BRGM

éléments d'analyse statistique

application à l'hydrologie

deuxième édition

D. Thiery

octobre 1989
R 30 173
EAU 4S 89

BUREAU DE RECHERCHES GÉOLOGIQUES ET MINIÈRES
SERVICES SOL ET SOUS-SOL
Département Eau
B.P. 6009 - 45060 ORLÉANS CEDEX 2 - France - Tél.: (33) 38.64.34.34

ELEMENTS D'ANALYSE STATISTIQUE
Application à l'Hydrologie

Dominique THIERY

PR 93.048.00263 - R 30173 EAU/4S/89

R E S U M E

Ce rapport présente un certain nombre de techniques statistiques élémentaires :

- étude d'un échantillon de valeurs et détermination de la fonction de distribution empirique,
- étude et méthode d'ajustement aux fonctions de distribution les plus classiques,
- calculs de régression linéaire,
- calcul des intervalles de confiances et pratique de tests statistiques (comparaison des moyennes ou de variance, analyse de variance),

Pour chadune de ces techniques, on s'est efforcé de préciser :

- les hypothèses d'application à respecter,
- les erreurs classiques à éviter.

Un exemple numérique a été traité à chaque fois pour rendre la méthode plus explicite.

S O M M A I R E

LISTE DES FIGURES

(En Annexe)

INTRODUCTION

Il existe un certain nombre de techniques qui permettent d'étudier un échantillon de valeurs pour en préciser sa distribution statistique et le comparer à un autre échantillon.

Un certain nombre de ces techniques sont très utilisées en hydrologie superficielle. Malheureusement, ces méthodes statistiques sont souvent utilisées de manière partielle, sans tenir compte des hypothèses d'application et donnent souvent lieu à des interprétations douteuses ou tendancieuses.

Le but de ce rapport est de rappeler un certain nombre de méthodes d'analyse statistiques élémentaires :

- en précisant les hypothèses d'application,
- en montrant comment calculer les intervalles de confiance de tous les résultats en fonction de la taille de l'échantillon,
- en attirant l'attention du lecteur sur les erreurs d'application les plus classiques.

Le texte qui suit a été rédigé à l'occasion des recyclages statistiques de novembre 1980 et mars 1981 à Orléans. Il reprend presque intégralement le texte de la note technique n° 80/16 après révision et corrections.

La deuxième édition, conçue à la suite du recyclage statistique de Septembre 1989 à Orléans, corrige et complète largement la première édition de 1981.

1. - ETUDE D'UN ECHANTILLON

1.1. GENERALITES ET DEFINITIONS

Soit une série de valeurs; par exemple :

* le débit journalier d'un cours d'eau,
* la pluie annuelle sur un bassin versant,
* le niveau piézométrique en un point,
* le prix de l'eau en différentes localités,
* la concentration en nitrates en différents points

Le but des statistiques est de voir s'il est possible de résumer les valeurs de la série par quelques paramètres caractéristiques, et de préciser la fréquence de ces différentes valeurs.

La série des valeurs disponibles constitue un *ECHANTILLON* comprenant un nombre limité de valeurs appelées *OBSERVATIONS*.

En général on ne s'intéresse pas particulièrement à l'*ECHANTILLON* disponible : (par exemple 5 valeurs de pluie annuelle) mais plutôt à la *POPULATION* d'où est tiré cet échantillon (la pluie annuelle).

Une *POPULATION* est un ensemble (théorique) constitué d'un nombre infini de valeurs. Cette population peut être décrite par quelques paramètres : par exemple par sa *MOYENNE*, son *ECART-TYPE* (voir plus loin) etc...

On s'intéresse le plus souvent à des populations *STABLES* c'est-à-dire dont les paramètres ne présentent pas de tendances.

Exemples :

* le prix de l'eau en différents points et à différentes dates ne constitue pas une population stable car en un point donné, le prix évolue (plus ou moins régulièrement) à la hausse en fonction du temps. Il y a donc une tendance qu'on peut éventuellement corriger en exprimant les prix en Francs (ou Dollars) constants.

* le débit d'un cours d'eau situé dans un bassin versant qui s'urbanise progressivement va évoluer. On ne pourra donc pas définir,par exemple la moyenne des débits.

* le débit mensuel d'un cours d'eau présente généralement une forte périodicité saisonnière. Le débit de chaque mois constitue cependant une population stable.

Les paramètres décrivant une population sont inconnus car pour les calculer il faudrait disposer de toutes les valeurs de la population or ces valeurs sont en nombre infini.

Exemple :

Quelle est la pluie annuelle moyenne à Orléans ? Pour connaître cette moyenne il faudrait disposer de toutes les valeurs de pluie depuis l'origine de la terre jusqu'à sa ... fin.

Pour étudier une population on étudie donc un échantillon formé par les observations disponibles. Pourquoi procède-t'on ainsi ?

* parce qu'on ne peut faire autrement,
* parce qu'on démontre que - si la population est stable - les paramètres estimés à partir d'un échantillon sont proches de ceux de la population. Ils sont d'autant plus proches que le nombre d'observations de l'échantillon est grand ("Loi des grands nombres").

On conçoit aisément que la moyenne de la pluie annuelle calculée sur 3 années d'observations a beaucoup plus de chances d'être éloignée de la moyenne vraie que celle calculée sur 50 années d'observations, En effet, sur les 3 années il peut y avoir par exemple 2 années très sèches (ou très humides) alors que sur 50 années il est très peu probable d'avoir 30 années sèches (ou humides).

Avant d'étudier un échantillon il faut tout d'abord vérifier :

a) que toutes ses valeurs proviennent de la population à étudier,
b) que cette population est stable,
c) que cette population est homogène.

Exemple : 1) Si on veut étudier les caractéristiques chimiques d'une nappe, il faut s'assurer que tous les prélèvements proviennent de la même nappe.

Exemple : 2) Voir plus haut : prix évoluant en fonction du temps.

Exemple : 3) Quand on étudie les débits de crue d'un cours d'eau il convient souvent de séparer les crues de printemps (fonte des neiges) de celles d'été (orages) qui proviennent de phénomènes essentiellement différents.

1.2. PARAMETRES STATISTIQUES D'UN ECHANTILLON

Dans ce paragraphe nous définirons seulement deux paramètres statistiques qui sont utilisés constamment et nous montrons comment les calculer.

Dans des chapitre ultérieurs nous montrerons comment apprécier la précision des paramètres obtenus.

La moyenne

C'est le paramètre définissant la tendance *CENTRALE* de l'échantillon :

$$m_x = \frac{1}{n} \Sigma\, x_i$$

Σ = signifie somme de tous les éléments x_i

n = nombre d'éléments de l'échantillon

m_x = moyenne des valeurs x_i

L'écart type

C'est le paramètre définissant la *DISPERSION* de l'échantillon autour de sa moyenne.

L'écart type a même dimension que les valeurs de l'échantillon (m^3/s, mm de pluie, etc...).

Plus l'écart type est grand et plus l'échantillon est étalé (ou dispersé). L'écart type est généralement noté σ (sigma) ou s :

$$\sigma_x = \sqrt{\frac{\Sigma (x_i - m)^2}{n - 1}}$$

(n = nombre d'éléments de l'échantillon)

(attention on divise par n-1 et non par n).

Calcul pratique :

a) Un certain nombre de calculatrices de poche permettent le calcul automatique de l'écart type, c'est l'idéal.

b) En développant l'expression de l'écart type on voit que l'on peut écrire :

$$\sigma_x = \sqrt{\frac{\Sigma x_i^2 - n. m_x^2}{n - 1}}$$

ce qui est beaucoup plus rapide à calculer, mais parfois nettement moins précis.

Un exemple de calcul est présenté dans le chapitre sur la régression linéaire (paragraphe 3.2.).

Coefficient de variation

C'est le rapport de l'écart type à la moyenne $C_v = \frac{\sigma x}{m_x}$ c'est un nombre sans *dimension* (donc indépendant des unités) qui ne présente de l'intérêt que dans la mesure ou la moyenne m_x a un sens physique (C_v n'a pas d'intérêt par exemple quand on étudie des charges qui sont définies par rapport à un repère arbitraire).

Il existe d'autres paramètres statistiques comme le coefficient d'asymétrie. Nous n'en parlerons pas ici. Il faut en effet disposer d'échantillons de grandes dimensions pour être en mesure de le définir avec précision.

1.3. ETUDE DESCRIPTIVE DE LA DISTRIBUTION STATISTIQUE D'UN ECHANTILLON

Une telle étude se propose de déterminer quelle est la distribution statistique des observations de l'échantillon c'est-à-dire :

* quel est le pourcentage d'observations comprises entre 2 valeurs (c'est-à-dire dans une *CLASSE* limitée par ces 2 valeurs)

ou encore :

* quel est le pourcentage d'observations supérieures à une certaine valeur.

La représentation graphique d'une telle étude est un *HISTOGRAMME* des fréquences qui permet d'en déduire un *HISTOGRAMME DES FREQUENCES CUMULEES.*

Avant d'expliquer comment tracer un histogramme il convient de définir ce que l'on appelera une *CLASSE* et une *FREQUENCE.*

* *LA CLASSE* [a, b] est l'ensemble des valeurs comprises entre a et b (inclus ou non suivant le cas).

* *LA FREQUENCE* de la *CLASSE* [a, b] est le pourcentage des observations qui appartiennent à la classe [a, b] .

S'il y a n_i observations dans la classe i la fréquence f_i de cette classe est :

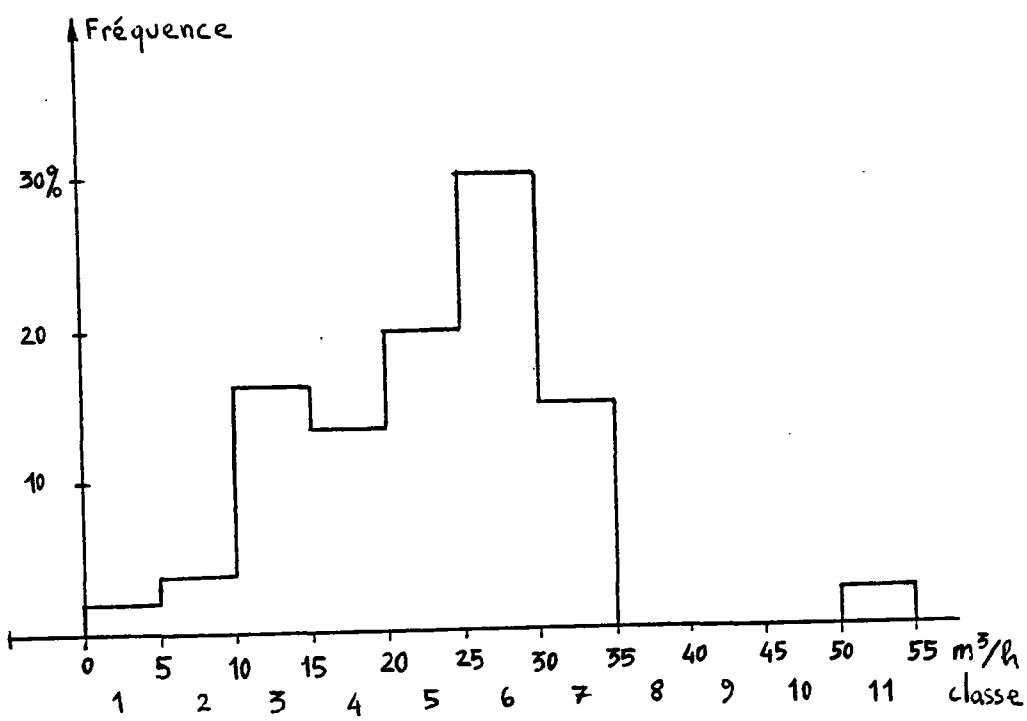
$$f_i = \frac{n_i}{N}$$

N : étant le nombre total d'observations

Construction d'un histogramme :

On détermine l'intervalle de variation, compris entre l'observation minimum et l'observation maximum, et on le divise en classes égales contenant un minimum d'environ 5 observations. Il est alors facile de calculer les "fréquences" de chaque classe. Il suffit alors de reporter en abscisse les classes et en ordonnée les fréquences correspondantes.

Exemple : Débits maximal des forages d'une région.



La classe ayant la plus forte fréquence est appelée le *Mode*, (ici c'est la classe de 25 à 30 m^3/h).

Quand il y a deux "pics" de fréquence, la répartition est *bimodale*. Un histogramme bimodal indique souvent que l'échantillon provient de deux populations différentes ayant chacune un mode :

Exemple : pour une rivière :

* les débits d'automne, d'origine pluviale,
* les débits de printemps, d'origine nivale.

Il est conseillé d'étudier alors séparément les deux populations.

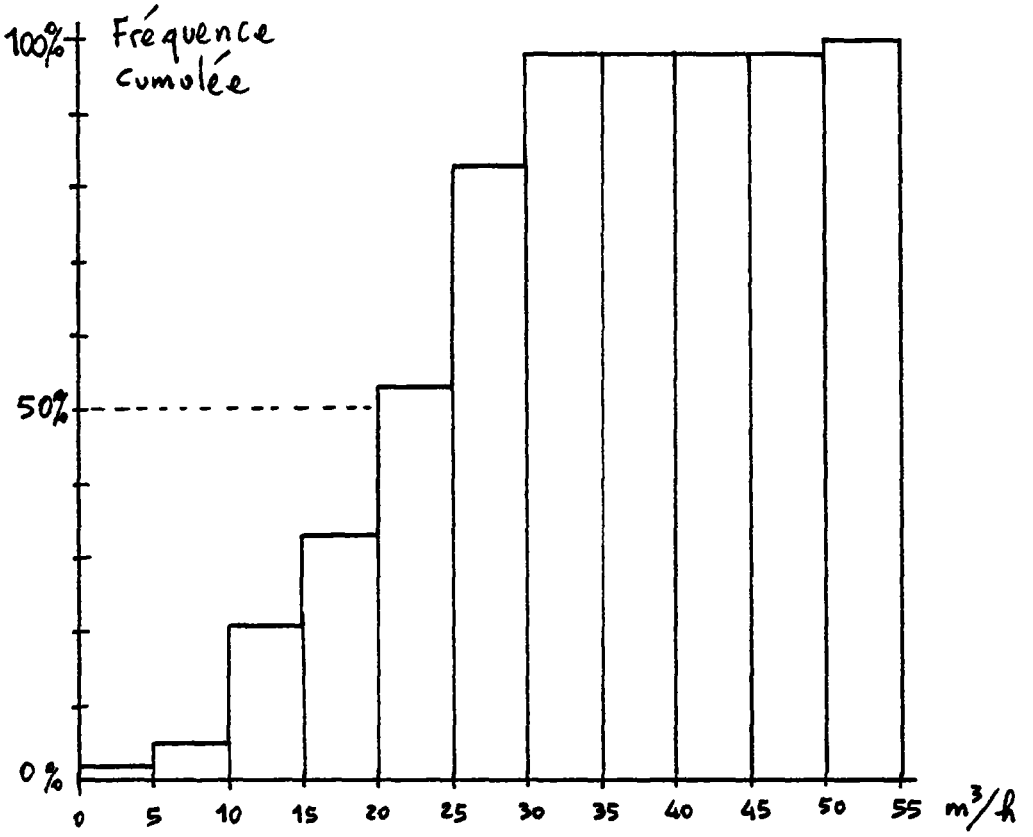
Remarque : la somme des fréquences est égale à 1 en effet toutes les valeurs sont affectées à une classe donc toutes les classes contiennent au total 100 % des valeurs.

Histogramme cumulé

On s'intéresse maintenant à la fréquence des valeurs comprises dans toutes les classes inférieures ou égales à la classe i.

C'est donc le pourcentage des valeurs *qui ne dépassent* pas la classe i. La fréquence cumulée est donc la *FREQUENCE DE NON DEPASSEMENT*.

On dira par exemple que 83 % des forages ont un débit maximal qui ne dépasse pas 30 m^3/h.Pour construire un histogramme cumulé, il suffit d'affecter à chaque classe la somme de toutes les fréquences jusqu'à cette classe. On reporte alors la fréquence cumulée en fonction de la classe.



Définition : La médiane est la valeur dont la fréquence de non dépassement est de 50 % (50 % des valeurs lui sont supérieures ; 50 % lui sont inférieures).

Dans notre exemple la médiane appartient à la classe $\left[20,\ 25\right]$ m^3/h.

1.4. DETERMINATION DE LA FONCTION DE DISTRIBUTION

L'histogramme dépend beaucoup du découpage et du nombre des classes qui est bien entendu arbitraire. L'histogramme cumulé dépend beaucoup moins de ce découpage et tend vers une limite stable quand le nombre de classes augmente jusqu'à ne conserver qu'au maximum une valeur par classe.

