test d'hypothèse :

On peut proposer l'organigramme suivant :

a) Collecte de données: $\Rightarrow$ Echantillon $\{x_i , i = 1 \text{ à } n\}$

b) Critique des données (cf. . Troisième Partie)

c) Tracé de la fonction de répartition empirique (aide au choix de la loi)

d) Choix d'une famille de loi (exemple : famille Gamma Incomplète)

e) Calcul des paramètres de la loi dans la famille retenue par une méthode de calage (Moments, Maximum de Vraisemblance ...), la méthode dépend parfois de la loi retenue.

f) *Choix d'un test d'ajustement (Test du Chi2, Kolmogorov ...)*

g) *Réponse du test :*

*Rejet de la loi et autre recherche (d'où retour en **d)** ,*

ou

Acceptation de la loi

Nous ne présenterons, à titre d'exemple, que quelques tests d'ajustement:

V-2) Test du Chi 2 (χ^2):

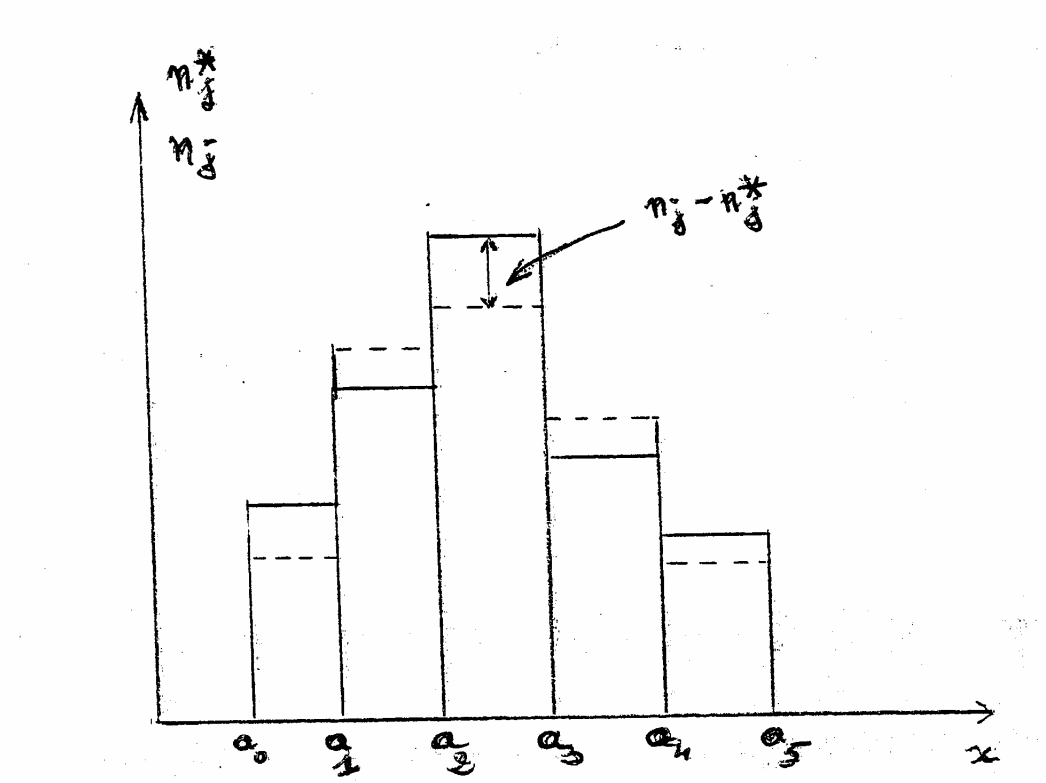
Le test du Chi 2 consiste à comparer un histogramme empirique (c'est à dire défini par les données de l'échantillon) à l'histogramme que donnerait la loi à tester. Nous voyons déjà qu'il nous faut définir l'histogramme par un choix de nombre de classes et de position de classes.

Soit $a_0, a_1 , a_2,...,a_j , ...a_{k+1}$ les limites de classes $C_{j+1}=[a_j , a_{j+1}]$ avec $a_j < a_{j+1}$

x_i appartient à la classe j si x_i est compris entre a_j et a_{j+1} avec égalité admise avec a_{j+1} ; d'où un comptage, nous permettant de définir les effectifs "empiriques" n_j de chaque classe j.

Si F(x) est la fonction de répartition à tester et n le nombre total de données, l'effectif n_j^* que donnerait cette loi pour la classe j serait :

$$n_j^* = n . [F(a_{j+1}) - F(a_j)]$$



On définit alors une distance entre ces deux histogrammes :

$$D = \sum_{j=1}^{k} \frac{(n_j - n_j^*)^2}{n_j^*} = \chi^2$$

Cette distance est d'autant plus grande que l'écart est grand entre les deux distributions. Elle n'est nulle que par le plus grand des hasards même si l'échantillon appartient à la loi (effet d'échantillonnage). Par ailleurs n^*_j n'est pas toujours entier.

⇒ Comment savoir si la distance calculée est vraiment trop grande ?
Et si cela est le cas, il faut alors rejeter l'hypothèse
d'appartenance de l'échantillon à cette loi particulière.

On démontre donc que, sous certaines hypothèses, D suit une loi de probabilité appelée la loi du Chi 2, loi à un seul paramètre qui, dans cette application de test, vaut :

$$n = k - 1 - p$$

où p est le *nombre de paramètres estimés* à partir de l'échantillon pour caler la loi
(*Exemple* p = 1 pour une loi de Poisson mais 2 pour une loi Gamma Incomplète).

Ce paramètre n est appelé nombre de ***degrés de liberté***.

Il sera alors possible de lire dans une *table du Chi 2*, la probabilité de dépasser la valeur de D_0 si l'hypothèse est exacte $Pr[D > D_0]$
Si cette probabilité au dépassement est faible : i.e. si la valeur obtenue pour D, soit D_0 , avait a priori peu de chance d'être atteinte ou dépassée
⇒ il peut être conseillé de rejeter l'hypothèse.

Conseils :

1)- Choix des classes :

Il est conseillé de définir des classes *équiprobables* selon la fonction F(x).

D'où pour k classes : on calcule la borne aj par :

$$F(aj) = \frac{j-1}{k}$$

2)- Nombre de classes :

Il est souhaitable que $n^*_j > 5$ ce qui détermine le nombre maximum de classes.

Remarques :

+ Choix du seuil de rejet de l'hypothèse :

Comme pour les tests d'homogénéité, cela dépend du problème et du coût des erreurs (ici 2 types d'erreurs sont possibles : accepter l'hypothèse alors qu'elle est fausse ou rejeter l'hypothèse alors qu'elle est vraie).

Une valeur classique de seuil de rejet utilisé en Hydrologie est de l'ordre de 10 ou 5 %, probabilité au dépassement de la valeur calculée du Chi 2.

+ Réponse du test :

Soit pour une définition donnée de classes, la valeur D de la distance du Chi 2. La loi du Chi 2 à k - 1 - p degrés de liberté donne une valeur P de Probabilité au non dépassement. Cela signifie que :

Si l'hypothèse d'appartenance de l'échantillon à la population définie par la fonction de répartition F(x) était exacte, il y aurait une probabilité P de trouver une distance supérieure ou égale à D. Autrement dit, si on se fixe un seuil S de rejet, et, si on testait un très grand nombre N d'échantillons réellement tirés de la loi F(x), on serait amené à en rejeter approximativement S*N.

La réponse est donc surtout intéressante si elle nous amène à rejeter nettement l'hypothèse. L'acceptation de l'hypothèse (cas où le Chi 2 est petit) nous dit simplement que l'échantillon présente un histogramme empirique qui n'est pas incompatible avec celui de la loi F(x); mais cela ne prouve pas que l'échantillon est effectivement tiré de cette loi….

Exemple d'Application complète :

Nous allons analyser les débits d'Octobre de 1913 à 1962 de la Loire à Blois (cf. tableau, les débits sont en m^3/s).

an	Q	an	Q	an	Q	an	Q	an	Q	an	Q
1913	425	1914	149	1915	120	1916	291	1917	187	1918	141
1919	85	1920	439	1921	52	1922	147	1923	119	1924	281
1925	125	1926	57	1927	239	1928	82	1929	120	1930	441
1931	143	1932	289	1933	590	1934	65	1935	214	1936	136
1937	92	1938	194	1939	358	1940	444	1941	125	1942	81
1943	333	1944	505	1945	54	1946	54	1947	36	1948	74
1949	30	1950	49	1951	107	1952	203	1953	131	1954	136
1955	58	1956	367	1957	59	1958	254	1959	73	1960	562
1961	74	1962	47								

Hypothèse I (à tester… !) : "ils sont tirés d'une loi Normale"

Etape 1 : Calage des 2 paramètres de la loi Normale.

Pour une loi Normale, la méthode des Moments et la méthode du Maximum de Vraisemblance donnent les mêmes résultats.

Dans cet exemple, les 2 paramètres sont la moyenne 188.5 m^3/s et l'écart type 150m^3/s.

Etape 2 : Calcul d'une distance Chi 2 :

Suivant les conseils précédents, on va prendre 8 classes équiprobables au sens de la loi Normale de moyenne et écart type égaux à ceux des données.

D'où les limites de classes aj, telles que
$F(a_j) = (j-1)/8$ et $n^*j = n/8 = 6.25$ (au passage notons que le Chi 2 ne pourra jamais être nul puisque le nombre d'individus par classe sera évidemment entier avec l'échantillon!).

$a_1 = - \infty$ $\quad a9 = + \infty$

Calculons par exemple a_2 : $F(a_2) = 1/8 = .125$

On trouve que dans une loi Normale centrée réduite, si $F(u) = .125$ $u = - 1.15$

d'où a_2 = Moyenne + (Ecart Type)*(-1.15), soit a2=16 m^3/s. On calcule ainsi toutes les autres bornes, on compte les effectifs empiriques par classes et on calcule le Chi 2; ce qui donne le tableau suivant :

Limites de classes		Effectifs n_j observés	Effectifs n^*_j dans la loi	$(n_j-n^*_j)^2$
- ∞	16	0	6.25	39
16	87.3	17	6.25	115.6
87.3	140.7	10	6.25	14.06
140.	188.5	5	6.25	1.56
188.	236.3	3	6.25	10.56
236.	289.7	4	6.25	5.06
289.	361	3	6.25	10.56
361	+ ∞	8	6.25	3.06

d'où Chi 2 = $\boldsymbol{D} = \sum_{j=1}^{8} \frac{\left(n_j - n_{ji}^*\right)^2}{n_j^*} = \chi^2$ = 31.9

Le nombre de paramètres estimé pour caler la loi Normale est de 2, le nombre de classes est de 8, d'où D, si l'hypothèse est exacte, suit une loi du Chi 2 à 8-1-2 = 5 degrés de liberté.

La probabilité de dépasser 31.9 dans une loi du Chi2 à 5 degrés de liberté est infime (de l'ordre de .000004) ⇒ d'où rejet de l'hypothèse de la loi Normale.

Etape 3 : changement d'hypothèse….!

Hypothèse II (toujours à tester…) : "ils sont tirés d'une loi Log-Normale (loi de Galton)"

La loi Log-Normale est la loi Normale après transformation logarithmique de la variable.

C'est grâce à l'allure de la distribution de l'échantillon et au fait que les débits sont plutôt le résultat de produits de variables (pluie par coefficient d'écoulement) que l'on tente cette hypothèse.

Reprenons l'étape précédente mais en travaillant sur les logarithmes des débits exprimés en m^3/s. D'où 2 paramètres à estimer pour la loi Normale sur les Log : la moyenne des Log 4.92 et l'écart type des Log .78.
Le tableau précédent est modifié de la façon suivante (les bornes sont exprimés en Log):

Limites de classes		nj	n*j	$(n_j-n^*_j)^2$
- ∞	4.03	7	6.25	.87
4.03	4.39	8	6.25	3.06
4.41	4.69	4	6.25	5.06
4.69	4.94	8	6.25	3.06
4.94	5.19	4	6.25	5.06
5.19	5.47	4	6.25	5.06
5.47	5.85	6	6.25	.06
5.85	+ ∞	9	6.25	7.56

d'où Chi 2 = D = 4.7 avec 5 degrés de liberté.

La probabilité de dépasser 4.7 dans une loi du Chi 2 à 5 degrés de liberté est de 67%. Autrement dit, si l'hypothèse d'appartenance de l'échantillon à une loi Log-Normale était vraie, il y aurait 67 % de chances de dépasser cette valeur, probabilité très élevée.

⇒Il n'y a donc pas lieu de rejeter l'hypothèse à partir de cette réponse du test.

Remarque I :

Si l'on avait fait le choix d'une loi Gamma Incomplète sur les mêmes données, le test du Chi 2 ne l'aurait pas rejetée non plus...

Donc le test ne choisit pas à votre place! Il donne des indications pour que ***vous*** acceptiez ou rejetiez votre hypothèse...

Remarque II : *Importance du choix de la loi sur cet exemple :*

Si l'on veut calculer l'étiage décennal, c'est à dire le débit que l'on a 9 chances sur 10 de dépasser, la loi Normale nous aurait donné:

-3.5 m3/s ....!

alors que la loi Normale sur les Log propose : 50 m3/s , ce qui parait plus correct….

V-3) Test de Kolmogorov Smirnoff :

Le principe général est le même mais la distance entre la distribution de l'échantillon et la distribution F(x) est définie comme le plus grand écart (en valeur absolue) entre $F(x_i)$ et $F^*(x_i)$:

$$D = Max\left|F(x_i) - F^*(x_i)\right|$$

On montre alors que dans le cas où l'échantillon est tiré de la loi F(x), cette distance D suit une loi de Probabilité dite de Kolmogoroff Smirnoff à un paramètre k égal au nombre n de données. On trouvera en annexe une table de la loi de Kolmogoroff-Smirnoff.

Exemple : *Application pratique sur les données précédentes :*

Pour la loi Normale, on trouve D = 15%.
Pour la loi Log-Normale, D = 6%
Or dans une table de Kolmogoroff Smirnoff, on trouve que dépasser D = .06 avec n = 50 a une probabilité bien supérieure à 20%; ⇒ on n'est donc pas tenté de rejeter l'hypothèse, puisque si elle était vraie, on aurait plus d'une chance sur 5 d'avoir D au moins aussi grand.

Pour D =15%, cette probabilité de dépasser D dans l'hypothèse d'une loi normale est plus réduite mais reste vraisemblable (d'où une réponse différente de celle du test du Chi2... : ici on accepterait les deux lois)

Conclusions sur les tests d'ajustement :

Nous n'avons présenté que 2 tests… or il en existe d'autres. Il faut donc retenir qu'ils ne sont qu'une aide, parmi d'autres, au choix des lois, mais qu'ils ne sont pas une arme absolue.

Au cours des séances de Travaux Dirigés, on pourra s'en rendre compte en travaillant sur des échantillons dont l'origine est garantie (parce qu'on les a fabriqués, par génération stochastique). Il arrivera sur ces exemples que l'on soit tenté de rejeter l'hypothèse alors qu'elle est vraie et inversement, ou que l'on hésite entre plusieurs lois.
Dans la pratique, l'expérience de l'analyste (compter 1à à 15 ans) tranchera parfois le débat !.

Le problème le plus délicat restera l'ajustement des valeurs extrêmes: selon la loi choisie, les résultats peuvent différer énormément (- dès que l'on travaille dans des probabilités faibles au non dépassement -) , or cela pourra avoir une incidence économique considérable…

VI- COMPLEMENTS THEORIQUES (*)

VI-1) La méthode des Moments Pondérés:

VI-1-a) *Préambule :*

Il faut bien comprendre que les ajustements que nous allons réaliser vont ensuite être utilisés pour prendre des décisions aux conséquences économiques significatives. C'est surtout vrai dans le domaine de la sécurité, d'où les recherches, et parfois les polémiques, à propos de ces méthodes de décision.

La méthode des moments " classique " présente un certain nombre d'inconvénients, notamment lorsqu'une donnée se trouve loin de la moyenne. En effet, dès le moment d'ordre 2, celui-ci va être très influencé par cette donnée surtout si l'échantillon est de petite taille.

Le terme isolé $\frac{1}{n}.(x_i - m_X)^2$ peut alors prendre un poids considérable.

Un autre inconvénient est que dans les estimateurs, de la moyenne ou de la variance, par exemple :

$$s_X^2 = \sum_{i=1}^{n} \frac{1}{n}.(x_i - m_X)^2 \quad \text{, censé estimer } \sigma_X^2 = \int_{-\infty}^{+\infty} (x - \mu_X)^2 . f(x).dx$$

on fait apparaître partout la même quantité 1/n à la place de $f(x_i).\Delta x_i$, dont on sent bien qu'elle ne devrait pas être la même partout.

Dit autrement, on pressent que les valeurs extrêmes de l'échantillon, bien que réellement observées, n'ont pas la même probabilité d'apparaître que des valeurs plus courantes.

On a déjà abordé cette difficulté dans les méthodes graphiques via la *probabilité empirique* associée ("plotting position").

Mais ici, on devrait travailler sur la densité de probabilité, ce qui est plus délicat que la probabilité au non dépassement. On va donc essayer d'y revenir, via une nouvelle sorte de moments.

VI-1-b) *Définition succincte des Moments Pondérés (par les probabilités):*

Comme dans la méthode des moments classiques, on va :

- définir des quantités que l'on peut exprimer *théoriquement* à l'aide des *paramètres* de la loi analytique
- et que l'on pourra estimer facilement à l'aide des *données* mais en faisant apparaître la *probabilité associée* à ces données.

Cela s'applique particulièrement à des ***lois facilement inversibles***, où l'on peut facilement exprimer x en fonction de F(x). Cette méthode a été introduite par Greenwood et al.(1979)

On définit alors des moments à 3 indices:

$$M_{l,j,k} = E\left[x^l . F^j(x).\{1 - F(x)\}^k\right] = \int_0^1 x(F)^l . F(x)^j . \{1 - F(x)\}^k . dF$$

où j, k, l peuvent être des réels quelconques...

On vérifie que si $j = k = 0$ *et l entier,*

alors on retombe sur les moments classiques, non centrés, d'ordre *l*.

En pratique, on n'utilisera que des $M_{1j\,0}$ ou $M_{10\,k}$.

Et toute l'astuce consistera:

- d'une part, à exprimer ces intégrales en fonction des paramètres
- d'autre part à les estimer numériquement à l'aide des données, c'est à dire des x_i *mais aussi* des $F(x_i)$ …!

On va le voir sur l'exemple de la loi de Gumbel (sachant que Masson et Lubes l'ont décrite par ailleurs pour la loi plus générale de Jenkinson 1991)

VI-1-c) *Application à la* ***loi de Gumbel:***

Dans ce cas, la loi de Gumbel:

$$F(x,\alpha,\beta) = e^{-e^{-.\left(\frac{x-\beta}{\alpha}\right)}} \quad \text{s'inverse en} \quad x(F) = \beta + \alpha.u_F = \beta + \alpha.\left[-Log\left(-LogF\right)\right]$$

et les moments pondérés par les probabilités M_{1j0} s'écrivent :

$$M_{1j0} = \int_0^1 x(F).F(x)^{j}.\mathrm{dF} = \int_0^1 \{\beta + \alpha.\left[-Log\left(-LogF\right)\right]\}.\ F^{j}.dF$$

L'intégration (-assez laborieuse pour le terme en -Log(-Log)...), fournit:

$$M_{1j0} = .\ \frac{\beta}{1+j} + \alpha.\frac{Log(1+j) + 0.5772}{1+j}$$

Si on calcule les deux moments d'ordre le plus bas, pour j = 0 et j = 1 , on trouve:

$$M_{100} = \beta + 0.5772.\alpha$$

et

$$2.M_{110} = \beta + \alpha.(Log2 + 0.5772)$$

Ayant ces deux relations entre deux moments et les paramètres α et β, on en tire aisément:

$$\alpha = \frac{2M_{110} - M_{100}}{Log2} \qquad \text{et} \qquad \beta = M_{100} - 0.5772.\alpha$$

Il reste à trouver une estimation empirique des M_{1j0}

$\Rightarrow$ Pour cela, il suffit de remplacer l'intégrale par une somme:

$$M_{1j0} = \int_0^1 x(F).F(x)^{j}.\mathrm{dF} \Rightarrow \hat{M}_{1j0} = \frac{1}{n}\sum_{i=1}^{n} x_i.\hat{F}(x_i)^{j}$$

où $\hat{F}(x_i)$ est la *probabilité empirique* associée à x_i :

$\Rightarrow$ par exemple $\hat{F}(x_i) = \frac{\text{i - 0.5}}{\text{n}}$ ou $\frac{\text{i - 0.35}}{\text{n}}$ selon le choix de l'utilisateur.

On notera que, pour j = 0: $\hat{\text{M}}_{100} = \frac{1}{n}\sum_{i=1}^{n} x_i = m_x$

et pour j = 1 $\hat{\text{M}}_{110} = \frac{1}{n}\sum_{i=1}^{n} x_i \cdot \hat{F}\left(x_{\text{i}}\right)_x$

d'où:

$$a = \frac{2\hat{M}_{110} - m_x}{Log2} \qquad \text{et} \qquad b = m_x - 0.5772.a$$

Note : Le gros avantages de cette méthode est que les valeurs observées de l'échantillon ne sont plus élevées à une puissance autre que 1 …

VI-1-d) *Exemple numérique:*

Dans cet exemple (cf. page ci-contre), on a pris ***deux fois le même échantillon***, ***sauf*** que l'on a changé une valeur et une seule, celle du maximum observé (dans l'échantillon).

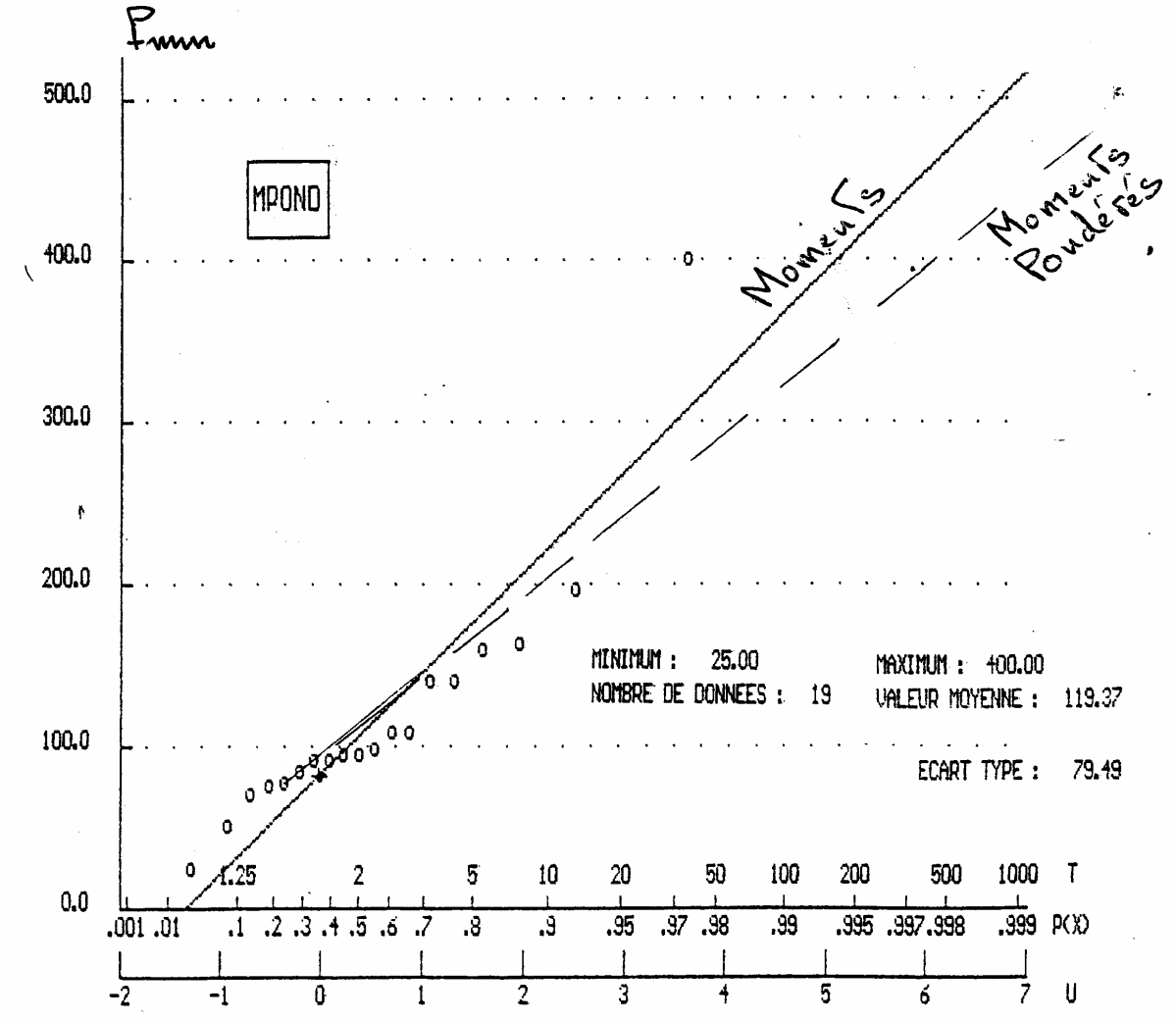
On a voulu montrer que l'estimation de la loi de Gumbel par la méthode des moments classiques était *sensible* et que l'idée globale que l'on se faisait de la distribution changeait nettement à cause de cette seule valeur.

Par contre, l'estimation par la méthode des moments pondérés est beaucoup moins sensible.

Evidemment, dans le cas où il n'y a pas de points trop excentrés dans l'échantillon, les deux méthodes donnent des résultats voisins (cas des données réelles testées ici).

i	F(i)=(i-0.35)/n	x(i)	F(i)*x(i)	données réelles
1	0.034	25	0.855	25
2	0.087	50	4.342	50
3	0.139	70	9.763	70
4	0.192	75	14.408	75
5	0.245	77	18.845	77
6	0.297	84	24.979	84
7	0.350	91	31.850	91
8	0.403	91	36.639	91
9	0.455	95	43.250	95
10	0.508	95	48.250	95
11	0.561	98	54.932	98
12	0.613	108	66.221	108
13	0.666	109	72.571	109
14	0.718	140	100.579	140
15	0.771	141	108.718	141
16	0.824	160	131.789	160
17	0.876	163	142.839	163
18	0.929	196	182.074	196
19	0.982	231	226.745	231
	Moyenne	110.47	69.455	
	ecart type	50.52		
	a moments	39.40		39.40
	X0 moments	87.74		87.74
	a mpond	41.03		41.03
	X0 mpond	86.79		86.79

i	F(i)=(i-0.35)/n	x(i)	F(i)*x(i)		données réelles
1	0.034	25	0.855		25
2	0.087	50	4.342		50
3	0.139	70	9.763		70
4	0.192	75	14.408		75
5	0.245	77	18.845		77
6	0.297	84	24.979		84
7	0.350	91	31.850		91
8	0.403	91	36.639		91
9	0.455	95	43.250		95
10	0.508	95	48.250		95
11	0.561	98	54.932		98
12	0.613	108	66.221		108
13	0.666	109	72.571		109
14	0.718	140	100.579		140
15	0.771	141	108.718		141
16	0.824	160	131.789		160
17	0.876	163	142.839		163
18	0.929	196	182.074		196
19	0.982	400	392.632	horsain artificiel	231
	Moyenne	119.37	78.186		
	ecart type	79.49			
	a moments	62.00		données réelles	39.40
	X0 moments	83.60		données réelles	87.74
	a mpond	53.39		données réelles	41.03
	X0 mpond	88.55		données réelles	86.79



VI-2) Intervalle de confiance des paramètres ou d'un quantile:

Compte tenu de l'échantillon disponible, il est intéressant de s'interroger sur l'incertitude échantillonnage qui affecte un paramètre, ou un quantile x(F).

Nous ne voulons pas alourdir cet exposé, et ce dernier aspect sera évoqué partiellement à propos de l'utilisation de la Loi de Gumbel dans la méthode du gradex.

Il faut toutefois garder à l'esprit que:

- dans la *méthode des moments* par exemple, on utilise des moments empiriques calculés sur l'échantillon. Or ceux-ci ne sont pas strictement égaux à ceux de la population.

Par exemple, si on prend différents échantillons de taille n:

leurs moyennes empiriques $m_x = \frac{1}{n}\sum_{i=1}^{n} x_i$ fluctuent

autour de la vraie moyenne de la population μ_X

avec un écart-type $\sigma_{\mu_x} = \frac{\sigma_X}{\sqrt{n}}$

De même pour l'écart-type empirique s_X, dont les estimations sur différents échantillons fluctuent

autour de la vraie valeur de la population σ_X

avec un écart-type $\sigma_{s_x} = \frac{\sigma_X}{\sqrt{n}}$

Et donc, quand on va utiliser ces moments empiriques pour estimer les paramètres α et β, les valeurs a et b obtenues varieront selon m_X et s_X, et donc selon l'échantillon...

- de même dans une *méthode graphique*, on va faire un choix quant à la droite qui intercepte au mieux les points. Celle-ci dépend déjà du *choix de la formule utilisée pour affecter les probabilités empiriques.*

Et si on a deux échantillons, ils ne fourniront pas strictement la même droite, et probablement pas celle qui correspond exactement à la population...

Ces incertitudes sur les valeurs obtenues pour les *paramètres* se transfèrent sur les résultats les plus utilisés en pratique: certains *quantiles* extrêmes.

Par exemple, pour la loi normale :

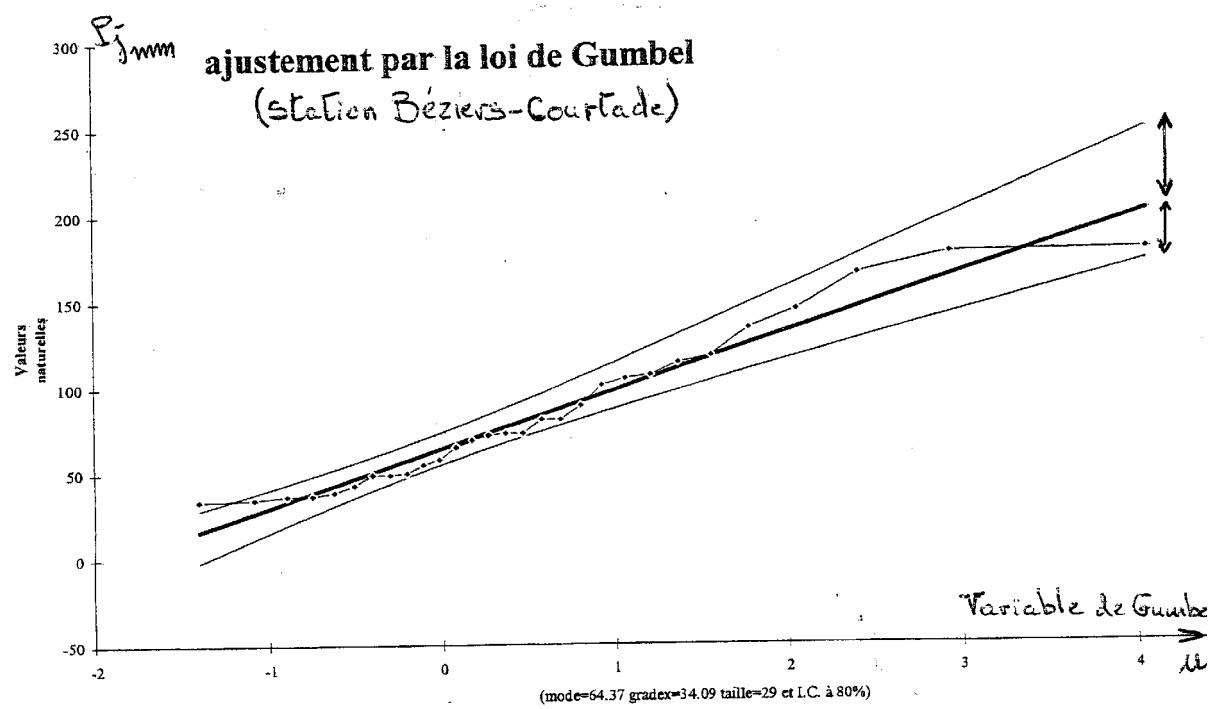
- l'incertitude sur la *pente* ne va pas trop affecter le centre du nuage de points, et donc le quantile $x_{.50}$, (ou x_2 en période de retour),
- mais beaucoup plus les extrêmes (x_{100} ou x_{1000} en période de retour).

De même pour la loi de Gumbel :$_T$

En fait, cette incertitude sur les quantiles est assez couramment utilisée pour la loi de Gumbel, qui sert notamment à traiter les valeurs « extrêmes » dans les problèmes de pluies ou de débits de projet.

On est donc tenter d'utiliser en **extrapolation** la loi ajustée, et à proposer des quantiles Q_T de grandes périodes de retour t = 500, 1000, 3000… ans ou plus.

Ces valeurs sont donc affectées d'une incertitude d'échantillonnage, notamment sur la pente de la droite. De plus cette incertitude n'est pas symétrique de part et d'autre de Q_T, donc l'écart type ne suffit pas à la caractériser. On en voit un exemple sur la figure ci dessous pour les pluies de Béziers La Courtade (1970-98) avec le logiciel HYDROLAB.



La théorie de l'échantillonnage est complexe et dépasse les besoins d'un ingénieur hydrologue. ***Mais il doit garder ce problème à l'esprit*** et savoir qu'il trouvera au besoin dans les ouvrages spécialisés des formules donnant l'incertitude de ces quantiles..

CONCLUSIONS

Nous pensons néanmoins avoir donné un bon aperçu des méthodes classiques d'ajustement probabiliste que l'ingénieur doit connaître. Mais ce n'est qu'un début, (déjà substantiel , n'est-ce pas…?), et il faudra peut-être le compléter à l'occasion. De plus, c'est un domaine en pleine évolution, même si certains développements récents ne font pas toujours l'objet d'un consensus immédiat ...

On se gardera donc de tout dogmatisme et, au besoin, on simulera des échantillons nombreux sur lesquels on testera le plus objectivement possible deux méthodes concurrentes avant d'en choisir une...

Courage, et bonne chance...!

BIBLIOGRAPHIE:

FIORENTINO M. and S. GABRIELE (1984)
A correction for the bias of maximum likelihood estimators of Gumbel parameters
J. of Hydrology, Vol 73, p. 39-49

Groupe CHADULE (1974)
Initiation aux méthodes statistiques en Géographie.
(Ouvrage collectif) Masson et Cie ed. 192 p. (probablement épuisé mais disponible en bibliothèque)

GREENWOOD J.A., LANDWEHR J.M. , and MATALAS N.C. (1979)
Probability weighted moments: definition and relations to parameters of several distributions expressible in inverse form
Water Ressources Research,, Vol. 15, N° 5, pp. 1049-54

GUMBEL E.J. (1958)
Statistics of Extremes
Columbia University Press - New York

HUBERT P. et H. BENDJOUDI (1998)
A propos de la distribution statistique des cumuls pluviométriques annuels : Faut-il en finir avec la normalité ?
Revue des Sciences de l'Eau

OMM (1983)
Guide des Pratiques Hydrologiques - Vol. II : Analyse, prévision et autres applications
Organisation Météorologique mondiale . Publi. N° 168 Genève

LUBES H., MASSON J.M., RAOUS P., TAPIAU M. (1994)
SAFARHY, Logiciel de calculs statistiques et d'analyse fréquentielle adapté à l'évaluation du risque en hydrologie.
Manuel de référence, ORSTOM, Univ. de Montpellier II

MASSON J.M. et H. LUBES (1991)
Méthodes des moments de probabilité pondérés: application à la loi de Jenkinson.
Hydrologie Continentale Ed. ORSTOM Vol. 6, N° 1, pp. 67-84

ROCHE M. (1965)
Hydrologie de Surface Ed. Gauthier-Villars PARIS

SLIMANI M. (1985)
Etude des pluies de fréquences rares à faible pas de temps sur la région Cévennes - Vivarais: estimation, relation avec le relief, et cartographie synthétique.
Thèse de l'Institut National Polytechnique de Grenoble.

VIALAR (1986)
Probabilités et Statistiques (5 fascicules)
Cours de l'Ecole Nationale de la Météorologie

2éme Partie: LIAISONS STOCHASTIQUES ENTRE VARIABLES

CHAPITRE IV : LA CORRELATION SIMPLE

Objectifs : 131

I) ASPECTS ANALYTIQUES: 133

I-1) Recherche de la meilleure droite d'estimation 133

I-2) Compléments sur droites de régression et Intervalles de confiance 139

I-3) Extensions aux cas non linéaires 142

II) ASPECTS PROBABILISTES: 145

II-1) Interprétation dans le cas d'une loi binormale 145
II-2) Effets de l'échantillonnage 151
II-3) Simulation stochastique 157

III) PIEGES DE LA CORRELATION 159

III-1) Pièges géométriques 159
III-2) Pièges de cofluctuation 160
III-3) Variables monotones 161
III-3) Variable influente cachée 161
III-4) Corrélation et liaisons de cause à effets 162

IV) APPLICATIONS PARTICULIERES: 165

IV-1) Reconstitution de données - extension de séries 165
IV-2) Traitements de données de mesures 169

2éme Partie: LIAISONS STOCHASTIQUES ENTRE VARIABLES

CHAPITRE I : LA CORRELATION SIMPLE

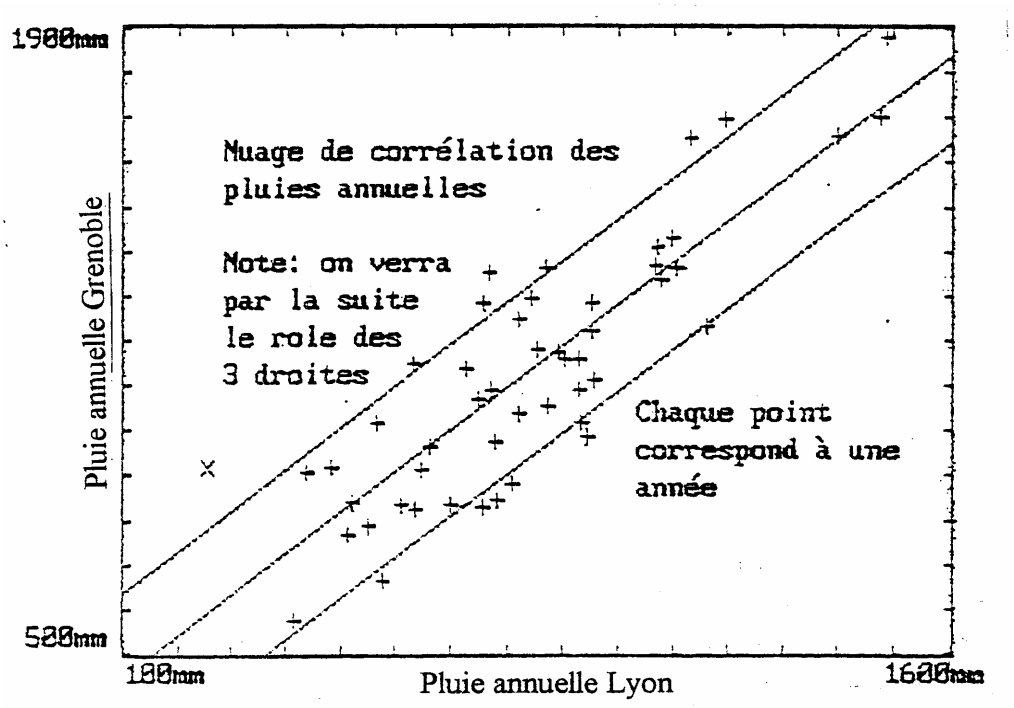
Objectifs :

Soit deux variables aléatoires X et Y:

(par exemple, X est la pluie annuelle à Lyon, et Y la pluie annuelle à Grenoble).

Celles-ci sont connues sur un *échantillon* de N observations. Si on porte sur un graphique (cf. Figure 1) les N points de coordonnées X_i, Y_i, i de 1 à N, on obtient quelque chose qui ressemble plus à un nuage de points qu'à un tracé pointilliste de courbe :

Figure 1:



et ceci pour diverses raisons:

- la liaison n'est pas toujours fonctionnelle (c'est le cas des pluies),
- les données sont entachées d'erreurs, etc...

On peut alors chercher à:

+ schématiser analytiquement cette liaison

(par exemple pour pouvoir facilement estimer une valeur de Y à partir d'une valeur de X)

+ caractériser la dépendance entre X et Y par une valeur numérique.

Les applications sont nombreuses et très importantes:

+ prévision (par exemple:
  prévision des apports de fusion nivale à partir des précipitations d'hiver)

+ contrôle et reconstitution de données (on va reconstituer Y à Grenoble, où
  des valeurs sont manquantes, à partir de Lyon, où la série est complète)

+ comparaison théorie-expérimentation.

Certes dans la pratique, on utilisera souvent plus de 2 variables (cf. chapitre II de cette 2ème Partie sur la corrélation *multiple*), mais il faut déjà bien comprendre le cas le plus simple de la corrélation entre 2 variables.

Notons que depuis une douzaine d'années, de nombreuses calculettes calculent tous les paramètres que nous allons décrire, et désormais ce sont les tableurs sur micro-ordinateurs qui offrent ces mêmes possibilités; il n'y a donc plus de problèmes matériels liés à des calculs fastidieux

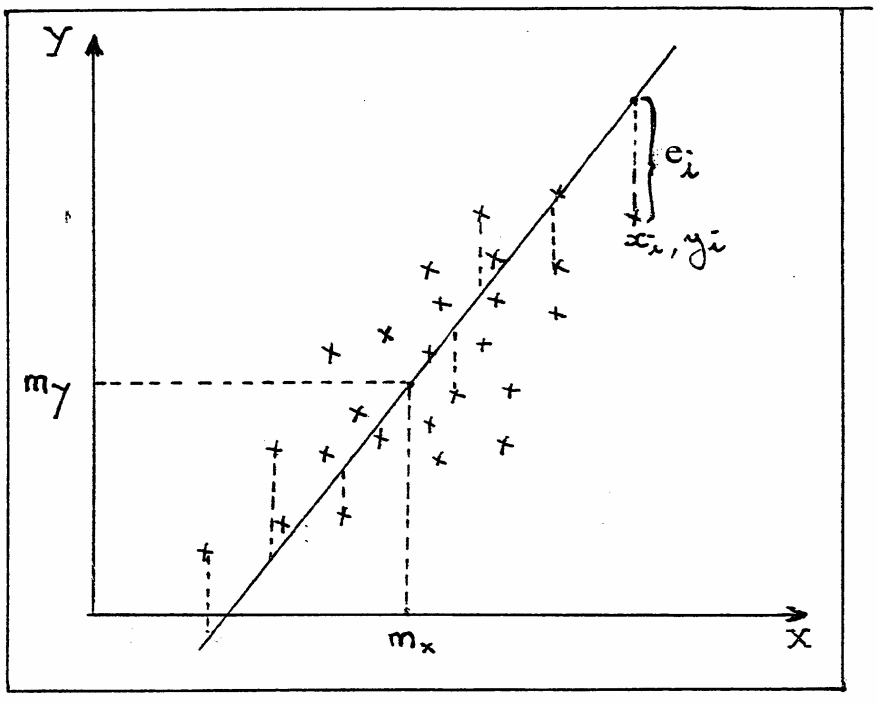


Figure 2:

I) ASPECTS ANALYTIQUES:

Nous allons commencer par la solution la plus simple, celle des liaisons linéaires.

