REVUE DE STATISTIQUE APPLIQUÉE

J. BERNIER
R. VERON

Sur quelques difficultés rencontrées dans l'estimation d'un débit de crue de probabilité donnée

Revue de statistique appliquée, tome 12, n° 1 (1964), p. 25-48
<http://www.numdam.org/item?id=RSA_1964__12_1_25_0>

SUR QUELQUES DIFFICULTÉS RENCONTRÉES DANS L'ESTIMATION D'UN DÉBIT DE CRUE DE PROBABILITÉ DONNÉE

J. BERNIER et R. VERON
Ingénieurs à la Division Hydrologie du Centre de Recherches
et d'essais de CHATOU E.D.F.

Les lois de probabilités connaissent diverses utilisations en hydrologie statistique.

D'une part, elles servent à décrire : C'est par la donnée de la loi de probabilité ajustée aux courbes de fréquences observées (pour des débits comme les crues, les étiages, les modules par exemple) que l'on caractérise finalement les renseignements rassemblés sur la répartition de ces différents débits.

Ainsi, après l'examen de très longues séries de modules on peut dire maintenant que, pour des fleuves comme le Rhin ou la Seine à Paris - c'est-à-dire des fleuves qui ont un bassin versant étendu -, la répartition des modules est bien représentée par une loi de Laplace-Gauss, ou loi normale.

Il semblerait aussi que pour des cours d'eau qui ont un bassin versant plus petit la loi de Galton représenterait de façon assez satisfaisante la loi de répartition des modules.

En aboutissant à ces différents énoncés, on décrit de façon commode la façon dont se répartissent les modules autour de leur moyenne générale qui sert d'élément de référence.

D'autre part, les lois de probabilités permettent d'aborder un autre domaine très important : celui de la prévision.

C'est cet aspect qui est au centre des travaux que nous avons entrepris pour l'estimation des quantiles, c'est-à-dire des débits de crue correspondant à des probabilités données.

En termes généraux, les problèmes relatifs aux quantiles se posent de la façon suivante :

Le débit X est une variable aléatoire définie par sa fonction de répartition :

$$F(X) = \text{Prob.}\ [X \leqslant x]$$

et on veut déterminer le quantile x_p tel que :

$$\text{Prob.}\ [X > x_p] = 1 - F(x_p) = p \qquad (1)$$

Nous allons tout d'abord préciser la signification que l'on peut donner à cette probabilité pour les petites valeurs de p, ce qui nous amènera à parler de la notion de durée de retour $T = 1/p$.

Nous examinerons ensuite les divers risques d'erreurs que comporte ce problème d'estimation :

1/ erreurs dûes à la plus ou moins bonne adéquation de la loi de probabilité choisie ;

2/ erreurs dûes à la nature des données expérimentales (manque de précision des mesures);

3/ erreurs d'échantillonnage. Ceci nous amenant à aborder la question des intervalles de confiance :

- d'un quantile
- de la probabilité affectée à un débit donné
- de la probabilité affectée à un débit observé.

I - INTERPRETATION DE LA PROBABILITE p

Pour une compréhension correcte du quantile x_p il importe de préciser le sens attribué à la probabilité p, surtout pour les très petites valeurs de p.

En effet, dans le cas de valeurs relativement élevées de p, 1/10 par exemple, un nombre suffisamment grand de réalisations indépendantes de la variable X peut-être obtenu dans un délai "relativement court" de quelques dizaines d'années et l'hydrologue n'aura pas de difficultés pour interprêter cette probabilité en termes de fréquences.

Faisons ici une parenthèse pour indiquer que l'indépendance des réalisations de X est essentielle : ce sont les liaisons entre débits journaliers successifs qui interdisent d'accorder un sens probabiliste aux fréquences calculées à partir de la courbe des débits classés.

Fréquence empirique et probabilité subjective

Une fois estimée la crue de probabilité p, les hydrologues l'utilisent en introduisant la notion de durée de retour :

La durée de retour T du quantile x_p étant définie par la relation

$$T = \frac{1}{p}$$

pour p = 1/10, 1/100, 1/1000, on définit ainsi :

- la crue dite décennale, valeur du débit dépassée en moyenne une fois tous les dix ans,
- la crue dite centenaire, valeur du débit dépassée en moyenne une fois tous les cent ans.
- la crue dite millénaire, valeur du débit dépassée en moyenne une fois tous les mille ans.

Le simple énoncé de ces définitions des crues, décennale, centenaire, millénaire, nous amène à préciser les deux interprétations qu'on peut donner au terme probabilité.

Si l'expression "la crue décennale représente la valeur du débit dépassée en moyenne tous les dix ans" ne surprend pas, par contre l'expression "la crue millénaire représente la valeur du débit dépassée en

moyenne tous les mille ans" paraît plus étonnante.

Dans le premier cas, nous pouvons donner à la formule un contenu auquel nous sommes habitués. Nous pouvons raisonner en termes de fréquences :

Dans un jeu de pile ou face par exemple, chacun admet rapidement que la fréquence d'apparition de "pile" ou de "face" tend vers $\frac{1}{2}$, ce que le statisticien traduit en disant que la probabilité d'apparition de "pile" ou de "face" est de $\frac{1}{2}$. On peut raisonner de cette manière dans le cas de la crue décennale. Ce n'est plus possible pour la crue millénaire. On ne peut plus penser en termes de fréquences, c'est la notion de "probabilité subjective" qui doit intervenir.

En fait, dans un grand nombre de cas on rattache, et ceci est justifié, la notion abstraite de probabilité à celle de fréquence empirique. Dans un cas comme celui de la crue millénaire, pour pouvoir raisonner en termes de fréquences, il faudrait disposer d'observations s'étalant sur une longue suite de millénaires pour lesquelles on pourrait calculer une moyenne, puis faire le raisonnement habituel.

Notons en passant que même si nos successeurs lointains disposent un jour d'une telle information, ils ne pourront, de toute façon, pas l'utiliser comme nous utilisons, dans la recherche de l'estimation de la crue décennale, les données relatives à une cinquantaine d'années. En effet, si l'on peut ne pas être affirmatif sur l'évolution du climat pendant une période assez courte, à l'échelle du millénaire cette évolution ne fait pas de doute. Ainsi du seul point de vue du climat, il est certain que les conditions d'occurrence des crues auront varié, ce qui complique le problème, car il faut tenir compte de cette évolution des conditions dans lesquelles les crues se produisent, évolution qui met en cause l'homogénéité des échantillons utilisés. Or, l'homogénéité des échantillons traités est, avec l'indépendance des observations, une des conditions essentielles d'application des méthodes les plus courantes de la statistique classique.

Ceci ne veut pas dire qu'on serait alors complètement désarmé en face d'un problème devenu insoluble. Simplement, il faudrait faire appel à des schémas beaucoup plus complexes, pour l'application desquels le problème n'est pas seulement celui de la "complexité", mais aussi et surtout celui de l'insuffisance de la quantité d'information dont dispose le statisticien, pour l'estimation d'un nombre accru de paramètres par exemple.

Il semble donc à certains hydrologues que les difficultés soient telles qu'elles remettent en cause le principe même de l'application du calcul des probabilités dans les problèmes de crues. Pour nous, il nous semble que l'on doive incriminer ici non pas le calcul des probabilités, mais le contenu concret que l'on donne communément à la notion de probabilité en la représentant par une fréquence empirique obtenue au cours d'une longue suite d'épreuves.

En fait, on peut aussi donner à la notion de probabilité un contenu subjectif et malgré l'absence d'une série d'épreuves, établir une "vraisemblance", un "certain degré de croyance rationnel" qu'un homme de bon sens accorde à des éléments "incertains" sur lesquels il ne possède qu'une information limitée imparfaite. Subjectif ne signifie pas alors arbitraire.

C'est ce qu'ont montré des statisticiens comme SAVAGE [1], en prouvant que, à condition d'admettre certains axiomes de cohérence très généraux concernant le comportement de l'homme devant l'incertitude, les probabilités subjectives se combinent naturellement d'une façon dont les théorèmes fondamentaux du calcul des probabilités rendent compte.

Ainsi pour la crue millénaire, nous ne disposons pas de l'information expérimentale qui nous permettrait de déterminer une quantité dont nous pourrions dire que sa fréquence d'apparition est effectivement 1/1000, mais les éléments dont nous disposons (valeurs observées sur une cinquantaine d'années) nous permettent néanmoins de nous fixer une valeur à laquelle nous sommes conduits à affecter un "certain degré de vraisemblance".

Comment utiliser la valeur ainsi déterminée ?

Dans l'état de développement actuel des techniques statistiques, les méthodes basées sur les probabilités subjectives ne sont pas assez développées pour qu'on puisse les utiliser fréquemment.

Dans la majorité des cas on sera donc conduit à utiliser des techniques existantes issues de la notion "fréquentiste" de la probabilité, tout en gardant présentes à l'esprit leurs insuffisances de façon à interprêter les résultats obtenus avec toute la prudence nécessaire.

LA NOTION DE DUREE DE RETOUR

L'usage répété d'une telle notion risque d'entraîner la croyance en une certaine régularité dans les apparitions successives du phénomène rare étudié.