A toute valeur X, (et non plus à toute classe i) on peut associer une fréquence de non dépassement (ou probabilité de non dépassement). La fonction F qui associe la fréquence de non dépassement à la valeur X est la *FONCTION DE DISTRIBUTION* de l'échantillon étudié.

L'étude de la distribution de la population consiste à essayer d'identifier la fonction de distribution de la population.

Calcul pratique

En pratique on peut procéder de deux façons :

a) calculer la fréquence cumulée F (X) pour chaque observation X,
b) ou répartir ces observations en classes et calculer F (X) pour chaque classe.

1ère méthode

- classer toutes les valeurs X par ordre croissant
- leur affecter leur numéro d'ordre j
- calculer la valeur *estimée* de F (X)

$$F(X) = \frac{2j - 1}{2N + 1}$$

N : étant le nombre total d'observations

- tracer F (X) en fonction de X (voir remarque 2)

Cette méthode est la plus appropriée quand on dispose d'un nombre de valeurs inférieur à 50.

2ème méthode

- diviser l'intervalle en p classes égales (contenant chacune au moins 3 ou 5 valeurs)
- compter le nombre K de valeurs dans chaque classe,
- pour chaque classe calculer la somme j des nombres K dans les classes inférieures ou égales
- calculer la valeur estimée de F (X) : X = borne supérieure de la classe

$$F(X) = \frac{2j - 1}{2N + 1}$$

X étant la valeur correspondant au milieu de chaque classe.

Remarque 1 : Quand on possède beaucoup de données, la 2ème méthode est beaucoup plus rapide car il est très long de classer un grand nombre d'observations.

Exemple : Pour classer N = 1000 observations, il faut au maximum 500 000 opérations ; pour les ranger en 50 classes, il en faut au maximum 50 000 soit 10 fois moins.

Remarque 2 : Lors de la construction de l'histogramme nous avons calculé la fréquence cumulée par $F(X) = \frac{j}{N}$

mais pour l'identification de la fonction de distribution on utilise

$$F(X) = \frac{2j - 1}{2N + 1}$$

pour la valeur maximale : j = N on trouve $F(X \max) = \frac{2N - 1}{2N + 1}$ qui est donc inférieure à 100 %

En effet, la valeur maximale observée est généralement inférieure à la valeur maximale possible de la population.

Remarque 3 : On utilise parfois F(X) = j/(N+1).

Définition : On appelle TEMPS DE RETOUR l'intervalle de temps MOYEN T séparant des évènements de fréquence de non-dépassement F. Il existe la relation suivante entre T et F :

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

Remarque 1 : Il convient de remarquer que T est un temps MOYEN. Par exemple si en 36 mois on observe, dans un cours d'eau, 9 crues supérieures à 15 m3/s on déduira que $T = \frac{36 \text{ mois}}{9 \text{ crues}} = 4$ mois mais il se peut très bien qu'on observe 2 ou 3 crues supérieures à 15 m3/s dans un même mois particulier.

Remarque 2 : Il est évident que la notion de temps de retour n'a de sens que quand on étudie des évènements successifs. Il n'y aurait pas lieu de calculer un temps de retour quand on étudie des variables indépendantes du temps (par exemple le débit maximal d'un forage dans une région).

En reportant sur un graphique les points définis par X, F(X) on peut déterminer la distribution empirique de l'échantillon en faisant passer une courbe lisse parmi les points. Il est alors possible d'INTERPOLER la fréquence de non-dépassement F de toute valeur X donnée ou réciproquement la valeur X correspondant à toute fréquence de non dépassement F donnée.

Cependant, en général, il ne faut pas s'arrêter à ce stade et il faut essayer de voir si la distribution identifiée correspond à une distribution connue définie par quelques paramètres. L'ensemble des observations pourra alors être représenté par ces seuls paramètres.

La structure de la distribution de la population n'étant en général pas connue a priori, il est très dangereux d'extrapoler la courbe en-dehors des valeurs observées. Il est en effet souvent possible d'ajuster de façon aussi satisfaisante à un même échantillon, plusieurs lois statistiques qui conduisent à des extrapolations extrêmement différentes.

On étudie généralement les lois de distribution en variable réduite : Dans ce but on pose :

$$u = \frac{x - mx}{\sigma}$$

- u = variable centrée et réduite
- x = variable aléatoire à étudier
- mx = moyenne de x
- σ = écart-type de x

la moyenne de u est alors : 0
son écart-type est : 1

Exemple 1 : Soient les débits mensuels du mois de septembre d'un cours d'eau pendant 25 ans.

Année	Débit (m^3/s)	Après classement m^3/s	Numéro	Fréquence de non dépassement %
1956	22	6	1	2.0
1957	40	6	2	5.9
1958	11	11	3	9.8
1959	24	11	4	13.7
1960	6	11	5	17.6
1961	55	11	6	21.6
1962	23	20	7	25.5
1963	160	21	8	29.4
1964	64	22	9	33.3
1965	21	22	10	37.3
1966	20	23	11	41.2
1967	45	23	12	45.1
1968	11	24	13	49.0
1969	23	25	14	52.9
1970	11	31	15	56.9
1971	11	40	16	60.8
1972	47	45	17	64.7
1973	87	45	18	68.6
1974	6	46	19	72.5
1975	85	47	20	76.5
1976	46	55	21	80.3
1977	25	64	22	84.3
1978	22	85	23	88.2
1979	45	87	24	92.2
1980	31	160	25	96.1

Cet exemple est représenté sur les figures 1 à 3 placées en annexe.

Exemple 2 : On dispose de 396 essais à l'air-lift de 396 forages. Il serait trop long d'appliquer la méthode précédente. On les répartit donc dans 11 classes de débits. On obtient ainsi :

numéro de la classe	1	2	3	4	5	6	7	8	9	10	11
limites de la classe m^3/h	0 5	5 10	10 15	15 20	20 25	25 30	30 35	35 40	40 45	45 50	50 55
nombre de forages	8	8	64	44	84	112	68	0	0	0	8
nombre cumulé	8	16	80	124	208	320	388	388	388	388	396

On calcule alors la fréquence cumulée F d'après le nombre cumulé de forages j suivant la formule :

$$F = \frac{2j - 1}{2 \times 396 + 1}$$

Cette fréquence est affectée à la borne supérieure de chaque classe. On obtient ainsi :

X (m^3/h)	2.5	7.5	12.5	17.5	22.5	27.5	32.5	37.5	42.5	47.5	52.5
F(X) (%)	1.9	3.9	20.1	31.1	52.3	80.6	97.7	97.7	97.7	97.7	99.7

Cet exemple correspond à l'histogramme donné dans le paragraphe 1.3.

F(X) est la fréquence cumulée associée à la classe de centre X ; c'est la fréquence de non dépassement de la borne supérieure de la classe de centre X.

2.- LES DISTRIBUTIONS STATISTIQUES

Nous n'aborderons ici que les distributions les plus courantes et les plus simples.

2.1. *LA DISTRIBUTION GAUSSIENNE (OU NORMALE)*

C'est la distribution la plus connue. Elle a un rôle très important en statistique car c'est la limite d'un certain nombre d'autres distributions (Student, CHI2 ...). Elle est représentée graphiquement par la fameuse "courbe en cloche".

On démontre que c'est la distribution que suit un phénomène aléatoire résultant de la SOMME d'un grand nombre de facteurs (aléatoires) indépendants et de même importance et ceci quelque soit la distribution statistique de chacun de ces facteurs.

⁂ Quelles sont les variables hydrologiques qui suivent en général une distribution Gaussienne ?

- La température (journalière, mensuelle ou annuelle)
- Le débit moyen annuel d'un cours d'eau
- La pluie annuelle
- L'ETP annuelle
- Toute variable qui est une SOMME ou une MOYENNE (ce qui revient au même).
- La partie centrale de toute distribution

⁂ Quelles sont les variables hydrologiques qui ne suivent absolument pas une distribution Gaussienne ?

- La pluiejournalière (ou mensuelle)
- Le débit journalier d'un cours d'eau
- Le nombre de jours sans pluie

  Et surtout :
- Le débit journalier maximal de chaque année
- La pluie journalière maximale de chaque année
- La transmissivité d'un aquifère
- En général tout ce qui est un extrême.

Propriétés :

. C'est un distribution symétrique par rapport à la médiane.
. La médiane est égale à la moyenne
. Elle est entièrement définie par 2 paramètres :
La moyenne et l'écart-type

La figure 4 donne un tableau des fréquences cumulées correspondant à la variable réduite :

$$u = \frac{x - mx}{\sigma}$$

$\left.\begin{array}{l} mx = \text{moyenne} \\ \sigma = \text{écart-type} \end{array}\right\}$

un examen de ce tableau permet de déduire les valeurs suivantes :

Pourcentage d'observations	Dans l'intervalle
68 %	$\bar{x} \pm 1.00\ \sigma$
80 %	$\bar{x} \pm 1.28\ \sigma$
90 %	$\bar{x} \pm 1.64\ \sigma$
95 %	$\bar{x} \pm 1.96\ \sigma$
99 %	$\bar{x} \pm 2.57\ \sigma$

Méthode d'ajustement

En pratique, pour vérifier si un échantillon peut être représenté par une distribution Gaussienne, on calcule sa fonction de distribution F(X) et on la trace sur un papier spécial appelé papier Gausso-Arithmétique. Ce papier est conçu de telle sorte qu'une distribution Gaussienne soit représentée par une droite : on peut donc voir immédiatement si une droite peut s'ajuster au "nuage" de points, exemple : figure 1 . On ne peut visiblement pas ajuster une droite.

Si on peut ajuster une droite, il est possible d'obtenir très rapidement les 2 paramètres qui définissent cette distribution.

$F = 50\ \% \longrightarrow$ médiane = moyenne mx

$F = 84\ \% \longrightarrow mx + \sigma$ donc $\sigma_x = X_{84\%} - X_{50\%}$

REMARQUE IMPORTANTE : Lors de l'ajustement il ne faut pas trop s'occuper des valeurs extrêmes (les 2 plus grandes et les 2 plus petites) car les valeurs sont par nature très aléatoires, ceci est valable pour toutes les distributions.

2.2. LA DISTRIBUTION LOG NORMALE (Galton-Gibrat)

C'est la même distribution que la précédente mais appliquée au logarithme (décimal ou Népérien) des observations.

* Quelles sont les variables hydrologiques qui suivent généralement une loi Log-Normale ?

- Les débits journaliers (non nuls) d'un cours d'eau
- Les pluies journalières (non nulles)
- La transmissivité d'une nappe
- Les débits mensuels des petits cours d'eau.

Comment voir si une distribution peut s'ajuster à une distribution Log-Normale ?

En traçant la fonction de distribution sur papier Gauss on voit que la variable est "plafonnée" pour les faibles fréquences et croît très rapidement pour les fortes fréquences : voir figure 1.

Méthode d'ajustement

Pour ajuster une distribution Log-Normale on utilise un papier Gausso-Log. Sur un tel papier une distribution Log-Normale est représentée par une droite : voir figure 2.

En regardant les valeurs correspondant à F = 50 % et 84 % on n'obtient plus la moyenne m_x et l'écart type σx de la variable x mais il est possible de les calculer :

On pose $y = \text{Ln}(x)$

$\begin{cases} F = 50\ \% \longrightarrow x_1 & \text{d'où on déduit } m_y = \text{Ln } x_1 \\ F = 84\ \% \longrightarrow x_2 & \text{d'où on déduit } \sigma y = \text{Ln } x_2 - \text{Ln } x_1 \end{cases}$ (Ln étant le logarithme Néperien)

on calcule alors :

$$\begin{cases} m_x = \exp\left(m_y + \sigma_y^2/2\right) \\ \sigma_x = m_x \sqrt{\exp(\sigma_y^2) - 1} \end{cases}$$ puis

exemple :

sur la figure 2 on lit ,

$\begin{cases} F = 50\ \% \longrightarrow x_1 = 28\ m^3/h \\ F = 84\ \% \longrightarrow x_2 = 68\ m^3/h \end{cases}$

$\begin{cases} m_y = 3.33 \\ \sigma_y = 4.22 - 3.33 = 0.89 \end{cases}$

$\begin{cases} m_x = 41.51\ m^3/h \\ \sigma_x = 45.62\ m^3/h \end{cases}$

Inversement si la distribution de la variable x est Log Normale le logarithme de la variable y = Ln x suit une loi Gaussienne dont les paramètres sont :

$$\begin{cases} \sigma_y = \sqrt{\text{Ln}\left(1 + \sigma^2_x / m^2_x\right)} \\ \text{et} \\ m_y = \text{Ln}\left(m_x\right) - \sigma_y^2/2 = \text{Ln}\left(\dfrac{mx}{\sqrt{1 + \sigma^2_x / m^2_x}}\right) \end{cases}$$

2.3. LA DISTRIBUTION DE GUMBEL

C'est une distribution souvent utilisée pour représenter des valeurs extrêmes c'est pourquoi on l'appelle parfois "Loi des valeurs extrêmes". Ce n'est cependant absolument pas un postulat et les valeurs extrêmes ne suivent pas forcément une distribution de Gumbel.

Méthode d'ajustement

Il existe un papier spécial dit papier Gumbel sur lequel la représentation d'une distribution de Gumbel est une droite (voir figure 3).

Pour calculer la moyenne et l'écart type on procède ainsi :

$$\begin{cases} F = 0.570 \longrightarrow x_1 \\ (T = 2.33) \\ F = 0.856 \longrightarrow x_2 \end{cases} \qquad \begin{cases} m_x = x_1 \\ \sigma_x = x_2 - x_1 \end{cases}$$

Propriétés

L'expression de la fréquence cumulée F est

$$\boxed{F = \exp\left(-\exp\left(-\frac{x-x_f}{g}\right)\right)}$$ x_f étant la valeur modale

$$\begin{cases} \sigma_x = 1.28\ g \\ m_x = x_f + 0.577\ g = x_f + 0.4500\ \sigma_x \end{cases}$$

La quantité g (qui est égale à 0.78 σ_x) est appelée GRADEX c'est à dire GRADIENT des valeurs EXTREMES.

NOTA. Si F proche de 100 % on peut écrire : Ln T ≃ (x - xf)/g

2.4. LA DISTRIBUTION DE STUDENT

C'est une distribution très utilisée pour les tests statistiques. Elle fait intervenir — en plus de la moyenne et de l'écart-type — le nombre de degrés de liberté (c'est à dire le nombre de variables indépendantes) qui est généralement noté ν (NU).

La figure 5 donne un tableau de la fréquence de non-dépassement correspondant à la variable réduite $t = \frac{x - mx}{\sigma}$ de ce tableau on peut extraire les valeurs approximatives suivantes :

Pourcentage d'observations	comprises entre : $mx \pm t.\sigma$										
	ν	1	2	3	4	5	6-7	8-9	10-13	14-27	>27
80 %	t=	3.1	1.9	1.6	1.5	1.5	1.4	1.4	1.4	1.3	1.3
95 %	t=	12.7	4.3	3.2	2.8	2.6	2.4	2.3	2.2	2.1	2.0

↑ loi Normale

ν = nombre de degrés de liberté

2.5. *LA DISTRIBUTION BINOMIALE*

C'est une distribution s'appliquant à des nombres entiers (distribution "discrète"). C'est la distribution du nombre k d'évènements de fréquence de non-dépassement F dans un échantillon de n valeurs. La probabilité P_k d'obtenir exactement k évènements de fréquence de non dépassement F dans un échantillon de n valeurs est :

$$P_k = C_n^k (1-F)^k F^{n-k} \quad \text{avec} \quad C_n^k = \frac{n!}{k!(n-k)!}$$

(P_k est la fréquence de la valeur k.)

Application :

Soit un échantillon de 10 années de débit d'un cours d'eau. Quel est la probabilité d'observer exactement k crues décennales ?

$$\begin{cases} n = 10 \\ F = 0.9 \end{cases}$$

on applique la formule pour k = 0 à 5. On trouve alors (en se souvenant que $C_n^0 = 1$) la probabilité d'observer 0, 1, ..., 5 crues au moins décennales dans une période de 10 ans, c'est-à-dire en fait le POURCENTAGE de périodes de 10 ans au cours desquelles on observe 0, 1, ... 5 crues au moins décennales

Nombre de crues décennales	Fréquence d'observations
0	34.87 %
1	38.74 %
2	19.37 %
3	5.74 %
4	1.12 %
5	0.15 %

On voit ici que bien qu'on observe EN MOYENNE une crue décennale tous les 10 ans, on n'observe exactement 1 crue décennale que dans 39 % des périodes de 10 ans, 0 crue décennale dans 35 % des périodes de 10 ans et 2 crues décennales ou plus dans 26 % des périodes de 10 ans.

