Essai d'une étude statistique des erreurs de calcul

Etienne Gorog

N° d'ordre :

THÈSE

présentée à

LA FACULTÉ DES SCIENCES DE L'UNIVERSITÉ DE GRENOBLE

pour obtenir

LE TITRE DE DOCTEUR DE SPÉCIALITÉ
Mathématiques Appliquées

par

Etienne GOROG

Licencié ès Sciences

ESSAI D'UNE ÉTUDE STATISTIQUE DES ERREURS DE CALCUL

Thèse soutenue le : *devant la Commission d'examen :*

MM. KUNTZMANN, *Président*
VAUQUOIS
GASTINEL

J'exprime ma reconnaissance à Monsieur le Professeur Kuntzmann. Qu'il me soit permis de lui témoigner aussi ma respectueuse admiration.

Je remercie également mes Maîtes, Messieurs Vauquois et Gastinel, dont les conseils m'ont été précieux.

INTRODUCTION

Nous nous proposons d'étudier, dans une première partie "l'erreur de calcul" finale qui résulte d'une succession d'opérations simples (sommes, produits).
Le nombre de ces opérations peut être très élevé.

Nous présentons pour chaque type d'opération une ou plusieurs théories probabilistes : dans le cas d'une somme nous développerons deux théories principales correspondant à deux voies d'approche différentes.

Les résultats obtenus sur des exemples pratiques sont ensuite comparés à ceux prédits par ces théories.

Dans une seconde partie nous traiterons le cas de la résolution approchée de systèmes d'équations différentielles. L'étude sera faite sur des systèmes qui satisfont à des conditions initiales et sont résolus par des méthodes à pas séparés.

Nous essaierons de montrer comment, et dans quelles conditions, il est possible de déterminer, au cours de l'intégration, "l'erreur de calcul" propagée.

I - ERREUR DE CALCUL

1) L'utilisation de toute machine à calculer, de la plus simple à la plus perfectionnée, engendre en général, au cours de l'exécution d'un calcul quelconque une certaine erreur.

Indépendamment de la méthode et de l'outil mathématique employés, cette erreur existe du fait de la nature même du nombre qui, transcendant, irrationnel ou simplement rationnel, peut comporter un nombre de chiffres illimité. Dans un calcul pratique, il n'est pas possible de l'utiliser tel quel. Il en est de même si le nombre des chiffres significatifs sans être illimité, est très élevé.

Nous nous intéresserons spécialement à l'erreur dûe au fait qu'un nombre de plus de k chiffres (nombre initial) doit être remplacé par un nombre qui ne comportera que k chiffres exactement (nombre approché qui servira à effectuer le calcul). Nous appellerons cette erreur erreur de chute lorsqu'elle est dûe à une perte complète de chiffres (coupure), et erreur d'arrondi lorsque le calculateur fait un arrondi automatique, après chaque opération des nombres qu'il manipule.

L'ensemble de ces deux types d'erreurs sera désigné par ERREUR de CALCUL.

2) Cette "erreur de calcul" dépend principalement :

a) <u>des caractéristiques de la machine</u>

- mémoire : nombre de positions
- système de numération employé

Notre étude portera sur le système décimal (base 10) et nous utiliserons le procédé de programmation appelé point décimal flottant (virgule flottante).

C'est le mode de travail qui, en pratique, est le plus souvent adopté.

Décomposition d'un nombre quelconque C différent de 0.

Nous l'écrirons : $C = \xi \cdot c \cdot 10^{\gamma}$

- ξ représente $\begin{cases} \text{soit la valeur } +1 \\ \text{soit la valeur } -1 \end{cases}$
- c est un nombre de la forme : $0,c^1\ c^2 \ldots c^m \ldots$

la suite des chiffres c^i peut être limitée ou illimitée.

Nous aurons nécessairement : $1 \leqslant c^1 \leqslant 9$

et $0 \leqslant c^i \leqslant 9$ pour $i \neq 1$

c'est à dire $0,1 \leqslant c < 1$

- γ représente un nombre positif ou négatif.

b) <u>des nombres</u>

S'ils sont en petite quantité et connus de façon précise, pour évaluer l'erreur de calcul, nous aurons besoin de la grandeur de chacun d'eux.

Lorsqu'ils seront nombreux, il sera alors nécessaire d'avoir une certaine connaissance de leur distribution en probabilité.

c) des opérations

- du nombre d'opérations
- de leur nature : l'erreur sur la somme (A+A+A) peut différer de l'erreur sur le produit (3xA)
- de l'ordre dans lequel s'effectuent les opérations :

exemple : (4 : 6) x 3 et (4 x 3) : 6.

II - EMPLOI D'HYPOTHESES PROBABILISTES

1) LOI DE LA VARIABLE A

Lorsque nous aurons à considérer une famille suffisamment grande de n nombres tels que $(A_1, A_2, \ldots A_n)$, nous supposerons que l'ensemble de ces n nombres forme un échantillon d'une loi de probabilité que nous nous serons donnée à priori.

Appelons A la variable aléatoire qui suit cette loi.

$$A = \xi \, . \, a \, . \, 10^{\alpha}$$

La loi de probabilité A sera parfaitement définie lorsque nous connaîtrons sans ambiguité les lois propres de chacune des variables ξ , a , α .

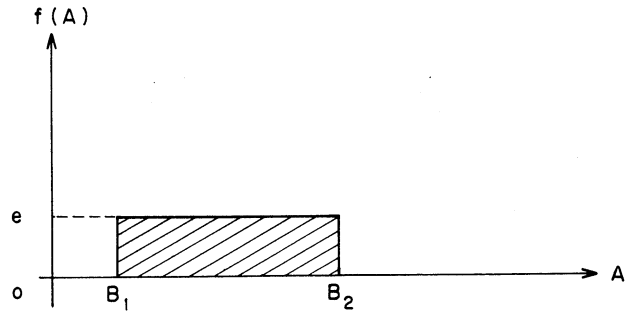
Exemple :

Donnons nous une loi uniforme pour A entre les bornes :

$B_1 = + 10^{\beta_1}$ et $B_2 = + 10^{\beta_2}$ $\qquad \beta_1 < \beta_2$

soit e la valeur $\dfrac{1}{B_2 - B_1}$

f (A) est la densité de répartition.



Les lois propres de ξ , a , et α seront, dans ce cas particulier, les suivantes :

a) $\xi = +1$ avec une probabilité égale à 1 (valeur certaine)

b) a . suit une loi uniforme entre 0,1 et 1

$E(a)^n$ est la valeur moyenne d'ordre n de la variable a .

$$E(a)^n = \frac{10}{9} \int_{0,1}^{1} t^n \, dt = \frac{10\,(1 - 10^{-(n+1)})}{9\,(n+1)}$$

c) α prendra la valeur : $\beta_1 + i$

avec la probabilité : $9\,e\,10^{\beta_1 + i - 1}$

pour : $1 \leq i \leq \beta_2 - \beta_1$

Nous avons naturellement :

$$\sum_{i=1}^{\beta_2 - \beta_1} 9\,e\,10^{\beta_1 + i - 1} = 1$$

Trois familles de fonctions, auxquelles nous pourrons nous référer par la suite, ont été choisies

notation :

B_1 et B_2 sont les bornes de la variable A $B_1 < B_2$

$E(A) = M \qquad E(A)^2 = Q^2$ (moments d'ordre 1 et 2)

$\sigma_A^2 = Q^2 - M^2$

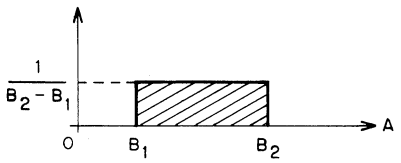
$f_k(A)$ est la densité de répartition de A

$$\int_{B_1}^{B_2} f_k(A) = 1 \qquad k = 1, 2, 3$$

Première famille :

$f_1(A)$ = constante

$$f_1(A) = \frac{1}{B_2 - B_1}$$



- Si les nombres sont tous arithmétiques : $0 \leq B_1 < B_2$
- Lorsque nous étudierons les opérations sur des nombres algébriques nous prendrons : $B_1 < 0 \qquad B_2 > 0$

$$M = \frac{B_1 + B_2}{2} \qquad Q^2 = \frac{B_1^2 + B_1 \cdot B_2 + B_2^2}{3}$$

Pour les lois de ξ, a, α, voir exemple page 5.

Deuxième famille :

$B_1 = b_1 \, 10^{\beta_1} \qquad 0,1 \leq b_1 < 1$

$B_2 = b_2 \, 10^{\beta_2} \qquad 0,1 \leq b_2 < 1 \qquad \xi = +1$

$A = a \, 10^{\alpha} \qquad 0,1 \leq a < 1$

Loi de α

$$\text{Prob.} \left\{ \alpha = \beta_1 + j \right\} = \frac{1}{\beta_2 - \beta_1 + 1} \qquad 0 \leq j \leq \beta_2 - \beta_1$$

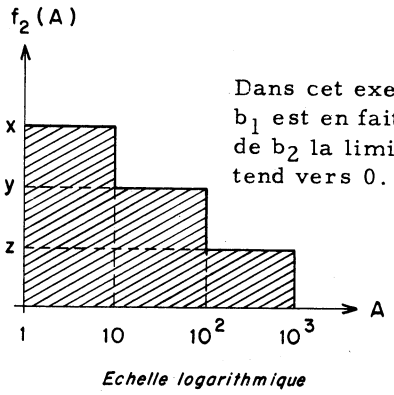
Loi de a

pour $\alpha = \beta_1$ loi uniforme entre (b_1 et 1)

pour $\alpha = \beta_1 + t$ loi uniforme entre (0, 1 et 1) $0 < t < \beta_2 - \beta_1$

pour $\alpha = \beta_2$ " " " (b_2 et 1)

Exemple : $B_1 = 1$ $B_2 = 10^3$



Dans cet exemple particulier, la valeur de b_1 est en fait la limite de $0,1 + \epsilon$ et celle de b_2 la limite de $1 - \epsilon$ lorsque ϵ positif tend vers 0.

$\beta_1 = 1$

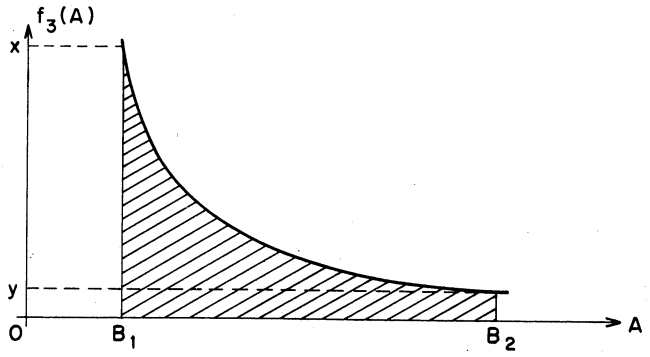
$\beta_2 = 3$

$x = \frac{1}{27}$ $y = \frac{1}{270}$ $z = \frac{1}{2700}$ $M = 203,5$ $Q^2 = 124.589$

Troisième famille :

$$f_3(A) = \frac{B_1 B_2}{B_2 - B_1} \quad \frac{1}{A^2}$$

a) $B_2 > B_1 > 0$



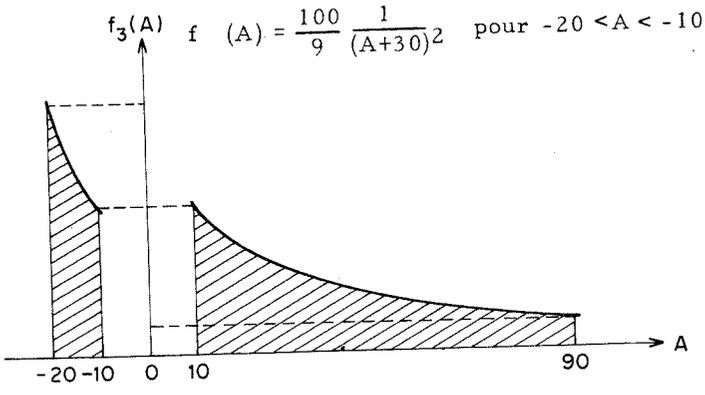
$$x = \frac{B_2}{B_1} \quad \frac{1}{B_2 - B_1} \qquad y = \frac{B_1}{B_2} \quad \frac{1}{B_2 - B_1}$$

$$M = B_1 B_2 \quad \frac{\log (B_2) - \log (B_1)}{B_2 - B_1} \qquad Q^2 = B_1 B_2$$

b) Lorsque les échantillons A_i ne sont pas tous positifs nous avons pris pour A la loi suivante :

$$f(A) = \frac{100}{9} \frac{1}{(A+10)^2} \quad \text{pour } 10 < A < 90$$

$$f(A) = \frac{100}{9} \frac{1}{(A+30)^2} \quad \text{pour } -20 < A < -10$$



$$M \# 5$$

$$Q^2 \# 386$$

2) LES 2 CLASSES DE NOMBRES

Nous définirons deux classes de nombre :

- la classe (I) étant l'ensemble de tous les nombres réels C qui, écrits sous la forme (1), sont tels que quel que soit $j > 0$ (j étant un nombre entier) $c^{p+j} = 0$ (nombre de chiffres significatifs $\leq p$)

- la classe (II) étant formée par l'ensemble des nombres réels C tels qu'il existe au moins une valeur de j pour laquelle $c^{p+j} \neq 0$ (nombre de chiffres significatifs $> p$)

3) ECHANTILLON ALEATOIRE L attaché à un nombre de la classe (II)

A tout nombre C de la classe (II) nous associerons un échantillon

aléatoire que nous noterons L (C) et qui sera défini par l'égalité suivante :

$$L(C) = \xi \cdot x_c \cdot 10^{\gamma - p}$$

Définition de x_c

x_c est un échantillon d'une variable aléatoire notée x dont la répartition est uniforme.

Nous examinerons 2 cas :

a) erreur de chute $\qquad 0 \leqslant x < 1$

$$E(x)^n = \int_0^1 t^n \, dt = \frac{1}{n+1}$$

b) erreur d'arrondi $\qquad -\frac{1}{2} \leqslant x < \frac{1}{2}$

$$E(x)^n = \int_{-0,5}^{+0,5} t^n \, dt = \frac{1 - (-1)^{n+1}}{(n+1)\, 2^{n+1}}$$

Remarques :

-L (C) peut s'écrire aussi $10^{-p} \frac{x_c}{c} C$

-L'application de L à tout nombre C de la classe (I) donnera, par définition, la valeur certaine zéro :

$$L(C) = 0 \qquad C \in (I)$$

Propriétés de L

1. $L(-A_k) = -L(A_k)$

11. Soit $h = 10^Z$ Z étant un nombre quelconque

$L(hC) = h\, L(C)$

iii. Etude de L $\left[C + L(C)\right]$ $\quad C \in (II)$

Par la suite, nous rencontrerons des expressions de cette forme :

$$L(C) = \xi_c \; x_c \; 10^{\gamma - p}$$

$$C + L(C) = \xi_c \; (c + x_c \; 10^{-p}) \; 10^{\gamma}$$

si $c + x_c \; 10^{-p} < 1$

$L(C)$ et L $\left[C + L(C)\right]$

seront 2 échantillons d'une même variable $\xi_c \; x \; 10^{\gamma - p}$

a) erreur de chute $\quad x_c < 1$

la probabilité d'avoir en général $\quad c < 1 - 10^{-p}$

est égale à $1 - 10^{-p}$

b) erreur d'arrondi $\quad x_c < \frac{1}{2}$

la probabilité d'avoir en général $\quad c < 1 - 0{,}5 \; 10^{-p}$

est égale à $1 - 0{,}5 \; 10^{-p}$

Résultats :

$$\text{Prob.} \left\{ E \left[L \left[C + L(C) \right] \right]^n = E \left[L(C) \right]^n \right\} > 1 - q$$

erreur de chute $\quad q = \frac{1}{10^p}$

erreur d'arrondi $\quad q = \frac{1}{2.10^p}$

Les mêmes résultats sont obtenus en étudiant des expression de la forme L $\left[C - L(C)\right]$

iv. Appelons S_k une somme arithmétique de k nombres.

" S'_k " algébrique "

" P_k un produit de k nombres .

En écrivant :

$$L(S_k) = 10^{-p} \frac{x_{s_k}}{s_k} S_k \qquad L(S'_k) = 10^{-p} \frac{x_{s'_k}}{s'_k} S'_k$$

$$L(P_k) = 10^{-p} \frac{x_{p_k}}{p_k} P_k$$

nous supposons que les nombres S_k , S'_k et P_k font partie de la classe (II)

a) Si l'un quelconque des nombres A_i , intervenant dans la somme arithmétique ou dans le produit, est un élément de la classe (II), nous considèrerons toujours S_k et P_k comme des éléments de la classe (II).

b) Si les nombres A_i (i = 1, 2, ... k) font tous partie de la classe (I) S_k , S'_k , P_k peuvent être soit de la classe (I) soit de la classe (II).

En général nous rencontrerons des sommes arithmétiques et surtout des produits qui seront des éléments de la classe (II) ; toutefois il faudra le vérifier avant de leur associer l'échantillon aléatoire L correspondant.

v. Soient: A_1 et A_2 ξ classe (II) avec $A_1 \neq A_2$

$$E\left[a_1 L(A_1) + a_2 L(A_2)\right] = E\left[s_2 L(S_2)\right]$$

en effet chacune de ces deux moyennes est égale à : $E\left[x\ 10^{-p}(A_1+A_2)\right]$

4) CORRECTION DE CALCUL

Notations :

Soient: Y la valeur numérique exacte de l'expression à calculer

$\tilde{Y}$ la valeur trouvée par le calculateur

L'erreur de calcul sur Y est $\tilde{Y} - Y$

En fait il sera plus commode d'utiliser la "correction de calcul" :

$$R(Y) = Y - \tilde{Y}$$

Hypothèse fondamentale sur la "correction de calcul"

Dans toute la théorie qui suivra nous considèrerons la "correction de calcul" sur un nombre C, R (C), comme l'échantillon aléatoire L (C) défini précédemment.

$$R(C) = L(C)$$

Prenons par exemple le cas où un nombre négatif ($C < 0$) subit une erreur de chute.

$$C = -(0, c^1\ c^2 \ldots c^p\ c^{p+1} \ldots c^m \ldots)\ 10^{\gamma}$$

$$\tilde{C} = -(0, c^1\ c^2 \ldots c^p)\ 10^{\gamma}$$

$$R(C) = C - \tilde{C}$$

$$R(C) = -10^{\gamma - p}\ (0, c^{p+1}\ c^{p+2} \ldots c^m \ldots)$$

Comparons à L (C)

$$L(C) = -10^{\gamma - p}\ (x_c)$$

Il est tout à fait raisonnable, et l'expérience le prouve, de faire correspondre à la suite de chiffres perdus ($c^{p+1} \ldots c^m$), que l'on ne connait en général pas, un nombre aléatoire (x_c) échantillon d'une variable dont la répartition est uniforme entre 0 et 1 (pour cet exemple).

Remarque :

Dans le cas très particulier où l'on est assuré que c est de la forme :

$$c = 0,c^1\ c^2 \ldots c^p \underbrace{0\ 0 \ldots 0}_{k}\ c^{p+k+1} \ldots c^m$$

c'est à dire $c^i = 0 \quad p < i \leq p + k$

l'hypothèse sur la correction de calcul devient :

$$R\ (C) = 10^{-k}\ L\ (C)$$

Propriétés de R

Soient $U_1\ U_2\ U_3$ ... des expressions représentant chacune un certain volume de calcul (qui nécessite la succession d'un certain nombre d'opérations).

$\widetilde{U}_i$ sera la valeur calculée par la machine électronique.

(élément de la classe (I)).

a) Correction de calcul sur une somme d'expressions U_i

Etude de la correction de calcul sur $U_1 + U_2$

Le calculateur exécutera tout d'abord les opérations relatives à l'expression U_1 , nous obtiendrons la valeur $\widetilde{U}_1$

Il exécutera ensuite celles relatives à U_2 , nous aurons la valeur $\widetilde{U}_2$

Il effectuera finalement la somme de U_1 et de U_2 dont le résultat pourra s'écrire $\widetilde{(\widetilde{U}_1 + \widetilde{U}_2)}$

$$R\ (U_1 + U_2) = U_1 + U_2 - \widetilde{(\widetilde{U}_1 + \widetilde{U}_2)}$$

$$R\ (U_1 + U_2) = U_1 + U_2 - \widetilde{U}_1 - \widetilde{U}_2 + (\widetilde{U}_1 + \widetilde{U}_2) - \widetilde{(\widetilde{U}_1 + \widetilde{U}_2)}$$

$$R\ (U_1 + U_2) = R\ (U_1) + R\ (U_2) + R\ (\widetilde{U}_1 + \widetilde{U}_2)$$

$\widetilde{U}_1 + \widetilde{U}_2$ représente un nombre.