Or, si on appelle N le nombre d'années séparant deux apparitions successives d'une valeur supérieure à x_p ; N est une variable aléatoire dont la loi s'écrit :

$$\text{Prob. } [N = n] = p\,(1 - p)^{n-1}$$

d'où :

$$\text{Prob. } [N \leqslant n] = 1 - (1 - p)^n \qquad (2)$$

si p est petit on peut remplacer cette formule par l'expression approchée suivante :

$$\text{Prob. } [N \leqslant n] = 1 - e^{-np} \qquad (3)$$

On peut voir effectivement que la valeur moyenne de N est bien égale à 1/p ; cependant des durées très inférieures sont loin d'être improbables comme le montre le tableau ci-après dans lequel figurent les valeurs de la probabilité Prob. $[N \leqslant n]$ calculées en utilisant la formule (3) :

n (années)	10	20	50	100	200	500	1 000	2 000
p = 0,01	0,095	0,181	0,393	0,632	0,865			
p = 0,001			0,049	0,095	0,181	0,393	0,632	0,865

SAVAGE : The foundations of statistics - Wiley and sons - New York 1954

Le phénomène peut être vu sous un autre angle en partant du nombre K d'observations supérieures à x_p relevées au cours d'une période fixe de N années. La variable aléatoire K obéit à la loi de POISSON :

$$\text{Prob. } [K \geqslant k] = \sum_{x=k}^{x=\infty} e^{-Np} \frac{(Np)^x}{x!}$$

Cette formule permet la construction du tableau ci-dessous dans le cas où la probabilité p est égale à 0,01.

p = 0,01 Valeurs de Prob. [K ⩾ k]

k \ N (années)	10	20	50	100
1	0,095	0,181	0,393	0,593
2	0,005	0,017	0,090	0,227
3	-	0,001	0,014	0,063
4	-	-	0,002	0,013
5	-	-	-	0,002

Ces différents résultats montrent que des évènements de probabilité très faible peuvent être observés, même de façon répétée, dans des délais assez courts.

II - CHOIX D'UNE LOI DE PROBABILITE, LES ERREURS D'ADEQUATION

Les considérations développées dans la suite sont très générales et s'appliquent à toute détermination d'un quantile. Cependant pour fixer les idées, nous nous réfèrerons uniquement à l'étude des maxima annuels des débits moyens journaliers.

La loi de probabilité des débits maximaux n'est pas exactement connue. Pratiquement on ne dispose que des observations relatives à une période de temps souvent assez courte et on doit en rechercher l'ajustement par une loi de probabilité de forme mathématique donnée. Le choix de cette loi peut-être guidé par des considérations théoriques basées, par exemple, sur l'étude du comportement asymptotique des lois des débits journaliers ou des courbes de débits classés. Nous nous bornerons ici à la recherche de la loi qui permet le meilleur ajustement aux observations, encore limiterons-nous le choix aux seules lois des valeurs extrêmes, GUMBEL ou FRECHET, dont les fonctions de répartition s'écrivent :

$$\boxed{F(x) = e^{-e^{-y}}} \qquad (5)$$

où y est une variable réduite liée au débit x par les formules :

$$y = \alpha (x - u) \quad \text{pour la loi de GUMBEL} \qquad (6)$$

$$y = \alpha (\text{Log } x - u) \quad \text{pour la loi de FRECHET} \qquad (7)$$

Le choix peut-être fondé sur la méthode classique qui consiste à comparer l'alignement, sur le papier à probabilité de GUMBEL, des points obtenus en affectant des probabilités expérimentales aux valeurs des échantillons observés. Un problème pratique se pose d'ailleurs à ce propos quant au choix de la formule à utiliser pour définir la probabilité expérimentale à attribuer à une observation. On pourra se reporter à l'analyse qui en a été faite par M. GUMBEL dans plusieurs de ses publications[1]. Nous nous bornerons à indiquer la solution généralement retenue : on préconise l'utilisation de la formule :

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

qui fait correspondre à la ième valeur d'un échantillon de n unités la probabilité :

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

Nous rappelons d'autre part, que le papier à probabilité de GUMBEL porte en abscisses l'échelle des probabilités calculées selon la formule (5) et en ordonnées l'échelle des débits x, à variation arithmétique pour la loi de GUMBEL et logarithmique pour la loi de FRECHET. Si l'échelle des débits est arithmétique la courbe de FRECHET présente une convexité tournée vers les probabilités croissantes.

Les figures I et II reproduisent deux ajustements effectués selon ce principe par M. PELLECUER, Ingénieur des Ponts & Chaussées à la 4ème Circonscription électrique à LIMOGES, pour les maximums annuels de la Corrèze à BRIVE et du Cher à TEILLET ARGENTY. Nous remercions vivement M. PELLECUER de nous avoir autorisés à puiser nos exemples dans son étude récente "Essai de détermination des durées de retour des crues d'Octobre 1960 dans le Massif Central".

On peut déterminer le choix d'un ajustement d'après l'examen de tels graphiques mais il n'est pas douteux que cette méthode empirique est très subjective ; un alignement observé peut être tout à fait fortuit. Il serait souhaitable de pouvoir utiliser un critère objectif comme le test de χ^2 de PEARSON par exemple. On se heurte alors à d'autres difficultés.

Rappelons le principe du test de PEARSON :

Le champ de variation du débit x est découpé en un certain nombre de classes (soit k ce nombre) et on confronte les fréquences absolues observées O_i des classes avec les valeurs théoriques T_i obtenues à partir de la loi de probabilité ajustée aux observations :

$$\chi^2 = \sum_{i=1}^{i=k} \frac{(O_i - T_i)^2}{T_i}$$

La loi de probabilité du χ^2 fait intervenir le nombre de degrés de liberté ν calculé en retranchant de k - 1 le nombre des paramètres dé-

---

(1) E.J. GUMBEL :"Statistics of extremes" - Colombia University Press N - Y - 1958 p. 29 à 34

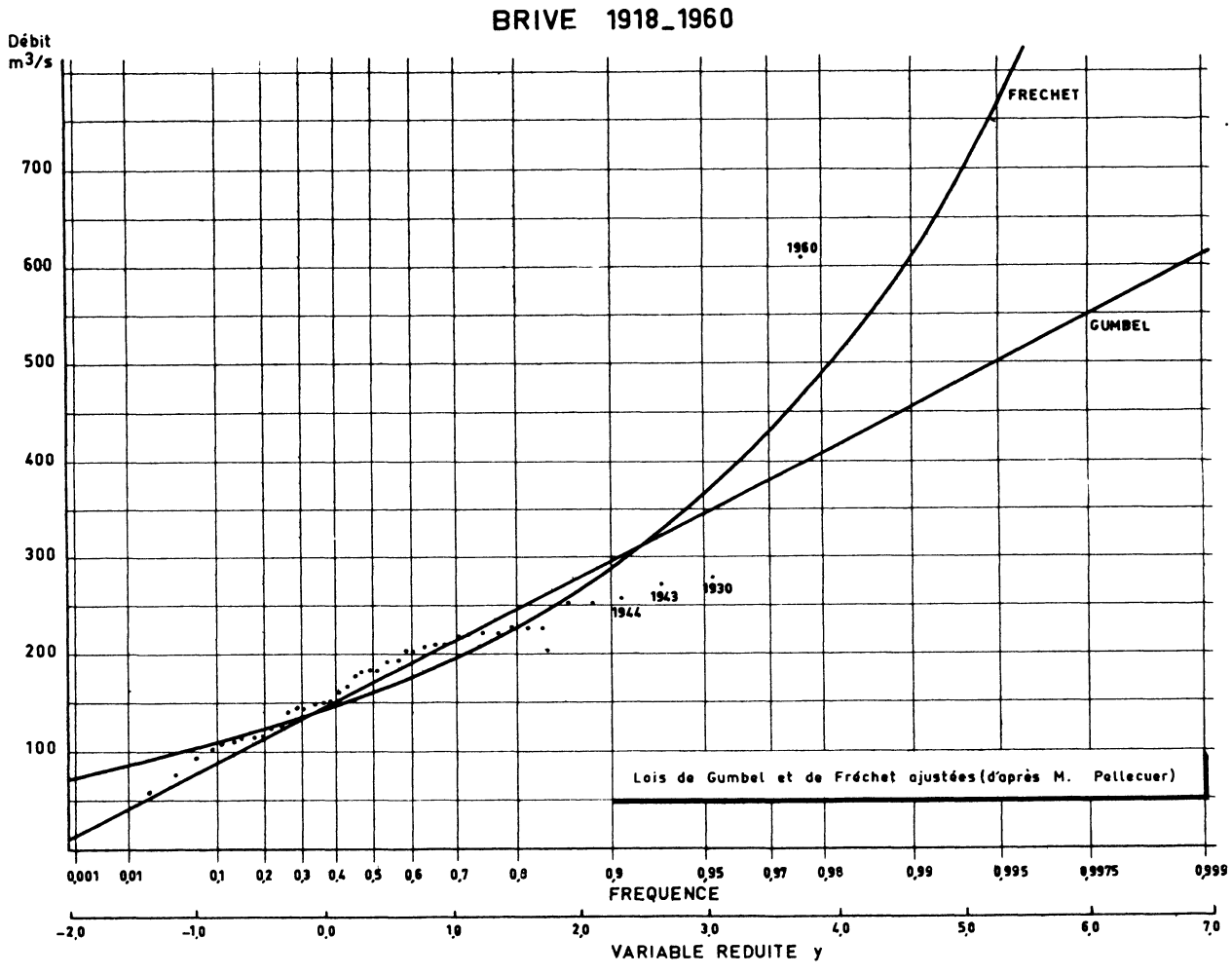

BRIVE 1918_1960
Débit m3/s
700
600
500
400
300
200
100
FRECHET
GUMBEL
1960
1944
1943
1930
Lois de Gumbel et de Fréchet ajustées (d'après M. Pollecuer)
0,001 0,01 0,1 0,2 0,3 0,4 0,5 0,6 0,7 0,8 0,9 0,95 0,97 0,98 0,99 0,995 0,9975 0,999
FREQUENCE
-2,0 -1,0 0,0 1,0 2,0 3,0 4,0 5,0 6,0 7,0
VARIABLE REDUITE y