- soit un échantillon de N observations formé de N couples $\{ x_i , y_i , i = 1 , N\}$
- soit m_X et m_Y les moyennes de X et Y sur l'échantillon,
  (encore notées parfois $\overline{X}$ et $\overline{Y}$)
- et soit s_X et s_Y les écart-types de X et Y estimés là aussi sur l'échantillon,

Hypothèse: pour l'instant: aucune...!

(si ce n'est qu'on suppose que les 2 écart-types sont non nuls: il serait absurde de chercher une liaison entre une constante et une variable ! ou entre 2 constantes).

I-1) RECHERCHE de la **MEILLEURE DROITE D'ESTIMATION** de Y ***à partir de*** X:

Attention:

On notera que dès le départ, on fait des rôles différents aux 2 variables Y et X, (cf. Figure 2), en comptant les écarts *parallèlement* à l'axe des Y.

Soit donc: $y = a.x + b$ l'équation de cette droite.

Les deux coefficients a et b sont des *paramètres*, que l'on va adapter pour que la droite représente au mieux la relation linéaire sur cet échantillon particulier.
Pour chaque point, on commet une erreur d'estimation, en estimant

$$y_i \quad \text{par} \quad y_i^* = a.x_i + b$$

celle-ci est appelée résidu et noté e_i :

$$e_i = y_i - a.x_i - b$$

Et nous allons chercher la "meilleure droite", ($\Rightarrow$ donc ses paramètres, ou ses *coefficients* a et b), au sens d'un certain *critère*.

Pas 1: **Choix du critère**: définition de "**meilleure**" dans "meilleure droite".

C'est un point important, car il est lié aux objectifs.

On pourrait dire que la "meilleure droite" est celle qui, *sur l'échantillon*, rend:

a) $\sum_{i=1}^{N} e_i$ minimum

c'est à dire que la somme algébrique des erreurs serait minimum...

Ceci est un peu absurde car on autoriserait alors de grandes erreurs, tant positives que négatives, pourvu qu'elles se compensent !

Plus rigoureusement, on peut vérifier que cette somme est d'ailleurs nulle pour toute droite passant par le centre de gravité des points $(\overline{X}, \overline{Y})$!

$\Rightarrow$ Cette droite ne serait donc pas unique..!

b) $\sum_{i=1}^{N} |e_i|$ minimum

Cette fois, c'est la somme des valeurs absolues des écarts que l'on voudrait minimiser. C'est intéressant mais compliqué..., comme d'ailleurs le critère suivant:

c) $\underbrace{\text{Maximum de } |e_i|}_{i=1 \grave{a} N}$ = Minimum

où c'est le maximum (en valeur absolue) des écarts que l'on voudrait minimiser, grâce aux paramètres a et b.

C'est là aussi compliqué, encore que cette méthode MiniMax ait une solution (algorithme de REMES)...

De plus, cela privilégie quelques points (en fait 3: les points les plus extérieurs au nuage) et ce résultat (la droite obtenue) est alors très lié à l'échantillon considéré et peut changer sensiblement en changeant un seul point.

d) $\sum_{i=1}^{N} e_i^2$ minimum

C'est la méthode dite des **moindres carrés**, méthode la plus utilisée, car elle est doublement intéressante. Sa solution est rapide et simple. Et elle est relativement robuste quand on change d'échantillon. Toutefois, pour des petits échantillons, elle est très sensible aux points un peu écartés.

$\Rightarrow$ C'est celle que nous retiendrons.

Pas 2: **Calcul des paramètres de la droite des moindres carrés.**

Notre critère est donc de minimiser, *pour l'échantillon considéré*, la quantité E, fonction des deux paramètres a et b:

$$E(a,b) = \sum_{i=1}^{N} e_i^2 = \sum_{i=1}^{N} \left(y_i - a.x_i - b\right)^2$$

E est donc une fonction de a et b, une fois l'échantillon donné (i.e. les $\{x_i , y_i\}$ fixés).
On sait qu'une des conditions nécessaires d'extremum est que les dérivées partielles de E (par rapport à a et b) soient nulles, donc :

$$\frac{\partial E(a,b)}{\partial a} = 0 \qquad \text{et} \qquad \frac{\partial E(a,b)}{\partial b} = 0$$

Si nous commençons par la seconde:

$$\frac{\partial E(a,b)}{\partial b} = -2.\sum_{i=1}^{N} \left(y_i - a.x_i - b\right) = 0 \quad \Leftrightarrow \quad \sum_{i=1}^{N} \left(y_i - a.x_i - b\right) = 0$$

ou encore:

$$\sum_{i=1}^{N} \left(y_i\right) - a.\sum_{i=1}^{N} \left(x_i\right) - b.\sum_{i=1}^{N} 1 = 0 \quad \Leftrightarrow \quad \frac{1}{N}\sum_{i=1}^{N} \left(y_i\right) - \frac{1}{N}.a.\sum_{i=1}^{N} \left(x_i\right) - b.\frac{1}{N}.\underbrace{\sum_{i=1}^{N} 1}_{=N} = 0$$

soit finalement l'équation (1):

$$m_y - a.m_x - b = 0 \quad \Leftrightarrow \quad m_y = a.m_x + b \qquad \text{(eq. 1)}$$

donc les valeurs moyennes m_x et m_y *estimées sur l'échantillon* vérifient exactement l'équation de la droite *optimisée sur l'échantillon*.

⇒***Résultat 1*** : la droite passe par le *centre de gravité* du nuage (m_x ,m_y).
Et on en déduit de plus que *la somme des résidus est strictement nulle.*

En effet:

$$\sum_{i=1}^{N} e_i = \sum_{i=1}^{N} \left(y_i - a.x_i - b\right) = \sum_{i=1}^{N} \left(y_i\right) - a.\sum_{i=1}^{N} \left(x_i\right) - b.\sum_{i=1}^{N} \left(1\right) = N.m_y - a.N.m_x - N.b = 0$$

Avant de traiter la première équation, on peut (✸ astuce ...!) lui intégrer le résultat déjà obtenu sur la seconde, puisque les deux doivent être vérifiées en même temps.
L' équation de E , en prenant en compte ce résultat est alors du type:

$$b = m_y - a.m_x \quad \Rightarrow \quad y_i = a.x_i + b \; = \; a.x_i + m_y - a.m_x = \; m_y + a.(x_i - m_x)$$

D'où l'expression qui reste à minimiser:

$$E(a,b) = \sum_{i=1}^{N} \left(y_i - a.x_i - b\right)^2 \quad \Rightarrow \quad E(a) = \sum_{i=1}^{N} \left[\left(y_i - m_y\right) - a.\left(x_i - m_x\right)\right]^2$$

soit: $$\frac{\partial E(a,b)}{\partial a} = -2.\sum_{i=1}^{N}\left[(y_i - m_y) - a.(x_i - m_x)\right].(x_i - m_x) = 0$$

ou encore: $$a = \frac{\sum_{i=1}^{N}(y_i - m_y).(x_i - m_x)}{\sum_{i=1}^{N}(x_i - m_x)^2} = \frac{\frac{1}{N-1}.\sum_{i=1}^{N}(y_i - m_y).(x_i - m_x)}{\frac{1}{N-1}.\sum_{i=1}^{N}(x_i - m_x)^2} \quad \text{(eq.2)}$$

où, en introduisant en haut et en bas le facteur $\frac{1}{N-1}$, l'on reconnaît au dénominateur la variance empirique de X soit s_X^2, et au numérateur le moment croisé ou covariance C_{XY}

Posons alors:

$$r_{xy} == \frac{\frac{1}{N-1}.\sum_{i=1}^{N}(y_i - m_y)(x_i - m_x)}{\sqrt{\frac{1}{N-1}.\sum_{i=1}^{N}(x_i - m_x)^2}.\sqrt{\frac{1}{N-1}.\sum_{i=1}^{N}(y_i - m_y)^2}}$$

r_{xy} est appelé le **coefficient de corrélation** entre x et y;
c'est la somme des produits des écarts aux moyennes respectives, divisée par le produit de la racine carrée de la somme des carrés des écarts aux moyennes respectives, c'est à dire divisée par le produit des écart-types.

Finalement, on trouve comme paramètre optimaux de la droite:

$$a = r_{xy}.\frac{s_y}{s_x} \qquad \text{et} \qquad b = m_y - a.m_x$$

et cette droite $$y = a.x + b$$
est appelée **droite de régression de X en Y**.

Elle est toujours définie et *unique pour un échantillon donné.*

***Pas 3**: **Q**ualité de l'estimation*

Il nous reste à savoir si cette droite nous permet d'estimer, pour les points $\{x_i,y_i\}$de l'échantillon, Y à partir de X sans trop d'erreur.

On sait déjà que l'erreur **moyenne** est strictement nulle *sur l'échantillon*.(i.e. si on ré applique la relation aux points qui constituent l'échantillon

$$m_e = \frac{1}{N}\sum_{i=1}^{N}(y_i - a.x_i - b) = \frac{1}{N}\sum_{i=1}^{N} y_i - a.\frac{1}{N}\sum_{i=1}^{N} x_i - b.\frac{1}{N}\sum_{i=1}^{N} 1 = m_y - a.m_x - m_y + a.m_x = 0$$

On peut aussi calculer l'écart type résiduel, c'est à dire l'écart type s_e des erreurs d'estimation; c'est déjà une première mesure de la qualité de la relation linéaire.

Mais la mesure la plus intéressante est sans aucun doute la comparaison entre l'écart type résiduel et l'écart type marginal s_y de la variable Y estimée.
Cet écart-type du résidu, puisque sa moyenne est nulle, s'écrit, à partir de la somme des carrés des résidus:

$$E(a,b) = \sum_{1}^{N} e_i^2 = \sum_{i=1}^{N}(y_i - a.x_i - b)^2$$

Celle-ci devient avec les valeurs *optimales* de a et b:

$$\frac{1}{N-1}\sum_{i=1}^{N} e_i^2 = \frac{1}{N-1}E(a,b) = \frac{1}{N-1}\sum_{i=1}^{N}\left[(y_i - m_y) - r_{xy}.\frac{s_y}{s_x}.(x_i - m_x)\right]^2$$

que l'on peut écrire, en développant le terme de gauche,

$$s_e^2 = \frac{1}{N-1}\sum_{i=1}^{N} e_i^2 = \frac{1}{N-1}\sum_{i=1}^{N}(y_i - m_y)^2 - 2.r_{xy}.\frac{s_y}{s_x}.\underbrace{\frac{1}{N-1}.\sum_{i=1}^{N}(y_i - m_y).(x_i - m_x)}_{r_{xy}.s_y.s_x} + r_{xy}^2.\frac{s_y^2}{s_x^2}.\underbrace{\frac{1}{N-1}.\sum_{i=1}^{N}(x_i - m_x)^2}_{s_x^2}$$

ou encore

$$s_e^2 = s_y^2 - 2.r_{xy}^2.s_y^2 + r_{xy}^2.s_y^2 = s_y^2.(1 - r_{xy}^2)$$

soit:

$$s_e = s_y.\sqrt{1 - r_{xy}^2}$$

⇒**<u>Résumé</u>**: Sur l'échantillon de n couples x_i, y_i, il existe:

+ une droite de régression de X en Y:

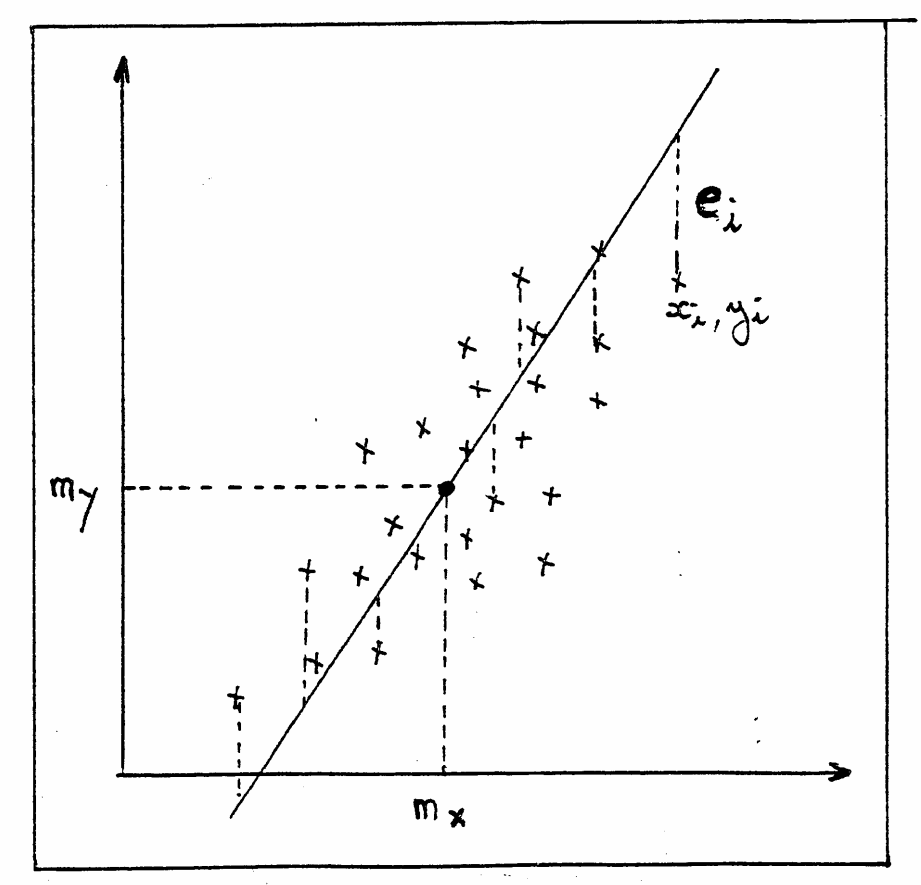
$$y = m_y + r_{xy} \cdot \frac{s_y}{s_x} \cdot (x - m_x)$$

donnant pour chaque point i de l'échantillon une estimation Y_i^* de Y_i

+ entachée d'une erreur e_i, de moyenne nulle et d'écart type:

$$s_e = s_y \cdot \sqrt{1 - r_{xy}^2}$$

Figure 2:



I-2) COMPLEMENTS sur DROITES de REGRESSION, et INTERVALLES de CONFIANCE:

a) les **DEUX** droites de régression**:**

On peut de même rechercher la meilleure estimation linéaire de X à partir de Y. On parlera alors de la droite de régression de Y en X, qui estime:

X* = a'.Y+b' en minimisant: $\sum_{i=1}^{N}(x_i^* - x_i)^2$

Cette droite a pour équation:

$$x = m_x + r_{xy}.\frac{s_x}{s_y}.(y - m_y) \quad \Leftrightarrow \quad y = m_y + \frac{1}{r_{xy}}.\frac{s_y}{s_x}.(x - m_x)$$

donnant pour chaque point de l'échantillon une estimation de X à partir de Y, de moyenne nulle et d'écart type:

$$s'_e = s_x.\sqrt{1 - r_{xy}^2}$$

De manière plus symétrique, on peut écrire que la régression de :

$Y \Leftarrow X$ $\qquad\qquad$ $X \Leftarrow Y$

fournit: $\frac{y - m_y}{s_y} = r_{xy}.\frac{x - m_x}{s_x}$ $\qquad\qquad$ $\frac{x - m_x}{s_x} = r_{xy}.\frac{y - m_y}{s_y}$

qui, une fois réécrites dans le repère classique y en fonction de x, donnent:

$$y = m_y + r_{xy}.\frac{s_y}{s_x}.(x - m_x) \qquad\qquad y = m_y + \frac{1}{r_{xy}}.\frac{s_y}{s_x}.(x - m_x)$$

Notons que ces deux droites ne sont confondues que ***si*** le coefficient de corrélation est égal à 1 ou -1, c'est à dire si les points sont strictement alignés, (-cas de la liaison linéaire exacte-).

Figure 3:

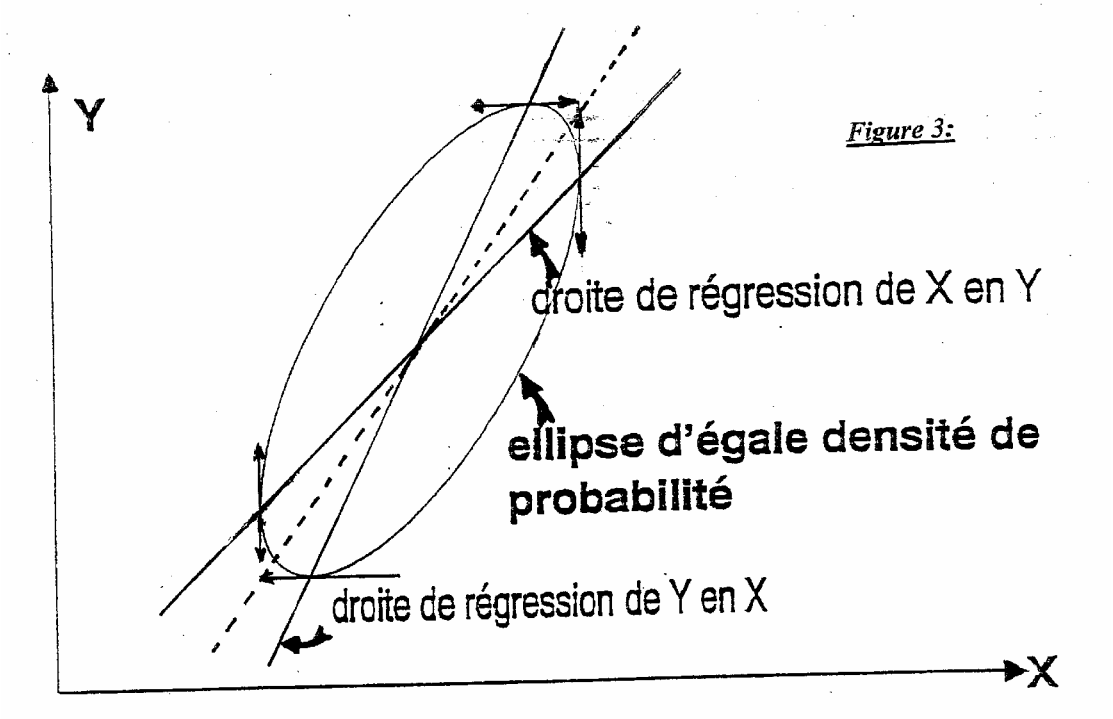


Figure 3:

b) Intervalle de confiance des résidus

On a donc "calé" la droite de régression sur les N couples, ce qui permet de calculer les N écarts e_i correspondant à chaque couple (x_i, y_i).
On peut alors considérer l'échantillon des écarts $\{e_i, i = 1,N\}$.
On sait déjà qu'il a par construction :

une moyenne nulle $m_e = 0$ et un écart-type $s_e = s_y \cdot \sqrt{1 - r_{xy}^2}$

Hypothèse:

On peut de plus *tester* s'il suit approximativement une loi *normale*:

Si l'hypothèse est vérifiée (en général, on l'admet sans vérification, quitte à la contester ensuite - cf. paragraphe sur l'homoscédasticité-) , on peut alors appliquer les propriétés de cette loi.

Par exemple, dans une loi normale:

80% des individus sont compris entre la moyenne et + ou - 1.28 écart-type

On va donc tracer, autour de la droite de régression, à une distance de + et - $1{,}28.s_e$, deux parallèles à cette droite et vérifier qu'approximativement, 80% des points ayant servis à l'ajustement sont contenus dans cette ***"bande de confiance"*** à 80%.

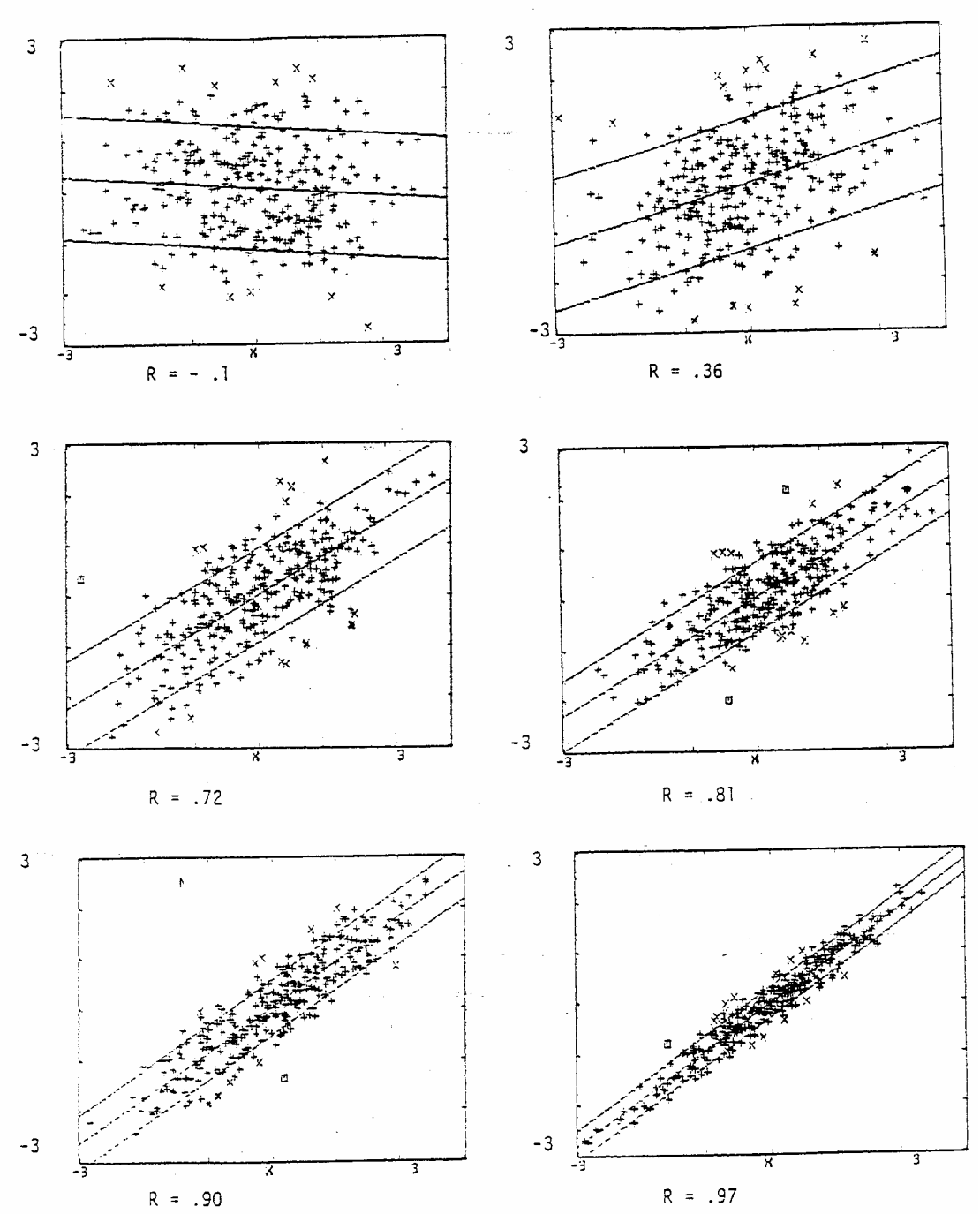


Figure 4: Exemples de nuages de corrélation entre 2 variables tirées de populations binormales de moyenne 0 et d'écart type 1, de différentes valeurs de corrélation. Sur chaque graphe, sont tracées la droite de régression de X en Y et les droites de confiance à 80% sur l'échantillon.

c) Relation entre l'*échantillon* et la *population*

Jusqu'à présent, on n'a considéré que l'échantillon de N couples disponibles. Mais celui-ci est en général extrait d'une population, potentiellement infinie, et qui a des caractéristiques bien définie.

On peut ainsi considérer que, sur la population infinie des X et des Y, on a une relation: :

$y = \alpha.x + \beta + \varepsilon$ ε gaussien N(0,σ_ε)

Et sur l'échantillon de N couples (x_i, y_i), on cale un modèle optimal

$y = a.x + b + e$

Mais si on prend un autre échantillon , on trouvera:

$y = a'.x + b' + e'$ avec a'≠ a , b'≠ b et $s_e \neq s_{e'}$

Donc il faudra aussi penser que, si on ajoute des couples à un échantillon, et si on ne recalcule pas a et b, la somme des résidus cessera d'être strictement nulle.

I-3) EXTENSIONS AUX CAS NON LINEAIRES (*):

Par linéaire, on indique que le système d'optimisation conduit à une expression linéaire des paramètres. Dans certains cas, on peut ajuster autre chose qu'une droite et arriver à un tel système linéaire.

a) cas linéarisables:

Un premier exemple est celui où la relation entre X et Y n'est pas linéaire, mais où les paramètres de calage interviennent linéairement.

Par exemple, si on pressent une fluctuation saisonnière entre le rayonnement ou de l'évaporation Y, et le nombre de jours *t* depuis le solstice d'hiver , on pourra chercher à caler une relation:

$$y = a.\sin\frac{2.\pi}{365}t + b$$

Cette relation n'est qu'approchée, par exemple, pour le rayonnement, à cause du masque des montagnes environnantes. C'est pourquoi on va caler statistiquement pour ne pas avoir à entrer dans le détail de ces influences " parasites ".

Mais à condition de définir x = $\sin\frac{2.\pi}{365}t$, on reste dans le cas de la corrélation simple.

Un autre exemple en est la *régression polynomiale,* où l'on cherche à caler par exemple:

y = a.x² + b.x + c en minimisant $E(a,b,c)=\sum_{i=1}^{N}\left(y_i - a.x_i^2 - b.x_i - c\right)^2$

ce qui conduit aux 3 équations:

$$\frac{\partial E}{\partial a}(a,b,c) = -2\sum_{i=1}^{N}\left(y_i - a.x_i^2 - b.x_i - c\right).x_i^2 = 0$$
$$\frac{\partial E}{\partial b}(a,b,c) = -2\sum_{i=1}^{N}\left(y_i - a.x_i^2 - b.x_i - c\right).x_i = 0$$
$$\frac{\partial E}{\partial c}(a,b,c) = -2\sum_{i=1}^{N}\left(y_i - a.x_i^2 - b.x_i - c\right) = 0$$

qui restent linéaire en a,b,c.
(En pratique, on considérera plus généralement que x et x² sont deux variables distinctes et on appellera alors un algorithme de corrélation multiple)

b) cas linéarisable par transformation:

Un autre cas peut concerner, par exemple, des fonctions puissances

y = a.x ^{b} ou y = a .e $^{b.x}$

Dans ce cas, l'optimisation de:

$$E(a,b)=\sum_{i=1}^{N}\left(y_i - a.e^{b.x_i}\right)^2$$

fournirait: $\frac{\partial E}{\partial a}(a,b) = -2.\sum_{i=1}^{N}\left(y_i - a.e^{b.x_i}\right)e^{b.x_i}$

et $\frac{\partial E}{\partial b}(a,b) = -2.\sum_{i=1}^{N}\left(y_i - a.e^{b.x_i}\right).x_i.e^{b.x_i}$

qui ne sont plus linéaires en a et b... Par contre, un simple passage en logarithme nous fournit:

y = a.e $^{b.x}$ $\Rightarrow$ Log y = Log a + b. x

mais **attention**...!
cette formulation va minimiser :

non pas $\sum_{i=1}^{N}\left(y_i - y_i^*\right)^2$ **mais** $\sum_{i=1}^{N}\left(Log\ y_i - Log\ y_i^*\right)^2$

Dans ce cas, les valeurs obtenues pour a et b ne seront pas optimales sur les valeurs brutes, par exemple pour de la prévision sur y; et il faudra éventuellement les affiner (cf. ci-après) par un algorithme itératif. Celui-ci cherchera à minimiser $\sum e_i^2$ en faisant varier a et b:

$\Rightarrow$ on pourra partir de a_0 et b_0 , qui minimise en fait $\sum_{i=1}^{N}\left(Log\ y_i - Log\ y_i^*\right)^2$

c) cas **non** linéarisable:

C'est le cas où même des transformations ne permettent pas de revenir à une fonction linéaire des paramètres.

Par exemple: $y = e^{a.x}.\cos bx$ qui donne $Log\ y = a.x + Log \cos bx$

Dans ce cas, il faut utiliser des *techniques itératives*, comme l'algorithme de MARQUARDT (1953), ce qui suppose une initialisation de a et b pour laquelle on n'a pas d'indication ...

II) ASPECTS PROBABILISTES:

Jusqu'ici, on s'était limité à l'analyse de l'échantillon disponible, soit N couples, même si on avait noté que cet échantillon était en fait un tirage (parmi d'autres possibles...) dans une population infinie.

On va maintenant faire *en plus* des *hypothèses probabilistes* sur la distribution ***conjointe*** de X et Y dans cette population, (- en supposant qu'elle est binormale -) et voir les interprétations que l'on peut en tirer.

II-1) INTERPRETATION dans le cadre d'une LOI BI-NORMALE

Nous allons supposer ici que le couple de variables X, Y appartient à une **loi Binormale**. (*Attention*: ceci est différent de dire que X et Y suivent séparément une loi normale - cf. contre-exemple).
Cette loi binormale est définie par sa densité de probabilité:

$$f(x,y) = \frac{1}{2\pi.\sigma_x.\sigma_y.\sqrt{1-\rho^2}}.e^{-\frac{1}{2}.\frac{1}{1-\rho^2}.\left\{\frac{(x-\mu_x)^2}{\sigma_x^2}-2.\rho.\frac{(x-\mu_x)(y-\mu_y)}{\sigma_x.\sigma_y}+\frac{(y-\mu_y)^2}{\sigma_y^2}\right\}}$$

avec:
$\rho = \rho_{XY}$ = coefficient de corrélation ***théorique***, sur la population complète entre X et Y.

Cela permet de dire que :

la **Probabilité** de tirer X entre [x et x+dx], **et** Y entre [y et y+dy],
est égale à **f(x,y).dx.dy**

Pour simplifier la suite, on va supposer les variables X et Y *standardisées*, c'est à dire:

centrées: $\mu_x = \mu_y = 0$ et *réduites:* $\sigma_x = \sigma_y = 1$

La **loi conjointe** de X et Y devient donc:

$$f(x,y) = \frac{1}{2\pi.\sqrt{1-\rho^2}}.e^{-\frac{1}{2}.\left\{\frac{x^2-2.\rho.x.y+y^2}{1-\rho^2}\right\}}$$

On va alors chercher la **loi conditionnelle de Y**, c'est à dire la loi de distribution de Y quand la valeur de X est connue (ou fixée).

On montre qu'elle s'exprime en général par:

$$h_x(y) = h(y|x) = \frac{f(x,y)}{g(x)}$$

c'est à dire que:

$$\text{loi conditionelle de y (sachant x)} \;=\; \frac{\text{loi conjointe de (x,y)}}{\text{loi marginale de x}}$$

On va donc calculer ces différents termes, dans le cas de la loi binormale.

a) Loi Marginale de X et Y(*)

(*on peut sauter tout de suite au résultat)*:

Loi Marginale de X:

C'est la loi de X, sans précision sur la valeur de Y, $\Rightarrow$ donc intégrée sur toutes les valeurs de y:

$$g(x) = \int_{-\infty}^{+\infty} f(x,y).dy = \frac{1}{2\pi.\sqrt{1-\rho^2}}.\int_{-\infty}^{+\infty} e^{-\frac{1}{2}.\left\{\frac{x^2-2.\rho.x.y+y^2}{1-\rho^2}\right\}}.dy$$

Ici, une *astuce de calcul* consiste à écrire que :

$$x^2 - 2.\rho.x.y + y^2 \;=\; (y - \rho.x)^2 + (1-\rho^2).x^2$$

d'où:

$$g(x) = \frac{1}{2\pi.\sqrt{1-\rho^2}}.\int_{-\infty}^{+\infty} e^{-\frac{1}{2}.\left\{\frac{x^2-2.\rho.x.y+y^2}{1-\rho^2}\right\}}.dy = \frac{1}{2\pi.\sqrt{1-\rho^2}}.\int_{-\infty}^{+\infty} e^{-\frac{1}{2}.\left\{\frac{(y-\rho.x)^2+(1-\rho^2).x^2}{1-\rho^2}\right\}}.dy$$

On peut alors isoler une exponentielle en x^2, qui ne dépend plus de y:

$$g(x) = \frac{1}{2\pi.\sqrt{1-\rho^2}}.\int_{-\infty}^{+\infty} e^{-\frac{1}{2}.\left\{\frac{(y-\rho.x)^2}{1-\rho^2}+x^2\right\}}.dy = = \frac{1}{2\pi.\sqrt{1-\rho^2}}.\int_{-\infty}^{+\infty} e^{-\frac{1}{2}.\left\{\frac{(y-\rho.x)^2}{1-\rho^2}\right\}}.\underbrace{e^{-\frac{1}{2}.x^2}}.dy$$

et donc sort de l'intégrale en y, ou plutôt en

$$u = \frac{y-\rho.x}{\sqrt{1-\rho^2}} \qquad \text{et donc} \qquad du = \frac{dy}{\sqrt{1-\rho^2}}$$

d'où:

$$g(x) = \frac{e^{-\frac{1}{2}.x^2}}{2\pi}.\int_{-\infty}^{+\infty} e^{-\frac{1}{2}.\left\{\frac{(y-\rho.x)^2}{1-\rho^2}\right\}}.\frac{dy}{\sqrt{1-\rho^2}} = \frac{e^{-\frac{1}{2}.x^2}}{2\pi}.\int_{-\infty}^{+\infty} e^{-\frac{1}{2}.u^2}.du = \frac{e^{-\frac{1}{2}.x^2}}{\sqrt{2\pi}}$$

Résultat:

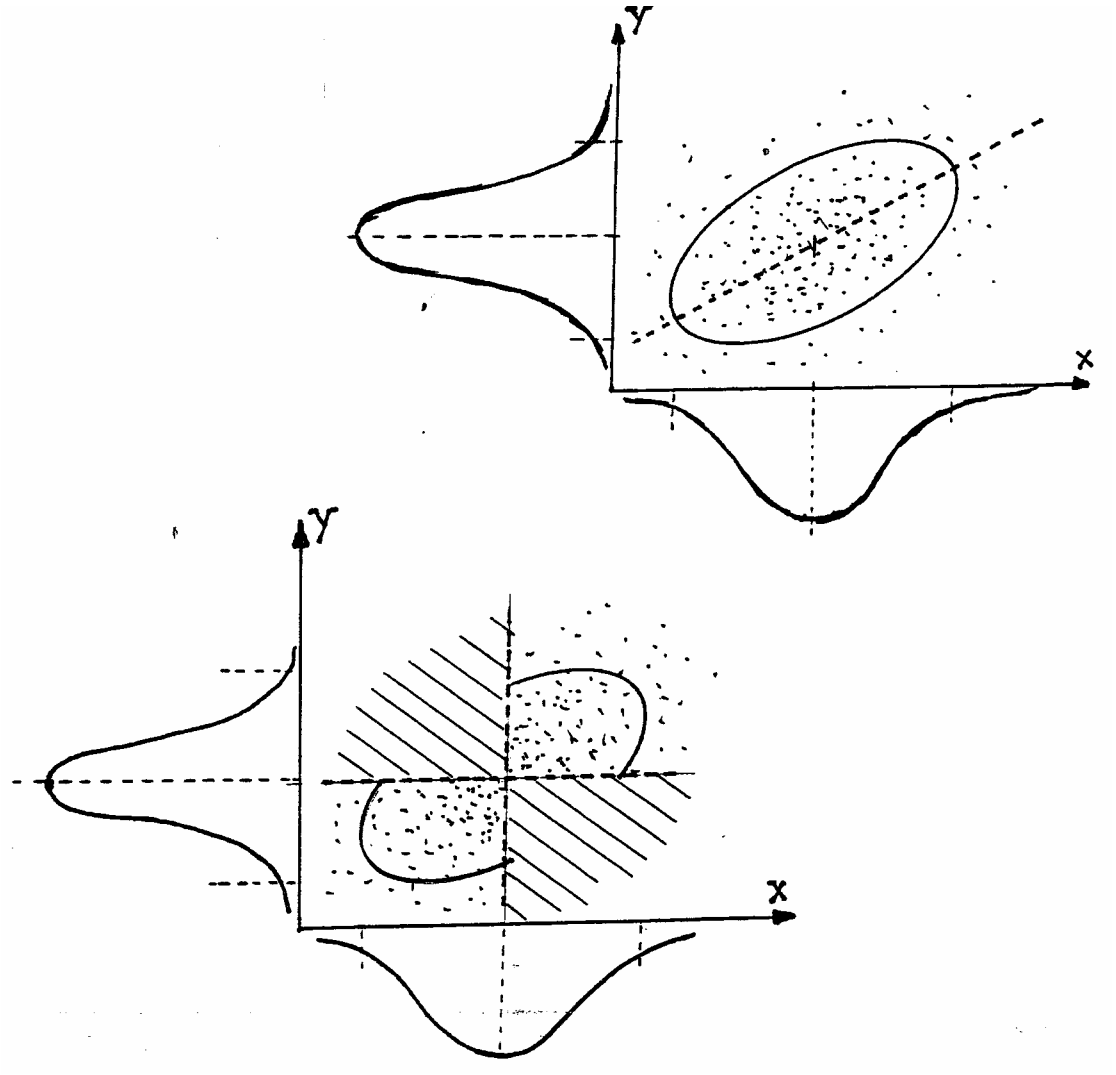
Donc la loi *marginale de X* dans une loi conjointe f(x,y) *binormale* est:

- une Loi Normale, ici N(0,1) parce que l'on est en centrée réduite,
- mais plus généralement $N(\mu_X,\sigma_X)$.

Loi Marginale de Y:

C'est de même la loi de distribution de Y, sans précision sur X. On montre de la même manière, en intégrant sur x, que c'est là aussi une Loi Normale $N(\mu_Y,\sigma_Y)$.

Figure 5:



b) Loi *conditionnelle* de Y sachant X:

C'est la distribution de Y quand X est fixé. Si on rappelle le résultat théorique:

$$\text{loi conditionelle de y (sachant x)} = \frac{\text{loi conjointe de (x,y)}}{\text{loi marginale de x}}$$

et si on l'applique ici maintenant que l'on connaît la loi marginale de X:

$$h_x(y) = \frac{f(x,y)}{g(x)} = \frac{\dfrac{e^{-\frac{1}{2}.\left\{\frac{x^2-2.\rho.x.y+y^2}{1-\rho^2}\right\}}}{2\pi.\sqrt{1-\rho^2}}}{\dfrac{e^{-\frac{1}{2}.x^2}}{\sqrt{2\pi}}} = \frac{\dfrac{e^{-\frac{1}{2}.\left\{\frac{(y-\rho.x)^2+(1-\rho^2).x^2}{1-\rho^2}\right\}}}{2\pi.\sqrt{1-\rho^2}}}{\dfrac{e^{-\frac{1}{2}.x^2}}{\sqrt{2\pi}}} = \frac{1}{\sqrt{2\pi}.\sqrt{1-\rho^2}}.e^{-\frac{1}{2}.\left\{\frac{(y-\rho.x)^2}{1-\rho^2}\right\}}$$

On constate alors, tous calculs faits, que *c'est encore une loi Normale*,
mais :

- qui n'est *pas centrée réduite*,
- puisque sa moyenne vaut **E[Y|X] = ρ.x**
- et sa variance est 1-ρ² ou, si Y n'est pas standardisée: **σ^2_Y.(1-ρ²)**

⇒ donc pour une valeur de X *fixée* → x_0, la moyenne *conditionnelle* de Y est l'estimé de Y pour X = x_0 par l'équation de régression, soit $y^* = \rho.x_0$

Et les valeurs de Y, autour de son espérance (- l'estimé par la régression) seront distribuées normalement autour de cette moyenne (- ici ce n'est pas une simple *constatation* sur l'échantillon de résidus, mais c'est un résultat *théorique* pour la loi binormale, -) , avec un résidu dont l'écart-type:

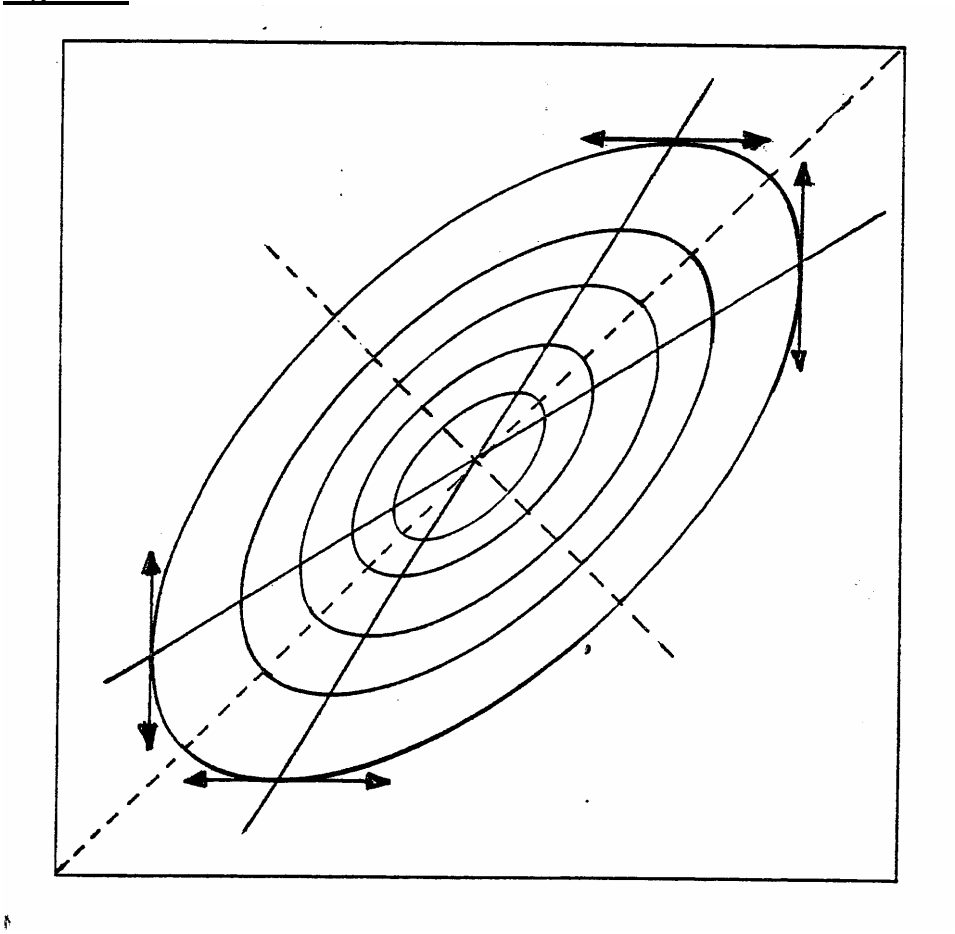
- ne dépend pas de la valeur x_0
- et est égal à l'écart-type résiduel calculé précédemment soit $\sigma^2_Y.(1-\rho^2)$.

c) Aspects géométriques:

On montre que pour cette distribution binormale, les courbes d'égale densité de probabilité f(X,Y) = cste sont dans le plan X, Y des ellipses d'autant plus allongées que la corrélation est bonne (ρ voisin de 1 ou -1).