Selon l'hypothèse fondamentale énoncée plus haut

$$R\ (\widetilde{U}_1 + \widetilde{U}_2) = L\ (\widetilde{U}_1 + \widetilde{U}_2)$$

En définitive :

$$R\ (U_1 + U_2) = R\ (U_1) + R\ (U_2) + L\ (U_1 + U_2)$$

Considérons la somme $U_1 + U_2 + U_3$

Le calculateur effectuera une première somme $\widetilde{(\widetilde{U}_1 + \widetilde{U}_2)}$ puis ajoutera $\widetilde{U}_3$ à la valeur trouvée ($\widetilde{U}_1$, $\widetilde{U}_2$, $\widetilde{U}_3$ auront été établies en temps utile).

$$R\ (U_1+U_2+U_3) = U_1+U_2+U_3 - \widetilde{(\widetilde{U}_1+\widetilde{U}_2)} + \widetilde{U}_3$$

$$R\ (U_1+U_2+U_3) = R\ (U_1)+ R\ (U_2) + R\ (U_1) + L\ (\widetilde{U}_1+\widetilde{U}_2) + L\ \left[\widetilde{(\widetilde{U}_1+\widetilde{U}_2)} + \widetilde{U}_3\right]$$

Généralisation : $(U_1 + U_2 + \ldots + U_n)$

Supposons que le calcul de la somme se fasse par adjonction successive des termes.

Tous les calculateurs électroniques sont en fait conçus pour réaliser de façon très simple ce procédé.

Soit V_i tel que $\begin{cases} V_{i+1} = \widetilde{V}_i + \widetilde{U}_i \\ V_2 = U_1 \end{cases}$

Nous obtiendrons alors la formule générale suivante :

$$R\ \left(\sum_{i=1}^{n} U_i\right) = \sum_{i=1}^{n} R\ (U_i) + \sum_{i=2}^{n} L\ (V_{i+1})$$

b) Correction de calcul sur un produit de n expressions U_i

Etude de la correction de calcul sur $U_1 . U_2$

$$R\ (U_1-U_2) = U_1 . U_2 - \widetilde{(\widetilde{U}_1 . \widetilde{U}_2)}$$

La décomposition est possible de 3 façons :

1) $R\ (U_1 . U_2) = U_1 R(U_2) + R(U_2)U_2 + L(\widetilde{U}_1 . \widetilde{U}_2) - R(U_1).R(U_2)$

2) $R(U_1.U_2) = \tilde{U}_1 R(U_2) + R(U_2)U_2 + L(\tilde{U}_1.\tilde{U}_2)$

3) $R(U_1.U_2) = U_1 R(U_2) + R(U_1)\tilde{U}_2 + L(\tilde{U}_1.\tilde{U}_2)$

$R(\tilde{U}_1.\tilde{U}_2) = L(\tilde{U}_1.\tilde{U}_2)$ $\quad$ $\tilde{U}_1.\tilde{U}_2$ est un nombre

La symétrie qui existe entre U_1 et U_2 se manifeste dans la première façon d'évaluer $R(U_1\ U_2)$

Correction de calcul sur $(U_1.U_2.U_3)$

$$R(U_1.U_2.U_3) = U_1.U_2 R(U_3) + R(U_1\ U_2)\tilde{U}_3 + L\left[\widetilde{(\tilde{U}_1.\tilde{U}_2).\tilde{U}_3}\right]$$

c'est à dire :

$$R(U_1.U_2.U_3) = U_1.U_2 R(U_3) + U_1 R(U_2)\tilde{U}_3 + R(U_1)\tilde{U}_2\tilde{U}_3 + L(\tilde{U}_1\ \tilde{U}_2).\tilde{U}_3 + L\left[\widetilde{(\tilde{U}_1.\tilde{U}_2).\tilde{U}_3}\right]$$

Généralisation sur un produit : $(U_1\ U_2 \ldots U_n)$

Nous supposerons à nouveau que : $\prod_{i=1}^{i=k} U_i$ est obtenu en formant le produit de $\prod_{i=1}^{i=k-1} U_i$ par U_k

pour $1 \leqslant k \leqslant n$

Soit W_i tel que $\begin{cases} W_{i+1} = \tilde{W}_i.\tilde{U}_i \\ W_2 = U_1 \end{cases}$

nous pourrons écrire :

$$R\left(\prod_{i=1}^{i=n} U_i\right) = \sum_{1}^{n} U_{i-n+1} \ldots U_{i-1}\ R(U_i)\ \tilde{U}_{i+1} \ldots \tilde{U}_{i+n-1} + \sum_{2}^{n} L(W_{i+1})\ \tilde{U}_{i+1} \ldots \tilde{U}_n$$

En adoptant la convention suivante :

$$U_i = 1 \text{ pour } \left[\begin{array}{l} i < 1 \\ i > n \end{array}\right.$$

PREMIERE PARTIE

ERREURS DE CALCUL DANS LES SOMMES ET PRODUITS

Nous étudierons ce problème :

- soit en établissant une borne stricte de l'erreur,
- soit en déterminant, grâce à la théorie des probabilités, une loi à laquelle devra satisfaire la "correction de calcul".

En fait nous essaierons d'en déterminer la valeur moyenne et l'écart type.

Ces deux points de vue se complètent : il sera souvent possible de trouver une borne de l'erreur mais, sauf dans certains cas exceptionnels, elle sera nettement supérieure à l'erreur véritable.
(d'autant plus que les opérations seront plus nombreuses).

Pour simplifier les calculs nous nous placerons tout d'abord dans le cas où les nombres A_i font tous partie de la classe (I) tandis que le résultat de la première opération, quelle qu'elle soit, est un nombre de la classe (II).
(Chapitres I, II, III)

CHAPITRE I

CORRECTION DE CALCUL SUR LE RESULTAT D'UNE SOMME ARITHMETIQUE DE n NOMBRES

Soit S_n cette somme

$$S_n = \sum_{i=1}^{i=n} A_i = s_n \, 10^{s_n}$$

Les expressions U_i utilisées précédemment sont alors réduites à des nombres A_i de la classe (I).

(le cas particulier étudié en remarque page 12, ne se présentera pas pour $n \leqslant 10^p$ en effet nous aurons alors $\alpha_{i+1} > s_i - p \quad 1 \leqslant i \leqslant n$)

Appliquons directement la formule générale relative à la correction de calcul sur une somme.

$$R(S_n) = \sum_{i=2}^{n} L(V_{i+1}) \quad \text{avec} \quad \begin{cases} V_{i+1} = \widetilde{V}_i + A_i \\ V_2 = A_1 \end{cases}$$

BORNE DE L'ERREUR

-Erreur de chute-

Quel que soit i $V_{i+1} \leqslant S_i$

$$\left| R(S_n) \right| \leqslant \sum_{i=2}^{n} \left| \frac{x_i}{s_i} \right| S_i \; 10^{-p}$$

$$0 \leqslant \frac{x_i}{s_i} < 10$$

Une première borne stricte de l'erreur sera donnée par :

$$S_i \leqslant S_n \qquad \boxed{R(S_n) < 10^{-p+1} S_n (n-1)}$$

Une seconde par :

$$S_i \leqslant iB_2 \qquad \boxed{R(S_n) < 10^{-p+1} B_2 \frac{n^2+n-2}{2}}$$

B_2 est la borne supérieure des A_i

- Erreur d'arrondi -

$$\left|\frac{x_i}{s_i}\right| < 5 \qquad (V_{i+1} \# S_i)$$

$S_i \leqslant S_n$ entraine $\boxed{|R(S_n)| < 10^{-p+1} S_n \frac{n-1}{2}}$

$S_i \leqslant i B_2$ entraine $\boxed{|R(S_n)| < 10^{-p+1} B_2 \frac{n^2+n-2}{4}}$

Dans chaque cas la plus petite de ces 2 bornes - celle que l'on utilisera - sera toujours beaucoup plus élevée que l'erreur véritable, et de ce fait n'offre qu'un intérêt relatif.

ETUDE STATISTIQUE

RAPPEL D'UN THEOREME FONDAMENTAL

Etant donné une variable aléatoire qui possède des moments d'ordre 1 et 2, la loi de probabilité de la somme centrée réduite des n échantillons aléatoires indépendants de cette variable, tend lorsque n augmente indéfiniment vers une loi normale (Laplace-Gauss) de densité f (x)

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

ETUDE DE LA VARIABLE ALEATOIRE (s_i)

Examinons l'ensemble de n expériences telles que la $k^{ième}$ par exemple consiste :

- à prélever i échantillons de la variable A
- à effectuer leur somme S_i^k et
- à ne prendre en considération que la valeur s_i^k correspondante, sachant que $S_i^k = s_i^k \ 10^{\varsigma_i^k}$

$$0,1 \leqslant s_i^k < 1$$

Nous noterons simplement par s_i et S_i les variables dont les échantillons sont respectivement :

$$(s_i^1 \ldots s_i^k \ldots s_i^m) \quad \text{et} \quad (S_i^1 \ldots S_i^k \ldots S_i^m)$$

a) Etude théorique.

$$E(S_n) = n M \qquad S_n = s_n \ 10^{\varsigma_n}$$

$$\sigma_{S_n} = \sqrt{n} \ \sigma_A$$

Posons : $E(s_n) = \mu_n$

La variable ς_n ne prend que des valeurs entières. Lorsque n est suffisamment élevé, ces valeurs (ς_n^k) deviennent quasi certaines.

Nous aurons en ce cas : $n M = 10^{\varsigma_n} \mu_n$

$$\sigma_{S_n} = 10^{\varsigma_n} \sigma_{s_n}$$

et $$\sigma_{s_n} = \frac{1}{\sqrt{n}} \ \frac{\sigma_A}{M}$$

Dans l'intervalle (0,1, 1) s_n converge en probabilité vers μ_n lorsque n augmente indéfiniment.

b) Etude pratique.

Après avoir fait des expériences d'ordre statistiques sur les variables aléatoires (s_i), nous constatons que, pour des lois différentes de A, la courbe de répartition de s_{10} a déjà une allure tout à fait gaussienne.

Les tableaux A_1 A_2 montrent que pour une loi uniforme de A cette constatation est nette dès la variable (s_5). Les bornes prises sont 0 et 10^3.

L'évolution vers une loi de type gaussien à faibles écarts est légèrement plus lente pour la loi de probabilité dont les bornes sont 1 et 10^3 prise dans la 2ème famille : tableaux B_1 B_2 ou dans la 3ème famille : tableaux $C_1 \ldots C_6$

$i > 10$ entraîne $s_i^k \neq \mu_i$

ETUDE DE L'ECHANTILLON ALEATOIRE L (V_{i+1})

$$V_{i+1} = \widetilde{V}_i + A_i = V_i + A_i - L(V_i)$$

Posons : $V_i = v_i \ 10^{\phi_i}$

$V_i + A_i = w_i \ 10^{\psi_i}$

2 cas sont possibles :

$\psi_i = \phi_i$

$\psi_i = \phi_i + 1$

Notons K_i la valeur $10^{\phi_i - \psi_i}$

L (V_i) et K_i L $(V_i + A_i)$ sont alors 2 échantillons d'une même loi uniforme.

L'étude de L (V_{i+1}) revient à celle de :

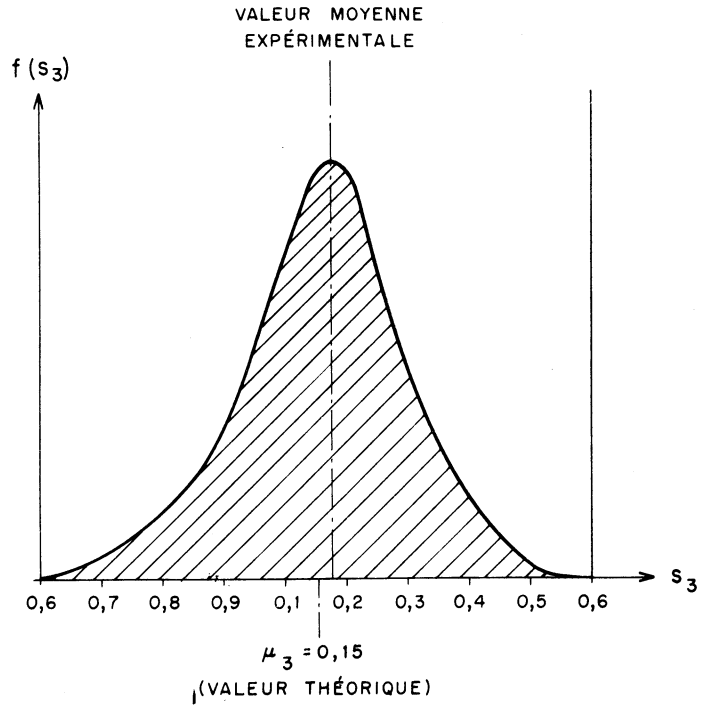
$$L\left[V_i + A_i - K_i L(V_i + A_i)\right]$$

Appliquons la propriété n° 3 de L

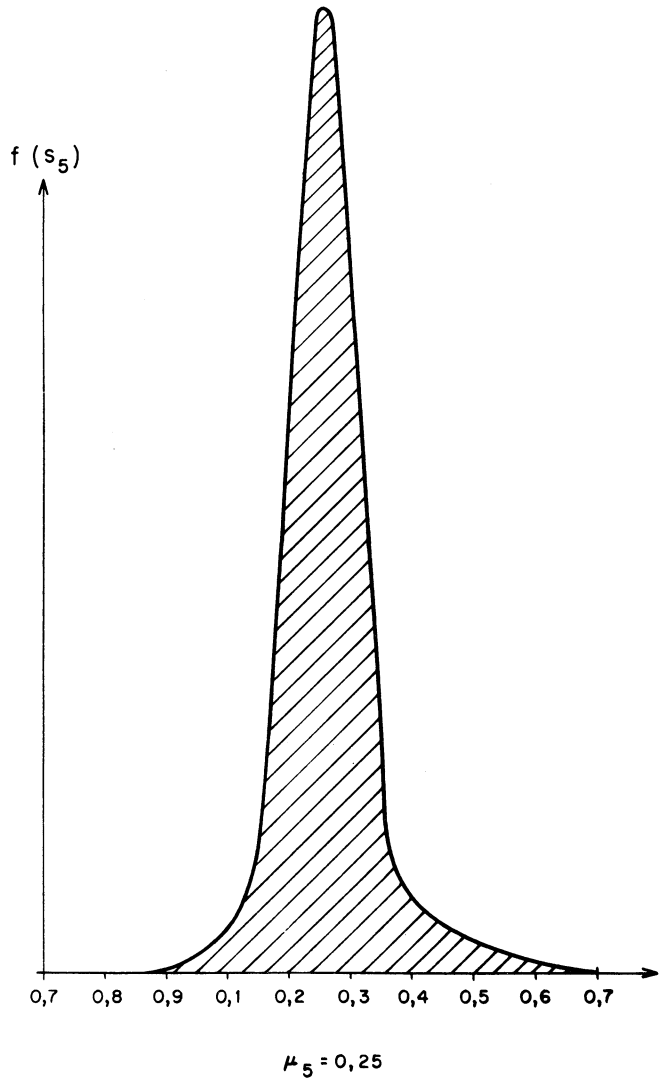
$$\text{Prob} \quad E\left\{\left[L(V_{i+1})\right]^n = E\left[L(V_i + A_i)\right]^n\right\} > 1 - q$$

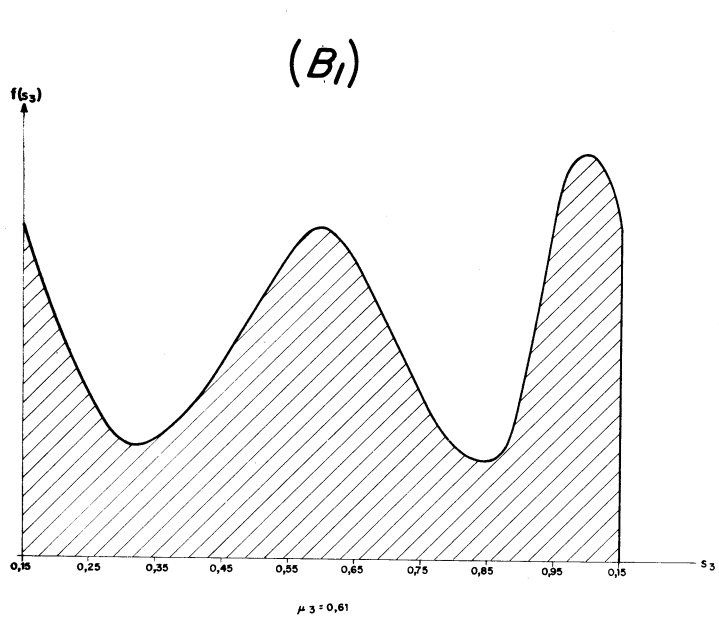
- erreur de chute $q = K_i \ 10^{-p}$
- erreur d'arrondi $q = 0{,}5 \ K_i \ 10^{-p}$

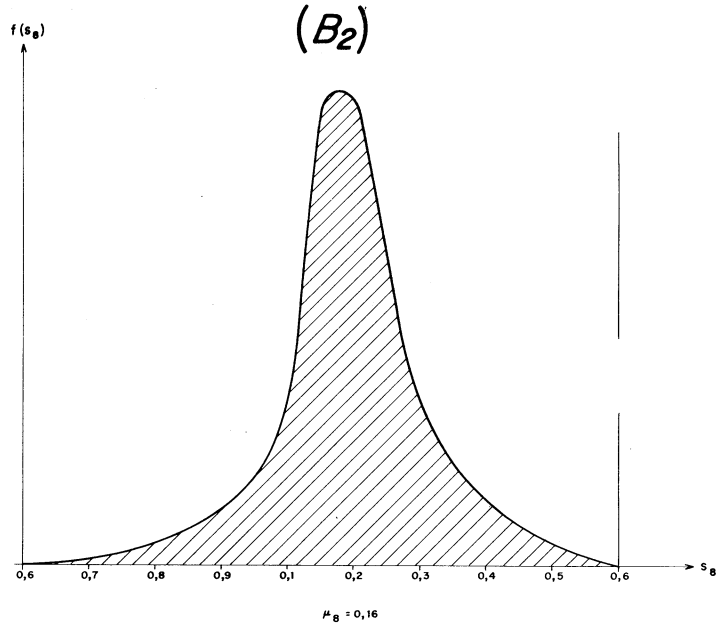
(A_1)

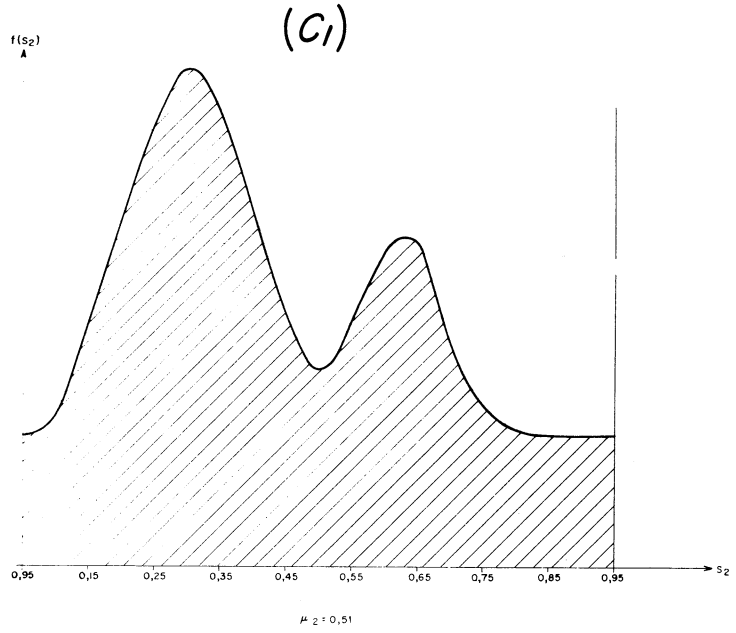

VALEUR MOYENNE
EXPÉRIMENTALE
f (s3)
s3
0,6 0,7 0,8 0,9 0,1 0,2 0,3 0,4 0,5 0,6
μ3 = 0,15
(VALEUR THÉORIQUE)