Propriétés

* moyenne = n (1-F)
ce qu'on vérifie bien dans notre exemple

$n = 10$, $F = 0.9$ } $m = 10 \times (1-0.9) = 1$ crue, par période de 10 ans en moyenne

* écart-type = $\sqrt{nF(1-F)}$ (0,95 dans notre exemple)

2.6. LA DISTRIBUTION DE POISSON

C'est la limite de la distribution binomiale quand n est grand et F proche de 100 % (c'est à dire pour les valeurs rares).

La probabilité P_k d'obtenir k évènements de fréquence de non dépassement F dans un échantillon de n observations est :

$$P_k = \frac{m^k e^{-m}}{k!} \quad \text{avec} \quad m = n\,(1-F)$$

expression beaucoup plus facile à utiliser que celle de la distribution binomiale.

Propriétés :

moyenne : $m = n(1-F)$

écart-type : $\sigma = \sqrt{n(1-F)} = \sqrt{m}$

⁂ Quelles sont les variables hydrologiques qui suivent une distribution de Poisson ?

- Le nombre de jours de pluie par mois, par an
- Le nombre de crues, de fréquences données, par an, par siècle.

Application : On reprend l'exemple précédent :

$F = 0.9$, $n = 10$ ans } quelle est la probabilité d'observer k crues décennales ?

$m = 1$ on trouve alors :

Nombre de crues	0	1	2	3	4	5
probabilité	36.79	36.79	18.39	6.13	1.53	0.31

valeurs proches de celles obtenues avec la distribution binomiale

2.7. LA DISTRIBUTION EXPONENTIELLE

Quelles sont les variables hydrologiques qui suivent une distribution exponentielle ?

- La durée d'un évènement :
  ex : ⁂ durée d'une averse
  ⁂ durée d'un épisode sans pluie
  ⁂ durée pendant laquelle un niveau (ou un débit est dépassé)

L'expression de la Fréquence de non dépassement est :

$$F(x) = 1-\exp(-\lambda x)$$

propriété : moyenne = $m_x = 1/\lambda$
écart type = $\sigma_x = \sqrt{2/\lambda}$

remarque : $Ln(T) = \lambda x$

2.8. *DISTRIBUTION DU CHI2 (prononcer KI-DEUX)*

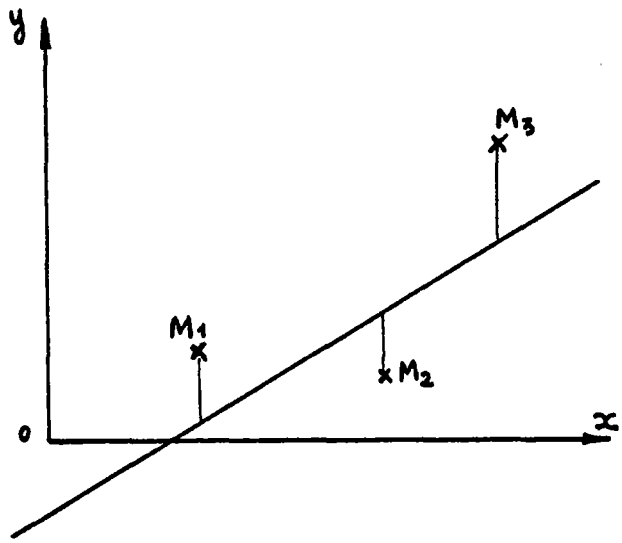
C'est une distribution très utilisée pour les tests statistiques. Elle dépend du nombre ν de degrés de liberté (voir table 6).

3.- LA REGRESSION LINEAIRE

3.1. *GENERALITES*

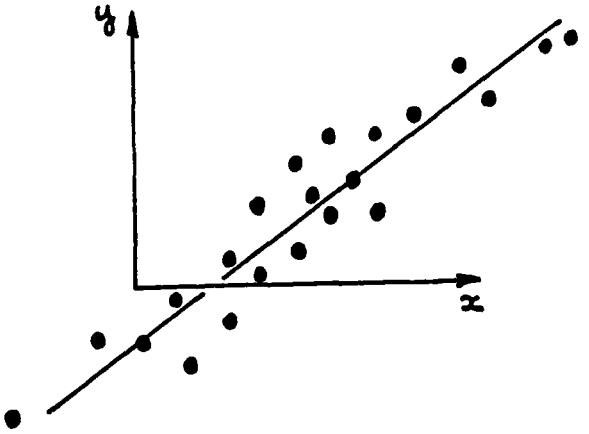
C'est une technique (parfois appelée à tort "correlation") qui permet d'étudier la liaison linéaire existant entre 2 échantillons de valeurs x et y.

La régression linéaire permet de déterminer la relation linéaire avec laquelle on peut "le mieux" calculer les valeurs de l'échantillon y à partir de celles de l'échantillon x. Ce n'est pas la même que celle qui permet de calculer x à partir de y.

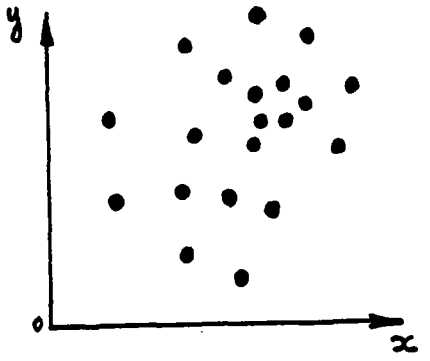


La méthode de calcul est celle des "moindrescarrés" qui consiste à rendre minimale la somme des CARRES des distances VERTICALES entre les points de coordonnées x_i, y_i et la droite représentant la relation linéaire $y = ax+b$.

En pratique il est donc INDISPENSABLE de vérifier, avant tout calcul, que le nuage de point a une forme allongée au milieu de laquelle il paraît possible de tracer une droite.

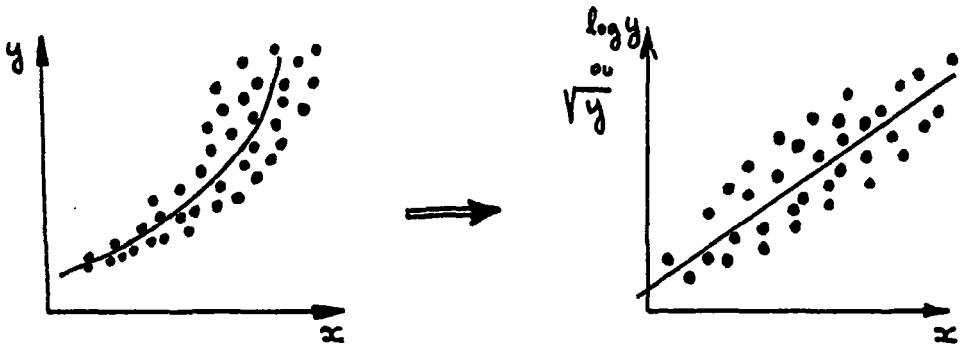


Possibilité de tenter une régression linéaire

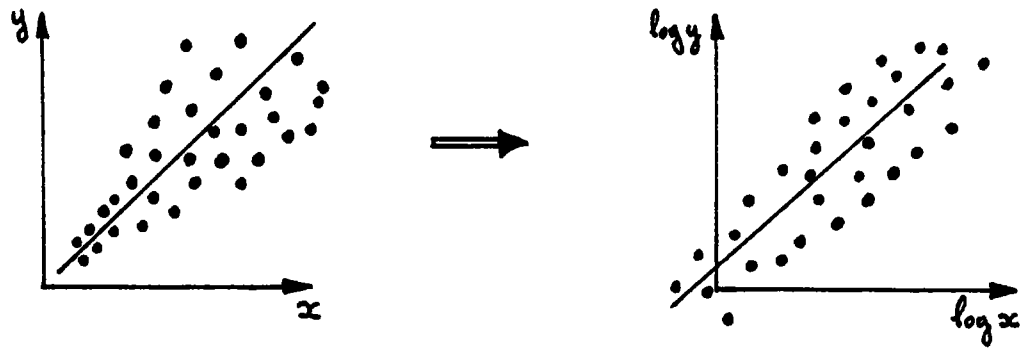


Impossible de tenter une régression

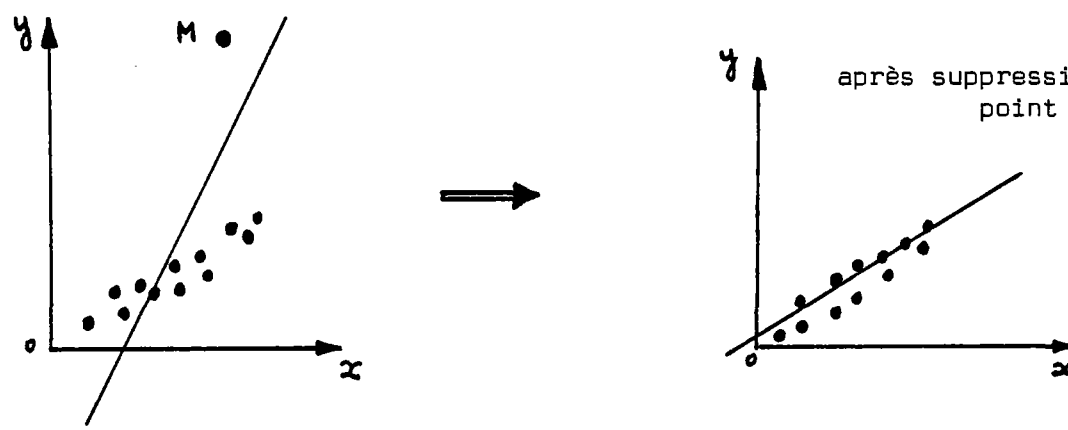
Si le nuage a une forme qui peut être approchée par une autre courbe qu'une droite, il est souvent possible, pour une tranformation des données d'obtenir un nuage plus linéaire. On est parfois guidé, par la nature des séries pour effectuer la transformation.



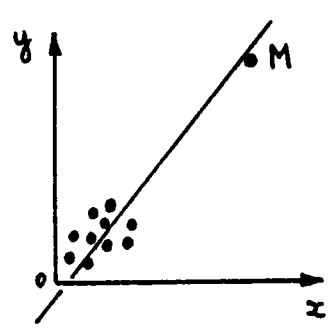
Si le nuage de points n'est pas "elliptique" mais a au contraire une dispersion qui augmente avec les fortes valeurs de x et de y on peut tenter de prendre la racine carrée (ou le logarithme à condition bien sûr que les valeurs soient toutes positives).



Dans certains cas, il conviendra de faire une critique des données pour éventuellement supprimer certaines valeurs erratiques qui par leur poids trop important risqueraient de fausser l'ajustement.



après suppression du point M



illusion d'un très bon ajustement (r élevé)

3.2. METHODE DE CALCUL

Il faut d'abord calculer le coefficient de correlation entre les valeurs de x et celles de y.

Calcul du coefficient de correlation

Remarque 1 : Un certain nombre de calculatrices de poche permettent d'effectuer ce calcul automatiquement.

Σx Σx^2	Σy Σy^2	Σxy
$m_x = \frac{1}{n}\Sigma x$	$m_y = \frac{1}{n}\Sigma y$	
$\sigma_x = \sqrt{\frac{1}{n-1}(\Sigma x^2 - n.m_x^2)}$	$\sigma_y = \sqrt{\frac{1}{n-1}(\Sigma y^2 - n.m_y^2)}$	$C = \frac{\Sigma xy}{n} - m_x.m_y$

$$r = \frac{C}{\sigma_x \cdot \sigma_y}$$ (C s'appelle la covariance)

Le coefficient de correlation r, est toujours compris entre -1 et +1.

$\left\{\begin{array}{l} r = +1 \text{ ou } -1 \text{ indique une relation linéaire parfaite} \\ r = 0 \quad \text{indique une relation linéaire nulle} \end{array}\right.$

comme on le voit dans le calcul de r, c'est une expression symétrique de x et y

$r(x,y) = r(y,x)$

Calcul des coefficients de régression

Un certain nombre de calculatrices de poche permettent ce calcul automatiquement

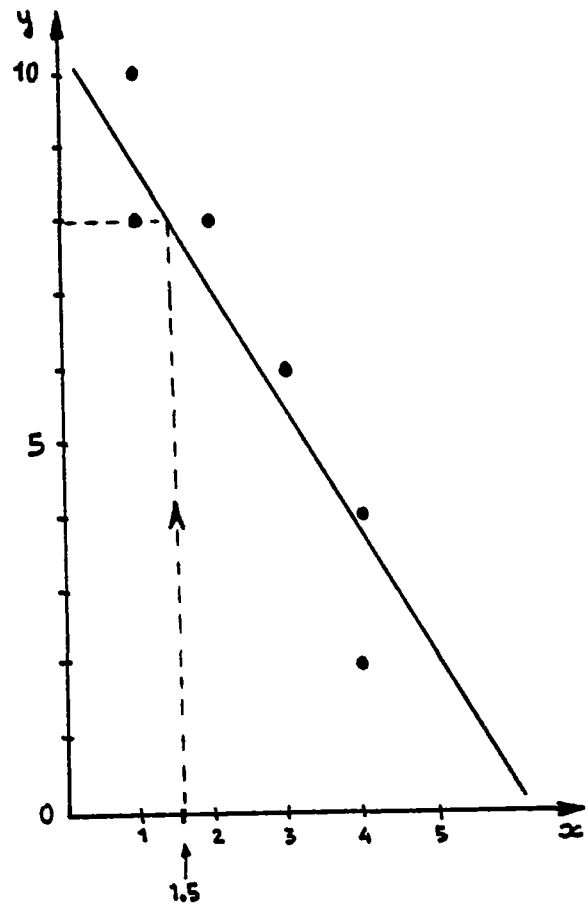
$y = ax + b$ avec :

$$\begin{cases} a = r.\sigma y/\sigma x & (1) \\ b = m_y - a.m_x & (2) \end{cases}$$

l'égalité (2) indique que la droite de régression passe par le centre de gravité (m_x, m_y) du nuage de points car $m_y = am_x + b$.

Application :

x	2	4	1	3	1	4
y	8	2	8	6	10	4



$n = 6$		
$\Sigma x = 15$	$\Sigma y = 38$	
$\Sigma x^2 = 47$	$\Sigma y^2 = 284$	$\Sigma xy = 76$
$mx = 2.50$	$my = 6.33$	
$\sigma x = 1.38$	$\sigma y = 2.95$	$C = -3.16$

$r = -0.78$
$a = -1.66$
$b = 10.48$

d'où l'ajustement $y = -1.66\ x + 10.48$

On peut alors calculer l'écart ε = y-ax-b pour chaque observation.

x	2	4	1	3	1	4
y	8	2	8	6	10	4
ε	0.84	-1.84	-0.82	0.50	1.18	0.16

La somme des écarts est égale à 0.02 donc quasiment égale à 0 (aux erreurs d'arrondis près) ce qui est normal.

3.3. *INTERPRETATION STATISTIQUE*

Jusqu'à ce stade nous n'avons fait aucune hypothèse sur les distributions statistiques de x et y. Il est toujours possible de calculer un coefficient de correlation entre 2 échantillons d'observations.

L'interprétation du coefficient de correlation nécessite cependant des hypothèses :

On suppose maintenant que :

- Les valeurs de y sont indépendantes les unes des autres.
- y suit une distribution à peu près Gaussienne
- L'écart ε = y-ax-b suit également une distribution à peu près Gaussienne. Cette distribution est la même pour toute valeur de x.

Le carré du coefficient de correlation est alors le pourcentage de la variance de y qui est expliquée par la liaison linéaire. Le reste est la variance résiduelle c'est la variance de l'écart (la variance est le carré de l'écart-type).

$$\underset{\text{variance totale}}{\sigma_y^2} = \underset{\text{variance expliquée}}{r^2 \cdot \sigma_y^2} + \underset{\text{variance résiduelle}}{\sigma_\varepsilon^2}$$

on en déduit donc que la variance résiduelle (variance de l'écart) est égale à $(1-r^2)\ \sigma_y^2$

Application : en reprenant l'exemple précédent :

$n = 6$

$r = 0.78$

$\sigma y = 2.95$

On suppose que les observations sont indépendantes. En première approximation la répartition peut être considérée comme Gaussienne. On en déduit donc :

variance expliquée = $r^2 = 61\%$

l'écart-type sans biais de l'écart est : $\sigma_\varepsilon = \sqrt{1-r^2} \cdot \sqrt{\frac{n-1}{n-2}} \cdot \sigma_y = 0.70\sigma_y = 2.06$

l'utilisation de la régression linéaire permet ici de réduire l'écart-type de la variable y à l'écart-type de l'écart qui est diminué d'environ 1/3.

Rappelons que cette interprétation statistique n'a de sens que si :

* Les observations sont indépendantes
* La variable et le résidu suivent des distributions à peu près Gaussienne. L'écart-type du résidu étant indépendant de x.