De même on montre que les droites de régression de X en Y , et de Y en X sont les diamètres conjugués des directions verticales et respectivement horizontales de ces ellipses, qui sont toutes homothétiques (cf. Figure 6).

Figure 6:



BILAN:

1) dans tous les cas:

+ critère retenu: Moindres carrés des erreurs d'estimation de Y par une fonction linéaire de X:. Ces écarts vérifient donc

$$e_i = y_i - y_i^* \qquad \sum_{i=1}^{N} e_i = 0 \qquad \sum_{i=1}^{N} e_i^2 = \text{minimum}$$

+ la droite cherchée a pour équation:

$$y = m_y + r_{xy} . \frac{s_y}{s_x} . (x - m_x)$$

+ l'écart type des erreurs vaut:

$$s_e = s_y.\sqrt{1-r_{xy}^2}$$

2) si, de plus, la distribution est binormale, alors les résultats sur la population sont:

+ Loi Marginale de Y: $N(\mu_y, \sigma_y)$

+ Loi conditionnelle de Y pour X = x_0 :

$$N\left\{\mu_y + \rho.\frac{\sigma_y}{\sigma_x}.(x_0 - \mu_x)\,,\, \sigma_y.\sqrt{1-\rho^2}\right\}$$

L'équation de régression s'écrit donc:

$$y = \alpha.x + \beta + \varepsilon \quad \text{avec} \quad \alpha = \rho.\frac{\sigma_y}{\sigma_x} \qquad \beta = \mu_y - \alpha.\mu_x$$

et le résidu ε est distribué selon une loi normale $N\left\{0\,,\, \sigma_y.\sqrt{1-\rho^2}\right\}$

Remarque évidente: (donc qui va sans dire, mais qui va encore mieux en le disant...!)

S'il n'y a *pas de corrélation* entre Y et X (r ou $\rho_{XY} = 0$), alors:

- la droite de corrélation de Y en X fournit comme estimé y* toujours la même valeur , à savoir $y^* = m_Y$.

Ceci doit sembler évident, puisque X n'expliquant rien de la variance de Y, notre meilleure estimation pour Y est son espérance E[Y], estimée elle-même par sa moyenne sur l'échantillon.

- cette droite de régression est une horizontale y = cste = m_Y.
- de même la régression de X en Y fournit la droite perpendiculaire

x = cste = m_X.

On s'en rappellera quand on utilisera la régression pour compléter des séries de données (cf. parag. IV-1).

II-2) EFFETS DE L'**ECHANTILLONNAGE (*)**

Le fait de considérer que l'ensemble de données disponibles n'est qu'un échantillon dans une population permet des raffinements dans l'interprétation..
En effet, si la *vraie* relation dans la population est:

$$y = \alpha.x + \beta + \varepsilon \quad \text{avec} \quad \alpha = \rho.\frac{\sigma_y}{\sigma_x} \quad \beta = \mu_y - \alpha.\mu_x$$

alors celle que l'on ajuste sur un échantillon de N couples s'écrit:

$$y = a.x + b + e \quad \text{avec} \quad a = r.\frac{s_y}{s_x} \quad b = m_y - a.m_x$$

avec: $a \neq \alpha$ et $b \neq \beta$ et a et b *fonction de l'échantillon* particulier.

On sait déjà que, comme pour toute population, les moyennes ont une variance d'échantillonnage:

$m_x \quad \Rightarrow \quad$ variance d'échantillonnage $\quad \sigma_{m_x} = \frac{\sigma_x}{\sqrt{N}}$

$m_y \quad \Rightarrow \quad$ variance d'échantillonnage $\quad \sigma_{m_y} = \frac{\sigma_y}{\sqrt{N}}$

et de même les écart-types:

$s_x \quad \Rightarrow \quad$ variance d'échantillonnage $\quad \sigma_{s_x} = \frac{\sigma_x}{\sqrt{2.N}}$

$s_y \quad \Rightarrow \quad$ variance d'échantillonnage $\quad \sigma_{s_y} = \frac{\sigma_y}{\sqrt{2.N}}$

Mais il est intéressant de considérer l'effet de l'échantillonnage sur α, β, ρ et ε.

a) Estimateurs non biaisés:

+ Coefficient de corrélation non biaisé:

Le coefficient de corrélation précédemment défini, r_{xy} ou encore r, est un estimateur *biaisé* du coefficient de corrélation ρ.

C'est à dire que si X et Y sont tirés d'une population où la corrélation est de ρ , des calculs du coefficient r sur un grand nombre d'échantillons de taille N vont donner des valeurs de r plutôt optimistes (en effet, on va optimiser sur chaque échantillon, notamment en utilisant dans le calcul les moyennes et écart-types *propres à* chaque échantillon).

$\Rightarrow$ on va donc chercher un ***estimateur non biaisé*****,** c'est à dire plus proche (en espérance mathématique) de celui de la population.

On démontre que cet estimateur vaut: $$r'^2 = \frac{r^2(N-1)-1}{N-2}$$

Cette valeur est d'autant plus différente de r que r^2 est faible et N petit.

N	r	r'	;	N	r	r'	;	N	r	r'
5	.6	.39	;	10	.6	.53		30	.8	.79
5	.8	.60	;	10	.8	.77		30	.9	.896
5	.9	.87	;	10	.9	.887		30	.95	.948
5	.95	.93	;	10	.95	.944				

+ Ecart type résiduel:

Rappelons que l'on a cherché à minimiser les résidus sur l'échantillon. D'où:

$$s_e = s_y \cdot \sqrt{1 - r_{xy}^2}$$

Mais ce qui nous intéresse en général, c'est d'appliquer le schéma de régression sur des données non issues de l'échantillon, que ce soit en reconstitution ou en prévision. On commettra alors des "erreurs", ou plutôt on observera des écarts, dont la variance aura une espérance mathématique *plus grande*, dans la plupart des cas, que celle optimisée sur l'échantillon.

C'est pourquoi, on définit l'***écart type résiduel non biaisé*****:**

$$s_e' = s_e \cdot \sqrt{\frac{N-2}{N-1}} \quad \text{et soit} \quad k = \frac{s_e'}{s_e}$$

On a par exemple les valeurs suivantes:

N	3	5	10	20	30	50	100
k	1.41	1.15	1.06	1.03	1.02	1.01	1.005

b) Distribution du coefficient de corrélation:

Soit ρ la valeur de la corrélation dans la population supposée binormale, et *r* la valeur calculée sur un échantillon de taille N :

***r** est une **variable aléatoire**,* dont le tirage dépend de l'échantillon,

et on montre que :

- si *N est grand* (>500), alors les estimations r de ρ sont approximativement normales de distribution N(ρ, σ_r), avec $\sigma_r = \frac{1-\rho^2}{\sqrt{N}}$

- si *N est petit*, alors c'est la variable transformée (*variable de FISCHER*):

$$Z = \frac{1}{2} Log \frac{1+r}{1-r}$$

qui suit une loi normale:

de *moyenne*: $\mu_Z = \frac{1}{2} Log \frac{1+\rho}{1-\rho}$ et d'*écart type*: $\sigma_Z = \frac{1}{\sqrt{N-3}}$

Cette distribution est utilisée pour:

+ tester l'hypothèse d'indépendance des variables (ρ = 0 ?)

+ définir un intervalle de confiance de r

+ tester la différence entre 2 calculs de r sur des échantillons différents, pour savoir si elle est significative ou non.

***Exemple**:*

Entre 2 variables on a trouvé r = .3 sur un échantillon de 10 valeurs indépendantes.
Question: Peut-on affirmer raisonnablement que ces 2 variables sont liées (même faiblement) ?.

Faisons l'hypothèse ρ = 0 et calculons la probabilité de trouver r supérieur à 0.3 sachant que ρ = 0:
Dans ce cas l'espérance de Z est :

$$\mu_Z = \frac{1}{2} Log \frac{1+\rho}{1-\rho} = \frac{1}{2} Log \frac{1}{1} = 0$$

et l'écart-type $\sigma_Z = \frac{1}{\sqrt{N-3}} = \frac{1}{\sqrt{7}} = .378$

On a trouvé une valeur de Z, sur l'échantillon, de: $Z = \frac{1}{2} Log \frac{1+r}{1-r} = \frac{1}{2} Log \frac{1+0.3}{1-0.3} = 0.310$

D'où la valeur de la variable normale centrée réduite correspondant à Z:

$$u = \frac{0.310 - 0.0}{0.378} = 0.82$$

On trouve dans une table de la loi normale que l'on a alors encore une probabilité de 20.7 % de dépasser cette valeur, dans l'hypothèse d'indépendance des variables. Cette probabilité est suffisamment élevée pour que l'on puisse accepter l'hypothèse d'indépendance (puisque si les données étaient indépendantes, on aurait presque une chance sur 5 de trouver un coefficient supérieur à .3).

On en conclut qu'avec 10 couples, un coefficient de corrélation de r = 0.3 n'est pas significatif (i.e. pas significativement différent de zéro).

c) Distribution des coefficients de régression:

On rappelle que, si la *vraie* relation dans la population est:

$$y = \alpha.x + \beta + \varepsilon$$

en fait, on ajuste sur l'échantillon de N couples :

$$y = a.x + b + e$$

Hypothèses: X et Y sont des variables binormales.

On montre alors de même que *si l'échantillon est grand:*

E[a] = α E[b] = β

et $\sigma_a = \frac{\sigma_y}{\sigma_x} \cdot \sqrt{\frac{1-\rho^2}{N}}$ $\sigma_b = \sigma_y \cdot \sqrt{\frac{1-\rho^2}{N}}$

De plus, leur distribution est gaussienne.

Par contre, dans le cas des petits échantillons, on montre que ces formules deviennent:

$$\sigma_a = \frac{\sigma_y}{\sigma_x} \cdot \sqrt{\frac{1-\rho^2}{N-2}} \qquad \sigma_b = \sigma_y \cdot \sqrt{\frac{1-\rho^2}{N-2}}$$

et que la variable de Student t:

$$t = \frac{a - \alpha}{\sigma_a}$$ suit une loi de Student à N-2 degrés de liberté.

En pratique, on ne connaît pas ρ mais seulement une estimation r, d'où :

$$\sigma_b = s_y.\sqrt{\frac{1-r^2}{N-2}}$$

ou encore, en remarquant que $a = r.\frac{s_y}{s_x}$, la variable estimée t devient:

$$t = \frac{a-\alpha}{a}.\sqrt{\frac{r^2.(N-2)}{1-r^2}}$$

et de même pour b: $$t = \frac{b-\beta}{b}.\sqrt{\frac{(N-2)}{(1-r^2).(s_x^2+m_x^2)}}$$

suit une loi de Student à N-2 degrés de liberté où s^2_X est l'estimateur non biaisé de la variance de x.

Applications:

+ Tester si la constante de l'équation de régression peut être considérée comme nulle (souvent utile): $b \Rightarrow E[b] = \beta \neq 0$?

+ tester si la différence entre 2 équations est significative ou non.

$$a,b \text{ et } a',b' \Rightarrow E[a] \neq E[a'] \text{ et } E[b] \neq E[b]' \text{ ?}$$

d) Estimation d'un intervalle de confiance de l'estimé de Y pour la population(*):

Nous avons vu que sur l'échantillon, la droite optimisée sur cet échantillon fournissait:

$$y_i = a.x_i + b + e_i = y_i^* + e_i \qquad \text{avec} \quad e_i \in N\left\{0\,,\, s_e = s_y.\sqrt{1-r^2}\right\}$$

Dans une première approche, (-la plus courante en pratique-), on fournit:

- pour estimé de y_i à l'abscisse x_i la valeur y_i* déduite de cette droite;
- **or** celle-ci n'est *optimale **que** pour cet échantillon.*

On fournit ensuite:

- un intervalle de confiance qui est sensé représenter l'incertitude due aux facteurs non contrôlés par x, et concentrés dans le résidu.

Ce faisant, on travaille comme si on avait trouvé les *vrais* coefficients α et β de la population, et comme si e_i était strictement identique à ε_i.

Mais en fait, si on prend un autre échantillon, on trouvera une *autre droite:*

$$y_i = a'.x_i + b' + e_i = y'^{*}_i + e'_i \quad \text{avec} \quad e'_i \in N\left\{0\,,\, s_{e'} = s'_y.\sqrt{1-r'^2}\right\}$$

et donc, pour la même valeur de x_i, une valeur y'^{*}_i qui est calculée avec des a' et b' légèrement différents à cause de l'échantillonnage.

Donc un "raffinement" intéressant consiste:

- à cerner la variation de l'estimé (y_i^*, y'^{*}_i, etc...), en fonction de l'échantillonnage,
- et donc d'estimer pour une valeur x_i, la valeur la plus probable de y_i , c'est à dire l'espérance des y^*_i , soit $E[y^*_i]$ (-et un intervalle de confiance correspondant-),
- en tenant compte de l'échantillonnage sur les coefficients de régression.

On montre que la valeur la plus probable compte tenu de l'échantillon observé est celle définie par l'équation calculée sur l'échantillon,(- le seul disponible-), mais que par contre, y peut s'écarter de cette valeur selon une *loi de Student*.

D'où, si t_p est la valeur de la variable de Student à N-2 degrés de liberté telle que:

$$\Pr ob\left[|t| \le t_p\right] = p$$

l'intervalle de confiance à p% de probabilité (par exemple 80%) de l'estimation de y est défini par:

$$y_i^* \pm \Delta y_i^* = \underbrace{m_y + a.(x_i - m_x)}_{y_i^*} \pm t_p . \frac{s_e}{\sqrt{N}} . \sqrt{1 + \frac{(x_i - m_x)^2}{s_x^2}}$$

On remarquera que cet intervalle de confiance, qui inclue la fluctuation de la droite des moindres carrés selon l'échantillon, augmente si on s'éloigne de la moyenne des x , donc du barycentre. On peut comprendre intuitivement que le nuage de l'échantillon, sous l'hypothèse binormale, est plus dense et mieux défini autour du barycentre qu'à la périphérie.

Si on prend en compte cette fluctuation de l'estimé y^*_i dans l'intervalle de confiance "total" que l'on fournit pour y_i , et qui alors prend en compte *à la fois:*

- l'incertitude due aux facteurs non corrélés à x
- ***et*** le fait que l'on ne dispose que d'un échantillon, donc que a et b ne correspondent pas exactement à α et β,

cette incertitude globale devient:

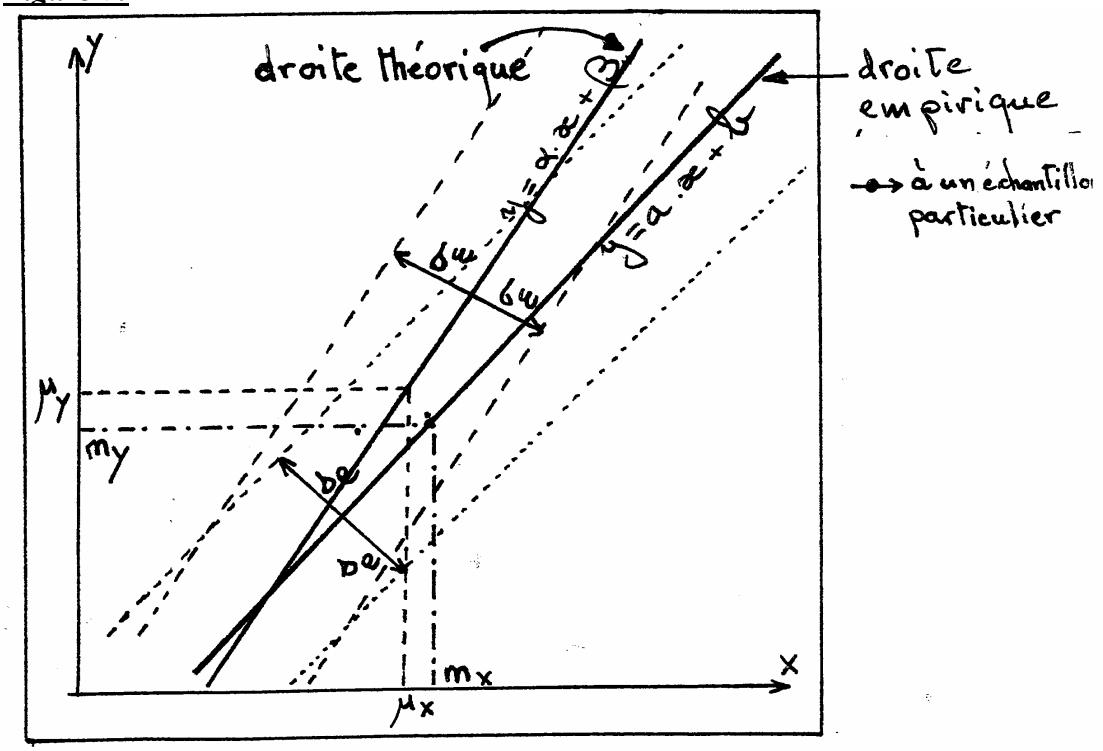
$$y_i \pm \Delta y_i = \underbrace{m_y + a.(x_i - m_x)}_{y_i^*} \pm t_p.s_e.\sqrt{1 + \frac{1}{N} + \frac{(x_i - m_x)^2}{N.s_x^2}}$$

On remarquera qu'il est supérieur à celui calculé couramment sur l'échantillon (en ignorant l'effet d'échantillonnage) et défini par:

$$y_i \pm \Delta y_i = \underbrace{m_y + a.(x_i - m_x)}_{y_i^*} \pm u_p.s_e$$

où u(p) est la variable centrée réduite de Gauss dont la valeur absolue n'est pas dépassée avec une probabilité p (par exemple u = 1.28 pour p = 80%).

Figure 7:



II-3) Simulation stochastique

Une façon de prendre conscience, de manière concrète, de ces effets consiste à réaliser de la simulation par génération stochastique. En général on effectue celle-ci dans un contexte de lois normales, mais des adaptations sont tout à fait possibles.

Pour cela, on part d'une corrélation calculée sur un échantillon, qui a fourni une relation:

$y_i = a.x_i + b + e_i = y_i^* + e_i$ avec $e_i \in N\{0, s_e = s_y.\sqrt{1-r^2}\}$

On prend cette relation comme référence (i.e. on suppose que c'est la relation qui vaut sur la population), et on va regarder ce que l'on peut obtenir sur des échantillons qui respectent exactement cette structure que l'on vient de "figer".
Par exemple, pour une taille d'échantillon souhaité P, on réalise successivement:

- *pas 1:* Tirage au hasard d'une valeur de x dans une loi $N(m_x, s_x)$
- *pas 2:* Calcul de la partie expliquée de y par $y^* = a.x.+b$
- *pas 3:* Tirage au hasard d'une valeur de e dans une loi $N(0, s_e)$
- *pas 4:* Calcul de la valeur de y par $y = y^*+e$ et retour au pas 1

et on itère P fois cette opération.

Sur l'échantillon obtenu, on recalcule la corrélation. Bien qu'elle ait été **générée** selon la structure $y = a.x + b + e$ avec une corrélation r, l'ajustement fournit $y = a'.x +b' + e'$ avec une corrélation r'. On répète cela pour différents échantillons générés de taille P et on peut ainsi mesurer l'incertitude sur a, b, ainsi que sur r et e, due à un échantillon de taille P.

Remarque:

Pour des détails sur le tirage aléatoire dans une loi préfixée, on se reportera par exemple au cours polycopié d'analyse numérique (Ch. Obled 1978).
La pratique préconisée ici est programmée dans le petit logiciel de corrélation CORSIM (proposé par Ph. Bois).

III) PIEGES DE LA CORRELATION

Avec l'avènement de calculettes puissantes et de logiciels largement diffusés, la corrélation est devenue banale, avec le risque de l'utiliser comme une technique "presse-bouton".

Or il y a des pièges à éviter. Et ils sont très nombreux: certains sont "classiques", mais d'autres moins évidents.

Nous évoquerons les plus courants rencontrés en hydro-climatologie.

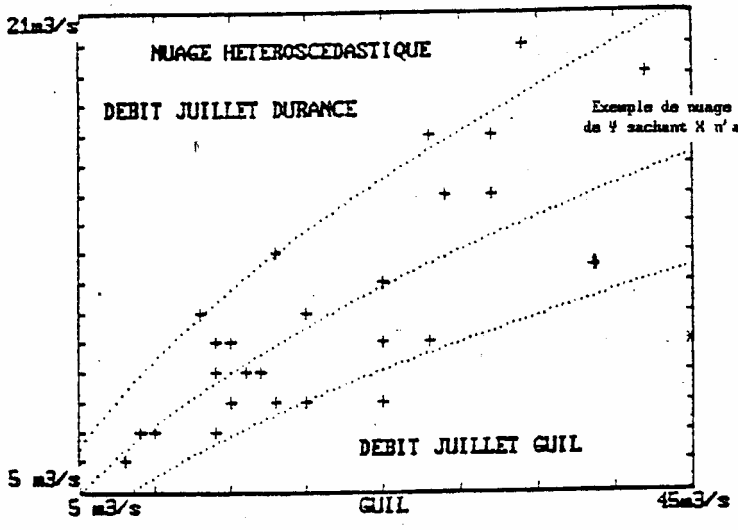
III-1) Pièges géométriques:

Ils sont dus à une forme particulière du nuage de points, et facilement décelables **si** *on prend la précaution de dessiner le nuage de points sur le plan X, Y.*

⇒ Ce sera donc une **règle** de ***toujours visualiser le nuage des observations***.

Exemples:

- Nuage *hétéroscédastique* (l'hypothèse binormale n'est pas vérifiée; contrairement à ce cas, vu en II, le résidu cette fois a une variance *fonction de X*).



Exemple de nuage hétéroscédastique : la loi conditionnelle de Y sachant X n'a manifestement pas un écart type constant

Figure 8 :

C'est le cas de nuage de corrélation entre deux stations de pluies journalières, de données journalières d'insolation, etc...

L'estimation de Y à partir de X doit alors se faire avec un résidu dont on ne peut pas considérer la variance comme indépendante de la valeur de X. Sinon, on fait une erreur sur les lois conditionnelles.

Solutions possibles:

Essayer de rendre la distribution plus binormale, par des transformations de variables (du type racine carrée, Log etc..).

III-2) Pièges de cofluctuation:

Très classiques en Hydrologie, en Géophysique...

Ils consistent à analyser très soigneusement et à utiliser comme potentiel prédictif une information qui est en fait triviale! Ils sont dus au fait que de nombreuses variables naturelles ont des composantes saisonnières, liées à la rotation de la terre autour du soleil.

Par exemple, si on calcule la corrélation entre la série des débits mensuels de l'Isère à Grenoble et la série des débits mensuels du Niger à Bamako au Mali, la corrélation est assez bonne.
Ceci n'est dû qu'aux variations saisonnières :

- il pleut en été sur le bassin du Niger -d'où des hautes eaux d'été-, à cause de la position du front de convergence tropicale
- tandis que sur l'Isère on assiste à une fusion nivale et glaciaire d'été.

Mais, à part cet effet saisonnier, il n'y a aucune relation physique entre les deux ...!, et une année donnée, il n'y a rien à gagner à tenter de s'appuyer sur les débits de l'Isère (-plutôt hauts en été -) pour prédire ceux du Niger (-eux aussi plutôt hauts en été-).

Solutions possibles:

Désaisonnaliser les variables, soit en travaillant par saisons, soit en enlevant de chaque variable la composante saisonnière en moyenne et écart type:

$$Q(\text{mois } i, \text{année } j) \Rightarrow q(i,j) = \frac{Q(i,j) - \overline{Q_i}}{s_{Q_i}}$$

avec $\overline{Q_i}$ = moyenne des mois i et s_{Q_i} = écart-type des mois i

III-3) Variables monotones:

Si X et Y sont des variables monotones (fonctions monotones d'une troisième variable, par exemple du temps), la corrélation sera toujours bonne même si ces variables n'ont aucune liaison physique.

Il s'agit d'échantillons où le couple X,Y est constitué de variables "fabriquées" de telle sorte qu'elles ne peuvent être que systématiquement croissantes ou décroissantes, (-souvent par le biais de cumuls -).

Exemple:

X(mois i, année j) = Volume de sédiments piégés par le barrage de Serre Ponçon depuis sa création jusqu'à ce mois i de l'année j

Y(mois i, année j) = Population de la Chine (actuellement c'est une variable monotone croissante).

La corrélation entre X et Y est alors très forte!, mais sans causalité physique aucune.

Solutions possibles:

Travailler sur des dérivées, ou en pratique, des **incréments:**

$y_i - y_{i-1}$ en fonction de $x_i - x_{i-1}$

C'est ainsi que l'on pourra constater que l'accroissement du volume de sédiments déposés dans Serre Ponçon n'est pas du à l'érosion induite par la population chinoise et n'est donc pas corrélé avec l'accroissement démographique chinois...!

III-4) Variable influente cachée:

Certaines corrélations peuvent paraître étonnantes.

Par exemple, il y a une bonne corrélation entre le nombre de morts de froid en hiver en France et la consommation de chauffage (-plus on chauffe donc, plus il y a de morts de froid..!-); on devine qu'une variable cachée (la température de l'hiver) a une influence primordiale.

Solutions:

Voir chapitre suivant 2ème Partie, Chap. III sur la Corrélation Multiple, paragraphe concernant la corrélation partielle.

III-5) Corrélation et liaisons de cause à effets:

Se rappeler qu'une bonne corrélation entre variables ne démontre pas l'existence obligatoire d'une liaison physique de cause à effets. Il ne s'agit que d'une constatation statistique. Seul le physicien peut trancher cette question.

Exemple I:

On constate en France une bonne corrélation entre le taux de boisement et les précipitations; mais n'en déduisons pas rapidement, comme on l'a parfois écrit, que la forêt augmente les précipitations!.

Il se peut que l'on ait simplement décidé de laisser se reboiser les zones trop arrosées, ou que ces zones arrosées soient plutôt situées en montagne donc peu accessibles pour la mécanisation de l'agriculture, etc... Mais la corrélation reste un fait observé.

Exemple II:

On propose un autre exemple (-fictif mais plausible-) de variable cachée pouvant entraîner une interprétation erronée.
On suppose que l'on a rassemblé des statistiques sur la longévité (durée de vie) en fonction de la consommation d'alcool. Cette enquête , "sponsorisée " par une grande marque de boissons alcoolisées, a couvert par exemple tout le continent américain.
Un premier calcul a conclu que, même si elle est assez modeste et surtout, très hétéroscédastique (cf. Figure 9) , la corrélation n'en est pas moins positive et significative ($\Rightarrow$ donc on vivrait d'autant plus longtemps que l'on consomme plus d'alcool …!)

Pourtant, en considérant plus attentivement l'origine et la répartition des individus, on s'aperçoit qu'ils sont organisés par sous-populations :

- les populations "développées ", qui ont d'ailleurs en moyenne une assez forte espérance de vie, des moyens économiques qui leur permettent de consommer beaucoup de spiritueux mais aussi de se faire soigner, une partie de la mortalité étant plutôt due aux accidents de la circulation.
- les populations dites "en développement ", qui ont une espérance de vie moindre et ne disposent que de moyens économiques modestes pour consommer, notamment de l'alcool, et pour se soigner.
- les populations "sous-développées et paupérisées" (tribus indiennes par exemple) , qui ont une espérance de vie très faible, n'ont pas vraiment accès aux soins et sont très vulnérables aux effets de l'alcool.

En fait, dans chacune de ces sous-populations prise séparément, la corrélation, et donc l'effet de l'alcool sur la longévité, … est comme on l'attend très négatif… C'est le regroupement abusif en une seule population qui fait apparaître une corrélation "irréaliste" en terme de causalité (donc au niveau de l'interprétation) , mais bien réelle au niveau du calcul strict…

Figure 9 :

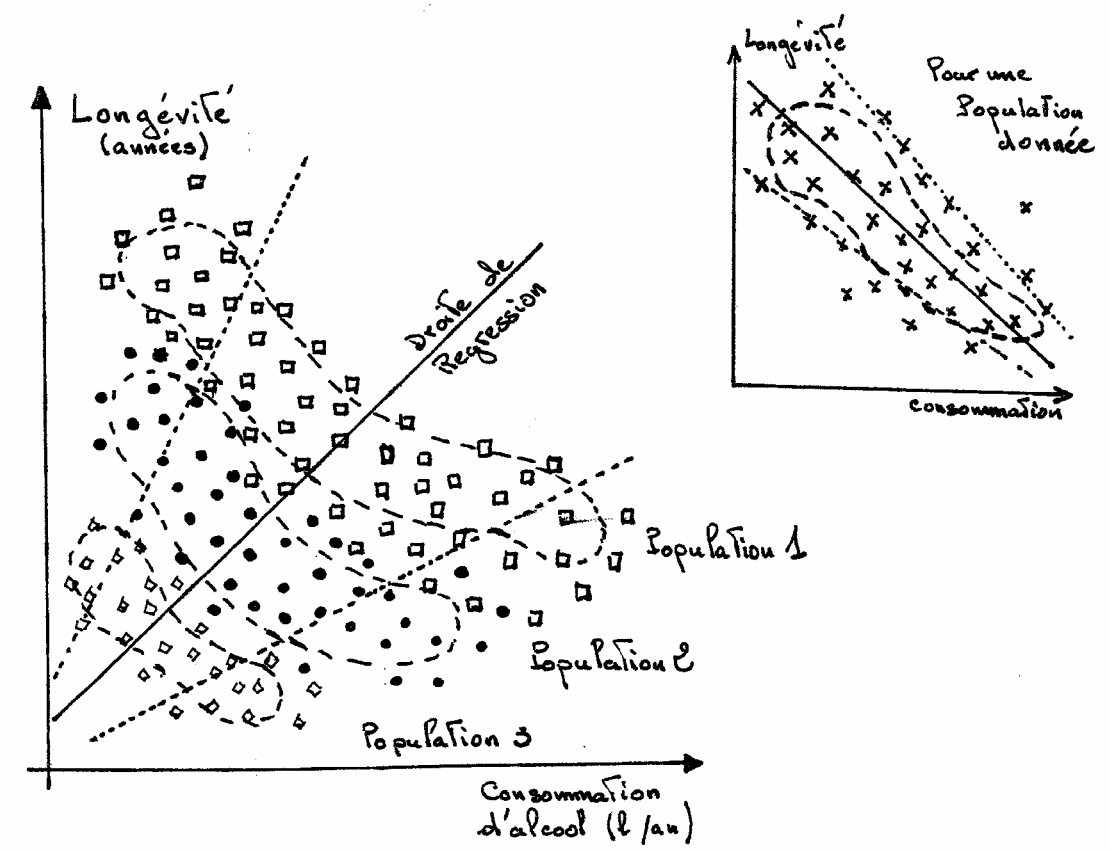
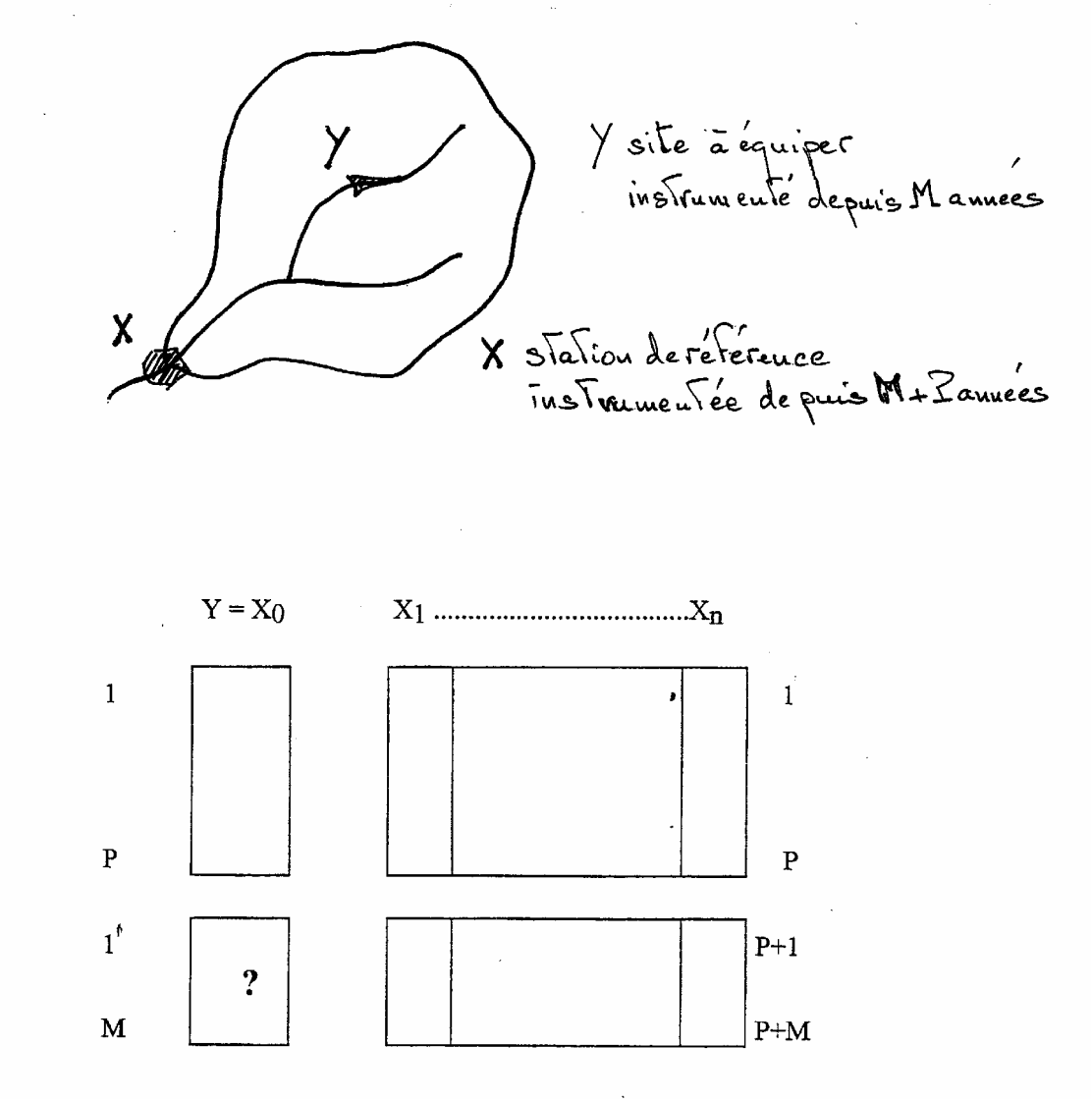


Figure 10 :


Y
X
Y site à équiper
instrumenté depuis M années
X station de référence
instrumentée depuis M+P années
Y = X0
X1 .....Xn
1
P
1'
M
?
1
P
P+1
P+M

IV) APPLICATIONS PARTICULIERES:

IV-1) RECONSTITUTION de DONNEES. EXTENSION de SERIES.

La corrélation est un outil très utile en hydrologie: on en verra un exemple approfondi en "Critique des données" (cf. 3ème Partie de ce cours d'analyse des données). Nous évoquons ici un problème fréquent aussi en hydrologie opérationnelle:

"Compléter une série courte à partir d'une série longue (extension de série)"

Exemple I:

C'est le cas par exemple si l'on veut dimensionner un ouvrage de stockage. Souvent, on a installé une station de mesure de débits sur le site de l'ouvrage seulement quand la décision d'étudier l'ouvrage a été prise $\Rightarrow$ de ce fait, la série collectée sur le site de l'ouvrage est souvent trop courte pour le dimensionner.(cf. Figure 10 ci-contre)

Mais ce site peut se trouver à proximité relative d'une station du réseau de base, exploitée depuis longtemps, mais qui ne draine pas forcément le même bassin... Il y a pourtant une certaine corrélation entre les deux, et ce serait intéressant de l'exploiter pour augmenter l'information disponible au site de l'ouvrage.

Exemple II:

Toujours pour un dimensionnement d'ouvrage, on dispose d'une série courte de débits, et donc de modules annuels, mais on a une longue série pluviométrique à proximité: peut-on étendre la série des modules de débits?

Hypothèses:

- on a des mesures communes sur **K** années aux stations X et Y
- mais une série de **N** années (N>K) à la station X (donc N-K années supplémentaires)

a) mise en oeuvre:

La corrélation entre les stations Y et X sur les K observations de la période commune fournit une équation de régression:

$$y_i^* = m_K(y) + r_K(x,y).\frac{s_K(y)}{s_K(x)}.\left[x_i - m_K(x)\right]$$

que l'on peut ensuite appliquer aux N-K valeurs observées de x_i de la période où elle est seule disponible, pour fournir N-K estimations y^*_i de la variable Y.

On peut maintenant s'interroger sur l'intérêt d'une telle pratique, proposée par Matalas et Jacobs (1964).

b) *Gain d'information* sur la moyenne μ_Y:

On dispose désormais de 2 estimations:

Estimation 1 : $\frac{1}{K}.\sum_{j=1}^{K} y_j \quad \Rightarrow \quad m_K(y)$

On sait que l'incertitude d'échantillonnage peut être exprimée par la variance théorique de cette estimation: $\sigma^2_{m_K} = \frac{\sigma^2_Y}{K}$

Estimation 2 : $\frac{1}{N}.\sum_{l=1}^{N} y_l = \frac{1}{N}.\sum_{j=1}^{K} yj + \frac{1}{N}.\sum_{i=1}^{N-K} y_i^* \quad \Rightarrow \text{m*(y)}$

Tous calculs faits, cette seconde estimation fournit:

$$m*(y) = m_K(y) + r_K(x,y)\frac{s_K(y)}{s_K(x)}.\left[m_{\underset{\Uparrow}{N}}(x) - m_{\underset{\Uparrow}{K}}(x) \right]$$

ou encore

$$m*(y) = m_K(y) + \frac{\text{N - K}}{\text{N}}.r_K(x,y).\frac{s_K(y)}{s_K(x)}.\left[m_{\underbrace{N-K}_{\Uparrow}}(x) - m_{\underset{\Uparrow}{K}}(x) \right]$$

en appelant $m_{N\text{-}K}(x)$ la moyenne des X calculée sur la période N-K où Y n'est pas connue.

La variance de cette estimation, calculée sur la série étendue a été proposée par Cochran (1953)

$$\sigma^2_{m*(y)} = \underbrace{\frac{\sigma^2_Y}{K}}_{\sigma^2_{m_K(y)}}.\left[1 + \left(1 - \frac{K}{N} \right).\left(\frac{1-(K-2).r_K^2}{K-3} \right) \right]$$

Le **gain de précision**, encore appelé l'**efficacité de l'extension** sur l'estimation de la *moyenne (car cette efficacité va différer selon le paramètre statistique que l'on considère… ici la moyenne),* s'exprime par:

$$E = 1 + \left(1 - \frac{K}{N} \right).\left[\frac{1-(K-2).r_K^2}{K-3} \right]$$

et s'interprète comme l'augmentation du nombre équivalent d'observations. Au lieu de K observations, la moyenne a une précision comparable à celle tirée de

$$N' = \frac{K}{E} \text{ observations, avec K < N' < N}$$

Interprétation intuitive:

Si la *corrélation* est *parfaite*, tant dans la population que dans l'échantillon, (r =1), alors on reconstitue parfaitement la série Y, donc on retrouve de fait $E = \frac{K}{N}$ d'où N' = N informations indépendantes (qui sont en fait $y^*_i \equiv y_i$).

Si au contraire, la *corrélation* est *nulle* tant dans la population que dans l'échantillon (r = 0), alors on ne reconstitue rien de la série Y. On remplace les valeurs manquantes (cf. la "remarque évidente " en fin du paragraphe II) , par N-K "estimations" qui ne sont que:

$y^*_i \equiv m_K(y)$ c'est à dire … la moyenne des seules valeurs observées !

⇒ donc on ajoute à la série des *K* valeurs de Y observés *N-K* fois *la moyenne de cette série...!* , et on croit avoir apporté de l'information…. !

En fait, on a fait pire que mieux, puisque l'on diminue la variabilité sans ajouter quoique ce soit, *mais* tout en pensant avoir des informations plus nombreuses...!.

Vérification : (≠ démonstration)

Si on reprend la formule: $E = 1 + \left(1 - \frac{K}{N}\right).\left[\frac{1-(K-2).r_K^2}{K-3}\right]$

et que l'on fait : $r_K = 1$

on trouve: $E = 1 + \left(1 - \frac{K}{N}\right).\left[\frac{1-(K-2)}{K-3}\right] = \frac{K}{N}$ d'où $N' = \frac{K}{E} = N$

et c'est bien le résultat attendu !

Si, par contre: $r_K = 0$

on trouve $E = 1 + \left(1 - \frac{K}{N}\right).\left[\frac{1}{K-3}\right]$

ce qui impose d'une part K>3 et E est toujours > 1 d'où N' < K ...!,
et …on a effectivement fait ***pire que mieux***...!

Comme $N' = \frac{K}{E}$, ⇒ l'opération ne vaut la peine que si E <1, pour avoir N' > K, et donc la limite, pour E = 1, correspond à:

$$E = 1 = 1 + \left(1 - \frac{K}{N}\right).\left[\frac{1-(K-2).r_K^2}{K-3}\right]$$

soit encore: $1-(K-2).r_K^2=0$ ou $r_K=\pm\sqrt{\frac{1}{K-2}}$

Et il faut que r_K soit supérieur à cette valeur pour améliorer l'estimation de la moyenne.

c) *Gain d'information* sur la variance σ_y:

On pourrait faire le même raisonnement sur l'estimation de la variance. En effet, on a la relation:

Var(Y) = Var. *expliquée par X* + Var. *résiduelle*

soit sur la population:

$$\sigma_y^2=\left(\rho_{xy}^2.\frac{\sigma_y^2}{\sigma_x^2}\right).\sigma_x^2+\left(1-\rho_{xy}^2\right).\sigma_y^2$$

Si on estime la variance y à partir de l'échantillon, on obtient $\mathrm{var}_K[Y]$

Mais on pourrait calculer avec une meilleure précision l'estimation de la variance de X soit :

$$\mathrm{var}_N(X)=\frac{1}{\underbrace{N}_{\Uparrow}-1}.\sum_{i=1}^{N}\left(x_i-m_{\underbrace{N}_{\Uparrow}}(x)\right)^2$$

donc on peut essayer d'utiliser cette meilleure estimation pour améliorer var[Y] par:

$$\mathrm{var}^*[Y]=\left[r_K^2(x,y).\frac{\mathrm{var}_K[Y]}{\mathrm{var}_K[X]}\right].\mathrm{var}_{\underbrace{N}_{\Uparrow}}[X]+\left[1-r_K^2(x,y)\right]\mathrm{var}_K[Y]$$

ce qui, après simplification, fournit:

$$\mathrm{var}^*[Y]=\mathrm{var}_K[Y]+r_K^2(x,y).\frac{\mathrm{var}_K[Y]}{\mathrm{var}_K[X]}\left[\mathrm{var}_{\underbrace{N}_{\Uparrow}}[X]-\mathrm{var}_K[X]\right]$$

Matalas et Jacobs (1964) proposent plutôt :

$$\mathrm{var}^*[Y]=r_K^2(x,y).\frac{\mathrm{var}_K[Y]}{\mathrm{var}_K[X]}\mathrm{var}_N[Y]+\left[1-\frac{N-3}{(K-3).(N-1)}\right]\left[\mathrm{var}_K[Y]-r_K(x,y)\sigma_K[Y]\sigma_K[X]\right]$$

On peut de même calculer la variance d'échantillonnage de cet estimateur et vérifier les conditions pour qu'il soit inférieur à $\mathrm{Var}_K(y)$ (Stedinger et Vogel 1985)

Attention:

Ces formules sont à utiliser avec précaution. Il est exclus de les justifier entièrement ici, et on conseille à l'utilisateur de se rapporter aux auteurs originaux s'il doit en faire un usage intensif.