Fig. 1

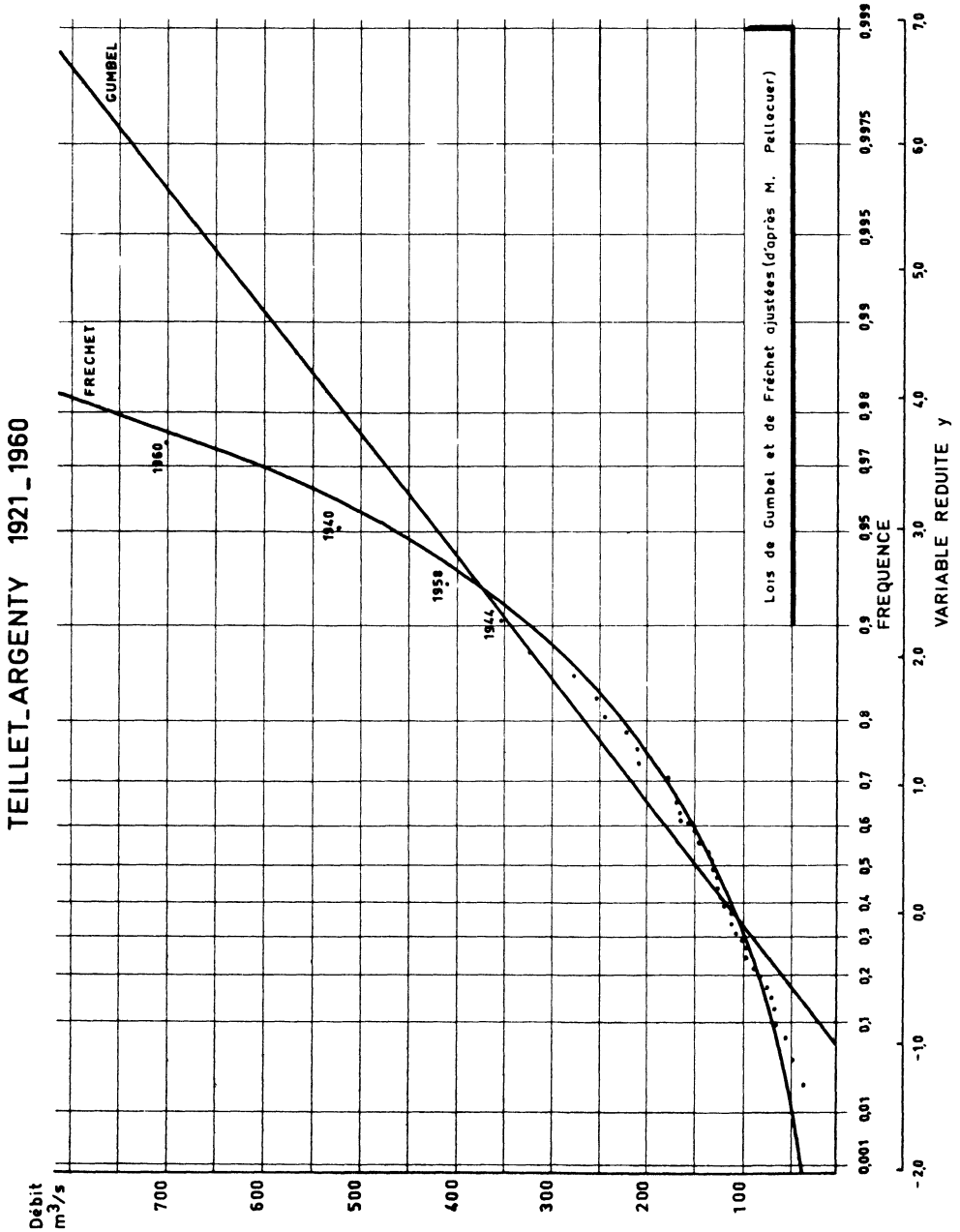

TEILLET_ARGENTY 1921_1960
Débit m³/s
GUMBEL
FRECHET
1960
1940
1958
1944
Lois de Gumbel et de Fréchet ajustées (d'après M. Pellegrin)
FREQUENCE
VARIABLE REDUITE y

Fig. 2

finissant la loi ajustée et qui sont estimés à partir des observations.

Pour les cas de BRIVE et de TEILLET-ARGENTY les calculs du χ^2 sont rassemblés dans les deux tableaux ci-après :

1) BRIVE (1918 - 1960)

Classes	O_i	T_i (GUMBEL)	$\frac{(O_i - T_i)^2}{T_i}$	T_i (FRECHET)	$\frac{(O_i - T_i)^2}{T_i}$
$x < 100$	3	6.1	1.575	2.5	0.100
$100 \leqslant x < 140$	9	8.6	0.019	13.3	1.390
$140 \leqslant x < 190$	11	11.1	0.001	13.4	0.430
$190 \leqslant x < 230$	14	6.7	7.956	5.5	13.136
$x \geqslant 230$	6	10.5	1.929	8.3	0.637
TOTAL	43	43.0	11.480	43.0	15.693

2) TEILET-ARGENTY (1921 - 1960)

Classes	O_i	T_i (GUMBEL)	$\frac{(O_i - T_i)^2}{T_i}$	T_i (FRECHET)	$\frac{(O_i - T_i)^2}{T_i}$
$x < 100$	11	13.0	0.308	13.8	0.568
$100 \leqslant x < 150$	12	7.0	3.571	11.0	0.091
$150 \leqslant x < 200$	6	6.0	-	5.7	0.016
$200 \leqslant x < 350$	7	10.1	0.951	6.1	0.133
$x \geqslant 350$	4	3.9	0.003	3.4	0.106
TOTAL	40	40.0	4.833	40.0	0.914

Dans chaque cas le nombre de degrés de liberté est égal à 2. Les tables de PEARSON donnent la probabilité Q de dépasser par hasard le χ^2 observé. Si cette probabilité est trop faible on rejette la loi ajustée. Dans les exemples précédents les valeurs de Q sont respectivement :

	BRIVE	TEILLET-ARGENTY
GUMBEL	0.003	0,09
FRECHET	0.0004	0.64

Les probabilités ainsi trouvées ne permettent pas de choix entre les deux lois. Quelles sont les raisons de ces résultats décevants qui se retrouvent d'ailleurs dans la plupart des applications du χ^2 aux problèmes de crues ?

Le χ^2 n'est pas un critère de choix mais un test d'adéquation, il ne permet pas de dire si l'ajustement est meilleur avec telle loi plutôt qu'avec telle autre, il nous permet seulement de savoir si l'échantillon observé est compatible avec la loi ajustée. Or, ce test est plus ou moins puissant. La puissance, qui se définit par la probabilité de conclure au rejet de la loi lorsque l'échantillon n'en est pas issu, est d'autant plus grande que la taille de cet échantillon est grande ; cette dernière circons-

tance ne se rencontre pas fréquemment dans les applications aux débits de crue. D'un autre côté la description de la loi des débits est d'autant meilleure que le découpage en classes est plus fin et par conséquent le nombre de classes plus élevé ; mais si on augmente le nombre des classes on diminue leur fréquence. On ne peut donc pas pousser le découpage trop loin car la loi du χ^2 n'est valable que si les fréquences théoriques des classes sont supérieures à une certaine limite (disons 5 pour fixer les idées). C'est très peu, et cela diminue les chances de conclure exactement au rejet de la loi testée même lorsque l'échantillon est issu d'une loi très différente.

Dans le cas des lois de GUMBEL et FRECHET, cette limitation du nombre de classes est d'autant plus fâcheuse que ces lois sont assez semblables dans la partie centrale des distributions et ne se distinguent que pour les grandes valeurs de la variable, là où il y a peu d'observations ; on aurait donc intérêt à multiplier les classes pour ces valeurs. Malheureusement la condition imposée aux fréquences demande que l'on regroupe ces classes extrêmes, les différences entre les lois s'atténuent considérablement avec ces regroupements. Enfin, le découpage en classes n'est pas exempt d'un certain arbitraire dans le choix des limites des classes et on pourrait avoir des résultats très différents selon les découpages utilisés. Cet arbitraire a d'autant plus d'importance que le nombre de classes est plus petit.