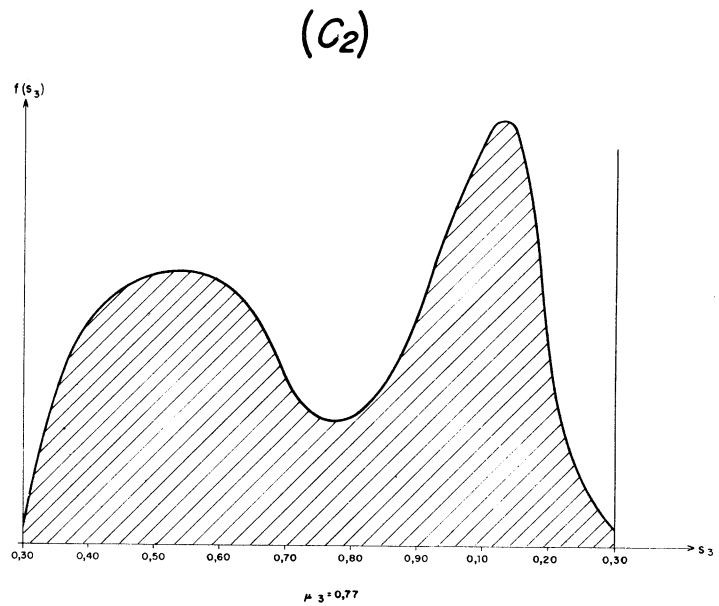
(A_2)


f (s5)
0,7 0,8 0,9 0,1 0,2 0,3 0,4 0,5 0,6 0,7
μ5 = 0,25

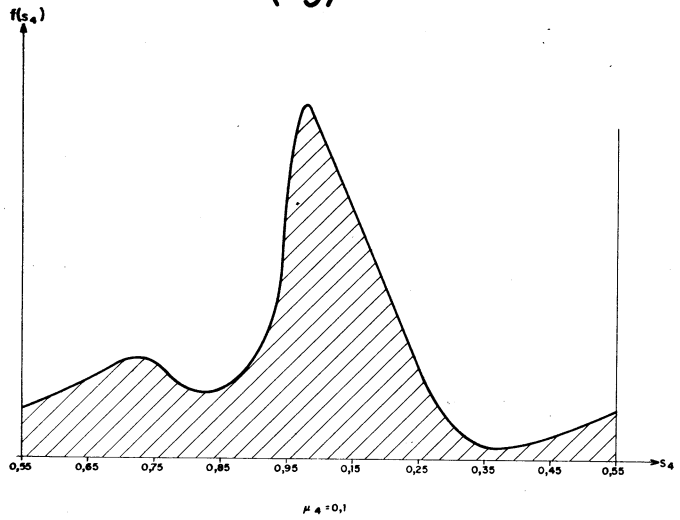

(B1)
f(s3)
0,15
0,25
0,35
0,45
0,55
0,65
0,75
0,85
0,95
0,15
s3
μ 3 = 0,61


(B2)
f(s8)
0,6
0,7
0,8
0,9
0,1
0,2
0,3
0,4
0,5
0,6
s8
μ8 = 0,16


(C1)
f(s2)
0,95
0,15
0,25
0,35
0,45
0,55
0,65
0,75
0,85
0,95
s2
μ 2 = 0,51

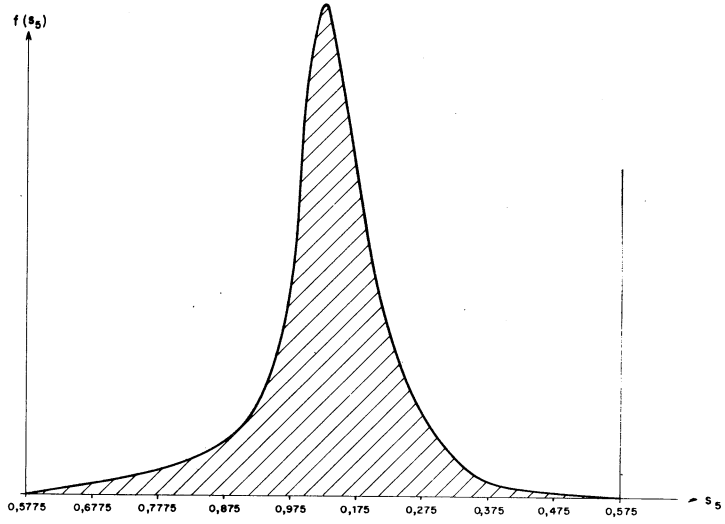

(C2)
f(s3)
0,30
0,40
0,50
0,60
0,70
0,80
0,90
0,10
0,20
0,30
s3
μ 3 = 0,77

(C3)

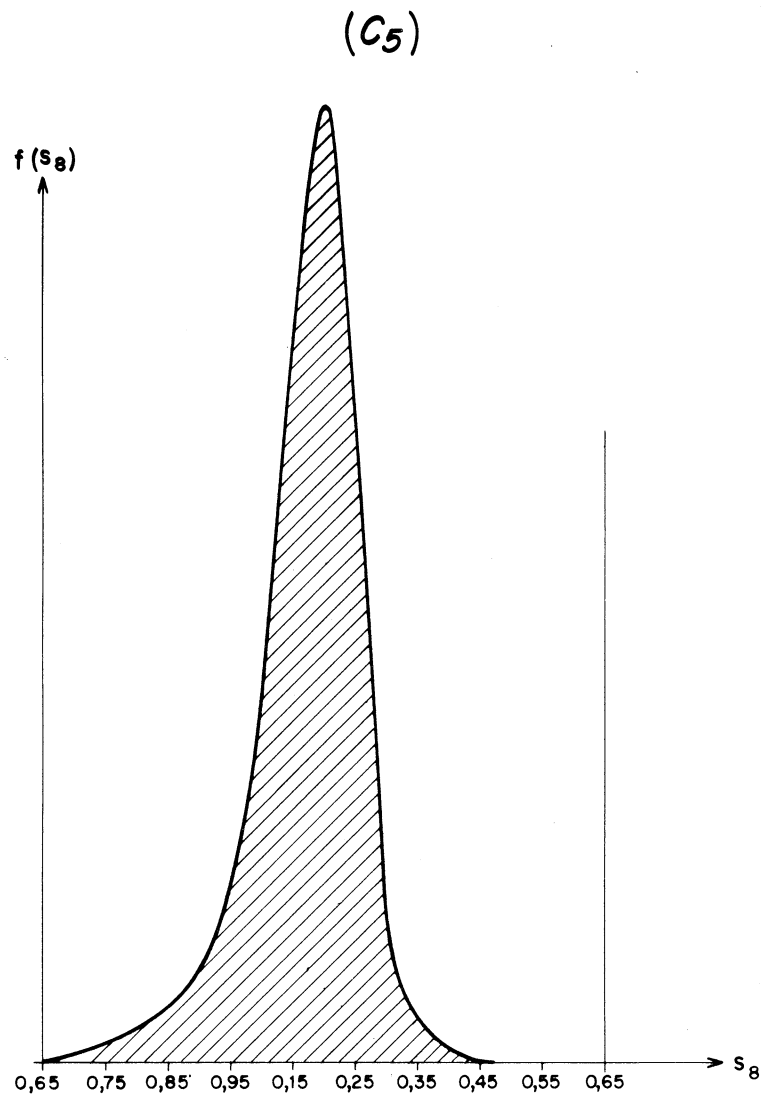


$\mu_4 = 0,1$

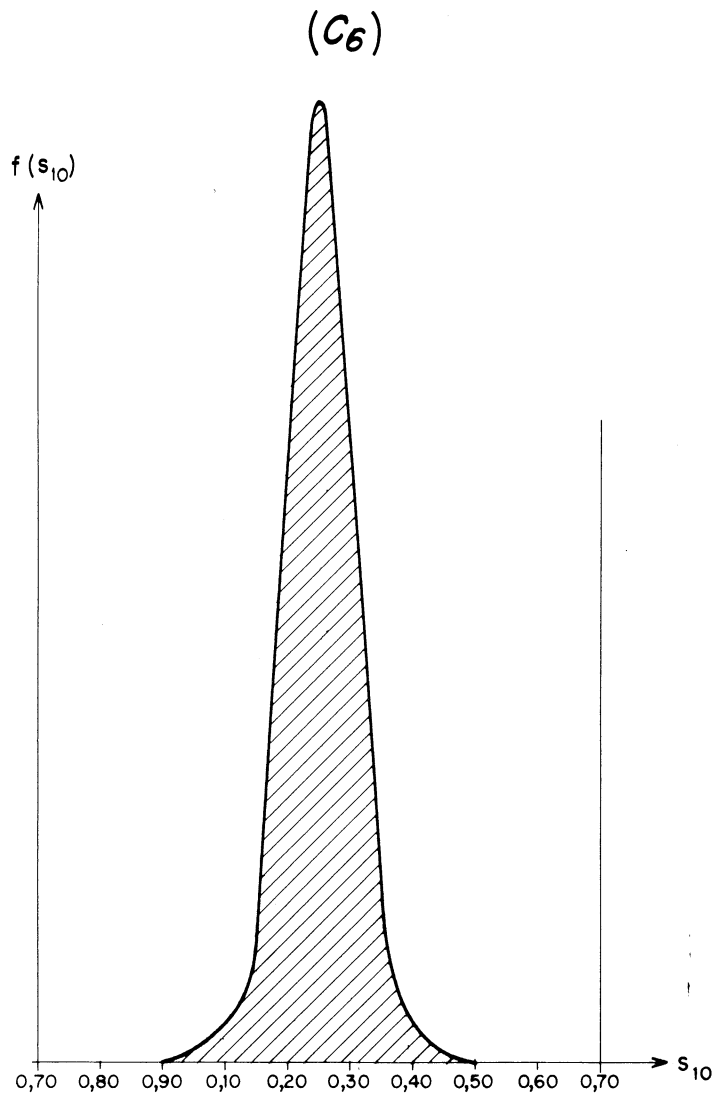
(C4)



$\mu_5 = 0,12775$



$\mu_8 = 0,204$



$\mu_{10} = 0,255$

En appliquant (i-1) fois cette même propriété on obtient finalement :

$$\text{Prob.} \quad E\left\{ \left[L\,(V_{i+1}) \right]^n = E \left[L\,(S_i) \right]^n \right\} \gg 1 - iq$$

Conséquences

$$L\,(V_{i+1}) \;\#\; L\,(S_i) \qquad (I_1)$$

$$\sum_{i=2}^{n} L\,(V_{i+1}) \;\#\; \sum_{i=2}^{n} L\,(S_i) \qquad (I_2)$$

Remarque

Si nq est de l'ordre de 1 (ou supérieur à 1) l'utilisation des formules établies ultérieurement est contestable : $L\,(V_{i+1})$ et $L\,(S_i)$ ne sont pas toujours 2 échantillons d'une même loi uniforme.

Prenons p = 9 par exemple les approximations (I_1) et (I_2) seront certainement bonnes si les sommes considérées comportent moins de 10^8 termes.

$$R\,(S_n) = \sum_{i=2}^{n} L\,(S_i)$$

$$L\,(S_i) = 10^{-p}\, x_i\, 10^{s_i} \qquad (I_3)$$

$$L\,(S_i) = 10^{-p}\, \frac{x_i}{s_i}\, S_i \qquad (I_4)$$

Nous évaluerons à présent la valeur moyenne et l'écart type de $R\,(S_n)$

Il est possible notamment de procéder de 2 façons :

- soit en utilisant la formule (I_3) : théorie (a)
- soit en utilisant la formule (I_4) : théorie (b)

THEORIE (a)

$$R(S_n) = 10^{-p} \sum_{i=2}^{n} x_i \, 10^{\varsigma_i}$$

Quelle que soit la loi de probabilité de A, quel que soit n ($n < 0,1\,q$) les valeurs ς_i ne sont pas indépendantes entre elles.

En effet l'ensemble des éléments de la forme $x_i \, 10^{\varsigma_i}$ se divise nécessairement en sous ensembles qui contiennent, chacun, les mêmes puissances de 10.

POSONS :

$$\begin{array}{ccccccc}
\varsigma_2 & = & \varsigma_3 & = \cdots\cdots = & \varsigma_{l_1+1} & = & \lambda + 1 \\
\varsigma_{l_1+2} & = & \varsigma_{l_1+3} & = \cdots\cdots = & \varsigma_{l_1+l_2+1} & = & \lambda + 2 \\
\vdots & & \vdots & & \vdots & & \\
\varsigma_{l_1+\cdots+l_{j-1}+2} & = & \varsigma_{l_1+\cdots+l_{j-1}+3} & = \cdots\cdots = & \varsigma_{l_1+\cdots+l_j+1} & = & \lambda + j \\
\vdots & & \vdots & & \vdots & & \\
\varsigma_{l_1+\cdots+l_{k-1}+2} & = & \varsigma_{l_1+\cdots+l_{k-1}+3} & = \cdots\cdots = & \varsigma_n & = & \lambda + k
\end{array}$$

l'indice i varie de 2 à n et l'indice j varie de 1 à k

Remarques

$$n = l_1 + l_2 + \quad + l_k + 1$$

$$l_j = 0 \quad \text{pour} \quad \begin{cases} j < 1 \\ j > k \end{cases}$$

l_j valeur de ς_i seront égales à $\lambda + j$ (le $j^{\text{ième}}$ sous ensemble se compose de l_j éléments).

Posons : $l_1 + l_2 + \cdots + l_{j-1} = \delta_j - 1$

$\delta_1 = 1$

$\delta_{j+1} = \delta_j + l_j \qquad (1 < j < k)$

$\delta_{k+1} = n$

Soit r_j un nouvel indice qui variera de 1 à l_j

La "Correction de Calcul" sur une somme arithmétique s'écrira sous la forme suivante :

$$R(S_n) = 10^{-p} \sum_i x_i \, 10^{s_i} = 10^{-p} \sum_j \left[10^{\lambda+j} \sum_{r_j} x_{(\delta_j + r_j)} \right]$$

DETERMINATION DE $(l_1 \ldots l_j \ldots l_k)$

1) k ne sera jamais très élevé.

Si nous effectuons une somme de 10^N termes nous aurons :

$N \leq k \leq N+1$ avec en général $N < p-1$

2) Hypothèse

Quels que soient les n échantillons A_i prélevés dans la répartition donnée, nous trouverons toujours les mêmes valeurs de l_j

3) Evaluation de δ_j

$$\delta_{j+1} \cdot M = \mu_{\delta_{j+1}} \, 10^{\lambda+j} < 10^{\lambda+j}$$

$$(\delta_{j+1} + 1) \cdot M = \mu_{(\delta_{j+1}+1)} \, 10^{\lambda+j+1} > 10^{\lambda+j}$$

Nous évaluerons δ_{j+1} par la relation : $\delta_{j+1} \cdot M + (\delta_{j+1} + 1) M = 2 \cdot 10^{\lambda+j}$

soit : $\delta_{j+1} = \frac{10^{\lambda+j}}{M} - \frac{1}{2}$

4) Détermination de l_j

$j = 1$ $\qquad \delta_2 - 1 = l_1 = \dfrac{10^{\lambda+1}}{M} - \dfrac{3}{2}$

$1 < j < k$ $\qquad \delta_{j+1} - \delta_j = l_j = \dfrac{10^{\lambda+j}}{M} - \dfrac{10^{\lambda+j-1}}{M}$

$$l_j = 9\,\frac{10^{\lambda+j-1}}{M}$$

$j = k$ $\qquad n - \delta_k = l_k = \dfrac{10^{\lambda+k-1}}{M}(10\mu_n - 1) + \dfrac{1}{2}$

$$l_j = n\,10^{j-1-k}\,C_j \quad \text{avec} \quad \begin{cases} C_j = \dfrac{10 - 7{,}5\,\mu_2}{\mu_n} & (j = 1) \\ C_j = \dfrac{9}{\mu_n} & (1 < j < k) \\ C_j = \dfrac{10\mu_n - 1 + \frac{5\mu_n}{n}}{\mu_n} & (j = k) \end{cases}$$

$\left(\frac{5\mu_n}{n}\right.$ terme négligeable$\left.\right)$

VALEUR MOYENNE DE R (S_n)

$$E(x_{\delta_j + r_j}) = E(x)$$

$$E(R) = E(x)\,10^{\lambda-p} \sum_{j=1}^{k} 10^j\, l_j$$

$$E(R) = n\,E(x)\,10^{\lambda-p}\,10^{1+k}\left(C_k + 10^{-2}\,C_{k-1} + \cdots + 10^{-2(k-1)}\,C_1\right)$$

Pour $k > 2$ la somme $10^{-4}\,C_{k-2} + \cdots + 10^{-2(k-1)}\,C_1$ est négligeable

devant $C_k + 10^{-2}\,C_{k-1}$

en effet $C_k + 10^{-2}\,C_{k-1} = 10 - \dfrac{0{,}911}{\mu_n} > 0{,}9$

et $10^{-4}\,C_{k-2} + \cdots + 10^{-2(k-1)}\,C_1 \quad \dfrac{10^{-4}}{\mu_n} < 0{,}001$

$$\boxed{E(R) = n\,E(x)\,10^{\lambda+k-p}\left[1 - \frac{0{,}0911}{\mu_n}\right]}$$

$$M_n = \mu_n\ 10^{\lambda+k}$$

Nous aurons sous une autre forme

$$E(R) = 10^{-p}\, M\, \frac{n^2}{2} \times \frac{2\, E(x)\, (\mu_n - 0{,}0911)}{\mu_n{}^2}$$

ECART TYPE DE R (S_n)

$$1 \leqslant j,t \leqslant k$$

$$1 \leqslant r_j \leqslant l_j$$

$$1 \leqslant s_t \leqslant l_t$$

$$R^2 = 10^{2\lambda-2p}\left\{\sum_j 10^{2j}\left[\sum_{r_j}(x_{\delta_j+r_j})^2+\sum_{r_j}\sum_{s_t\neq r_j}(x_{\delta_j+r_j})(x_{\delta_j+s_t})\right]+\sum_j\sum_{t\neq j}10^{j+t}\left[\sum_{r_j}\sum_{s_t}(x_{\delta_j+r_j})(x_{\delta_t+s_t})\right]\right\}$$

$$\begin{cases} E(x_{\delta_j+r_j})^2 = E(x)^2 \\ E(x_{\delta_j+r_j})(x_{\delta_t+s_t}) = \left[E(x)\right]^2 \quad \text{pour} \quad \delta_j + r_j \neq \delta_t + s_t \end{cases}$$

$$E(R)^2 = 10^{2\lambda-2p}\left\{E(x)^2\sum_j l_j\, 10^{2j} + \left[E(x)\right]^2\sum_j l_j(l_j-1)\,10^{2j} + \left[E(x)\right]^2\sum_j\sum_{t\neq j}10^{j+t}\, l_t\, l_j\right\}$$

$$\left[E(R)\right]^2 = 10^{2\lambda-2p}\left\{\left[E(x)\right]^2\left(\sum_j l_j^2\, 10^{2j} + \sum_j\sum_{t\neq j} l_j \cdot l_t\, 10^{t+j}\right)\right\}$$

$$\sigma_R^{\,2} = 10^{2\lambda-2p}\quad \sigma_{(x)}^{\,2}\sum_j l_j\, 10^{2j}$$

$$\sigma_R^{\,2} = 10^{2\lambda-2p}\quad \sigma_{(x)}^{\,2}\, n\, 10^{2k-1}\left(C_k + 10^{-3}C_{k-1} + \cdots + 10^{-3(k-1)}C_1\right)$$

Pour $k > 2$ la somme $10^{-6}\, C_{k-2} + \quad + 10^{-3(k-1)}\, C_1$ est négligeable devant $C_k + 10^{-3}\, C_{k-1}$

$$\boxed{\sigma_R^{\,2} = n\ \sigma_{(x)}^{\,2}\ 10^{2(\lambda+k-p)}\left[1 - \frac{0{,}0991}{\mu_n}\right]}$$

Dans le cas de l'erreur de chute, les résultats peuvent s'écrire sous la forme suivante :

$$\begin{cases} E(R) = 10^{-p} M \frac{n^2}{2} \left(\frac{\mu_n - 0,0911}{\mu_n^2} \right) \\ \sigma_R^2 = 10^{-2p} M^2 \frac{n^3}{12} \left(\frac{1 - 0,0991}{\mu_n^3} \right) \end{cases} \qquad (I_5)$$

En effet : $E(x) = \frac{1}{2}$ $\sigma_x^2 = \frac{1}{12}$

THEORIE (b)

$$R(S_n) = 10^{-p} \sum_{i=2}^{n} \frac{x_i}{s_i} (A_1 + A_2 + \cdots + A_i)$$

Nous avons vue que nous ne modifierons que très faiblement le problème en remplaçant chaque échantillon s_i par la valeur μ_i correspondante.

$$R(S_n) \neq 10^{-p} \sum_{i=2}^{n} \frac{x_i}{\mu_i} (A_1 + A_2 + \cdots + A_i)$$

HYPOTHESES DE TRAVAIL

Prenons comme hypothèse de travail :

1) Que chaque valeur μ_i est, elle même, un échantillon d'une certaine variable aléatoire que nous appellerons μ comprise entre 0, 1 et 1.

2) Que les variables x ; μ ; et A sont indépendantes entre elles.