IV-2) Traitement de Données de Mesures.

Il est fréquent que cette méthode (la régression) soit utilisée pour traiter des données de mesures.
Il arrive notamment que certains couples $\{x_i, y_i\}$ soient considérés comme plus fiables que d'autres. En d'autres termes, on a une *"mesure" de qualité pour l'observation i,* et on voudrait en tenir compte dans la corrélation en donnant plus de poids à ce couple.

Astuce:

Si on suppose que la qualité varie de 1 à 10, on peut fabriquer un nouvel échantillon de taille N' dans lequel on *duplique* 10 fois les couples très fiables et où l'on ne fait apparaître qu'une fois un couple peu fiable. Cela donnera au premier un poids de 10 dans les calculs.

On utilise alors un programme de corrélation classique: le couple dupliqué 10 fois attirera plus à lui la droite de corrélation que celui qui n'apparaît qu'une fois.
Par contre, tous les résultats d'échantillonnage seront erronés car le programme croira disposer de N' >>N observations.
Mais c'est un bon truc préliminaire...

Une approche plus théorique consiste à accorder à chaque couple $\{x_i, y_i\}$ un poids ω_i et à calculer dans ce contexte la droite de corrélation pondérée.
Un cas fréquemment rencontré est celui où:

- le couple $\{x_i, y_i\}$ est le résultat de la répétition P fois de la même mesure,
- dont on a fait ensuite la moyenne pour fournir le couple $\{x_i, y_i\}$. Dans ce cas, il est à peu près équivalent soit de mettre les mesures individuelles, soit de dupliquer P fois le couple $\{x_i, y_i\}$, c'est à dire de lui donner un poids P.

On peut raffiner en tenant compte, pour donner un poids au couple $\{x_i, y_i\}$, de la *variance observée* sur les P mesures de Y , mais même aussi de X, car on n'est pas toujours sûr de pouvoir se repositionner à la même abscisse exactement pour chaque mesure.

Pour ces aspects de l'utilisation de la corrélation en traitements des mesures, on se reportera à des ouvrages spécialisés comme celui de *CETAMA*

BIBLIOGRAPHIE

CETAMA (1986)
Statistique appliquée à l'exploitation des mesures.
Commission d'établissement des méthodes d'analyses du Commissariat à l'Energie Atomique. 2 ème Edition. Masson ed. 444 p.

COCHRAN W.G. (1953)
Sampling techniques . John Wiley Ed. New York

JOHNSTON (1974)
Econometric methods . John Wiley Ed. New York

MARQUARDT D. (1963)
An algorithm for least-squares estimation of non-linear parameters
J. Soc. Indust. Appl. Math, Vol 11, N°2

MATALAS N.C et JACOBS B. (1964)
A correlation procedure for augmenting hydrologic data
US Geol. Survey; Professional papers 434-E, 7 p

MORAN M.A. (1974)
On estimators obtained from a sample augmented by multiple regression
Water . Res. , vol 10, N°1, pp. 81-85

OBLED Ch. (1978)
Méthodes Numériques pour l'Ingénieur Hydraulicien.
Cours polycopié ENS Hydraulique Grenoble , 200 p.(Dernière réédition 1992)

ROCHE M. (1965)
Hydrologie de Surface. Gauthier Villars Ed. Paris.

VIALAR J. (1955)
Calcul des Probabilités et Statistiques- T III: Statistique, contingence et corrélation.
Dir. de la Météo. Nationale. Ecole Nationale de la Météorologie Toulouse. (Réédition 1986).

VOGEL R.M. and STEDINGER J.R. (1985)
Minimum variance streamflow record augmentation procedures
Water Ressour. Res. , vol 21, N°5, pp. 715-723

YEVJEVICH V.(1972)
Probability and Statistic in Hydrology
Water resources Publications, Fort Collins, Colorado U.S.A.

2éme Partie: LIAISONS STOCHASTIQUES ENTRE VARIABLES

CHAPITRE V : LA CORRELATION LINEAIRE MULTIPLE

I) OBJECTIFS ET NOTATIONS 175

II) CARACTERISTIQUES DE LA CORRELATION: 177

II-1) Critère d'optimisation 177
II-2) Calcul des coefficients de régression en variables centrées réduites et en variables brutes 177

III) LA CORRELATION PARTIELLE 181

III-1) Objectifs 181
III-2) Calcul des coefficients de corrélation partielle 184

IV) ESTIMATIONS SANS BIAIS: 186

IV-1) Coefficient de corrélation multiple débiaisé 186
IV-2) Fluctuations d'échantillonnage 187

V) CAS DE 2 VARIABLES EXPLICATIVES 190

VI) RAPPELS IMPORTANTS SUR LES NOTATIONS 191

VII) DIVERS ALGORITHMES INTERESSANTS 191

VIII) EXEMPLE DE CALCUL 193

2éme Partie: LIAISONS STOCHASTIQUES ENTRE VARIABLES

CHAPITRE V: LA CORRELATION LINEAIRE MULTIPLE

Objectifs:

On cherche à estimer une variable X_1 (que l'on appellera Variable à expliquer), par un lot de p-1 variables X_2, X_3,...,X_p, (appelées Variables explicatives) par l'intermédiaire d'une liaison linéaire du type:

$$\hat{X}_1 = \sum_{j=2}^{j=p} b_{1j,2\ldots p}\ X_j + c \qquad (1)$$

Les caractéristiques de cette liaison linéaire (valeurs optimales des coefficients, qualité de la liaison) seront estimées à partir d'un échantillon de n observations, i de 1 à n..

***Note importante** : En anglais, la variable à expliquer s'appelle « dependent variable » et les variables explicatives « independent variable » si bien que certaines personnes croient dur comme fer que cette méthode ne s'applique que si les variables « explicatives » sont indépendantes, ce qui est totalement faux*.

Applications:

+ Reconstitution de données manquantes
+ Modèles de prévision (étiages, crues, etc...)
+ Contrôle de données
+ etc..

Une étape importante du travail sera de proposer éventuellement des changements de variables, à partir des variables brutes, afin que la variable à expliquer puisse raisonnablement être expliquée par une liaison linéaire des variables explicatives.

V-I) Notations:

Soit: X_1 la variable à expliquer
X_2, X_3,..., X_p, les p-1 variables explicatives
$X_j(i)$ est la valeur de la variable X_j dans l'observation i de l'échantillon de taille n.
r_{jk} coefficient de corrélation linéaire (ou encore coefficient de corrélation totale entre X_j et X_k), calculé sur l'échantillon.

R est la matrice de ces coefficients de corrélation totale; c'est une matrice symétrique semi définie positive.

$$R = \begin{pmatrix} 1 & r_{1\,2} & \ldots. & r_{1\,j} & \ldots & r_{1\,p} \\ r_{2\,1}=r_{1\,2} & 1 & \ldots & r_{2J} & \ldots & r_{2\,p} \\ \ldots & \ldots & & \ldots & \ldots & \ldots \\ r_{1\,j} & r_{2\,j} & \ldots & 1 & \ldots & r_{j\,p} \\ \ldots & \ldots & \ldots & .. & \ldots & \ldots \\ r_{1\,p} & r_{2\,p} & & r_{j\,p} & \ldots & 1 \end{pmatrix}$$

Δ est le déterminant de R
$\Delta_{j\,k}$ le mineur j, k de R
$\delta_{j\,k}$ le terme j k de la matrice R^{-1}, matrice inverse de R.

V-II) Caractéristiques de la corrélation:

II-1) Critère d'optimisation:

Nous retiendrons d'emblée (cf. corrélation simple) le critère des moindres carrés des écarts d'estimation, c'est à dire que nous cherchons à minimiser, sur l'échantillon :

$$\sum_{i=1}^{i=n}\left(X_1(i)-\hat{X}_1(i)\right)^2 \text{ avec : } \quad \hat{X}_1(i)=\sum_{j=2}^{j=p} b_{1j,2..p}X_j(i)+c \qquad (2)$$

Ce critère est assez bon et permet des calculs rapides.

II-2) Calcul des coefficients de régression en Variables centrées réduites et en Variables brutes

Pour des raisons de simplification de présentation, nous travaillerons sur des variables centrées réduites en effectuant des transformations linéaires simples ; en outre, dans la pratique, cela est conseillé car on peut ainsi comparer les coefficients de régression entre eux puisqu'ils ont même dimension et sont relatifs à des variables de même variance.

$$X_j(i) \quad \rightarrow \quad x_j(i)=\frac{X_j(i)-\bar{X}_j}{s_{X_j}} \text{ avec } \overline{X_j} \text{ moyenne des } X_j \text{ sur l'échantillon et } s_{Xj} \text{ écart type}$$

des mêmes valeurs,

Cette transformation est biunivoque si aucun écart type n'est nul. Les nouvelles variables x_j ont des moyennes nulles sur l'échantillon et des écart types égaux à 1 sur l'échantillon, de plus, elles n'ont pas de dimension.

Nous cherchons donc le terme constant γ et les p-1 coefficients $\beta_{1j,\,2....p}$ appelés coefficients de régression (en variables centrées réduites) de x_1 avec x_j , compte tenu de x_2 , x_3 , ..., x_p qui minimisent :

$$S=\sum_{i=1}^{i=n}(x_1(i)-\beta_{1\,2,3..p}x_2(i)-\beta_{1\,3,2...p}x_3(i)-...-\beta_{1j,2..p}x_j(i)-...-\beta_{1p,2...p-1}x_p(i)-\gamma)^2$$

soit :

$$S=\sum_{i=1}^{i=n}\varepsilon^2_{1,2\,3...p}(i)$$

$\varepsilon_{1,2\,3....p}(i)$ est le résidu de $x_1(i)$ expliqué par $x_2, x_3,\ldots,x_p$

S est donc une fonction de p paramètres:

- les p-1 coefficients de régression $\beta_{1k,2...p}$ (k de 2 à p)
- le terme constant γ

Calcul du terme constant en variables centrées réduites:

Minimisons S sur l'échantillon par rapport à γ :

$$\frac{\partial S}{\partial \gamma}=0$$

soit :

$$\sum_{i=1}^{i=n}(x_1(i)-\beta_{1\,2,3..p}x_2(i)-\beta_{1\,3,2...p}x_3(i)-...-\beta_{1j,2..p}x_j(i)-...-\beta_{1p,2...p-1}x_p(i)-\gamma)\ =0$$

comme les variables sont centrées réduites:

$$\sum_{i=1}^{i=n}x_j(i)=0 \text{ pour tout j}$$

donc $\gamma = 0$

la somme des résidus sur l'échantillon, est donc nulle; autrement dit, l'erreur moyenne est nulle (en valeurs algébriques) et l'hyperplan passe par le centre de gravité.

Calcul des coefficients de régression en Variables centrées réduites:

On a à résoudre le système de p-1 équations à p-1 inconnues (les p-1 $b_{1k,2.....p}$):

$$\frac{\partial S}{\partial \beta_{1j,2...p}}=0 \quad \text{pour j = 2, 3 ,..., p}$$

d'où p-1 équations:

$$\sum_{i=1}^{i=n}x_j(i)\left[x_1(i)-\sum_{j=2}^{j=p}\beta_{1j,2....p}x_j(i)\right]=0 \quad \text{pour j=2 à p}$$

que l'on peut écrire avec les notations précédentes:

$$\sum_{i=1}^{i=n}x_j(i)*\varepsilon_{1,2...p}(i)=0 \text{ pour j=2 à p}$$

On en déduit, les variables x_j et $\varepsilon_{1,2.....p}$ étant centrées, que la corrélation entre le résidu et toute variable explicative est strictement nulle sur l'échantillon.

Le système s'écrit de façon plus classique:

$$\sum_{i=1}^{i=n}x_1(i)*x_2(i)=\sum_{i=1}^{i=n}\left[x_2^2(i)\beta_{1\,2,2...p}+.....+\beta_{1k,2...p}x_2(i)*x_k(i)+...+\beta_{1p,2...p}x_2(i)*x_p(i)\right]$$

……..

$$\sum_{i=1}^{i=n}x_1(i)*x_j(i)=\sum_{i=1}^{i=n}\left[x_j(i)x_2(i)\beta_{1\,2,2...p}+.....+\beta_{1j,2..p}x_j^2(i)+...+...+\beta_{1p,2...p}x_j(i)*x_p(i)\right]$$

……

$$\sum_{i=1}^{i=n}x_1(i)*x_p(i)=\sum_{i=1}^{i=n}\left[x_p(i)x_2(i)\beta_{1\,2,2...p}+..+\beta_{1j,2...p}x_p(i)x_j(i)+.....+\beta_{1p,2...p}x_p^2(i)\right]$$

Or la corrélation entre $x_j$et x_k est la même que celle entre X_j et X_k, soit r_{jk}, puisque toute transformation linéaire laisse invariant le coefficient de corrélation totale entre les variables.

$$r_{jk}=\frac{1}{n}\sum_{i=1}^{i=n}x_j(i)x_k(i)$$

car les variables x_j sont centrées réduites; d'où un système linéaire de p-1 équations à p-1 inconnues (les $\beta_{1k,2\ldots p}$ k de 2 à p).

$r_{1\,2} = 1\ \beta_{1\,2,\ldots.} + r_{2\,3}\ \beta_{1\,3,\ldots..} + \ldots\ldots\ldots\ldots + r_{2\,p}\ \beta_{1\,p,\ldots..}$
$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$
$r_{1\,j} = r_{2j}\ \beta_{1\,2,\ldots..} + \ldots\ldots\ldots\ldots\ldots + r_{jk}\ \beta_{1k,\ldots..} + \ldots\ldots\ldots + r_{jp}\ \beta_{1\,p,\ldots..}$
$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$
$r_{1p} = r_{2p}\beta_{1\,2,\ldots..} + \ldots\ldots\ldots\ldots\ldots + r_{pk}\beta_{1k,\ldots..} + \ldots\ldots\ldots + 1*\beta_{1\,p,\ldots..}$

On retrouve que les coefficients de ces équations sont les termes de la matrice R de corrélation.

Donc si $\Delta_{1\,1}$ n'est pas nul (ce qui est le cas le plus courant), mais il existe des contrexemples, comme une variable fonction linéaire d'autres, telle la température moyenne et les températures min et max)

$$\boxed{\beta_{1j,2\ldots p} = -\frac{\Delta_{1j}}{\Delta_{11}} = -\frac{\delta_{1j}}{\delta_{11}}}$$

Coefficients de régression et terme constant en Variables Brutes:

Comme $X_j(i) = \overline{X}_j + x_j(i) * s_{X_j}$

si $b_{1\,j,2\ldots p}$ est le coefficient de régression de X_1 avec X_j compte tenu de X_2, X_3,...,X_p, c'est à dire en variables brutes, on a la relation :

$$b_{1j,2..p} = \beta_{1j,2..p}\frac{s_{X_1}}{s_{X_j}}$$

et le terme constant c vaut:

$$c = \overline{X}_1 - \sum_{j=2}^{j=p}\frac{s_{X_1}}{S_{X_j}}\overline{X}_j\beta_{1j,2..p}$$

D'où l'équation de régression en variables brutes:

$$\frac{\hat{X}_1(i) - \overline{X}_1}{s_{X_1}} = \sum_{j=2}^{j=p}\beta_{1j,2..p}\frac{X_j(i) - \overline{X}_j}{s_{X_j}}$$

avec, rappelons le:

$$\beta_{1j,2\ldots p} = -\frac{\Delta_{1j}}{\Delta_{11}} = -\frac{\delta_{1j}}{\delta_{11}}$$

Le calcul est donc simple, il suffit d'inverser la matrice de corrélation, matrice semi définie positive.

II-3) Qualité de la liaison:

Il nous reste à mesurer la qualité de cette estimation; Le plus simple est de calculer la corrélation linéaire entre X_1 et son estimé par l'équation de régression. Ce coefficient de corrélation totale entre X_1 et son estimé par l'équation de régression sera appelé **coefficient de corrélation multiple entre X_1 et le lot de variables explicatives X_2, X_3,....,X_p.**

Nous le noterons $R_{1,2\ 3....p}$ *(notez la place de la virgule en indice !)*

Les variables x_j étant centrées réduites:

De façon analogue au coefficient de corrélation totale :Variance($\varepsilon_{1,2\ ...p}$) = 1 - $R^2_{1,2...p}$

or :

$$\varepsilon_{1,2...p}(i) = x_1(i) - \sum_{j=2}^{j=p} \beta_{1j,2...p} x_j(i)$$

D'où la variance du résidu (ce résidu a une moyenne nulle):

$$\frac{1}{n}\sum_{i=1}^{i=n}\left[x_1(i) - \sum_{j=2}^{j=p} \beta_{1j,2..p} x_j(i) \right]\left[x_1(i) - \sum_{j=2}^{j=p} \beta_{1j,2..p} x_j(i) \right]$$

Or le résidu n'est corrélé avec aucune variable explicative, d'où la variance vaut:

$$\frac{1}{n}\sum_{i=1}^{i=n}[x_1(i)\ ]\left[x_1(i) - \sum_{j=2}^{j=p} \beta_{1j,2..p} x_j(i) \right]$$

$$= Variance(x_1) - \sum_{j=2}^{j=p} \beta_{1j,2..p} r_{1j} = 1 + \sum_{j=2}^{j=p} \frac{\Delta_{1j}}{\Delta_{11}} r_{1j} = \frac{\Delta}{\Delta_{11}}$$

D'où : $1 - R^2_{1,2...p} = \frac{\Delta}{\Delta_{11}}$

Ou encore : $\mathbf{R^2_{1,2..p} = 1 - \frac{1}{\delta_{11}}}$

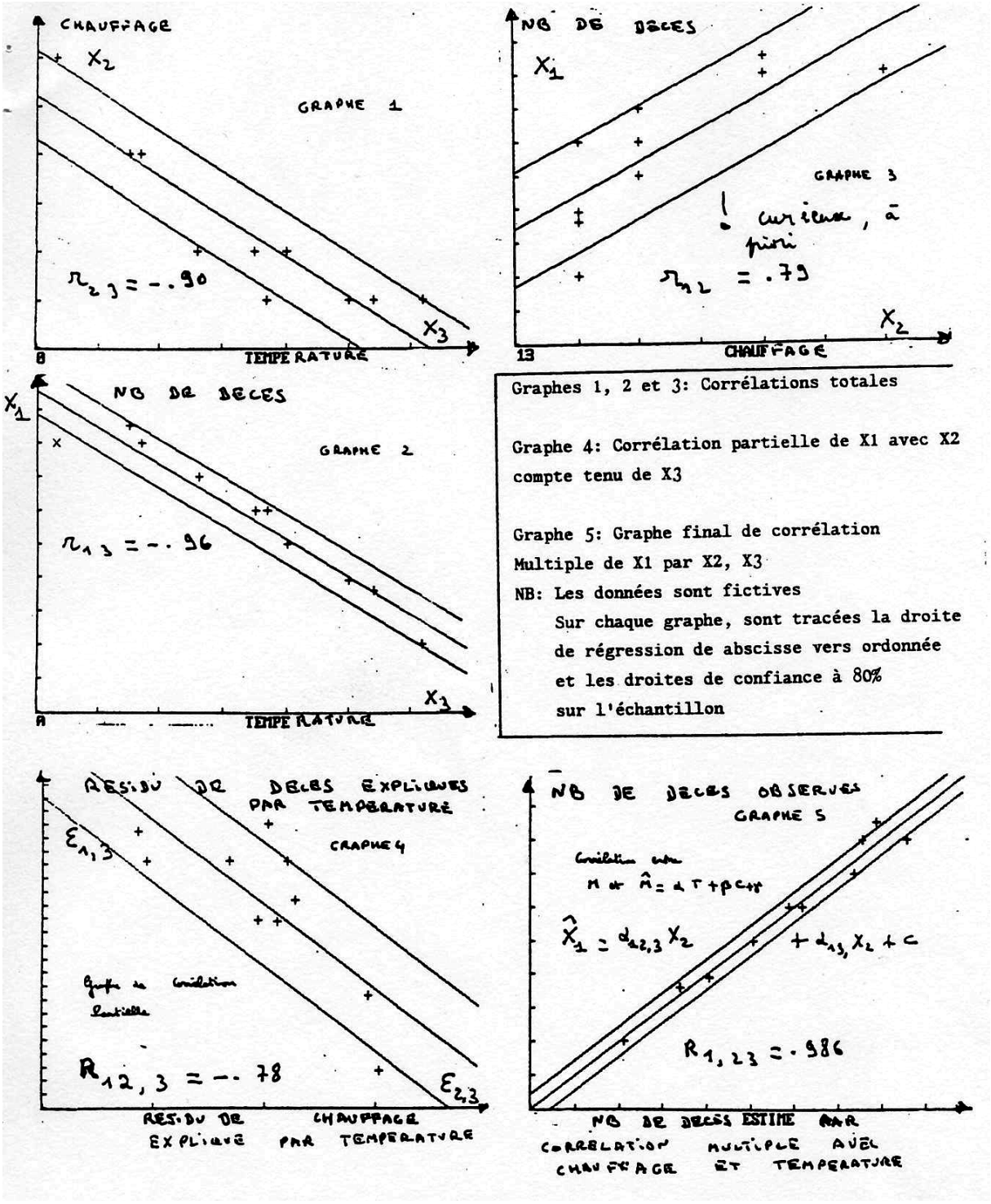
Expression plus usitée, car il est plus facile d'inverser une matrice que de calculer des déterminants.

V-III Corrélation partielle:

III-1) Objectifs:

La plupart des phénomènes sont causés par plusieurs variables plus ou moins liées et il est souvent difficile d'évaluer l'influence réelle d'une variable sur le phénomène à cause de la complexité des relations entre variables.

Prenons l'exemple des décès par cause de froid en hiver en France (exemple volontairement agressif, mais assez proche de la réalité). La corrélation entre le nombre de décès et la consommation de chauffage (cf Fig.) est positive ; plus on chauffe, plus on meurt!. Il est évident qu'il faut faire intervenir la température de l'hiver; la corrélation Chauffage-Température est élevée et négative. Plus il fait froid, plus on se chauffe. La relation Décès-Température est également bonne et négative, plus il fait froid et plus il y a de décès.



Graphes 1, 2 et 3: Corrélations totales

Graphe 4: Corrélation partielle de X1 avec X2 compte tenu de X3

Graphe 5: Graphe final de corrélation Multiple de X1 par X2, X3

NB: Les données sont fictives
Sur chaque graphe, sont tracées la droite de régression de abscisse vers ordonnée et les droites de confiance à 80% sur l'échantillon

Il est intéressant de savoir quelle est l'influence de la consommation du chauffage sur le nombre de décès, compte tenu de la température.

Si on possédait de très nombreuses observations, on pourrait regrouper les hivers de températures voisines et calculer pour ces hivers la corrélation Consommation-Décès et refaire ce calcul pour différentes températures.

Malheureusement, on ne possède qu'un nombre assez restreint d'observations.

L'idée est alors la suivante:

But: Chercher l'influence de X_2 sur X_1**, compte tenu de X_3**

Exemple: $X_1(i)$ = Décès de l'hiver i
$X_2(i)$ = Chauffage de l'hiver i
$X_3(i)$ = Température de l'hiver i

Méthode:

1) Retirons de X_2 l'influence de X_3. Autrement dit, on va écrire qu'une partie de X_2, X'_2 celle qui nous intéresse n'est pas expliquée (au sens de la corrélation) par X_3:

$X_2 = a X_3 + b + X'_2$

Nous venons tout simplement d'écrire l'équation de régression de X_3 en X_2.
X'_2 partie de X_2 non expliquée par X_3 est donc le résidu de la régression de X_3 en X_2.
$X_2 = \varepsilon_{2,3}$

2) Faisons le même travail pour X_1

La partie de X_1 non expliquée par X_3 est donc le résidu de la régression de X_3 en X_1.
Soit $\varepsilon_{1,3}$

3) La relation entre X_1 et X_2 compte tenu de X_3 est la relation entre la partie de X_1 non expliquée par X_3, soit $\varepsilon_{1,3}$ et la partie de X_2 non expliquée par X_3 soit $\varepsilon_{2,3}$. C'est ce que nous cherchons.

Nous calculerons donc la corrélation entre le résidu de X_1 par X_3 et le résidu de X_2 par X_3 et donnerons à ce coefficient le nom de **corrélation partielle de X_1 avec X_2, compte tenu de X_3.**

Dans le cas précédent, on aboutit à une corrélation partielle entre les décès et le chauffage, compte tenu de la température ; elle est négative (plus on chauffe, moins il y a de décès à température donnée) alors que la corrélation totale Décès-Chauffage était positive (plus on chauffe, plus il y a de morts.

La corrélation partielle est donc un outil très puissant:

+ pour le physicien (sens réel des relations entre 2 variables compte tenu des autres)

+ pour le choix des variables explicatives (une variable explicative ayant une corrélation partielle faible avec la variable à expliquer compte tenu des autres variables explicatives est de peu d'intérêt dans le schéma, même si la corrélation totale entre ces 2 variables est forte).

III-2) Calcul du coefficient de corrélation partielle $R_{1j,2...p}$

Cherchons la corrélation partielle entre x_1 et x_j compte tenu des variables x_2, x_3.....,x_p , (sauf xj évidemment) :

+ **Calcul du résidu de x_1 par x_2, x_3.....,x_p (sauf x_j)**:

$$\varepsilon_{1,2..p\ sauf\ j} = x_1(i) - \sum_{k=2}^{k=j-1} \beta_{1k,2...,j-1,j+1,...p} x_k(i) - \sum_{k=j+1}^{k=p} \beta_{1k,2...,j-1,j+1,...p} x_k(i)$$

où les $\beta_{1k,...}$sont les coefficients de régression de x_1 avec x_k pour le paquet des p-2 variables x_2,.....,x_{j-1}, x_{j+1},.....,x_p, coefficients différents de ceux d'avec le paquet total des variables explicatives x_2 à x_p .

+ **Résidu de x_j par x_2, x_3,....., x_{j-1}, x_{j+1},.....,x_p** :

$$\varepsilon_{j,2..p\ sauf\ j} = x_j(i) - \sum_{k=2}^{k=j-1} \beta_{jk,2...,j-1,j+1,...p} x_k(i) \quad - \sum_{k=j+1}^{k=p} \beta_{jk,2...,j-1,j+1,...p} x_k(i)$$

+ **Régression entre ces 2 résidus**:

$\varepsilon_{1,\ 2..p\ sauf\ j}(i) = b\ \varepsilon_{j\ ,2..p\ sauf\ j}(i) + \varepsilon_{1j\ ,2..p\ sauf\ j}(i)$ (3)

où le dernier terme est le résidu de la corrélation entre le résidu de x_1 par x_2,.....x_p sauf x_j et le résidu de x_j par x_2,.....x_p sauf x_j.

De la même façon, on pourrait écrire en intervertissant 1 et j:

$\varepsilon_{j,\ 2..p\ sauf\ j}(i) = b'\ \varepsilon_{1\ ,2..p\ sauf\ j}(i) + \varepsilon_{j1\ ,2..p\ sauf\ j}(i)$

équation de l'autre droite de régression. Nous avons vu que le produit des 2 termes multiplicatifs des 2 équations de régression est égal au carré du coefficient de corrélation, qui est le coefficient dont nous cherchons la valeur.

Ce calcul ainsi présenté serait laborieux; aussi allons nous nous ramener à des calculs déjà effectués.

(3) s'écrit également:

$\varepsilon_{1j,\ 2..p\ sauf\ j}(i) = \varepsilon_{1\ ,2..p\ sauf\ j}(i) - b\ \varepsilon_{j\ ,2..p\ sauf\ j}(i)$

Multiplions les 2 membres par $x_m(i)$ et faisons en la somme de i =1 à n

$$\sum_{i=1}^{n} x_m(i)\varepsilon_{1j,2...p(sauf\ j)}(i) = \sum_{i=1}^{n} x_m(i)\varepsilon_{1,2...p(sauf\ j)}(i) - b\sum_{i=1}^{n} x_m(i)\varepsilon_{j,2...p(sauf\ j)}(i) \quad (4)$$

si m est différent de 1 ou m différent de j, x_m est une variable explicative de chacune des corrélations ayant défini les résidus; or nous avons vu qu'une variable explicative n'est pas corrélée avec le résidu :

$$\sum_{i=1}^{n} x_m(i)\ \varepsilon_{1,2...p} = \sum_{i=1}^{n} x_m(i)\ \varepsilon_{j,2...p} = 0$$

Donc, si m est différent de 1 ou de j:

$$\sum_{i=1}^{n} x_m(i)\ \varepsilon_{1j,...} = 0$$

Or (4) s'écrit:

$$\sum_{i=1}^{n} x_m(i)\left[x_1(i) - \sum_{k=2}^{p\ sauf\ j} \beta_{1k,2...p\ sauf\ j} x_k(i) - b\left(x_j(i) - \sum_{k=2}^{p\ sauf\ j} \beta_{jk,2..p\ sauf\ j} x_k(i) \right) \right] = 0$$

On retrouve le système des p-1 équations du calcul des coefficients de régression de x_1 avec toutes les variables explicatives. b est donc le coefficient de régression de x_1 avec x_j compte tenu du lot x_2 à x_p des variables explicatives. Donc:

$$b = -\frac{\delta_{1j}}{\delta_{11}}$$

On pourrait faire de même pour calculer b' :

$$b' = -\frac{\delta_{1j}}{\delta_{jj}}$$

Or $R^2_{1j,2.....p \text{ sauf } j}$ est égal à bb'

D'où:

le coefficient de corrélation partielle entre x_1 et x_j, compte tenu de x_2, x_3,.....x_{j-1}, x_{j+1},.....,x_p a pour valeur :

$$R^2_{1j,2...p(\text{sauf } j)} = \frac{\delta^2_{1j}}{\delta_{11}\delta_{jj}}$$

On montre que le signe de $R_{1j,2.....p}$ est le signe du coefficient de régression de x_1 avec x_j compte tenu de x_2,.....x_p, soit le signe de :

$$-\frac{\delta_{1j}}{\delta_{11}}$$

V-IV Estimations sans biais

Tout ce que nous avons vu est strictement exact sur l'échantillon, mais ce qui nous intéresse souvent, c'est d'estimer au mieux, à partir d'un échantillon, les valeurs dans la population. Comme pour la corrélation simple, on montre que le coefficient de corrélation multiple et les coefficients de corrélation partielle précédemment définies sont biaisés; ils surestiment (en espérance mathématique) les valeurs dans la population. En effet, ils s'ajustent au mieux et notamment prennent comme moyennes des variables les moyennes au sein de l'échantillon et non les moyennes de la population (souvent inconnues) ; or si on se rappelle le théorème de Huyghens, le moment d'ordre 2 par rapport à un axe est minimum quand cet axe passe par le centre de gravité. Or, c'est ce que l'on fait en calculant les variances, notamment.

Aussi est il honnête de calculer les coefficients débiaisés, c'est à dire des coefficients qui, en moyenne, sont plus proches de ceux de la population.

IV-1) -Coefficient de corrélation multiple débiaisé:

Si n est le nombre d'observations supposées indépendantes et p le nombre total de variables (y compris la variable à expliquer), la valeur $R^*_{1,2.....p}$ du coefficient de corrélation multiple débiaisé est la suivante :

$$R^*_{1,2.....p} = \sqrt{\frac{(n-1)R^2_{1,2...p} - (p-1)}{n-p}}$$

Exemples :

Valeurs du coefficient de corrélation multiple débiaisé en fonction de la valeur du coefficient et du nombre d'observations et de variables.

		n=20	n=40	n=80
		R^*	R^*	R^*
R=.8	p=5	.74	.77	.79
R=.8	p=10	.56	.73	.77
R=.8	p=15	-	.66	.75
R=.95	p=5	.936	.944	.947
R=.95	p=10	.90	.934	.943
R=.95	p=15	.79	.921	.939

Conseils:

* éviter d'avoir un nombre de variables explicatives supérieur à la moitié du nombre d'observations.

* vérifier toujours si le résultat est donné en valeurs biaisées ou débiaisées, surtout si le nombre d'observations n'est pas très grand vis à vis du nombre de variables.

En effet, une mauvaise tendance naturelle est d'accroître le nombre de variables explicatives pour augmenter la corrélation.

IV-2) Fluctuations d'échantillonnage

Le problème est le suivant: on suppose que l'échantillon est tiré d'une certaine population, si on extrait de cette population plusieurs échantillons, les résultats de corrélation (R, les coeff. de régression...) vont être différents d'un échantillon à l'autre. Il est intéressant de connaître comment peuvent fluctuer ces différents coefficients.

Nous n'examinerons que le cas d'observations indépendantes et de variables normales.

+ Coefficient de corrélation multiple:

On montre que, avec les notations précédentes:

$$\frac{n-p}{p-1} * \frac{R^2_{1,2\ldots p}}{1-R^2_{1,2..p}} = F$$

F suit une loi de Fisher Snedecor à 2 paramètres de valeurs respectives p-1 et n-1. Cela est utile pour tester si la valeur R est significativement différente de zéro.

Exemples:

1) n=20 p=10 R=.6 (valeur biaisée)

On trouve F=.625 . F suit une loi de Fisher Snedecor avec 19 et 10 degrés de liberté (qui sont les 2 paramètres de la loi de Fisher). Or dans une table de cette loi, on trouve que la probabilité de dépasser une telle valeur est de 82%; autrement dit si l'échantillon avait été tiré d'une population sans corrélation, on aurait eu 82% de chance de tirer une valeur au moins aussi forte. La corrélation obtenue est donc probablement due au hasard.

2) n=20 p=10 R=.8 (valeur biaisée)

On trouve cette fois F=1.97 qui n'a que 13% de chance d'être dépassé; il y a donc de fortes chances que la corrélation obtenue sur l'échantillon ne soit pas le simple fait du hasard.

+ Fluctuation du coefficient de corrélation partielle:

On montre que:

$$t = \frac{R_{1j,2....p}}{\sqrt{1 - R_{1j,2....p}^2}} \sqrt{n - p}$$

suit une loi de Student à n-p degrés de liberté (seul paramètre de la loi de Student). Ceci permet de tester l'intérêt d'une variable explicative, compte tenu des autres: si la valeur de ce terme est faible, la variable n'a pas d'intérêt, toujours compte tenu des autres. On lira donc dans une table de Student la probabilité d'être plus grand.

+ fluctuations des coefficients de régression:

Variance:

Considérons une population où l'équation de régression s'écrit (en variables centrées réduites):

$$\hat{x}_1 = \beta_{12,2...p} x_2 + ... + \beta_{1p,2...p} x_p$$

Si on tire plusieurs échantillons de taille n et que l'on effectue sur chaque échantillon un calcul de corrélation multiple, on va trouver des résultats différents:

Echantillon 1:

$$\hat{x}_1^1 = \beta_{12,3...p}^1 x_2 + ... + \beta_{1p,3...p}^1 x_p$$

Echantillon k:

$$\hat{x}_1^k = \beta_{12,3...p}^k x_2 + ... + \beta_{1p,3...p}^k x_p$$

Les coefficients de régression ne seront pas égaux d'un échantillon à l'autre. Il est intéressant de savoir comment ils peuvent fluctuer (sous certaines hypothèses). On montre que:

$$\text{Variance de } \beta_{1j,2...p} = \beta_{1j,2..p}^2 \frac{1 - R_{1j,2...p}^2}{R_{1j,2...p}^2} * \frac{1}{n - p}$$

C'est à dire que le coefficient de régression de la variable x_j (compte tenu des autres) est d'autant plus stable:

- que la corrélation partielle de x_1 avec celle ci, compte tenu des autres est forte
-que la taille de l'échantillon est grande par rapport au nombre de variables explicatives.

Cas limite (fréquent en Hydrologie):

Si on prend 2 variables explicatives bien corrélées, la corrélation partielle de x_1 avec l'une d'elles, compte tenu des autres est faible, même si elle est bien corrélée (au sens de la corrélation totale) avec la variable à expliquer. Son coefficient de régression est donc fort instable et on peut même aboutir à des changements de signe d'un échantillon à l'autre (ce qui physiquement peut paraître curieux. En fait, il n'y a cohérence que sur l'ensemble des coefficients de régression.

Si par exemple, on cherche la corrélation des cumuls annuels de pluie mesurés sur le toit de l'ENSHMG avec comme variables explicatives les données de METEO France Saint Martin d'Hères et les données du pluviographe du CEA, on pourra trouver à la limite une relation du type :
Cumul ENSHMG=1.2*Météo France - .15 CEA +10 (mm). Ne pas en conclure que plus il pleut au CEA, moins il pleut à ENSHMG

Covariance des coefficients de régression:

Les coefficients de régression ne sont pas indépendants entre eux, il est possible de calculer les covariances (ce qui sort du cadre de ce manuel d'initiation).

V-V) Cas de 2 Variables explicatives:

C'est le cas le plus simple que l'on peut résoudre facilement avec une calculette.
Soit $r_{1.2}$, $r_{1.3}$, $r_{2.3}$ les 3 coefficients de corrélation totale.
On montre que:

le coefficient de corrélation multiple (biaisé) de X_1 avec X_2 et X_3 a pour expression:

$$R^2_{1,23} = \frac{r^2_{12} + r^2_{13} - 2r_{13}r_{12}r_{23}}{1 - r^2_{23}}$$

On voit sur cet exemple que R est d'autant plus fort que la corrélation entre variables explicatives est faible (à $r_{1.2}$ et $r_{1.3}$ constants).

Corrélation partielle entre X_1 et X_2 compte tenu de X_3:

$$R_{12,3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{(1-r^2_{23})(1-r^2_{13})}}$$

Corrélation partielle entre X_1 et X_3 compte tenu de X_2 :

$$R_{12,3} = \frac{r_{13} - r_{12}r_{23}}{\sqrt{(1-r^2_{23})(1-r^2_{12})}}$$

On remarquera que la valeur, comme le signe de $R_{1.2,3}$ n'ont rien à voir avec la valeur et le signe de $r_{1.2}$. Par exemple:

Avec pour les 3 cas : $r_{1\,2} = .9$ $r_{2\,3} = .9$
Cas 1: $r_{1\,2} = .7$ $R_{1\,2,3} = -.58$
Cas 2: $r_{1\,2} = .8$ $R_{1\,2,3} = -.05$
Cas 3: $r_{1\,2} = .9$ $R_{1\,2,3} = .47$

Coefficients de régression en variables centrées réduites:

$$\beta_{12,3} = \frac{r_{12} - r_{13}\,r_{23}}{1 - r^2_{23}}$$

$$\beta_{13,2} = \frac{r_{13} - r_{12}\,r_{23}}{1 - r^2_{23}}$$

V-VI) RAPPELS IMPORTANTS SUR LES NOTATIONS ET ANALOGIE AVEC LES DERIVEES PARTIELLES ET TOTALES :

$\mathbf{r_{jk}}$ est le coefficient de corrélation totale entre X_j et X_k
$\mathbf{R_{1,2...p}}$ est le coefficient de corrélation multiple de X_1 expliqué par X_2, …, X_p
$\mathbf{R_{1\ 2,3....p}}$ est le coefficient de corrélation partielle entre X_1 et X_2, compte tenu de X_3,…, X_p
$\mathbf{b_{1\ 2}}$ est le coefficient de régression de X_2 pour expliquer X_1 sans tenir compte d'autres variables
$\mathbf{b_{1\ 2,,3...p}}$ est le coefficient de régression de X_2 pour expliquer X_1 en tenant compte de X_3,…, X_p
Notez bien la place de la virgule dans les listes d'indices.

Les termes « totale » et partielle » correspondent tout à fait au sens que l'on donne entre les différentielles totales et dérivées partielles en Mathématiques.

V-VII) DIVERS ALGORITHMES INTERESSANTS :

VII-1) Sélection de variables explicatives

Bien souvent, on a le choix entre de nombreuses variables explicatives plus ou moins corrélées. Par exemple, si l'on cherche à expliquer la fusion nivale journalière d'un petit bassin en période de fonte, on pourra, de manière physique, dire qu'elle dépend de la température moyenne journalière mais aussi de la température max et de la température min, et de l'insolation et de la nébulosité et du rayonnement et du vent etc.. Ces variables sont plus ou moins liés.

Une méthode classique, qui sera développée en cours est la **sélection progressive ascendante pas à pas**(il en existe d'autres) :

On va procéder pas à pas :

Pas 1 : On prend parmi les variables explicatives possibles la plus utile au sens de la corrélation, c'est simple, c'est celle qui a le plus grand coefficient de corrélation totale avec la variable à expliquer.

Pas 2 : On cherche alors, parmi les variables explicatives restantes, la plus utile : c'est celle qui a le plus grand coefficient de corrélation partielle avec la variable à expliquer compte tenu de la première variable explicative retenue On calcule le coefficient de corrélation multiple débiaisé et on teste si le coefficient de corrélation partielle est significatif (à l'aide de la variable de Student précédemment décrite). On la garde si cela en vaut la peine et on continue.

Pas k : On a retenu k-1 variables explicatives significativement intéressantes et on cherche alors parmi les variables explicatives non encore retenues, celle qui le plus grand coefficient de corrélation partielle avec la variable à expliquer, compte tenu des variables explicatives déjà retenues. On teste si cela vaut la peine de l'ajouter en testant ce coefficient.

Arrêt : on s'arrête quand l'ajout d'une variable n'améliore rien et même fait baisser le coefficient de corrélation multiple débiaisé.

Piège : Le test que l'on fait n'est pas très adroit car on teste si la variable à ajouter vaut la peine, sans tenir compte du nombre de variables que l'on pouvait ajouter. Nous n'avons pas trouvé de test résolvant ce problème. Pour donner une image : un inspecteur veut connaître

rapidement le niveau d'une classe en Géographie, il peut prendre au hasard un élève et lui poser une question, mais souvent il demande à l'enseignant de lui désigner un élève au hasard pour lui poser la question. L'enseignant va évidemment désigner le meilleur élève en Géographie. Le résultat ne sera pas forcément le même !.

Conseils : Ne conserver que des variables vraiment utiles afin d'avoir un modèle simple et robuste.

VII-2) Validations

Par un échantillon mis en réserve :

Si l'on possède un assez grand nombre d'observations, il est prudent de caler le modèle sur une partie de l'échantillon et de valider le modèle sur une autre partie de l'échantillon n'ayant pas servi à caler le modèle (c'est évidemment plus méchant mais réaliste).

Par la méthode des « résidus supprimés ou Validation croisée(logiciel STATISTICA) :
On prend un échantillon de taille n et on enlève la première observation ; on cale le modèle sur les n-1 observations restantes et on l'applique à la première observation qui n'a pas servi au calage.
On refait cela pour chacune de n observations et finalement on a n résidus, certes calculés avec n modèles différents mais voisins.
Cette procédure est remarquable pour détecter des observations « bizarres » et elle donne bien ce que l'on obtient en opérationnel. En outre, elle est très rapide.