Le choix entre les lois de GUMBEL et FRECHET est donc malaisé il est souvent guidé par des considérations subjectives. Il nous semble qu'on ne saurait trouver de solutions en ce domaine que par l'application répétée des méthodes statistiques précédentes à un très grand nombre de cas. Cela permettrait peut-être de dégager certaines règles de choix basées sur les caractéristiques hydrologiques et météorologiques des bassins versants. De toutes façons, le choix n'est pas exempt d'un certain risque d'erreurs. Ces erreurs d'adéquation ont malheureusement une très grande importance pour le calcul d'un quantile car les comportements des deux lois des extrêmes envisagées ici sont très différents dans la zone des petites valeurs de la probabilité p.

III - LES ERREURS DE MESURE

Il ne faut pas oublier que les erreurs de mesure (singulièrement dans l'étude des crues) peuvent jouer un rôle important. D'ailleurs les erreurs d'adéquation et les erreurs de mesure sont souvent dépendantes. En effet, les erreurs de mesure sont d'autant plus grandes que les valeurs des débits sont plus fortes, en particulier parce que ces dernières sont estimées dans la zone de la courbe de tarage définie, non par des résultats de jaugeages complets, mais obtenue par des règles d'extrapolation souvent assez arbitraires. Ces erreurs de mesure auront donc tendance à accroître indûment la dispersion des débits maximaux annuels. Or, la loi de FRECHET se caractérise par une dispersion relative plus grande que celle de la loi de GUMBEL. Ces circonstances peuvent entraîner au choix trop fréquent de la loi de FRECHET au détriment de la loi de GUMBEL.

Nous avons commencé à aborder l'examen pratique de cette question au moyen de la méthode des échantillons fictifs. Nous avons donc contitué des échantillons de valeurs $z = x + \varepsilon$ pour en étudier ensuite la répartition.

VARIABILITE DES LOIS DE GUMBEL AJUSTEES SUR DES ECHANTILLONS DE 30 VALEURS TIRES D'UNE MEME POPULATION

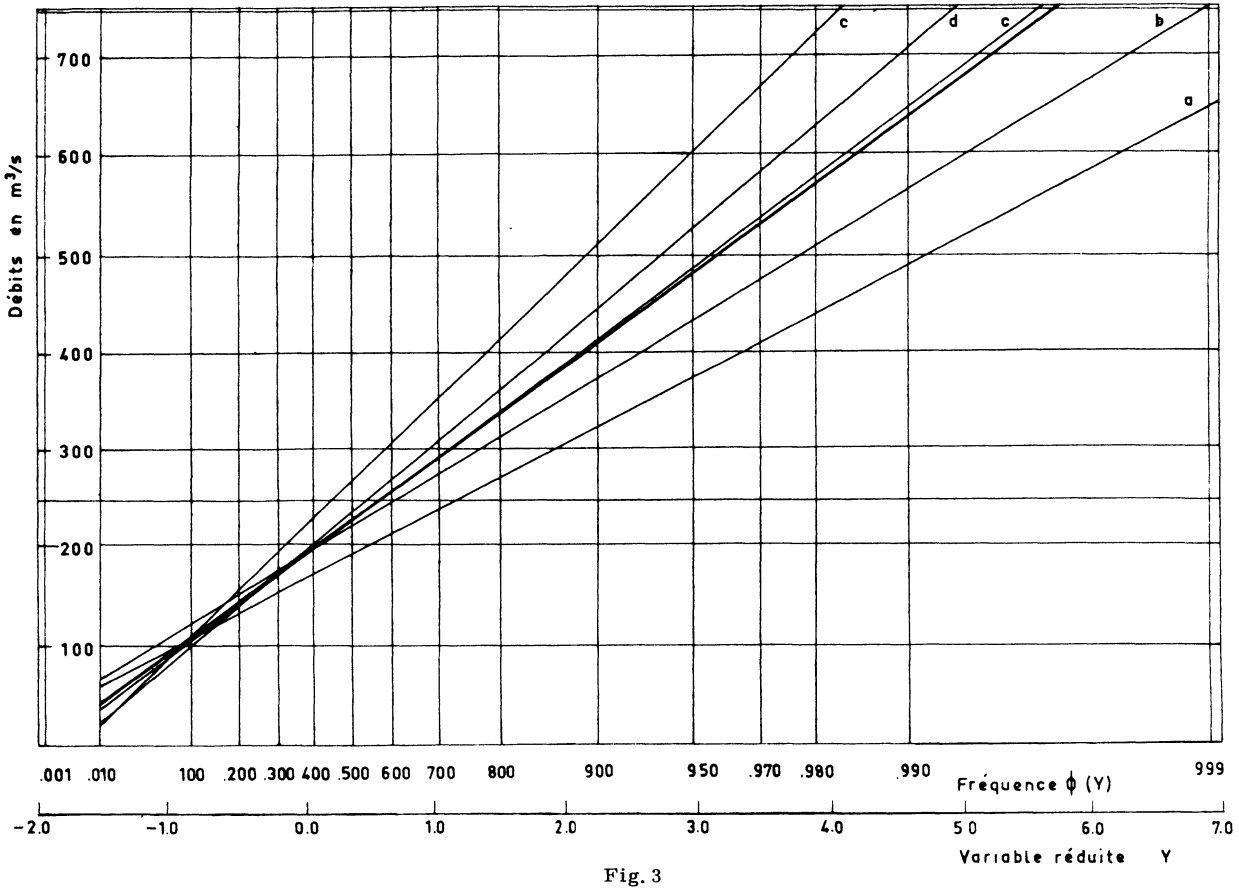


Fig. 3

HISTOGRAMME DES VALEURS ESTIMÉES DE LA CRUE MILLENAIRE

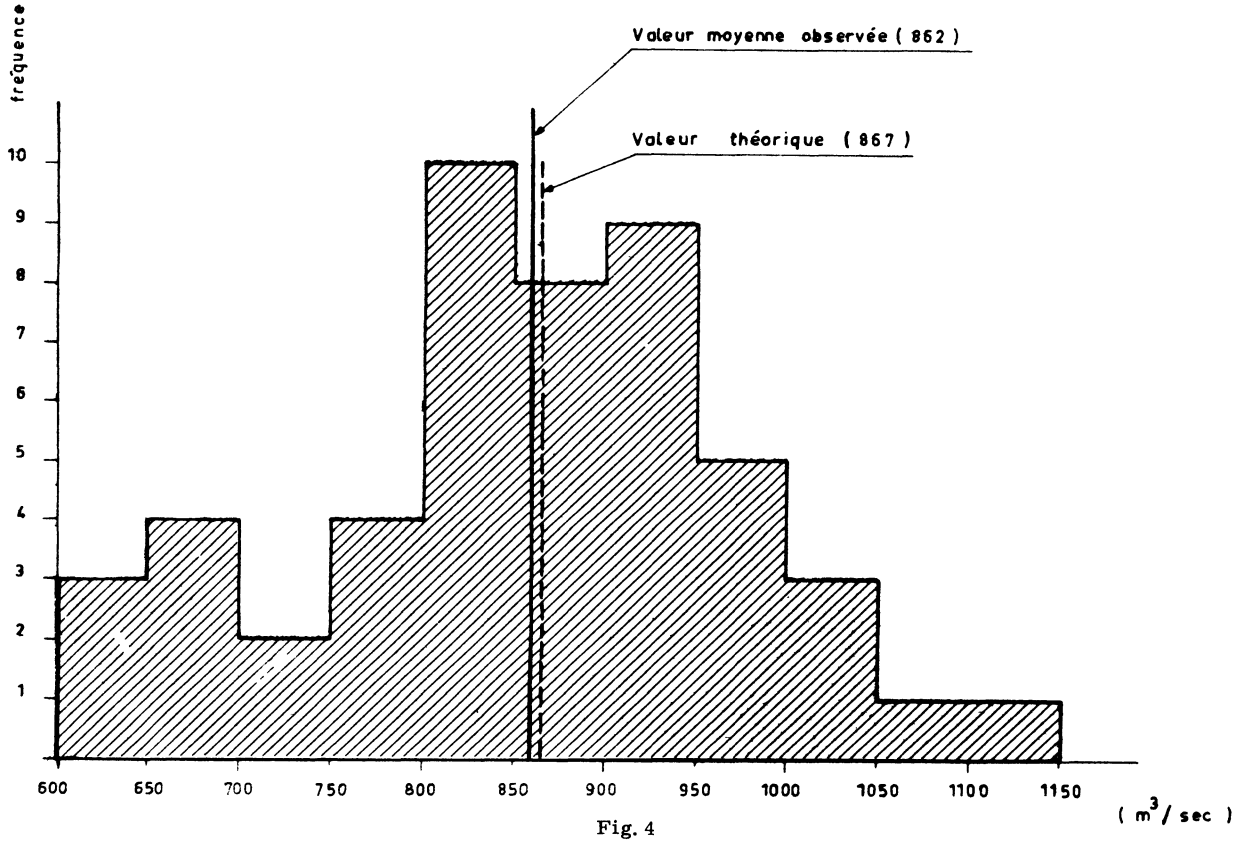


Fig. 4

Les valeurs de X sont tirées au sort dans une loi de GUMBEL :

d'espérance mathématique $E(X) = 250\ m^3/sec.$

d'écart type $\sigma(X) = 125\ m^3/sec.$

Les valeurs de ε sont tirées au sort dans une loi normale :

d'espérance mathématique $E(\varepsilon) = a\,X$

d'écart type $\sigma(\varepsilon) = b\,X$

Les valeurs de z ainsi obtenues symbolisent les débits réels. Les valeurs de ε représentent les erreurs de mesure.