LOI DE PROBABILITE DE μ

$S_i = s_i \, 10^{s_i}$ $\qquad$ i $M = \mu_i \, 10^{s_i}$ $\qquad$ $\mu_i = E(s_i)$

Une valeur et une seule de i , soit l , sera déterminée par :

$$\begin{cases} s_l = s_n - 1 \\ s_{l+1} = s_n \end{cases} \qquad n \; M = \mu_n \, 10^{s_n}$$

Ceci revient à écrire $\begin{cases} l. \; M < 10^{s_n - 1} \\ (l+1) \; M > 10^{s_n - 1} \end{cases}$

$$l \neq \frac{10^{s_n - 1}}{M} = \frac{n}{10 \, \mu_n}$$

a) Considérons l'ensemble des valeurs discrètes ($\mu_2 \ldots \mu_l$)

Chaque valeur μ_i ($2 \leq i \leq l$) se comporte comme un échantillon d'une v.a. continue μ_I dont la loi serait uniforme entre 0,1 et 1.

($\mu_2 \ldots \mu_l$ constitue un échantillonnage presque parfait de la variable μ_I)

b) Examinons par contre l'ensemble des valeurs discrètes ($\mu_{l+1} \ldots \mu_n$)

Chacune d'entre elles ($l+1 \leq i \leq n$) peut être assimilée à un échantillon d'une v.a. continue μ_{II} dont la loi serait uniforme entre 0,1 et μ_n

Exemple :

Supposons que l'on ait à effectuer une somme de 678 nombres. La moyenne théorique est 678 x 4,285 = 2905,25

Ici $l = 233$

μ_2	μ_3	$\mu_4 \ldots\ldots\ldots \mu_{23}$	$\mu_{24} \ldots\ldots\ldots\ldots \mu_{233}$
0,857	0,12855	0,1694 0,985	0,10284 0,998405

La loi de μ_I , dont les échantillons sont $\mu_2 \ldots\ldots \mu_{233}$ est uniforme entre 0,1 et 1.

μ_{234}	μ_{235}	μ_{678}
0,100271	0,1006995	0,290525

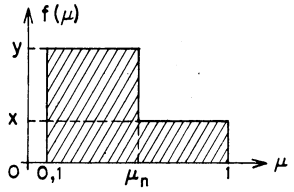
La loi de μ_{II} est uniforme entre 0,1 et 0,290525 (μ_n)

c) Loi de μ

La loi de μ tenant compte de μ_I et μ_{II} sera représentée par le schéma suivant :

$$\frac{(1-0,1)\,x}{1} = \frac{(\mu_n - 0,1)\,(y-x)}{n-1} = \frac{1}{n}$$

$$x = \frac{1}{9\,\mu_n} \qquad y = \frac{10}{9\,\mu_n}$$



$$E\left(\frac{I}{\mu}\right)^n = y \int_{0,1}^{n} \left(\frac{I}{t}\right)^n dt + x \int_{\mu_n}^{1} \left(\frac{I}{t}\right)^n dt$$

On trouve

$$\boxed{E\left(\frac{1}{\mu}\right) = \frac{23 + 9\ \mathrm{Log}\ \mu_n}{9\,\mu_n}} \qquad \boxed{E\left(\frac{1}{\mu}\right)^2 = \frac{11\mu_n - 1}{\mu_n^2}}$$

Soient
$$\begin{cases} m = E\left(\frac{x}{\mu}\right) = E(x)\ E\left(\frac{1}{\mu}\right) \\ q^2 = E\left(\frac{x}{\mu}\right)^2 = E(x)^2\ E\left(\frac{1}{\mu}\right)^2 \end{cases}$$

VALEUR MOYENNE DE R (S_n)

$$E(R) = 10^{-p} \sum_{i=2}^{n} E(x_i)\ E\left(\frac{1}{\mu_i}\right)\ E(A_1 + A_2 + \cdots + A_i)$$

$$\begin{cases} E(x_i) = E(x) \\ E\left(\frac{1}{\mu_i}\right) = E\left(\frac{1}{\mu}\right) \\ E(S_i) = i\,M \end{cases}$$

$$E(R) = 10^{-p}\ m.\ M.\ \sum_{i=2}^{n} i$$

$$\boxed{E(R) = 10^{-p}\ m.\ M\ \frac{n^2 + n - 2}{2}}$$

ECART-TYPE de R (S_n)

$$2 \leqslant k,\ l \leqslant n$$

$$R^2 = 10^{-2p}\left[\sum_k \left(\frac{x_k}{\mu_k}\right)^2 (S_k)^2 + \sum_k \sum_{l \neq k} \left(\frac{x_k}{\mu_k}\right)\left(\frac{x_l}{\mu_l}\right)(S_k)(S_l)\right]$$

$$E(R)^2 = 10^{-2p}\left[q^2 \sum_k \left(Q^2 k + M^2 k(k-1)\right) + m^2 \sum_k \left[\sum_{l<k}\left(Q^2 l + (k-1) l M^2\right) + \sum_{l>k}\left(Q^2 k + k(l-1) M^2\right)\right]\right]$$

$$\left[E(R)\right]^2 = 10^{-2p}\, m^2\, M^2 \left(\sum_k k^2 + \sum_k \sum_{l \neq k} kl\right)$$

$$\sigma_R^{\,2} = 10^{-2p}\left[C_1\, q^2\, Q^2 + C_2\, q^2\, M^2 + C_3\, m^2\, Q^2 - C_4\, m^2\, M^2\right]$$

$$C_1 = \sum_k k = \frac{1}{2}(n^2 + n - 2)$$

$$C_2 = \sum_k k(k-1) = \frac{1}{3}(n^3 - n)$$

$$C_3 = \sum_k \left[\sum_{l<k} l + \sum_{l>k} k\right] = \frac{1}{3}(n^3 - 7n)$$

$$C_4 = \sum_k \left[k^2 + \sum_{l<k} l + \sum_{l>k} k\right] = \frac{1}{6}(4n^3 + 3n^2 - 13n - 6)$$

Pour n assez grand $n > 50$ les résultats sont les suivants :

$$\boxed{\sigma_R^{\,2} = 10^{-2p}\,\frac{n^3}{3}\left[M^2 (q^2 - m^2) + m^2 (Q^2 - M^2)\right]} \qquad I_6$$

et $$E(R) = 10^{-p}\, M\, \frac{n^2}{2}\, m$$

ETUDE EXPERIMENTALE

Elle a été faite dans le cas de l'erreur de chute. L'erreur d'arrondi présente moins d'intérêt, quant à la vérification de la formule donnant m, puisque dans ce cas m est nul.

Nous avons effectué k fois les 2 sommes S_n et $\widetilde{S_n}$ en prenant, à chaque fois, n échantillons différents d'une même loi f (A).

$$R_j = S_n^j - \widetilde{S_n^j}$$

Les k "corrections de calcul" obtenus sont : $R_1, R_2, \ldots\ldots, R_k$

Leur valeur moyenne est : $\frac{1}{k} \sum_{j=1}^{k} R_j$

et leur écart type $\frac{1}{k} \sum_{j=1}^{k} (R_j)^2 - \left[\frac{1}{k} \sum_{j=1}^{k} \right]^2$

1) VALEURS MOYENNES

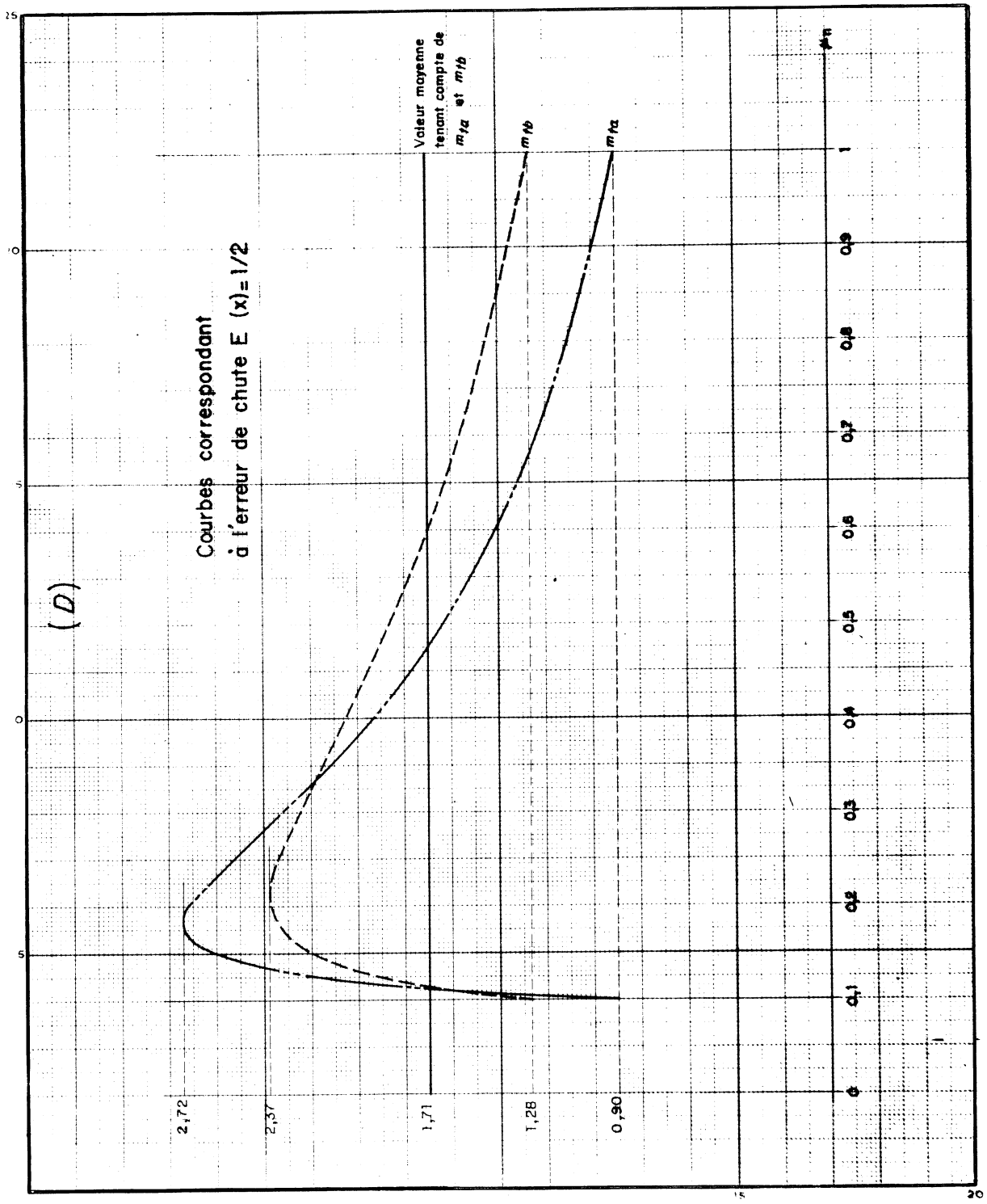
$$E(R) = 10^{-p} M \frac{n^2}{2} . m$$

La théorie (a) donne : $m = \frac{\mu_n - 0,0911}{\mu_n^2}$ que nous noterons m_{ta}

La théorie (b) donne : $m = \frac{23 + 9 \text{ Log } \mu_n}{18 \mu_n}$ que nous noterons m_{tb}

Le tableau D représente m_{ta} et m_{tb} en fonction de μ_n

Nous comparerons m_{ta} et m_{tb} à la valeur m_{pr} trouvée lors des expériences. *Voir page 39*


Courbes correspondant
à l'erreur de chute E (x)= 1/2
(D)
Valeur moyenne
tenant compte de
m_{1a} et m_{1b}
m_{1b}
m_{1a}
2,72
2,37
1,71
1,28
0,90
μ_n
1
0,9
0,8
0,7
0,6
0,5
0,4
0,3
0,2
0,1
0
25
20
15
10
5
0
5
10
15
20

FAMILLE	n	p	k	B_1	B_2	m_{ta}	m_{pr}	m_{tb}	r_{ta}	r_{pr}	r_{tb}
I	1000	5	20	0	10^3	1,64	1,69	1,86	0,0332	0,059	0,0692
II	137	5	20	1	10^3	2,4	2,29	2,27	0,15	0,2	0,29
II	826	5	20	1	10^3	2,68	2,53	2,29	0,12	0,10	0,14
III	90	4	10	10	10^2	2,6	2,55	2,35	0,021	0,032	0,033
III	800	4	10	10	10^2	2,7	2,3	2,36	0,008	0,0137	0,012
III	800	5	10	10	10^2	2,7	2,6	2,36	0,008	0,014	0,012

2) ECARTS TYPE

Nous appellerons

$$\sigma^2_{R_{ta}} = 10^{-2p} \frac{n^3}{3} \quad \frac{M^2 (\mu_n - 0,0991)}{4 \mu_n^3}$$

$$\sigma^2_{R_{tb}} = 10^{-2p} \frac{n^3}{3} \left[M^2 (q^2 - m^2) + m^2 (Q^2 - M^2) \right]$$

$$\sigma^2_{R_p} = \frac{1}{k} \sum_{j=1}^{k} (R_j)^2 - \left[\frac{1}{k} \sum_{j=1}^{k} R_j \right]^2$$

La comparaison des formules donnant E (R) ayant été faite sur le coefficient m, dont on cherche le plus possible à préciser la valeur, nous comparerons à présent r_{ta} et r_{tb} à r_{pr}. *Voir page 39*

$$r_{ta} = \frac{\sigma_{R_{ta}}}{10^{-p} M \frac{n^2}{2}} \qquad r_{tb} = \frac{\sigma_{R_{tb}}}{10^{-p} M \frac{n^2}{2}} \qquad \frac{\sigma_{R_p}}{10^{-p} M \frac{n^2}{2}} = r_{tr}$$

CONCLUSIONS

1) Les valeurs moyennes de la "correction de calcul" sur des sommes arithmétiques obtenues en pratique, dans les différents cas traités, sont proches de celles prévues par les deux théories.

Toutefois la première théorie exposée (théorie (a)) donne peut être des résultats légèrement meilleurs.

2) Pour les écarts type les résultats sont en général moins satisfaisants.

La théorie (b) se rapproche le plus de la réalité car elle tient compte de l'écart type de la variable A.

$\sigma_{R_{ta}}$ et $\sigma_{R_{tb}}$ donnent plutôt un ordre de grandeur qu'une évaluation exacte de l'écart type réel de R (S_n).

Lors de l'utilisation des formules il serait bon de prendre un seuil de signification suffisamment élevé.

3) On peut donner des formules plus générales en prenant :

- pour m la moyenne de toutes les valeurs m_{ta} et m_{tb} possibles.
- pour q la moyenne des valeurs q (μ_n) - théorie (b) -

Notons λ le rapport $\frac{Q}{M}$ (caractéristique de la répartition des nombres A_i). On obtient :

- Erreur de chute -

$$\begin{cases} E(R) = 0{,}856\ 10^{-p}\ S\ n \\ \sigma_R = 10^{-p}\ S \sqrt{n}\ \lambda \end{cases}$$

- Erreur d'arrondi -

$$\begin{cases} E(R) = 0 \\ \sigma_R = 0{,}53\ 10^{-p}\ S \sqrt{n} \left(\sqrt{1 + \frac{3\lambda^2}{2n}} \right) \end{cases}$$

Remarque

Les formules des écarts type données par la théorie (a) sont les mêmes avec $\lambda = 1$.

CHAPITRE II

CORRECTION DE CALCUL SUR UNE SOMME ALGEBRIQUE DE n NOMBRES

Soit S_n la somme $\sum_{i=2}^{i=n} A_i$ $\qquad A_i = \xi_{a_i}\ a_i\ 10^{a_i}$

La formule générale donnera à nouveau

$$R(S_n) = \sum_{2}^{n} L(V_{i+1})$$

avec les conventions $\begin{cases} V_{i+1} = \widetilde{V_i} + A_i \\ V_2 = A_1 \end{cases}$

BORNE DE L'ERREUR

Soit B le maximum de $|B_1|$ et de $|B_2|$ (bornes de la variable A)

Nous aurons quel que soit i

$$|A_i| \leq B. \qquad |V_{i+1}| \leq i\,B$$

$$|R(S_n)| \leq 10^{-p} \sum_{i=2}^{n} \left|\frac{x_i}{s_i}\right| i\,B$$

Une borne stricte de l'erreur sera donnée par :

erreur de chute $\qquad \boxed{|R(S_n)| < B\ 10^{-p+1}\ \frac{n^2+n-2}{2}}$

erreur d'arrondi $\qquad \boxed{|R(S_n)| < B\ 10^{-p+1}\ \frac{n^2+n-2}{4}}$

ETUDE STATISTIQUE

ETUDE DE LA VARIABLE ALEATOIRE (s_i)

1) La loi donnée f (A) est telle que M $\neq$ 0

Définie dans l'intervalle (0, 1 ; 1) s_n converge en probabilité vers μ_n lorsque n augmente indéfiniment.

Mais on peut constater, lors des expériences, que l'allure "gaussienne" de la variable s_i est obtenue moins rapidement que dans le cas d'une somme arithmétique.

Résultat pratique :

$i > 40$ entraîne en général $s_i \neq \mu_i$

voir tableaux E_1 et E_2 .

2) M = 0

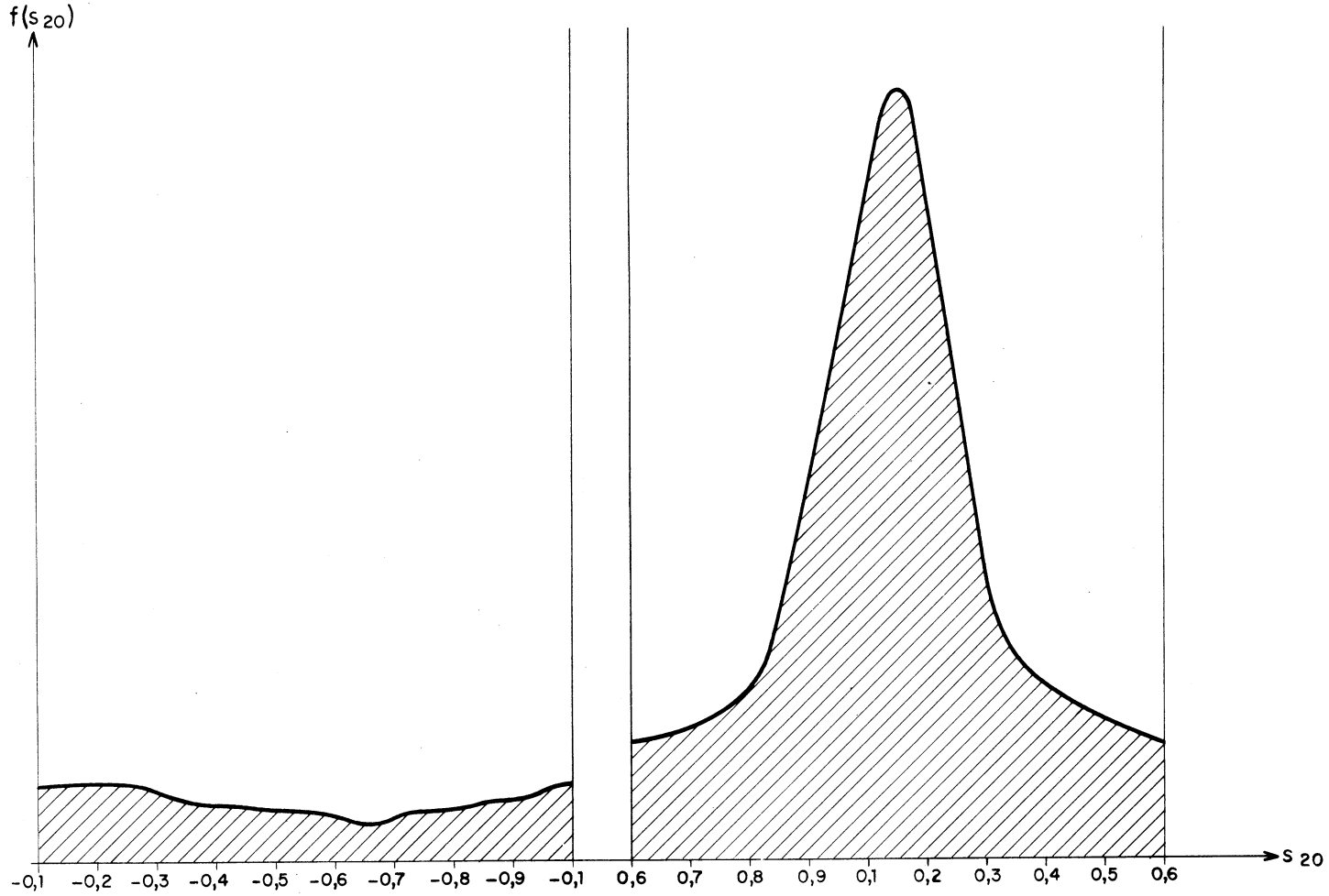
Les valeurs moyennes μ_i sont théoriquement toutes nulles.