Si ces hypothèses ne sont pas vérifiées, on peut essayer de s'y ramener :

- en rendant la variable Gaussienne par transformation

- en ne sélectionnant que des observations indépendantes.

4.- TESTS STATISTIQUES ET CALCUL DES INTERVALLES DE CONFIANCE

4.1. INTERVALLE DE CONFIANCE D'UNE MOYENNE

Soit un échantillon de n valeurs INDEPENDANTES ; on calcule m_x et σ_x comme il a été expliqué plus haut. Pour définir la précision avec laquelle on connait m_x on calcule l'Intervalle de Confiance (I.C.) qui a un certain pourcentage de chance de contenir la vraie valeur. On choisit généralement l'intervalle de confiance à 95 %;parfois l'intervalle de confiance à 80 %. C'est à dire que si on fait cette opération pour chaque échantillon disponible de n valeurs, 95 % (ou 80%) des intervalles de confiances calculés contiendront la vraie moyenne m de la population. Réciproquement on peut dire (en simplifiant) qu'on a un risque de 5 % (ou 20 %) seulement que la vraie moyenne soit en-dehors de l'intervalle calculé.

Méthode pratique :

La variable m_x suit une distribution de Student à n-1 degrés de liberté d'écart-type : $\frac{\sigma x}{\sqrt{n}}$

on utilise donc la table de Student (figure 5) et on lit à la ligne $\nu = n-1$ la valeur de t correspondant à la précision cherchée (95 %).

(Rappelons que dès que $n \geq 15$ on a $t_{80\%} = 1.3$ et $t_{95\%} = 2.0$)

L'intervalle de confiance est alors :

$$\boxed{m_x \pm t. \frac{\sigma x}{\sqrt{n}}}$$

exemple : la moyenne sur 5 ans de la pluie annuelle à une station est de 800 mm. L'écart-type calculé sur les 5 ans est de 200 mm.

$$\begin{cases} n = 5 \\ \sigma x = 200 \text{ mm} \end{cases}$$

dans la figure 5, on lit à la ligne n-1 = 4. On trouve, pour 95 %, t = 2.78. On en déduit que l'intervalle de confiance à 95 % est :

$$800 \pm 2.78. \frac{200}{\sqrt{5}} = 800 \pm 249 \text{ mm} = [551, 1049] \text{ mm}$$

Si le nombre d'année n'est plus 5 mais 30, on trouve pour ces mêmes valeurs : t = 2.05 soit

$$800 \pm 2.05 . \frac{200}{\sqrt{30}} = [725, 875] \text{ mm}$$

on voit que l'intervalle de confiance a nettement diminué ce qui est normal car 30 années permettent de définir une moyenne bien mieux que 5 années.

Rappelons que ce calcul n'est valable que si les n valeurs sont INDEPENDANTES.

4.2. *INTERVALLE DE CONFIANCE D'UN ECART-TYPE*

Soit un échantillon de n valeurs INDEPENDANTES sur lequel on calcule un écart-type σ_x. On montre que :

$(n-1) \frac{\sigma_x^2}{\sigma^2}$ suit une distribution du χ^2 (CHI^2) à n-1 degrés de liberté

(σ étant l'écart-type de la population).

L'intervalle de confiance à 95 % a donc pour limites σ_{min}, σ_{max} définis par :

$$\sigma^2_{min} = \frac{(n-1)\sigma_x^2}{\chi^2(97,5\%)} \quad \text{et} \quad \sigma^2_{max} = \frac{(n-1)\sigma_x^2}{\chi^2(2,5\%)}$$

on calcule σ^2 inconnu d'après σ_x^2 connu.

il suffit alors de prendre la racine carrée pour obtenir σ_{min} et σ_{max}

Application :

$\begin{cases} n=5 \\ \sigma_x = 200 \end{cases}$

sur la ligne $n-1 = 4$ de la figure 6 , on lit : $\begin{cases} \chi^2(2.5\%) = 0.48 \\ \chi^2(97.5\%) = 11.1 \end{cases}$

soit $\sigma^2_{min} = \frac{4x200^2}{11.1}$; $\sigma^2_{max} = \frac{4x200^2}{0.48}$

c'est à dire $120 < \sigma_x < 575$ mm

On remarque que la distribution n'est plus symétrique. Si n = 30, on trouve de la même manière

$159 < \sigma_x < 269$ mm

quand n devient assez grand la distribution devient symétrique (et Gaussienne) et l'intervalle de confiance à 95 % est défini approximativement par :

$$\sigma_x \pm 2. \frac{\sigma x}{\sqrt{2n}}$$

pour n = 30 on trouve, par cette méthode $148 < \sigma_x < 252$

4.3. INTERVALLE DE CONFIANCE D'UN QUANTILE D'UNE REPARTITION GAUSSIENNE

Soit la valeur x_F de fréquence de non dépassement F :

$x_F = m_x + u_F.\sigma_x$ (u_F = variable réduite = $\frac{x_F - m_x}{\sigma_x}$)

pour calculer son intervalle de confiance on admet que les variances de m_x et σ_x sont indépendantes.

$$var\ (x_F) = var\ (m_x) + u_F^2\ var\ (\sigma_x)$$

on utilise l'expression de l'écart-type d'un écart-type ;
on obtient ainsi

$$var\ (x_F) = \frac{\sigma_x^2}{n} + u_F^2\ \frac{\sigma_x^2}{2n} \quad \text{soit}$$

$$\boxed{\sigma_{x_F} = \frac{\sigma_x}{\sqrt{n}} \sqrt{1 + \frac{u_F^2}{2}}}$$

x_F suit approximativement une loi normale.

Application :

$\left.\begin{array}{l} m = 800 \text{ mm} \\ \sigma = 200 \text{ mm} \\ n = 30 \text{ années} \end{array}\right\}$ quel est l'intervalle de confiance de la valeur décennale ?

$F = 0,9$. En utilisant une table de Gauss (figure 4), on obtient : $u_F = 1.28$ d'où $x_F = 800 + 1.28 \times 200 = 1056$ mm

$$\sigma_{xF} = \frac{200}{\sqrt{30}} \sqrt{1 + \frac{1.28^2}{2}} = 49.25 \text{ mm}$$

pour un seuil de confiance de 95 % - u = 1.96

soit :

$$959 < x_{90\%} < 1153 \qquad \text{à 95 \%}$$

4.4. INTERVALLE DE CONFIANCE D'UN COEFFICIENT DE CORRELATION

Soit n couples d'observations (x,y) INDEPENDANTES et provenant de populations suivant des distributions à peu près Gaussiennes. On a calculé un coefficient de correlation r.

Quel est son intervalle de confiance ?

La figure 7 permet de répondre immédiatement à cette question sans aucun calcul.

Application 1 :

$\left.\begin{array}{l} n = 20 \\ r = 0.70 \end{array}\right\}$ on trouve sur l'abaque l'intervalle de confiance à 95 % : [0.37, 0.86]

Application 2 : dans le calcul de régression traité précédemment on a trouvé :

$r = -0.78$ pour $n=6$

on déduit donc de l'abaque :

$-0.96 \leqslant r \leqslant +0.06$ (à 95 %)

l'intervalle de confiance à 95 % encadre la valeur 0. On ne peut donc pas affirmer (à 95 %) que le coefficient de correlation des populations dont est tiré l'échantillon est différent de 0.

La figure 8 donne, en fonction du nombre $\nu = n-2$ de degrés de liberté la valeur minimale (de la valeur absolue) de r pour qu'on puisse considérer qu'il est significativement différent de 0 (à 95 %).

Ex. : Pour $n-2 = 4$ on trouve $r = 0.81$

Application 3 : A titre d'exemple on a construit, à l'aide de nombres au hasard, 5 échantillons de 3 couples (x,y) tirés d'une population de *moyenne nulle* d'écart-type = 1 et de coefficient de correlation nul entre x et y.

Les résultats sont les suivants :

	1	2	3	4	5
m_x	-0.500	-0.023	-0.170	0.463	-0.740
m_y	-0.820	-0.307	-0.363	0.460	-0.077
σ_x	0.840	1.075	0.928	0.705	1.280
σ_y	0.649	0.685	0.966	0.234	1.339
r	-0.272	0.121	-0.708	-0.291	0.806
a	-0.210	0.077	-0.737	-0.096	0.843
b	-0.925	-0.305	-0.238	0.504	0.547

On voit que sur les 5 coefficients de correlation on obtient, par le jeu du hasard :

$r_3 = -0.708$

et

$r_5 = 0.806$

alors qu'en fait le coefficient de correlation est nul, de même les écarts-types s'échelonnent de 0.234 à 1.339 alors que l'écart-type de la population est égale à +1.

4.5. *INTERVALLE DE CONFIANCE DES COEFFICIENTS DE REGRESSION*

Après un calcul de régression on a trouvé des coefficients a et b. Quel est l'intervalle de confiance de ces coefficients ? En particulier a et b sont ils significativement différents de 0.

Pour calculer ces intervalles de confiance, il faut que les hypothèses suivantes soient vérifiées :

- Les écarts doivent être indépendants.
- Les écarts doivent avoir une distribution à peu près Gaussienne ; leur écart-type doit être indépendant de x.

L'écart-type des coefficients est alors :

$$\sigma_a = \sqrt{\frac{1-r^2}{n-2}} \quad \frac{\sigma_y}{\sigma_x}$$

$$\sigma_b = \sigma_y \sqrt{\frac{1-r^2}{n-2}} \cdot \sqrt{1+m_x^2 / \sigma_x^2}$$

$$\left(\sigma_b = \sigma_y \sqrt{\frac{1-r^2}{n-2}} \quad \text{si x est centré}\right)$$

On voit que les écarts-types sont d'autant plus petits, et les coefficients sont donc d'autant mieux déterminés, que :

r est grand (meilleure relation linéaire)

n est grand (grand nombre d'observations)

σ_x est grand (x est plus variable)

Les coefficients a et b suivent une distribution de Student à n-2 degrés de liberté de moyenne a (ou b) et d'écart-type σ_a (ou σ_b).

Méthode pratique :

On utilise la table de Student (figure 5) et on lit à la ligne $\nu = n-2$ la valeur de t correspondant à la précision choisie (généralement 95 % parfois 80 %). Les intervalles de confiance sont alors :

$$a \pm t.\sigma_a$$

$$b \pm t.\sigma_b$$

rappelons que dès que $n \geqslant 15$ on a : $t_{80\ \%} = 1.3$ et $t_{95\ \%} = 2.0$

remarque : Quand n est assez grand (15 ou 20) on voit qu'on a approximativement :

$$\begin{cases} \sigma_a = \sqrt{\dfrac{1-r^2}{n}} \cdot \dfrac{\sigma_y}{\sigma_x} \simeq \dfrac{\sigma_\varepsilon}{\sqrt{n}\,\sigma_x} \\ \sigma_b = \sqrt{\dfrac{1-r^2}{n}}\,\sigma_y \simeq \dfrac{\sigma_\varepsilon}{\sqrt{n}} \end{cases} \qquad \text{si x est centré}$$

la distribution de Student est alors approximativement Gaussienne.

Application

Dans l'exemple traité précédemment :(§ 3.2.) :

$r = -0.78$

$\sigma_x = 1.38$

$\sigma_y = 2.95$

$a = -1.66$

$b = 10.48$

$n = 6$

$m_x = 2.50$

On calcule :

$$\sigma_a = \sqrt{\frac{1-0.78^2}{4}} \cdot \frac{2.95}{1.38} = 0.67$$

$$\sigma_b = 2.95 \sqrt{\frac{1 - 0.78^2}{4}} \sqrt{1 + 2.50^2 / 1.38^2} = 1.91$$

la table de Student donne, pour $\nu = 6-2 = 4$ degrés de liberté, et pour un intervalle de confiance à 95 % : $t = 2.78$ d'où

$$-3.52 < a < 0.20$$
$$5.17 < b < 15.79$$

N.B. $t_{régression} = -2.48$

on voit ainsi que a n'est pas significativement différent de 0 au seuil de 95 % car l'intervalle encadre 0.

4.6. *INTERVALLE DE CONFIANCE D'UNE PREVISION*

Soit un échantillon de n couples de valeurs (x,y) ; on a ajusté sur ces valeurs une relation linéaire $y = ax+b$. Soit une nouvelle valeur x_p pour laquelle on calcule y_p. Quel est l'intervalle de confiance sur cette valeur calculée y_p ?

La valeur, y_p suit une distribution de Student à n-2 degrés de liberté de moyenne y_p et l'écart-type,

$$\sigma_{yp} = \sigma_\varepsilon \sqrt{1 + \frac{1}{n} + \frac{u_p^2}{(n-1)}} \qquad \text{(approximativement)}$$

$$\text{avec } \sigma_\varepsilon = \sqrt{(1-r^2)\ \frac{n-1}{n-2}}\ \sigma_y \qquad \text{(écart type de l'écart)}$$

$$\text{et } u_p = \frac{x_p - m_x}{\sigma_x} \qquad \text{(variable réduite)}$$

On voit que dès que n est suffisamment grand (20 ou 30 valeurs) et pour des valeurs raisonnables de u_p (de -2 à +2) l'écart-type se réduit approximativement à

$$\sigma_{yp} = \sigma_\varepsilon = \sqrt{1 - r^2}\ \sigma_y$$

Application :

avec les données de l'exemple précédent :

$r = -0.78$ $\quad a = -1.66$ $\quad t = 2.78$

$\sigma_x = 1.38$ $\quad b = 10.48$

$\sigma_y = 2.95$ $\quad n = 6$

$m_x = 2.50$

soit $x_p = 1.5 \Longrightarrow y_p = -1.66 \times 1.5 + 10.48 = 7.99$

$$\begin{cases} \sigma_\varepsilon = 2.06 \\ u_p = \dfrac{1.5-2.5}{1.38} = 0.72 \end{cases}$$

d'où $\sigma_{yp} = 2.06 \sqrt{1 + \frac{1}{6} + \frac{0.72^2}{5}} = 2.32$

d'où y_p est dans l'intervalle $7.99 \pm 2.78 \times 2.32$

soit $1.54 < y_p < 14.44$

Le même calcul, avec les mêmes valeurs, mais pour n = 30 donnerait : t = 2.05 d'où :

$4.04 < y_p < 11.94$ c'est à dire un intervalle beaucoup plus étroit.

4.7. COMPARAISON DE 2 MOYENNES

Soit deux échantillons de valeurs INDEPENDANTES provenant de distributions plus ou moins Gaussiennes dont les caractéristiques sont les suivantes :

	échantillon 1	échantillon 2
nombre de valeurs	n_1	n_2
moyenne	m_1	m_2
écart-type	σ_1	σ_2

On suppose que ces échantillons sont tirés de populations qui ont même variance. Peut on admettre que leurs MOYENNES sont égales ?

N.B. Le test décrit dans le paragraphe 4.8. permet de vérifier si on peut accepter l'hypothèse que les 2 populations ont même variance.

Application en hydrologie :

- On a modifié l'emplacement d'un pluviomètre à une certaine date. On a n_1 années d'enregistrement avant et n_2 après la date de changement. Les moyennes m_1 et m_2 des pluies annuelles des 2 périodes peuvent-elles être considérées comme égales ? c'est à dire : y-a-t'il eu une influence systématique.

- Idem avec une station de jaugeage.

- Dans une région, le débit annuel par unité surface a une valeur moyenne m_1. Un autre cours d'eau indépendant observé pendant n_2 années a une moyenne m_2.

Peut-on considérer que ce nouveau cours d'eau a même débit par unité de surface ?

Principe du test :

On calcule $d = m_1 - m_2$. Si les moyennes sont égales cette DIFFERENCE suit une distribution de Student à n_1+n_2-2 degrés de liberté, de moyenne 0 et d'écart-type :

$$\sigma_d = \sigma \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

avec $\sigma = \sqrt{\frac{(n_1-1)\sigma_1^2 + (n_2-1)\sigma_2^2}{(n_1-1) + (n_2-1)}}$ = écart type calculé sur toutes les observations

On calcule donc son intervalle de confiance à l'aide d'une table de Student.

Méthode pratique :

On calcule la différence des moyennes $d = |m_1 - m_2|$

On calcule σ puis σ_d. On lit dans une table de Student la valeur t correspondant au seuil fixé (par exemple 95 %) pour un nombre de degrés de liberté égal à $n_1 + n_2 - 2$.

On compare alors la différence d (en valeur absolue !) à $t.\sigma_d$. Si la différence est supérieure, alors les 2 échantillons ont une moyenne significativement différente de 0.

Application

Soit deux périodes d'observation d'un pluviomètre avant et après changement d'emplacement (le pluviomètre est transporté à une altitude un peu plus élevée) peut-on attribuer la différence de moyenne au changement d'emplacement ?