Pour conclure :

La méthode de corrélation linéaire multiple est une méthode rapide, honnête et fiable. Mais, rappelons q'il faut d'abord réfléchir (à partir de la connaissance des phénomènes):

- sur la forme de la liaison

En effet, il ne faut pas d'emblée prendre la forme linéaire, quitte à faire des transformations de variables pour proposer une forme linéaire sur les transformées. Par exemple, si on cherche à expliquer un volume de crue par la pluie, la durée de la pluie, le débit de base avant la crue, il est évident qu'un modèle linéaire est maladroit ; mieux vaut prendre un modèle multiplicatif Puissance qui correspond mieux aux lois de l'hydrologie.

- sur le choix et le nombre de variables explicatives :

Même en utilisant des logiciels performants, ne conserver que peu de variables explicatives mais utiles et ayant un sens. Pour donner un exemple, si on cherche à voir s'il y a une liaison entre les paramètres statistiques des pluies extrêmes et le relief, on aboutit vite à plus de cent variables explicatives possibles ; par le fait du hasard la corrélation multiple non débiaisée va être bonne. D'où l'intérêt de rester prudent en n'en prenant que peu mais vraiment utiles et en faisant de la validation.

V-VIII) EXEMPLE COMPLET :

Exemple : Prévision de crues

A) Objectifs :

A-1) Hydrologique :

Etude des relations Pluie Débit de crues d'un petit bassin versant soumis à de fortes pluies

A-2) Méthodologique :

Utilisation de la corrélation multiple

B) Documents :

Source des données : EdF. On dispose de 26 épisodes de crues d'une rivière des Cévennes (Sud de la France) ainsi que des données correspondantes de pluies horaires d'une station bien représentative du bassin. Dans cette région, les pluies peuvent être très fortes en quelques heures et les crues sont quasiment immédiates.

Crue N°	Pluie	Durée	Qbase	Point	IMX	TRETA	VOL24
1	163	12	0	138	48	7	5.2
2	61	14	28	120	10	10	6.3
3	26	4	13	60	5	2	3.1
4	43	6	15	70	11	3	4.4
5	124	18	47	520	21	10	24.6
6	51	9	9	85	13	4	3.5
7	36	17	17	75	6	6	5.0
8	76	14	60	650	21	13	27.0
9	47	8	210	970	24	2	28.9
10	47	7	135	315	13	5	18.9
11	41	6	35	115	7	4	7.2
12	68	10	1	65	8	5	1.9
13	87	7	10	435	33	6	13.2
14	39	30	8	108	5	18	3.5
15	79	31	14	275	12	26	11.3
16	54	10	100	275	12	6	14.3
17	54	21	13	110	7	14	5.4
18	90	24	10	165	10	14	7.9
19	24	9	69	118	8	1	8.7
20	39	13	69	245	7	12	12.8
21	131	8	5	1600	70	6	30.5
22	64	18	3	45	9	10	2.5
23	27	8	35	105	8	5	5.6
24	101	19	30	560	35	3	22.1
25	151	27	70	540	14	19	28.9
26	52	13	14	100	9	4	6.2
	en mm	en heures	en m3/s	en m3/s	en mm	en heures	en hm3

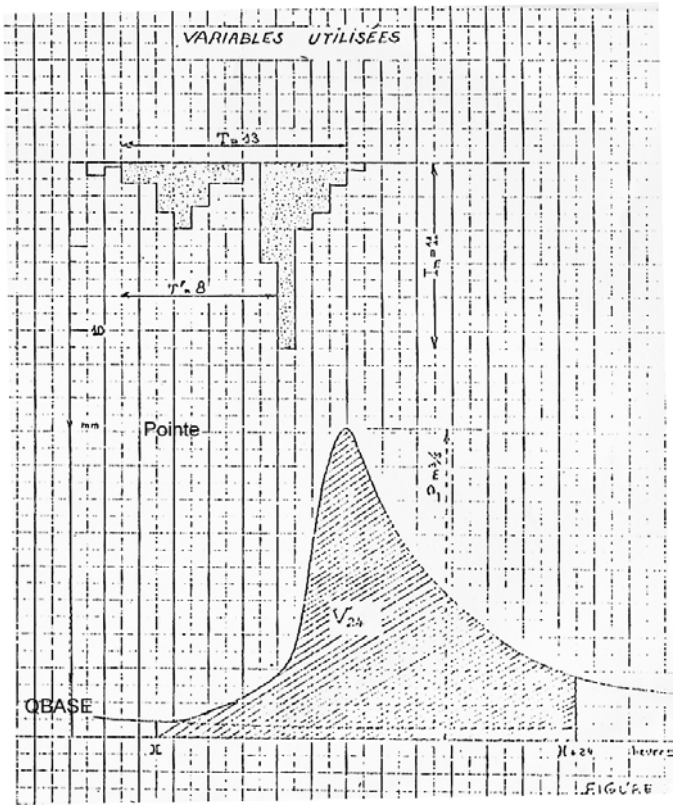
C) Description du problème :

On se propose d'établir deux modèles de prévision des crues (l'un pour les volumes, l'autre pour les débits de pointe). Une bonne méthode consiste à voir, dans une première étape, si les variables mesurées permettent d'expliquer, au sens statistique la variable à expliquer. Si ce n'est pas le cas, inutile d'essayer d'établir un modèle de prévision qui lui, n'utilisera que les variables connues au moment de la prévision et des variables plus ou moins bien prévues. Nous nous intéresserons à cette première étape.

Pour chaque épisode on connaît :

- le numéro de la crue
- Pluie : la pluie totale de l'épisode en mm
- Durée : la durée de la pluie en heures
- Qbase : le débit en m3/s de la rivière avant la pluie (donne une idée de la saturation du bassin
- Point : le débit de pointe en m^3/s
- IMAX : la pluie maximale horaire de l'épisode en mm
- TRETA : le temps en heures séparant le début de la pluie de la pluie horaire la plus forte (indique si la pluie la plus forte est tombée au début ou à la fin, important en hydrologie).
- VOL24 : le volume en 24 heures de la crue en hm^3

Note : le graphique joint explicite les variables



C-1) Construction des modèles :

En utilisant un logiciel de corrélation multiple linéaire à sélection de variables, proposer un schéma d'explication de la pointe et un autre du volume à partir des variables explicatives fournies. Ce schéma devra être directement utilisable avec une simple calculette et fournir :

- la valeur la plus probable (en m3/s ou en hm3, selon le modèle)

- l'intervalle de confiance à 80% de la variable à expliquer, ou mieux, les valeurs de probabilité au non-dépassementt de 10% et 90%, ceci, en valeurs brutes.

C-2) Applications

Appliquer vos modèles à deux cas bien différents, en donnant pour chaque réponse la valeur la plus probable et son intervalle de confiance à 80%, c'est à dire la valeur qui a 10% de chances de ne pas être dépassée et celle qui a 10% de chances d'être dépassée, ceci en valeurs brutes :

C-2-a) Pluie moyenne sur sol assez saturé :

Pluie=90 mm
Durée=8heures
Qbase= 10 m3/s
IMAX= 25 mm/h
TRETA= 5 heures

C-2-b) Pluie totale forte mais sur sol assez sec :

Pluie=160 mm
Durée=13 heures
Qbase= .2 m3/s
IMAX=40 mm/h
TRETA= 8 heures

Avant de faire des calculs savants, essayer d'estimer à l'œil les résultats.

Quelques conseils :

- Changement de variables : les logiciels simples ne traitent que des cas linéaires, aussi est-il peut être judicieux de travailler plutôt sur des variables transformées que sur les variables brutes pour construire un modèle plus réaliste d'un point de vue hydrologique.

- Ne conserver que quelques variables explicatives, les plus intéressantes (cela se verra avec les coefficients de corrélation partielle ou avec des tests sur les valeurs des variables de Student des variables explicatives).

- Coupez l'échantillon en deux parties et refaire les calculs. Il est possible que vous aboutissiez à des schémas différents. Est ce grave et pourquoi des variables explicatives parfois différentes ? .

- Utiliser la méthode de validation croisée pour voir ce que vous donnerait un schéma appliqué à des observations n'ayant pas servi au calage.

- Vérifier que les résidus du modèle sur les variables transformées sont à peu près gaussiens pour pouvoir calculer un intervalle de confiance.

Corrélation Multiple Exemple : Prévision de crues
Correction rapide :

C-1) Construction des modèles :

a) Transformation des variables :

Il est évident qu'un modèle linéaire du type POINT (ou VOL24) = Somme pondérée des variables explicatives n'a pas de grande valeur hydrologique ; par contre, on peut penser que le volume, comme la pointe sont à peu près fonction de produits des mêmes variables explicatives, élevées à une certaine puissance. Ceci dans une première approche, car, par exemple, pour la pluie on pourrait penser que c'est une fonction de la pluie diminuée d'une certaine quantité. Quant au débit de base, on peut penser qu'il donne une idée de l'état de saturation du bassin et qu'ainsi, il peut intervenir comme un facteur multiplicatif s'il est élevé à une certaine puissance.

Aussi va t on travailler sur les logarithmes népériens des variables pour pouvoir utiliser un modèle simple de corrélation multiple linéaire. Ce qui explique que par la suite les noms des variables commenceront par un L, car il s'agit des Logarithmes Népériens des données brutes. On reviendra à la fin sur un modèle multiplicatif puissance où les coefficients de régression du modèle logarithmique sont les exposants des variables correspondantes.

Tous les calculs suivants seront donc effectués sur les Log des variables.

b) Modèle explicatif du volume en 24 h :

On a tout d'abord effectué le calcul sans sélection de variables, avec comme variables explicatives :

LPLUI
LDUREE
LQB
LIMAX
LTRET

Et comme variable à expliquer : LVOL

On obtient de bons résultats (un coefficient de détermination élevé), mais certaines variables explicatives ont un coefficient de corrélation partielle, compte tenu des autres variables explicatives, non significatif. En effet, certaines variables explicatives sont corrélées et il serait inutile et maladroit de les conserver toutes.

Aussi, a-t-on refait le même calcul mais en choisissant une procédure de sélection ascendante ; c'est à dire qu'à chaque pas de calcul on ajoute une variable explicative et on s'arrête lorsqu'il n'y a plus de variable explicative intéressante, compte tenu de celles déjà retenues. Le test est effectué sur la variable de Student qui teste l'hypothèse nulle, à savoir quelle est la probabilité d'obtenir un coefficient de corrélation partielle au moins aussi fort avec une variable qui n'aurait rien à voir avec le problème. En général, on prend un seuil de l'ordre de 5%.

Sur les 26 observations (cf. Tableau 1), on obtient :

TABLEAU 1 :Synthèse Régression de la Var. Dépendante :LVOL (pcru.sta)
R= .961 R²= .924 R² Ajusté= .914
F(3,22)=89.258 p<.00000 Err-Type de l'Estim.: .24393

	BETA	Err-Type de BETA	B	Err-Type de B	t(22)	niveau p
OrdOrig.			-2.79428	.492778	-5.67048	.000011
LQB	.801467	.063508	.43401	.034391	12.61995	.000000
LIMAX	.441376	.089402	.51683	.104685	4.93700	.000061
LPLU	.373953	.094119	.58842	.148097	3.97318	.000644

Note : les B sont les coefficients de régression, les BETA sont les coefficients de régression en variables centrées réduites, Err-Type sont les écart types d'estimation. Le F est la variable de Fischer Snedecor calculée à partir du coefficient de détermination ; la loi de probabilité de F a deux paramètres fonction du nombre de variables du modèle et de la taille de l'échantillon. Il est ici très élevé. Le t() est la variable de Student, calculée à partir de la corrélation partielle et du nombre d'observations ; le niveau p est la probabilité d'avoir un meilleur coefficient de corrélation partielle (qui n'apparaît pas sur ce tableau) pour une variable indépendante. Ici toutes les variables retenues sont hautement significatives.

LVOL=-2.79+0.43401*LQB+0.51683*LIMAX+.58842*LPLU

En Log Népériens sur les unités préalablement définies.

Le coefficient de détermination non biaisé (c'est à dire celui qui tient compte du nombre de variables retenues et de la taille de l'échantillon) est le suivant :
R^2 = .914 , ce qui est très bon
L'écart type résiduel est de 0.244 (toujours en Log)

Tableau 2 : Valeurs Prévues & Résidus (pcru.sta)
Var. Dépendante : LVOL

	Valeur Observée	Valeur Prévue	Résidus	Standard Val.Prév	Standard Résidus	Err.Type Val.Prév	Mahalns. Distance	Résidus Supprim.	Cook Distance
1	1.6487	1.6021	.04660	-.70	.19	.1631	10.216	.0843	.013
2	1.8405	2.2609	-.42033	.13	-1.72	.0573	.420	-.4449	.046
3	1.1314	1.0134	.11800	-1.43	.48	.0987	3.133	.1411	.014
4	1.4816	1.8335	-.35189	-.41	-1.44	.0642	.771	-.3781	.042
5	3.2027	3.2866	-.08381	1.41	-.34	.1055	3.715	-.1031	.008
6	1.2528	1.8142	-.56143	-.43	-2.30	.0650	.814	-.6043	.109
7	1.6094	1.4250	.18441	-.92	.76	.0747	1.383	.2035	.016
8	3.2958	3.0920	.20381	1.17	.84	.0746	1.374	.2248	.020
9	3.3638	3.4344	-.07058	1.60	-.29	.1223	5.325	-.0943	.009
10	2.9392	2.9258	.01337	.96	.05	.0825	1.896	.0151	.000
11	1.9741	1.9092	.06489	-.31	.27	.0661	.875	.0700	.002
12	.6419	.5887	.05314	-1.97	.22	.1239	5.483	.0716	.006
13	2.5802	2.6191	-.03890	.58	-.16	.0881	2.300	-.0447	.001
14	1.2528	1.0738	.17896	-1.36	.73	.0895	2.402	.2068	.024
15	2.4248	2.2064	.21838	.06	.90	.0634	.727	.2342	.016
16	2.6603	2.8359	-.17560	.85	-.72	.0735	1.309	-.1931	.014
17	1.6864	1.6548	.03160	-.63	.13	.0686	1.016	.0343	.000
18	2.0669	2.0341	.03276	-.16	.13	.0885	2.331	.0377	.001
19	2.1633	1.9881	.17523	-.21	.72	.1087	4.000	.2186	.040
20	2.5494	2.1744	.37509	.02	1.54	.0744	1.366	.4136	.067
21	3.4177	2.9686	.44910	1.02	1.84	.1358	6.787	.6508	.552
22	.9163	1.1861	-.26985	-1.22	-1.11	.0850	2.076	-.3072	.048
23	1.7228	1.7628	-.04005	-.49	-.16	.0955	2.871	-.0473	.001
24	3.0956	3.2394	-.14383	1.35	-.59	.0909	2.510	-.1670	.016
25	3.3638	3.3658	-.00196	1.51	-.01	.1590	9.658	-.0034	.000
26	1.8245	1.8117	.01289	-.43	.05	.0535	.243	.0135	.000
Min.	.6419	.5887	-.56143	-1.97	-2.30	.0535	.243	-.6043	.000
Max.	3.4177	3.4344	.44910	1.60	1.84	.1631	10.216	.6508	.552
Moy	2.1580	2.1580	.00000	.00	.00	.0912	2.885	.0089	.041
Méd	2.0205	2.0111	.02249	-.18	.09	.0866	2.188	.0247	.014

- Si on regarde le tableau 2, qui donne les résultats, observation par observation, on note (distance de Cook) que l'observation 21 est un peu éloignée du nuage de points (c'est celle qui correspond à la crue de 1600 m3/s).

- Un résidu est un peu fort (-2.3 en résidu normé) pour l'observation 6

- un « résidu supprimé », c'est à dire un résidu d'une observation à laquelle on applique non pas le modèle calé sur l'ensemble des observations, mais le modèle calé sur toutes les observations sauf celle à laquelle on s'intéresse est un peu fort ; il s'agit encore de l'observation 21 de la crue de 1600 m3/s.

Il y aurait lieu de vérifier que ces données sont bonnes.

Sur des demi-échantillons :

Les calculs ont été repris en coupant en deux parties l'échantillon de départ (observations 1 à 13 et observations 14 à 26, cf. tableaux 3 et 4). On obtient les résultats suivants, en ce qui concerne les équations (cste et coefficients de régression) :

Tableau 3 : Obs. 1-13 Synthèse Régression de la Var. Dépendante :LVOL (pcru.sta)
R= .96594970 R²= .93305882 R² Ajusté= .91074509
F(3,9)=41.815 p<.00001 Err-Type de l'Estim.: .27083

	BETA	Err-Type de BETA	B	Err-Type de B	t(9)	niveau p
OrdOrig.			-3.080	.971496	-3.17085	.011350
LQB	.911253	.103795	.43771	.049857	8.77934	.000010
LPLU	.388912	.177102	.69441	.316219	2.19598	.055705
LIMAX	.338371	.162106	.43137	.206658	2.08734	.066469

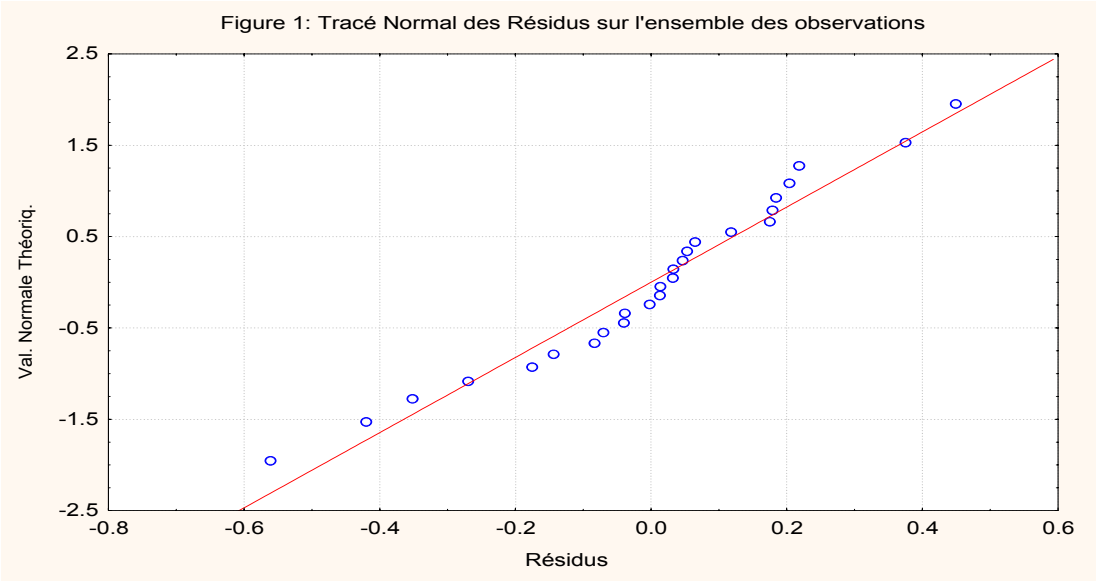
Tableau 4 : Obs. 14-26 Synthèse Régression de la Var. Dépendante :LVOL (pcru.sta)
R= .97076230 R²= .94237944 R² Ajusté= .92317258
F(3,9)=49.065 p<.00001 Err-Type de l'Estim.: .21467

	BETA	Err-Type de BETA	B	Err-Type de B	t(9)	niveau p
OrdOrig.			-1.971	.397409	-4.96037	.000780
LIMAX	.896089	.083911	.95338	.089275	10.67907	.000002
LQB	.670608	.085901	.45426	.058188	7.80679	.000027
LTRET	.299166	.087136	.25658	.074732	3.43331	.007469

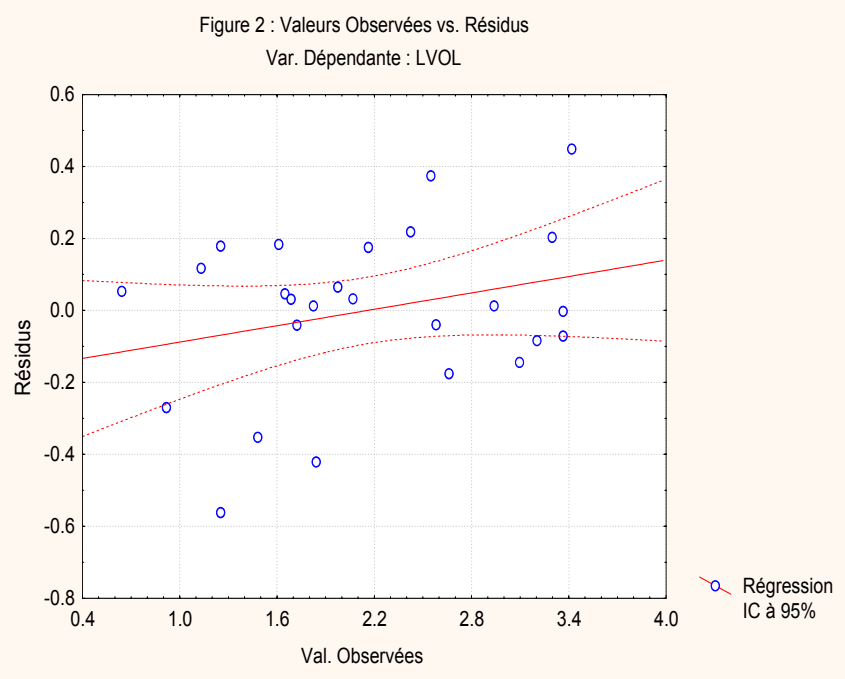
Soit en résumé :

Echantillon :	cste	LIMAX	LQB	LPLU	TRETA	R^2 débiaisé
Obs. 1-13		-3.08	.431	.438	.694	0 .911
Obs. 14-26	-1.97	.953	.454	0	.256	.923
Obs. 1-26(rappel)	-2.79	.517	.434	.589	0	.914

On observe que les modèles diffèrent. Rappelons ainsi que les coefficients de régression dépendent évidemment de la variable qu'ils pondèrent mais aussi des variables explicatives retenues. Seul le coefficient de régression de LQB, compte tenu des autres est assez stable ; en effet, cette variable est peu corrélée (elle ne l'est que pour des raisons d'échantillonnage) avec les autres variables explicatives. Quant aux résidus, on peut vérifier qu'ils sont assez bien gaussiens, (cf. figure 1). Le tracé de la figure 1 est en fait un papier de Gauss sur lequel une loi de Gauss est représentée par une droite, ce qui est presque le cas.



Il faut aussi vérifier que les résidus ne sont pas fonction de la variable à expliquer, ce qui est à peu près le cas (cf. figure 2) :



c) Modèle explicatif de la pointe :

Ce sont évidemment les mêmes variables explicatives avec le même changement de variables (Log Népérien).

Modèle global (toutes observations) :

Dans les tableaux 5 et 6 apparaissent les résultats les plus intéressants. On note :

- l'observation 21 (celle de la crue de 1600 m3/s) est toujours assez éloignée du nuage de points.
-
- Un résidu »supprimé » un peu fort (celui correspondant à la même observation 21 »

Modèles sur sous échantillons :

Comme précédemment, on a refait le calcul par sélection ascendante de variables sur les deux sous échantillons. Les résultats sont les suivants (cste, coefficients de régression et coefficient de détermination non biaisé) :

	Cste	LIMAX	LQB	LTRET	LPLU	R^2
Obs. 1-26	..947	1.093 .	3492	2943	0	.849
Obs. 1-13	0.00813 .7451	.384	0	. 533	.838	
Obs. 14-26	.3223	1.263	.374	.384	0	.876

Comme précédemment, certaines variables explicatives sont différentes ; en effet, certaines sont corrélées entre elles. On constate que le coefficient de régression de LQB est à peu près constant, du fait que c'est une variable explicative théoriquement non corrélée avec les autres variables explicatives.

Résumé des résultats :

- sur les volumes :

Si l'on veut donner l'intervalle de confiance à 80%, il faut tout d'abord examiner la fonction de répartition des résidus ; ici, on trouve que les résidus sur les Log sont à peu près gaussiens, si bien qu'il faut ajouter ou retrancher de la valeur estimée sur les Log $1.28\sigma_\varepsilon$.
Soit : valeur la plus probable en Log : $LVOL=Cste+\Sigma a_j X_j$
Valeur à 10% au non dépassement = $Cste+\Sigma a_j X_j-1.28\sigma_\varepsilon$
Valeur à 90% au non dépassement = $Cste+\Sigma a_j X_j+1.28\sigma_\varepsilon$

Si on revient en valeurs brutes : $VOL=e^{cste*}$Produit des $X_i^{a_i}$, valeur la plus probable
En posant $k= e^{cste}$ et $k'=e^{1.28\sigma_\varepsilon}$ Valeur à 10% =k'VOL et Valeur à 90% = (1/k')VOL

Numériquement :
$e^{cste}=e^{-2.79428}=0.061159$ et $e^{1.28\sigma_\varepsilon}=1.37$

Soit $VOL=0.06159QB^{0.434}IMX^{0.517}PLU^{0.588}$ (en unités définies au début)

Et valeur à 10% au non dépassement = 1.37*VOL

Valeur à 90% au non dépassement = (1/1.37)*VOL

- sur la pointe :

Même raisonnement : on trouve :
Valeur la plus probable : $POINTE=2.578IMX^{1.093}QB^{0.3492}TRET^{0.294}$ (en unités définies au début)
Et valeur à 10% au non dépassement = 1.60*POINTE
Valeur à 90% au non dépassement = (1/1.60)*POINTE

Interprétation physique des modèles :

Elle est simple pour QB, débit de base qui donne une idée de l'état de saturation du bassin ; en effet, si le débit de base avant la crue est assez fort, c'est qu'il a plu auparavant. On s'aperçoit que l'intensité maximale intervient dans les deux modèles mais que pour la pointe c'est plus la forme du hyétogramme que le total de la pluie qui intervient. Enfin, on constate que le volume est mieux expliqué que la pointe.

Amélioration possible :

On pourrait essayer par tâtonnements de travailler en retirant de la pluie et de l'intensité maximale une certaine quantité.

Résultats numériques :

		Cas 1		Cas 2
PLUI (mm)		90		160
Durée (heures)		8		13
Qbase (en m3/s)	10		.2	
IMX (en mm)		25		40
TRETA (en heures)		5		8

	Valeur à 10%	Valeur la plus probable	Valeur à 90%
VOL (hm^3) :			
Cas 1 :	9	12	17
Cas 2 :	3.2	4	5.6
POINTE (m3/s) :			
Cas 1 :	194	311	497
Cas 2 :	95	152	243

Le cas 1 est assez proche de la crue N° 13 et le cas 2 de la crue N° 1

Résultats donnés par le logiciel STATITCF à la disposition des élèves

Fichiers de données d'entrée : CRUREA (valeurs brutes) et LCRU (données en Log décimaux)

Les variables en Log sont précédées de la lettre L

3éme Partie: CRITIQUE DES DONNEES

CHAPITRE VI :

SOURCES D'ERREUR EN HYDROMETEOROLOGIE
et
TECHNIQUES ELEMENTAIRES DE DETECTION

I) - SOURCES D'ERREUR EN HYDROMETEOROLOGIE: 205

I-1) Erreurs dues au capteur 207
I-2) Changement des conditions d'environnement 207
I-3) Les erreurs liées aux conditions de la mesure 208
I-4) Traitements et transcriptions 208
I-5) Récapitulation des types d'erreur 210
I-6) Votre contribution ? 211

II) - TECHNIQUES ELEMENTAIRES DE DETECTION: 213

II-1) Analyse graphique 213
II-2) Contrôles de rupture (en monovariable sur la seule série disponible) 214
II-3) Contrôles de séquence (en monovariable sur la seule série disponible)
219
II-4) Compléments et exemples: 221

III) – CONTRÔLE PAR STATION TEMOIN : 224
méthodes des simples et doubles cumuls

III-1) La pratique des doubles cumuls 224
III-2) Aspects théoriques 226
III-3) Compléments et exemples 228
III-4) Limites et adaptation de ces méthodes 232

CONCLUSIONS 234

3ème Partie - **CHAPITRE VI :**

SOURCES D'ERREUR EN HYDROMETEOROLOGIE
et
TECHNIQUES ELEMENTAIRES DE DETECTION

I -) LES SOURCES D'ERREUR EN HYDROMETEOROLOGIE

Il serait prétentieux ici de vouloir être exhaustif, d'abord parce que les sources d'erreurs sont nombreuses et déconcertantes (les erreurs les plus triviales n'étant jamais exclues…!). D'autre part, elles sont souvent liées à la variable considérée, laquelle possède évidemment son capteur spécifique, mais aussi son propre protocole de mesure voire de transcription. Enfin, là où longtemps la transcription a été manuelle, les systèmes d'acquisition électroniques , sur site ou par télétransmission, qui sont apparus dans les années 1970, génèrent eux aussi des erreurs spécifiques.

Nous nous limiterons donc aux erreurs les plus couramment rencontrées dans les variables hydrométéorologiques (pluies, débits). Nous évoquerons aussi quelques variables parfois utilisées en complément (températures, rayonnement). Mais il est évident que les techniques présentées pourront facilement être adaptées quand on étudiera par exemple :

- des niveaux piézometriques
- des chroniques de vent (qui est alors un vecteur !), etc...

Le but de ces analyses critiques est d'abord de détecter les ***valeurs individuelles "anormales"***, puis de décider, pour ces individus isolés, si la valeur est plausible ou au contraire suspecte et risque d'être le résultat d'une erreur.

L'autre but est de décider si l'ensemble des données, souvent organisé en une série chronologique, est ***homogène au cours du temps*** et peut être traité comme tel pour le calcul de paramètres statistiques.

De même, la question peut se poser de savoir si cette série chronologique, même apparemment homogène, est ***cohérente avec d'autres séries*** de variables corrélées avec celle-ci.

En effet, si le principal souci est souvent d'avoir une série homogène, il arrive aussi que l'on cherche à détecter de vraies hétérogénéités afin d'étayer des hypothèses de changements climatiques ou de cycles .

Fig VI-1

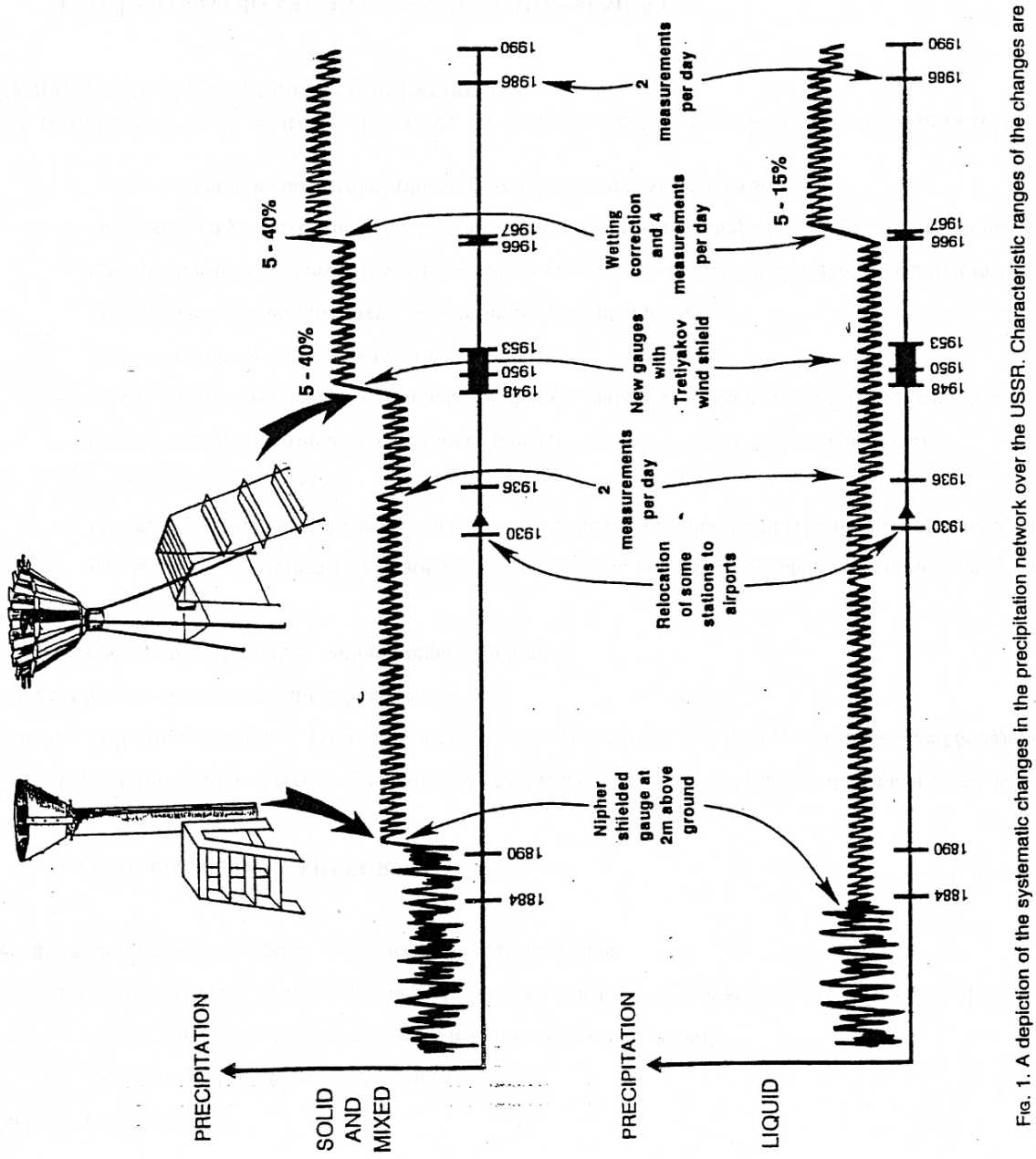


Fig. 1. A depiction of the systematic changes in the precipitation network over the USSR. Characteristic ranges of the changes are provided where possible.

On citera par exemple:

- changement dans le régime des pluies au Sahel
- cycle biennal dans les pluies ou dans les niveaux des rivières (Nil)

Là encore il faudra être très prudent et s'assurer que l'hétérogénéité détectée n'est pas seulement le résultat d'un problème de capteur ou de changement de protocole.

I-1) Erreurs dues au CAPTEUR

Un certain nombre d'erreurs peuvent être expliquées soit par la défaillance du capteur, soit par une mauvaise utilisation de celui-ci. D'où la nécessité de bien connaître le fonctionnement de l'***ensemble capteur - enregistreur*** qui sert à l'acquisition de la donnée.

Le capteur peut avoir été modifié, volontairement ou non.

Par exemple :

- un pluviomètre installé à 1 m du sol a été mis sur un toit, ou déplacé de quelques centaines de mètres.
- le capteur lui-même a été changé: passage d'un cône de réception de pluviomètre ou pluviographe de 2 000 à 1 000 ou 400 cm^2.
- une échelle limnimétrique a été démontée puis refixée, mais avec un ou deux cm d'écart.
- un capteur de rayonnement n'a jamais été changé, mais le corps noir a vieilli (et suggère, à tort, une légère diminution du rayonnement).
- un piézomètre, ou un limnimètre, s'est progressivement colmaté…
- pour les mesures de neige: il y a présence de cônes de fusion autour des perches ou nivomètre. Au contraire, on peut avoir un effet de "gâteau" sur des coussins à neige ou des lysimètres, qui mesurent ou collectent alors une surface plus large que leur simple surface au sol.

On montre à titre d'exemple (cf. page ci-contre), l'effet des changements d'appareils sur l'homogénéité d'une chronique pluviométrique.

I-2) Changement des CONDITIONS d'ENVIRONNEMENT

Outre des déplacements importants de capteur :

- transfert de la station météo d'Eybens à St Martin d'Hères (environ 6 Km)
- transfert d'une station basse à un endroit plus élevé (quelques 100 m en altitude)

On notera aussi les changements d'*environnement* autour d'un capteur en place:

Construction d'un bâtiment à proximité du capteur

- plus insidieux..! : développement de la végétation à proximité du capteur (rideau d'arbres à proximité d'un pluviomètre, broussailles dans le lit d'une rivière sous un capteur de niveau à ultrasons), changement d'état du sol (pelouse devenant parking), etc...
- travaux de recalibrage dans le lit d'une rivière à proximité d'une station (qui elle est inchangée)
- développement urbain autour d'une station météo (température, rayonnement), etc...

I-3) Erreurs liées à certaines conditions de la MESURE

Ce sont les plus difficiles à détecter, car elles ne se produisent *pas systématiquement*, mais dans certaines occasions, parfois aléatoires :

- dans un pluviomètre totalisateur, qui collecte la précipitation sous forme liquide ou solide, la capacité de collecte va dépendre de cette forme de précipitation
  (*ex :* 90% de la pluie, mais 50 à 80 % de la neige seulement, à cause de la sensibilité au vent).
  Or les données finales ne contiennent plus d'information sur la forme de la précipitation ou sur la présence / absence de vent
- les mesures de rayonnement supposent un appareil propre: or il peut être couvert de rosée, de pluie voire de neige (et donner quand même une "mesure").
  De même pour un anémomètre qui sera couvert de givre, mais qui tournera quand même!
- l'électronique (ou la mécanique) peut avoir une réponse variable selon la température (cas des sondes piézométriques de mesure de niveaux…), mais celle-ci n'est pas enregistrée en parallèle...

I-1) Erreurs dans les TRAITEMENTS et TRANSCRIPTIONS

Ce sont les erreurs liées aux dépouillements et aux transferts de l'information. (On ne dira jamais assez le temps que l'on perd par exemple à remettre en temps absolu les passages heure d'hiver / heure d'été!)

On citera, (parmi d'autres...!):

- le cas des ***cumuls aléatoires*** dans les séries pluviométriques :
  Faute d'avoir pu relever l'appareil,
  $\Rightarrow$ une pluie tombée sur les jours j et j + 1 est entièrement affectée à j + 1 :

Exemple :

08h ← j → 08h ← j+1 → 08h donne (à tort…!): 08h ← j → 08h ← j+1 → 08h

22 mm 31 mm **0** mm **53** mm !

A l'inverse, dans un pluviographe enregistreur mais non chauffant, une précipitation neigeuse tombée en une fois va fondre, une fois le beau temps revenu, sur les jours suivants et les faire apparaître , à tort, comme des jours pluvieux

- Cas des enregistrements limmigraphiques (niveaux) à transformer en débits par une ***courbe de tarage.***

  - Celle-ci a changé au cours du temps (modification de la section) mais on utilise toujours la vieille courbe de tarage.
  - Ou on change soudainement d'algorithme pour caler la courbe de tarage, et celle-ci en pratique se modifie "fortement" ...
  - Ou encore on a différentes courbes de tarage selon les époques mais on ne sait pas exactement quand (pour quelle crue?) il faut passer de l'une à l'autre, etc...

- On change la calibration d'un pluviographe (correction selon l'intensité mesurée, surtout dans les fortes valeurs > 40 mm/h)

De plus et surtout, il y a possibilité d'erreur à chaque *nouvelle transcription* :

- du diagramme de l'appareil au bordereau envoyé à l'administration centrale,
- du report par station au report par mois ou par année,
- du passage du document papier à l'acquisition sur support informatique, etc...

(cf. aussi des exemples en II-1)

Les *changements de protocoles* :

- passage de données *moyennes*, intégrées sur le pas de temps, à des données "*instantanées*" lues à l'heure de la mesure (ou inversement...)

Exemples:

cas des débits horaires : débit *moyen* 8h-9h ou débit *instantané* lu à 9h ?

cas du rayonnement ou du vent :
mesure pendant 1 minute à 9 h ou cumul de 8h à 9h ?

Signalons aussi qu'en cas de panne de l'appareil, il peut être souhaitable de compléter la série en "bouchant" la période manquante.

Mais ces ***données reconstituées*** doivent toujours être signalées, car c'est parfois la procédure utilisée qui crée elle-même une hétérogénéité (cf. reconstitution de données par corrélation et la perte de variance correspondante - in 2ème Partie Chap IV, parag III-6)

I-5) Récapitulation des Types d'Erreurs

Les différentes sources signalées donnent matière à des erreurs de différents types :

- erreurs **ponctuelles** : point “aberrant”, erreur de lecture ou de transcription

- erreurs **aléatoires** selon situation météo: valeur erronée en cas de vent, ou par régime de Sud Est, ou en période d’automne (à cause des feuilles) etc...

- erreur **systématique “brutale”**, (à partir d’une certaine date) :
  - additive : changement de position d’un appareil, décalage d’échelle, etc...
  - multiplicative : changement de surface d’un cône de pluviomètre, changement de calibration, de loi de tarage, etc...

- erreur s**ystématique “progressive”**: Détarage d’un appareil par vieillissement. Modification de l’environnement par croissance de la végétation ou urbanisation, etc...

<u>I-6)</u> Votre Contribution:

Nous laissons un peu de place ci-dessous pour y noter ***vos propres expériences***…!
Aujourd'hui, ce chapitre peut vous ennuyer par son côté "exhaustif" ou anecdotique. Vous préféreriez sans doute, avec raison, un beau calcul formel sur un cas particulier de la loi de Navier Stokes (Mais si, mais si...!).

Pourtant, si vous devez réaliser des études hydrologiques, ou climatiques, vous ne manquerez pas de rencontrer des sources nouvelles et inattendues d'hétérogénéités qui ne figurent pas dans ce récapitulatif...!

N'hésitez pas à nous les signaler…! Cela enrichira le bêtisier...

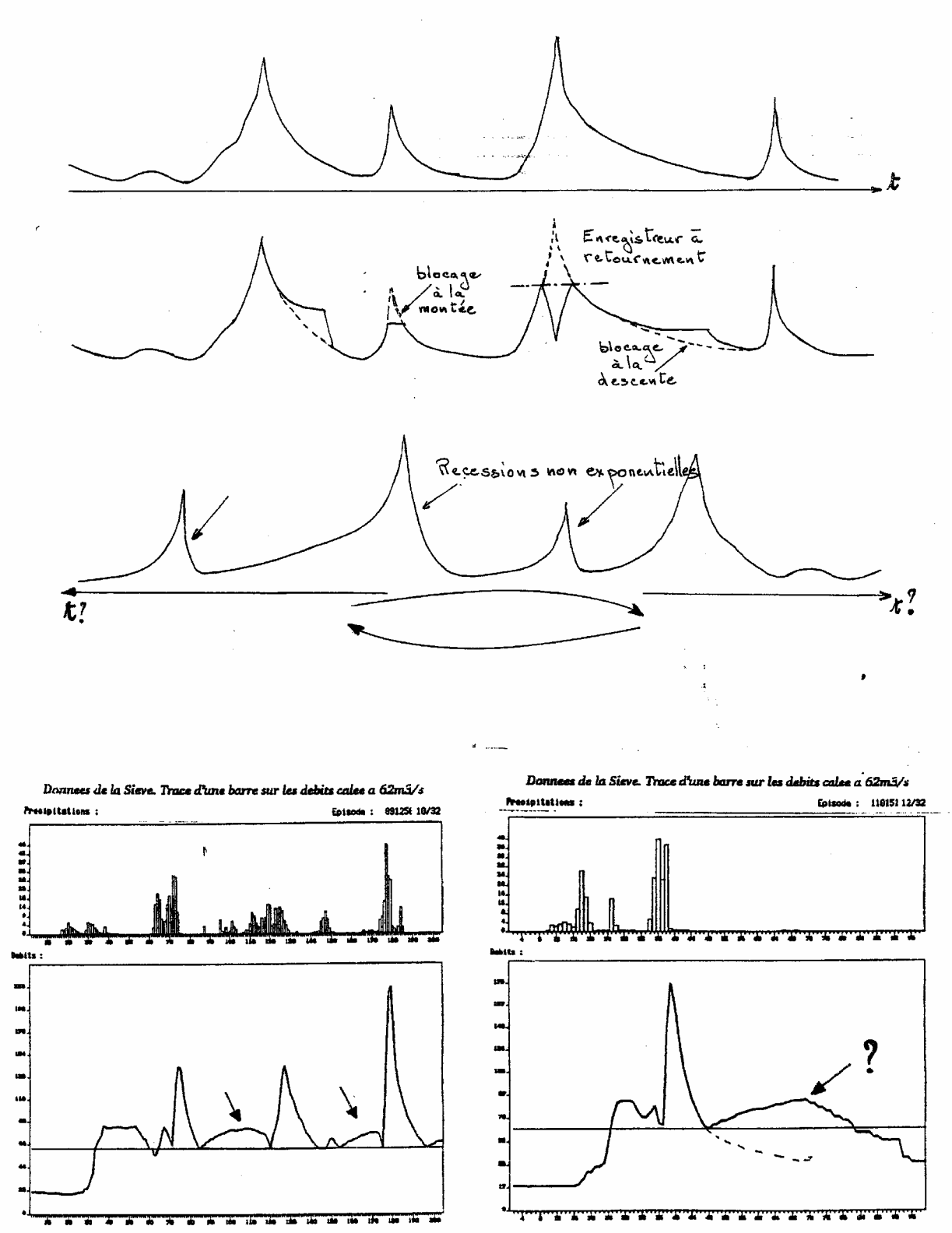

t
Enregistreur à retournement
blocage à la montée
blocage à la descente
Recessions non exponentielles
t?
t?
Donnees de la Sieve. Trace d'une barre sur les debits calee a 62m3/s
Precipitations :
Episode : 89125(10/32
Debits :
Donnees de la Sieve. Trace d'une barre sur les debits calee a 62m3/s
Precipitations :
Episode : 116151 12/32
Debits :
?

