UNIVERSITÉ de NANCY FACULTÉ des SCIENCES

PRÉSENTATION et MISE AU POINT d'un TEST NON PARAMETRIQUE

GÉNÉRATIONS de NOMBRES PSEUDO-ALÉATOIRES

THESE

pour l'OBTENTION du DOCTORAT de SPÉCIALITÉ
MATHEMATIQUES (3ème Cycle)

SOUTENUE DEVANT LE JURY EN OCTOBRE 1962 par

François GIANNESINI

Jury : Mr J. LEGRAS , Président

Melle HUET)
Mr M. DEPAIX) Examinateurs

UNIVERSITÉ de NANCY FACULTÉ des SCIENCES

PRÉSENTATION et MISE AU POINT d'un TEST NON PARAMETRIQUE

GÉNÉRATIONS de NOMBRES PSEUDO-ALÉATOIRES

THESE

pour l'OBTENTION du DOCTORAT de SPÉCIALITÉ
MATHEMATIQUES (3ème Cycle)

SOUTENUE DEVANT LE JURY EN OCTOBRE 1962 par

François GIANNESINI

Jury : Mr J. LEGRAS, Président

Melle HUET) Examinateurs
Mr M. DEPAIX)

UNIVERSITE de NANCY
FACULTE des SCIENCES

Doyen : M. ROUBAULT
Assesseur : M. AUBRY

Doyens honoraires : MM. CORNUBERT, DELSARTE, URION

Professeurs honoraires : MM. GUTTON, CROZE, RAYBAUD, LAFFITTE, LERAY, JOLY, LAPORTE, EICHHORN, CAPELLE, GODEMENT, DUBREIL, L. SCHWARTZ, DIEUDONNE, de MALLEMANN, LONGCHAMBON, LETORT, DODE, GAUTHIER, GOUDET, OLMER, CORNUBERT, CHAPELLE, GUERIN, CHEVALLIER.

Maîtres de conférences honoraires : MM. RAUX, LIENHART

PROFESSEURS

MM.			
URION	Chimie biologique	LIONS	M.M.P.
DELSARTE	Analyse supérieure	SUHNER	Physique expérimentale
ROUBAULT	Géologie	HILLY	Géologie
VEILLET	Biologie animale	LE GOFF	Génie Chimique
ECHEVIN	Botanique	CHAPON	Chimie biologique
BARRIOL	Chimie théorique	HEROLD	Chimie industrielle
BIZETTE	Physique	SCHWARTZ	Exploit. minière
GUILLIEN	Electronique	GAYET	Physiologie
GIBERT	Chimie Physique	MANGENOT	Phytopathologie
HERVE	Calcul dif. et intég.	MALAPRADE	Chimie
LEGRAS	Mécanique rationnelle	HADNI	Physique
DAVID	Chimie organique	DELAMARE-DEBOUTEVILLE	Zoologie
BOLFA	Minéralogie		
NICLAUSE	Chimie	BONVALET	Mécanique physique
FAIVRE	Physique appliquée	Mme ROIZEN	Physique
AUBRY	Chimie minérale	KERN	Minéralogie
DUVAL	Chimie	BASTICK	Chimie
COPPENS	Radiogéologie	DUCHAUFOUR	Pédologie
FRUHLING	Physique	NEEL	Chimie ind. organique

MAITRES de CONFERENCES

MM.			
WERNER	Botanique	VUILLAUME	Psychophysiologie
GARNIER	Agronomie	PLAN	Mécanique physique
REGNIER	Physico-chimie	Mme BASTICK	Chimie MPC (Epinal)
WEPPE	Minéralogie appl.	GUDEFIN	Physique
BERNARD	Géologie	HORN	Physique
RENARD	Physique théor. &nucl.	CERF	Mathématiques générales
CONDE	Zoologie	FRENTZ	Biologie animale
GOSSE	Génie chimique	Mme HERVE	Math. (propédeutique)
CHAMPIER	Physique	AUROUZE	Paléontologie
GAY	Chimie biologique	FELDEN	Physique M.P.C. (Epinal)
ROCCI	Géologie		

CHARGES D'ENSEIGNEMENT

MM.			
MARI	Chimie (I.S.I.N.)	EYMARD	Math. (propédeutique)
LAFON	Physique (I.S.I.N.)	DANYSZ	Mécanique physique (I.S.I.N.)

Secrétaire Principal : M. CARON

Que Monsieur le Professeur J. LEGRAS, veuille bien trouver ici l'assurance de ma gratitude pour toutes les facilités qu'il m'a offertes durant les deux années que j'ai passées au Centre de Calcul, et Monsieur M. DEPAIX, Chargé d'Enseignement, l'expression de tous mes remerciements pour avoir constamment dirigé mon travail.

Enfin, je remercie Messieurs les Professeurs qui ont bien voulu me faire l'honneur de composer le Jury.

I N T R O D U C T I O N

1. Contrôle de l'homogénéité d'une fabrication -

Il arrive fréquemment dans l'industrie que l'on ait à s'assurer de l'invariance des caractéristiques d'un produit, à la sortie d'une chaîne de fabrication. En général, il s'agit simplement de déceler une usure excessive d'un des éléments de la chaîne, mais on peut aussi chercher à modifier un des éléments du processus de fabrication, sans que cela entraîne une modification de la production.

Il est nécessaire d'effectuer alors un contrôle d'homogénéité. Voici les méthodes qui sont le plus couramment employées dans ce but.

a) Etant donné un échantillon prélevé à la sortie de la chaîne, et son histogramme représentatif, il s'agit, moyennant l'allure de cet histogramme, d'ajuster cet échantillon à une loi de probabilité. Or, si la nature de cette loi peut être suggérée par l'histogramme, ses paramètres sont inconnus et il est nécessaire de les estimer à partir de l'échantillon.

Si ce premier point peut être réalisé, il sera toujours possible par la suite, certaines modifications étant intervenues dans le processus de fabrication, de prélever un deuxième échantillon, et de tenter le même ajustement. Le résultat de cette deuxième opération permettra de conclure quant à l'homogénéité de la fabrication.

Ce procédé, en théorie très séduisant, est malheureusement inapplicable dans la plupart des cas. En effet, rien ne permet de prévoir que la population est déterminée par une loi de probabilité simple. En fait, cette loi est bien souvent une superposition de lois simples, qu'il est extrêmement difficile de mettre en évidence. On trouvera un exemple de ce type de difficultés dans le travail effectué au Centre de Calcul par Monsieur Colas.

b) Un deuxième procédé consiste à considérer simultanément deux échantillons et à vérifier l'homogénéité en appliquant un test du χ^2. On est alors conduit à comparer des fréquences observées à des fréquences théoriques que l'on estime à partir des valeurs contenues dans les deux échantillons.

Malheureusement, on retrouve dans ce procédé tous les inconvénients inhérents au test du χ^2 : La répartition en classes, et le nombre d'éléments contenus dans chaque classe. En effet, le choix des classes dépend le plus souvent de l'utisateur, et il existe des exemples montrant qu'une modification dans la définition des classes influe dans une grande mesure sur les conclusions du test.

De plus, dans chaque classe, la variable définie ne peut être assimilée à un χ^2 que par passage à la limite. Sans parler du désaccord qui existe entre les différents auteurs sur le nombre d'éléments que doit contenir une classe pour que le passage à la limite puisse être effectué, on constate que le nombre de valeurs contenues dans les échantillons considérés doit être important, ce qu'il n'est pas toujours possible de réaliser.

Dans les deux méthodes que nous venons de voir, nous avons été conduits à estimer certaines valeurs inconnues : paramètres de la loi théorique ou probabilités théoriques. Ces méthodes relèvent des tests dits "tests paramétriques".

c) Un dernier procédé consiste à utiliser un des "tests non paramétriques". Nous n'aurons alors à considérer que les valeurs respectives composant deux échantillons, sans aucune référence à des valeurs théoriques qu'il faudrait estimer.

Le test de Gnédenko appartient à cette dernière catégorie.

2. Plan suivi dans l'étude du test de Gnédenko -

Nous commencerons par l'étude théorique du test en exposant la méthode utilisée par Gnédenko lui-même, puis en donnant une méthode plus moderne, faisant intervenir la notion d' images de probabilité, notion due à Monsieur Tortrat. Nous donnerons alors les résultats pour des valeurs différentes de n, nombre de valeurs contenues dans les échantillons considérés.

Puis nous chercherons la limite de la loi de Gnédenko lorsque le nombre n devient très grand, on sait que la loi limite ainsi obtenue est la loi de Kolmogorroff-Smirnoff. Nous déterminerons ensuite la valeur de n, à partir de laquelle cette loi limite est pratiquement utilisable.

Nous recommencerons le même travail, non plus sur deux échantillons, mais en essayant d'ajuster un échantillon à une loi théorique.

Cela nous conduira à prélever dans une loi théorique un ensemble de valeurs au hasard donc à disposer sur un axe une série de points, dont les abcisses seront équiréparties entre 0 et 1.

3. Génération de nombres pseudo-aléatoires -

Nous donnerons un procédé de génération de nombres pseudo-aléatoires, nous justifierons ce procédé par application des théorèmes de Weyl et de Riesz, nous exposerons alors les vérifications effectuées sur l'ensemble des nombres ainsi obtenus.

4. Ayant alors la possibilité de prélever un échantillon dans une loi théorique, nous pourrons comparer l'efficacité du test de Gnédenko, à celle du test du χ^2 par exemple.

E T U D E D U T E S T D E G N E D E N K O

I - Position du Problème

II - Fonction de répartition des écarts maximum de deux répartitions empiriques

III - Démonstration de Gnédenko

IV - Autre démonstration

V - Loi limite

VI - Résultats

VII - Conclusion

E T U D E D U T E S T D E G N E D E N K O

I - Position du problème -

Considérons les résultats de deux séries de mesures :

(1) $x_1, x_2, \ldots\ldots\ldots x_n$

(2) $y_1, y_2, \ldots\ldots\ldots y_m$

et cherchons à tester l'hypothèse : les suites ont la même fonction de répartition.

Lors d'un contrôle industriel, on considère d'habitude que la loi est normale. En réalité, la fonction de répartition véritable est généralement inconnue, d'où la nécessité de rechercher, pour tester l'homogénéité, des méthodes qui ne dépendent pas d'hypothèses sur la nature des fonctions de répartition des variables aléatoires étudiées.

L'étude qui suit est due à une idée de Kolmogoroff largement développée par Smirnoff.

1. Théorème de Smirnoff.

- Soient (1) et (2) les résultats de mesures indépendants de variables aléatoires ayant la même fonction de répartition continue.

- Soit $k_1(x)$ le nombre des x_k inférieurs à x, et $k_2(x)$ le nombre des y_k inférieurs à x.

- Considérons :

$$D(m, n) = \sup \left| \frac{k_1(x)}{n} - \frac{k_2(x)}{m} \right|$$

Alors, si le rapport $\frac{n}{m}$ admet une limite finie a lorsque n tend vers l'infini :

$$\lim_{n \longrightarrow \infty} \text{Prob} \left[\sqrt{\frac{mn}{m+n}} \; D(m, n) < x \right] = \sum_{k=-\infty}^{+\infty} (-1)^k e^{-2k^2x^2}$$

Il s'agit évidemment d'un théorème asymptotique inapplicable pratiquement. Il faut donc trouver une loi exacte donnant $\text{Prob} \left[\sqrt{\frac{mn}{m+n}} \; D(m,n) < x \right]$ en fonction de x, m et n, et étudier la vitesse de convergence de cette loi exacte vers la loi limite.

Or dans le cas où m = n, on peut effectivement trouver cette loi exacte, moyennant l'hypothèse que la fonction de répartition de la variable examinée est continue.

2. Utilisation pratique des résultats.

- Dans tout ce qui suivra, nous prendrons donc m = n, c'est-à-dire que nous prendrons toujours des échantillons comportons le même nombre de valeurs.