(i $M = \mu_i \; 10^{s_i} = 0$), alors que le calcul effectif donnera toujours une valeur s_i^k telle que :

$$0,1 \leq s_i^k < 1$$

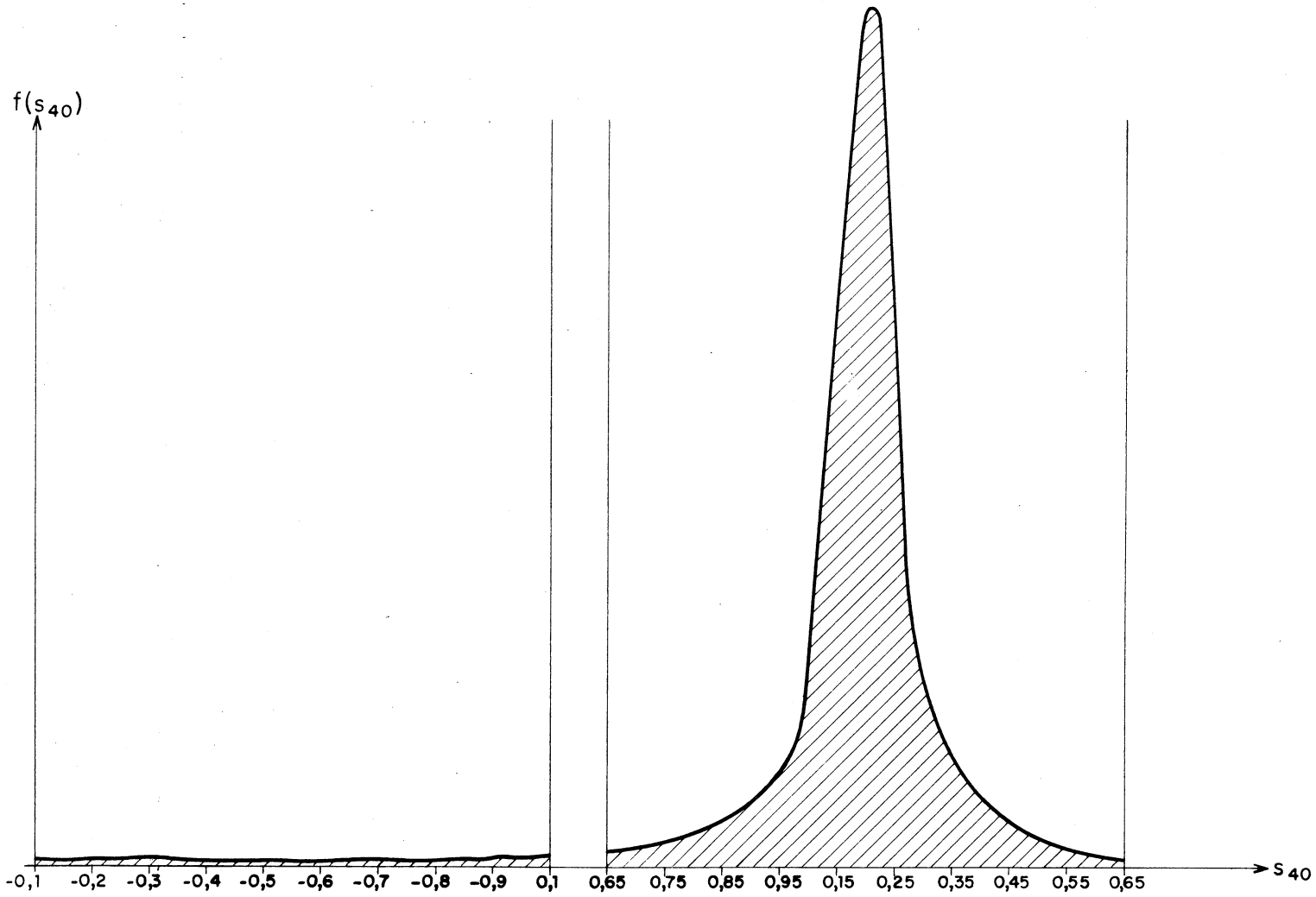
Nous n'avons aucune information théorique sur la loi de probabilité de s_i .

(E_1)



$\mu_{20} = 0{,}11$

(E_2)


f(s40)
-0,1
-0,2
-0,3
-0,4
-0,5
-0,6
-0,7
-0,8
-0,9
0,1
0,65
0,75
0,85
0,95
0,15
0,25
0,35
0,45
0,55
0,65
S40

$\mu_{40} = 0{,}22$

ETUDE DE L'ECHANTILLON ALEATOIRE L (V_{i+1})

$$V_{i+1} = \widetilde{V}_i + A_i = V_i + A_i - L(V_i)$$

Nous poserons à présent

$$\begin{cases} A_i = \xi_{a_i} \cdot a_i \cdot 10^{\alpha_i} \\ V_i = \xi_{v_i} \cdot v_i \cdot 10^{\phi_i} \\ V_i + A_i = \xi_{w_i} \cdot w_i \cdot 10^{\psi_i} \end{cases}$$

Les échantillons L (V_i) et K_i L $(V_i + A_i)$ sont, au signe près, 2 échantillons d'une même loi (uniforme) si nous attribuons à K_i la valeur $10^{\phi_i - \psi_i}$

Nous aurons à nouveau :

$$\text{Prob.} \left\{ E\left[L(V_{i+1}) \right]^n = E\left[L(V_i + A_i) \right]^n \right\} > 1 - q$$

erreur de chute $\quad q = k_i \, 10^{-p}$

erreur d'arrondi $\quad q = 0{,}5 \, k_i \, 10^{-p}$

Le problème est de savoir si q est toujours suffisamment petit pour qu'il soit possible d'affirmer que les échantillons L (V_{i+1}) et L (S_i) proviennent de la même loi.

Considérons de façon tout à fait générale une opération $V_i + A_i$ et comparons la valeur ψ_i obtenue à celle de ϕ_i.

1) $\xi_{v_i} \cdot \xi_{a_i} > 0$

Nous aurons nécessairement : $\psi_i \geq \phi_i$

En effet : $\psi_i = \phi_i$ ou $\psi_i = \phi_{i+1}$

C'est le cas d'une somme arithmétique.

2) $\xi_{v_i} \cdot \xi_{a_i} < 0$

a)	$\alpha_i > \phi_i$	$\psi_i \geq \phi_i$	$\psi_i = \alpha_i$ $\psi_i = \alpha_i - 1$
b)	$\phi_i > \alpha_i$	$\psi_i \geq \phi_i - 1$	$\psi_i = \phi_i$ $\psi_i = \phi_i - 1$
c)	$\phi_i = \alpha_i$	$\psi_i = \phi_i$	$\left\{ v_i^1 \neq a_i^1 \right\}$
		$\psi_i = \phi_i - 1$	$\left(\begin{array}{l} v_i^1 = a_i^1 \\ v_i^2 \neq a_i^2 \end{array} \right\}$
		$\psi_i = \phi_i - t$	$\left(\begin{array}{l} v_i^1 = a_i^1 \\ \vdots \\ v_i^t = a_i^t \\ v_i^{t+1} \neq a_i^{t+1} \end{array} \right\}$

Nous pouvons affirmer que le cas qui se présentera presque toujours sera : $\psi_i \geq \phi_i - 1$ c'est à dire : $K_i \leq 10$ et $q < 10^{-(p-1)}$

Remarques

1) Pour un problème particulier, seule une étude préliminaire de la répartition des nombres A_i en fonction des possibilités d'avoir toutes les conditions suivantes satisfaites : ($\xi_v . \xi_a < 0$, $\phi_i = \alpha_i$, $v_i^t = a_i^t$ $t \geq 1$) permettrait de donner une conclusion catégorique.

2) Lorsque n est nettement inférieur à 10^{p-1} ($n q \ll 1$) l'approximation de l'échantillon L (V_{i+1}) par l'échantillon L (S_i) est valable pour tout $i \leq n$

Consequence

$$R(S_n) \# \sum_{i=2}^{i=n} L(S_i)$$

avec $S_i = \xi_{s_i} . s_i . 10^{s_i}$

EVALUATION DES VALEURS MOYENNES ET ECARTS TYPE

1) $M \neq 0$

Nous aurons les mêmes formules que dans le cas d'une somme arithmétique.

- la théorie (a) donne $\boxed{E(R) = 10^{-p} \; M \frac{n^2}{2} \quad E(x) \frac{\mu_n - 0,0911}{\mu_n^2}}$

- la théorie (b) donne $\boxed{E(R) = 10^{-p} \; M \frac{n^2}{2} \quad E(x) \frac{23 + 9 \text{ Log } \mu_n}{9 \mu_n}}$

- en ce qui concerne l'écart type seule la formule correspondant à la théorie (b) peut être utilisée dans le cas algébrique :

$$\boxed{\sigma_R^2 = 10^{-2p} \frac{n^3}{3} \left[M (q^2 - m^2) + m^2 (Q^2 - M^2) \right]}$$

En effet, la formule donnée par la théorie (a) :

$$\sigma_R^2 = n \; \sigma_x^2 \quad 10^{2(\lambda + k - p)} \left[1 - \frac{0,0911}{\mu_n} \right]$$

ne fait pas intervenir l'écart type de A.

2) $M = 0$

Les théories (a) et (b) donnent : $\boxed{E(R) = 0}$

Pour l'écart type (théorie (b) $\boxed{\sigma_R^2 = 10^{-2p} \left[\frac{n^3}{3} m^2 Q^2 + \frac{n^2}{2} q^2 Q^2 \right]}$

Remarques

$\frac{n^2}{2} q^2 Q^2$ qui apparait dans cette dernière formule est l'un des termes que l'on néglige dans le cas d'une somme arithmétique n étant suffisamment grand. *Voir page 36*

La théorie (a) donne $\sigma_R^2 = 0$ ce qui est évidemment toujours faux ($B_1 \neq B_2$).

Détermination des coefficients m et q (lorsque M = 0)

$$\begin{cases} m = E(x) \cdot E\left(\frac{1}{s}\right) \\ q^2 = E(x)^2 \cdot E\left(\frac{1}{s}\right)^2 \end{cases}$$

En désignant par (s^k) la variable aléatoire dont l'ensemble de n - 1 échantillons est constitué par $(s^k_2 ; s^k_3 \ldots s^k_n)$

Le seul renseignement certain sur m et q est :

$$|m| < |E(x)| \cdot 10$$

$$q^2 < E(x)^2 \cdot 10^2$$

ce qui entraîne

- pour l'erreur de chute -

$$\boxed{\sigma_R^2 < \frac{25}{3} 10^{-2p} n^3 . Q^2}$$

- pour l'erreur d'arrondi - $\begin{cases} m = 0 \\ q^2 = \frac{1}{12} \end{cases}$

$$\boxed{\sigma_R^2 < \frac{25}{6} 10^{-2p} n^2 . Q^2}$$

Ces bornes établies pour les écarts type sont au moins, en général, trois ou quatre fois plus grandes que les écarts type réels.

Hypothèse supplémentaire nécessaire

Dans le cas d'une répartition uniforme de A nous prendrons comme hypothèse supplémentaire que la loi de s est, elle même, uniforme entre 0,1 et 1. *Voir page 43*

ETUDE EXPERIMENTALE

Les expériences ont été faites de la même façon que pour les sommes arithmétiques, dans le cas des erreurs de chute.

Nous avons comparé m_{ta} et m_{tb} à m_p . *Voir page 51*

Les valeurs $\sigma_{R_{ta}}$ ne figurent plus dans le tableau des résultats expérimentaux, n'ayant plus aucun rapport avec celles des écarts type obtenus dans la pratique $\sigma_{R_{pr}}$

CONCLUSIONS

Lorsqu'on applique les formules donnant la correction de calcul sur une somme algébrique, les résultats sont aussi bons et semblent même meilleurs que pour une somme arithmétique.

Dans le cas particulier où M = 0 l'hypothèse d'uniformité en ce qui concerne la variable (s) est d'autant plus valable que l'écart type de A est grand.

Famille (III) b $M = 5$ $Q^2 = 386$

n	p	k	m_{ta}	m_{pr}	m_{tb}	r_{pr}	r_{tb}
400	4	20	2,7	2,5	2,36	0,36	0,44
700	4	20	2,12	2,06	2,14	0,368	0,367
9000	5	1	1,78	1,87	1,95		
9000	4	1	1,78	1,85	1,95		

Famille (I) $M = 0$ $B_1 = -10^3$ $B_2 = 10^3$

n	p	k	$E_{(R_{ta})}$	$E_{(R_{pr})}$	$E_{(R_{tb})}$	$\sigma_{R_{pr}}$	$\sigma_{R_{tb}}$
100	4	15	0	- 12	0	34	42
700	4	15	0	+ 88	0	560	780

CHAPITRE III

CORRECTION DE CALCUL SUR LE RESULTAT D'UN PRODUIT DE n NOMBRES

Soit P_n le produit $P_n = \prod_{i=1}^{n} A_i = p_n 10^{\gamma_n}$

La formule générale relative à la correction de calcul sur un produit donne

$$R(P_n) = \sum_{i=2}^{n} L(W_{i+1})\, A_{i+1} \ldots A_{n+1}$$

avec toujours les conventions suivantes $\begin{cases} W_{i+1} = \widetilde{W_i} \cdot A_i \\ W_2 = A_1 \end{cases}$ $\qquad A_{n+1} = 1$

BORNE DE L'ERREUR

Quel que soit i $\qquad |W_{i+1}| \neq |A_1 . A_2 \ldots A_i|$

$$\left| R(P_n) \right| \neq \sum_{i=2}^{n} |P_n| \left| \frac{L(A_1 . A_2 \ldots A_i)}{A_1 . A_2 \ldots A_i} \right|$$

$$\frac{L(P_i)}{P_i} = \frac{x_i}{p_i} 10^{-p}$$

$$\left| R(P_n) \right| \neq |P_n| \, 10^{-p} \sum_{i=2}^{n} \frac{x_i}{p_i}$$

Une borne stricte de l'erreur sera :

pour l'erreur de chute

$$0 \leqslant \frac{x_i}{p_i} < 10 \quad \text{quel que soit } i$$

$$\boxed{\left|R(P_n)\right| < 10^{1-p}(n-1)\left|P_n\right|}$$

pour l'erreur d'arrondi

$$-5 < \frac{x_i}{p_i} < 5$$

$$\boxed{\left|R(P_n)\right| < 5.\,10^{-p}(n-1)\left|P_n\right|}$$

ETUDE STATISTIQUE

ETUDE DE LA VARIABLE ALEATOIRE (p_i)

1) Etude théorique $\quad P_i^k = p_i^k \; 10^{\gamma_i^k}$

P_i et p_i sont les variables aléatoires dont les échantillons sont respectivement :

$(P_i^1 \ldots P_i^k \ldots P_i^m)$ et $(p_i^1 \ldots p_i^k \ldots p_i^m)$

$$0,1 \leqslant p_i < 1$$

Soient k_1 et k_2, 2 nombres fixes tels que :

$$0,1 < k_1 < k_2 < 1$$

supposons que les bornes de la variable A sont :

$$B_1 = 1 \text{ et } B_2 = 10^l \qquad (l \text{ quelconque})$$

$1 < P_n < 10^{nl}$ $\qquad (1 < 10\,k_1 < 10^{nl}\; k_2 < 10^{nl})$

La définition de la variable p_n implique :

$$\text{Prob.} \left\{ k_1 < p_n < k_2 \right\} = \begin{cases} \text{Prob} \left\{ 10\,k_1 < P_n < 10\,k_2 \right\} \\ + \\ \text{Prob} \left\{ 10^2 k_1 < P_n < 10^2 k_2 \right\} \\ + \\ \vdots \\ + \\ \text{Prob} \left\{ 10^i k_1 < P_n < 10^i k_2 \right\} \\ + \\ \vdots \\ + \\ \text{Prob} \left\{ 10^{nl}\, k_1 < P_n < 10^{nl}\, k_2 \right\} \end{cases}$$

En effet le domaine favorable à l'évènement :

$$P_n \text{ compris entre } \begin{cases} 10\ k_1 \text{ et } 10\ k_2 \\ \underline{\text{ou}} \\ 10^2\, k_1 \text{ et } 10^2\, k_2 \\ \vdots \\ \underline{\text{ou}} \\ 10^{nl} k_1 \text{ et } 10^{nl} k_2 \end{cases} \quad \text{(somme logique)}$$

est le domaine favorable à l'évènement p_n compris entre k_1 et k_2

Considérons à présent Log (P_n) et Log (p_n)

$$\text{Notons} \quad \begin{cases} K_1 = -\text{ Log } (k_1) \\ K_2 = -\text{ Log } (k_2) \end{cases} \qquad 0 < K_2 < K_1 < \text{Log } (10)$$

$$(i-1)\ \text{Log}(10) < i\ \text{Log}(10) - K_1 < i\ \text{Log}(10) - K_2 < i\ \text{Log}(10)$$

$$\text{Prob.} \left\{ K_2 < -\text{Log}(p_n) < K_1 \right\} = \sum_{i=1}^{nl} \text{Prob.} \left\{ i\,\text{Log}(10) - K_1 < \text{Log}(P_n) < i\,\text{Log}(10) - K_2 \right\}$$

Lorsque n est suffisamment grand nous savons que

$$\text{Log } (P_n) = \sum_{i=1}^{n} \text{Log } (A_i)$$

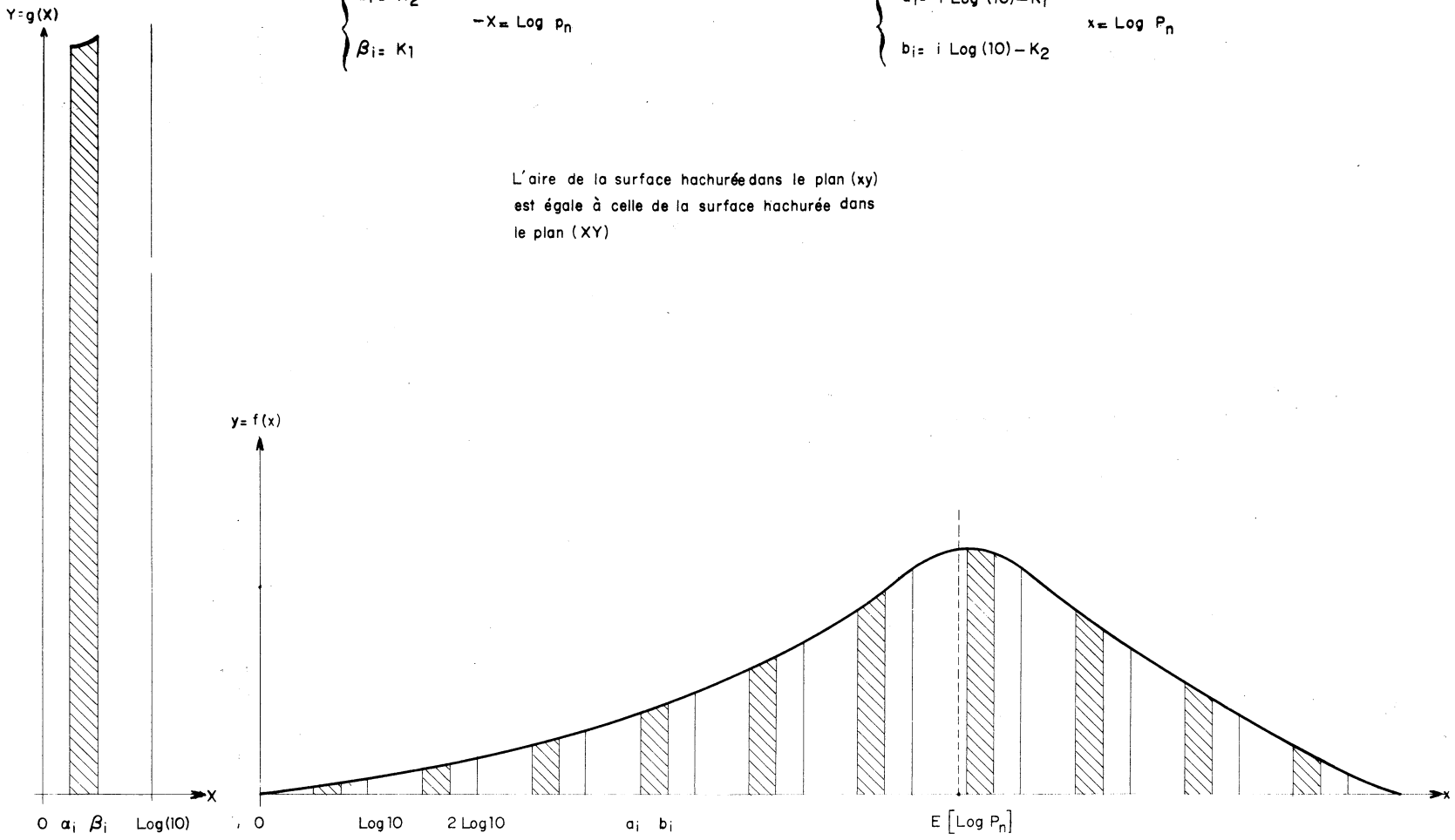
tend vers une loi de Gauss (avec n) *(Voir figure F, page 55)*

(F)

$$\begin{cases} \alpha_i = K_2 \\ \beta_i = K_1 \end{cases} \quad -X = \text{Log } p_n$$

$$\begin{cases} a_i = i\,\text{Log}(10) - K_1 \\ b_i = i\,\text{Log}(10) - K_2 \end{cases} \quad x = \text{Log } P_n$$

L'aire de la surface hachurée dans le plan (xy) est égale à celle de la surface hachurée dans le plan (XY)



La courbe représentative de cette loi ($n > N$) possède une certaine symétrie par rapport à la valeur moyenne $E\left[\text{Log}\,(P_n)\right]$

Il se peut qu'elle ne soit pas complètement symétrique dans l'intervalle $0, n\,\text{Log}\,10$ - cela dépend de la loi de probabilité de la variable Log (A) - mais elle le sera nécessairement dans toute la partie prépondérante de la courbe.