On pourrait bien entendu faire d'autres hypothèses sur la loi de variation de ε et sur la façon dont ses caractéristiques sont liées à X. Celles que nous avons retenues n'entraînent ni calculs difficiles ni manipulation de tables compliquées.

Après les essais que nous avons effectués pour quelques valeurs de a et de b, il ne nous est pas encore possible de tirer des conclusions sur le sens des distorsions provoquées par les erreurs de mesure telles que nous les avons envisagées. Nous avons simplement constaté des écarts pour les grandes valeurs de débit, écarts dont certains seraient susceptibles d'infléchir le choix de la loi de répartition des maxima annuels. Nous ne pouvons pas préciser que ces distorsions "favoriseraient" en général la loi de Fréchet.

Notons que le fait que ces écarts portent tout particulièrement sur les grandes valeurs des débits a une grande importance pour nous, puisque c'est aux crues catastrophiques que nous nous intéressons.

IV - LES ERREURS D'ECHANTILLONNAGE

La loi de probabilité donne la répartition des fréquences de la population composée de tous les maxima annuels possibles observables. Les valeurs observées au cours de la période étudiée ne constituent en fait qu'un échantillon tiré au hasard parmi la population totale. La loi ajustée à l'échantillon ne coïncide pas avec la loi exacte de la population, elle peut s'en écarter plus ou moins : les écarts sont aléatoires et constituent les erreurs d'échantillonnage.

Supposons que la loi exacte des maximums annuels relevés à une certaine station de jaugeage soit une loi de GUMBEL définie par son espérance mathématique et son écart-type :

$$E(X) = 250\ m^3/s \qquad (9)$$

$$\sigma(X) = 250\ m^3/s \qquad (10)$$

et supposons que l'on ait observé sur une période de 30 ans les valeurs de l'échantillon noté a) sur le graphique III ci-joint. Cet échantillon est tiré au sort dans la loi initiale. Cependant on aurait tout aussi bien pu obtenir des échantillons différents tels que ceux qui sont notés b) c)... sur le graphique III et qui sont également tirés au sort dans la loi initiale. Les droites de GUMBEL ajustées sont différentes et les valeurs estimées de la crue millénaire par exemple varient dans une très large fourchette.

Le graphique IV montre le polygone de fréquences construit à partir

des estimations de la crue millénaire calculées sur 50 échantillons tirés au sort dans la population précédente. On constate la très grande dispersion de l'estimation autour de la vraie valeur de la crue millénaire. Les deux graphiques concrétisent ces erreurs d'échantillonnage dont il est important de tenir compte dans le calcul d'un quantile.

Calcul de l'intervalle de confiance d'un quantile

Plaçons-nous dans le cadre de la loi de GUMBEL ; la loi de FRECHET s'y ramène par une transformation logarithmique. D'après les formules (5) et (6) le quantile x_p peut s'écrire :

$$x_p = u + \frac{1}{\alpha} y(p) \tag{11}$$

où y (p) est tel que :

$$1 - e^{-e^{-y(p)}} = p \tag{12}$$

Il est commode de transformer la formule (11) de façon à faire apparaître des paramètres ayant une signification plus directe comme l'espérance mathématique m et l'écart-type σ :

$$x_p = m + \sigma \lambda(p) \tag{13}$$

avec :

$$\lambda(p) = \frac{y(p) - E(y)}{\sigma(y)} \tag{14}$$

y et λ sont donc liés linéairement.

Pour l'estimation à partir des n observations x_i, m et σ sont remplacés respectivement par la moyenne et l'écart-type empiriques calculés sur l'échantillon des x_i :

$$\bar{x} = \frac{\Sigma x_i}{n}$$

$$s = \sqrt{\frac{\Sigma (x_i - \bar{x})^2}{n - 1}}$$

d'où le quantile estimé :

$$\hat{x}_p = \bar{x} + \lambda_p s \tag{15}$$

$\hat{x}_p$, calculé sur l'échantillon, est, nous l'avons vu, une grandeur aléatoire. On démontre que la loi de probabilité de la quantité :

$$T = \frac{\hat{x}_p - x_p}{s} \text{ (variable "studentisée")} \tag{16}$$

est indépendante des paramètres définissant la loi de GUMBEL (u et α ou m et σ). Pour cela, on introduit la variable t telle que :

$$T = \lambda + t, \qquad t = \frac{\bar{z} - \lambda}{s(z)}$$

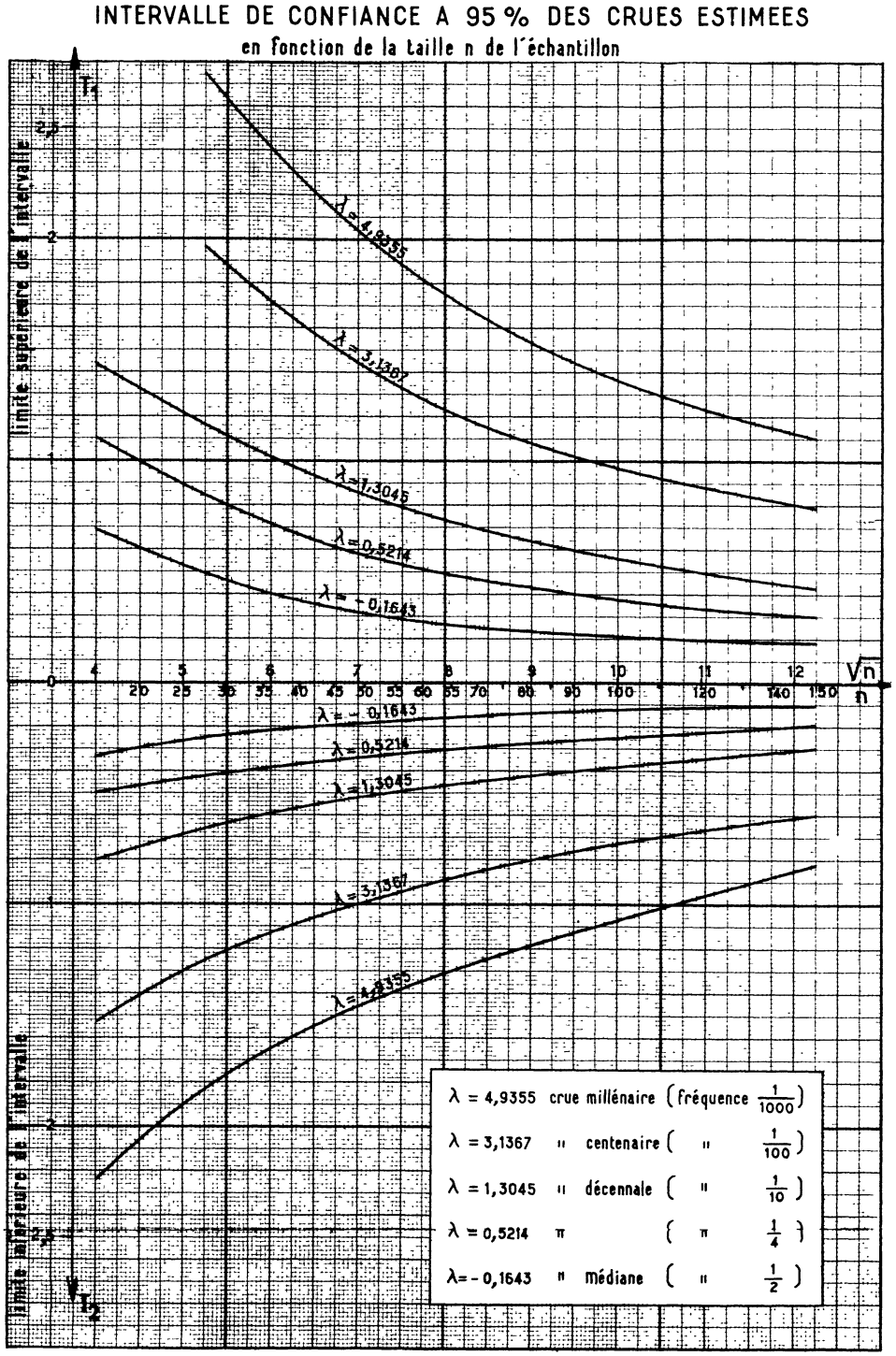

INTERVALLE DE CONFIANCE A 95 % DES CRUES ESTIMEES
en fonction de la taille n de l'échantillon
limite supérieure de l'intervalle
limite inférieure de l'intervalle
T1
T2
λ = 4,9355
λ = 3,1367
λ = 1,3045
λ = 0,5214
λ = - 0,1643
√n
n
λ = 4,9355 crue millénaire (fréquence 1/1000)
λ = 3,1367 " centenaire (" 1/100)
λ = 1,3045 " décennale (" 1/10)
λ = 0,5214 " (" 1/4)
λ = - 0,1643 " médiane (" 1/2)