- Nous poserons alors $D_n = D(n,n)$

Exemple :

n=6 (1) 6,46 6,42 6,52 6,44 6,36 6,40

(2) 6,26 6,38 6,46 6,34 6,26 6,32

- Nous allons chercher non pas D_n, mais

$$n D_n = \max \left| k_1(x) - k_2(x) \right|$$

⇐ Pour cela, rangeons nos 2 échantillons par valeurs croissantes :

				6,36		6,40	6,42	6,44	6,46	6,52
6,26	6,26	6,32	6,34		6,38				6,46	

- Alors $k_1(x) - k_2(x)$ vaut successivement :

$$-1; -2; -3; -4; -3; -4; -2; -1; -1; 0$$

- Donc dans ce cas $n D_n = 4$

En se reportant aux tables données, pour n = 6, on constate que la probabilité d'un tel écart ou d'un écart supérieur vaut environ 0,143.

Naturellement si n augmente, la méthode devient plus précise.

Sur le plan purement pratique, donnons quelques résultats extraits des tables de résultats :

On pourra parler d'une modification de la fonction de répartition avec une probabilité supérieure à 0,95 si pour un échantillon de n valeurs, la grandeur $n D_n$ atteint on dépasse les valeurs précisées dans le tableau.

n :	9	10	12	15	20	25	30	35	50	75	100
$n D_n$:	6	6	7	8	9	10	11	12	14	17	19

II - Fonction de répartition des écarts maximum de deux répartitions empiriques -

1. Enoncé des problèmes fondamentaux.

Soient $x_1, x_2 \ldots x_n$, $y_1, y_2 \ldots y_n$

les résultats de deux séries d'observations différentes faites sur des variables aléatoires ayant la même fonction de répartition $F(x)$ que nous supposons continue.

Construisons les fonctions de répartition empiriques

$$S_n(x) = \frac{k_1(x)}{n} \qquad \text{et} \qquad T_n(x) = \frac{k_2(x)}{n}$$

Introduisons :

$$D_n^+ = \max_{-\infty < x < +\infty} \left\{ S_n(x) - T_n(x) \right\}$$

$$D_n^- = -\min_{-\infty<x<+\infty}\left\{S_n(x) - T_n(x)\right\} = \max_{-\infty<x<+\infty}\left\{T_n(x) - S_n(x)\right\}$$

et $$D_n = \max\left\{D_n^-, D_n^+\right\} = \max_{-\infty<x<+\infty}\left|S_n(x) - T_n(x)\right|$$

Nous allons d'abord chercher les répartitions

$$\varphi_n^+(x) = \text{Prob}\left|\sqrt{\frac{n}{2}}\ D^+ < x\right| \qquad \varphi_n^-(x) = \text{Prob}\left|\sqrt{\frac{n}{2}} D_n^- < x\right|$$

$$\varphi_n(x) = \text{Prob}\left|\sqrt{\frac{n}{2}}\ D_n < x\right|$$

et $$\varphi_n(x,y) = \text{Prob}\left|\sqrt{\frac{n}{2}}\ D_n^- < x\ ;\ \sqrt{\frac{n}{2}}\ D_n^+ < y\right|$$

2. Premières égalités.

Disposons les résultats des deux séries par grandeur croissante, dans une série unique :

$$z_1 < z_2 < z_3 \cdots\cdots < z_{2n}$$

Les probabilités de disposition des x_i et y_i dans cette suite sont les mêmes. En effet, comme x_i et y_i ont la même fonction de répartition $F(x)$, la probabilité d'apparition de chacune des $(2n)!$ dispositions possibles est égale à :

$$1 = \int_{-\infty}^{+\infty}\left[\int_{z_1}^{+\infty}\int_{z_2}^{+\infty}\cdots\cdots\int_{z_{2n-1}}^{+\infty} dF(z_{2n})\, dF(z_{2n-1})\cdots dF(z_2)\right] dF(z_1)$$

En sommant de droite à gauche :

$$\int_{z_{2n-1}}^{+\infty} d\,F(z_{2n}) = 1 - F(z_{2n-1})$$

$$\int_{z_{2n-2}}^{+\infty} \left[1-F(z_{2n-1})\right] d\,F(z_{2n-2}) = \left[-\frac{(1-F(z))^2}{2} \right]_{z_{2n-2}}^{+\infty}$$

$$= \frac{(1 - F(z_{2n-2}))^2}{2}$$

d'où la valeur finale de l'intégrale : $I = \frac{1}{(2n)!}$

Associons à z_k la variable ξ_k $\begin{cases} = +1 \text{ si } z_k \text{ est l'un des } x_i \\ = -1 \text{ si } z_k \text{ est l'un des } y_j \end{cases}$

Examinons alors :

$S_0 = 0$

$S_1 = \xi_1$

$S_2 = \xi_1 + \xi_2$

.....

$S_{2n} = \xi_1 + \xi_2 + \ldots\ldots \xi_{2n}$

Par définition, D_n^+ est égale au quotient par n de la plus grande différence entre le nombre d'observations de la première série et celui de la deuxième série comprise dans $(-\infty, x)$ quand x varie de $-\infty$ à $+\infty$.

Ainsi $n D_n^+ = \max_{1 \leq k \leq 2n} S_k$

De même $n D_n = \max_{1 \leq k \leq 2n} |S_k|$

et $n D_n^- = - \min_{1 \leq k \leq 2n} S_k$

3. Méthode de recherche.

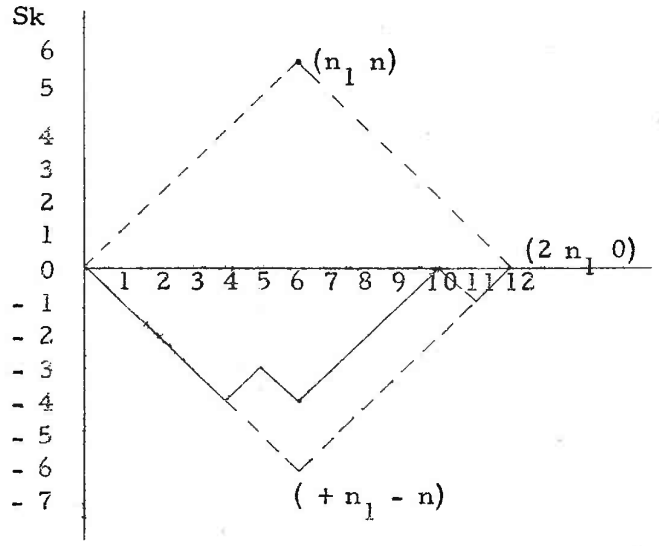
Utilisons l'illustration :

Une particule est au temps $t = 0$ dans la position S_0. Elle subit aux instants $t = 1, 2, \ldots 2n$ des chocs aléatoires qui peuvent la déplacer de + 1 ou de - 1, suivant les valeurs de ξ_k au temps $t = k$. Il y aura n chocs positifs et n chocs négatifs. Dans le plan t, S chaque déplacement sera représenté par le vecteur $(1, \xi_k)$. Chaque trajectoire joint (0,0) à (2n,0).

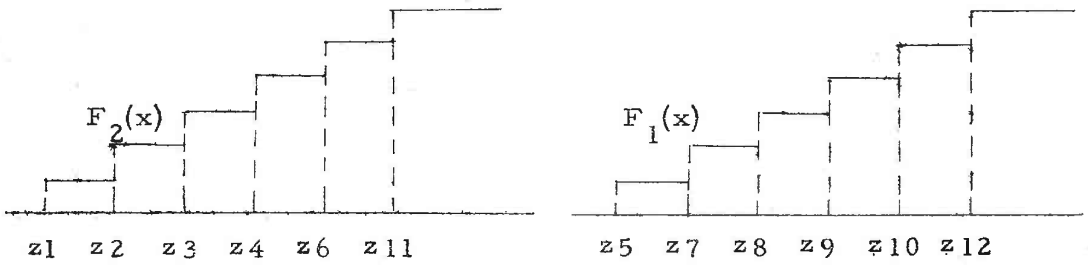
Représentons la trajectoire correspondant à l'exemple déjà donné :

				6,36		6,40	6,42	6,44	6,46	6,52
6,26	6,26	6,32	6,34		6,38				6,46	

$S_k = - 1; - 2; - 3; -4; - 3; - 4; - 3; - 2; - 1; 0; - 1; 0$



On peut en déduire les fonctions de répartitions empiriques (les valeurs des abcisses restent inconnues).



Le nombre des trajectoires possibles est C_{2n}^{n} ; c'est en effet le nombre de façons de répartir n chocs positifs sur 2n positions. Toutes ces trajectoires étant également probables, la probabilité de chacune d'elle est

$$\frac{1}{C_{2n}^{n}} = \frac{(n!)^2}{(2n)!}$$

Donc pour trouver :

$\text{Prob}\left[n D_n^+ < s\right]$ il suffit de calculer le nombre des trajectoires qui se trouvent sous la droite $S = s$

$\text{Prob}\left[n D_n < s\right]$ il suffit de calculer le nombre des trajectoires comprises entre les droites $S = -s$ et $S = s$

$\text{Prob}\left[n D_n^- < s \; ; \; n D_n^+ < t\right]$ il suffit de calculer le nombre des trajectoires qui se trouvent entre les droites $S = s$ et $S = t$.

- Remarquons que pour tout s et t vérifiant les égalités : $c \leqslant s < c+1 \qquad d \leqslant t < d+1$

On a :

$$\text{Prob}\left[n D_n^+ < s\right] = \text{Prob}\left[n D_n^+ < c\right]$$

$$\text{Prob}\left[n D_n < s\right] = \text{Prob}\left[n D_n < c\right]$$

$$\text{Prob}\left[n D_n^- < c \; ; \; n D_n^+ < t\right] = \text{Prob}\left[n D_n^- < c; \; n D_n^+ < d\right]$$

4. Théorème sur les écarts unilatéraux.

Théorème 1. Si les résultats des observations (1) et (2) sont indépendants et soumis à une même fonction de répartition continue, on a :

$$\varphi_n^+(x) = \begin{cases} 0 & \text{pour } x \leqslant 0 \\ 1 - \dfrac{C_{2n}^{n-c}}{C_{2n}^{n}} & \text{pour } 0 < x \leqslant \sqrt{\dfrac{n}{2}} \\ 1 & \text{pour } x > 1 \end{cases}$$

Avec $c = -\left[-x\sqrt{2n} \right]$; c est l'opposé de la partie entière de $-x\sqrt{2n}$; exemple $x\sqrt{2n} = 3,5$
$c = 4$

Démonstration :

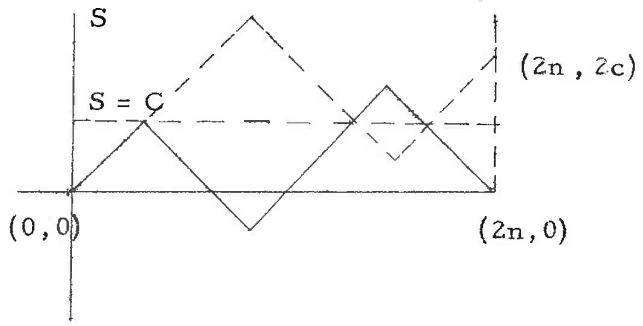
$$\varphi_n^+(x) = \text{Prob}\left[\sqrt{\frac{n}{2}}\, D_n^+ < x\right] = \text{Prob}\left[nD_n^+ < x\sqrt{2n}\right]$$

$$= \text{Prob}\left[nD_n^+ < c\right]$$

Il suffit donc de calculer le nombre de trajectoires situées sous la droite $S = c$; on, comme

$$\text{Prob}\left[n\, D_n^+ < c\right] = 1 - \text{Prob}\left[n\, D_n^+ \geqslant c\right]$$

il suffit de calculer le nombre de trajectoires qui touchent ou qui coupent la droite $S = c$.



A chaque trajet de (0,0) à 2n,0) coupant $S = c$, correspond un trajet identique au précédent jusqu'à $S = c$, puis symétrique par rapport à $S = c$. Un tel trajet joint $S = c$ au point (2n, 2c).

Le nombre de trajets est : $C_{2n}^{n+c} = C_{2n}^{n-c}$

D'où la formule figurant dans le Théorème donné.

On constate que c peut au plus être égal à n, ce qui donne l'intervalle de variation de x.