Lorsque n augmente les intervalles de longueur Log (10) (et de longueur $K_1 - K_2$) sont de plus en plus nombreux dans la partie symétrique de la courbe (l'écart type ne varie qu'avec $\sqrt{n}$).

Définissons une variable X dans l'intervalle K_2 , K_1

$$K_2 \leqslant X \leqslant K_1$$

et soit $x_i = i\,\text{Log}\,(10) - X$

Pour chaque valeur de i fixée la variable x_i est définie dans l'intervalle $i\,\text{Log}\,(10) - K_1$, $i\,\text{Log}\,(10) - K_2$

Posons $g\,(X_1) = \sum_{i=1}^{n\iota} f\,(i\,\text{Log}\,(10) - X_1)$ pour une valeur X_1 de X

Dans le plan (X, g (X)) traçons la courbe :

$$g\,(X) = \sum_{1}^{n\iota} f(x_i)$$

$\int_{K_2}^{K_1} g(X)\,dX$ est la somme de toutes les aires $\int_{i\,\text{Log}\,(10)-K_1}^{i\,\text{Log}\,(10)-K_2} f(x)\,dx$

La fonction g (x) tend, lorsque n augmente indéfiniment, à être constante dans l'intervalle K_2 , K_1.

La densité de probabilité de-Log (p_n) tend donc à être uniforme entre K_2 , K_1 et, de façon plus générale, entre 0, Log10.

En effet considérons la somme $\sum_{1}^{n\,1} (x_i)$

1) - les termes correspondant aux portions de courbes extérieures à l'intervalle prépondérant sont rapidement négligeables.

2) - on peut associer deux à deux les termes dont les portions de courbe sont symétriques : par exemple f (x_i) et f (x_j).

3) - lorsque n augmente :

f (x_i) tend vers t_i X + f (i Log (10) - K_1)

f (x_j) tend vers - t_i X + f (j Log (10) - K_1)

dont la somme est constante pour tout X

(t_i étant indépendant de la valeur X)

La limite de g (X) est $\frac{1}{\text{Log (10)}}$ $\left(\frac{1}{\text{Log (10)}} \times \int_{\text{Log (0, 1)}}^{\text{Log (1)}} dx = 1 \right)$

soit $X = \text{Log } (z)$ $\quad dX = \frac{1}{z}\, dz$

La limite de la loi de probabilité de la variable p_n est une loi dont la densité de répartition est :

$$f(p) = \frac{1}{\text{Log (10)}} \quad \frac{1}{p}$$

$$E\left(\frac{1}{p}\right) = 3,9116$$

$$E\left(\frac{1}{p}\right)^2 = 21,497$$

2) Etude purement statistique de la variable p_i

L'expérience nous montre que les v.a. (p_i) se comportent sensiblement de la même façon.

- Les échantillons ont été fabriqués avec 4 décimales:

$$1,000 \leqslant A_i \leqslant 9,999 \qquad \begin{cases} B_1 = 1 \\ B_2 = 10 \end{cases}$$

1) - en prenant une fonction de la famille (I) on trouve les résultats représentés par le tableau (G)

2) - divers produits de i échantillons de la famille (III) donnent les lois figurant sur le tableau (H)

3) - toutes ces courbes de répartitions statistiques sont très proches de $\frac{1}{\text{Log}(10)} \quad \frac{1}{p}$ (tableau I)

Remarque

Le fait d'approcher si rapidement, pour n = 10 par exemple, la loi limite établie par la théorie ($n \to \infty$) n'est pas tellement étonnant : c'est une conséquence de l'allure gaussienne très vite obtenue, pour la densité de répartition d'une somme d'échantillons.

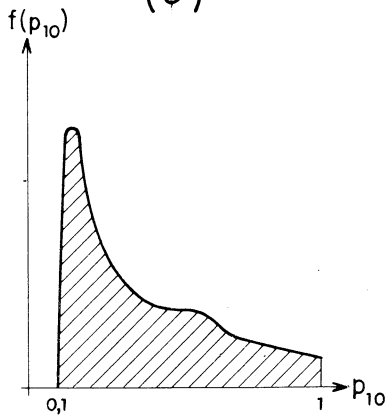
Ces échantillons étant dans la mesure du possible, équiprobables vis à vis de f (A)

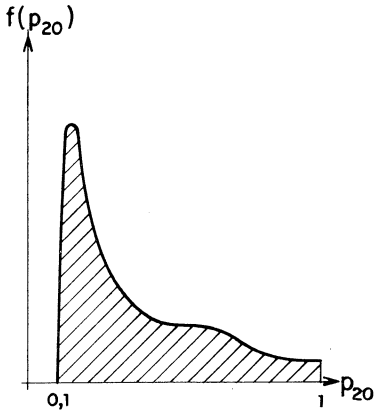
ETUDE DE L'ECHANTILLON ALEATOIRE L (W_{i+1})

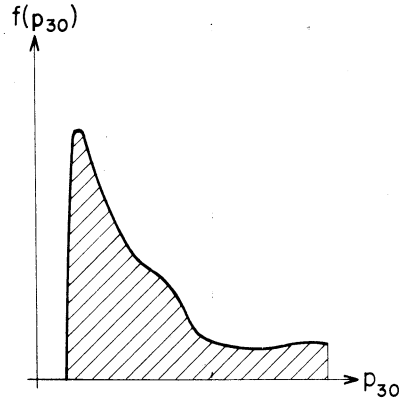
$W_{i+1} = \widetilde{W}_i \cdot A_i = W_i A_i - L(W_i) A_i$

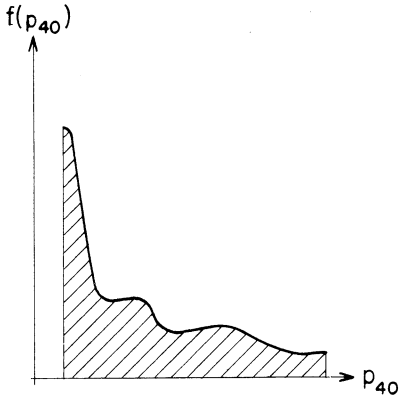
$\left| L(W_i) A_i \right|$ et $K \left| L(W_i A_i) \right|$ sont 2 échantillons d'une même loi si $0,1 < K < 10$

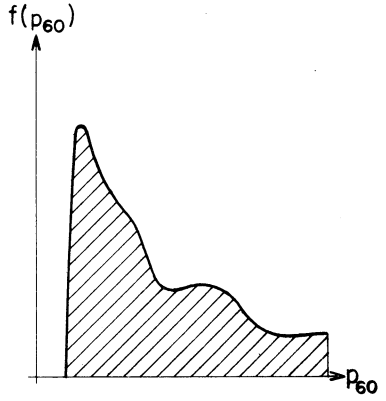
(G)



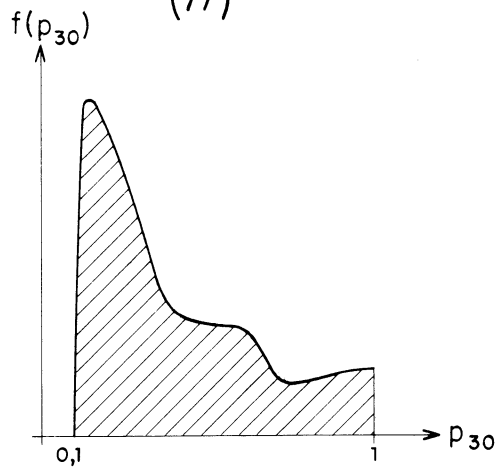


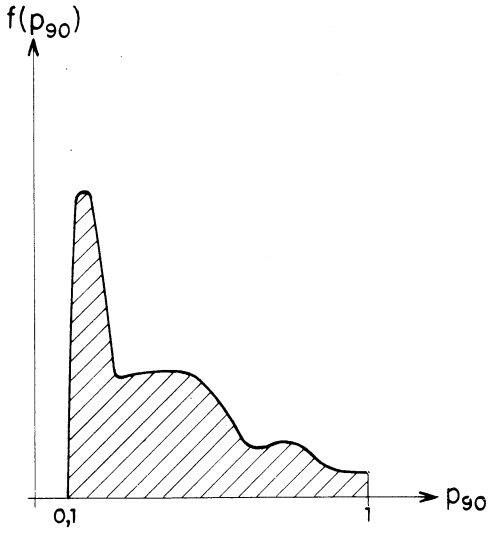


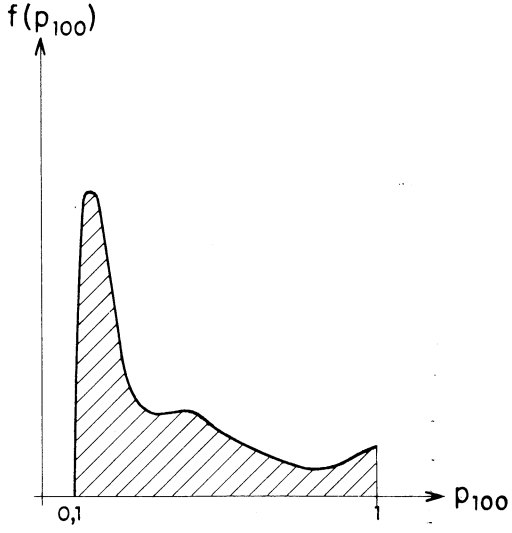




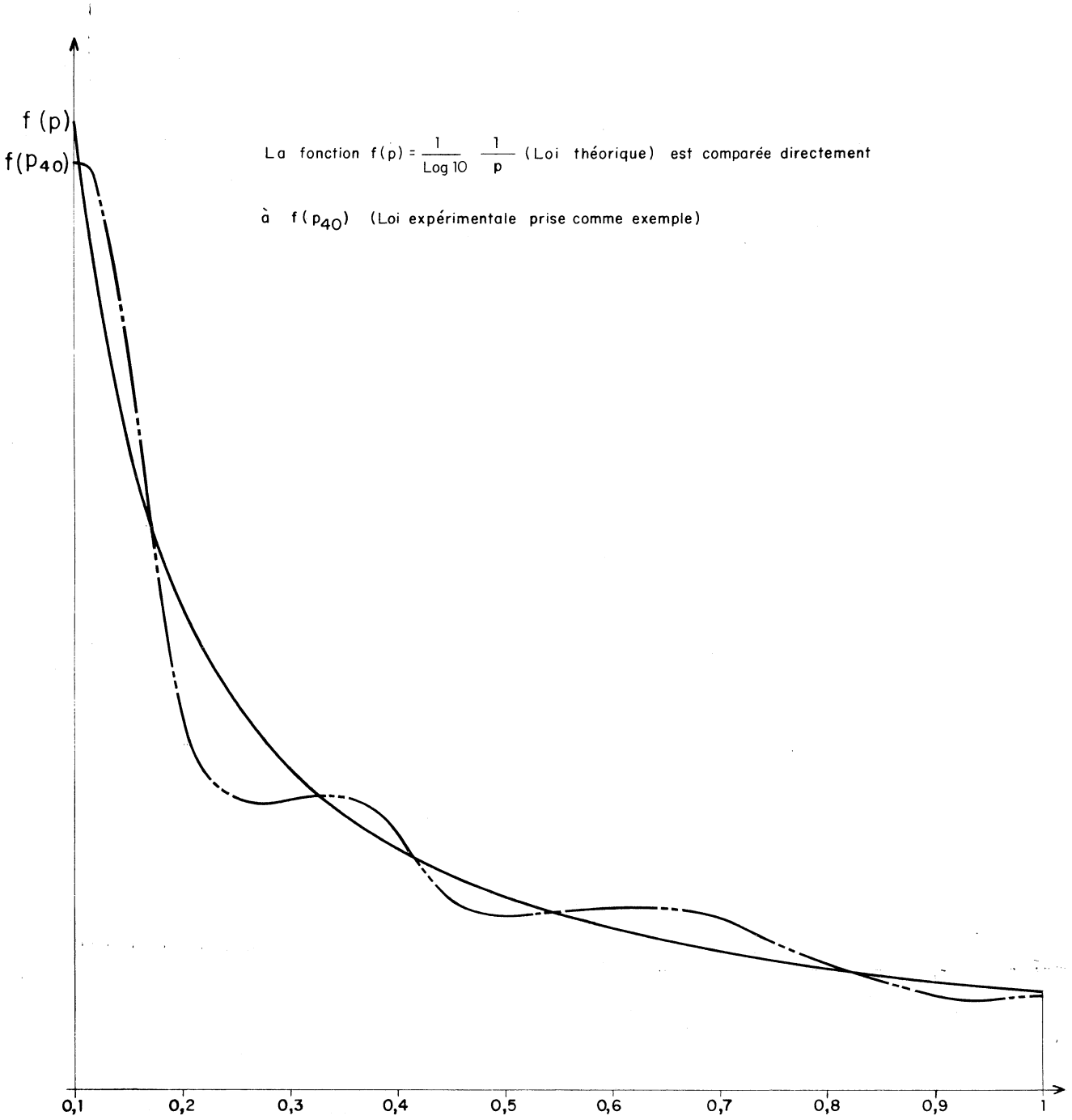
(H)







(I)



La fonction $f(p) = \frac{1}{\text{Log } 10} \frac{1}{p}$ (Loi théorique) est comparée directement

à $f(p_{40})$ (Loi expérimentale prise comme exemple)

Nous aurons à nouveau :

$$\text{Prob.} \quad \left\{ E \left[L (W_{i+1}) \right]^n = E \left[L (W_i \cdot A_i) \right]^n \right\} > 1 - q$$

erreur de chute $\quad q = 10^{-p+1}$

erreur d'arrondi $\quad q = 0,5\ 10^{-p+1}$

et finalement pour p assez grand

$$\boxed{\sum_{i=2}^{n} L (W_{i+1})\ A_{i+1} \ldots A_n \quad \# \quad \sum_{i=2}^{n} L (P_i)\ A_{i+1} \ldots A_n}$$

$$R (P_n) = P_n \sum_{i=2}^{n} \frac{L (P_i)}{P_i}$$

$$R (P_n) = 10^{-p}\ P_n \sum_{i=2}^{n} \frac{x_i}{p_i}$$

soient $\quad E \left[\frac{x_i}{p_i} \quad \frac{x_j}{p_j} \right] = \begin{cases} m^2 & (i \neq j) \\ q^2 & (i = j) \end{cases}$

Hypothèses

1) $\quad f (p_i) = f (p_j) = \frac{I}{\text{Log } (10)} \quad \frac{I}{p}$ pour $i \neq j$

ce qui est conforme à l'étude faite sur (p_i)

2) Les variables (x) et (p) sont indépendantes :

$$\begin{cases} m = E (x)\ E \left(\frac{1}{p}\right) \\ q^2 = E (x)^2\ E \left(\frac{1}{p}\right)^2 \end{cases}$$

Notons r (P_n) la correction de calcul relative sur P_n

$$r (P_n) = \frac{R (P_n)}{P_n} = 10^{-p} \sum_{2}^{n} \frac{x_i}{p_i}$$

EVALUATION DES VALEURS MOYENNES ET ECARTS TYPE

Valeurs moyennes

$$E(r) = m\ 10^{-p}\ (n-1)$$

Ecarts type

$$r^2 = 10^{-2p} \sum_i \left(\frac{x_i}{p_i}\right)^2 + \sum_i \sum_{j \neq i} \left(\frac{x_i}{p_i} \cdot \frac{x_j}{p_j}\right)$$

$$\sigma_r^2 = (q^2 - m^2)\ 10^{-2p}\ (n-1)$$

erreur de chute

$$\begin{cases} m = 1,9558 \\ q^2 - m^2 = 3,34 \end{cases}$$

erreur d'arrondi

$$\begin{cases} m = 0 \\ q^2 - m^2 = 1,7914 \end{cases}$$

ETUDE EXPERIMENTALE

Les expériences ont été faites à nouveau dans le cas de l'erreur de chute.

Elles consistent à vérifier les valeurs attribuées à m et à $q^2 - m^2$.

Nous avons effectué k fois les deux produits P_n et $\tilde{P}_n$, en prenant à chaque fois n échantillons différents d'une même loi.

soit $$r_j = \frac{P_n{}^j - \tilde{P}_n{}^j}{P_n{}^j}$$

Les k "corrections de calcul relatives" obtenues sont $(r_1\ r_2 \ldots r_k)$

Leur valeur moyenne est $$E(r_{pr}) = \frac{1}{k} \sum_{j=1}^{k} r_j \quad \text{et}$$

leur écart type est $$\sigma^2_{r_{pr}} = \frac{1}{k} \sum_{j=1}^{k} (r_j)^2 - \left[\frac{1}{k} \sum_{j=1}^{k} \right]^2$$

Les valeurs de m et $q^2 - m^2$ données par l'expérience sont finalement :

$$m_{pr} = \frac{10\,P}{n-1} E(r_{pr})$$

$$q^2_{pr} - m^2_{pr} = \frac{10^2\,P}{n-1} \sigma^2_{r_{pr}}$$

Les valeurs théoriques sont notées m_t et $q_t^2 - m_t^2$

FAMILLE	n	p	k	B_1	B_2	m_t	m_{pr}	$q_t^2 - m_t^2$	$q_{pr}^2 - m_{pr}^2$
I	30	4	15	1	10	1,956	2,02	3,34	3,21
I	60	4	15	1	10	1,956	1,9	3,34	3,4
III	50	4	20	1	10	1,956	1,95	3,34	3,43
III	100	4	20	1	10	1,956	1,954	3,34	2,9

CHAPITRE IV

Examinons à présent le cas où les nombres A_i de départ sont des éléments de la classe (I) mais où les résultats des premières opérations ne sont pas nécessairement des nombres faisant partie de la classe (II).

SOMME ARITHMETIQUE

$$S_1, S_2, \ldots S_g \quad \in \text{ classe (I)}$$

$$S_{g+1}, \ldots S_n \quad \in \text{ classe (II)}$$

$$R(S_n) = \sum_{g+1}^{n} L(S_i) \qquad (L(S_i) = 0 \text{ pour } i \leqslant g)$$

Il est possible de se ramener au cas précédemment étudié, et utiliser les formules établies,

en posant
$$\begin{cases} S^x_i = S_i = s_i \; 10^{\delta i} & \text{pour} \quad i > g \\ S^x_i = s_i + x_i \; 10^{\delta i} & \text{pour} \quad i \leqslant g \end{cases}$$

$$\sum_{g+1}^{n} L(S_i) = \sum_{2}^{g} L(S^x_i) + \sum_{g+1}^{n} L(S_i) - \sum_{2}^{g} L(S^x_i)$$

$$R(S_n) = \sum_{2}^{n} L(S^x_i) - \sum_{2}^{g} L(S^x_i)$$

$$R(S_n) = R(S^x_n) - R(S^x_g) \qquad \begin{cases} S^x_i \in \text{ classe (II)} \\ 2 \leqslant i \leqslant n \end{cases}$$

ce qui donne pour la valeur moyenne de l'erreur

$$E(R) = 10^{-p_M} \frac{n^2 \cdot m(\mu_n) - g^2 \cdot m(\mu_g)}{2}$$

Si le nombre moyen de chiffres significatifs de la variable A ($a\ 10^{\alpha}$) est connu, soit e, la valeur de g sera évaluée par : $g = \frac{10^{p-e}}{E(a)}$

exemple : f (A) est uniforme entre 1 et 10^3 tous les échantillons A_i sont des nombres entiers.

$$\text{Prob.} \left\{ e = j \right\} = \frac{0,910^{j}}{999} \quad (j = 1, 2, 3)$$

Valeur moyenne de e : 2,899

$$g = 2\ 10^{p-e}$$

pour $n < 2\ 10^{p-e} \longrightarrow R(S_n) = 0$

SOMME ALGEBRIQUE

$S_1, S_2, \ldots, S_g \in$ classe (I)

$S_{g+1} \in$ classe (II)

Il peut arriver que S_{g+i} fasse partie de la classe (I) $i > 1$ mais si nous raisonnons sur des valeurs moyennes nous aurons à nouveau :

$$R(S_n) = - R(S'_g) + R(S'_n) \quad \text{avec}$$

$$g \neq 10^{p-e} \frac{1}{E(a)}$$

PRODUIT

$P_1, P_2, \ldots P_g \in$ classe (I)

$P_{g+1}, \ldots\ldots P_n \in$ classe (II)

$$E(r) = m\ 10^{-p}\ (n - g)$$

$$\sigma_r^{\ 2} = (q^2 - m^2)\ 10^{-2p}\ (n - g)$$

e étant le nombre moyen de chiffres significatifs des échantillons A_i

$$g = \frac{p}{e + \log_{10}\ [E(a)]} \qquad \left(g < \frac{p}{e - 1}\right)$$

DEUXIEME PARTIE

ERREURS DE CALCUL
DANS LA RESOLUTION DES SYSTEMES DIFFERENTIELS

Les "erreurs de calcul" qui résultent de la résolution numérique d'un certain problème sont elles-même, très souvent, solution d'un problème de même type que celui étudié.