Fig. 5

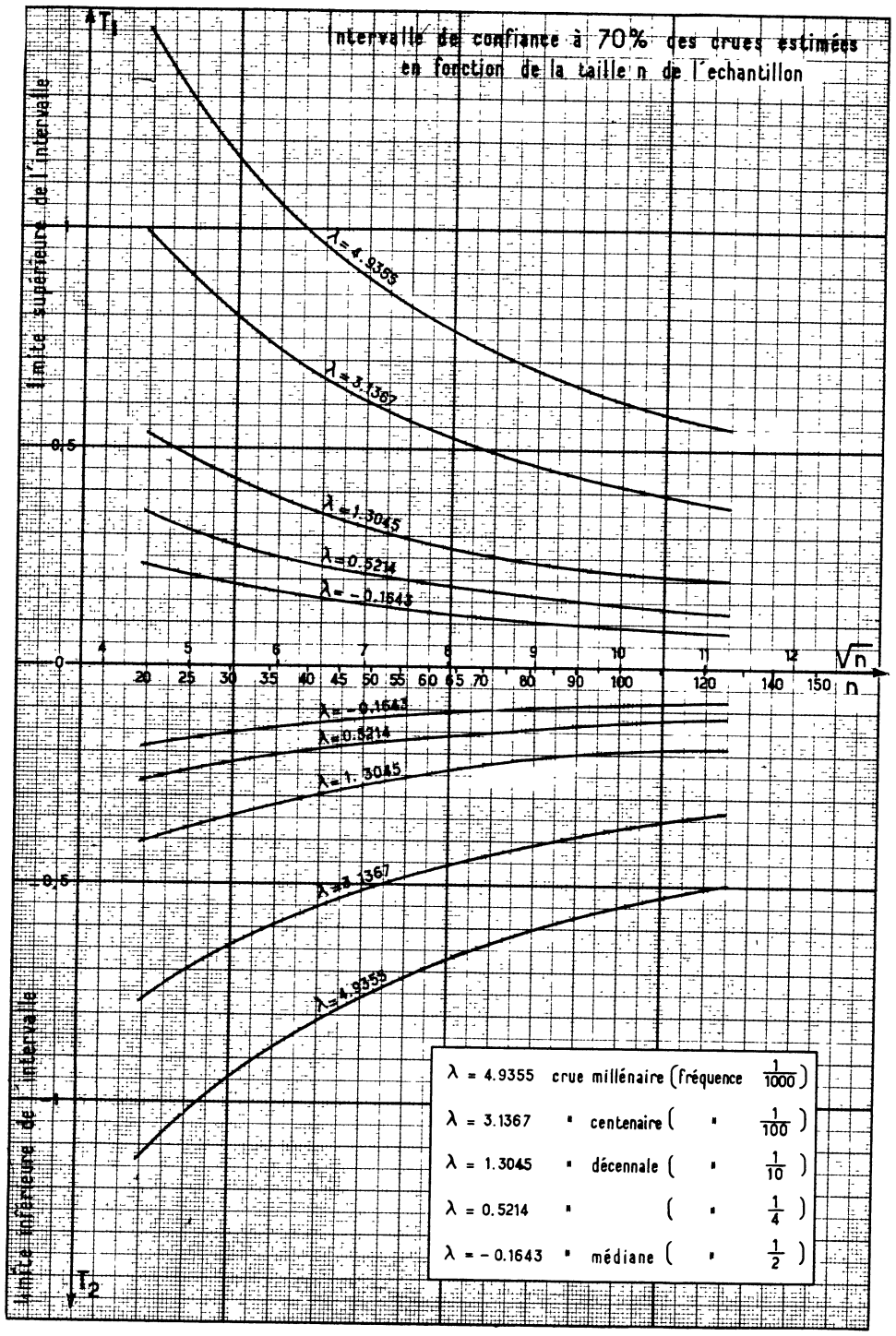

Intervalle de confiance à 70% des crues estimées
en fonction de la taille n de l'échantillon
limite supérieure de l'intervalle
T_1
$\lambda = 4.9355$
$\lambda = 3.1367$
$\lambda = 1.3045$
$\lambda = 0.5214$
$\lambda = -0.1643$
0,5
0
4 5 6 7 8 9 10 11 12 $\sqrt{n}$
20 25 30 35 40 45 50 55 60 65 70 80 90 100 120 140 150 n
$\lambda = -0.1643$
$\lambda = 0.5214$
$\lambda = 1.3045$
$\lambda = 3.1367$
$\lambda = 4.9355$
-0,5
limite inférieure de l'intervalle
T_2
λ = 4.9355 crue millénaire (fréquence $\frac{1}{1000}$)
λ = 3.1367 " centenaire (" $\frac{1}{100}$)
λ = 1.3045 " décennale (" $\frac{1}{10}$)
λ = 0.5214 " (" $\frac{1}{4}$)
λ = − 0.1643 " médiane (" $\frac{1}{2}$)

Fig. 6

$\bar{z}$ et s(z) représentent la moyenne et l'écart-type de n variables issues d'une loi de GUMBEL de moyenne nulle et d'écart-type unité. Pour t donné :

$$\text{Prob. } [t \leqslant t_o] = \text{Prob. } \left[\frac{\bar{z} - \lambda}{s(z)} \leqslant t_o \right]$$

On obtient avec la même probabilité la relation :

$$\bar{z} - t_o \, s \, (z) < \lambda$$

De l'étude du rapport :

$$T = \frac{\hat{x}_p - x_p}{s}$$

on est ainsi passé à l'étude de la variable :

$$Z = z - t_o \, s(z)$$

Dans cette formule t_o joue le rôle d'un paramètre. Connaissant Z on déduira t, or si on ne peut déterminer exactement la loi de Z, on peut en trouver une approximation très suffisante.

En effet, Z est une fonction continue de la moyenne et de l'écart-type empiriques. La distribution de Z est donc asymptotiquement normale.

Cependant, lorsque le paramètre t_o est grand (fortes crues), l'approximation normale n'est pas suffisante. La cause d'erreur est essentiellement la dissymétrie de la loi de s(z). Il est donc nécessaire d'utiliser une correction tenant compte de cette dissymétrie, ce qui est obtenu, de façon classique, en utilisant le développement de GRAM-CHARLIER limité aux premiers termes de la loi de Z, ce qui conduit finalement, pour un quantile fixé et un seuil de probabilité donné, à une relation entre la taille de l'échantillon : n et les limites de l'intervalle de confiance pour la crue estimée.

On peut donc déterminer T_1 et T_2 tels que :

$$\text{Prob. } [T_1 \leqslant T \leqslant T_2] = \alpha \qquad (17)$$

et l'intervalle de confiance pour x_p défini par :

$$\hat{x}_p - T_2 \, s \leqslant x_p \leqslant \hat{x}_p + T_1 \, s \qquad (18)$$

recouvrira la vraie valeur du quantile x_p avec la probabilité α.

Les abaques V et VI donnent, pour $\alpha = 0,95$ et $\alpha = 0,70$. les valeurs de T_1 et T_2 en fonction de la taille de l'échantillon n pour différents quantiles.

Pratiquement, on utilisera l'intervalle à 70 % car si l'intervalle à 95 % correspond à une probabilité plus forte de recouvrir la vraie valeur de x_p, il est en général d'amplitude trop grande (singulièrement pour la loi de FRECHET) pour avoir une signification pratique ; cependant, il aura son utilité pour résoudre un problème soulevé dans la suite de cet exposé.