- Théorème 2 : Sous les hypothèses du Théorème 1, on a :

$$\varphi_n(x) = \begin{cases} - \ 0 & \text{pour } x < -\frac{1}{\sqrt{2n}} \\ - \sum_{k=-\left[\frac{n}{c}\right]} (-1) \ \frac{C_{2n}}{C_{2n}^{n}} & \text{pour } -\frac{1}{\sqrt{2n}} < x < \sqrt{\frac{n}{2}} \\ - \ 1 & \text{pour } x > \sqrt{\frac{n}{2}} \end{cases}$$

- Théorème 3 : Sous les hypothèses du Théorème 1, on a :

$$\varphi_n(x,y) = \begin{cases} 0 \text{ pour } \min(x,y) < -\frac{1}{\sqrt{2n}} \\ 1 \text{ pour } \min(x,y) > \sqrt{\frac{n}{2}} \end{cases}$$

et pour les autres valeurs de x et de y. :

$$\varphi_n(x,y) = \frac{1}{C_{2n}^{n}} \left\{ \sum_{k=-\left[\frac{n}{c+d}\right]}^{\left[\frac{n}{c+d}\right]} C_{2n}^{n-k(c+d)} - \sum_{k=1}^{\left[\frac{n+c}{c+d}\right]} C_{2n}^{n+c-k(c+d)} - \sum_{k=1}^{\left[\frac{n+d}{c+d}\right]} C_{2n}^{n+d-k(c+d)} \right\}$$

où $c = -\left[-x\sqrt{2n}\right]$ et $d = -\left[-y\sqrt{2n}\right]$

Il est à remarquer que les théorèmes 1 et 2 sont des cas particuliers du théorème 3.
En effet

$$\varphi_n^+(x) = \varphi_n\left(\sqrt{\frac{n}{2}}, x\right)$$

$$\varphi_n(x) = \varphi_n(x, x)$$

Nous donnerons donc uniquement la démonstration du théorème 3.

$$\varphi_n(x,y) = \text{Prob}\left[\sqrt{\frac{n}{2}} D_n^- < x \; ; \; \sqrt{\frac{n}{2}} D_n^+ < y\right]$$

$$= \text{Prob}\left[n D_n^- < c \; ; \; n D_n^+ < y\right]$$

$\varphi_n(x,y)$ représente donc le rapport du nombre de trajectoires situées entre les droites $S = -c$, que nous appellerons droite (c), et $S = d$, que nous appellerons droite (d), au nombre total de trajectoires.

III - Démonstration de Gnédenko -

Divisons l'ensemble $\mathcal{M}$ des trajectoires en sous-ensembles disjoints :

- a_0 : trajectoires qui ne coupent ni (c) ni (d)
- a_1 : trajectoires qui coupent (c) mais qui ne coupent pas (d)
- b_1 : trajectoires qui ne coupent pas (c) mais qui coupent (d)

- a_2 : trajectoires qui coupent (c) puis (d) et n'atteignent plus (c)

  b_2 : trajectoires qui coupent (d) puis (c) et n'atteignent plus (d)

On a alors : $\mathcal{M} = a_0 + \sum_{i=1} a_i + b_i$

Cependant la puissance de tels ensembles est difficile à évaluer. Nous les remplacerons donc par d'autres sous-ensembles, ainsi définis.

- A_1 : trajectoires coupant (c) au moins une fois

  B_1 : trajectoires coupant (d) au moins une fois

- A_2 : trajectoires coupant (c) au moins une fois puis (d)

  B_2 : trajectoires coupant (d) au moins une fois puis (c)

- A_3 : trajectoires coupant (c) au moins une fois puis (d) puis (c) au moins une fois.

.....

Avant de chercher le nombre d'éléments contenus dans chacun de ces ensembles, remarquons que $\mathcal{M}$ ne s'exprime pas aussi simplement en fonction des A_i et B_i (qui ne sont pas des ensembles disjoints), qu'en fonction des a_i et b_i. Aussi nous allons commencer par chercher une relation entre ces deux types de sous-ensembles.

Des définitions données, il résulte que :

$$A_1 = a_2 + \sum_{i=2} (a_i + b_i) \qquad B_1 = b_1 + \sum_{i=2} (a_i + b_i)$$

$$A_2 = a_2 + \sum_{i=3} (a_i + b_i) \qquad B_2 = b_2 + \sum_{i=3} (a_i + b_i)$$

.....

$$A_j = a_j + \sum_{i=j+1} (a_i+b_i) \qquad B_j = b_j + \sum_{i=j+1} (a_i+b_i)$$

.....

Alors $$A_{2i-1} - A_{2i} = a_{2i-1} + \sum_{k=2i} (a_k+b_k) - a_{2i} - \sum_{k=2i+1} (a_k+b_k)$$

$$A_{2i-1} - A_{2i} = a_{2i-1} + b_{2i}$$

De même

$$B_{2i-1} - B_{2i} = a_{2i} + b_{2i-1}$$

D'où

$$A_{2i-1} + B_{2i-1} - A_{2i} - B_{2i} = a_{2i-1} + b_{2i-1} + a_{2i} + b_{2i}$$

Donc $$\mathcal{M} = a_0 + \sum_{i=1} (A_{2i-1} + B_{2i-1} - A_{2i} - B_{2i})$$

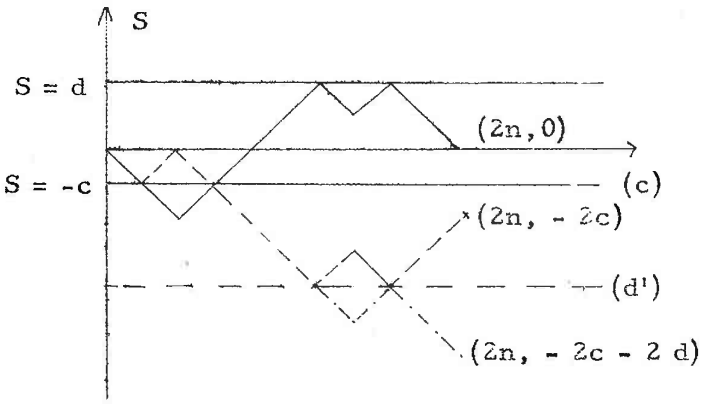
Dans cette somme nous cherchons le nombre d'éléments contenus dans l'ensemble a_0. Nous connaissons le nombre d'éléments contenus dans $\mathcal{M}$, il reste à évaluer le nombre d'éléments contenus dans A_i et B_i.

Le calcul a déjà été fait pour A_1 : C_{2n}^{n-c}

de même pour B_1 : C_{2n}^{n-d}

Calculons le nombre de trajectoires contenues dans A_2.

Pour cela appelons (d') la droite symétrique de (d) par rapport à (c).



Formons alors le trajet suivant :

- La trajectoire jusqu'à (c)
- Le trajet symétrique par rapport à (c) de la partie comprise entre (c) et (d)
- Le trajet symétrique par rapport à (d') du symétrique par rapport à (c) de la partie de la trajectoire comprise entre (d) et l'axe horizontal.

On aboutit alors, au point (2n, - 2c - 2d) et le nombre des trajectoires joignant ce point à l'origine est : $C_{2n}^{n-(c+d)}$

On aura de même pour A_{2i-1} $C_{2n}^{n-ic-(i-1)d}$ car on coupe (c) i fois et (d) i-1 fois

A_{2i} $C_{2n}^{n-ic-id}$

B_{2i-1} $C_{2n}^{n-(i-1)c-id}$

B_{2i} $C_{2n}^{n-ic-id}$

Rappelons que le nombre de trajectoires contenues dans est C_{2n}^{n} ; alors le nombre de trajectoires de a_0 est

$$(a_0) = C_{2n}^{n} - \sum_{i=1} C_{2n}^{n-ic-(i-1)d} + C_{2n}^{n-(i-1)c-id} - 2C_{2n}^{n-ic-id}$$

$$(a_0) = C_{2n}^{n} + 2\sum_{i=1} C_{2n}^{n-i(c+d)} - \sum_{i=1} C_{2n}^{n+c-i(c+d)} - \sum_{i=1} C_{2n}^{n+d-i(c+d)}$$

Or $$2\sum_{i=1}^{\left[\frac{n}{c+d}\right]} C_{2n}^{n-i(c+d)} = \sum_{i=-\left[\frac{n}{c+d}\right]}^{\left[\frac{n}{c+d}\right]} C_{2n}^{n-i(c+d)} - C_{2n}^{n}$$

d'où :

$$\varphi_n(x,y) = \frac{1}{C_{2n}^{n}} \left[\sum_{i=-\left[\frac{n}{c+d}\right]}^{\left[\frac{n}{c+d}\right]} C_{2n}^{n-i(c+d)} - \sum_{i=1}^{\left[\frac{n+c}{c+d}\right]} C_{2n}^{n+c-i(c+d)} - \sum_{i=1}^{\left[\frac{n+d}{c+d}\right]} C_{2n}^{n+d-i(c+d)} \right]$$

IV - Autre démonstration -

Sur une trajectoire, le passage d'un état au suivant peut se caractériser par la fonction génératrice de passage : $Y = \frac{1}{2}\left(u + \frac{1}{u}\right)$. Ce qui donne le processus suivant :

Au bout de k opération, l'extrêmité de la trajectoire est un point d'abcisse k et d'ordonnée $i \leqslant k$. Représentons un tel point par u^i ; si la probabilité d'aboutir en un tel point est p_i, nous représenterons l'état examiné par $p_i u^i$. Après l'opération suivante, le nouvel état sera caractérisé par

$$p_i u^i \times \frac{1}{2}\left(u + \frac{1}{u}\right) = \frac{1}{2} p_i u^{i+1} + \frac{1}{2} p_i u^{i-1}$$

Ce qui signifie que la trajectoire aboutira au point $(k+1 ; i+1)$ avec la probabilité $\frac{1}{2}p_i$ ou au point $(k+1 ; i-1)$ avec la probabilité $\frac{1}{2}p_i$

La probabilité affectée à l'origine sera 1 puisque c'est le point de départ obligatoire nous noterons donc l'état initial par $g^0(u) = 1 \times u^0 = 1$

Après n opération l'état de la trajectoire sera donc donné par

$$g^n(u) = g^0(u) \times Y^n$$

$$g^n(u) = g^0(u) \times \frac{1}{2^n} \left(u + \frac{1}{u}\right)^n$$

$$g^n(u) = \frac{1}{2^n} \left[C_n^0 u^n + C_n^1 u^{n-2} + \ldots\ldots + C_n^p u^{n-2p} + \ldots. + C_n^n u^{-n}\right]$$

Ceci indique que le point représentatif de l'extrêmité sera le point $(n, n-2p)$ affecté de la probabilité $\frac{Cn^p}{2^n}$. Or, la probabilité d'arriver en un point est évidemment égale au rapport du nombre de trajectoires aboutissant en ce point, au nombre total de trajectoires. Or, le nombre total de trajectoires après n opérations est 2^n. On en déduit que le nombre total de trajectoires aboutissant au point (n,i) après opérations sera donné par le coefficient du terme en u^i dans le développement de $2^n g^n(u)$.

Tout ceci n'est évidemment valable que lorsque les trajectoires ne sont soumises à aucune contrainte.
Essayons maintenant de traduire le fait que toutes les trajectoires sont comprises entre les droites (c) et (d).

Pour cela nous exprimerons que toute trajectoire touchant la droite (c) , par exemple, est absorbée par cette barrière. Voici comment sera figurée cette absorption.

Construisons toutes les trajectoires ayant pour origine le point (0,0) ; puis leurs symétriques par rapport à la barrière absorbante. Ainsi une trajectoire aboutissant en un point P de cette barrière, y rencontrera sa symétrique. Pour exprimer que cette trajectoire est absorbée, il faut affecter au point P une probabilité nulle. Si la trajectoire considérée arrive au point P avec une probabilité (p), il faut donc que la trajectoire symétrique y arrive avec une probabilité égale à (-p). Or ceci peut être réalisé simplement. En effet, si p_0 est la probabilité affectée à l'origine (0,0), il suffira d'affecter à son symétrique par rapport à la barrière absorbante la probabilité $(-p_0)$.

L'absorption sur les différentes barrières se traduit donc par une transformation de la répartition initiale de probabilité $g^0(u) = 1 \times u^0$, imitée du phénomène des images dans deux miroirs plans et parallèles. Cette méthode est due à Monsieur Tortrat. Calculons donc la nouvelle répartition initiale de probabilité $G^0(u)$.