Les caractéristiques sont les suivantes :

Avant	Après
n_1 = 12 ans	n_2 = 7 ans
m_1 = 800 mm	m_2 = 1000 mm
σ_1 = 200 mm	σ_2 = 250 mm

On vérifie bien que les pluies annuelles :

- sont indépendantes,
- suivent une distribution à peu près Gaussienne.

on calcule σ = 218.95 mm d'où σ_d = 104.13 mm.Pour 12 + 7 - 2 = 17 degrés de liberté et au seuil de 95 % la table de Student donne t = 2.11(figure 5).

On calcule alors t . σ_d = 2.11 x 104.13 = 219.72 mm .

La différence $m_2 - m_1$ = 200 mm est inférieure à 219.72 mm , on ne peut donc pas affirmer que les 2 moyennes sont significativement différentes.

Du point de vue purement statistique on ne peut donc pas considérer cette différence comme anormale vu la taille des échantillons cependant ... Si la différence de pluie survient justement à partir de la date de déplacement du pluviomètre, on peut quand même se poser des questions.

Si on effectue le même calcul pour n_1 = 20 ans et n_2 = 10 ans, on trouve σ_d = 87.17 mm et t = 2.05 soit t . σ_d = 172.55 mm . Les moyennes sont alors significativement différentes au seuil de confiance de 95 %.

4.8. COMPARAISON DE 2 VARIANCES

Soit deux échantillons INDEPENDANTS de n_1 et n_2 valeurs indépendantes provenant de populations suivant des distributions à peu près Gaussiennes et dont les écarts-types sont σ_1 et σ_2 .(pas forcément mêmes moyennes).

Peut-on considérer que ces écarts-types sont significativement différents (à un seuil de confiance donné) ?

On montre que le rapport R^2 des carrés des écarts-types suit une distribution de FISCHER - SNEDECOR (figure 9 et 10)ou distribution F à $n_1 - 1$ et $n_2 - 1$ degrés de liberté.

Méthode pratique

On calcule $R^2 = \dfrac{\sigma_1^2}{\sigma_2^2}$

On calcule l'intervalle de confiance (à 95 % par exemple) de R^2 en utilisant la figure 9.

$$\boxed{\frac{1}{F_{\nu 2,\nu 1}} < R^2 < F_{\nu 1,\nu 2}} \quad \text{avec} \begin{cases} \nu 1 = n_1 - 1 \\ \nu 2 = n_2 - 1 \end{cases}$$

Il convient de bien faire attention à l'ordre des nombre de degrés de liberté car :

$$F_{\nu 1,\nu 2} \neq F_{\nu 2,\nu 1}$$

Application

Soit deux pluviomètres *INDEPENDANTS* (c'est à dire suffisamment éloignés pour présenter seulement un faible coefficient de corrélation). On observe les valeurs suivantes :

n_1 = 12 ans $\qquad$ n_2 = 7 ans

σ_1 = 200 mm $\qquad$ σ_2 = 250 mm

Les valeurs limites admissibles du rapport R^2 si les variances sont égales sont :

$$\left(\frac{1}{F_{6,11}} , F_{11,6} \right) \text{ soit, } \left(\frac{1}{3,88} , 5.42 \right) = (0.26 , 5.42)$$

Or le rapport R^2 est égal à $(\frac{200}{250})^2 = 0.64$

Il est compris entre les 2 valeurs extrêmes.

On ne peut donc pas affirmer les 2 variances (donc les 2 écarts-types) sont différentes au seuil de 95 %. Si $n_1 = 30$ et $n_2 = 20$, on trouve :

$$0.45 < R^2 > 2.39$$

on voit donc que la variance varie considérablement d'un échantillon à un autre même quand le nombre de valeurs est relativement élevé. (test utile pour autre test nécessitant égalité des variances).

4.9. COMPARAISON DE PLUSIEURS MOYENNES
(Analyse de variance)

Soit un nombre k d'échantillons comprenant chacun un nombre différent de valeurs. Pour chaque échantillon on peut calculer une moyenne m_k (et un écart-type σ_k) .

La méthode décrite ci-dessous permet de décider (à un certain pourcentage de confiance) si on peut considérer que les échantillons ont même moyenne et si des différences observées sont dues aux aléas des échantillons.

Quelles sont les applications en hydrologie ?

Cette méthode est utilisée pour des tests d'homogénéité par exemple pour des données pluviométriques, pour des débits par unité de surface, pour des données chimiques etc...

Il faut cependant être prudent et vérifier que les hypothèses suivantes sont vérifiées :

- les échantillons doivent être INDEPENDANTS . On ne pourra donc pas comparer des séries de pluie ou de débits en des points proches présentant donc un fort coefficient de corrélation ;

- chaque échantillon doit avoir une répartition à peu près Gaussienne ;

- les variances des populations doivent pouvoir être considérées comme égales.

Méthode de calcul

On compare en fait l'écart-type des moyennes à l'écart-type des observations à l'aide d'un test de FISCHER - SNEDECOR. Soit les échantillons suivants :

Echantillon	1	2	3	.....	k
Nombre de valeurs	n_1	n_2	n_3	.....	n_k
Moyenne	m_1	m_2	m_3	.....	m_k
Ecart-type	σ_1	σ_2	σ_3	.....	σ_k

$n = n_1 + n_2 + \ldots + n_k$ = nombre total d'observations

$n = \dfrac{n_1 m_1 + n_2 m_2 + \ldots + n_k m_k}{n}$ = moyenne totale

$\sigma^2 = \dfrac{(n_1-1)\,\sigma_1^2 + (n_2-1) + \ldots + (n_k-1)\,\sigma_k^2}{n-k}$ (variance totale)

$\sigma_m^2 = \dfrac{n_1\,(m_1-m)^2 + n_2\,(m_2-m)^2 + \ldots + n_k\,(m_k-m)^2}{k-1}$ (variance interne)

On calcule alors : $R^2 = \dfrac{\sigma^2 m}{\sigma^2}$

plus les moyennes sont semblables plus R^2 sera petit.

Une table de FISCHER - SNEDECOR donne la limite *supérieure* de R^2 si les échantillons proviennent de populations qui ont même moyenne.

Cette limite est $F_{\nu 1,\ \nu 2}\ (\alpha)$

avec
$\nu 1 = k - 1$
$\nu 2 = n - k$
α = niveau de confiance.

donc : si $R^2 > F_{\nu 1, \nu 2}\ (\alpha)$ les échantillons qui proviennent de populations n'ont pas des moyennes égales.

On remarque qu'on effectue le test que "d'un seul coté" car R^2 est toujours plus grand que 1 quand les échantillons proviennent de populations qui n'ont pas des moyennes égales. On utilisera donc une table à 95% (et non 97.5%).

Application

Soit les débits obtenus à l'air lift pour des forages réalisés dans 5 régions différentes. Peut-on considérer que les débits moyens sont les mêmes dans chaque région ?

Un examen des données montre qu'elles ne suivent pas une répartition Gaussienne. En prenant le logarithme (décimal) des débits on se ramène cependant à des répartitions Gaussiennes.

Les caractéristiques des données sont les suivantes :

Région	1	2	3	4	5
Moyenne (en log)	1.0	1.1	0.5	1.4	1.5
Ecart-type	0.6	0.9	0.7	1.1	1.3
Nombre d'observations	30	20	30	15	20

$n = 30 + 20 + 30 + 15 + 20 = 115$

$k = 5$

$m = 1/115 \cdot (30x1 + 20x1.1 + 30x0.5 + 15x1.4 + 20x1.5) = 1.026$

$$\sigma^2 = \frac{29 \times 0.6^2 + 19 \times 0.9^2 + 29 \times 0.7^2 + 14 \times 1.1^2 + 19 \times 1.3^2}{115 - 5} = 0.81$$

$$\sigma_m^2 = 1/(5-1) \cdot \left[30(1-1.026)^2 + 20(1.1-1.026)^2 + \ldots + 20(1.5-1.026)^2\right] = 3.755$$

d'où $R^2 = \frac{\sigma^2 m}{\sigma^2} = 4.63$

Or une table de FISCHER avec $\alpha = 5\%$ (attention pas 2.5 %) figure 10 donne $F_{4,110}$ (5%) = 2.46

R^2 étant supérieur à 2.46 , on peut affirmer (avec seulement 5% de chance de se tromper) que les débits moyens de chaque région ne sont *pas* égaux.

On remarque en effet que la région n°3 semble avoir une moyenne nettement plus faible. Si on refait le même calcul en retirant la région n°3 on trouve :

$n = 85$ $\quad \sigma^2 = 0.924$ $\quad R^2 = 1.36$

$k = 4$ $\quad \sigma^2 m = 1.26$ $\quad F_{3,80}$ (5%) = 2.73

$m = 1.21$

Ces 3 régions n'ont donc pas des moyennes significativement différentes.

CHAPITRE 5

COMPLEMENTS A LA PREMIERE EDITION

5.1 - REGRESSION DES "MOINDRES DISTANCES"

On a vu qu'un calcul de régression faisait jouer un rôle dissymétrique au variables x et y. Si on cherche à calculer y à partir de x, on n'obtiendra pas la même droite (sur le même diagramme x, y) que si on cherche à calculer x à partir de y.

En particulier, si le coefficient de corrélation r est nul on obtiendra dans le premier cas une droite horizontale et dans le deuxième cas une droite verticale.

Il existe un certain nombre de cas où l'on désire étudier la relation entre x et y et ajuster la "meilleure" droite au milieu du nuage de points. Cette droite est celle dont tous les points sont à la moindre distance. Elle minimise donc les distances (sur les perpendiculaires à cette droite) d'où son nom de "moindres distances" (on trouve parfois... on ne sait trop pourquoi le terme de "moindres rectangles"). Cette droite correspond à celle que tracerait instinctivement un dessinateur en essayant de passer au mieux au milieu des points. Elle fait jouer un rôle symétrique à x et y, mais elle n'est pas optimale pour calculer y à partir de x puisque la droite optimale est la droite de régression. En fait les 2 droites sont confondues si le coefficient de corrélation r est égal à +1 ou -1;

La droite des moindres distances s'écrit :

$$y = ax + b$$

avec

$$a = \pm \frac{\sigma_y}{\sigma_x} \qquad \text{le signe est celui de r}$$

$$b = m_y - a.m_x$$

on note immédiatement que la droite est indépendante du coefficient de corrélation r ; elle passe par le point moyen (m_x, m_y) et qu'elle se confond avec la droite de régression $r = r.\sigma_x/\sigma_y$ si r = 1 ou -1.

Si les variables x ou y sont normées ou ont un même écart type la pente a est égale à 1 (ou -1). On montre en effet que c'est la première composante principale.

La droite des moindres distance fait jouer un rôle symétrique à x et y mais on ne peut pas faire une interprétation statistique simple sur l'erreur commise.

5.2 - CORRELATION DOUBLE

Un cas particulier de corrélation avec plusieurs variables explicatives est la corrélation "double" avec deux variables explicatives x_1 et x_2 intercorrellés : on note les moyennes et les écarts-types de y/x_1 et x_2 respectivement my, m_1, m_2 et σy, σ_1, σ_2.

On note les coefficients de corrélations :

r_1 (y, x_1), r_2 (y, x_2), r (x_1, x_2) et R entre y observé et y calculé.

On écrit la relation :

$$y = a_1.x_1 + a_2.x_2 + b$$

et on obtient les coefficients suivants :

$$a_1 = \frac{\sigma_y}{\sigma_1} \cdot \frac{r_1 - r.r_2}{1 - r^2}$$

$$a_2 = \frac{\sigma_y}{\sigma_2} \cdot \frac{r_2 - r.r_1}{1 - r^2}$$

$$b = m_y - a_1.m_1 - a_2.m_2$$

$$R^2 = \frac{r_1^2 - 2.r.r_1.r_2 + r_2^2}{1 - r^2}$$

Il est ainsi possible de calculer facilement tous les termes a_1, a_2, b et R^2 en effectuant 3 calculs de coefficients de corrélation simple. Il convient de remarquer que a_1 et a_2 sont différents des coefficients A_1 et A_2 qu'on obtiendrait par régression simple respectivement avec x_1 et x_2. Ces coefficients A_1 et A_2, égaux à $r_1\sigma_y/\sigma_1$ et $r_2\sigma_y/\sigma_2$ seraient identique à a_1 et a_2 seulement si les variables x_1 et x_2 étaient indépendantes (r = o).

Les coefficients de corrélation partielle sont respectivement :

$$r_{p1} = \frac{r_1 - r \cdot r_2}{\sqrt{1-r^2}\ \sqrt{1-r_2^2}}$$

$$r_{p2} = \frac{r_2 - r \cdot r_1}{\sqrt{1-r^2}\ \sqrt{1-r_1^2}}$$

Ce sont ces coefficients de corrélation partielle qui montrent les relations réelles entre variables.

Exemple d'application 1 :

Une enquête (réalisée dans les années 60) entre les variables suivantes :

y = fréquentation des salles de cinéma,

x_1 = possession d'un téléviseur (encore assez rare en 1960) aurait donné le résultat suivant :

$r_1 (y, x_1) = - 0.40$

une conclusion hâtive avait conclu que la possession d'un téléviseur dissuadait d'aller au cinéma car le coefficient de corrélation r_1 est négatif.

Une enquête plus approfondie a fait intervenir la variable x_2 = présence de petits enfants dans le foyer :

$r_2 (y, x_2) = -0,85$ (quand on a des petits enfants on va moins au cinéma)

et

$r (x_1, x_2) = + 0.80$ (quand on a des petits enfants on a plus souvent la télévision que quand on est célibataire ou sans enfants).

Les calculs montrent alors que :

$\begin{cases} \pi_{p_1} (y, x_1) = +0,89 \\ \pi_{p_2} (y, x_2) = 0,96 \end{cases}$ la télévision incite nettement à aller au cinéma (et réciproquement)

π_{p_1} a le signe opposé de r_1. On démontre dans cet exemple que ce sont les petits enfants qui empêchent d'aller souvent au cinéma et non la possession d'un téléviseur qui aurait même un rôle légèrement inverse.

Exemple d'application n°2 :

$\begin{cases} y = \text{débit annuel d'un cours d'eau} \\ x_1 = \text{présence de forêt de sapin} \\ x_2 = \text{altitude} \end{cases}$

$\begin{cases} r_1 (y, x_1) = + 0.70 \\ r_2 (y, x_2) = + 0.90 \\ r (x_1, x_2) = + 0.80 \quad \text{(relation forêt de sapin et altitude)} \end{cases}$

Un examen sommaire montre que $r_1 > 0$, donc la présence de forêt de sapins favoriserait les forts écoulements annuels (peut être à cause d'une plus faible évaporation...).

En fait, on trouve :

$\begin{cases} r_{p_1} = - 0,08 & \text{effet de la forêt quasi nul} \\ r_{p_2} = + 0,79 & \text{effet très marqué de l'altitude.} \end{cases}$

5.3 - DUREE DE VIE D'UN PROJET

Risque de défaillance	Durée de vie du projet 10	25	50	100 ans
1 %	910	2440	5260	9100
10 %	95	238	460	

Le tableau ci-dessus donne des temps de retour qu'il faut choisir pour une durée de vie donnée d'un projet et un risque donné :

Exemple :

Projet de 100 ans de durée de vie ; risque de défaillance admis 1 % il faut dimensionner l'ouvrage à la crue de 9100 ans.

5.4 - ETUDE D'UNE VARIABLE AU-DESSUS D'UN SEUIL

C'est l'étude d'une loi "tronquée".

On a :

P_o = probabilité de ne pas dépasser le seuil

On étudie alors les valeurs au-dessus du seuil (qui se produisent dans (1 - P_o) cas. Ces valeurs x suivent une loi $P_1(x)$ et sont caractérisées par une moyenne n et un écart-type s. On a alors :

$$P(x) = P_o + (1 - P_o).\ P_1\ (x)$$

Exemple :

* 25 % de forages à débit nul

* quand le débit n'est pas nul, la moyenne et l'écart type du débit (en logarithme décimal) sont :

m = 0.3

δ = 1

loi normale

On en déduit :

x	u	$P_1(x)$	P(x)
- 1.66	- 1.96	2.5 %	27 %
- 0.98	- 1.28	10 %	32.5 %
0.3	0	50 %	62.5 %
1.58	1.28	90 %	92.5 %
2.26	1.96	97.5 %	98.1 %

5.5 - COMPOSITION DES 2 LOIS DE PROBABILITE

Il arrive qu'on soit obligé de séparer des évènements qui suivent 2 lois de probabilité différentes :

Exemple :

* Loi de distribution des crues d'été : F été

* Loi de distribution des crues de printemps : F prin

A un évènement x on cherche à associer la fréquence de non dépassement F, connaissant F été et F prin.

On a : F été. F prin de ne pas dépasser x ni au printemps ni en été.

d'où $\boxed{F = F_{été} \cdot F_{prin}}$

On a F été. (1 - F prin) de ne pas dépasser l'été mais de dépasser le printemps.