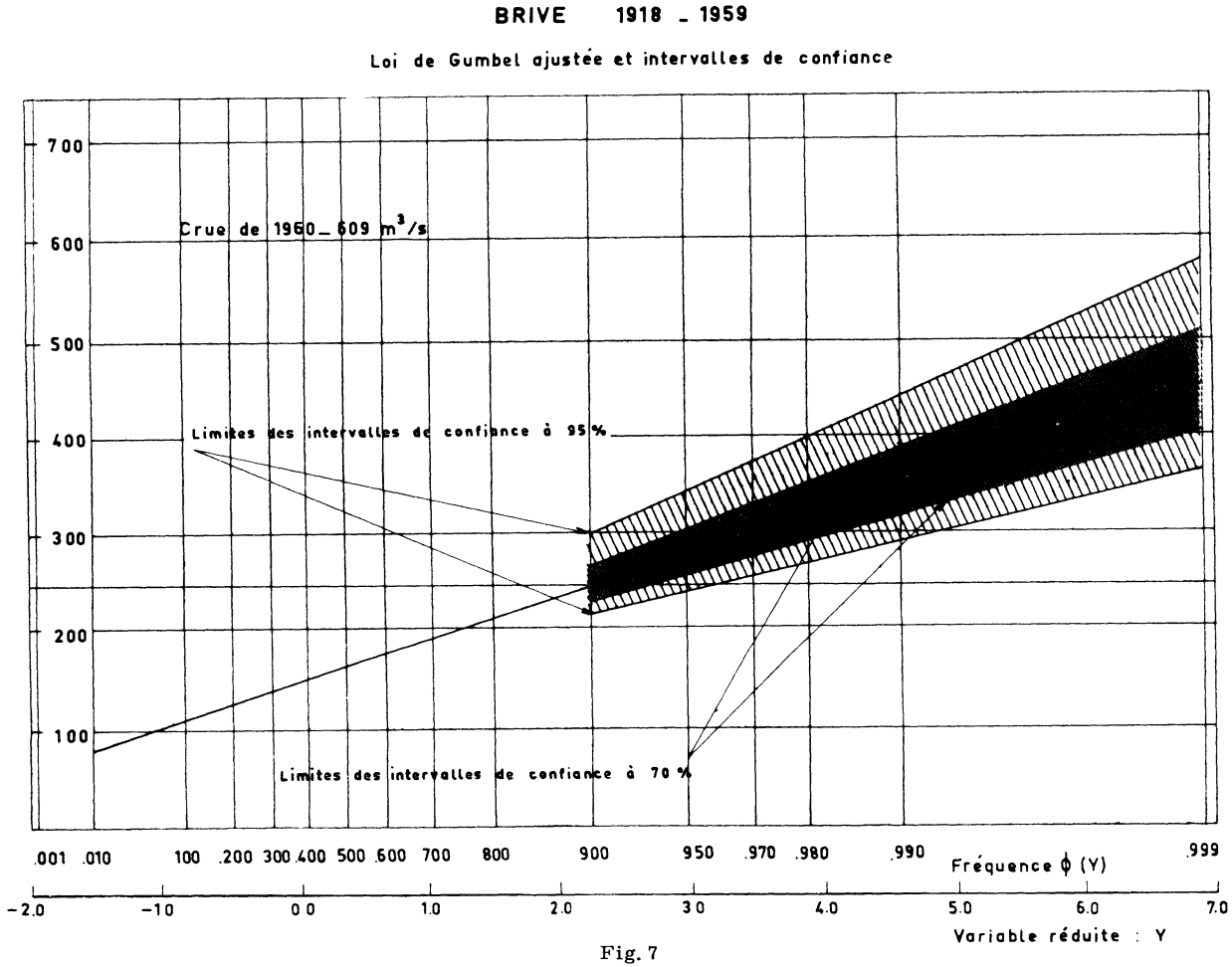

BRIVE 1918 _ 1959
Loi de Gumbel ajustée et intervalles de confiance
Crue de 1960_ 609 m³/s
Limites des intervalles de confiance à 95 %
Limites des intervalles de confiance à 70 %
700
600
500
400
300
200
100
.001 .010 100 .200 300 .400 500 .600 700 800 900 950 .970 .980 .990 .999
Fréquence ϕ (Y)
-2.0 -1.0 0.0 1.0 2.0 3.0 4.0 5.0 6.0 7.0
Variable réduite : Y

Fig. 7

La figure VII reproduit la loi ajustée par M. PELLECUER aux maxima annuels de la Corrèze à BRIVE observés pendant la période 1918-1959. Les limites inférieures et supérieures des intervalles de confiance sont alignées. C'est une propriété générale qui pourrait se démontrer directement.

Remarquons que les droites ainsi construites ne limitent pas un domaine de confiance à 70 % ou à 95 % de la droite de GUMBEL théorique toute entière, mais simplement les intervalles de confiance pour chaque crue de probabilité donnée.

Intervalle de confiance de la probabilité affectée à un débit donné

C'est le problème inverse du précédent : On se fixe une valeur x_o du débit et on veut estimer p_o tel que :

$$p_o = \text{Prob.}\ [X > x_o]$$

La loi ajustée aux observations fournit une estimation ponctuelle de p_o mais il est important d'en déterminer également un intervalle de confiance.

La loi de GUMBEL s'exprime de façon indépendante des paramètres au moyen des variables réduites y ou λ. Déterminer une valeur de p c'est déterminer une valeur de y ou λ et réciproquement.

Comme il existe une relation linéaire entre λ et y (formule 14) nous raisonnerons uniquement sur λ.

Si nous connaissions m et σ, la valeur λ correspondant au débit x_o serait :

$$\lambda_o = \frac{x_o - m}{\sigma}$$

Or, on estime m et σ par $\bar{x}$ et s, on ne connaît donc de λ_o que son estimation :

$$l_o = \frac{x_o - \bar{x}}{s}$$

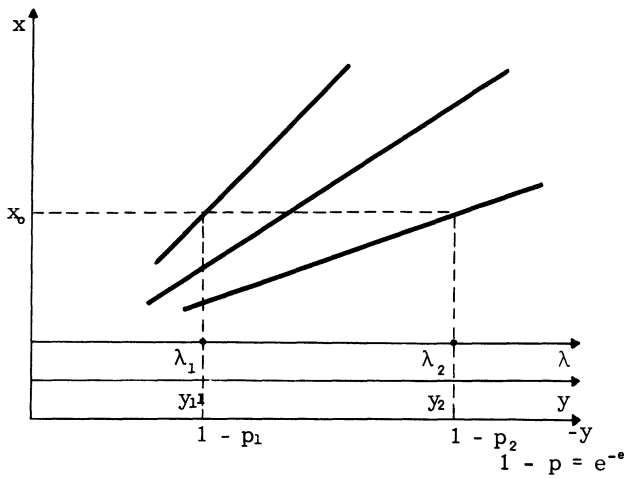
l_o est fonction de l'échantillon, c'est donc une variable aléatoire dont on démontre que la loi de probabilité ne dépend que de λ_o. Cette loi permet la détermination d'un intervalle de confiance (λ_1, λ_2) qui recouvre la vraie valeur λ_o avec une probabilité égale à α.

Pratiquement, cet intervalle se détermine graphiquement à partir des droites limites des intervalles de confiance au seuil α des quantiles. Il suffit de lire les abscisses sur l'échelle des probabilités des points d'intersection des droites calculées pour le seuil α avec la droite d'ordonnée constante égale à x_o.

Remarquons que les valeurs lues sur l'échelle des probabilités ne sont pas les valeurs de p mais celles de 1 - p.

On peut appliquer cette méthode pour déterminer un intervalle de confiance de la crue de 1960 à partir de la loi de GUMBEL ajustée aux débits de 1918 à 1959 :

$$x_o = 609\ m^3/s$$



On obtient :

- pour l'intervalle à 70 % : $p_1 = 1,2 \times 10^{-4}$ $p_2 = 4 \times 10^{-6}$
- pour l'intervalle à 95 % : $p_1 = 5,5 \times 10^{-4}$ $p_2 = 4 \times 10^{-7}$

Intervalle de confiance de la probabilité affectée à un débit observé

La méthode précédente est valable lorsque le débit x_0 est donné à priori. Or, dans l'exemple de BRIVE, x_0 est la dernière valeur observée, c'est même le maximum de la période d'observation de 1918 à 1960 ; la réalisation de cet évènement apporte donc une information supplémentaire sur p_0. D'ailleurs, ces circonstances se rencontrent fréquemment en pratique car l'étude des débits maximaux annuels n'est souvent reprise qu'au moment de l'apparition d'une crue très importante.

On dispose donc de n + 1 débits observés dont le dernier x_{n+1} dépasse très largement les observations antérieures. On peut se demander si x_{n+1} et les autres débits constituent un échantillon homogène dans le cadre d'une certaine loi de probabilité.

La méthode exposée au paragraphe précédent permet la construction d'un intervalle de confiance a priori au seuil α à partir de la loi ajustée aux n premiers débits. Soient p_1 et p_2 les limites de cet intervalle.

On peut construire un autre intervalle basé sur les considérations suivantes : la probabilité que la valeur x_0 n'ait été observée qu'à partir de la (n + 1)ième année s'écrit :

$$(1 - p_0)^{n+1} \simeq e^{-(n+1) p_0} \qquad (19)$$

on peut alors déterminer un intervalle (p'_1, p'_2) tel que toute valeur de p comprise entre ces limites affecte à l'évènement observé une probabilité non négligeable.

Prenons p'1 tel que :

$$e^{-(n+1)p'_1} = \frac{1 - \alpha}{2}$$

soit :

$$p'_1 = \frac{-1}{n+1} \operatorname{Log} \frac{1-\alpha}{2} \tag{20}$$

et p'_2 tel que :

$$e^{-(n+1)p'_2} = 1 - \frac{1-\alpha}{2} = \frac{1+\alpha}{2}$$

soit :

$$p'_2 = \frac{-1}{n+1} \operatorname{Log} \frac{1+\alpha}{2} \tag{21}$$

Les limites déterminées par (20) et (21) fournissent un intervalle de confiance au seuil α pour le paramètre p_o. Cet intervalle est d'ailleurs construit en ne faisant aucune hypothèse sur la forme de la loi de probabilité.