L'image de $g^0(u)$ dans (c) sera $-u^{-2c}$

L'image de $g^0(u)$ dans (d) sera $-u^{2d}$

L'image de u^{-2c} dans (d) sera $+u^{2d+2c}$

.....

Nous prendrons donc :

$$G^0(u) = \ldots\ldots - u^{2c+4d} + u^{2c+2d} - u^{2d} + 1 - u^{-2c} + u^{-2c-2d} - u^{-2c-4d} \ldots\ldots$$

Ainsi, d'après tout ce que nous venons de rappeler, nous aurons le nombre de trajectoires aboutissant au point $(2n, 0)$ en prenant le coefficient du terme en u^0 dans le développement du produit :

$$2^{2n} \; Y^{2n} \; G^0(u)$$

$$2^{2n} \; Y^{2n} = \frac{2^{2n}}{2^{2n}} \left(u + \frac{1}{u}\right)^{2n}$$

$$= u^{2n} + C_{2n}^{1} u^{2n-2} + \ldots + C_{n}^{i} u^{2n-2i} + \ldots + C_{2n}^{n} + \ldots \; C_{2n}^{n+j} u^{-2j} + \ldots + u^{-2n}$$

Cherchons alors le coefficient du terme en u^0.

a) Termes positifs :

$$-- \; C_{2n}^{i} \; u^{2n-2i} \times u^{-2k(c+d)}$$

qui entraîne : $(n-i)-k(c+d)=0$

$$i = n-k(c+d)$$

i variant de 0 à n, k ne pourra prendre que les valeurs comprises entre 0 et $\left[\frac{n}{c+d}\right]$

Le nombre de ces termes est donné par

$$\sum_{k=0}^{\left[\frac{n}{c+d}\right]} C_{2n}^{n-k(c+d)}$$

-- $C_{2n}^{n+j}\ u^{-2j} \times u^{2k(c+d)}$

qui entraîne $j=d(c+d)$; k sera alors compris entre 1 et $\left[\frac{n}{c+d}\right]$

Le nombre de ces termes est donné par $\sum_{k=1}^{\left[\frac{n}{c+d}\right]} C_{2n}^{n+k(c+d)}$

Ceci peut aussi s'écrire $\sum_{k=-\left[\frac{n}{c+d}\right]}^{-1} C_{2n}^{n+k(c+d)}$

Et en groupant les deux résultants , on voit que le nombre de termes positifs est donné par

$$\sum_{k=-\left[\frac{n}{c+d}\right]}^{\left[\frac{n}{c+d}\right]} C_{2n}^{n-k(c+d)}$$

b) Termes négatifs :

-- $C_{2n}^{n+j}\ u^{-2j} \times u^{2d+2k(c+d)}$

qui entraîne $d + k\,(c+d) = j$

donc k varie entre $\left[\frac{1-d}{c+d}\right] = 0$ et $\left[\frac{n-d}{c+d}\right]$

Le nombre de ces termes est $\sum_{k=0}^{\left[\frac{n-d}{c+d}\right]} C_{2n}^{n+d+k(c+d)}$

-- $C_{2n}^{i}\ u^{2n-2i} \times u^{-2c-2k(c+d)}$

qui entraîne : $n - i = c + k(c+d)$

k varie alors entre $-\left[\frac{c}{c+d}\right] = 0$ et $\left[\frac{n-c}{c+d}\right]$

Le nombre de ces termes est $\sum_{k=0}^{\left[\frac{n-c}{c+d}\right]} C_{2n}^{n-c-k(c+d)}$

Modifions alors les 2 résultats en posant k+1=l

- le deuxième deviendra : $\sum_{l=1}^{\left[\frac{n+d}{c+d}\right]} C_{2n}^{n+d-l(c+d)}$

en effet : $n-c-k(c+d) = n+d-l(c+d)$

si $k_m = \left[\frac{n-c}{c+d}\right]$ $\quad k_m \leqslant \frac{n-c}{c+d} < k_m+1$

$$k_m+1 \leqslant \frac{n-c}{c+d} +1 < k_m+2$$

$$l_m \leqslant \frac{n+d}{c+d} < l_m+1$$

d'où $l_m = \left[\frac{n+d}{c+d}\right]$

- le premier deviendra :

$$\sum_{k=0}^{\left[\frac{n-d}{c+d}\right]} C_{2n}^{n+d+k(c+d)} = \sum_{l=1}^{\left[\frac{n+c}{c+d}\right]} C_{2n}^{n+c-l(c+d)}$$

On retrouve ainsi les résultats de la 1ère démonstration.

V - Loi limite -

Ayant établi la loi régissant le test de Gnédenko, lorsque le nombre d'éléments composant les échantillons a une valeur finie, on peut se demander ce que devient cette loi lorsque ce nombre devient très grand.

Nous avons vu que

$$\text{Prob}\left[n\,D_n < c\right] = \sum_{k=-\left[\frac{n}{c}\right]}^{k=\left[\frac{n}{c}\right]} (-1)^k \frac{C_{2n}^{n-kc}}{C_{2n}^{n}}$$

En supposant n suffisamment grand, nous pouvons appliquer la formule de Sterling :

$$\text{Prob}\left[n\,D_n < c\right] = \sum_{-\infty}^{+\infty} (-1)^k \frac{(2n)!}{(n-kc)!(n+kc)!}\frac{(n!)^2}{(2n)!}$$

$$\text{Prob}\left[n\,D_n\ c\right] = \sum_{-\infty}^{+\infty} (-1)^k \frac{\left(\frac{n}{e}\right)^{2n}\ 2\ n}{\frac{n-k_c}{e}^{n-k_c}\ \frac{n+k_c}{e}^{n+k_c}\ 2\ \sqrt{n^2-k^2c^2}}$$

$$= (-1)^k \frac{n^{2n+1}}{n^{2n+1}\left[1-\frac{k^2c^2}{n^2}\right](n-kc)^{-kc}(n+kc)^{kc}\sqrt{1-\frac{k^2c^2}{n^2}}}$$

Posons alors $c = x\sqrt{2n}$

$\text{Prob}\left[n\,D_n < c\right]$ devient $\text{Prob}\left[\sqrt{\frac{n}{2}}\,D_n < x\right]$

Si nous appelons alors F(x) la loi limite, nous aurons :

$$F(x) = \lim_{n\to\infty} \sum_{-\infty}^{+\infty} \frac{(-1)^k}{\left[1-\frac{2k^2x^2}{n}\right]^n \left[1-\frac{kx\sqrt{2}}{\sqrt{n}}\right]^{-kx\sqrt{2n}} \left[1+\frac{kx\sqrt{2}}{\sqrt{n}}\right]^{-kx\sqrt{2n}} \sqrt{1-\frac{2k^2x^2}{n}}}$$

$$F(x) = \sum_{-\infty}^{+\infty} \frac{(-1)^k}{e^{-2k^2x^2} \times e^{4k^2x^2}}$$

$$F(x) = \sum_{-\infty}^{+\infty} (-1)^k \; e^{-2k^2x^2}$$

d'où

$$\underset{n \to \infty}{\text{Prob}} \left[\sqrt{\frac{n}{2}} \, D_n < x \right] = \sum_{-\infty}^{+\infty} (-1)^k \; e^{-2k^2x^2}$$

VI - Résultats.

1. Convergence vers la loi limite .

La loi limite est définie par :

$$\text{Prob}\left[\sqrt{\frac{n}{2}}\ D_n < x\right] = \sum_{-\infty}^{+\infty} (-1)^k\ e^{-2k^2x^2}$$

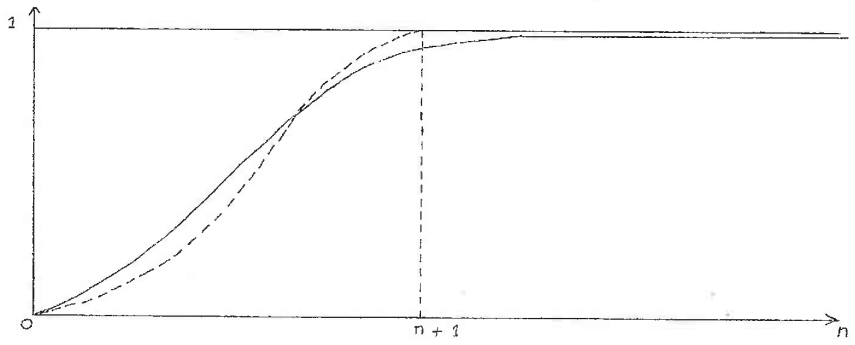
Les courbes de Gnédenko sont définies par :

$$\text{Prob}\left[n\,Dn < c\right]$$

Dans ces conditions si on appelle u l'abcisse d'une courbe relative à u fini, et x l'abcisse de la loi limite, on passe d'une courbe à l'autre par le changement de variable :

$$u = x\ \sqrt{2n}$$

Il est peut être intéressant de signaler le mode de convergence. En effet, pour u fini, la courbe atteint strictement la valeur 1. Or la courbe limite est asymptote à la droite d'ordonnée 1 et vaut 0 au point 0 ; ce qui nous donne la disposition suivante :

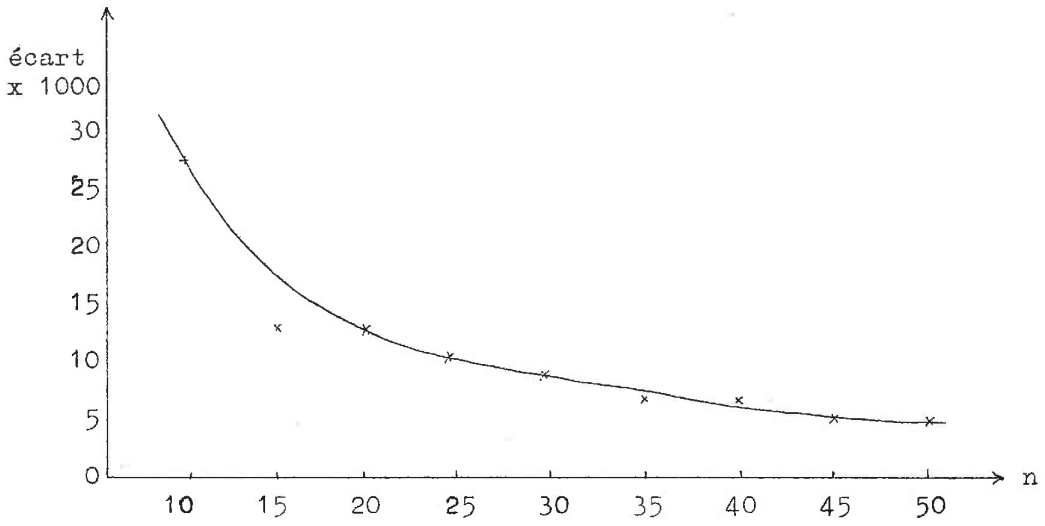


La courbe relative à u fini est en pointillé, la courbe limite en trait plein.

Pour avoir une idée précise de la convergence, on a calculé les écarts maximum, pour des valeurs successives de u, et on trouve :

n	10	15	20	25	30	35	40	45	50
écart x 1000	27	13	13	11	9	7	7	6	5

Ce que l'on peut traduire par la courbe :



On constate que à 1/100 près, la courbe limite est atteinte pour n = 30, encore ces écarts sont-ils atteints pour des valeurs de n très faibles, donc sur la partie la moins intéressant. En fait, si on ne considère les courbes qu'au-dessus d'nne probabilité de 50 %, on peut diviser par 2 les écarts signalés.

2. Comparaison avec le test du χ^2

Tous les résultats qui suivent sont relatifs à des calculs menés sur des variables normales.

Nous avons utilisé le χ^2 comme test d'homogénéité sous la forme suivante :

1ère série	2ème série	
		CLASSES
μ_i	ν_i	
m	n	TOTAL

$$\chi^2 = \frac{(m+n)^2}{m\,n}\left(\sum_i \frac{i^2}{\mu_i+\nu_i} - \frac{m^2}{m+n}\right)$$

Cette façon de calculer χ^2 est en effet beaucoup plus pratique pour le calcul automatique.