Figure VI – 2

II -) TECHNIQUES ELEMENTAIRES DE DETECTION

II-1) Analyse Graphique

Une première approche consiste à ***scruter les données***, de manière si possible automatique (par exemple: dépassement d'un seuil), pour détecter les valeurs douteuses ou aberrantes.

Exemple : durée journalière d'insolation : 33 h, à la place probablement de 3,3 h …!.

_____

Toutefois, la scrutation, à l'oeil, de tableaux de chiffres est fastidieuse, et la scrutation automatique, elle, est souvent trop grossière (par dépassement de seuil, on alerte tout le temps si le seuil choisi est faible) et surtout on ne détecte pas les ***séquences*** anormales.

Pour cela, il est préférable d'utiliser la capacité d'apprentissage de l'oeil et les connaissances qualitatives en traçant le ***graphique des données***.

Exemple: sur des données de débit (cf. figure VI – 2 ci-contre):

On constate "aisément" que des données, bien qu'appartenant à une gamme de variation raisonnable, ont une "allure" inhabituelle dans leur organisation temporelle. Sur les dépouillements de débits ci-joints, on constate:

- *courbe 1:* des débits d'allure "raisonnable" à l'œil

- *courbe 2:* des paliers inexpliqués (surtout en l'absence de pluies qui auraient pu soutenir les débits, d'où la nécessité de les mettre en regard...), ou des pointes récurrentes et de même niveau qui correspondent à une erreur de dépouillement sur des appareils dits "à retournement" (type OTT).

- *courbe 3:* une erreur de dépouillement car bien que toutes les valeurs soient correctes, il apparaît que, dans l'acquisition à la table à digitaliser, l'axe des temps a été inversé..! Mais un oeil exercé ne peut accepter que les décrues aient une forme aussi inhabituelle.

- *graphes 4 et 5* : données réelles de débit de la Sieve (affluent de l'Arno). On a tracé une horizontale pour le niveau 62 m3/s. Manifestement, il y a un retournement mécanique du stylet (appareil type OTT), qui n'est pas pris en compte au dépouillement: les récessions sont irréalistes...

C'est ainsi que sur les pluviogrammes à enregistrement sur papier, on reconnaissait aussi assez facilement l'allure liée à un appareil partiellement bouché (cf. Cours de Météo-Climato). Ce n'est plus aussi évident aujourd'hui avec les enregistrements électroniques totalisés sur des pas de temps assez conséquents (horaire par

exemple). Par contre, le nombre d'impulsions de pluies liées à des parasites (à tous les sens du terme: électromagnétiques, mais aussi mulots attaquant les câbles..!) va croissant…

Dans le cas de données spatialement réparties (précipitations, niveaux piézométriques), il est bon aussi de ***tracer une carte***, même succincte, car la présence d'un "trou" au milieu d'une zone globalement pluvieuse indiquera soit un appareil bouché, soit un décalage temporel important pour cette station. De même pour un piézomètre colmaté ou proche d'un pompage clandestin, etc... !

C'est l'esquisse du contrôle multivariable que nous verrons ensuite.

II-2) Contrôles de RUPTURE en monovariable (i.e. sur la seule série disponible)

Le cas monovariable est le cas le plus défavorable, car on ne dispose pas de "référence" à quoi se comparer, et l'information disponible, sur laquelle on va s'appuyer, est de fait suspectée d'être partiellement douteuse. On pratique d'abord des tests statistiques qui recherchent un changement "brutal et définitif" de propriétés statistiques.

a) Test des valeurs aberrantes *(isolées)*

On peut par exemple calculer la moyenne et l'écart-type de la série, et tester chaque écart à la moyenne correspondant à chaque observation.

Exemple 1 :

Soit des températures moyennes annuelles, dont la distribution peut être raisonnablement considérée a priori comme gaussienne :

θ °C	3.6	4.6	4.8	3.9	5.6	4.6	6.5	5.7	5.7	4.4	7.3
Obs:	1	2	3	4	5	6	7	8	9	10	11

Dans ce cas, on calcule la moyenne et l'écart-type empirique de l'échantillon:

$$m_x = \overline{X} = 5.15 \qquad s_x = 1.12$$

On peut alors calculer les valeurs centrées réduites :

-1.39	- .50	- .32	-1.12	.40	- .50	1.2	.49	.49	- .67	1.92

On constate alors une valeur "fortement" positive 1.92 (mais aussi une autre négative -1.39) dont la probabilité d'occurrence est certes faible (cf. votre table de la loi de Gauss standard) ....

Mais que faut-il en conclure ...?

Exemple 2 : Supposons maintenant que nous fassions un contrôle en "temps réel", à l'arrivée des données. On ne dispose pour l'instant que de 10 valeurs:

3.6	4.6	4.8	3.9	5.6	4.6	6.5	5.7	5.7	4.4
1	2	3	4	5	6	7	8	9	10

On calcule la moyenne et l'écart type empirique sur ces 10 valeurs:

$$m_x = \overline{X} = 4.90 \qquad s_x = 0.92$$

et les valeurs réduites deviennent alors :

-1.46	- .37	- .15	-1.13	.72	- .37	- .50	1.70	.83	.83	- .59

On a là encore des valeurs de faible probabilité (- 1.46, + 1.70) sur lesquelles il est délicat de conclure.

Par contre, si on transmet une nouvelle valeur de 7,3° C, celle-ci a une valeur centrée (par rapport aux moments de l'échantillon *antérieur*), de :

u = 2.57 …! probabilité associée ***0.0051*** soit une chance sur 200...!

ce qui doit immédiatement faire réagir (vérification de capteur, confirmation de la transmission, etc...).

Mais on voit bien dans cet exemple (cas n° 1) que, *une fois la valeur douteuse incluse dans l'échantillon*, il devient difficile de la détecter...

b) Cas de changements "brusques" :

On appelle ainsi un changement significatif de caractéristique de la série. On en donne des exemples ci-dessous:

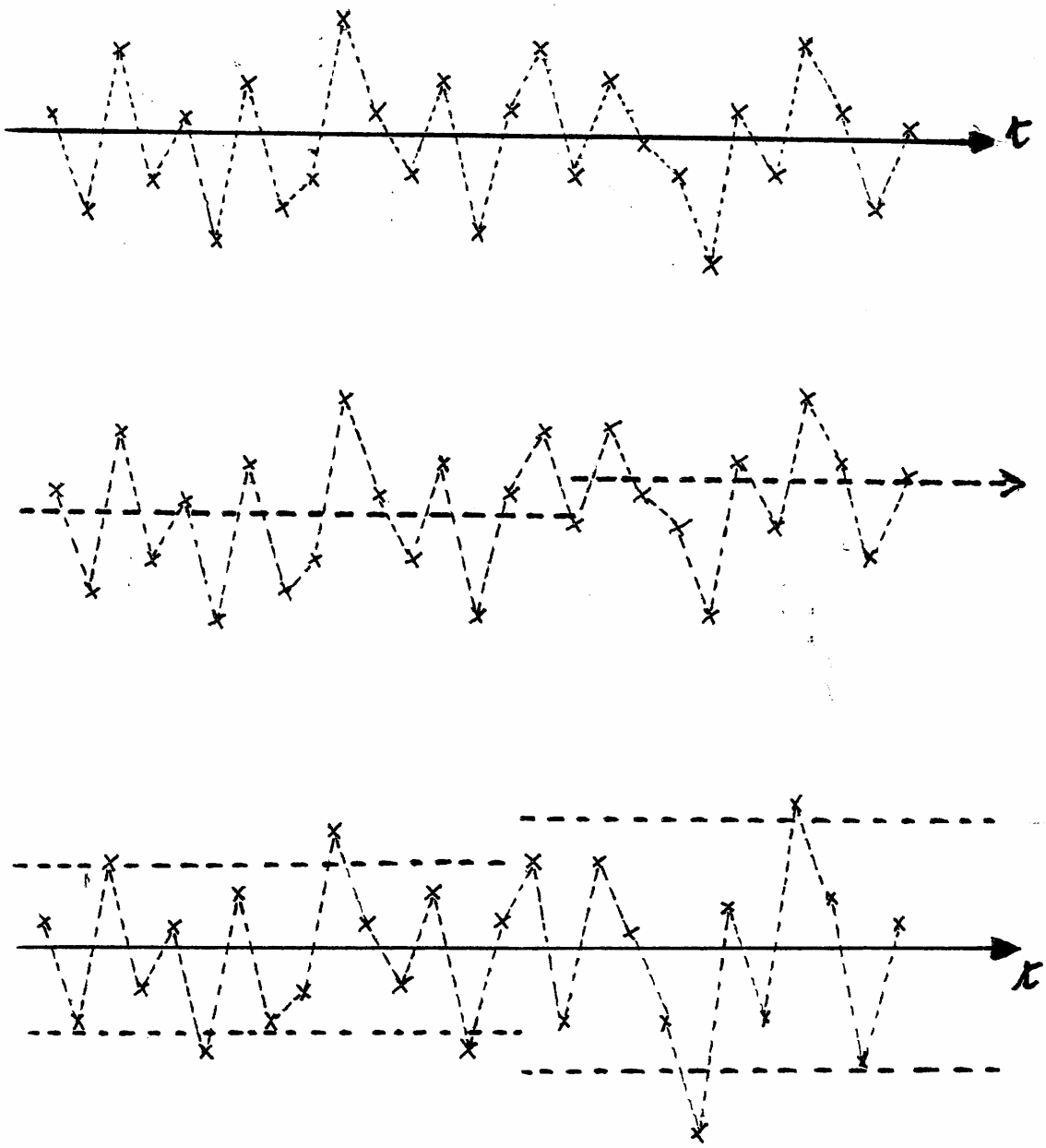


Figure VI - 3

On peut alors, *si on a l'intuition d'une date de changement*, tester si les moyennes m_1 et m_2 ont significativement changé entre les deux périodes.

Le ***Test de Student*** permet de tester si 2 échantillons sont bien issus de la même population de variance théorique σ et de même moyenne théorique μ.

Pour cela, on calcule les 2 moyennes :

$$m_1 = \frac{1}{n_1} \sum_{i=1}^{n_1} x_i \qquad m_2 = \frac{1}{n_2} \sum_{j=n_1+i}^{n_1+n_2} x_i$$

mais aussi les écarts types :

$$s_1 = \sqrt{\frac{1}{n_1 - 1} \sum_{i=1}^{n_1} (x_i - m_1)^2} \qquad s_2 = \sqrt{\frac{1}{n_2 - 1} \sum_{j=n_1+1}^{n_1+n_2} (x_j - m_2)^2}$$

et, *la variance étant supposée identique* en théorie sur les 2 échantillons, on estime la variance globale par :

$$s^2 = \frac{(n_1 - 1)\, s_1^2 + (n_2 - 1)\, s_2^2}{n_1 + n_2 - 2}$$

On sait qu'alors, la variable :

$$t = \frac{(m_1 - m_2)}{s} \sqrt{\frac{n_1 . n_2}{n_1 + n_2}}$$

suit une loi de Student à $u = n_1 + n_2 - 2$ degrés de liberté. (cf. Chap. II de la I^{ère} Partie)

Selon la valeur de t, et donc de la probabilité d'apparition d'une telle valeur, on décide s'il est plausible ou non de considérer que m_1 et m_2 soient 2 estimations de la ***même*** moyenne μ.

De même, on pourrait tester ***(Test de Fisher)*** si les 2 échantillons sont issus de populations normales ayant même variance*, leurs moyennes étant supposées identiques*.

Dans ce cas, la variable $F = \frac{s_1^2}{s_2^2}$ suit une loi de Fisher à:

$u_1 = n_1 - 1$, $u_2 = n_2 - 1$ degrés de liberté

(où encore $Z(\upsilon_1, \upsilon_2) = \frac{1}{2} \operatorname{Log} \frac{s_1^2}{s_2^2}$ suit une loi normale)

On peut multiplier ces tests (cf. Dictionnaire de Statistique. E. Morice, Dunod éditeur 1968), mais on remarquera qu'*il faut d'abord choisir la date de "rupture" présumée* (ou multiplier de manière combinatoire les essais pour essayer de la cerner... !)

c) Utilisation de connaissances physiques sur les variables.

Exemple du rayonnement solaire.

L'exemple considéré concerne des séries journalières de rayonnement solaire (énergie globale incidente en cal/cm^2 ou J/m^2 recueillie chaque jour).

Chaque jour, la valeur mesurée R_j est inférieure (cas de nébulosité partielle ou totale) ou égale à un maximum R_{jmax} qui dépend :

- de la date considérée j, pour des raisons astronomiques / géométriques
- du site de mesure et de son environnement, (masque dû à des bâtiments, des montagnes),

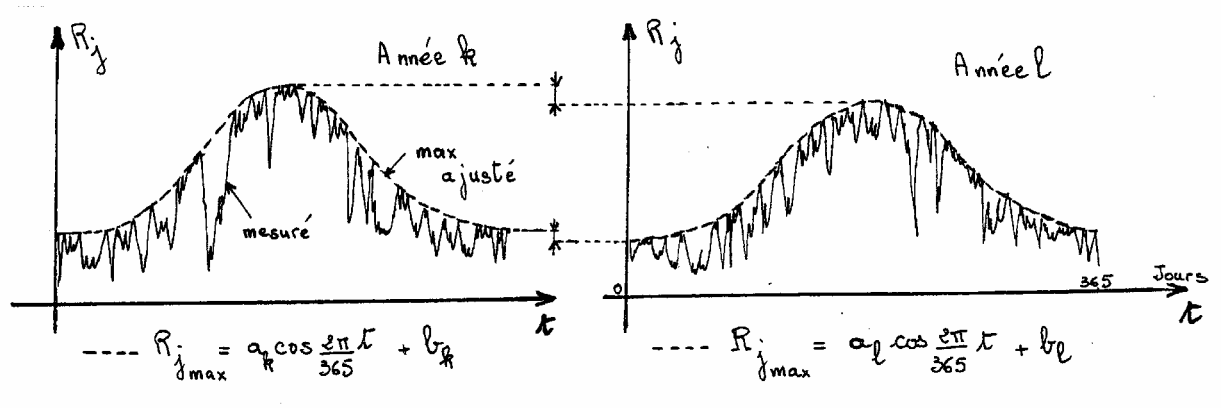
mais ce maximum devrait se retrouver *identique* d'une année à l'autre, toutes choses égales par ailleurs.

Par contre, s'il y a modification de l'environnement ou du capteur, les maximums enregistrés en porteront la trace.

Comme les données avec nébulosité partielle ne sont pas utilisables (les données de nébulosité sont peu précises et peu fiables car dépendant de l'observateur), on a considéré les seules valeurs maximales. Pour cela, on a cherché, chaque année, une courbe <u>enveloppe</u> du rayonnement maximal. Celle-ci peut être approchée par une sinusoïde qui pour l'année k aura une expression :

$$R_{jmax} = a_k \sin 2\Pi \frac{t}{365} + b_k$$

(en prenant pour t la date en jours à partir du solstice d'hiver par exemple).



<u>Figure VI-4:</u>

On peut alors tester, d'une année k à une autre *l* , si les valeurs a_k/a_l et b_k/b_l sont significativement différentes ou non.

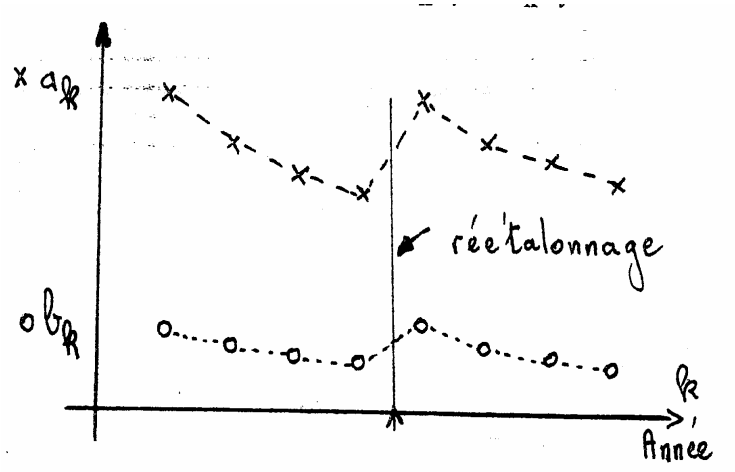


Fig VI-5

C'est ainsi que sur une série de 25 ans, (à Davos Weisfluhjoch - Suisse), on avait mis en évidence un cycle de 4 ans, qui correspondait au réétalonnage régulier des appareils...!.

II-3) Contrôles de SEQUENCE en monovariable (sur la seule série disponible)

On peut aussi chercher à tester l'organisation temporelle des données. Pour cela, on porte en abscisse le temps écoulé k, et en ordonnées le cumuls des valeurs correspondantes Y_k:

$$Y_k = \sum_{i=1}^{k} x_i$$

(Ce cumul peut se faire soit dans le sens normal (passé vers présent), soit en remontant le temps si l'on *"présume"* que les données récentes sont de meilleure qualité (condition de collecte bien connues) et que l'on préfère corriger les données anciennes…

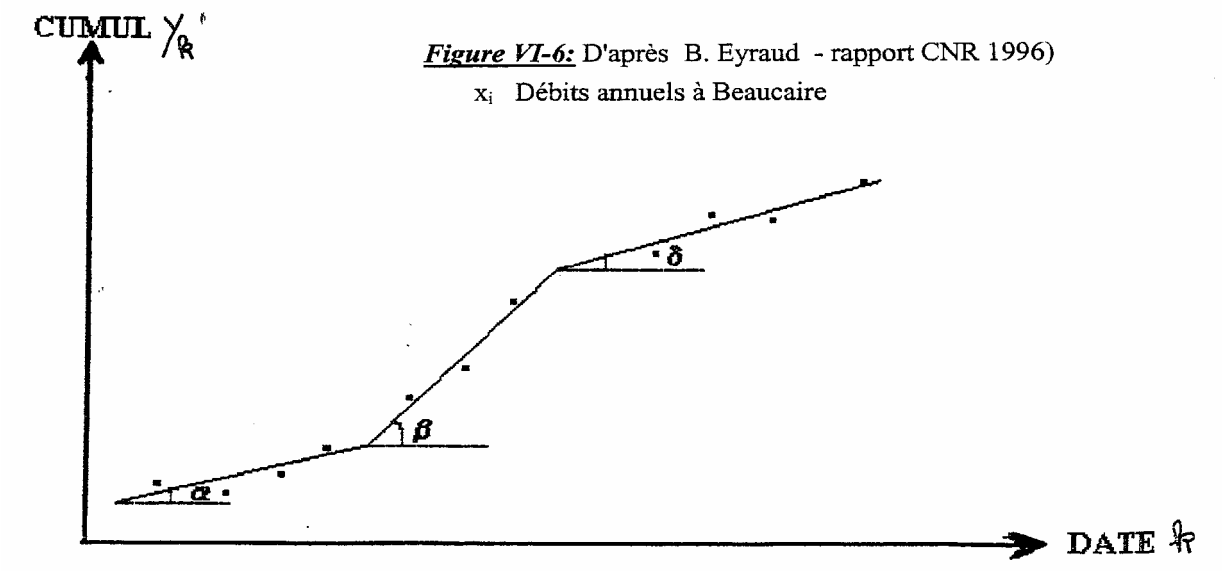


Figure VI-6: D'après B. Eyraud - rapport CNR 1996)
x_i Débits annuels à Beaucaire

L'idée de la méthode est que, *si* les mesures restent stables dans le temps, (aux fluctuations d'échantillonnage statistique près), les points de mesure devraient osciller de part et d'autre de la droite qui joint le premier point au dernier de la série.

Si par contre ils se répartissent selon différents segments de droite, on peut l'interpréter comme le signe de séquences (par exemple sèches et humides…) ou comme une dérive de l'appareil (déplacement, changement de l'appareil ou de son environnement…)

Dans une variante, on norme chaque donnée par la moyenne de la variable et on trace :

$$\{ \text{k}, \; Y_k = \sum_{i=1}^{k} \frac{x_i}{m_x} \; \}$$

et dans ce cas, les valeurs oscillent autour de la constante 1.

Une formalisation de cette dernière méthode a été proposée par D. Buishand, et reprise par Naden et Bayliss à l'Institut d'Hydrologie (Wallingford).

Elle considère une série $\{x_1, x_2, \ldots, x_N\}$, et elle calcule la variable intermédiaire S_k:

$$\text{pour k =1 à N} \qquad \{ \text{k}, \; S_k = \frac{1}{S_x . \sqrt{N}} \sum_{i=1}^{k} (x_i - m_x) \; \}$$

On a d'abord considéré des bandes de confiance *constantes* $\forall$ k.

On a ensuite utilisé le fait que la variable S_k est distribuée, pour l'abscisse k, selon une loi normale de moyenne $E[S_k] = 0$ et de variance $Var [S_k] = k.(N-k)/N$, ce qui donne des intervalles de confiance autour de l'axe des k constitués par des ellipses.

C'est une démarche analogue (quoique postérieure), à celle de Ph. Bois (1976), que l'on verra au chapitre VII suivant.

2224

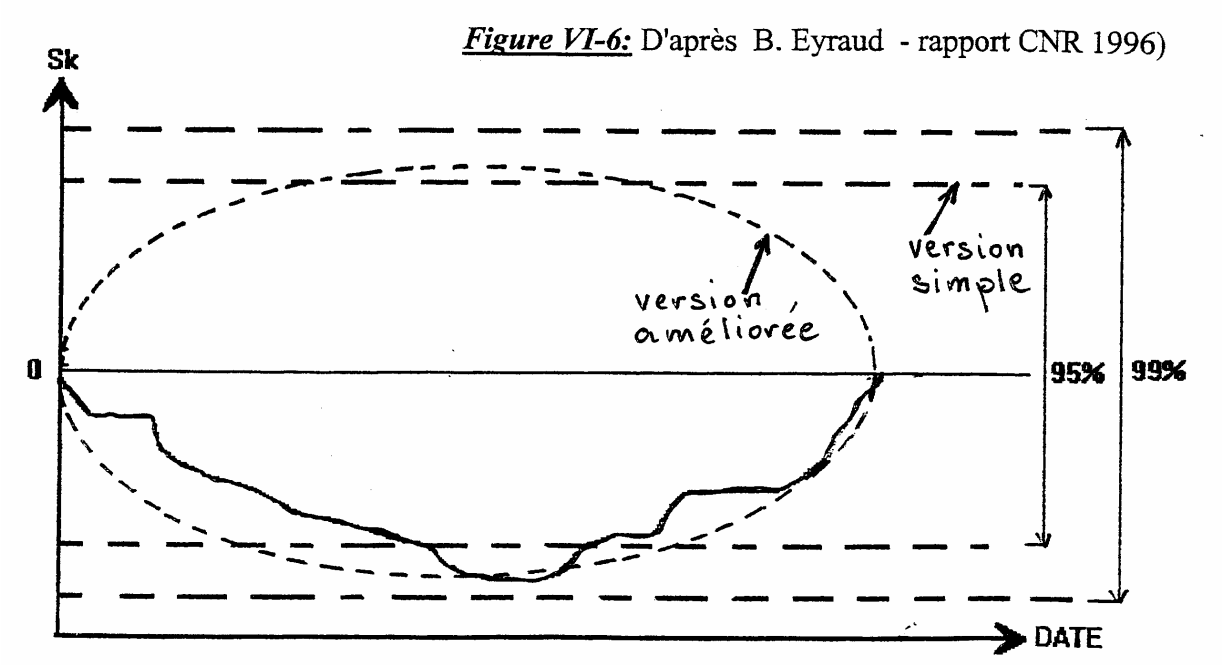


Figure VI-6: D'après B. Eyraud - rapport CNR 1996)

On peut aussi faire des tests de signes sur les écarts à la moyenne: on calcule la moyenne m_x et la séquence { k, x_k- m_x } et on analyse les séquences de + et de -, de manière à détecter des pseudo cycles, ou une rupture dans la série, (mais il n'est jamais sûr qu'elle soit unique...)

On trouvera un certain nombre de ces test décrits en détail dans un N° Spécial de la revue du CERESTA (1986).

II-4) Compléments et exemples:

a) Problème du choix de la probabilité limite de rejet :

On conçoit que si la probabilité calculée dans ces tests est très faible, c'est que l'hypothèse de même provenance est peu probable. Mais à partir de quelle valeur faut il rejeter cette hypothèse...?

Si on choisit une probabilité très faible comme seuil de rejet, on risque d'accepter assez souvent l'hypothèse de même provenance, alors qu'elle est fausse; si au contraire, on est très sévère en n'acceptant l'hypothèse que si la probabilité est forte, on risque de rejeter souvent l'hypothèse alors qu'elle est vraie.

Le seuil dépend donc du problème, surtout en contrôle de qualité. En Hydrologie le seuil de 5 ou 10% est le plus souvent retenu; cela veut dire que l'on rejette dans 5 (ou 10) % des cas l'hypothèse de même provenance au sens des moyennes et écart types, alors qu'elle est vraie.

De plus rappelons que ce seuil est un seuil d'alerte, qui a pour but **d'*attirer l'attention*** *et de* ***déclencher une enquête plus approfondie***.

Exemple d'application :

On possède une longue série de débits de la ***Loire à Blois*** (depuis 1863); ces débits ont été en partie reconstitués à partir des niveaux observés régulièrement et de courbes de tarage *estimées.*
En coupant la série chronologique en différents sous échantillons, on obtient les résultats suivants :

Période :	**Moyenne** *(en m3/s)*	**Ecart type** *(en m3/s)*
1863-1887	356	97
1888-1912	360	106
1913-1937	397	94
1938-1962	315	111
1863-1937	371	98

Comparons par exemple la période 1863-1937 à la période 1938-1962.

On a vérifié que les données sont assez bien représentées par une loi Normale.

$m_1 = 371$ $s_1 = 98$ m3/s $n_1 = 75$

$m_2 = 315$ $s_2 = 111$ m3/s $n_2 = 25$

Mais ces écarts sont-ils significatifs ?.

Comparaison entre les moyennes :

$$t = \frac{(m_1 - m_2)}{s} \sqrt{\frac{n_1 . n_2}{n_1 + n_2}} \qquad \text{avec} \qquad s^2 = \frac{(n_1 - 1)\, s_1^2 + (n_2 - 1)\, s_2^2}{n_1 + n_2 - 2}$$

soit ici $t = 4.8$ avec $n = n_1 + n_2 - 2 = 98$ degrés de liberté (paramètre de la loi de Student).

Cette valeur est élevée; dans une table de la loi de Student, on peut lire que la probabilité au dépassement est très faible, de l'ordre de .000003 !

$\Rightarrow$ On ne peut donc pas accepter l'hypothèse d'homogénéité.

Pour l'hydrologue, cela peut venir des cause suivantes :

+ Biais sur les données (d'où un travail de vérification et d'enquête)

+ évolution hydrologique (d'où une étude régionale sur d'autres rivières voisines)

Comparaison des variances :

Dans ce cas, la variable $F = \frac{s_1^2}{s_2^2}$ suit une loi de Fisher à

$\upsilon_1 = n_1 - 1$, $\upsilon_2 = n_2 - 1$ degrés de liberté

(où encore $Z(\upsilon_1, \upsilon_2) = \frac{1}{2} \text{Log} \frac{s_1^2}{s_2^2}$ suit une loi normale)

Ici, les variances observées sont assez proches, la valeur de F de Fischer-Snedecor vaut:

$$F = (111/ 98)^2 = 1.28 .$$

et les 2 paramètres de la loi de Fischer-Snedecor valent :

$n_1 - 1 = 24$ et $n_2 - 1 = 74$ degrés de liberté.

Or dans une table de Fischer Snedecor (cf. 1ère Partie - Chap III), on trouve que:

la probabilité pour que $F > 1.28 = 21\ \%$.

On peut donc considérer que les deux variances ne sont pas significativement différentes, puisque si on tirait au hasard des échantillons provenant vraiment de la même population Normale, on dépasserait cette valeur dans un cas sur cinq environ.

Exercice proposé : Etude des températures moyennes annuelles à Messeix.

On a vérifié que ces températures annuelles peuvent être considérées comme gaussiennes.

Les moyennes et écart-types des données selon les périodes sont les suivantes :

Période :	Moyenne	Ecart type
1933 1949	8.81 °C	.77°C
1950 1967	8.20 °C	.67°C

Ces données sont elles homogènes ?

III-) CONTROLE PAR STATION TEMOINS :

METHODES DES *DOUBLES CUMULS*

On a vu la difficulté de critiquer des séries en l'absence d'autres sources d'information. Dans le c) du paragraphe précédent, on a utilisé une information exogène constituée par des connaissances physiques, déterministes, sur le phénomène. (Dans ce cas l'existence d'un maximum, dont on connaît la variation saisonnière, et qui devrait se retrouver inchangé d'une année à l'autre).

Dans le cas d'autres variables, comme les pluies, on ne dispose pas de telle informations physico déterministes.

Par contre, on "sait" que, "statistiquement", celles-ci ont un comportement "régional" dominant, et que 2 stations proches devraient avoir, sur le long terme, un comportement identique..

III-1) La pratique des doubles cumuls

On considère les 2 séries initiales x et y, observées sur N périodes successives (N années, N mois de Janvier, etc...).Et on construit les 2 variables cumulées:

$$X_i = \sum_{l=1}^{i} x_i \qquad \text{et} \qquad Y_i = \sum_{l=1}^{i} y_i$$

	y_1	Y_1	x_1	X_1
temps ↓	y_2	Y_2	x_2	X_2
	.	.	.	.
	y_i	Y_i	x_i	X_i
	.	.	.	.
	y_N	Y_N	x_N	X_N

d'où la nouvelle série de couples (X_i, Y_i), que l'on pointe sur un graphique.

Si on se place au point i, l'augmentation de x_{i+1} , c'est à dire l'incrément que va connaître

$X_{i+1} = X_i + x_{i+1}$ peut être :

- forte : mais alors elle le sera aussi pour y_{i+1} donc pour Y_{i+1}, si x et Y cofluctuent assez fortement
- faible : mais idem..., elle le sera aussi pour y_{i+1} donc pour Y_{i+1}, si x et Y cofluctuent assez fortement

Donc, globalement la trajectoire des points X,Y ne devrait pas se modifier sensiblement.

Si, à partir d'un certain moment, la relation entre x et y change (par exemple, y est systématiquement augmenté de 5 à chaque observation), alors la trajectoire des (X,Y) se modifie pour retrouver un autre équilibre d'où une "cassure" à ce changement de régime.

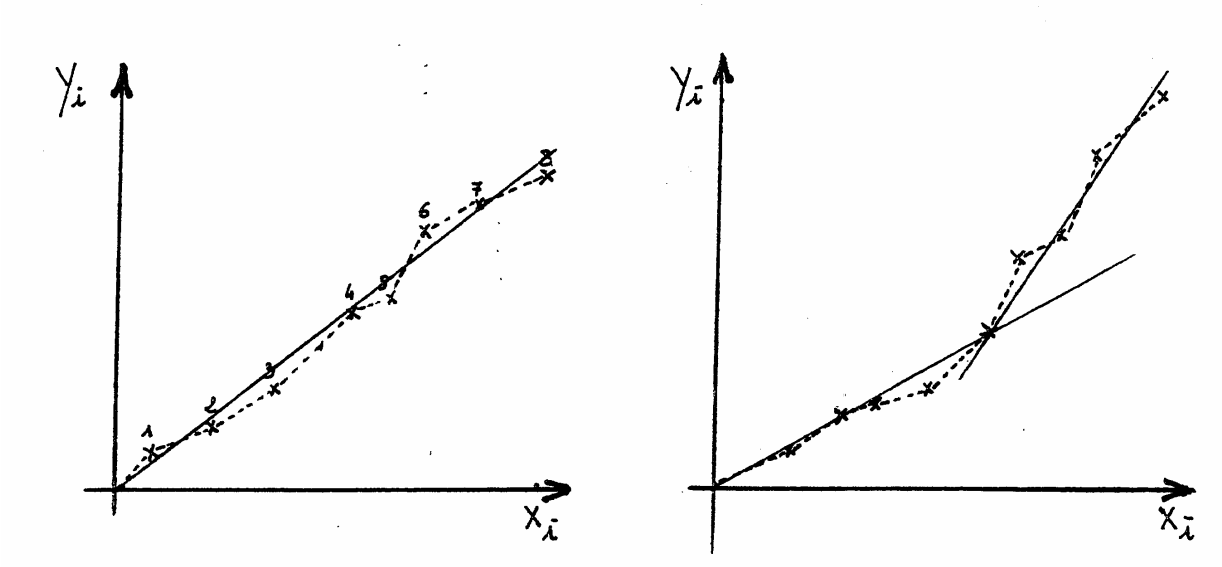


Figure VI-8:

Exemples : (de la vie courante).

- comparer l'argent de poche mensuel de 2 garnements du même âge, et détecter la date de divorce des parents, ou le retour de l'oncle d'Amérique.

-comparer les factures de téléphone de 2 familles voisines, et détecter quand l'une s'est équipée d'un minitel, ou d'un autre équipement qui incite à utiliser le téléphone...

- comparer les degrés-jour et la consommation de boisson gazeuse, et détecter le début d'une campagne de publicité (ou l'apparition d'une rumeur de pollution par le benzène) dans son effet sur la seconde variable, etc...

On comprend tout de suite que la méthode :

- n'accepte *pas de valeurs négatives* (le point courant x,y reviendrait en arrière !)
- *et* marche d'autant mieux que les *2 variables* x et y sont plus *corrélées.*

Par contre, on sent aussi que si les 2 variables ont une partie constante très forte par rapport à leur variation :

Exemple : x varie de 10 000 à 10 020
y varie de 12 000 à 12 030

la sensibilité de la méthode sera plus faible que si :

Exemple : x varie de 10 à 30
y varie de 10 à 50

En fait il ne faudrait prendre que la partie "variable" de x et y.

Ceci sera formalisé au paragraphe suivant. Pour l'instant, disons qu'il est souhaitable que les coefficients de variation:

$$C_x = \frac{s_x}{m_x} \quad \text{et} \quad C_y = \frac{s_y}{m_y} \quad \text{soient tels que} \quad C_x, C_y > 0{,}2$$

***Remarque* :**

Certains auteurs préconisent d'ailleurs de travailler sur des transformées de variables.

Exemple : $x \rightarrow 3 + \frac{x - m_x}{\sigma_x}$ (qui variera de 1 à 5 environ).

III-2) Aspects théoriques :

Si on considère que les 2 stations x et y sont constituées :

-

d'un terme qui représente la tendance régionale soit w, variable aléatoire que l'on suppose centrée réduite.

- et d'un terme aléatoire propre à la station

et que

- les corrélations de x et y avec la composante régionale w soient r_x , r_y, éventuellement différentes.

$$x \xrightarrow{r_x} w \qquad y \xrightarrow{r_y} w$$

Alors on peut écrire, pour l'observation i :

$$\frac{x_i - m_x}{s_x} = r_x\, w_i + \upsilon_i \cdot \sqrt{1 - r_x^2}$$

$$\frac{y_i - m_y}{s_y} = r_y\, w_i + \eta_i \cdot \sqrt{1 - r_y^2}$$

avec υ_i et η_i des variables aléatoires centrées réduites ($m_\upsilon = m_\eta = 0$; $\sigma_\upsilon = \sigma_\eta = 1$).

Dans ce cas :

$$X_l = \sum_{i=1}^{l} x_i = \sum_{i=1}^{l} \left(m_x + s_x r_x \,.\, w_i + s_x \upsilon_i \sqrt{1 - r_x^2} \right)$$

$$X_l = l \,.\, m_x + s_x \sum_{i=1}^{l} \left(r_x \,.\, w_i + \upsilon_i \sqrt{1 - r_x^2} \right)$$

et de même :

$$Y_l = l \,.\, m_y + s_y \sum_{i=1}^{l} \left(r_y \,.\, w_i + \eta_i \sqrt{1 - r_y^2} \right)$$

Si on divise les 2 variables pour trouver la pente:

$$\frac{Y_l}{X_l} = \frac{m_y \,.\, l + s_y \sum_{i=1}^{l} \left(r_y\, w_i + \eta_i \sqrt{1 - r_y^2} \right)}{m_x \,.\, l + s_x \sum_{i=1}^{l} \left(r_x\, w_i + \upsilon_i \sqrt{1 - r_x^2} \right)}$$

$$= \frac{m_y}{m_x} \cdot \frac{1 + C_y \sum_{i=1}^{l} \dfrac{r_y\, w_i + \eta_i \sqrt{1 - r_y^2}}{l}}{1 + C_x \sum_{i=1}^{l} \dfrac{r_x\, w_i + \upsilon_i \sqrt{1 - r_x^2}}{l}}$$

Si on considère de plus que les corrélations de x et y avec la composante régionale r_x et r_y sont élevés, proches de 1, (c'est une approximation, mais dans le cas contraire on ne compare pas les variables !) alors:

$$\frac{Y_l}{X_l} \cong \frac{m_y}{m_x} \cdot \frac{1 + C_y \cdot r_y \cdot \sum_{i=1}^{l} \frac{w_i}{l}}{1 + C_x \cdot r_x \cdot \sum_{i=1}^{l} \frac{w_i}{l}}$$

On notera d'abord que la pente sur une période *l* tend à être égale au rapport des moyennes.

On peut constater que le terme $\sum_{i=1}^{l} \frac{w_i}{l}$ est en espérance nul, avec des maxima de l'ordre de 1.

On constate aussi que le terme fluctuant lié à la station x est affecté d'un facteur d'amplification $\mathbf{C_X \cdot r_X}$

$\Rightarrow$ Il sera d'autant plus sensible à une modification du comportement de la station (par exemple : changement de moyenne m_X, ou changement de corrélation r_X avec la tendance régionale) que C_X est élevé.

On avait d'ailleurs vu qu'il était souhaitable que $C_X > 0.2$. Mais comme d'autre part, on ne veut pas de valeurs x_i négatives, on ne pourra guère aller au-delà de $C_X = 0.5$ à 0.6.

III-3) Compléments et exemples:

Exemple I : On montre ici un exemple d'étude (J. Lavabre Cemagref Aix en P) concernant la station pluviométrique de l'aéroport de Marignane. Bien qu'il s'agisse de la station "principale" du département, on avait quelques doutes car elle avait été déplacée à l'occasion des travaux d'extension de l'aéroport. On a donc décidé de la comparer à une "vieille" station du réseau, celle d'Aubagne (Cf. M. Pagnol – "Le château de ma mère").

On fournit (page suivante) les données annuelles (en mm) et la comparaison par doubles cumuls sur les 2 stations (Figure VI-9).

On pourra faire l'exercice, ainsi que celui du Chapitre VII sur le cumul des résidus.

Année	64	65	66	67	68	69	70	71	72	73	74	75	76	77	78	79
Aubagne	516	512	639	269	756	675	474	667	853	646	680	605	929	778	718	691
Marignane	479	453	532	212	623	563	450	561	902	738	627	535	824	644	546	604

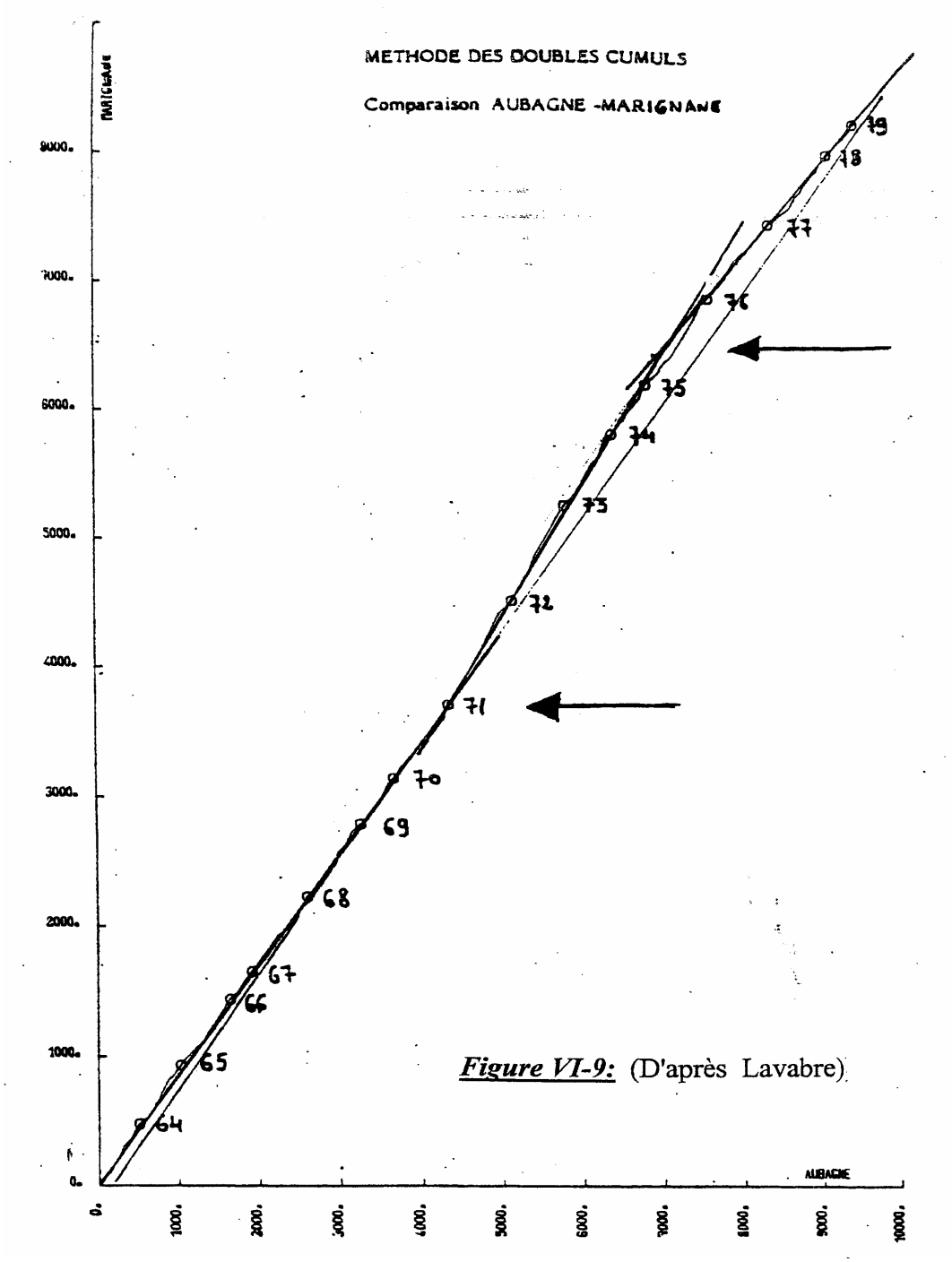


Figure VI-9: (D'après Lavabre)

Exemple II :

Cet exemple est tiré d'un rapport d'études CNR (B. Eyraud 1996) sur les débits annuels et mensuels du Rhône.