Nous le constaterons aussi pour la "correction de calcul propagée" due à la résolution pratique de systèmes d'équations différentielles satisfaisants à des conditions initiales.

La résolution du nouveau problème auquel nous sommes conduits, ne nécessite pas en général, la précision du problème initial.

Nous pourrons utiliser un pas d'intégration plus grand que celui qui a été pris pour le système donné.

NOTATIONS

a	représente un scalaire. $\underline{a}$ un vecteur.
n	le nombre de pas que nécessite l'intégration.
h	le pas utilisé.
$\underline{y}(t_n)$	est la valeur exacte de la solution à l'abscisse t_n.
$\underline{y}_n(t)$	la valeur de la solution établie par la méthode employée.
$\tilde{\underline{y}}_n(t)$	sa valeur numérique calculée (donnée par le calculateur)
$\underline{R}(\underline{y}_n) = \underline{y}_n - \tilde{\underline{y}}_n$	est la correction de calcul propagée. Cette fonction, que précisément nous nous proposons d'évaluer, dépend de l'abscisse t.
m	est l'ordre du système différentiel.

Les notations précédentes sont conservées, toutefois nous noterons une fonction $a(x) + b(y)$ calculée numériquement, par : $\widetilde{a(x) + b(y)}$
(cette écriture sous entend : $\left[\widetilde{\tilde{a}(\tilde{x}) + \tilde{b}(\tilde{y})}\right]$)

Le système est donné sous la forme résolue suivante :

$$\frac{d\underline{y}}{dt} = \underline{Y}(\underline{y}, t) \qquad t_n = a + nh$$
$$\underline{y}(a) = \underline{y}_o$$

avec $\underline{Y} \begin{cases} Y^1 \\ Y^2 \\ \\ Y^m \end{cases}$ $\underline{y} \begin{cases} y^1 \\ y^2 \\ \\ y^m \end{cases}$

Nous nous plaçons dans le cas général de résolution des systèmes par une méthode à pas séparés.

La formule d'intégration approchée s'écrit :

$\underline{y}_{n+1} = \underline{y}_n + h \ \underline{F} \ (\underline{y}_n , h)$

Lorsque la méthode utilisée est d'ordre q nous savons que F doit satisfaire à :

$$\underline{F} \ (y) = \underline{Y} \ (y) + \frac{h}{2} \ \underline{Y}' + \ldots + \frac{h^{q-1}}{q!} \ \underline{Y}^{(q-1)}$$

$\underline{Y}^{(k)}$ représente la dérivée totale d'ordre k de $\underline{Y}$ par rapport à t .

I - CALCUL DE $\underline{R} \ (\underline{y}_n)$

L'évaluation de la "correction de calcul" sur la solution suppose que le système a été préalablement intégré.

$$\underline{R} \ (\underline{y}_{n+1}) = \underline{R} \left[\underline{y}_n + h \ \underline{F} \ (\underline{y}_n, h) \right]$$

$$\underline{R} \ (\underline{y}_{n+1}) = \underline{R} \ (\underline{y}_n) + \underline{R} \left[h \ \underline{F} \ (\underline{y}_n, h) \right] + \underline{R} \left[\tilde{\underline{y}}_n + h \ \widetilde{\underline{F}} \ (\underline{y}_n, h) \right]$$

(correction de calcul sur une somme) d'autre part :

$$\underline{R} \left[h \ \underline{F} \ (\underline{y}_n, h) \right] = \underline{R} \left[h \ \underline{F} \ (\tilde{\underline{y}}_n, h) \right] + h \left[\underline{F} \ (\underline{y}_n, h) - \underline{F} \ (\tilde{\underline{y}}_n, h) \right]$$

Posons :

$$\underline{G} \ (\underline{F}, \tilde{\underline{y}}_n, h) = \underline{R} \left[h \ \underline{F} \ (\tilde{\underline{y}}_n) \right] + \underline{R} \left[\tilde{\underline{y}}_n + h \ \widetilde{\underline{F}} \ (\underline{y}_n, h) \right]$$

$$\underline{R} \ (\underline{y}_{n+1}) = \underline{R} \ (\underline{y}_n) + h \left[\underline{F} \ (\underline{y}_n, h) - \underline{F} \ (\tilde{\underline{y}}_n, h) \right] + \underline{G} \left[\underline{F}, \tilde{\underline{y}}_n, h \right]$$

$\underline{G} \left[t_n \right]$ est "la correction de calcul locale" (sur le $n^{ième}$ pas).

APPROXIMATIONS

La condition essentielle pour que l'étude de ces problèmes soit possible est : h suffisamment petit. Ce qui est conforme à la réalité.

Les termes en h^2, h^3, ... etc, lorsque ce sera légitime, seront négligés.

$$\underline{R}\left[\underline{y}_{n+1}(t)\right] = \underline{R}\left[\underline{y}_n(t)\right] + h\left[\underline{Y}\left(\underline{y}_n + R(\underline{y}_n)\right) - \underline{Y}(\underline{y}_n)\right] + \underline{G}\left[\underline{Y}, \underline{y}_n, h\right]$$

$$\underline{R}\left[\underline{y}_{n+1}(t)\right] = \underline{R}\left[\underline{y}_n(t)\right] + h\, J(t_n)\, \underline{R}\left[\underline{y}_n(t)\right] + \underline{G}\left[\underline{Y}, \underline{y}_n, h\right]$$

$$\underline{R}'_t\left[\underline{y}_n(t)\right] = J(t_n)\, \underline{R}\left[\underline{y}_n(t)\right] + \frac{1}{h}\, \underline{G}\left[\underline{Y}\left(\underline{y}_n(t)\right) \underline{y}_n(t)\right]$$

J est la matrice fonctionnelle du système initial.

La correction de calcul propagée est donnée par le vecteur solution :

$$\underline{R}(\underline{y}_n) = \frac{1}{h} \int_a^{t_n} M(t,u)\, \underline{G}(u)\, du + \underline{R}(\underline{y}_o)$$

Sachant que : $\begin{cases} M'_t(t,u) = J(t)\, M(t,u) \\ M(u,u) = I \end{cases}$

(M est la matière de propagation du système).

$$\underline{R}\, h\, \underline{Y}(\underline{y}_n) = \underline{Y}(\underline{y})\, \underline{R}(h) + h\, \underline{R}\, \underline{Y}(\underline{y}) + \underline{R}(h\, \underline{Y}(\underline{y}_n))$$

(correction de calcul sur un produit)

Prenons un pas de la forme $h = 10^{-z}$ z entier positif.

$\underline{R}(h) = 0$ $\quad \underline{R}\left[h\, \tilde{\underline{Y}}\right] = 0$ et $\underline{R}\left[\tilde{\underline{y}} + \widetilde{(h\, \tilde{\underline{Y}})}\right] = \underline{R}\left[\tilde{\underline{y}} + h\, \tilde{\underline{Y}}\right]$

$\underline{G}(u)$ peut donc s'écrire : $h\, \underline{R}\left[\underline{Y}(\tilde{\underline{y}}(u))\right] + \underline{R}\left[\tilde{\underline{y}}(u) + h\, \tilde{\underline{Y}}(\tilde{\underline{y}}(u))\right]$

Les indices k, l varient de 1 à m

Les indices i, j varient de 1 à n

Nous noterons la $k^{ième}$ composante de la solution $\underline{y}_i$ au pas i :

$$\tilde{y}^k (a+ih) \qquad \text{par} \qquad \tilde{y}_i^k$$

et $$Y^k \left[\tilde{\underline{y}} (a+ih) , a+ih \right] \quad \text{par} \quad Y_i^k$$

Pour un couple de valeurs (i, k) déterminé, nous aurons :

$$G_i^k = h R \left[Y_i^k \right] + R \left[\tilde{y}_i^k + h \tilde{Y}_i^k \right]$$

G_i^k et G_y^l sont 2 échantillons aléatoires indépendants ($i \neq j$) (mais peuvent éventuellement provenir d'une même loi).

HYPOTHESES FAITES SUR LA FONCTION G (u)

1) Nous connaissons une valeur B telle que,

$\left| G_i^k \right| \leq B$ quels que soient i et k

$\left| G^k (u) \right| \leq B$ pour $a \leq u \leq t_n$

2) Considérons les 2 égalités :

$$\begin{cases} E (G_i^k) = E (G_j^l) \\ E (G_i^k)^2 = E (G_j^l)^2 \end{cases}$$

Elles sont vérifiées :

soit pour i, j, k, l quelconques

le système différentiel sera du type (α)

soit pour k = l i et j étant quelconques

le système différentiel sera du type (β)

Les autres cas sont exclus.

REMARQUES

1) La seconde hypothèse restreint le domaine d'application des formules donnant la correction de calcul propagée ; elle suppose d'autre part, que l'on puisse évaluer les valeurs moyennes d'ordre 1 et 2 de $G(Y_i^k)$
(ce qui n'est pas toujours possible).

2) Il arrive fréquemment :

- d'une part, que $h\,R(Y_i^k)$ soit négligeable devant $R(\tilde{y}_i^k + h\,\tilde{Y}_i^k)$
  h étant petit
  $G_i^k \# R(\tilde{y}_i^k + h\,\tilde{Y}_i^k)$

- d'autre part, que l'on ait :

$$\underline{y}_i^k = 0,\ c^1\ c^2 \ldots c^p\ 10^{\gamma}$$
$$h\ \underline{\tilde{Y}}_i^k = 0,\ b^1\ b^2 \ldots b^p\ 10^{\gamma-1}$$
avec $0 < l < p$
$$\underline{R}(\underline{\tilde{y}}_i^k + h\ \underline{\tilde{Y}}_i^k) = 0,\ b^{p-l+1}\ b^{p-l+2} \ldots b^p\ \ 10^{\gamma-p}$$

L'hypothèse fondamentale sur la correction de calcul s'écrit en ce cas :

$R(\tilde{y}_i^k + h\,\tilde{Y}_i) = L(y_i^k)$

Conséquence : $G_i^k \# L(y_i^k)$

3) Lorsque le système différentiel envisagé ne rentre pas dans les types types (α) et (β) mais qu'il est possible de donner une bonne approximation de $E(G^k)$ et de $E(G^k)^2$, nous utiliserons en prenant suffisamment de précautions les résultats d'un système de type (β). (voir 4ème exemple, page 82)

Dorénavant nous noterons par $\underline{R}\ (t_n)$ ou plus simplement par

$\underline{R}_n = \underline{R}\ \left[y_n\ (t_n)\right]$ la correction de calcul propagée au $n^{ième}$ pas.

II - CAS DE RESOLUTION D'UNE EQUATION DIFFERENTIELLE.

Les types (α) et (β) sont confondus (m = 1)

$$R\ (t_n) = \frac{1}{h} \int_a^{a+nh} m\ (t_n, u)\ G\ (u)\ du + R\ (a)$$

Supposons que y_o fasse partie de la classe I :

$$R\ (y_o) = R\ (a)$$

La résolution de $\begin{cases} m'_t\ (t,u) = J\ (t)\ m\ (t,u) \\ m\ (u,u) = 1 \end{cases}$

donne $\qquad m\ (t,u) = e^{\int_u^t J\ (x)\, dx}$

posons $\qquad f\ (t) = \int_a^t e^{\int_u^t J(x)\, dx}\ du$

f (t) sera déterminée numériquement par la solution de l'équation différentielle

$$\begin{cases} f'(t) = 1 + J\ (t)\ f\ (t) \\ f\ (a) = 0 \end{cases} \qquad (\mathrm{II}_o)$$

BORNE DE L'ERREUR

$$\boxed{\left| R\ (y_n) \right| < \frac{1}{h}\ B\ f\ (t_n)} \qquad (\mathrm{II}_1)$$

ETUDE STATISTIQUE

Valeur moyenne

$$\boxed{E\left[R(t_n)\right] = \frac{1}{h} E(G) * f(t_n)} \qquad (II_2)$$

Ecart type

Soient $D(t_n) = E\left[R(t_n) - E\left[R(t_n)\right]\right]^2$

$\sigma^2_G = E\left[G - E(G)\right]^2$

$$R_{n+1} = R_n + h J R_n + G(t_n)$$

$$E(R_{n+1}) = E(R_n) + h J E(R_n) + E\left[G_n\right]$$

$$R_{n+1} - E(R_{n+1}) = R_n - E(R_n) + h J \left[R_n - E(R_n)\right] + G_n - E(G_n)$$

$$D_{n+1} = D_n + 2 h J D_n + 2 E\left[R - E(R)\right]\left[G - E(G)\right] + 2 h J E\left[R - E(R)\right]\left[G - E(G)\right] + \sigma^2_{G_n}$$

$$E\left[R_n(t) - E(R_n(t)\right]\left[G(t) - E\left[G(t)\right]\right] \neq 0$$

$$E\left[R\,G\right] = E(R)\,E(G) = f(t)\left[E(G)\right]^2 \qquad a \leq u < t_n$$

les variables R et G sont pratiquement indépendants

$$\begin{cases} D'_t(t) = 2 J D(t) + \frac{1}{h} \sigma^2_G \\ D(a) = 0 \end{cases} \qquad (II_3)$$

La solution de cette équation différentielle donne l'écart type de R $D(t_n)$

III - RESOLUTION DES SYSTEMES DIFFERENTIELS $m > 1$

$$R(t_n) = \frac{1}{h} \int_a^{t_n} M(t,u)\, \underline{G}(u)\, du$$

BORNE DE L'ERREUR

$$R^k(t_n) = \frac{1}{h} \int_a^{a+nh} \sum_l M_{kl}(t,u)\, G^l(u)\, du$$

$$\left|R^k(t_n)\right| < \frac{B}{h} \int_a^{t_n} \sum_l \left|M_{kl}(t,u)\right| du$$

posons $z^k(t_n) = \int_a^{t_n} \sum_l \left| M_{kl}(t,u) \right| du$

$$z_t^{k'}(tn) = \sum_l \left| M_{kl}(u,u) \right| \frac{dx}{dx} + \int_a^t \sum_l \frac{J \left| M_{kl}(t,u) \right|}{t} du$$

$$\begin{cases} z_t^{k'} \leq 1 + \sum_l \left| J_{kl} \right| z^l \\ z^k(0) = 0 \end{cases}$$

La solution du système d'ordre m

$$\begin{cases} \underline{g}'(t) = \underline{i} + \left| J \right| \underline{g}(t) \\ \underline{g}(0) = \underline{0} \end{cases} \qquad (II_4)$$

dont la solution satisfait à $g^k(t_n) \geq z^k(t_n) > 0$ permet d'écrire :

$$\boxed{\left| R^k(t_n) \right| \leq \frac{1}{h} B\, g^k(t_n)} \qquad (II_5)$$

$\underline{i}$ est le vecteur dont les m composantes sont égales à 1

J est la matrice dont les éléments sont les valeurs absolues des éléments de J

ETUDE STATISTIQUE

Valeur moyenne

a) Le système différentiel est du type (α)

$$E\left[R^k(t_n)\right] = \frac{E(G)}{h} \sum_l \int_a^{t_n} M_{kl}(t,u)\, du$$

$$\boxed{E\left[\underline{R}(t_n)\right] = \frac{E(G)}{h} \underline{f}(t_n)} \qquad \text{où}$$

$\underline{f}(t_n)$ est déterminée par $\begin{cases} \underline{f}'(t) = \underline{i} + J\,\underline{f}(t) \\ \underline{f}(0) = \underline{0} \end{cases}$ (II_6)

b) Le système différentiel est du type (β)

1) $$E\left|R^k(t_n)\right| = \frac{1}{h} \sum_l E(G^l) \int_a^{t_n} M_{kl}(t,u)\,du$$

Posons $$S_{kl}(t,a) = \int_a^t M_{kl}(t,u)\,du$$

$$\boxed{E\left[\underline{R}(t_n)\right] = \frac{1}{h} S\, E(\underline{G})} \qquad (II_7)$$

$$E(\underline{G}) = \begin{cases} E(G^1) \\ \\ E(G^m) \end{cases}$$

S (t) est donnée par $$\begin{cases} S'_t = I + J(t)\,S(t) \\ S(a) = 0 \end{cases} \qquad (II_8)$$

(système différentiel d'ordre m^2)

2) Soit $$P_G = \max_k \left|E(G^k)\right|$$

$$\boxed{\left|E\left(R^k(t_n)\right)\right| \leq \frac{P_G}{h}\, g^k(t_n)}$$

$\underline{g}(t_n)$ est déterminé par II_4

(système différentiel d'ordre m)

Si la matrice J (t) possède des éléments positifs et des éléments négatifs $\frac{P_G}{h}\, g^k(t_n)$ sera une borne très élevée de $\left|E(R^k)\right|$

(le phénomène nous échappe)

Ecart type

Soient $$D^{kl}(t_n) = E\left[R^k(t_n) - E\left[R^k(t_n)\right]\right]\left[R^l(t_n) - E(R^l(t_n))\right]$$

$$\delta_{kl}\,\sigma^2_{kl} = E\left[G^k - E(G^k)\right]\left[G^l - E(G^l)\right]$$

Le calcul analogue à celui de l'écart type dans le cas d'une équation différentielle donne :

$$(D^{kl})'_t = \sum_s (J_{ls}\ D^{sk} + J_{ks}\ D^{sl}) + \frac{\delta_{kl}\ \sigma_{kl}^2}{h}$$

D^{kl} satisfait à :

$$\begin{cases} D'(t) = J(t)\ D(t) + D(t)\ J^T(t) + Q \\ D(a) = 0 \end{cases}$$

(système différentiel d'ordre m^2)

$$Q = I\ \frac{\sigma^2}{h} \qquad \text{(type } \alpha\text{)}$$

$$Q = \frac{1}{h} \begin{bmatrix} \sigma_1^2 & & 0 \\ & \sigma_2^2 & \\ 0 & & \sigma_m^2 \end{bmatrix} \qquad \text{(type } \beta\text{)}$$

IV - ETUDE EXPERIMENTALE

Elle consiste :

- à intégrer une première fois le système différentiel avec un calculateur qui fonctionne normalement, c'est-à-dire en utilisant toutes les positions de ses mémoires soit m . On obtient $\underline{y}\ (t_n)$

- à intégrer une seconde fois exactement le même système ; mais le nombre des positions utilisables, dans toutes les mémoires du calculateur, est réduit à p $p < m$ ce qui donne $\underline{\tilde{y}}\ (t_n)$

Les expériences ont été faites dans le cas de l'erreur de chute. L'étude statistique a été réalisée en intégrant k différentes fois le système de la façon suivante :

A chaque étape du calcul on soustrait automatiquement, du nombre initial (a) placé dans la mémoire, à partir de la position p , un nombre aléatoire compris entre 0 et 1 (simulation d'erreur de chute)

Nous obtenons $\tilde{y}_s\ (t_n)$ $\quad 1 \leqslant s \leqslant k.$

Les k corrections de calcul propagées fournies par l'expérience sont $y\ (t_n) - \tilde{y}_s\ (t_n)$.

Leur moyenne et leur écart type sont notés respectivement par $E\ (\underline{R}_{pr})$ et $\sigma_{\underline{R}_{pr}}$ (les valeurs théoriques par $E\ (\underline{R})$ et $\sigma_{\underline{R}}$)

La méthode d'intégration utilisée est celle de Runge Kutta d'ordre 4. Nous nous sommes placés dans des conditions telles que l'erreur de méthode soit minime comparativement à l'erreur de calcul, et n'intervienne pas dans l'évaluation de R_{pr} .

Les courbes théoriques sont dessinées en "trait plein" les courbes en "trait mixte" représentent les résultats pratiques.

Expérience I.