Nous avons utilisé 3 séries de 50 variables normales donc $m = n = 50$

Voici les résultats :

Séries 1 - 2 (1)
$$\begin{cases} \chi^2 = 1{,}849 \\ \text{Prob}\left[\chi^2 \geqslant 1{,}849\right] = 76{,}1\ \% \\ nDn = 7 \\ \text{Prob}\left[nDn \geqslant 7\right] = 71{,}7\ \% \end{cases}$$

Séries 2 - 3
$$\begin{cases} \chi^2 = 1{,}903 \\ \text{Prob}\left[\chi^2 \geqslant 1{,}903\right] = 75{,}2\ \% \\ nDn = 8 \\ \text{Prob}\left[nDn \geqslant 8\right] = 55\ \% \end{cases}$$

Séries 1 - 3
$$\begin{cases} \chi^2 = 3{,}016 \\ \text{Prob}\left[\chi^2 \geqslant 3{,}016\right] = 55{,}9\ \% \\ nDn = 7 \\ \text{Prob}\left[nDn \geqslant 7\right] = 71{,}7\ \% \end{cases}$$

Nous avons alors pris les résultats (1) des comparaisons des séries 1 et 2 et nous avons décalé la série 2 en ajoutant systématiquement un nombre δ aux valeurs proposées.

En choisissant un décalage qui ne bouleverse pas le contenu des classes déterminées dans le calcul (1) on a pu obtenir :

$$\begin{cases} \chi^2 = 1{,}849 \\ \text{Prob}\left[\chi^2 \geqslant 1{,}849\right] = 76{,}1\ \% \\ nDn = 9 \\ \text{Prob}\left[nD \geqslant 9\right] = 39{,}6\ \% \end{cases}$$

On s'est alors demandé lequel des deux tests décèlerait le premier la manoeuvre de décalage, moyennant un seuil de probabilité donné.

Séries 1 - 2
décalage $\delta = 0,45$

$\chi^2 = 9,384$
$\text{Prob}\left[\chi^2 \geqslant 9,384\right] = 5,36\ \%$
$nDn = 14$
$\text{Prob}\left[nDn \geqslant 14\right] = 4\ \%$

Donc au seuil de 95 % le test nDn donne le signal d'erreur plus rapidement que le test du χ^2.

Séries 1 - 2
décalage $\delta = 0,58$

$\chi^2 = 9,54$
$\text{Prob}\left[\chi^2 \geqslant 9,54\right] = 4,6\ \%$
$nDn = 17$
$\text{Prob}\left[nDn \geqslant 17\right] = 0,6\ \%$

La remarque précédente reste donc vraie au seuil de 99 %.

Il est, en outre, à remarquer que tous ces calculs sont effectués sur des variables normales. On est donc dans le cas le plus favorable pour le test du χ^2. Il est probable que les écarts seraient encore plus grands si nous avions utilisé une loi très dissymétrique.

Posons alors la même question pour des échantillons dont le nombre d'éléments vaudra successivement $n = 20$ $n = 30$ $n = 40$

Pour $n = 40$, sans décalage :

$\chi^2 = 0,920$
$\text{Prob}\left[\chi^2 \geqslant 0,920\right] = 63,85\ \%$
$nDn = 6$
$\text{Prob}\left[nDn \geqslant 6\right] = 76,6\ \%$

Avec décalage :

$n = 40$

Au seuil de 95 %

$\chi^2 = 5,992$
$\text{Prob}\left[\chi^2 \geqslant 5,992\right] = 19\ \%$
$nDn = 13$
$\text{Prob}\left[nDn \geqslant 13\right] = 2,9\ \%$

Au seuil de 99 %

$\chi^2 = 15,56$
$\text{Prob}\left[\chi^2 \geqslant 15,56\right] = 0,6\ \%$
$nDn = 15$
$\text{Prob}\left[nDn \geqslant 15\right] = 0,6\ \%$

Même procédé avec n = 30

$n = 30$

Au seuil de 95 % :

$\chi^2 = 6,3$

$\text{Prob}\left[\chi^2 \geqslant 6,3\right] = 10,5\ \%$

$nD_n = 11$

$\text{Prob}\left[nD_n \geqslant 11\right] = 3,46\ \%$

Au seuil de 99 % :

$\chi^2 = 7,74$

$\text{Prob}\left[\chi^2 \geqslant 7,74\right] = 2,13\ \%$

$nD_n = 13$

$\text{Prob}\left[nD_n \geqslant 13\right] = 0,66\ \%$

$n = 20$

Au seuil de 95 % :

$\chi^2 = 2,74$

$\text{Prob}\left[\chi^2 \geqslant 2,74\right] = 27,66\ \%$

$nD_n = 9$

$\text{Prob}\left[nD_n \geqslant 9\right] = 3,36\ \%$

Au seuil de 99 % :

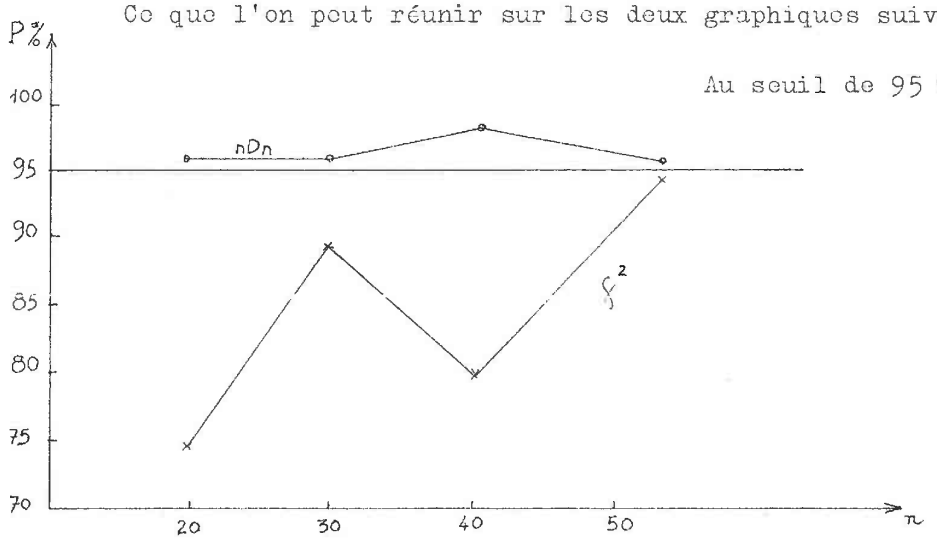
$\chi^2 = 11,04$

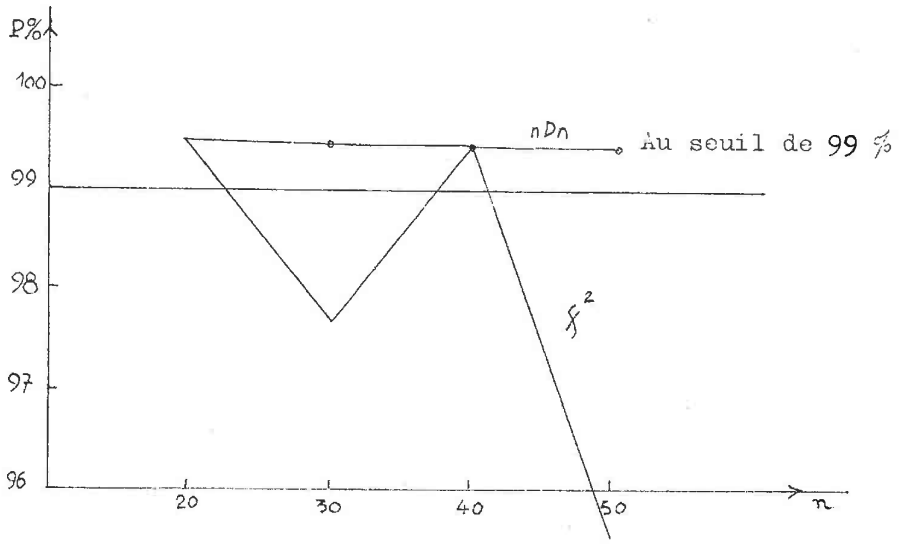
$\text{Prob}\left[\chi^2 \geqslant 11,04\right] = 0,5\ \%$

$nD_n = 11$

$\text{Prob}\left[nD_n \geqslant 11\right] = 0,4\ \%$

Ce que l'on peut réunir sur les deux graphiques suivants :





Test de comparaison de moyennes.

Reprenons n = 50

a) en considérant $\bar{x} - \bar{y}$ comme une variable normale de moyenne nulle et d'écart type

$$\sigma = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

On trouve les intervalles de confiance suivants :

au seuil de 95 % $|\bar{x} - \bar{y}| \leqslant 0{,}37$

au seuil de 99 % $|\bar{x} - \bar{y}| \leqslant 0{,}491$

b) un test de Student sur la variable $|\bar{x} - \bar{y}|$ donne des résultats équivalents :

au seuil de 95 % $|\bar{x} - \bar{y}| \leqslant 0{,}3795$

au seuil de 99 % $|\bar{x} - \bar{y}| \leqslant 0{,}497$

Rappelons que pour le test de Gnédenko les écarts admissibles sont :