(1 - F été). F prin de dépasser l'été mais pas le printemps.

(1 - F été) (1 - F prin) de dépasser l'été et le printemps.

Le temps de retour associé à F est déterminé par :

$$F = (1 - 1/T_{été})(1 - 1/T_{prin}) \simeq 1 - \frac{1}{T_{été}} - \frac{1}{T_{prin}}$$ si les 2 temps de retour sont grands.

$$F = 1 - (T_{été} + T_{prin}) / (T_{été} \cdot T_{prin})$$

$$T = 1/(1-F) = 1/\left[1 - 1 + (T_{été} + T_{prin})/(T_{été} \cdot T_{prin})\right]$$

$$\boxed{T \simeq \frac{T_{été} \cdot T_{prin}}{T_{été} + T_{prin}}}$$

Exemple : T été = 20 ans
T prin = 50 ans
} $T \approx 14$ ans

5.6. INTERVALLE DE CONFIANCE DU RAPPORT DE DEUX VARIANCES

On sait que si on dispose de 2 échantillons indépendants d'une même population de variance :

	Echantillon 1	Echantillon 2
Nbre d'observations	n_1	n_2
variance	$\sigma^2_{x_1}$	$\sigma^2_{x_2}$

on peut calculer l'intervalle de confiance du rapport R^2 des 2 estimations de σ^2

$$\frac{1}{F_{p\,\nu_2,\nu_1}} < \frac{\sigma^2_{x_1}}{\sigma^2_{x_2}} < F_{p\,\nu_1,\nu_2}$$

avec

$\nu_1 = n - 1$

$\nu_2 = n - 2$

F = loi de Fischer

p = probabilité

On peut ainsi tester si σ_{x_1} et σ_{x_2} peuvent provenir de la même population (ou de populations ayant même variances).

Un autre problème est celui où l'on dispose de 2 échantillons provenant de 2 populations ayant des variances différentes σ_1^2 et σ_2^2. On cherche alors l'intervalle de confiance du rapport $R^2 = \sigma_1^2/\sigma_2^2$

On a comme précédemment :

$$\frac{1}{F_{p\,\nu_2,\nu_1}} < \frac{\sigma^2_{x_1}/\sigma_1^2}{\sigma^2_{x_2}/\sigma_2^2} < F_{p\,\nu_1,\nu_2}$$

On en déduit alors (en multipliant par $\sigma^2_{x_2} / \sigma^2_{x_1}$

$$\frac{\sigma^2_{x_2}}{\sigma^2_{x_1}} \cdot \frac{1}{F_{p\,\nu_2,\nu_1}} < \frac{\sigma_2^2}{\sigma_1^2} < \frac{\sigma^2_{x_2}}{\sigma^2_{x_1}} \cdot F_{p\,\nu_1,\nu_2}$$

soit (en inversant l'inégalité) :

$$\boxed{\frac{\sigma^2_{x_1}}{\sigma^2_{x_2}} \cdot \frac{1}{F_{p\,\nu_1,\nu_2}} < \frac{\sigma_1^2}{\sigma_2^2} < \frac{\sigma^2_{x_1}}{\sigma^2_{x_2}} \cdot F_{p\,\nu_2,\nu_1}}$$

Attention à l'ordre des paramètres (ν_1 et ν_2 qui est inversé par rapport à l'ordre précédent.

Application :

n_1 = 12 ans $\quad$ n_2 = 7 ans

σ_{x_1} = 200 ans $\quad$ σ_{x_2} = 250 mm

$$R^2 = \sigma_{x_1}^2 / \sigma_{x_2}^2 = 0.64$$

1 - on montre que si les échantillons proviennent de la même population, on peut avoir (à 95 %):

$$\frac{1}{F_{6,11}} < R^2 < F_{11,6}$$

soit

$$0.26 < R^2 < 5.42$$ ce qui est bien le cas.

2 - si on suppose que les échantillons proviennent de 2 populations différentes (on a l'impression de $\sigma_1 < \sigma_2$), on montre que l'intervalle de confiance de σ_1^2/σ_2^2 est donné par :

$$\frac{0.64}{F_{11,6}} < \sigma_1^2/\sigma_2^2 < 0.64.F_{6,11}$$

$$0.12 < \sigma_1^2/\sigma_2^2 < 2.48$$

5.7. INTERVALLE DE CONFIANCE D'UN POURCENTAGE OBSERVE

Soit un échantillon de n observations parmi lesquelles n répondent à un certain critère. On définit naturellement le pourcentage de valeurs répondant au critère ("pourcentage de succès") par :

$$p = m/n$$

On cherche l'intervalle de confiance de p c'est-à-dire l'intervalle qui a une certaine probabilité (par exemple 95 %) de contenir le pourcentage p de la population d'où est tiré l'échantillon.

Les limites inférieures et supérieures de l'intervalle de confiance sont respectivement .

$P_i < p < p_s$ à 95 %

avec p_i : UBETA (P_i, m, n-m+1)

p_s : UBETA (P_s, m, n-m+1)

avec P_i = 2.5 %, P_s = 97.5 % pour un intervalle de confiance à 95 %.

UBETA = fonction inverse de la fonction statistique BETA.

L'abaque de la figure 11 permet de déterminer immédiatement sans aucun calcul les valeurs p_i et p_s de l'intervalle de confiance à 95 %.

Par exemple l'abaque montre qu'avec 50 observations, l'intervalle de confiance à 95 % d'un pourcentage de 75 % est donné par 63 % $< p <$ 84 %

Exemple d'application :

Soient 4 échantillons de forages forés à 4 profondeurs données, pour lesquels on observe un certain pourcentage de succès défini par l'obtention d'un débit minimal donné :

Profondeur	nbre de forages	nbre de succès	Taux de succès
50	100	30	30 %
60	50	13	26 %
70	30	5	17 %
80	20	5	25 %
TOTAL	200	53	27 %

Peux-on affirmer, comme on semble le "voir" qu'il y a une décroissance du taux de succès :

On fait l'hypothèse que les forages proviennent d'une même population de taux de succès 53/200 = 27 %.

En utilisant l'abaque et en rentrant 0.27 en ordonnée on trouve :

N	Limite de taux de succès qu'on peut observer dans 95 % des cas			Taux de succès observé
100	18 %	—	35 %	30 %
50	17 %	—	38 %	26 %
30	14 %	—	42 %	17 %
20	12 %	—	45 %	25 %

On ne peut donc PAS affirmer que les taux décroissent avec la profondeur (contrairement à la première impression).

Exemple 2 :

Sondage de n = 1000 individus.

Un candidat à 40 % d'intervention de vote à la date t1 puis à 36 % d'intentions de vote à la date t_2. Peut-on affirmer qu'il y a une baisse?

On fait l'hypothèse d'une même population inchangée, d'où le taux moyen p = (40 + 36)/2 = 38 %.

L'abaque indique, en rentrant en ordonnée avec 0.38 les limites de l'intervalle de confiance à 95 % :

35 % < p < 41 %.

On ne peut PAS affirmer, au seul vu des chiffres, qu'il y a eu une modification des intentions de vote de la population.

5.8. TEST D'AJUSTEMENT DU CHI² (prononcer KI - 2)

Ce test permet de vérifier si le choix d'une fonction de distribution (probabilité cumulée) est plausible. On calcule pour celà un index dépendant des écarts entre le nombre d'observations observé n_{obs} et le nombre théorique d'observation n_{th} dans des classes choisies à priori.

$$S = \sum_{Classes} \frac{(n_{obs} - n_{th})^2}{n_{th}}$$

La limite supérieure de S est un χ^2 (chi deux) à k - p - 1 degrés de liberté, p étant le nombre de paramètres de la loi théorique, k étant le nombre de classes.

Exemple : débits annuels de l'Aveyron à Rodez pendant 45 ans ; moyenne m = 6.9 m³/s ; écart-type 2.33 m³/s ajusté sur une loi normale.

1 - On choisit par exemple 7 classes équiprobables. Les limites de ces classes correspondant à des fréquences cumulées de 100 % = 14,3 % soit 45/7 = 6.43 observations.

2 - Un papier Gauss, ou une table de Gauss, donne immédiatement les limites des 7 classes :

Limites	0	4.4	5.6	6.5	7.3	8.2	9.39	infini
nombre d'observ.	7	8	6	4	5	8	7	

3 - On calcule :

$$S = \sum \left[(n_{obs} - 6.43)^2 / 6.43 \right] = 2.1$$

4 - Dans une table de Chi-deux, pour un nombre de degrés de libertés de 7 - 2 - 1 = 4 degrés, on relève pour un seuil deconfiance de 90 % :

$$\chi^2 = 7,8$$

S est très inférieur à 7.8 donc les écarts par rapport à la théorie sont très faibles. On a donc une très bonne adequation.

N.B. Le test du χ^2 est souvent peu sensible quand on ne dispose pas de grands échantillons.

A N N E X E

TABLES ET FIGURES

Figure 1 : Distribution empirique sur papier Gauss

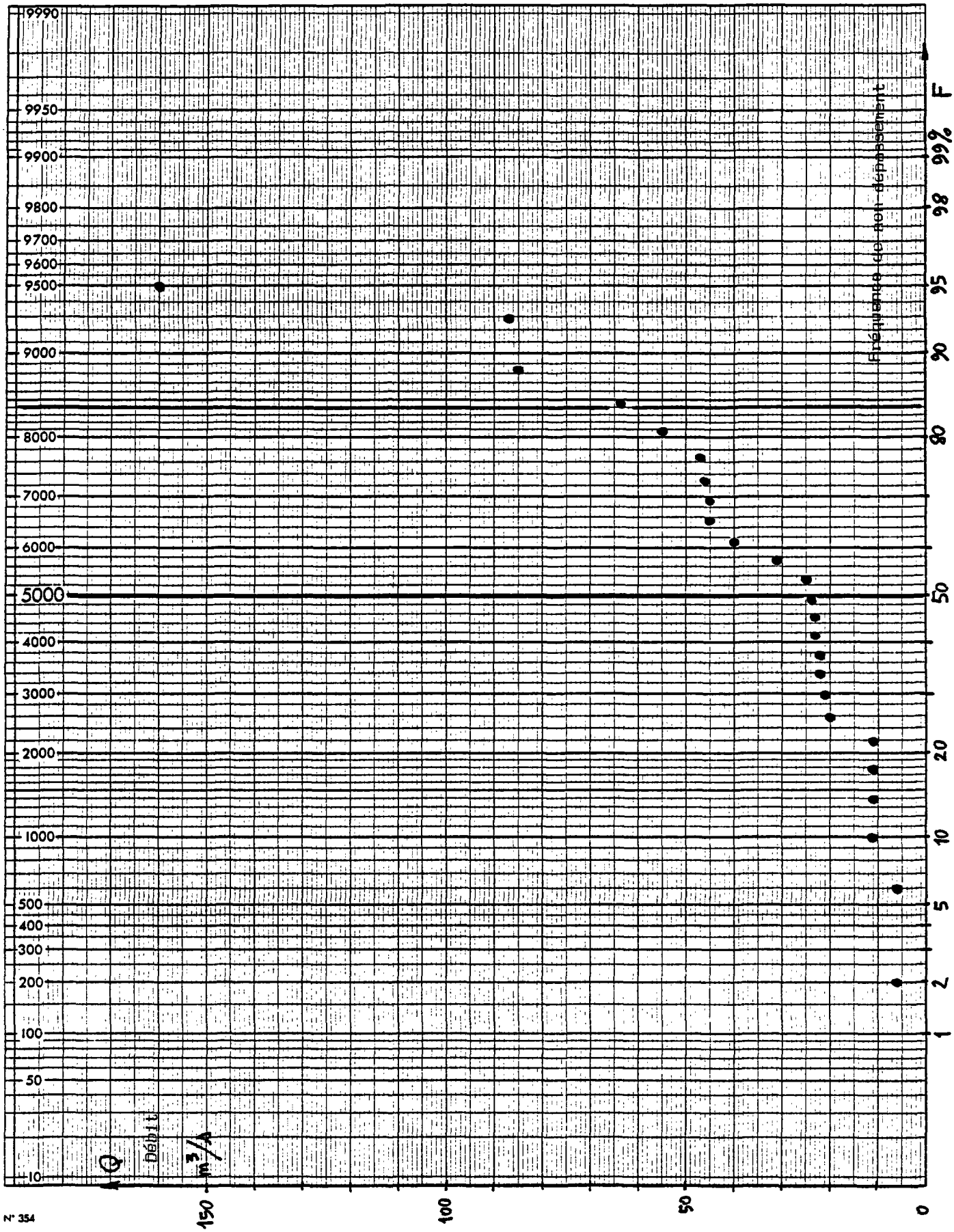
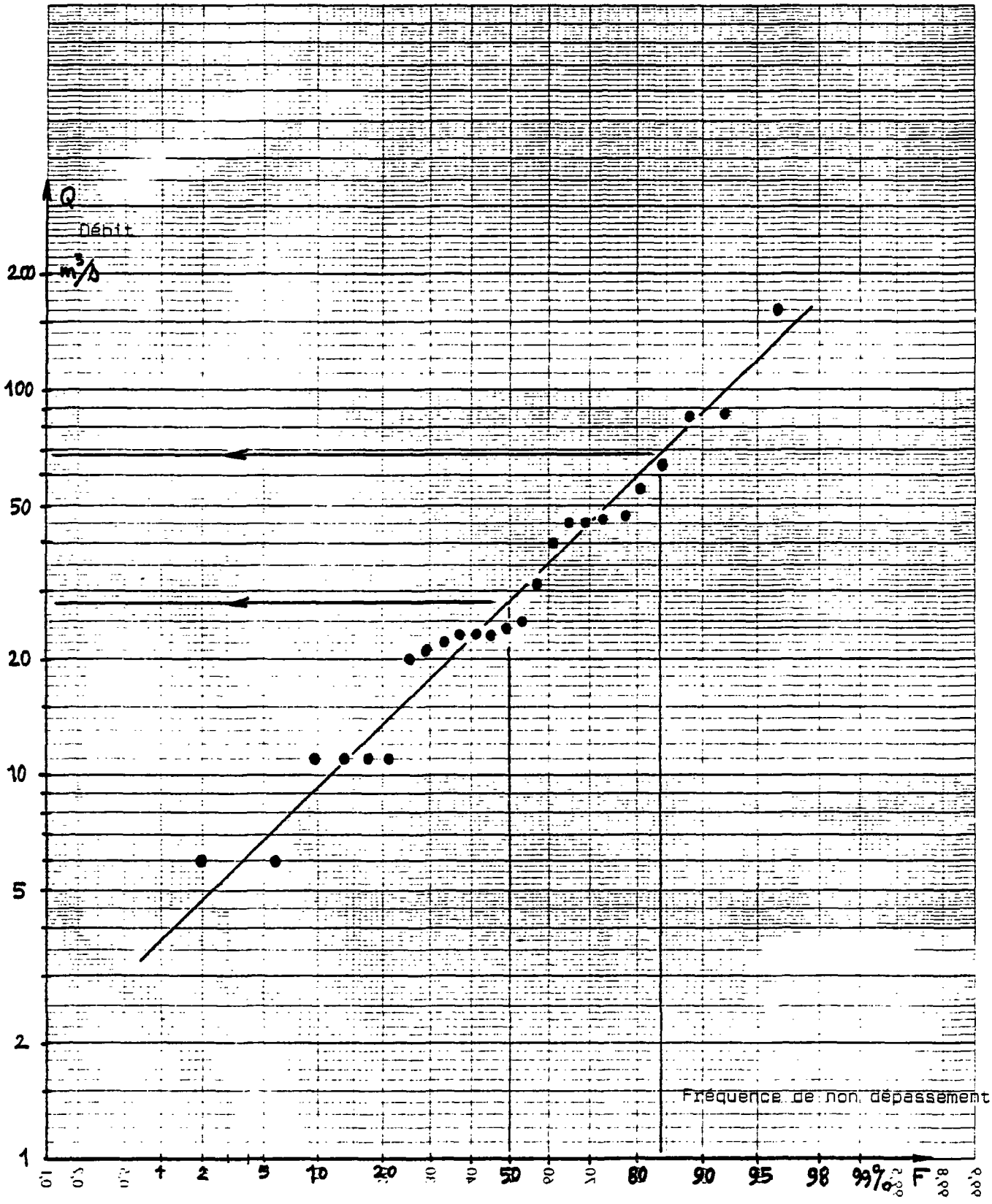


Figure 2 : Distribution Empirique sur papier Log-Normal (ou Gausso-Log)