L'intersection éventuelle des intervalles (p_1, p_2) et (p'_1, p'_2) permet de s'assurer que la valeur observée x_o reste compatible avec les observations antérieures dans le cadre de la loi ajustée à celles-ci. Dans le cas où les deux intervalles se coupent, le centre de la partie commune peut fournir une estimation ponctuelle de p_o.

Les deux intervalles sont indépendants ; de cela il résulte que la probabilité pour que la vraie valeur de p_o soit située à la fois sur les deux intervalles est égale à α^2 :

si $\alpha = 0,7$ $\quad \alpha^2 = 0,49$

si $\alpha = 0,95$ $\quad \alpha^2 \simeq 0,90$

La première probabilité semble trop petite, c'est pourquoi il est préférable d'utiliser dans ce problème un seuil à 95 %.

Dans le cas de BRIVE, on a $n + 1 = 43$ et

$$p'_1 = 0,0859 \qquad p'_2 = 0,0006$$

Avec la première méthode nous avons trouvé :

$$p_1 = 0,00055 \qquad p_2 = 0,0000004$$

Les deux intervalles sont disjoints et la loi de GUMBEL ajustée aux débits de 1918 à 1959 donne à la crue de 1960 une probabilité trop faible, incompatible avec l'apparition effective de cette crue.

La méthode que nous venons d'exposer nous semble préférable à celle qui consiste à ajuster une loi de probabilité, d'une part à la série complète des $n + 1$ débits, d'autre part à la série des n premiers débits. Les deux intervalles que l'on peut calculer à partir de chaque ajustement ne sont pas indépendants.

CONCLUSION

En conclusion nous pouvons dire que la statistique n'apporte pas de recettes susceptibles de conduire de façon mécanique à un résultat certain.

Les modes de raisonnement employés ont leurs règles propres qu'il ne faut pas oublier : ainsi l'estimation d'un quantile ou d'une probabilité doit toujours être donnée avec un intervalle de confiance correspondant au seuil de probabilité choisi. D'autre part, il faut se méfier de certains mots communs au vocabulaire statistique et au vocabulaire habituel, de la physique par exemple. Ces mots ne doivent pas être interprêtés de la même façon dans les deux "domaines". Il en va ainsi pour le mot loi : Il ne faut pas se laisser entraîner à lui donner en hydrologie statistique le caractère déterministe qu'il a dans bien des cas en physique classique.

L'intérêt des méthodes que nous employons est de nous permettre de chiffrer dans une certaine mesure l'imprécision de nos réponses. Cette imprécision est liée au stade de développement des connaissances dans les domaines où nous travaillons et à la nature de nos informations.

Les données dont dispose l'hydrologue contiennent une certaine quantité d'information que nous nous efforçons d'extraire de la façon la plus complète possible, mais il y a une limite qui ne dépend pas des méthodes employées. Déplacer cette limite ne relève plus de l'hydrologie statistique, mais de l'hydrologie expérimentale grâce à laquelle d'autres analyses statistiques deviendront à nouveau possible.

En tout état de cause, l'amélioration des résultats obtenus par les méthodes de prévision statistique est liée aux efforts faits tant dans le domaine de l'hydrologie expérimentale que dans le domaine de la statistique mathématique.

Il importe autant d'améliorer les techniques d'observation et de mesure que de raffiner encore les schémas probabilistes qui servent au traitement statistique des "données" que constituent les résultats de ces mesures.

DISCUSSION

(Président ; M. DELAPORTE)

M. le Président remercie et félicite M. VERON de son brillant exposé.

Au sujet de l'emploi du test χ^2 M. SNEYERS rappelle que, pour en faire une application objective, il convient d'abandonner le partage du domaine de définition de la loi en classes d'égale amplitude au profit d'un partage en classes d'égale probabilité (quantiles). Cf. Gumbel, "On the reliability of the χ^2 test", publié il y a quelques années.

M. NORMAND rappelle que M. VERON a souligné tout le danger qu'il y a à employer l'expression "temps de retour" pour des débits de probabilité annuelle $\frac{1}{100}$ ou $\frac{1}{1\,000}$, certains profanes risquant de lui donner un sens déterministe ; M. NORMAND suggère donc, dans un souci de logique du langage, de caractériser une période de 100 ans ou de 500 ans par le débit qui a une chance sur deux, par exemple, d'être dépassé une ou plusieurs fois en 100 ou 500 ans. Une telle notion serait aussi "parlante" et prêterait moins à des interprétations abusives que celle de "temps de retour".

M. VERON croit tout de même à l'utilité et à la commodité de la notion de temps de retour et pense que l'on devrait toujours associer les estimations que

l'on donne à quelques valeurs de Prob. [N < n] prises dans un tableau semblable à celui figurant au chapitre I du mémoire de MM. BERNIER et VERON.

M. BERNIER précise la notion de la durée de retour : quel que soit le choix du débit de crue caractéristique pris en compte pour le dimensionnement d'un barrage, il existe toujours un arbitraire qui ne peut-être levé que par la prise en compte des éléments économiques faisant intervenir les investissements, d'une part et les coûts des dommages occasionnés par les crues, d'autre part.

M. RODIER rappelle que M. VERON a insisté, à juste titre à son avis, sur l'intérêt d'améliorer la précision avec laquelle sont déterminés les débits, c'est-à-dire sur la mise au point de la courbe d'étalonnage.

M. RODIER illustre cet intérêt par l'exemple suivant ; sur le Konkouré en Guinée, M. RODIER a estimé une première fois la crue millénaire avec une courbe d'étalonnage dont le point le plus élevé était un peu inférieur à la valeur médiane du maximum annuel. Or, une mesure ultérieure de débit de la crue décennale faite, non sans difficulté, certes, sur ce fleuve a permis de constater une erreur de 30 % sur la courbe d'étalonnage, d'où résultait une différence voisine de 100 % sur la crue millénaire. M. RODIER souhaite donc que de semblables jaugeages de vérification soient faits le plus souvent possible en période de forte crue, en dépit des difficultés réelles que cela comporte.

M. RODIER met aussi en garde contre les erreurs graves dûes à de mauvais échantillonnages : dans un histogramme provenant du choix "au hasard" de divers échantillons, les valeurs extrêmes encadrant la valeur la plus probable recherchée, par exemple la moyenne annuelle interannuelle, correspondent à des écarts souvent inadmissibles et qui peuvent être évités. Souvent, en effet, on dispose de relevés sur une douzaine ou une quinzaine d'années mais on sait très bien qu'une partie très importante de ces années est dans une période exagérément humide ou exagérément sèche par rapport aux variations générales d'hydraulicité que l'on connaît de façon qualitative. On sait donc que l'échantillon est "fort" ou "faible", une légère réduction par exemple pourra donner plus de chance d'avoir un échantillon "haut placé", réduction qui pourrait être guidée par les données pluviométriques.

M. VERON remercie M. RODIER d'avoir abordé un point important sur lequel il n'avait pas suffisamment insisté, dans son exposé : le rôle de l'expérience des utilisateurs d'estimations.

M. SNEYERS demande s'il existe une méthode imposant une limite au nombre de paramètres à estimer pour ajuster une loi de probabilité à un échantillon de taille donnée.

M. ROCHE pense que l'on peut avoir une idée de cette limite en considérant comment la signification du test varie, dans l'ajustement ; avec la valeur du degré de liberté.

M. le Président suggère de trouver des méthodes d'estimation des paramètres qui utilisent au mieux l'information dont on dispose. La difficulté en hydrologie réside dans la faiblesse du nombre d'années d'observation par rapport à la durée incomparablement plus longue sur laquelle on désire faire des estimations : c'est donc la recherche d'optimum d'efficacité qui doit orienter la méthode d'estimation des paramètres.

M. BERNIER souligne que le choix du nombre de paramètres à prendre en compte dans un ajustement ne peut résulter que de l'expérience accumulée au cours de plusieurs années d'applications pratiques de multiples lois.

En ce qui concerne les phénomènes hydrologiques, il semble que, compte tenu de l'information disponible et de la valeur de cette information, on ne puisse pas dépasser 3 paramètres pour les séries d'une quarantaine d'années dont on dispose généralement.

M. ROCHE pense que l'estimation d'un débit de crue sur un cours d'eau est étayée par celle de la connaissance géographique, information "dans l'espace" qui, sans pouvoir être chiffrée, peut quelquefois compléter l'information "dans le temps", sur les crues anciennes, évoquée par M. RODIER.

M. le Président, en qualité de Président de la Société de Statistique de Paris, conclut en remerciant et félicitant la Société Hydrotechnique de France de la très haute tenue d'une communication comme celle-ci, parce qu'elle allie des méthodes fines du calcul des probabilités et de la Statistique mathématique avec des problèmes concrets particulièrement difficiles, tels que le comportement asymptotique des crues, et aussi avec le risque d'erreurs de mesures provenant du profil de la rivière et de ses modifications au cours du temps.