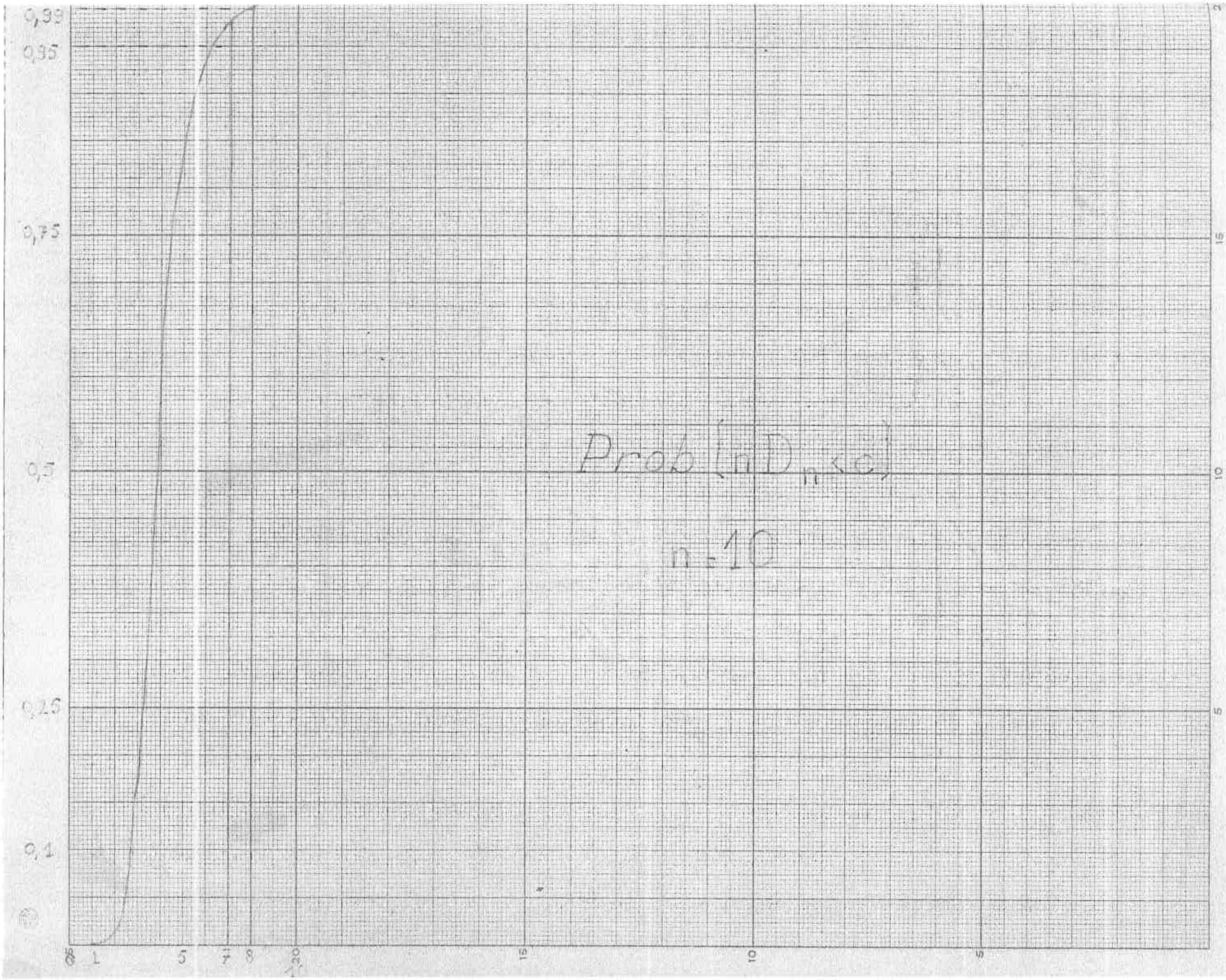
au seuil de 95 % : 0,45

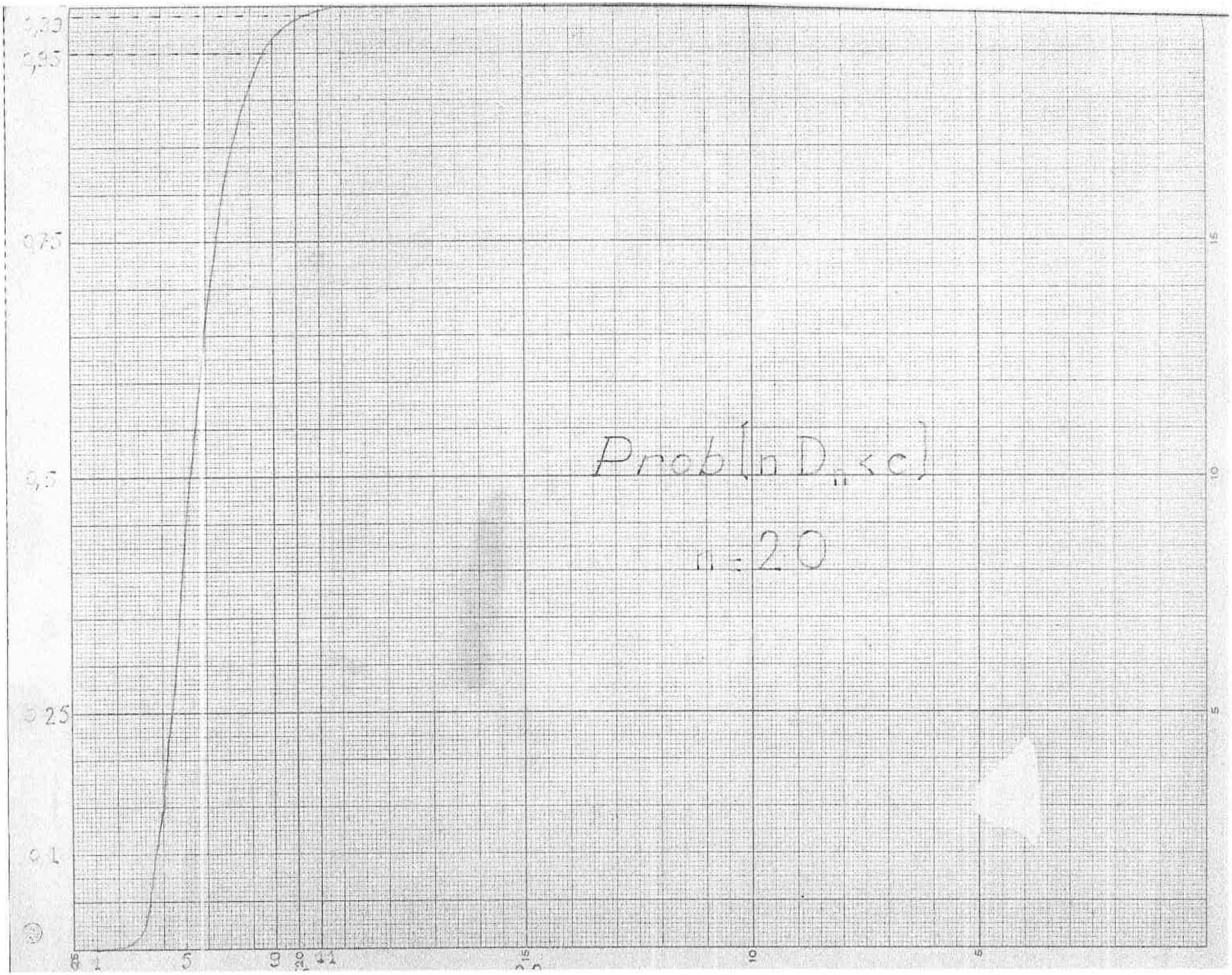
au seuil de 99 % : 0,58

VII - Conclusions

De tous les essais réalisés, on peut conclure que le test de Gnédenko n'est sûrement pas inférieur au test du χ^2 lorsqu'il s'agit de vérifier l'homogénéité de deux échantillons. Or il faut souligner que sa mise en oeuvre pratique est beaucoup plus simple que celle du χ^2. On peut en particulier concevoir un programme simple et rapide pour calculateur électronique.

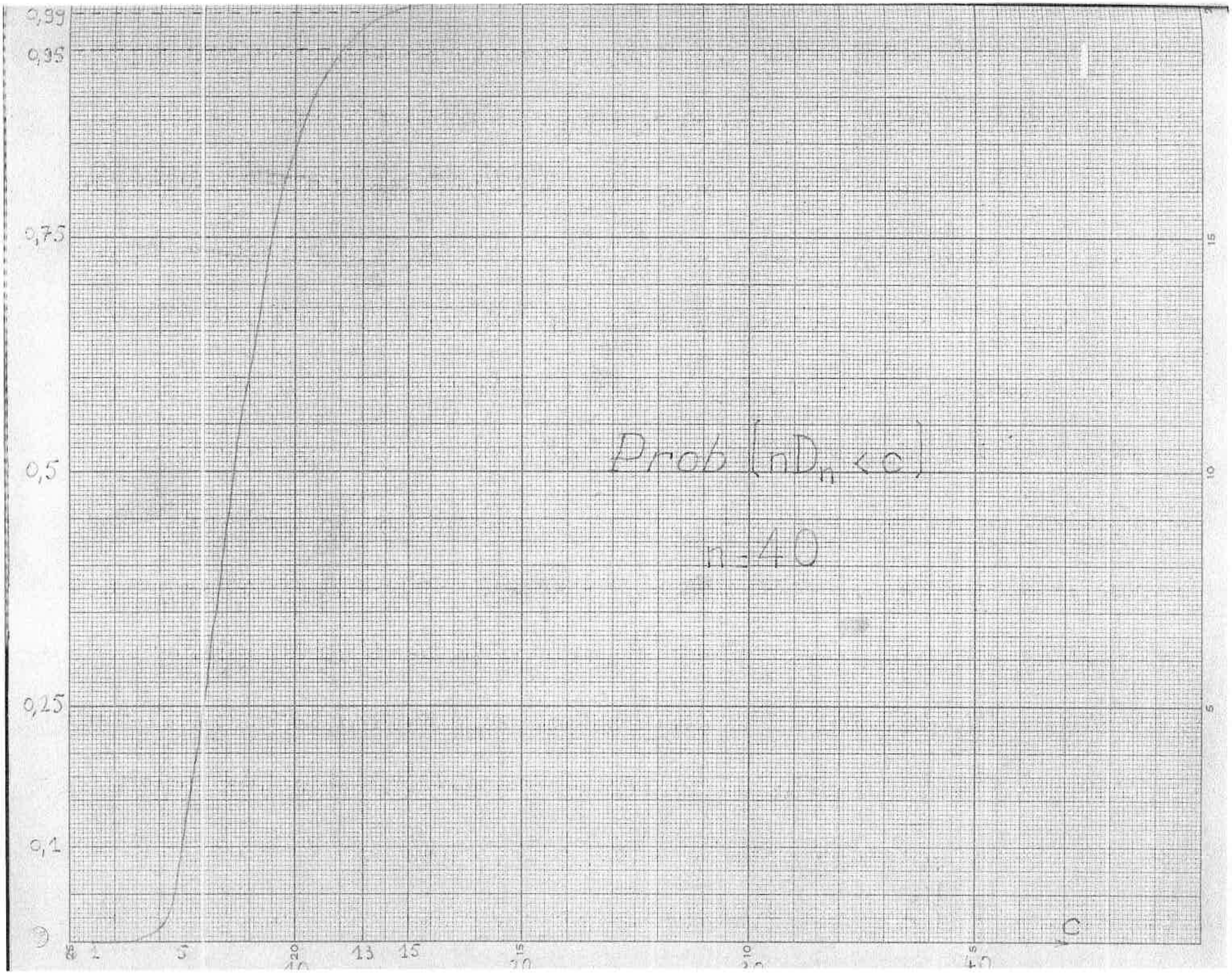
Cependant, et c'est là son gros défaut, ce test de Gnédenko n'est pas applicable aux lois discontinues pour lesquelles on est bien obligé de revenir au χ^2.


Prob [nD n<c]
n = 10
0,99
0,95
0,75
0,5
0,25
0,1
0
1
5
7
8

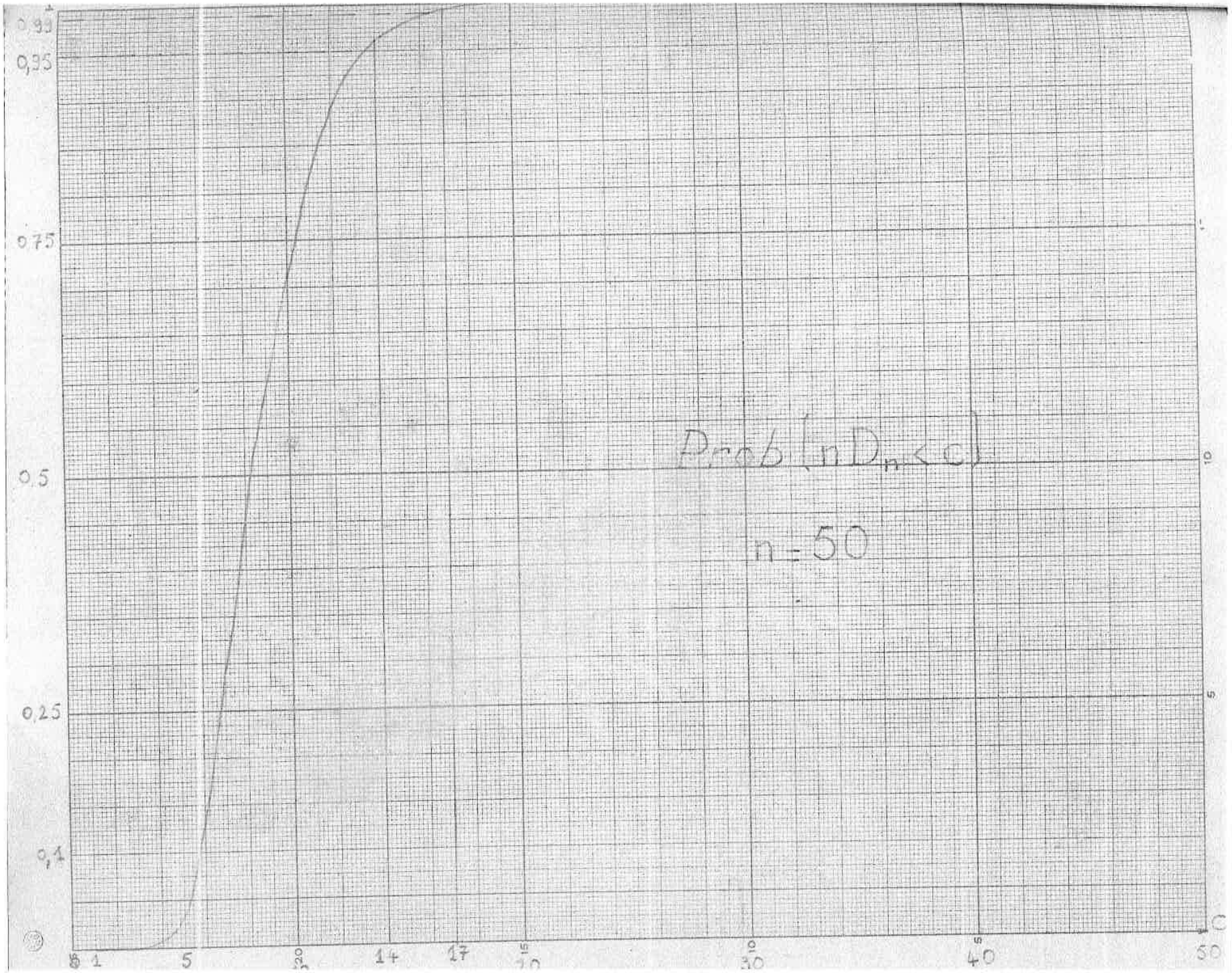


Prob[n Dn < c]
n = 20
0,99
0,95
0,75
0,5
0,25
0,1
1
5
9
10
11

15
10
5
0,95
0,85
0,75
0,6
0,25
0,1
5
10
15
20
25
C



Prob [nD_n < c]
n = 40
0,99
0,95
0,75
0,5
0,25
0,1
1
5
10
13
15
c


0,99
0,95
0,75
0,5
0,25
0,1
Prob [nD_n < c]
n = 50
1
5
20
14
17
15
10
30
5
40
50
1
5
10