On montre d'abord l'analyse en simple cumul pour les 3 stations de Ternay, Valence et Beaucaire prises isolément.

On y constate des cassures que l'on peut attribuer soit à une rupture d'homogénéité, soit à un phénomène climatique commun aux 3 stations. Par contre on ne peut en rejeter une plutôt qu'une autre....

Mais dans ce cas , l'analyse en double cumuls permet de trancher (cf. Figure VI-11 ci-après).

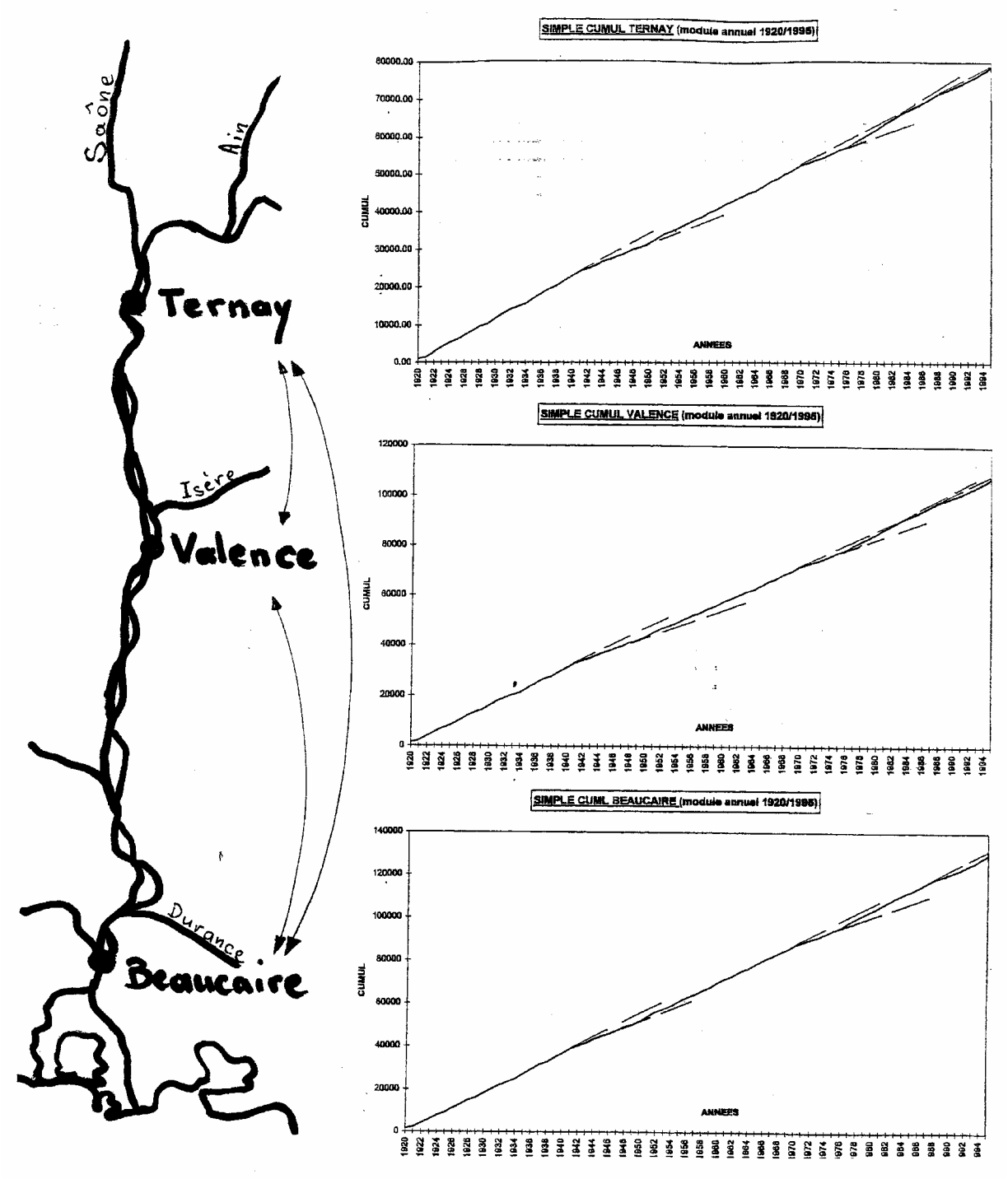


Figure VI-10: D'après B. Eyraud - rapport CNR 1996)

Figure VI-11- a)

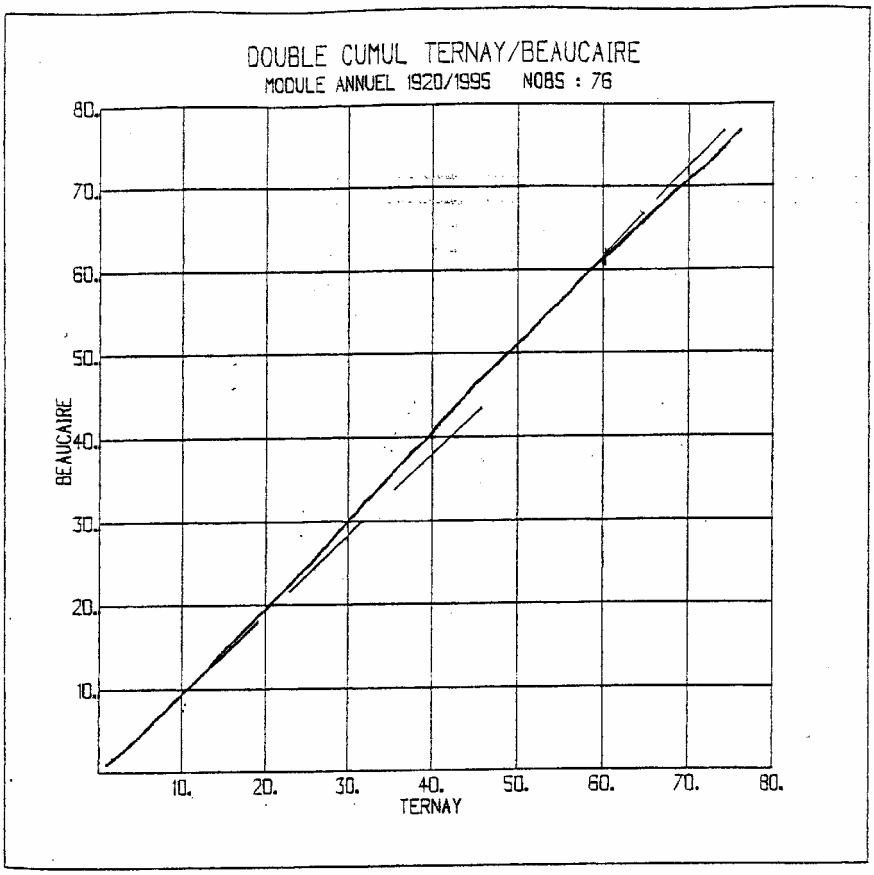


Figure VI-11- b)

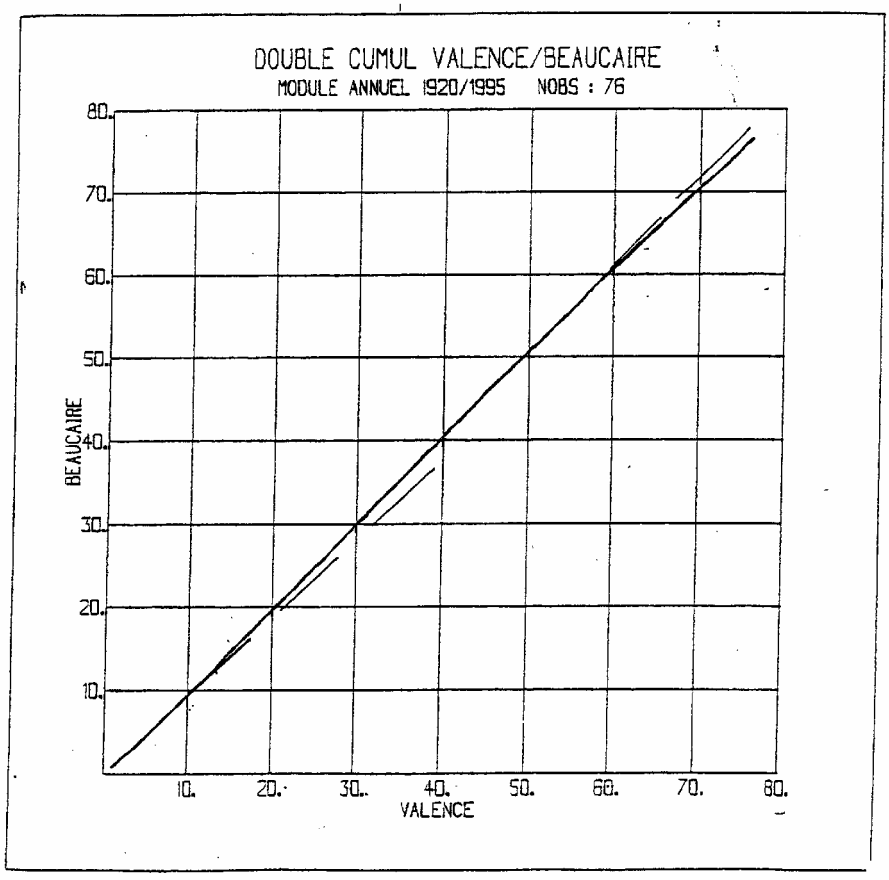
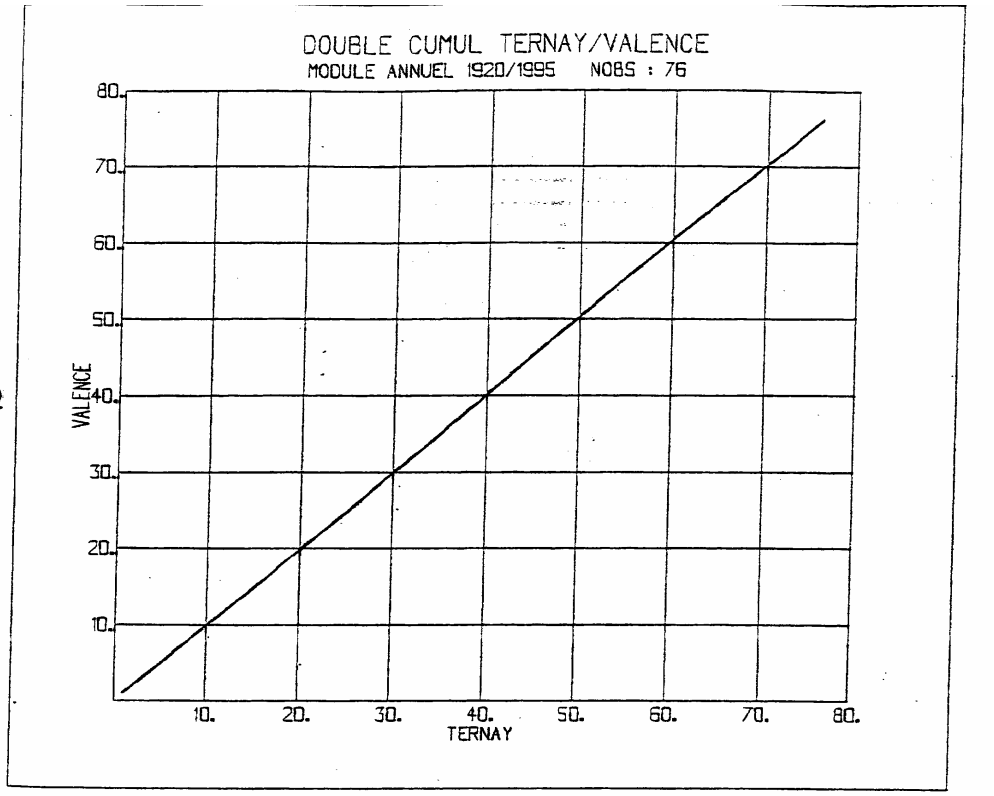


Figure VI-11- c)



En effet, si les cumuls entre Ternay et Beaucaire et entre Valence et Beaucaire présentent une cassure suspecte, la comparaison entre Ternay et Valence est très satisfaisante:

⇒ C'est donc Beaucaire seule qui est suspecte.

III-4) Limites et adaptation de ces méthodes:

Les méthodes présentées ci-dessus tablent toutes sur le caractère séquentiel des données, et le fait qu'*à partir d'une certaine date*, toutes les données suivantes ont été affectées par un changement de fonctionnement.

Il arrive pourtant que l'*hétérogénéité* ne soit pas organisée ainsi mais soit **conditionnée par une situation particulière**, qui apparaît de manière intermittente et qu'il faut identifier.

Exemple 1:

On dispose pour une station de jaugeage d'une courbe de tarage qui est extrapolée au delà du dernier débit jaugé Qm (niveau Hm) par une méthode **1**.

A partir d'une certaine date D, dans le cadre d'une rationalisation informatique, on décide que l'extrapolation sera faite par la méthode **2**, qui va donc pour la même hauteur d'eau H > Hm fournir un débit différent de la méthode 1.

Par contre, pour les niveaux inférieurs à Hm, les courbes coïncident à peu près. ⇒ On aura donc une hétérogénéité à partir de la date D **mais** pour les seuls débits supérieurs à Qm...!

On donne ici un exemple pour le ruisseau de la Vence: différents jaugeages sont disponibles jusqu'à des débits d'environ 1 m^3/s. Pour l'extrapolation, on hésite ensuite entre un polynôme du second degré (parabole) et une exponentielle. Pour la ***même*** chronique de hauteurs, voilà ce que l'on obtient:

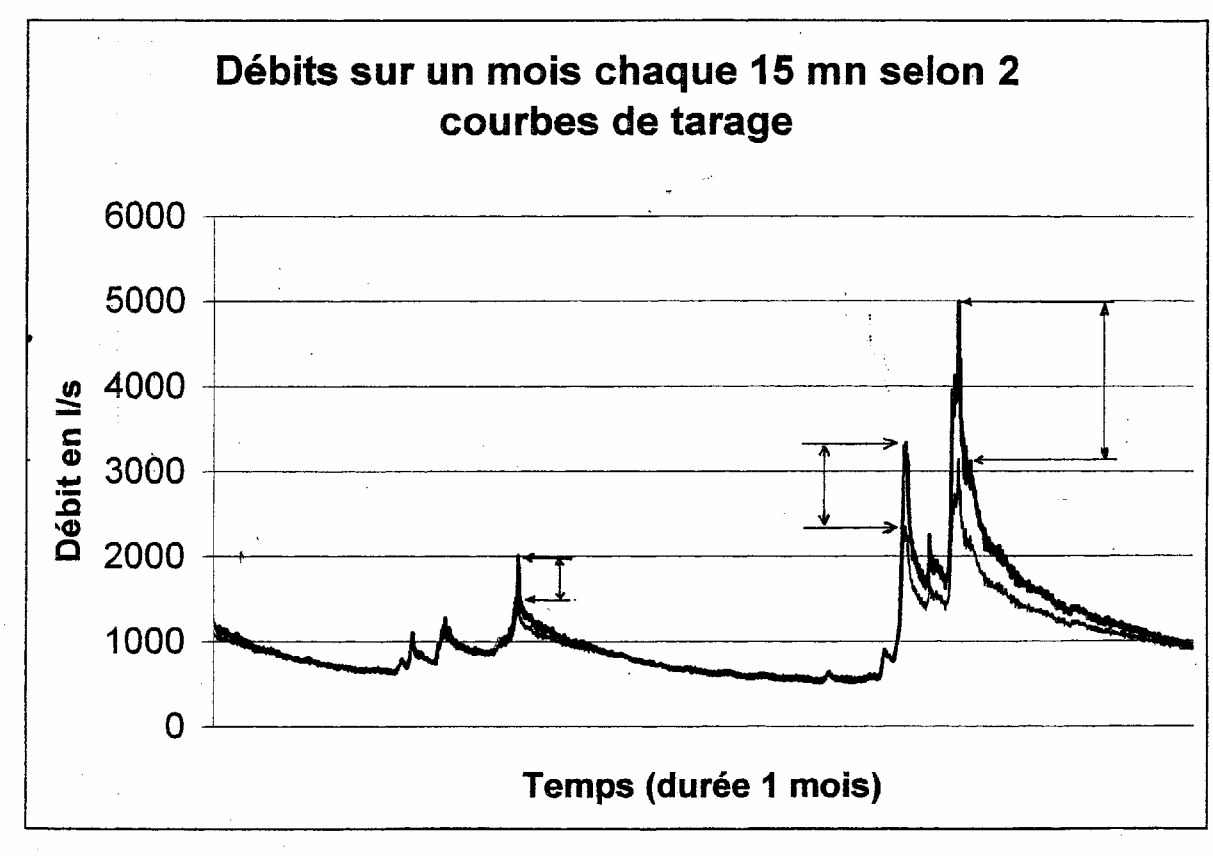


Figure VI-12

Exemple 2 :

Un cas analogue s'est produit dans une étude de **valeurs extrêmes de précipitations hivernales**. L'objectif était de dimensionner un système automatique de déneigement par chauffage électrique en tarif de nuit. Le système devait être capable de faire fondre en une nuit la précipitation maximale (hivernale) annuelle huit années sur dix. Il fallait donc disposer d'une bonne estimation de la loi du maximum annuel de précipitation journalière.

Cette précipitation était mesurée au pluviomètre ("au seau", relevé tous les matins à 8h). Comme il s'agissait souvent de neige, cette neige, accumulée dans le seau, était préalablement fondue et l'eau liquide vidée dans l'éprouvette de mesure.

La hauteur du seau étant approximativement 30 cm, la neige collectée était relativement protégée une fois dans le seau, et ce jusqu'à une hauteur de 30 cm soit environ 30 mm d'équivalent en eau.

Par contre, au delà, la neige:

- soit , par temps calme, s'accumulait en gâteau au dessus du seau et y restait (mais la surface de captation devenait incertaine...)
- soit s'accumulait en "gâteau" au dessus du seau mais pouvait être balayée par du vent survenant après la chute et avant le relevé
- soit ne pouvait, en cas de vent pendant la chute, s'accumuler dans le seau quand celui-ci était plein, faute de place "abritée" ...

On a ainsi pu constater un biais très fort des mesures au delà de 30 mm/jour, bien qu'il y ait de nombreuses mesures supérieures à cette valeur...

Mais le critère d'anomalie à prendre en considération était en fait plus complexe :

"plus de 30 mm/j **et** vent fort pendant ou après la chute de neige..."

Le dimensionnement, effectué à partir de la série hétérogène, était nettement sous-estimé et conduisait à des fréquences de défaillances bien supérieures à 8 années sur 10...!

CONCLUSIONS:

On donne ci-contre un organigramme , proposé par la CNR, des différentes pratiques à mettre en œuvre pour critiquer les données (certaines méthodes : cumul des résidus seront vues au chapitre VII suivant).

Evidemment ce protocole peut être amélioré et doit surtout être adapté selon la nature, le pas de temps, etc.. des données dont on dispose, ainsi que du *temps d'analyse* que l'on peut y consacrer.

Mais on se rappellera qu'il vaut toujours mieux en faire un peu trop avant, que découvrir trop tard qu'on aurait du y consacrer plus...

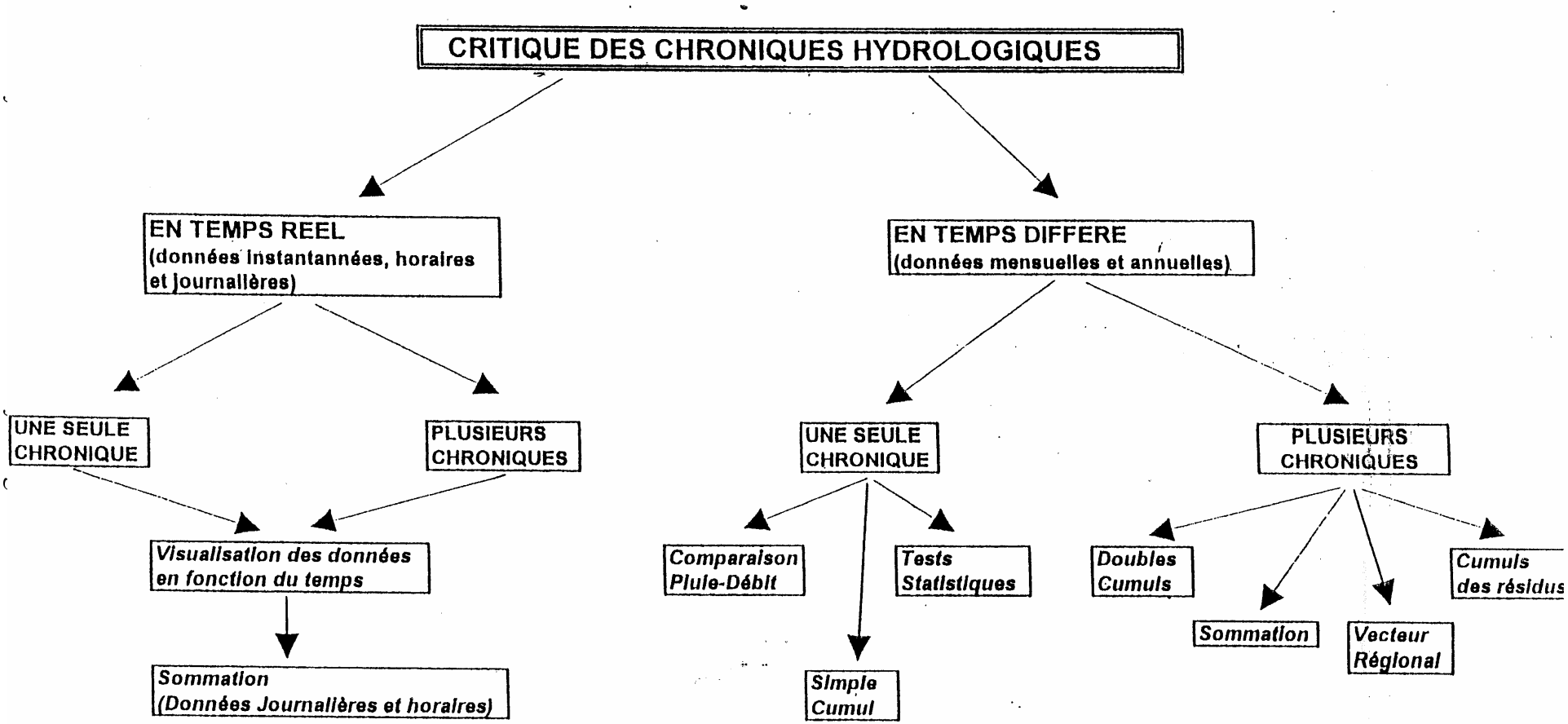


FIGURE : Critique des chroniques hydrologiques

BIBLIOGRAPHIE

CERESTA 1986

Aide Mémoire pratique des Techniques Statistique.

Revue de Statistique Appliquée Vol XXXIV N° spécial

C.N.R (Compagnie Nationale du Rhône) Différents rapports, dont

La Critique des données Hydrologiques,

Par B. EYRAUD, sous la direction de MM. D. JOUVE et B. ROSSE 1996

DALMEN E.R. and M.J. HALL 1990

Tests for stationnarity and relative consistency.

Manuel de présentation 60 p. + 1 disquette . Water Ressource Publications

P.O. Box 26 00 26 Highlands Ranch Co 80 126 0026 USA

LANG M. 1995

Les chroniques en hydrologie.

Thèse de l'Université J. Fourier. Grenoble (Mai 1995)

MESTRE O 2000

Méthodes statistiques pour l'homogénéisation de longues séries climatiques.

Thèse de l'Université Paul Sabatier Toulouse (Septembre 2000)

MORICE E. 1968

Dictionnaire de Statistique. Dunod éditeur

3 ème PARTIE - CRITIQUE DES DONNEES

CHAPITRE VII :

CONTRÔLE DE SERIES PAR CORRELATION ET CUMULS DES RESIDUS

I) ASPECT INTUITIF EN CORRELATION 239

II) ASPECT INTUITIF DE L'APPROCHE PAR CUMUL DES RESIDUS 243

III) PRESENTATION THEORIQUE " SIMPLE " : ELLIPSE GLOBALE 248

IV) PRESENTATION THEORIQUE COMPLETE: ELLIPSES INTERMEDIAIRES 250

V) BIBLIOGRAPHIE 257

VI) EXEMPLE EN SIMULATION D'ERREUR 259

VI) MISE EN ŒUVRE COMPLETE (PROBLEME DE LA REFERENCE*) (en cours de rédaction)*

23 ème PARTIE - CHAPITRE VII :

CONTRÔLE DE SERIES PAR CORRELATION ET CUMULS DES RESIDUS

I) Aspect intuitif en corrélation :

a) Les hypothèses restent les mêmes que dans la méthode des doubles cumuls, à savoir
- que l'on dispose de 2 informations :
  - station **Y à tester**
  - station **"témoin" X**

  qui sont raisonnablement liées, donc cofluctuent.
- et d'autre part que les données respectives

  X_i, $i = 1...N$ et Y_i, $i = 1...N$ constituent des *séries chronologiques*, c'est à dire que X_{i+1} est postérieur à X_i, et de même pour Y $\{i = 1...N\}$.

Par contre, pour Y, on soupçonne une hétérogénéité qui fait que la liaison entre Y et X pourrait avoir changé après une certaine date.

Note :
L'organisation séquentielle des X_i (resp. des Y_i) peut être "forte": par exemple il s'agit de données annuelles successives : X_{1981}, X_{82} ....X_{89}, X_{90} ....
Dans ce cas, le pas de temps est fixe $\Delta t = 1$ an et les données absolument en séquence.

Mais on pourrait aussi considérer :

X = total d'un épisode pluvieux (qui peut durer 2 ou 3, 4, 5 jours), de même Y.
avec X_1 = épisode du 3 au 5 Février 80
X_2 = épisode du 11 et 12 Mai 81, etc...

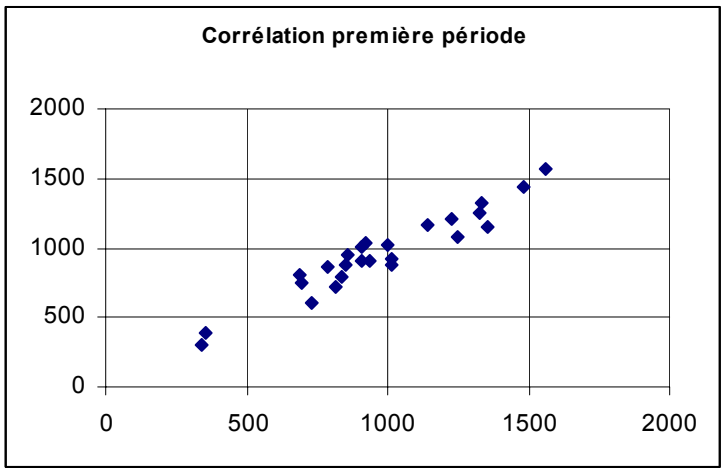
Il suffit alors que les données soient simplement *ordonnées* dans le temps :

date de X_1 < date X_2 <..... < date X_i < ...
et " " Y_1 Y_2 Y_i

pour tester l'apparition d'une hétérogénéité entre 2 épisodes, ou plutôt à partir d'un moment p dans la série 1 .... N.

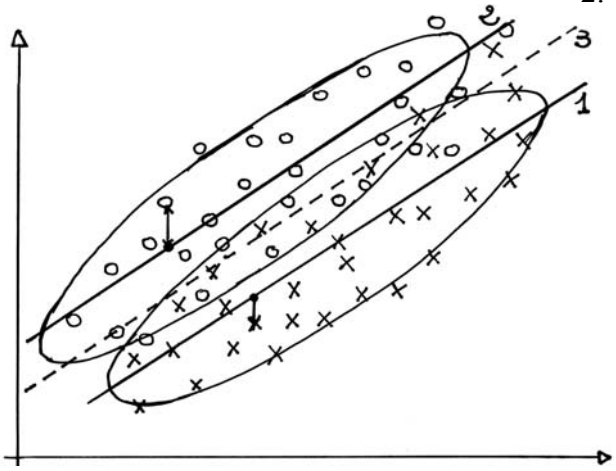
b) On peut alors imaginer que, sur la première période de 1 à p, on avait une corrélation:

$$Y = \mathbf{c_1}.\, X + \mathbf{d_1} + \boldsymbol{\nu} \qquad (1)$$

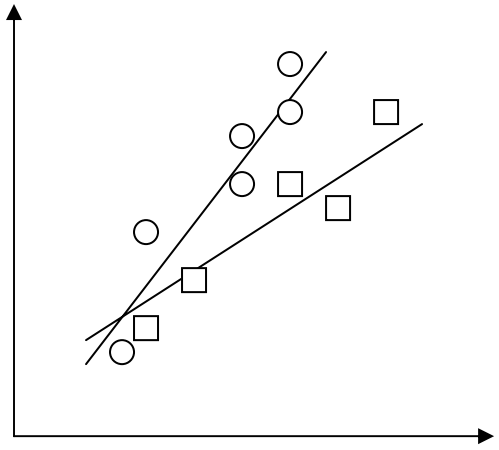


et que sur la seconde période de p + 1 à N on a une corrélation différente:

$$Y = \mathbf{c_2}.\, X + \mathbf{d_2} + \boldsymbol{\eta} \qquad (2)$$



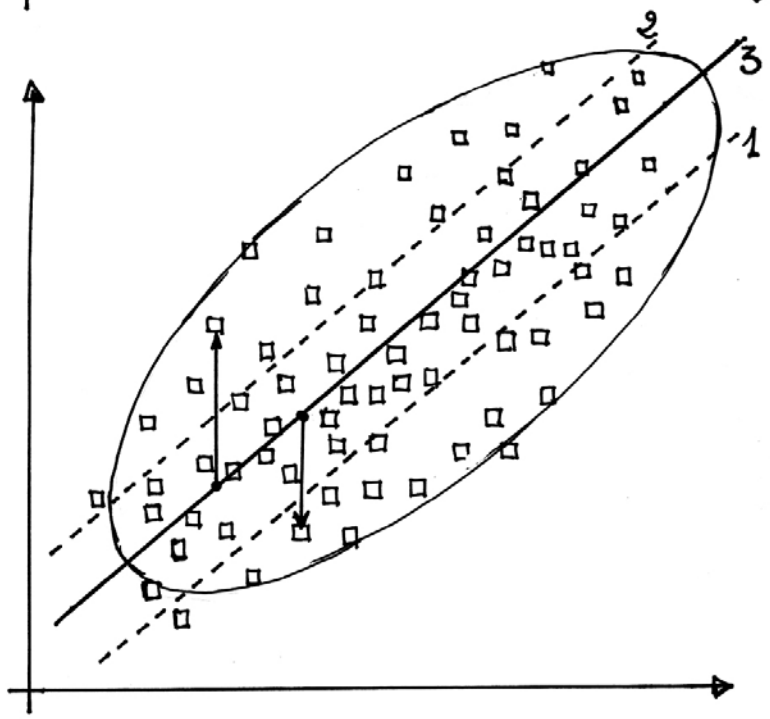
Cas A



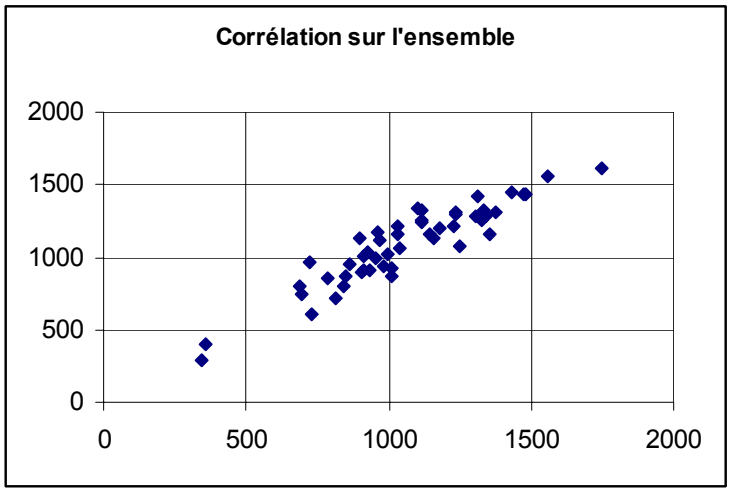
Cas B

la différence pouvant porter soit sur l'ordonnée à l'origine (A) soit sur la pente (B) , etc...

Dans le cas le plus classique, on dispose a priori, quand on considère la série *globale* {i = 1...p, p+1...= N} d'un nuage généralement moins bien corrélé (cf. figure suivante).



Si on représente le cas A par exemple (décalage d'ordonnée, pouvant correspondre à un décalage dans la série Y à partir de la date p) : on a alors un nuage plus étalé, qui donnera une corrélation inférieure (3), avec une plus grande variance des résidus e.



Toutefois, un oeil exercé (joint à un esprit perspicace !) "pourrait" constater, par exemple en codant différemment les points *antérieurs* et *postérieurs* à la date p, l'apparition de 2 nuages distincts... C'est assez peu probable d'y parvenir par hasard, mais par contre, on comprend bien que:

- pendant toute la première période 1 {i = 1...p}, les résidus tendront à être plutôt au-dessous de la droite "moyenne" globale (3) .

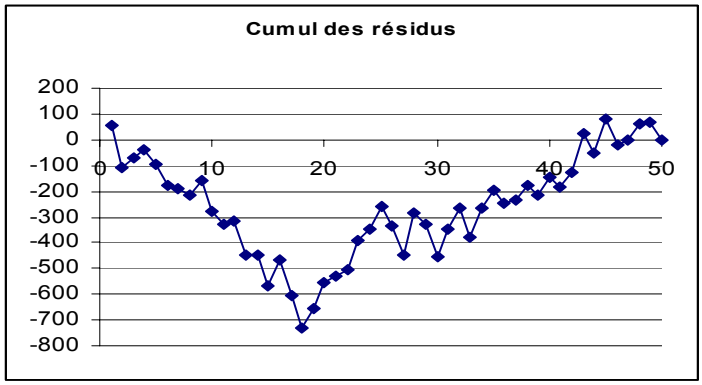
- tandis que durant la seconde période 2 {i = p+1...N}, ils seront plutôt au-dessus (sans que cela empêche quelques résidus d'être négatifs quand même, i.e en dessous de la droite (3)

Et une façon de faire apparaître cette organisation, sans connaissance a priori de la date p, est de ***cumuler ces résidus en séquence***.

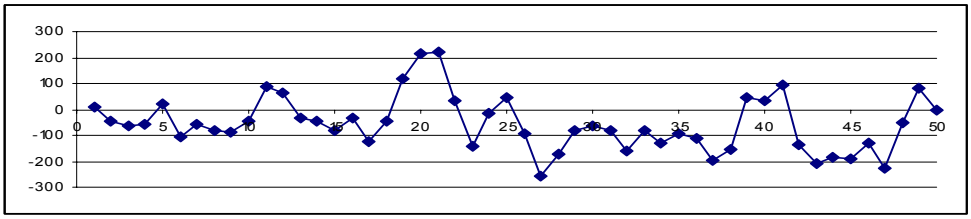
On constatera alors que dans la corrélation globale:

$$Y_i = \mathbf{a}.X_i + \mathbf{b} + \mathbf{e}_i \quad \{i=1...N\}, \quad \text{le cumul} \quad SE_i = \sum_{l=1}^{i} e_l \quad \text{est:}$$

- plutôt une somme de termes e_i négatifs dans la première période {i = 1 ...p}, et donc va en décroissant,
- alors qu'au delà de p, les e_l deviennent plutôt positifs, et leur cumul va revenir en croissant vers 0 (cf. remarque ci-dessous à propos de ce retour strict à 0), d'où une *allure particulière* :



Alors que l'on attendait plutôt une forme d'évolution moins systématique, plus aléatoire, comme ci après.



Toutefois, il est difficile d'anticiper ce que peut être une allure "normale" pour ce cumul des résidus (on pense parfois, à tort, à un bruit blanc, très chaotique), notamment du fait de la contrainte due à la corrélation sur un échantillon:

$$\sum_{i=1}^{N} e_i = SE_N = 0$$

C'est le but des paragraphes suivants que de s'en faire une idée.

II) Aspect intuitif de l'approche par cumul des résidus :

a) Si on a deux variables liées, au niveau de la *population* complète, par une corrélation *théorique*:

$$Y = \alpha . X + \beta + \varepsilon$$

on lui associe une série de résidus e, en théorie indépendants, de moyenne m_e et de variance:

$$\sigma_\varepsilon = \sigma_y \sqrt{1 - \rho^2_{XY}} .$$

En pratique, **sur un échantillon**, on *estime* , grâce à la technique des moindres carrés, une corrélation, et donc ses paramètres:

$$Y_i = \mathbf{a}.X_i + \mathbf{b} + \mathbf{e_i} \qquad \{i = 1...N\}$$

$\Rightarrow$ d'où une série de N résidus: $e_i \quad \{i = 1...N\}$

dont la moyenne m_e est **strictement** nulle : $m_e = 0$

Et c'est là qu'il y a une *"anomalie",* en ce sens qu'un résultat dont on attend qu'il ne soit vérifié qu' en **espérance,** (ou sur de grands échantillons), l'est en fait *rigoureusement*, sur chaque N-échantillon, (même si N petit).

Cette contrainte : *"Somme des résidus strictement égale à 0 "*

fait que les résidus ne seront pas tout à fait indépendants

(*Exemple :* si on en donne N-1, le $N^{ème}$ se déduit immédiatement !)

b) Pourtant, ignorons momentanément cette contrainte, et supposons que les résidus sont simplement *indépendants*, de moyenne nulle en espérance et de même loi, par exemple N $(0, s_e)$.

Dans ce cas, que devrait être un comportement *"normal"* pour le cumul des résidus ?

$$SE_t = \sum_{i=1}^{t} e_i$$

L'application des règles simples de calcul des probabilités nous indique que :

$$E[SE_t] = E\left[\sum e_i\right] = \sum \underbrace{E[e_i]}_{=0} = 0$$

Donc en espérance, SE_t est nul $\forall t$.

Et en variance :

$$\sigma^2_{SE_t} = E\left[(SE_t - E[SE_t])^2\right] = E\left[(SE_t)^2\right] = E\left[\left(\sum_{i=1}^{t} e_i\right)^2\right] = E\left[\left(\sum_{i=1}^{t} e_i^2 - \sum\sum e_i . e_j\right)^2\right]$$

$$\sigma^2_{SE_t} = \sum_{i=1}^{t} E[e_i^2] + \sum_{i=1}^{t}\sum_{i=1}^{t} \underbrace{E[e_i e_j]}_{=0}$$

= 0 car supposés *indépendants*

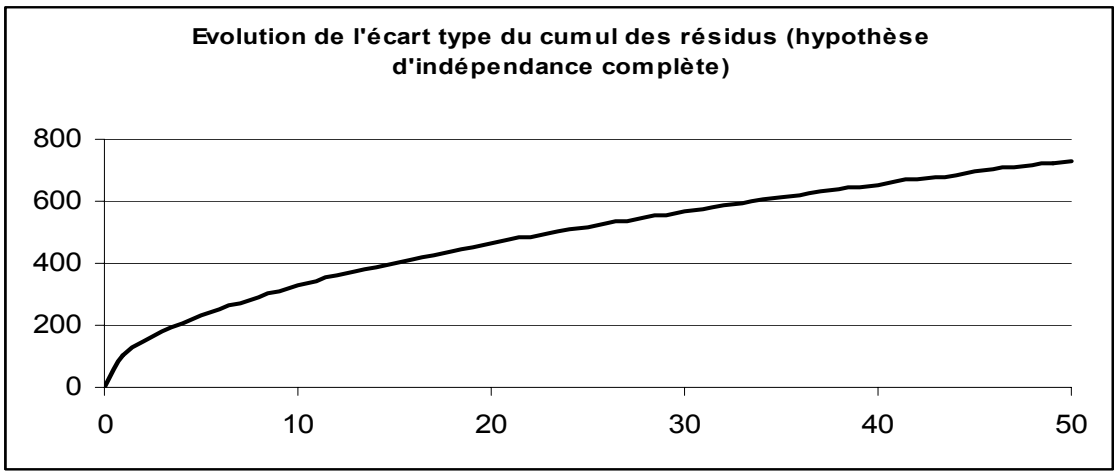
D'où, sous hypothèse d'indépendance : $\sigma^2_{SE_t} = \sum_{i=1}^{t} \underbrace{E[e_i^2]}_{\sigma_e^2} = t \cdot s_e^2$

ou encore :

l'écart type du Cumul des Résidus dépend de l'écart type du résidu sur le N-échantillon par:

$$\sigma_{SE_t} = s_e \cdot \sqrt{t}$$

ce qui signifie que, au fur et à mesure que l'on cumule les variables aléatoires e_i, la variable aléatoire SE_t voit son écart-type augmenter en $\sqrt{t}$.

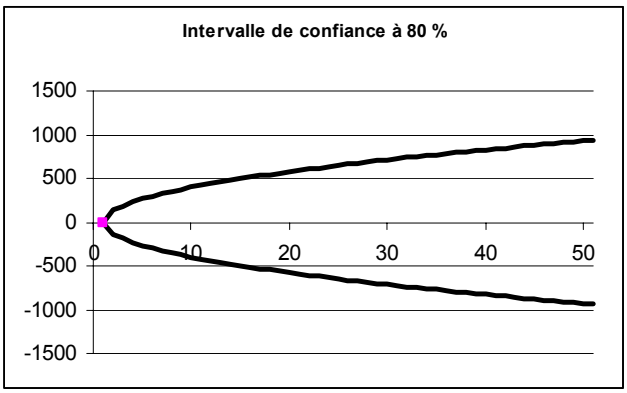


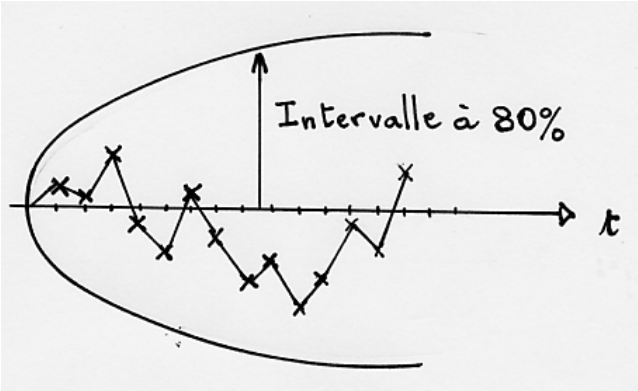
$\Rightarrow$ Autrement dit, au bout de t pas de temps, SE_t varie autour de 0 avec un écart type $s_e \cdot \sqrt{t}$.