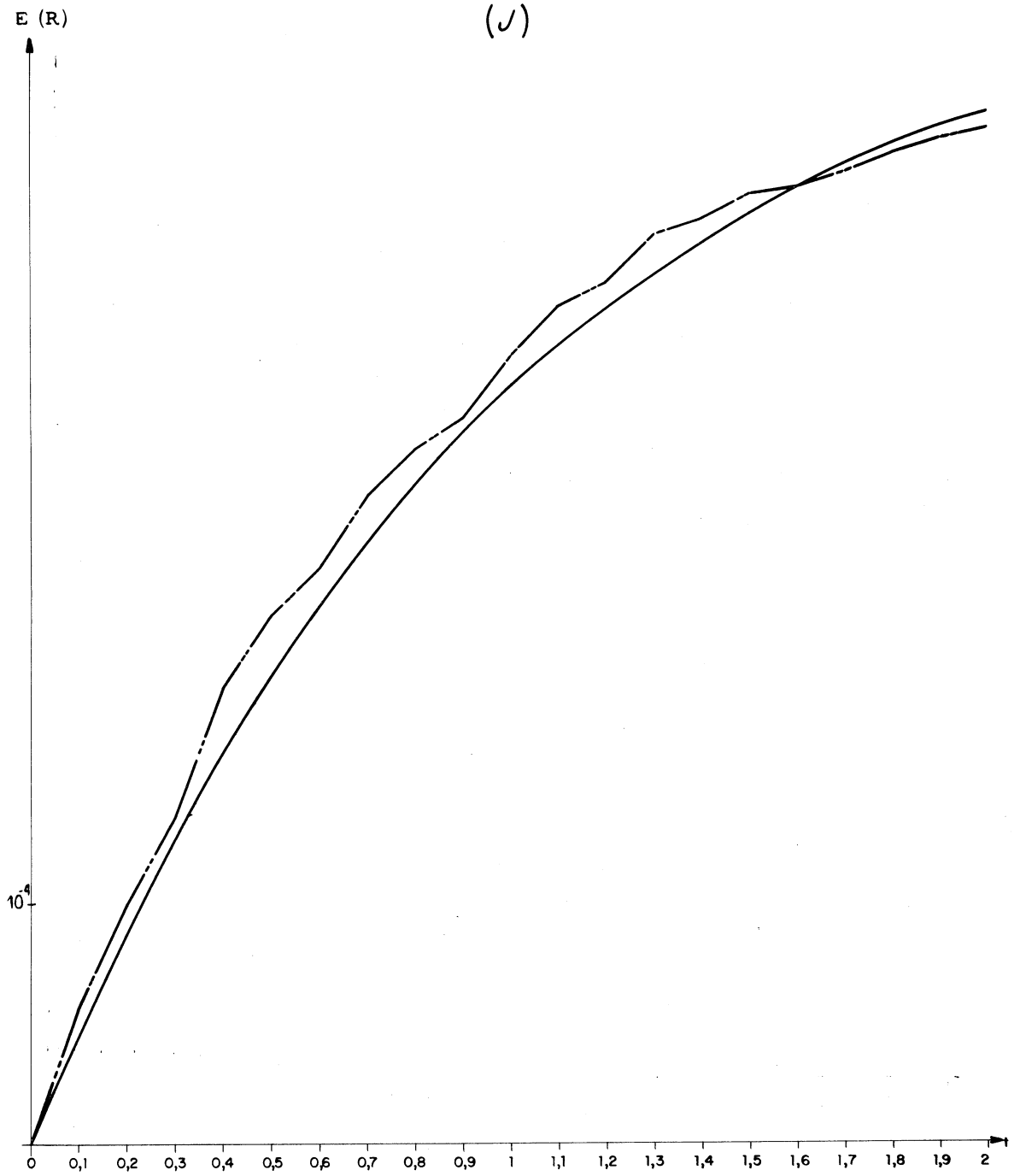
$p = 5 \quad h = 10^{-2}$

$y'_t\ (t) = -\ y\ (t)$ $\quad$ Voir tableau J (k = 10)

$y\ (0) = 1$

a) L'intervalle d'intégration est (0, 2)

- la solution y est de la forme a 10^0 dans cet intervalle
- " h Y " " a 10^{-2} " " "


E (R)
(J)
10^{-9}
0
0,1
0,2
0,3
0,4
0,5
0,6
0,7
0,8
0,9
1
1,1
1,2
1,3
1,4
1,5
1,6
1,7
1,8
1,9
2

h R (Y) est négligeable devant R ($\tilde{y}$ + h $\tilde{Y}$)

R ($\tilde{y}$ + h $\tilde{Y}$) = 2 (y) = x 10^{-p}

Les conditions de la remarque 2 page 74 , sont satisfaites :

G = x 10^{-p}

b) Le système (Π_o) donne f (t) = I - y (t)

et E (R) = 0,5 10^{-3} (I - y)

Voir Π_2 page 76

Expérience II.

p = 5 h = 10^{-2}

y'_t (t) = - $\frac{I}{3}$ t y Voir tableau K (k = 10).

y (0) = 1

L'intervalle d'intégration est (0 ; 3,75)

a) G = x 10^{-p}

b) E (R) = 0,5 10^{-3} f (t)

Expérience III.

p = 6 h = 10^{-2}

y'_t (t) = $\frac{5}{3}$ t + $\frac{4}{7}$ y Voir tableau L (k = 3)

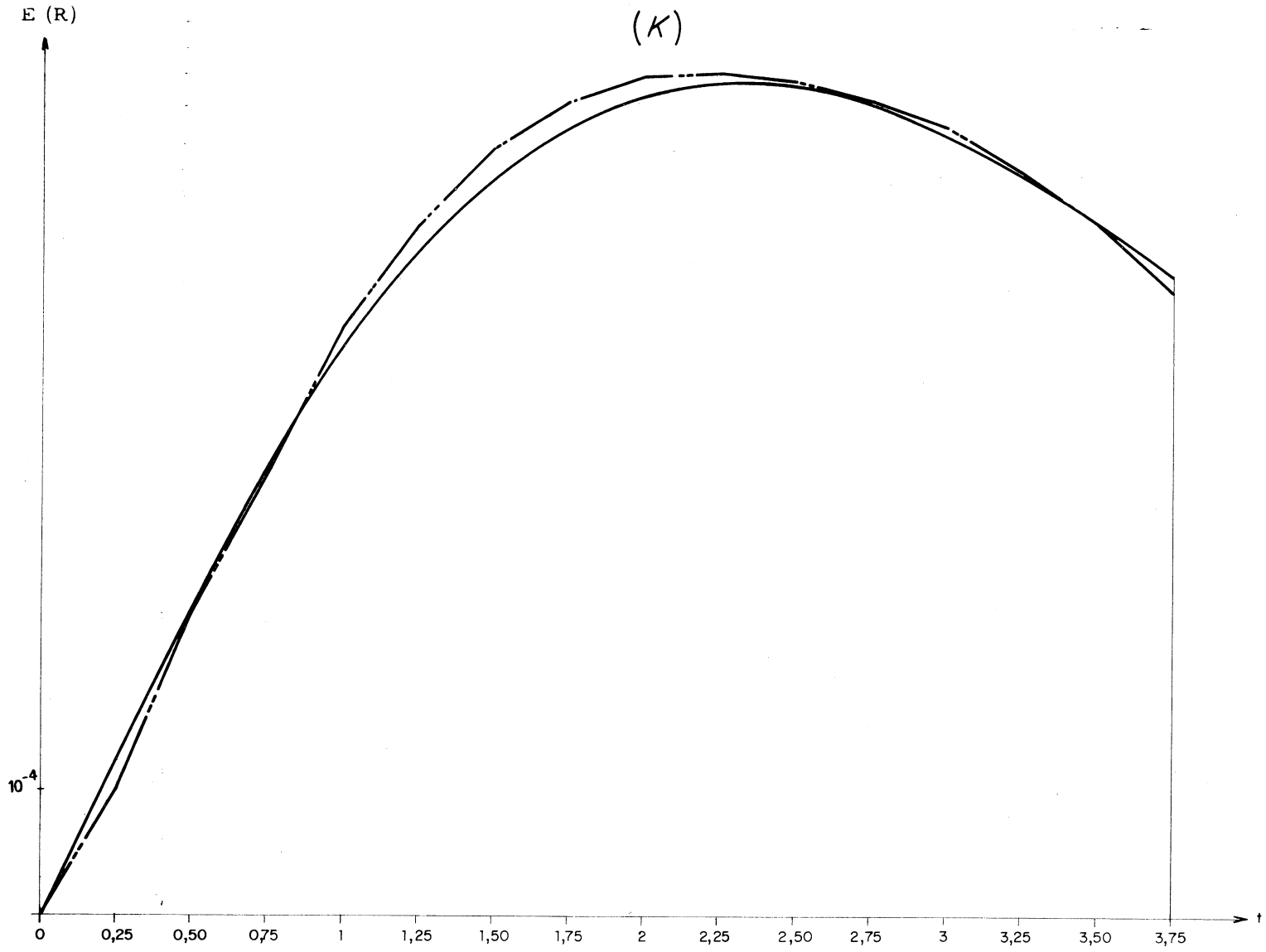
y (0) = 1

L'intervalle d'intégration est (0, 1)

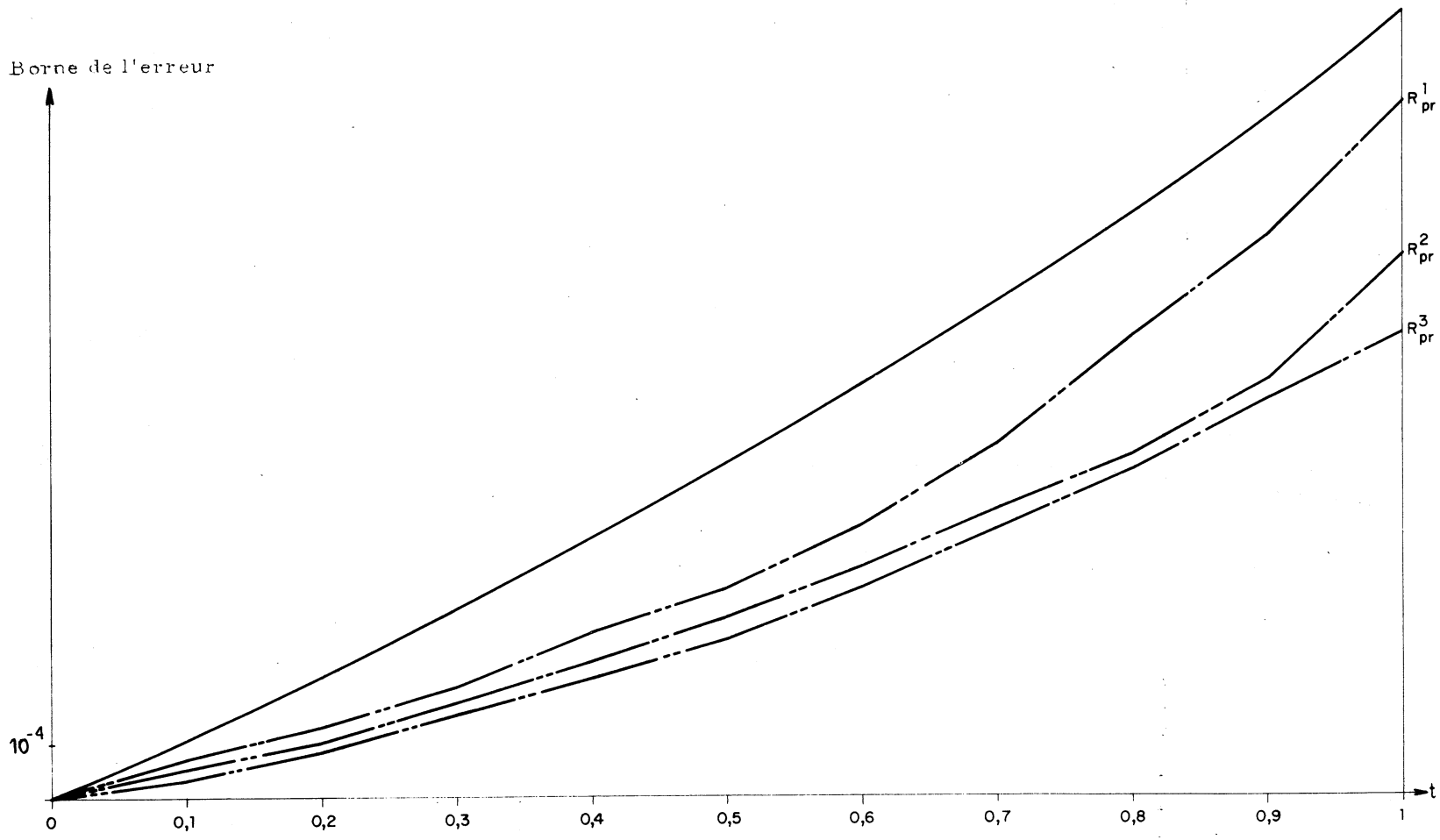
a) h R (Y) + R ($\tilde{y}$ + h $\tilde{Y}$) < 10,6 10^{-p}

b) R (y_n) < 10,6 10^{-4} f (t)

Voir Π_1 page 75


E (R)
(K)
10^{-4}
0
0,25
0,50
0,75
1
1,25
1,50
1,75
2
2,25
2,50
2,75
3
3,25
3,50
3,75
t

(7)


Borne de l'erreur
R^1_{pr}
R^2_{pr}
R^3_{pr}
10^{-4}
0
0,1
0,2
0,3
0,4
0,5
0,6
0,7
0,8
0,9
1
t

Expérience IV.

$p = 5 \qquad h = 10^{-3}$

$y' = y + a z \qquad y(0) = 1$

$z' = z + a y \qquad z(0) = 0$

Voir tableau M
($a = 0,77 \ldots 7$)

$R^1(t_n) = y_n - \tilde{y}_n \qquad R^2(t_n) = z_n - \tilde{z}_n$

L'intervalle d'intégration est (0 ; 1)

Dans cet intervalle nous aurons, sauf pour les premiers pas d'intégration : $10^{-2} < z < 10^{-1}$

tandis que y est toujours de la forme a 10^{-1}

a) Nous prendrons donc :

$G^1 = L(y) = x\ 10^{1-p}$

$G^2 = L(z) = x\ 10^{-1-p}$

le système étudié n'est pas vraiment du type (β)

b) $E(R^1) = 0,5\ 10^{-2}\ (10\ S_{11} + 10^{-1}\ S_{12})$

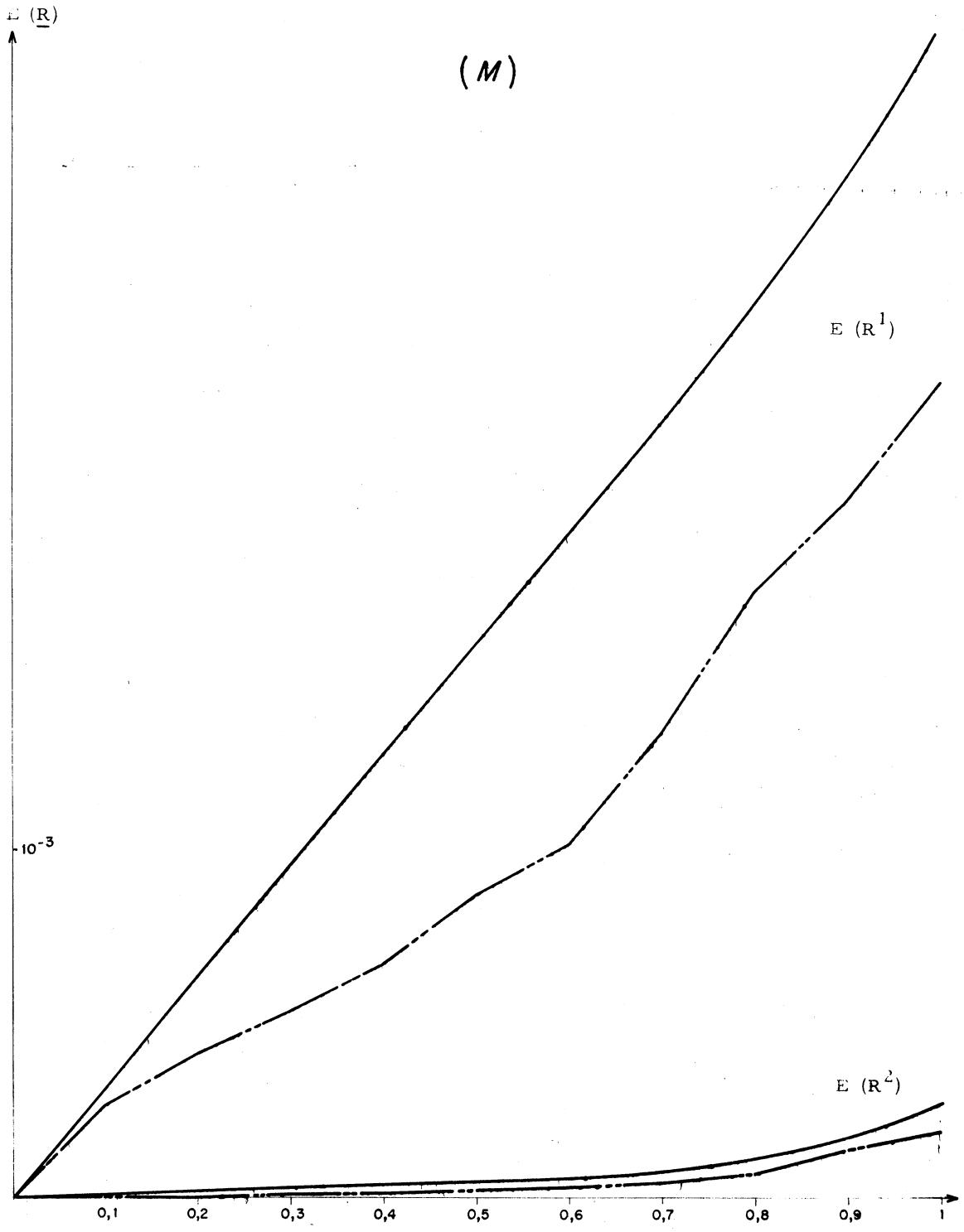
$E(R^2) = 0,5\ 10^{-2}\ (10\ S_{21} + 10^{-1}\ S_{22})$ $\quad \Pi_7$

S est solution du système $\Pi_8 \qquad J = \begin{bmatrix} 1 & a \\ a & 1 \end{bmatrix}$

CONCLUSION

Lorsque la variation de la "correction de calcul locale sur un pas" n'est pas trop grande dans l'intervalle d'intégration considéré (hypothèses p.73 vérifiées) l'évaluation des valeurs moyennes et écarts type de la "correction de calcul propagée" ne présente aucune difficulté.

Les nouveaux systèmes différentiels que l'on doit résoudre sont d'ordre m^2 si l'on veut des résultats satisfaisants. Cette solution est très onéreuse mais il est inutile de conserver le pas d'intégration correspondant à la résolution du système donné. (On prendra un pas nettement plus grand).


E (R)
(M)
E (R¹)
E (R²)
10⁻³
0,1
0,2
0,3
0,4
0,5
0,6
0,7
0,8
0,9
1

BIBLIOGRAPHIE

1. Dahlquist, G., "Convergence and Stability in the Numerical Integration of Ordinary Differential Equations", Mathematica Scandinavica (1956) : 33 - 53

2. Gill, S. "A process for the Step-by-Step Integration of Differential Equations in an Automatic Digital Computing Machine", Proc. Cambridge Philos. Soc. 47 (1951) : 96 - 108.

3. Henrici, P. "Theoretical and Experimental Studies on the Accumulation of Error in the Numerical Solution of Initial Value Problems for Systems of Ordinary Differential Equations", UNESCO/NS/ICIP/ A.1.13.

4. Mikulaschkova Renata. "Erreur d'arrondissement dans le calcul numérique, du point de vue statistique", Pokroky Math. Phys. Ast. 2 (6), 697 - 707 (1957).

5. Rutishauser, H. "Uber die Instabilitat von Methoden zur Integration gewohnlicher Differentialgleichungen", Z. angew. Math. Physik 3 (1952) : 65 - 74.

6. Todd, J. "Notes on Numerical Analysis, I. Solution of Differential Equations by Recurrence Relations", Math. Tables and Other Aids to Computation 4 (1950) : 39 - 44.

7. Wilkinson, J.H. "Rounding Errors in Algebraic Processes" UNESCO/NS/ICIP/A.1.8.

TABLE DES MATIERES

L'impression de ce travail a été effectué par les Services d'Etudes et Laboratoires de la Compagnie IBM France, dont je prie la Direction, en les personnes de Monsieur J. Jeanniot, et de Monsieur M. Papo, de croire à ma sincère reconnaissance.

Je tiens aussi à remercier le personnel qui a contribué à cette réalisation avec compétence et gentillesse.

VU :

Grenoble le
Le Président de la Thèse

VU :

Grenoble le
Le Doyen de la Faculté des Sciences

VU et permis d'imprimer :

Grenoble le
Le Recteur de l'Académie de Grenoble