G E N E R A T I O N D E

N O M B R E S P S E U D O - A L E A T O I R E S

I - Suites uniformément denses

II - Procédé de génération

III - Résultats pratiques

IV - Contrôles effectués sur les suites de nombres pseudo-aléatoires

V - Conclusion

GENERATION DE NOMBRES PSEUDO - ALEATOIRES

I - Suites uniformément denses -

1. Généralités.

Soit $f(x)$ une fonction bornée intégrale, définie sur le segment $(0,1)$. L'intégrale

$$I = \int_0^1 f(x)\, dx$$

qui semble n'avoir aucun rapport avec le calcul des probabilités peut cependant être considéré comme l'espérance mathématique d'une variable aléatoire x uniformément distribuée entre 0 et 1.

D'après la loi des grands nombres, si $x_1, \ldots, x_n$ sont les résultats de N épreuves successives sur X, la moyenne arithmétique

$$S_n = \frac{1}{N}\left[f(x_1) + f(x_2) + \ldots + f(xn)\right]$$

tend en probabilité vers l'intégrale I lorsque N tend vers l'infini. On peut considérer $x_1, \ldots, x_n$ comme les résultats d'épreuves successives sur N variables aléatoires indépendantes uniformément distribuées sur $(0,1)$.

Désignons par ∇ l'écart type de $f(x)$ défini par

$$\nabla^2 = \int_0^1 \left[f(x) - I\right]^2 dx = \int_0^1 \left[f(x)\right]^2 dx - I^2$$

On sait que S_n est une variable aléatoire qui a pour moyenne stochastique I, pour écart-type $\frac{\sigma}{\sqrt{N}}$ et que $\frac{\sqrt{N}}{\sigma}(S_n - I)$ est une variable asymptotiquement distribuée suivant une loi normale réduite.

Si N est suffisamment grand :

$$\text{Prob}\left[\left|S_n - I\right| > h\right] = \frac{2}{\sqrt{2\pi}} \int_{\frac{n\sqrt{N}}{\sigma}}^{\infty} e^{-\frac{t^2}{2}}\, dt$$

2. Nombres aléatoires.

Essayons alors de construire une suite de "nombres aléatoires". Il s'agit de nombres aléatoires au sens restreint, c'est-à-dire de nombres dont la suite ordonnée puisse être considérée comme le résultat d'un ensemble d'épreuves sur une variable aléatoire uniformément répartie entre 0 et 1.

Notons que si X a pour fonction de répartition F(x), la variable Y = F(x) est uniformément répartie distribuée entre 0 et 1. En effet, puisque F(x) est non décroissante :

$$\text{Prob}\left[X < x\right] = \text{Prob}\left[Y < F(x)\right] = F(x)$$

et si on pose $F(x) = y$, on a bien, :

$$\text{Prob}\left[X < y\right] = y$$

On peut donc toujours ramener une variable aléatoire quelconque à une variable uniformément distribuée entre 0 et 1.

Pour construire une succession de valeurs numériques qui soient le résultat d'épreuves indépendantes sur une variable aléatoire uniformément répartie entre 0 et 1, on peut recourir à un jeu statistique convenablement choisi. Il n'est d'ailleurs pas facile de trouver un jeu statistique sui soit à la fois assez précis, assez rapide et assez garanti pour donner avec un nombre de décimales adapté aux machines à calculer modernes plusieurs milliers de nombres successifs.

Cependant d'autres méthodes ont été proposées pour calculer des nombres aléatoires. Ces méthodes reposent sur des calculs de nature arithmétiques dans lesquels le hasard ne joue aucun rôle. Ces calculs sont organisés de telle sorte que la suite x_1, x_2 .... qu'ils fournissent soit compliquée et donne l'impression du hasard. Les nombres obtenus sont ensuite testés par des procédés statistiques.

Malheureusement les méthodes arithmétiques sont souvent difficiles à étudier du point de vue mathématique. Comme les tests statistiques ne sont pas des preuves absolues de la validité des hypothèses, les méthodes arithmétiques généralement employées restent très incertaines.

Nous utiliserons cependant un procédé pour lequel l'étude peut être complète. On notera que dès le moment où l'on utilise des théorèmes précis, les méthodes utilisées n'ont plus rien de statistique sauf l'apparence.

3. Théorèmes de Henri Weyl.

- Définition :

Soit x_1, x_2,... x_n....; une suite de nombres compris entre 0 et 1. Considérons un intervalle $(\alpha\ \beta)$ tel que

$0 < \alpha < \beta < 1$. Supposons que parmi les N premiers termes de la suite il y en ait N' dans l'intervalle $(\alpha\ \beta)$. Si le rapport $\frac{N'}{N}$ tend, lorsque $N \to \infty$, vers une limite égale à $\beta - \alpha$, et cela quel que soit l'intervalle $(\alpha\ \beta)$ choisi, on dit que la suite x_n est uniformément dense, ou équirépartie sur (0,1).

- Théorème 1.

Pour que la suite x_n soit équirépartie (0,1) il faut et il suffit que pour toute fonction f(x) intégrale au sens de Riemann sur (0,1) on ait :

$$\int_0^1 f(x)dx = \lim_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N-1} f(x_n)$$

- Théorème 2.

Pour que la suite x_n soit équirépartie sur (0,1), il faut et il suffit que pour tout entier $l \neq 0$ on ait :

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N-1} e^{2i\pi l x_n} = 0$$

On définit d'une manière analogue une suite de points $P_1, P_2, \ldots P_n, \ldots$ équirépartie dans l'hypercube fondamental C de l'espace à n dimensions, c'est-à-dire l'hypercube dont les points P ont des coordonnées toutes comprises entre 0 et 1. Soit $\mathcal{D}$ un domaine quarrable à l'intérieur de C, de volume V. Supposons que parmi les N points P_n, il y en ait

N' à l'intérieur de $\mathcal{D}$. Si le rapport $\frac{N'}{N}$ tend, lorsque N tend vers l'infini, vers une limite égale à V, et cela quel que soit le domaine $\mathcal{D}$, on dit que la suite P_n est équirépartie dans C. On démontre alors :

- Théorème 3 :

Pour que la suite P_n soit équirépartie dans l'hypercube C, il faut et il suffit que, pour toute fonction f(P) intégrable au sens de Riemann dans l'hypercube C, on ait :

$$\int_C f(P)dv = \lim_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N-1} f(P_n)$$

Il est à remarquer que le théorème 1 est l'équivalent du théorème statistique de (1) . Cependant dans un cas la convergence du second membre vers le premier se fait au sens ordinaire, dans l'autre cas, elle se fait en probabilité, ce qui est moins précis.
La raison physique de ces théorèmes est la même : la suite x_n tend à remplir d'une "façon homogène" le segment (0,1), la suite P_n tend à remplir "régulièrement" l'hypercube C. La moyenne des valeurs de f en des points régulièrement répartis est une approximation de la valeur moyenne théorique, c'est-à-dire de l'intégrale. Le sens du mot "remplir" ne doit naturellement pas être interprété de façon trop concrète. La suite x_n est dense dans le segment (0,1) mais l'ensemble des x_n est dénombrable donc de mesure nulle. Cependant si on prend des intervalles d'amplitude α et de centre les x_n de la suite, la réunion de ces intervalles recouvrira tout le segment (0,1).

II - Procédé de génération -

a) Description.

Choisissons arbitrairement un nombre x_0, soumis seulement à la condition

$$0 \leqslant x_0 < 1$$

puis un entier $N > 1$. Nous définirons alors une suite de nombres $x_0\ x_1\ x_2 \ldots\ x_n \ldots$ par la relation :

$$x_n = N\, x_{n-1} + \Theta - \left[N\, x_{n-1} + \Theta \right]$$

dans laquelle les crochets représentent la partie entière de le quantité considérée.

Nous pourrons considérer que $0 \leqslant \Theta < 1$.

Nous allons alors montrer que la suite déterminée est équirépartie sur (0,1), sous certaines conditions.

b) Justification théorique.

Henri Weyl a montré que la suite obtenue en faisant N=1 est équirépartie quel que soit x_0 et pour presque tout Θ.

Si N est un entier plus grand que 1 nous allons montrer que la suite obtenue est équirépartie pour tout Θ et presque tout x_0.

Pour cela nous démontrerons une extension du théorique de Henri Weyl :

Théorème : Soit f(x) une fonction intégrable au sens de Lebesgue. Alors pour $x_0\ x_1\ x_2 \ldots x_n \ldots$ de la suite

déterminée au a) :

$$\frac{1}{N} \sum_{n=0}^{N-1} f(x_n) \longrightarrow \int_0^1 f(x)dx \qquad \text{lorsque } N \longrightarrow \infty$$

Ceci pour presque toute valeur de x_0.

Nous nous servirons pour la démonstration du théorème ergodique suivant, dû à F. Riesz :

- Etant donné un ensemble mesurable Ω , de mesure finie ou infinie, cette mesure étant définie par une intégrale au sens de Lebesgue, ou plus généralement au moyen d'une répartition de masses positives. Désignons alors par T une transformation ponctuelle, univoque, de sur lui-même ; et supposons que T conserve la mesure, c'est-à-dire que si E étant un ensemble mesurable, TE son transformé, et E' l'ensemble des points P dont l'image est dans TE, alors les ensembles E' et TE ont même mesure. Dans ces conditions, si $f_1(P)$ est une fonction intégrale, et si $f_k(P) = f_1(T^{k-1}P)$, la moyenne arithmétique des fonctions f_1, $f_2 \ldots f_n$ converge presque partout, lorsque $n \to \infty$, vers une fonction intégrable F(P) invariante presque partout par rapport à T.

En outre dans le cas où Ω est de mesure finie

$$\int_\Omega F(P) = \int_\Omega f_1(P) \ .$$

Pour nous Ω est l'intervalle (0,1). La transformation T est définie par :

$$Tx = Nx + \Theta - \left[N x + \Theta\right]$$

Chaque point x de Ω est l'image par T de N points - En effet

$$T \frac{k+x-\Theta}{N} = x \begin{cases} k=0,1,\ldots N-1 & \text{si } x \geqslant \Theta \\ k=1,2,\ldots N & \text{si } x < \Theta \end{cases}$$

Pour montrer que T conserve la mesure, il suffit de prouver que :

$$\int_0^1 f(Tx)dx = \int_0^1 f(x)dx$$

Grâce à la périodicité de la fonction $x - x$ nous avons $\int_0^1 f(Tx)dx = \int_0^1 f(Nx+\theta - [Nx+\theta])\,dx$

$$\int_0^1 f(Tx)dx = \frac{1}{N}\int_0^N f(y+\theta-[y+\theta])\,dy$$

$$= \int_0^1 f(y+\theta-[y+\theta])dy$$

$$= \int_\theta^{1+\theta} f(y-[y])\,dy$$

$$= \int_0^1 f(y)dy$$

Nous sommes donc bien dans les conditions du théorème de Riesz. Démontrons le théorème proposé.