On peut même considérer un intervalle de confiance à 80 % par exemple et dire que :

SE_t a 80 % de chance de se trouver entre $\pm 1.28\ s_e \cdot \sqrt{t}$.

C'est l'équation d'une parabole horizontale :





c) Effet de la contrainte imposée au niveau du N-échantillon : $\sum_{i=1}^{N} e_i = 0$

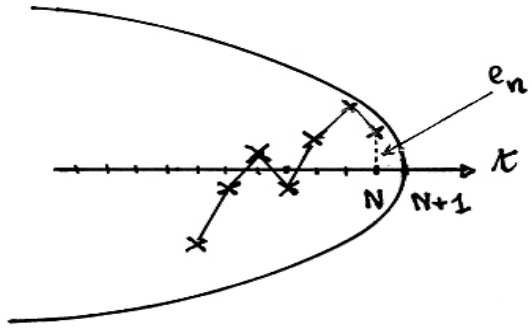
On "sent" bien que quand on commence à cumuler les résidus, la contrainte de "retour à zéro" :
$\sum_{i=1}^{N} e_i = 0$, ne se fait pas trop sentir, mais qu'elle sera de plus en plus présente au fur et à mesure que l'on se rapproche de i = N/2, indice au delà duquel on tend à revenir vers 0.
(et plus encore ensuite, puisque si on connaît SE_{N-1}, on en déduit sans aléas aucun : $e_N = SE_{N-1} \Rightarrow$ le dernier résidu n'est même pas aléatoire).
Donc, l'intervalle de confiance du cumul SE_t (par la formule $s_e.\sqrt{t}$) est probablement <u>*surestimé*</u> quand on s'éloigne de 0.

Rappelons aussi que l'on doit strictement *respecter l'ordre d'app*arition, l'organisation séquentielle des e_i, puisque ce que nous testons <u>au fond</u>, c'est la vraisemblance d'une telle séquence temporelle.
(Par exemple est-il "vraisemblable" qu'ils soient au début tous > 0 puis à la fin tous < 0 ...?).

Par contre, le sens du cumul (du début à la fin ou inversement) n'a pas vraiment d'importance (Attention, on ne peut cependant pas les brasser ou faire $SE_1 = e_5$, $SE_2 = e_5 + e_{27}$ etc...).

Mais on peut tout aussi bien considérer le cumul à rebours, à reculons, et regarder comment peut varier :

$$RE_1 = e_N \qquad RE_2 = e_N + e_{N-1} \qquad ... \qquad RE_t = \sum_{i=N}^{t} e_i$$



Et là aussi on aura un intervalle de confiance dans lequel doit varier RE_t, qui aura la forme :

- d'une parabole horizontale
- orientée vers les t négatifs.

Remarque :

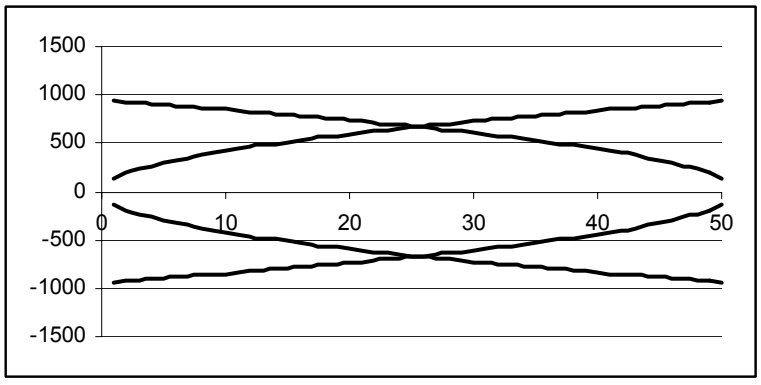
$$RE_{N+1} = 0 \rightarrow RE_1 = 0 \quad mais \quad RE_N = e_n \neq 0$$

$$SE_0 = 0 \rightarrow SE_N = 0 \quad mais \quad SE_1 = e_1 \neq 0$$

$$\Rightarrow \text{ SE} \neq \text{ symétrique de RE}$$

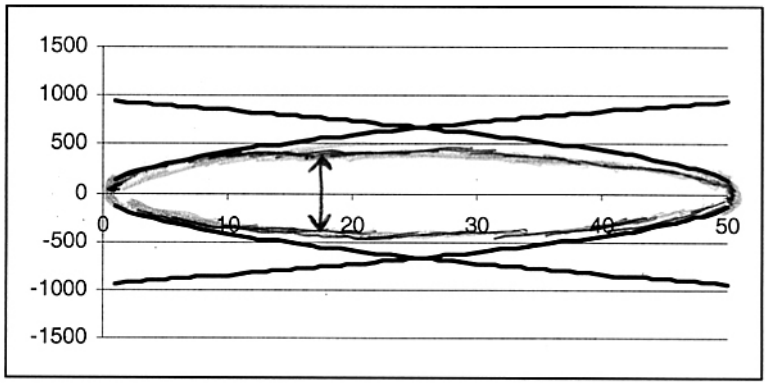
d) On peut donc *"cerner"*, définir une première enveloppe pour le (ou "les" , car RE et SE légèrement différents) cumuls des résidus:

C'est la combinaison de 2 paraboles



dont on sait que c'est une approximation acceptable au voisinage des extrémités, mais que la contrainte $\sum_{i=0}^{N} e_i = 0$ est de plus en plus sensible quand on s'en éloigne, et tend à ramener vers l'axe des x.

On “ sent ” donc que l’enveloppe *probable* est plus petite que l’intersection des 2 paraboles, bien que tangente à celles-ci aux extrémités.



Une autre façon de le dire est de considérer que dans la formule (1)

$$\sigma^2_{SE\,t} = \sum \mathrm{E}\left[\mathrm{e}_\mathrm{i}^2\right] + \sum\sum \underbrace{\mathrm{E}\left[\mathrm{e}_\mathrm{i}.\mathrm{e}_\mathrm{j}\right]}_{\neq 0}$$

$\neq 0$ car dépendants

le second terme n’est en fait pas nul (la contrainte $\sum_{i=1}^{N} e_i = 0$ fait que les résidus sont liés) et même qu’il est plutôt négatif… En effet :

Si un e_i est très grand, les autres e_j devront être de signe opposé

$\Rightarrow$ pour assurer la nullité finale de $\sum_{i=1}^{N} \mathrm{e}_\mathrm{i} = 0$.

Donc $\mathrm{E}\left[\mathrm{e}_\mathrm{i}.\mathrm{e}_\mathrm{j}\right] < 0$ et, en fait, on peut montrer qu’il est approché par :

$$-\frac{\sigma^2_\varepsilon}{\mathrm{N}-1} \text{ ou ici } -\frac{\mathrm{s}_\mathrm{e}^2}{\mathrm{N}-1}$$ (*cf.* *Compléments*).

$\Rightarrow$ Donc l’enveloppe d’acceptation prend la forme d’une ***ellipse*** incluse et tangente aux 2 paraboles. On va le démontrer plus rigoureusement ci-après.

III) Présentation théorique "simple" : (Ellipse globale)

a) On va pour cela utiliser des propriétés relativement connues sur l'échantillonnage.

On sait que dans une population **infinie** de moyenne μ et d'écart type σ, si on tire un échantillon de taille k $\{x_1, x_2, ... x_k\}$, la moyenne *empirique* :

$$m_k = \frac{1}{k}\sum_{i=1}^{k} x_i$$ a pour espérance: $E[m_k] = \mu$,

mais surtout pour variance : $\text{var}[m_k] = \frac{\sigma^2}{k}$.

Si maintenant on a une population **finie** de taille N et que l'on tire, **sans remise,** un échantillon de k individus.

On calcule $m_k = \frac{1}{k}\sum_{i=1}^{k} x_i$ et on montre que $E[m_k] = \mu$, **mais** $\text{var}[m_k] = \frac{N-k}{N-1}.\frac{\sigma^2}{k}$.

Vérification : si k = N var $[m_k] = 0$ car $m_k = \mu$ puisqu'on a pris toute la population!

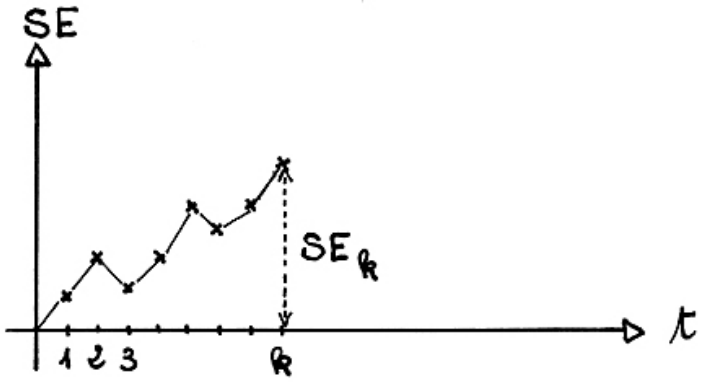
si k = 1 var $[m_k] = s^2$ c'est normal, car on a tiré 1 seul individu isolé.

b) Si on applique maintenant ces résultats à notre cas.

On a une population de N résidus e_i, de moyenne $m_N = 0$ et de variance s_e^2 connue $(s_y.\sqrt{1-r_{xy}^2})$.

Si on considère encore la variable SE_t, cumul des résidus, on a :

$$SE_k = \sum_{i=1}^{k} e_i = k.\frac{1}{k}.\sum_{i=1}^{k} e_i = k.m_k$$



Donc en espérance : $E[SE_k] = E[k.m_k] = k.E[m_k] = 0$

Et

$$\text{var}\left[SE_k\right] = \text{var}\left[k.m_k\right] = k^2\ \text{var}\left[m_k\right]$$

$$= k^2 \cdot \frac{N-k}{N-1} \cdot \frac{s_e^2}{k}$$

$$\text{var}\left[SE_k\right] = k \cdot \frac{N-k}{N-1} \cdot s_e^2$$

ou encore : $$\sigma\left[SE_k\right] = s_e \sqrt{\frac{k(N-k)}{N-1}}$$

$\Rightarrow$ C'est l'équation d'une ***ellipse*** $y = \alpha\sqrt{x.(N-x)}$ entre 0 et N.

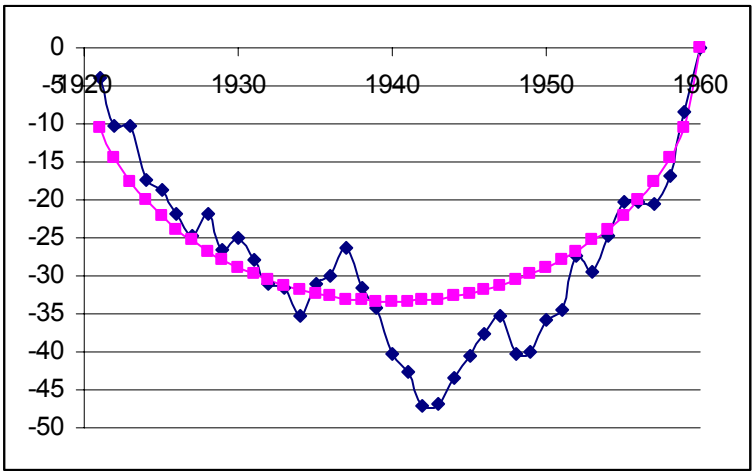
$$x = 0 \rightarrow y = 0 \qquad x = N \rightarrow y = 0$$

$$x = \frac{N}{2} \rightarrow y = \pm\ \alpha \frac{N}{2}$$

et *si on prend les échelles* telles que $\alpha = 1$ alors :

$$x = \frac{N}{2} \rightarrow y = \pm \frac{N}{2} \Rightarrow \text{c'est un cercle!}$$

exemple d'analyse de températures moyennes annuelles correctes mais avec injection d'une erreur quasi cste à partir de la mi période. Il s'agit de l'ellipse au seuil de 99% pour chaque valeur de i.



Remarque I : en nombre entier, si N est pair il y a un maximum, sinon, N impair, il y a 2 maxima.

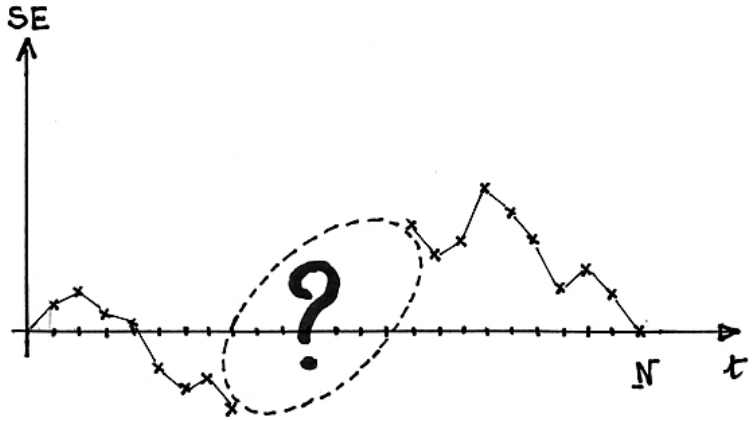
Remarque II : On peut comparer avec la parabole du § III.2 : $SE_t = s_e . \sqrt{k}$

IV) Présentation théorique complète : (Ellipses intermédiaires)(*)

a) Il s'agit là de la théorie complète, qu'il est tout à fait possible d'omettre en première lecture. On se propose de considérer l'intervalle de confiance de n'importe quel tronçon de la courbe *"cumul des résidus"*.

Par exemple, on suppose la courbe *connue* jusqu'à m $SE_m = \sum_{i=1}^{m} e_i$

et *connue* aussi de N à p (à reculons) $SEp = SE_N - \sum_{j=N}^{p+1} e_j$

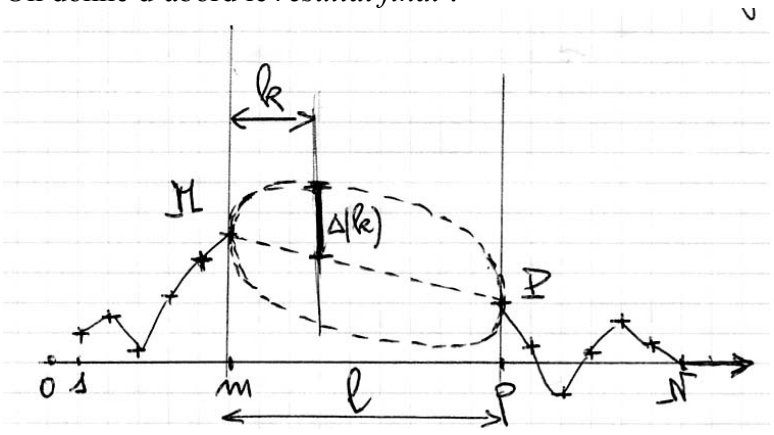


Et on se pose la question :

Quelle est l'enveloppe admissible pour les fluctuations du cumul SE entre M et P ?

(par exemple on sait déjà que les données sont correctes sur les périodes $1 \rightarrow m$ et $p \rightarrow N$).

On donne d'abord le *résultat final* :



Entre les points M et P, l'enveloppe acceptable pour le degré de probabilité f, associé à la variable normale standard α(f), est donnée par sa 1/2 largeur ? (k) :

$$\Delta(k) = \alpha(f) \cdot s_e \cdot \sqrt{\frac{N}{N-1} \frac{k(l-k)}{l}}$$

avec : k = 0,1,...*l* et α (f) = variable normale standard correspondant à la fréquence (1-f/2), ou

à l'intervalle de confiance 1-f/2 :(intervalle à 68 % α(f) = 1 , à 80 % α(f) = 1,28)

⇒ Δ (k) est l'équation d'une ellipse, dont MP est l'un des diamètres.

Remarque : On ne tracera pas systématiquement ces ellipses entre 2 points quelconques. En général, on tracera l'ellipse globale, décrite en §.III, plus quelques ellipses partielles (deux ou trois) quand on aura une configuration "bizarre" de la courbe "cumul des résidus", même si celle-ci est pourtant bien contenue dans l'enveloppe globale.

En général aussi, cela s'appuiera sur un programme conversationnel où il suffira de pointer M et P pour obtenir l'ellipse partielle.

Nous donnons maintenant une justification théorique plus complète des ellipses intermédiaires. *(Elle peut tout à fait être ignorée en première lecture).*

b)(*) On va d'abord se ramener au problème du § .III où l'on traitait le cas de l'enveloppe globale, et où l'on considérait l'ensemble des résidus.

Ici, on pose le problème de la façon suivante.

Si on connaît les point M et P, donc **si** les m premiers résidus, et, donc leur somme :

$$\sum_{i=1}^{m} \varepsilon_i$$ = segment **MM'** connu

et de même les N-p derniers, donc :

$$\sum_{i=m+1}^{p} \varepsilon_i$$ = segment **M'M"** connu

alors dans quelle enveloppe peuvent varier les résidus et surtout leur cumul entre les points M et P ? (On suppose là encore qu'il n'y a pas d'hétérogénéité et qu'ils ont tous les mêmes propriétés statistiques).

Comme dans le cas général (1 à N), leur somme (m+1 à p) est évidemment connue :

$$\sum_{k=m+1}^{p} \varepsilon_k = \underbrace{\sum_{i=1}^{N} e_i}_{=0} - \underbrace{\sum_{i=1}^{m} e_i}_{=MM'} - \underbrace{\sum_{i=p+1}^{N} e_i}_{=M'M''} = MM''$$

Donc l'espérance des résidus e_k compris entre M et P est strictement égale à :

$$E[e_k] = \frac{MM''}{l} \text{ avec } l = p - m$$

Par contre, on ne sait rien de leur *varianc*e, qui est sans doute voisine de s_e, mais > ou < , on ne sait pas… ?
Appelons-là en théorie (sur la population) s'_e que l'on pourrait estimer sur l'échantillon ici par:

$$s_e'^2 = \frac{1}{l-1} \sum_{i=m+1}^{P} (e_i - E[e_i])^2$$

mais ce n'est pas notre but d'utiliser cette estimation puisque, peut-être, elle est polluée par une anomalie.

c) On utilisera donc non pas la variance **empirique** s'_e entre i = m +1 et i = p, mais plutôt son espérance. (la moyenne si on faisait beaucoup d'essais) i.e. l'espérance de ce que peut être la variance des seuls l = p-m résidus quand :

- les m premiers et les N - p derniers sont fixés
- et que tout est *"normal",* c'est à dire ces l résidus intermédiaires sont bien issus de la même population que les m premiers et N - p derniers.

Dans ce cas, on a vu au § III-a des résultats théoriques sur ce que peut être la moyenne de ces l résidus (pris parmi une population finie de N), mais pas ce que pouvait être la variance s'.

On montre que la variance s' d'un échantillon sans remise de l individus parmi N est, en espérance:

$$E[s'^2] = \frac{N}{N-1} \cdot \frac{l}{l-1} \cdot \sigma^2$$

où σ est la variance de la population totale de N individus.

Ici, la variance des N résidus de la régression est connue et strictement égale à :

$$s_e^2 = \frac{1}{N}\sum \varepsilon_i^2 = s_y^2(1 - r_{xy}^2)$$

$\Rightarrow$ D'où pour un échantillon de *l* résidus, une variance $s_e^{'2}$ qui variera selon les échantillons, mais en espérance vaudra :

$$E\left[s_e^{'2}\right] = \frac{N}{N-1}.\frac{l-1}{l}.s_e^2$$

d) Si on revient maintenant à la population finie des *l* résidus entre M et P. Sa moyenne est fixée, connue, et sa variance est connue en espérance s_e.

Si on prend un échantillon de k (parmi *l*) au hasard, on peut calculer sa moyenne :

$$me_k = \frac{1}{k}\sum_{i=1}^{k} e_i \qquad\qquad e_i = \in \left[e_{m+1}.....e_p\right]$$

L'espérance de cette moyenne serait évidemment, comme en III-3,

$$E\left[me_k\right] = E\left[e_i\right]_{i\in[m+1,p]} = \frac{MM''}{l}$$

Mais surtout, la variance de cette moyenne serait, en espérance :

$$Var\left[me_k\right] = \frac{l-k}{l-1}.\frac{\sigma' e}{k}$$

e) Si on considère maintenant le cumul des k premiers résidus de cette sous-population. Cela nous mène en C, et on appelle C' le point correspondant sur le segment MP.

$$CC' = \sum_{i=m+1}^{j} e_i - \frac{k}{l}MM'' = \sum_{i=1}^{k} e_{m+i} - \frac{k}{l}MM''$$

En *espérance*, le point *courant* du cumul C s'éloigne du segment MP de :

$$E\left[CC'\right] = E\left[k.e_i\right] - \frac{k}{l}MM'' = kE\left[e_i\right] - \frac{k}{l}MM'' = k\frac{MM''}{l} - \frac{k}{l}MM'' = 0$$

car, d'après (d)), $E\left[me_k\right] = \frac{MM''}{l}$

Donc en *espérance*, en moyenne C parcourt MP.

f) Quant à la *variance* de cet écart CC', alors :

$$Var[CC'] = E\left[\left\{\underbrace{\sum_{i=m+1}^{m+k}(e_i - \frac{k}{l}MM'')}_{CC'} - \underbrace{0}_{E[CC']}\right\}^2\right]$$

$$Var[CC'] = E\left[\left\{\sum_{i=m+1}^{m+k} e_i - k.\frac{MM''}{l}\right\}^2\right] = E\left[\left\{\underbrace{k.(\frac{1}{k}\sum_{i=m+1}^{m+k} e_i)}_{1} - \underbrace{k.\frac{MM''}{l}}_{2}\right\}^2\right]$$

où

- le premier terme (1) est une *estimation* de me_k,
- tandis que le terme (2) est l'*espérance* de me_k (cf . parag**. b)**)

Donc :

$$\begin{aligned} var[CC'] &= E[\{k.me_k - kE[me_k]\}] \\ &= k^2\, E\left[\{me_k - E[me_k]\}^2\right] = k^2\, var[me_k] \end{aligned}$$

$$or\ \text{d'après c)} : var[me_k] = \frac{l-k}{l-1}.\frac{\sigma'_e}{k}$$

$$d'où : \qquad var[CC'] = k.\frac{l-k}{l-1}.\sigma'_e$$

g) Comme on ne connaît pas σ'_e, (- la variance de la population finie des ***l*** résidus entre M et P, ou ce qu'elle devrait être quand tout est "normal"), on remplace s'_e par son espérance dans la population finie des N résidus *observés* soit [cf. **b)**] :

$$E[s'^2_e] = \sigma'^2_e = \frac{N}{N-1}.\ \frac{l-1}{l}s_e^2$$

où s^2_e est connue (variance empirique des N résidus de la corrélation), d'où finalement :

$$\begin{aligned} var[CC'] &= k.\frac{l-k}{l-1}.\frac{l-1}{l}.\frac{N}{N-1}.s_e^2 \\ &= \frac{N}{N-1}.\frac{k(l-k)}{l}.s_e^2 \quad \Rightarrow \text{résultat donné en a)} \end{aligned}$$

Remarque : On insistera sur les conditions que l'on s'impose :

- On **ne** considère **pas** que la corrélation entre la variable à tester Y et la variable témoin X est une estimation, sur un N-échantillon, de $y = \alpha x + \beta + \varepsilon$

  En fait, donc pour laquelle il y aurait une variance d'erreur σ^2_e dont on a une estimation biaisée s^2_e.

  On travaille en fait sur le seul N-échantillon considéré, en admettant que les N résidus ont une variance *non aléatoire* s^2_e $\Rightarrow$ donc on regarde cet échantillon de N couples comme une population finie.

  - Par contre, les *sous-échantillons* de *l* individus sont eux supposés "aléatoires". Et une fois un tel échantillon de *l* résidus sélectionné, on considèrera sa moyenne comme connue mais pas sa variance. Donc on se pose la question :

    Que pourrait être la variance d'un l-échantillon,où les l individus sont pris (en séquence) parmi N, et dont la moyenne m' est connue ?

Compléments : **corrélation entre résidus.**

a) on a vu que la variance de la somme des résidus jusqu'au $k^{ième}$ était :

$$\mathrm{var}\left[\sum_{i=1}^{k} e_i\right] = E\left[\left\{\sum_{i=1}^{k} e_i - E\left[\sum_{i=1}^{k} e_i\right]\right\}^2\right] = E\left[\left\{\sum_{i=1}^{k} e_i - \sum_{i=1}^{k} \underbrace{E[e_i]}_{=0}\right\}^2\right] = E\left[\left\{\sum_{i=1}^{k} e_i\right\}^2\right] = 0$$

$$= E\left[\sum_{i=1}^{k} e_i^2 + \sum_{i\neq} \sum_{j} e_i e_j\right] = E\left[\sum_{i=1}^{k} e_i^2\right] + E\left[\sum_{i\neq} \sum_{j} e_i e_j\right]$$

$$= \sum_{i=1}^{k} E[e_i^2] + \sum_{i\neq} \sum_{j} E[e_i . e_j]$$

Donc :

$$\mathrm{var}\left[\sum_{i=1}^{k} e_i\right] = k\,\mathrm{var}[e_i] + k(k-1)\,\mathrm{cov}(e_i, e_j) \qquad (1)$$

Ceci est vrai quel que soit k, notamment pour k = N :

$$\mathrm{var}\left[\sum_{i=1}^{N} e_i\right] = N\ \mathrm{var}[e_i] + \ N(N-1)\,\mathrm{cov}(e_i, e_j)$$

Or dans ce cas, $\sum_{i=1}^{N} e_i$ est strictement nul (par construction, donc non aléatoire),

donc sa variance est nulle , et il reste:

$$\mathrm{cov}(e_i, e_j) = -\frac{\mathrm{var}[e_i]}{N-1} = -\frac{s_e^2}{N-1}$$

car ici $\Rightarrow$ Var [e_i] sur N = s^2_e, i.e. la variance des résidus calculée sur l'échantillon de calage de la corrélation.

et donc:

$$r(e_i, e_j) = \frac{\mathrm{cov}(e_i, e_j)}{\mathrm{var}[e_i]} = \frac{-1}{N-1}$$

b) On peut alors retrouver les résultats du III.3 sur l'enveloppe globale. Reprenant la formule (1), on a :

$$\mathrm{var}\left[\sum_{i=1}^{k} e_i\right] = k\ \mathrm{var}[e_i] + k(k-1)\,\mathrm{cov}(e_i, e_j)$$

mais on a vu que :

$$\mathrm{cov}(e_i, e_j) = -\frac{\mathrm{var}[e_i]}{N-1} = -\frac{s_e^2}{N-1}$$

d'où :

$$\mathrm{var}\left[\sum_{i=1}^{k} e_i\right] = k\ \mathrm{var}[e_i] - k(k-1).\ \frac{\mathrm{var}[e_i]}{N-1} = k.s_e^2.\left[1 - \frac{k-1}{N-1}\right] = k.s_e^2.\left[\frac{N-1-k+1}{N-1}\right] = k.s_e^2.\left[\frac{N-k}{N-1}\right]$$

où s^2_e est la variance connue sur les N.

Et on vérifie bien que cette variance est nulle pour k= 0 , mais aussi pour k = N.

V) Bibliographie

Bois Ph. 1986. *Contrôle des séries hydrologiques corrélées par étude du cumul des résidus*. Deuxièmes journées hydrologiques de l'ORSTOM p 89-100.

Hubert P., **1997**. *Change-points in hydrometeorological time series. Proc. Conf. Applications of time series analysis in Astronomy and Meteorology*. Chapman and Hall, Rap, Priesley and Lessi Editors, 399-412

Lang M., **1996**. *Les chroniques en hydrologie: modélisation comparée par un système de gestion de base de données relationnel et orienté objet, traitement de base et intervalles de confiance des quantiles de crues, techniques d'échantillonnage par la méthode du renouvellement. C*EMAGREF HHLY , Université Joseph Fourier Grenoble I, 1995, 296 p.

Mestre O. , 2000. *Méthodes statistiques pour l'homogénéisation de longues séries climatiques*. Thèse de Mathématiques Appliquées-Statistiques de l'Université Paul Sabatier de Toulouse. 19 septembre 2000, 226 pages.

WMO (World Meteorological Organisation), 2000. *Detecting trend and other changes in hydrological data*, WCDMP-45, WMO/TD 1013.

VI) Exemple (avec erreurs simulées)

L'exemple est tiré d'un exercice où l'on a pu contrôler l'erreur puisqu'elle a été introduite par nos soins !.

Documents :

Données de températures moyennes annuelles à Genève et au Grand Saint Bernard (station de très haute altitude bien surveillée) ; origine des données : documents suisses. Fichier complet

Description du problème :

On cherche à contrôler finement les données contenues dans le fichier Critique_Donnees_Temperature_Exo.XLS relatives aux moyennes annuelles de températures sous abri à Genève et au Grand Saint Bernard. Pour des raisons pédagogiques, on a introduit dans certains tableaux des erreurs connues.

La station de Genève est située dans un milieu urbain ; par contre la station du Grand Saint Bernard est une des stations les plus élevées d'Europe située près du col du Grand Saint Bernard à plus de 2000 m d'altitude et loin de toute agglomération.

Erreurs non ponctuelles mais petites :

On veut contrôler les données en utilisant le fait que les stations de Genève et du Grand Saint Bernard sont corrélées; pour des raisons pédagogiques, on a créé une série fictive de données à Genève appelée « Genève faux », en introduisant manuellement des erreurs. Le fichier à utiliser s'appelle Geneve_faux_1_periode.xls

C-2-1) Utilisez tout d'abord la méthode des doubles cumuls sur les séries historiques et sur la série erronée. Voyez-vous quelque chose même sur la série erronée. Peut être faudra t il faire un changement de variables pour utiliser de façon raisonnable la méthode classique et ancienne des doubles cumuls ;

C-2-2) Utilisez maintenant la méthode du cumul des résidus sur ces deux séries en dessinant les cumuls des résidus et les ellipses à 99% correspondantes. On trouvera des références à cette méthode dans :
Bois Philippe, Contrôle des séries chronologiques par étude du cumul des résidus. Colloques et séminaires ORSTOM Montpellier 16-17 septembre 1986 pages 89-99.

Correction rapide :

C-2) Erreurs plus complexes :

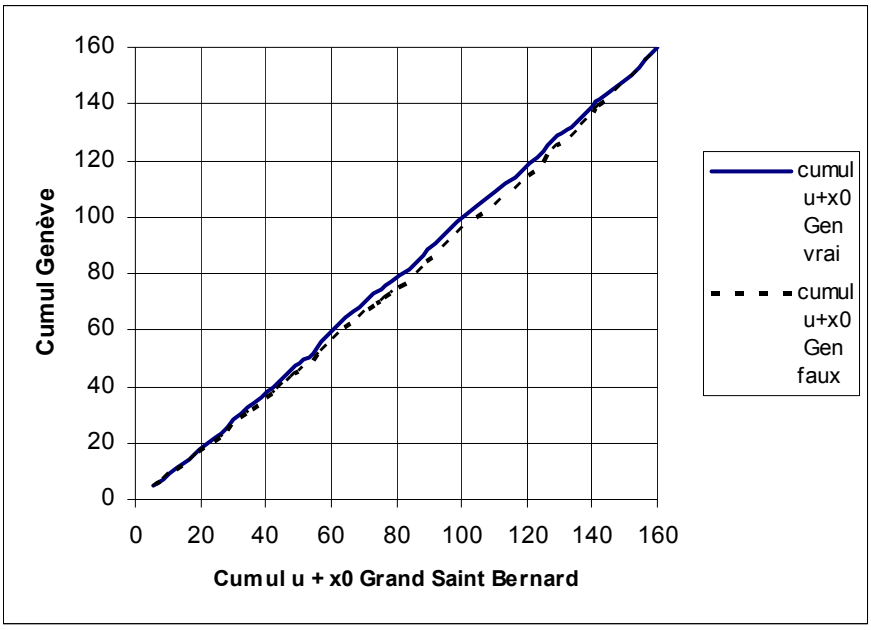
C-2-1) Méthode des doubles cumuls

Si on utilise directement la méthode classique des doubles cumuls (dubble mass en anglais), méthode surtout utilisée pour l'étude des pluies, il y a un petit problème, car les données du Grand Saint Bernard ne fluctuent pas du tout comme celles de Genève à cause de la moyenne légèrement négative. Même en utilisant la série fausse de Genève, on ne voit rien de spécial.

Aussi est il conseillé dans ce cas de faire une transformation linéaire sur les données du type :

$X_i = (x_i - \text{moyenne}(x))/(\text{écart type des x}) + \text{cste}$ avec une cste de 3 à 4 ; cette transformation a l'avantage que l'on travaille ainsi sur des variables positives de même moyenne, même écart type. On fait ensuite les cumuls sur ces variables transformées.

Réponse : dans le cas de la série fausse de Genève, il est difficile de deviner quelque chose.



C-2-2) Méthode du cumul des résidus (appelée méthode des ellipses) :

Références : Bois Philippe, Contrôle des séries chronologiques par étude du cumul des résidus. Colloques et séminaires ORSTOM Montpellier 16-17 septembre 1986 pages 89-99.

On montre que si deux variables X et Y sont gaussiennes, corrélées et stationnaires d'ordre deux (c'est à dire que les espérances mathématiques des moyennes et écart type ne dépendent pas de la date), le cumul Z des résidus ε de la régression de X en Y :

$$Z_j = \sum_{i=1}^{j} \varepsilon_i \quad avec \quad \varepsilon_i = Y_i - \widehat{Y}_i$$

On montre alors que ce cumul a une espérance mathématique nulle quel que soit j ,

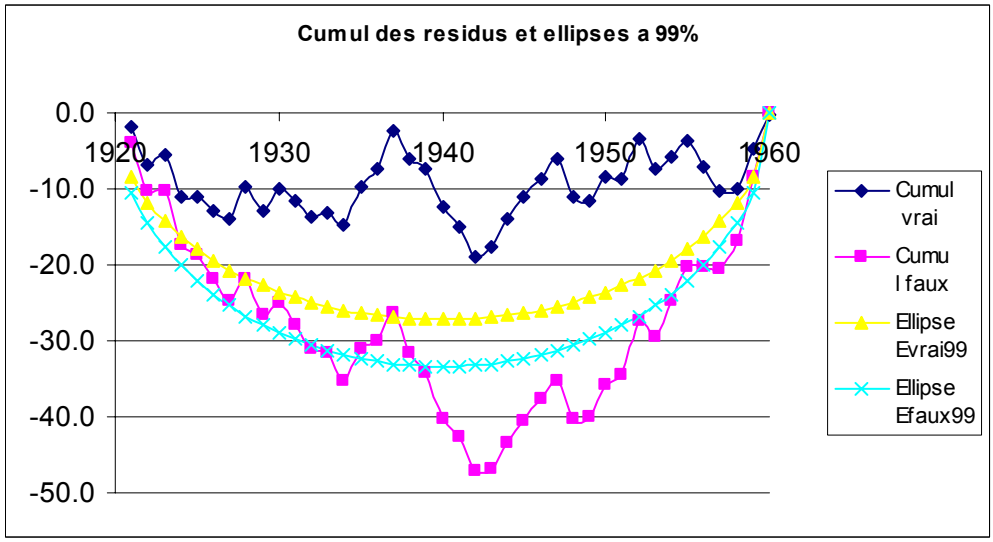
mais une variance qui dépend de j et que l'on approche par :

$$Variance(\quad Z_j \quad) \quad = \quad (1-r^2)\frac{j(n-j)}{n-1}Variance(Y)$$

Cette variance a son maximum pour j=n/2 (en valeur entière), n étant la taille de l'échantillon. L'idée est donc la suivante : on trace l'ellipse de confiance t(confiance)*Ecart type(Z_j) en fonction de j. On conseille de prendre un intervalle de confiance de 99%, ce qui correspond à t=+- 2.56. Si le tracé du cumul sort significativement de l'ellipse, il faut étudier de près la raison.

Dans le cas des données historiques, le cumul des résidus ne sort pas de l'ellipse.

Dans le cas du fichier erroné de Genève, le cumul sort nettement de l'ellipse, alors que l'erreur introduite est une addition de l'ordre de 0.3 à 0.5 °C aux données historiques de Genève à partir de la moitié de la période.



Réponse : Les données ne sont pas stationnaires.

C-2-3) Recherche du type d'anomalie :

Pour poursuivre, comme on a l'impression qu'il y a deux périodes avant et après 1942 (point max. du cumul), on peut calculer les corrélations correspondant à ces deux périodes :

Période :	r^2	A	s_A	B	s_B	Moy. Gen.	Ecart type Gen.	Moy. GSB	Ecart type GSB
1921-1941	.551	.66	.15	110.5	1.5	101	4.8	-14.4	5.4
1941-1960	.479	.52	.18	113.7	1.5	108.6	4.9	-9.9	6.5

A est le coefficient de régression de Genève expliquée par GSB, B la constante de l'équation de régression S_A et S_B les écart types d'estimation de ces valeurs (cf. formules dans certains cours ou résultats de certains logiciels).

On constate que (B(1921-1941)-B(1941-1960))/S_B=2 est beaucoup plus grand que (A(1921-1941)-A(1942-1960))/S_B=.3, c'est à dire que la différence fondamentale entre ces deux équations de régression porte essentiellement sur le terme constant de la régression.

<u>Réponse</u> : il semble qu'à partir des années 1940, il y ait eu un décalage des données. Dans la réalité, il faudrait faire une enquête ; mais ici, on retrouve bien le fait (cf. tableaux EXCEL) que les données dites fausses de Genève sont les données vraies de Genève aux quelles on a ajoute quelques dixièmes de degrés à partir de 1941.

Annexes : Quelques tables

Remarques : même sur Excel quelques tables sont accessibles par fonction

Table de Student
Table du Chi2

Probabilité:	0.9	0.7	0.5	0.2	0.1	0.05	0.02	0.01
n :								
1	0.158	0.510	1.000	3.078	6.314	12.706	31.821	63.656
2	0.142	0.445	0.816	1.886	2.920	4.303	6.965	9.925
3	0.137	0.424	0.765	1.638	2.353	3.182	4.541	5.841
4	0.134	0.414	0.741	1.533	2.132	2.776	3.747	4.604
5	0.132	0.408	0.727	1.476	2.015	2.571	3.365	4.032
6	0.131	0.404	0.718	1.440	1.943	2.447	3.143	3.707
10	0.129	0.397	0.700	1.372	1.812	2.228	2.764	3.169
15	0.128	0.393	0.691	1.341	1.753	2.131	2.602	2.947
20	0.127	0.391	0.687	1.325	1.725	2.086	2.528	2.845
25	0.127	0.390	0.684	1.316	1.708	2.060	2.485	2.787
30	0.127	0.389	0.683	1.310	1.697	2.042	2.457	2.750
35	0.127	0.388	0.682	1.306	1.690	2.030	2.438	2.724
40	0.126	0.388	0.681	1.303	1.684	2.021	2.423	2.704
60	0.126	0.387	0.679	1.296	1.671	2.000	2.390	2.660
80	0.126	0.387	0.678	1.292	1.664	1.990	2.374	2.639
100	0.126	0.386	0.677	1.290	1.660	1.984	2.364	2.626
200	0.126	0.386	0.676	1.286	1.653	1.972	2.345	2.601
400	0.126	0.386	0.675	1.284	1.649	1.966	2.336	2.588
600	0.126	0.386	0.675	1.283	1.647	1.964	2.333	2.584
800	0.126	0.385	0.675	1.283	1.647	1.963	2.331	2.582
1000	0.126	0.385	0.675	1.282	1.646	1.962	2.330	2.581
2000	0.126	0.385	0.675	1.282	1.646	1.961	2.328	2.578
4000	0.126	0.385	0.675	1.282	1.645	1.961	2.327	2.577
8000	0.126	0.385	0.675	1.282	1.645	1.960	2.327	2.576
10000	0.126	0.385	0.675	1.282	1.645	1.960	2.327	2.576

TABLE De t DE STUDENT

n est le nombre de degrés de liberté et P la probabilité au dépassement

Probabilité au dépassement:	0.99	0.95	0.9	0.5	0.2	0.1	0.05	0.02	0.01
1	0.000	0.004	0.02	0.45	1.64	2.71	3.84	5.41	6.63
2	0.020	0.103	0.21	1.39	3.22	4.61	5.99	7.82	9.21
3	0.11	0.35	0.58	2.37	4.64	6.25	7.81	9.84	11.34
4	0.30	0.71	1.06	3.36	5.99	7.78	9.49	11.67	13.28
5	0.55	1.15	1.61	4.35	7.29	9.24	11.07	13.39	15.09
6	0.87	1.64	2.20	5.35	8.56	10.64	12.59	15.03	16.81
7	1.24	2.17	2.83	6.35	9.80	12.02	14.07	16.62	18.48
8	1.65	2.73	3.49	7.34	11.03	13.36	15.51	18.17	20.09
9	2.09	3.33	4.17	8.34	12.24	14.68	16.92	19.68	21.67
10	2.56	3.94	4.87	9.34	13.44	15.99	18.31	21.16	23.21
11	3.05	4.57	5.58	10.34	14.63	17.28	19.68	22.62	24.73
12	3.57	5.23	6.30	11.34	15.81	18.55	21.03	24.05	26.22
13	4.11	5.89	7.04	12.34	16.98	19.81	22.36	25.47	27.69
14	4.66	6.57	7.79	13.34	18.15	21.06	23.68	26.87	29.14
15	5.23	7.26	8.55	14.34	19.31	22.31	25.00	28.26	30.58
16	5.81	7.96	9.31	15.34	20.47	23.54	26.30	29.63	32.00
17	6.41	8.67	10.09	16.34	21.61	24.77	27.59	31.00	33.41
18	7.01	9.39	10.86	17.34	22.76	25.99	28.87	32.35	34.81
19	7.63	10.12	11.65	18.34	23.90	27.20	30.14	33.69	36.19
20	8.26	10.85	12.44	19.34	25.04	28.41	31.41	35.02	37.57
21	8.90	11.59	13.24	20.34	26.17	29.62	32.67	36.34	38.93
22	9.54	12.34	14.04	21.34	27.30	30.81	33.92	37.66	40.29
23	10.20	13.09	14.85	22.34	28.43	32.01	35.17	38.97	41.64
24	10.86	13.85	15.66	23.34	29.55	33.20	36.42	40.27	42.98
25	11.52	14.61	16.47	24.34	30.68	34.38	37.65	41.57	44.31
26	12.20	15.38	17.29	25.34	31.79	35.56	38.89	42.86	45.64
27	12.88	16.15	18.11	26.34	32.91	36.74	40.11	44.14	46.96
28	13.56	16.93	18.94	27.34	34.03	37.92	41.34	45.42	48.28
29	14.26	17.71	19.77	28.34	35.14	39.09	42.56	46.69	49.59
30	14.95	18.49	20.60	29.34	36.25	40.26	43.77	47.96	50.89

Valeur de Chi2 En fonction du nombre de degrés de liberté (de 1 à 30) et de la probabilité au dépassement