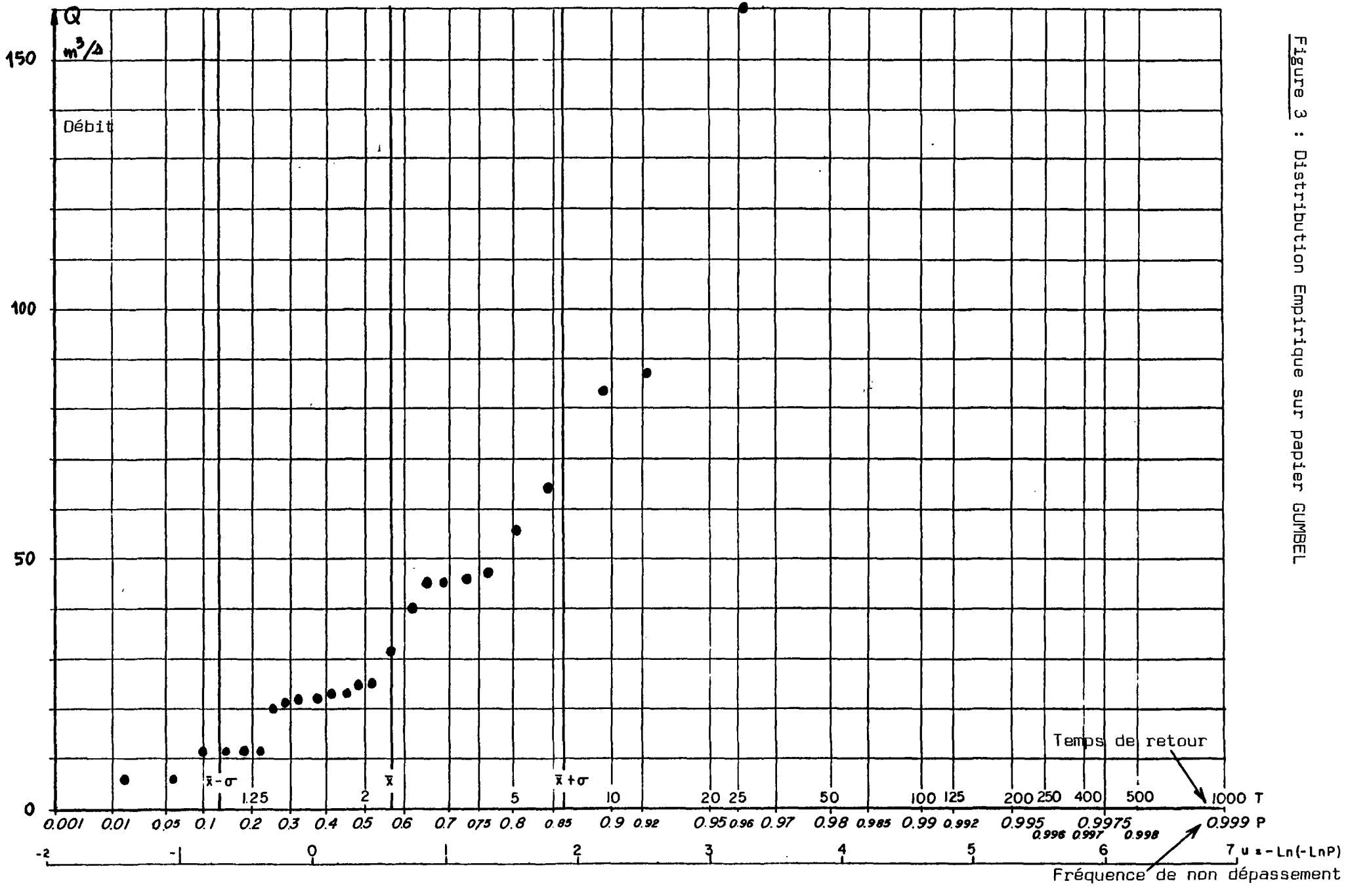
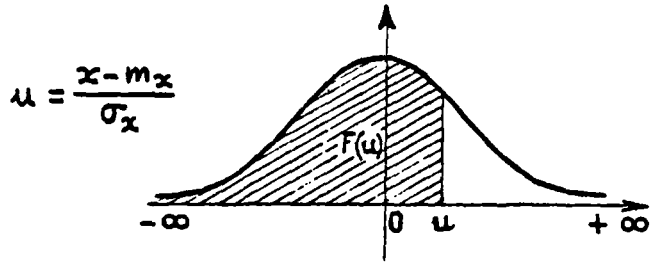


Figure 3 : Distribution Empirique sur papier GUMBEL

Figure 4

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

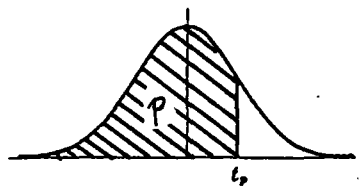
u	3,0	3,1	3,2	3,3	3,4	3,5	3,6	3,8	4,0	4,5
F(u)	0,99865	0,99904	0,99931	0,99952	0,99966	0,99976	0,999841	0,999928	0,999968	0,999997

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

Figure 5

DISTRIBUTION t de STUDENT
en fonction du nombre ν de degrés de liberté
(aire en grisé = p)



			95%	90%	80%		50%			
ν	$t_{0,995}$	$t_{0,99}$	$t_{0,975}$	$t_{0,95}$	$t_{0,90}$	$t_{0,80}$	$t_{0,75}$	$t_{0,70}$	$t_{0,60}$	$t_{0,55}$
1	63,66	31,82	12,71	6,31	3,08	1,376	1,000	0,727	0,325	0,158
2	9,92	6,96	4,30	2,92	1,89	1,061	0,816	0,617	0,289	0,142
3	5,84	4,54	3,18	2,35	1,64	0,978	0,765	0,584	0,277	0,137
4	4,60	3,75	2,78	2,13	1,53	0,941	0,741	0,569	0,271	0,134
5	4,03	3,36	2,57	2,02	1,48	0,920	0,727	0,559	0,267	0,132
6	3,71	3,14	2,45	1,94	1,44	0,906	0,718	0,553	0,265	0,131
7	3,50	3,00	2,36	1,90	1,42	0,896	0,711	0,549	0,263	0,130
8	3,36	2,90	2,31	1,86	1,40	0,889	0,706	0,546	0,262	0,130
9	3,25	2,82	2,26	1,83	1,38	0,883	0,703	0,543	0,261	0,129
10	3,17	2,76	2,23	1,81	1,37	0,879	0,700	0,542	0,260	0,129
11	3,11	2,72	2,20	1,80	1,36	0,876	0,697	0,540	0,260	0,129
12	3,06	2,68	2,18	1,78	1,36	0,873	0,695	0,539	0,259	0,128
13	3,01	2,65	2,16	1,77	1,35	0,870	0,694	0,538	0,259	0,128
14	2,98	2,62	2,14	1,76	1,34	0,868	0,692	0,537	0,258	0,128
15	2,95	2,60	2,13	1,75	1,34	0,866	0,691	0,536	0,258	0,128
16	2,92	2,58	2,12	1,75	1,34	0,865	0,690	0,535	0,258	0,128
17	2,90	2,57	2,11	1,74	1,33	0,863	0,689	0,534	0,257	0,128
18	2,88	2,55	2,10	1,73	1,33	0,862	0,688	0,534	0,257	0,127
19	2,86	2,54	2,09	1,73	1,33	0,861	0,688	0,533	0,257	0,127
20	2,84	2,53	2,09	1,72	1,32	0,860	0,687	0,533	0,257	0,127
21	2,83	2,52	2,08	1,72	1,32	0,859	0,686	0,532	0,257	0,127
22	2,82	2,51	2,07	1,72	1,32	0,858	0,686	0,532	0,256	0,127
23	2,81	2,50	2,07	1,71	1,32	0,858	0,685	0,532	0,256	0,127
24	2,80	2,49	2,06	1,71	1,32	0,857	0,685	0,531	0,256	0,127
25	2,79	2,48	2,06	1,71	1,32	0,856	0,684	0,531	0,256	0,127
26	2,78	2,48	2,06	1,71	1,32	0,856	0,684	0,531	0,256	0,127
27	2,77	2,47	2,05	1,70	1,31	0,855	0,684	0,531	0,256	0,127
28	2,76	2,47	2,05	1,70	1,31	0,855	0,683	0,530	0,256	0,127
29	2,76	2,46	2,04	1,70	1,31	0,854	0,683	0,530	0,256	0,127
30	2,75	2,46	2,04	1,70	1,31	0,854	0,683	0,530	0,256	0,127
40	2,70	2,42	2,02	1,68	1,30	0,851	0,681	0,529	0,255	0,126
60	2,66	2,39	2,00	1,67	1,30	0,848	0,679	0,527	0,254	0,126
120	2,62	2,36	1,98	1,66	1,29	0,845	0,677	0,526	0,254	0,126
∞	2,58	2,33	1,96	1,645	1,28	0,842	0,674	0,524	0,253	0,126

Figure 6

Table de distribution de χ^2 (Loi de K. Pearson)

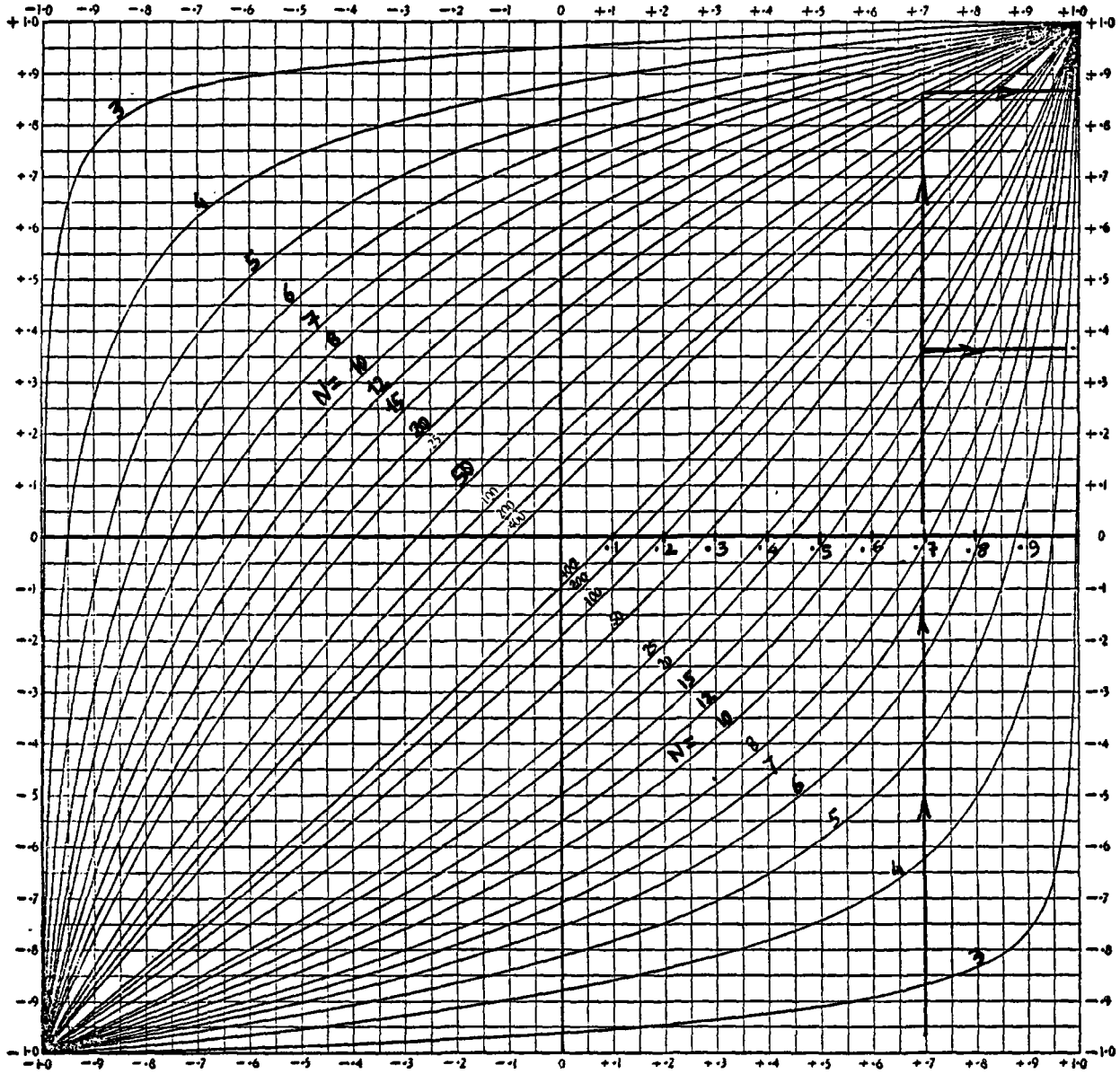
Valeurs de χ^2 ayant la probabilité d'être dépassées

P	1%	2.5%	5%	10%	90%	95%	97.5%	99%	χ^2
ν \ α	0,990	0,975	0,950	0,900	0,100	0,050	0,025	0,010	0,001
1	0,0002	0,0010	0,0039	0,0158	2,71	3,84	5,02	6,63	10,83
2	0,02	0,05	0,10	0,21	4,61	5,99	7,38	9,21	13,82
3	0,12	0,22	0,35	0,58	6,25	7,81	9,35	11,34	16,27
4	0,30	0,48	0,71	1,06	7,78	9,49	11,14	13,28	18,47
5	0,55	0,83	1,15	1,61	9,24	11,07	12,83	15,09	20,52
6	0,87	1,24	1,64	2,20	10,64	12,59	14,45	16,81	22,46
7	1,24	1,69	2,17	2,83	12,02	14,07	16,01	18,47	24,32
8	1,65	2,18	2,73	3,49	13,36	15,51	17,53	20,09	26,13
9	2,09	2,70	3,33	4,17	14,68	16,92	19,02	21,67	27,88
10	2,56	3,25	3,94	4,87	15,99	18,31	20,48	23,21	29,59
11	3,05	3,82	4,57	5,58	17,27	19,67	21,92	24,72	31,26
12	3,57	4,40	5,23	6,30	18,55	21,03	23,34	26,22	32,91
13	4,11	5,01	5,89	7,04	19,81	22,36	24,74	27,69	34,53
14	4,66	5,63	6,57	7,79	21,06	23,68	26,12	29,14	36,12
15	5,23	6,26	7,26	8,55	22,31	25,00	27,49	30,58	37,70
16	5,81	6,91	7,96	9,31	23,54	26,30	28,84	32,00	39,25
17	6,41	7,56	8,67	10,08	24,77	27,59	30,19	33,41	40,79
18	7,01	8,23	9,39	10,86	25,99	28,87	31,53	34,80	42,31
19	7,63	8,91	10,12	11,65	27,20	30,14	32,85	36,19	43,82
20	8,26	9,59	10,85	12,44	28,41	31,41	34,17	37,57	45,32
21	8,90	10,28	11,59	13,24	29,61	32,67	35,48	38,93	46,80
22	9,54	10,98	12,34	14,04	30,81	33,92	36,78	40,29	48,27
23	10,20	11,69	13,09	14,85	32,01	35,17	38,08	41,64	49,73
24	10,86	12,40	13,85	15,66	33,20	36,41	39,37	42,98	51,18
25	11,52	13,12	14,61	16,47	34,38	37,65	40,65	44,31	56,62
26	12,20	13,84	15,38	17,29	35,56	38,88	41,92	45,64	54,05
27	12,88	14,57	16,15	18,11	36,74	40,11	43,19	46,96	55,48
28	13,57	15,31	16,93	18,94	37,92	41,34	44,46	48,28	56,89
29	14,26	16,05	17,71	19,77	39,09	42,56	45,72	49,59	58,30
30	14,95	16,79	18,49	20,60	40,26	43,77	46,98	50,89	59,70

Lorsque $\nu > 30$ on peut admettre que la quantité $\sqrt{2\chi^2} - \sqrt{2\nu - 1}$ suit la loi normale réduite

si $\nu > 30$ $\qquad \chi^2 \simeq \dfrac{[u + \sqrt{2\nu - 1}]^2}{2}$

Figure 7 : Intervalle de confiance à 95 % d'un coefficient de correlation



ρ coefficient de correlation de la population

r coefficient de correlation de l'échantillon

* L'abaque donne l'intervalle de confiance à 95% [ρ_1, ρ_2] correspondant à r mesuré

* L'abaque donne également la limite supérieure et inférieure de r (au risque de de 2.5 % pour chacun) correspondant à ρ donné

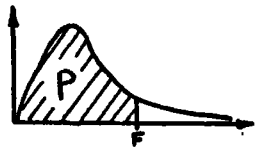
N = Nombre de valeurs de l'échantillon

Figure 8 : Valeur au-dessus de laquelle un coefficient de correlation mesuré est significativement différent de 0.

	90 %	95 %	98 %	99 %
N-2 = ν	P = 0,10	P = 0,05	P = 0,02	P = 0,01
1	0.98769	0.996917	0.9995066	0.9998766
2	0.90000	0.95000	0.98000	0.990000
3	0.8054	0.8783	0.93433	0.95873
4	0.7293	0.8114	0.8822	0.91720
5	0.6694	0.7545	0.8329	0.8745
6	0.6215	0.7067	0.7887	0.8343
7	0.5822	0.6664	0.7498	0.7977
8	0.5494	0.6319	0.7155	0.7646
9	0.5214	0.6021	0.6851	0.7348
10	0.4973	0.5760	0.6581	0.7079
11	0.4762	0.5529	0.6339	0.6835
12	0.4575	0.5324	0.6120	0.6614
13	0.4409	0.5139	0.5923	0.6411
14	0.4259	0.4973	0.5742	0.6226
15	0.4124	0.4821	0.5577	0.6055
16	0.4000	0.4683	0.5425	0.5897
17	0.3887	0.4555	0.5285	0.5751
18	0.3783	0.4438	0.5155	0.5614
19	0.3687	0.4329	0.5034	0.5487
20	0.3598	0.4227	0.4921	0.5368
25	0.3233	0.3809	0.4451	0.4869
30	0.2960	0.3494	0.4093	0.4487
35	0.2746	0.3246	0.3810	0.4182
40	0.2573	0.3044	0.3578	0.3932
45	0.2428	0.2875	0.3384	0.3721
50	0.2306	0.2732	0.3218	0.3541
60	0.2108	0.2500	0.2948	0.3248
70	0.1954	0.2319	0.2737	0.3017
80	0.1829	0.2172	0.2565	0.2830
90	0.1726	0.2050	0.2422	0.2673
100	0.1638	0.1946	0.2301	0.2540

ν = N-2

N = Nombre de valeurs de l'échantillon



$$F_{\nu_1,\nu_2}(P) = \frac{1}{F_{\nu_2,\nu_1}(1-P)}$$

P = 0,975
Q = 1 - P = 0,025

FISCHER - SCHNEDECOR

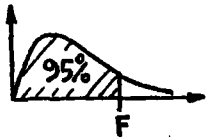
ν_2 \ ν_1	1	2	3	4	5	6	7	8	10	12	20	40	60	100	∞
1	648	800	864	900	922	937	948	957	969	977	993	1006	1010	1013	1018
2	38,5	39,0	39,2	39,2	39,3	39,3	39,4	39,4	39,4	39,4	39,4	39,5	39,5	39,5	39,5
3	17,4	16,0	15,4	15,1	14,9	14,7	14,6	14,5	14,4	14,3	14,2	14,0	14,0	14,0	13,9
4	12,2	10,6	9,98	9,60	9,36	9,20	9,07	8,98	8,84	8,75	8,56	8,41	8,36	8,32	8,26
5	10,0	8,43	7,76	7,39	7,15	6,98	6,85	6,76	6,62	6,52	6,33	6,18	6,12	6,08	6,02
6	8,81	7,26	6,60	6,23	5,99	5,82	5,70	5,60	5,46	5,37	5,17	5,01	4,96	4,92	4,85
7	8,07	6,54	5,89	5,52	5,29	5,12	4,99	4,90	4,76	4,67	4,47	4,31	4,25	4,21	4,14
8	7,57	6,06	5,42	5,05	4,82	4,65	4,53	4,43	4,30	4,20	4,00	3,84	3,78	3,74	3,67
9	7,21	5,71	5,08	4,72	4,48	4,32	4,20	4,10	3,96	3,87	3,67	3,51	3,45	3,40	3,33
10	6,94	5,46	4,83	4,47	4,24	4,07	3,95	3,85	3,72	3,62	3,42	3,26	3,20	3,15	3,08
11	6,72	5,26	4,63	4,28	4,04	3,88	3,76	3,66	3,53	3,43	3,23	3,06	3,00	2,96	2,88
12	6,55	5,10	4,47	4,12	3,89	3,73	3,61	3,51	3,37	3,28	3,07	2,91	2,85	2,80	2,72
13	6,41	4,97	4,35	4,00	3,77	3,60	3,48	3,39	3,25	3,15	2,95	2,78	2,72	2,67	2,60
14	6,30	4,86	4,24	3,89	3,66	3,50	3,38	3,29	3,15	3,05	2,84	2,67	2,61	2,56	2,49
15	6,20	4,76	4,15	3,80	3,58	3,41	3,29	3,20	3,06	2,96	2,76	2,58	2,52	2,47	2,40
16	6,12	4,69	4,08	3,73	3,50	3,34	3,22	3,12	2,99	2,89	2,68	2,51	2,45	2,40	2,32
17	6,04	4,62	4,01	3,66	3,44	3,28	3,16	3,06	2,92	2,82	2,62	2,44	2,38	2,33	2,25
18	5,98	4,56	3,95	3,61	3,38	3,22	3,10	3,01	2,87	2,77	2,56	2,38	2,32	2,27	2,19
19	5,92	4,51	3,90	3,56	3,33	3,17	3,05	2,96	2,82	2,72	2,51	2,33	2,27	2,22	2,13
20	5,87	4,46	3,86	3,51	3,29	3,13	3,01	2,91	2,77	2,68	2,46	2,29	2,22	2,17	2,09
21	5,83	4,42	3,82	3,48	3,25	3,09	2,97	2,87	2,73	2,64	2,42	2,25	2,18	2,13	2,04
22	5,79	4,38	3,78	3,44	3,22	3,05	2,93	2,84	2,70	2,60	2,39	2,21	2,14	2,09	2,00
23	5,75	4,35	3,75	3,41	3,18	3,02	2,90	2,81	2,67	2,57	2,36	2,18	2,11	2,06	1,97
24	5,72	4,32	3,72	3,38	3,15	2,99	2,87	2,78	2,64	2,54	2,33	2,15	2,08	2,02	1,94
25	5,69	4,29	3,69	3,35	3,13	2,97	2,85	2,75	2,61	2,51	2,30	2,12	2,05	2,00	1,91
26	5,66	4,27	3,67	3,33	3,10	2,94	2,82	2,73	2,59	2,49	2,28	2,09	2,03	1,97	1,88
27	5,63	4,24	3,65	3,31	3,08	2,92	2,80	2,71	2,57	2,47	2,25	2,07	2,00	1,94	1,85
28	5,61	4,22	3,63	3,29	3,06	2,90	2,78	2,69	2,55	2,45	2,23	2,05	1,98	1,92	1,83
29	5,59	4,20	3,61	3,27	3,04	2,88	2,76	2,67	2,53	2,43	2,21	2,03	1,96	1,90	1,81
30	5,57	4,18	3,59	3,25	3,03	2,87	2,75	2,65	2,51	2,41	2,20	2,01	1,94	1,88	1,79
40	5,42	4,05	3,46	3,13	2,90	2,74	2,62	2,53	2,39	2,29	2,07	1,88	1,80	1,74	1,64
60	5,29	3,93	3,34	3,01	2,79	2,63	2,51	2,41	2,27	2,17	1,94	1,74	1,67	1,60	1,48
120	5,15	3,80	3,22	2,89	2,67	2,51	2,39	2,30	2,15	2,05	1,82	1,61	1,52	1,45	1,31
∞	5,02	3,69	3,12	2,79	2,57	2,41	2,29	2,19	2,05	1,94	1,71	1,48	1,39	1,30	1,00

ν_1

ν_2

Revue de Statistique Appliquée, 1973 - Tables statistiques

Figure 9 : Table de FISCHER-SCHNEDECOR P = 97.5 % (pour intervalles de confiance des 2 côtés)



P = 0,95
Q = 1 - P = 0,05

ν_2 \ ν_1	1	2	3	4	5	6	7	8	10	12	20	40	60	100	∞
1	161	200	216	225	230	234	237	239	242	244	248	251	252	253	254
2	18,5	19,0	19,2	19,2	19,3	19,3	19,4	19,4	19,4	19,4	19,4	19,5	19,5	19,5	19,5
3	10,1	9,55	9,28	9,12	9,01	8,94	8,89	8,85	8,79	8,74	8,66	8,59	8,57	8,55	8,53
4	7,71	6,94	6,59	6,39	6,26	6,16	6,09	6,04	5,96	5,91	5,80	5,72	5,69	5,66	5,63
5	6,61	5,79	5,41	5,19	5,05	4,95	4,88	4,82	4,74	4,68	4,56	4,46	4,43	4,41	4,37
6	5,99	5,14	4,76	4,53	4,39	4,28	4,21	4,15	4,06	4,00	3,87	3,77	3,74	3,71	3,67
7	5,59	4,74	4,35	4,12	3,97	3,87	3,79	3,73	3,64	3,57	3,44	3,34	3,30	3,27	3,23
8	5,32	4,46	4,07	3,84	3,69	3,58	3,50	3,44	3,35	3,28	3,15	3,04	3,01	2,97	2,93
9	5,12	4,26	3,86	3,63	3,48	3,37	3,29	3,23	3,14	3,07	2,94	2,83	2,79	2,76	2,71
10	4,96	4,10	3,71	3,48	3,33	3,22	3,14	3,07	2,98	2,91	2,77	2,66	2,62	2,59	2,54
11	4,84	3,98	3,59	3,36	3,20	3,09	3,01	2,95	2,85	2,79	2,65	2,53	2,49	2,46	2,40
12	4,75	3,89	3,49	3,26	3,11	3,00	2,91	2,85	2,75	2,69	2,54	2,43	2,38	2,35	2,30
13	4,67	3,81	3,41	3,18	3,03	2,92	2,83	2,77	2,67	2,60	2,46	2,34	2,30	2,26	2,21
14	4,60	3,74	3,34	3,11	2,96	2,85	2,76	2,70	2,60	2,53	2,39	2,27	2,22	2,19	2,13
15	4,54	3,68	3,29	3,06	2,90	2,79	2,71	2,64	2,54	2,48	2,33	2,20	2,16	2,12	2,07
16	4,49	3,63	3,24	3,01	2,85	2,74	2,66	2,59	2,49	2,42	2,28	2,15	2,11	2,07	2,01
17	4,45	3,59	3,20	2,96	2,81	2,70	2,61	2,55	2,45	2,38	2,23	2,10	2,06	2,02	1,96
18	4,41	3,55	3,16	2,93	2,77	2,66	2,58	2,51	2,41	2,34	2,19	2,06	2,02	1,98	1,92
19	4,38	3,52	3,13	2,90	2,74	2,63	2,54	2,48	2,38	2,31	2,16	2,03	1,98	1,94	1,88
20	4,35	3,49	3,10	2,87	2,71	2,60	2,51	2,45	2,35	2,28	2,12	1,99	1,95	1,91	1,84
21	4,32	3,47	3,07	2,84	2,68	2,57	2,49	2,42	2,32	2,25	2,10	1,96	1,92	1,88	1,81
22	4,30	3,44	3,05	2,82	2,66	2,55	2,46	2,40	2,30	2,23	2,07	1,94	1,89	1,85	1,78
23	4,28	3,42	3,03	2,80	2,64	2,53	2,44	2,37	2,27	2,20	2,05	1,91	1,86	1,82	1,76
24	4,26	3,40	3,01	2,78	2,62	2,51	2,42	2,36	2,25	2,18	2,03	1,89	1,84	1,80	1,73
25	4,24	3,39	2,99	2,76	2,60	2,49	2,40	2,34	2,24	2,16	2,01	1,87	1,82	1,78	1,71
26	4,23	3,37	2,98	2,74	2,59	2,47	2,39	2,32	2,22	2,15	1,99	1,85	1,80	1,76	1,69
27	4,21	3,35	2,96	2,73	2,57	2,46	2,37	2,31	2,20	2,13	1,97	1,84	1,79	1,74	1,67
28	4,20	3,34	2,95	2,71	2,56	2,45	2,36	2,29	2,19	2,12	1,96	1,82	1,77	1,73	1,65
29	4,18	3,33	2,93	2,70	2,55	2,43	2,35	2,28	2,18	2,10	1,94	1,81	1,75	1,71	1,64
30	4,17	3,32	2,92	2,69	2,53	2,42	2,33	2,27	2,16	2,09	1,93	1,79	1,74	1,70	1,62
40	4,08	3,23	2,84	2,61	2,45	2,34	2,25	2,18	2,08	2,00	1,84	1,69	1,64	1,59	1,51
60	4,00	3,15	2,76	2,53	2,37	2,25	2,17	2,10	1,99	1,92	1,75	1,59	1,53	1,48	1,39
120	3,92	3,07	2,68	2,44	2,29	2,17	2,08	2,01	1,91	1,83	1,65	1,49	1,42	1,36	1,25
∞	3,84	3,00	2,60	2,37	2,21	2,10	2,01	1,94	1,83	1,75	1,57	1,39	1,32	1,24	1,00

ν_1

ν_2

Figure 10 : Table de FISCHER-SCHNEDECOR P = 95 % (à utiliser pour des tests d'un seul côté)

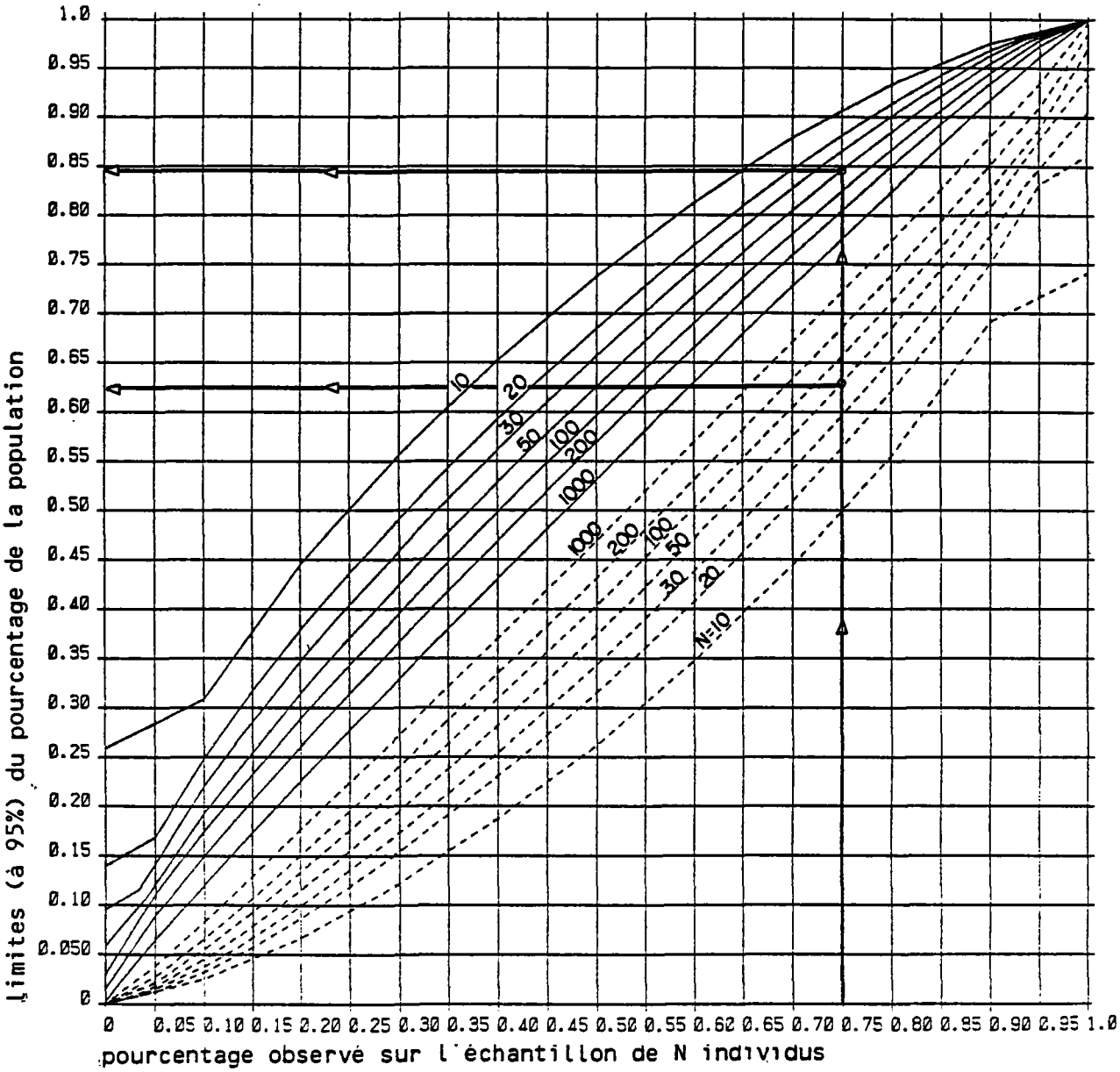


Fig.11- Intervalles de confiance à 95% d'un pourcentage ou d'une fréquence empirique