- D'après le théorème ergodique :

$$-\ \frac{1}{N}\sum_{k=0}^{N-1} f_k(x_0) \longrightarrow F(x_0) \qquad \text{presque partout lorsque } N\to\infty$$

$$-\int_0^1 F(x)\,dx = \int_0^1 f(x)dx$$

et nous voulons démontrer le résultat suivant :

$$\frac{1}{N} \sum_{k=0}^{N-1} f(x_k) \longrightarrow \int_0^1 f(x)\,dx \qquad \text{presque partout lorsque } N \longrightarrow \infty$$

c'est-à-dire

$$\frac{1}{N} \sum_{k=0}^{N-1} f_k(x) \longrightarrow \int_0^1 f(x)\,dx \qquad \text{presque partout lorsque } N \longrightarrow \infty$$

Il suffit donc de démontrer que

$$F(x) = \int_0^1 F(x)\,dx$$

Pour cela supposons que $F(x)$ admette un développement en série de Fourier

$$F(x) = \sum_{k=-\infty}^{+\infty} C_k \exp(2i\pi kt)$$

Etudions alors le coefficient C_{Nk}

$$C_{Nk} = \int_0^1 F(t) \exp(-2i\pi Nkt)dt$$

Or, d'après le théorème de Riesz :

$$F(t) = F(Tt)$$

donc :

$$C_{Nk} = \int_0^1 F\left[Nt+\theta-\left[Nt+\theta\right]\right] \exp(-2i\pi Nkt)\,dt$$

posons $Nt + \theta = y$

$$C_{Nk} = \frac{1}{N} \int_{\theta}^{N+\theta} F(y-[y]) \exp(-2i\pi k(y-\theta))dy$$

Or, $F(y-[y])$ ainsi que $\exp(-2i\pi k(y-\theta))$ sont périodiques et de période 1. D'où :

$$C_{Nk} = \int_{\theta}^{1+\theta} F(y-[y]) \exp(-2i\pi k(y-\theta))\, dy$$

$$C_{Nk} = \int_{\theta}^{1} F(y) \exp(-2i\pi k(y-\theta))\, dy + \int_{1}^{1+\theta} F(y-1) \exp(-2i\pi k(y-\theta))\, dy$$

dans la deuxième intégrale, posons $y-1=z$

$$C_{Nk} = \int_{0}^{1} F(y)\exp(-2i\pi k(y-\theta))\, dy + \int_{0}^{\theta} F(z) \exp(-2i\pi k(z-\theta))\, dz$$

$$C_{Nk} = \int_{0}^{1} F(y) \exp(-2i\pi k(y-\theta))\, dy$$

$$C_{Nk} = C_k \exp(2i\pi k\theta)$$

Faisons r itérations et prenons les modules

$$|C_k| = |C_p| \quad \text{avec} \quad p = N^r k$$

Or, si $p \to \infty$ $|C_p| \to 0$. Donc si k reste fixe et si r augmente $|C_p| \to 0$. Ceci entraîne $|C_k| = 0$ si $k \neq 0$.

On en déduit : $\varphi(x) = C_0$

d'où $$\varphi(x) = C_0 = \int_0^1 \varphi(x)\, dx = \int_0^1 f(x)\, dx$$

Or $$\varphi(x) = \lim_{N \to \infty} \frac{1}{N} \sum_{k=0}^{N-1} f_k(x) \quad \text{presque partout}$$

donc $$\int_0^1 f(x) = \lim_{N \to \infty} \frac{1}{N} \sum_{k=0}^{N-1} f(x_k) \quad \text{presque partout}$$

Ceci démontre que la suite x_k ainsi définie est équirépartie sur le segment (0,1).

III - Résultats pratiques -

1. Une remarque s'impose dès l'abord à l'utilisateur. La série de nombres que nous allons obtenir est calculée sur l'ordinateur I B M 650. Les nombres obtenus seront donc

tous des nombres à dix chiffres, donc des nombres rationnels. Or la théorie précédente supposait que le premier nombre de la suite x_0 était un irrationnel ce qui, naturellement, est impossible à réaliser. De plus, puisque tous les nombres obtenus sont à dix chiffres, nous n'obtiendrons au maximum que 10^{10} nombres différents, d'où la certitude d'apparition d'une période, incompatible avec la notion de hasard. Le calcul de cette période s'avère à peu près impossible, du moins dans le procédé utilisé. Cependant sur une soixantaine de séries différentes que nous avons étudiées, jamais nous n'avons pu constater l'apparition de cette période.

2. Il existe en fait un deuxième inconvénient beaucoup plus important. Il existe en effet une période dans l'apparition du dernier chiffre de la série des nombres obtenus. Voici comment on peut mettre ce phénomène en évidence.

Pour simplifier l'exposé, définissons d'abord une relation d'équivalence en disant que $a \sim b$ si les derniers chiffres de a et de b sont égaux. Ainsi :

$5 \sim 15 \sim 25 \sim 35 \ldots$

Les premiers nombres de la série obtenus sont :

$x_0, x_1, x_2, x_3 \ldots\ldots$

On constate facilement que :

$$x_1 = N x_0 + \Theta(1) - \left[N x_0 + \Theta(1)\right]$$

$$x_2 = N^2 x_0 + \Theta(1 + N) - \left[N^2 x_0 + \Theta(1 + N)\right]$$

$$x_3 = N^3 x_0 + (1 + N + N^2) - \left[N^3 x_0 + \Theta(1 + N + N^2)\right]$$

$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$

$$x_p = N^p x_0 + \Theta \sum_{k=0}^{p-1} N^k \left[- N^p x_0 - \Theta \sum_{k=0}^{p-1} N^k\right]$$

Attachons nous alors à l'étude du dernier chiffre des nombres ainsi formés. Les parties entières deviennent inutiles et nous pouvons écrire :

$$x_1 \sim N\, x_0 + \Theta(1)$$
$$x_2 \sim N^2\, x_0 + \Theta(1 + N)$$
$$x_3 \sim N^3\, x_0 + \Theta(1 + N + N^2)$$
$$x_4 \sim N^4\, x_0 + \Theta(1 + N + N^2 + N^3)$$
$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$
$$x_8 \sim N^8\, x_0 + \Theta(1+N+N^2+N^3+N^4+N^5+N^6+N^7)$$
$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$

Nous avons pris au départ :

$N = 7^9$ donc $N \sim 7$. On constate alors que

$N^4 \sim 1$

$N^5 \sim N$

$N^6 \sim N^2$

$N^7 \sim N^3$

$N^8 \sim N^4$ Par suite :

$$x_8 \sim N^4\, x_0 + 2\,\Theta(1 + N + N^2 + N^3)$$

On peut donc penser à rapprocher x_8 de x_4 , qui sont équivalents au facteur 2 près, qui multiplie le terme en Θ . Or, pour $N \sim 7$, $1+N+N^2+N^3 \sim 0$ donc $x_8 \sim x_4$ d'où une période de 4 dans l'apparition du dernier chiffre. On peut essayer alors de changer N de façon que $1+N+N^2+N^3 \not\sim 0$; ceci n'est possible que pour $N \sim 2, 4, 6$ ou 8 ; mais alors les puissances successives obtenues sont toutes paires, et

le facteur 2 qui apparaît dans x_8 ne modifie rien et la période de 4 subsiste.

Il reste donc à prendre $N \sim 1$. On voit alors que

$$x_1 \sim x_0 + \theta$$

$$x_2 \sim x_0 + 2\theta$$

.....

$$x_p \sim x_0 + p\theta$$

- Or les multiples de θ présentent une période de 10, ce qui n'est pas plus avantageux que la période de 4 que nous obtenions précédemment.

- Il va sans dire que cette période sur l'apparition du dernier chiffre se répercute sur les chiffres précédents, et on constate pratiquement que les 4 ou 5 derniers chiffres des nombres obtenus sont inutilisables. Ce grave inconvénient, que nous n'avons trouvé mentionné nulle part, nous a conduit à adopter un procédé de génération légèrement différent.

On forme toujours :

$$x_{n+1} = N\,x_n + \theta - \left[N\,x_n + \theta\right]$$

mais on modifie θ de façon éléatoire au cours du calcul.

Nous avons essayé deux procédés :

- tous les huit nombres, on modifie θ pour la suite, en prenant comme nouvelle valeur celle du 1er nombre de la liste des huit qui viennent d'être calculés.

- ou bien ,après avoir calculé quatre nombres, on prend comme nouveau θ la somme du premier et du quatrième nombre de la liste qui vient d'être formée.

C'est finalement le deuxième procédé que nous avons adopté, bien que les deux soient à peu près équivalents.

IV - Contrôles effectués sur les nombres pseudo-aléatoires -

Nous avons essayé trois types de contrôles sur les nombres obtenus :

- Calculs de coefficients d'auto-corrélation.
- Tests de fréquence;
- Tests graphiques qui se ramènent à un test du X^2.

1. Coefficients d'auto-corrélation.

Nous avons extrait deux échantillons dans une liste de nombres au hasard, et calculé le coefficient de corrélation entre ces deux échantillons.

Ces deux échantillons seront définis par deux paramètres que nous appellerons N et Δ

$N = 1\,000$; $\Delta = 100$ signifie que nous disposons d'une liste de 1 000 nombres au hasard ($N = 1\,000$), et que nous nous servons de deux échantillons dont le premier est composé des nombres 1 à 900, et le second des nombres 100 à 1 000 de la liste totale.

$N = 1\,000$; $\Delta = 200$ signifie que nous disposons d'une liste de 1 000 nombres au hasard, et que nous travaillons sur deux échantillons dont le premier est composé des nombres 1 à 800 et le second des nombres de 200 à 1 000 de la liste totale.

Rappelons encore une propriété du coefficient de corrélation :

si ρ est le coefficient théorique, et r sa valeur calculée, on peut définir deux nouvelles variables :

$$\zeta = \frac{1}{2} \log \frac{1+r}{1-r} \qquad \zeta = \frac{1}{2} \log \frac{1+\rho}{1-\rho}$$

Si ρ est petit, ζ est alors une variable normale avec :

$$E(\zeta) = \zeta + \frac{\rho}{2(n-1)} \qquad D^2(\zeta) = \frac{1}{n-3}$$

Puisque nous cherchons à réaliser l'indépendance des échantillons, nous aurons donc :

$$E(\zeta) = 0 \qquad \text{et} \quad D(\zeta) \# \frac{1}{\sqrt{n}}$$

n étant le nombre d'éléments des échantillons considérés.

On peut ainsi calculer, à un seuil donné, calculer les intervalles de confiance de ζ et revenir à r par des tables spéciales.

Voici les résultats obtenus :

N	Δ	ς	r	Intervalle de confiance de r au seuil de 95 %
1 000	100	0	- 0,0252	0,0666
	150	0	- 0,0592	0,0672
	200	0	0,0088	0,0693
	250	0	- 0,0446	0,0715
	300	0	0,0622	0,0740
	350	0	0,0058	0,0768
	400	0	0,0267	0,0799
	450	0	0,0522	0,0834
	500	0	0,0349	0,0894
1 000	100	0	- 0,0253	0,0666
	150	0	- 0,0360	0,0672
	200	0	0,0631	0,0693
	250	0	0,0394	0,0715
	300	0	0,0275	0,0740
	350	0	- 0,0235	0,0768
	400	0	- 0,0646	0,0799
	450	0	0,0198	0,0834
	500	0	0,0173	0,0894

N	Δ	ϑ	r	Intervalle de confiance de r au seuil de 95 %
1 000	100	0	- 0,0278	0,0666
	150	0	0,0748	0,0672
	200	0	- 0,0094	0,0693
	250	0	0,0059	0,0715
	300	0	- 0,0535	0,0740
	350	0	- 0,0032	0,0768
	400	0	- 0,0294	0,0799
	450	0	- 0,0689	0,0834
	500	0	0,0163	0,0894

Pour ces trois tableaux les calculs ont été effectués sur 1 000 nombres, les trois séries correspondent à trois valeurs initiales x_0 différentes : $x_0 \neq 0$; $x_0 \neq 0,5$; $x_0 \neq 1$. On constate que tous les résultats sont acceptables au seuil de 95 %, sauf un seul, souligné dans le dernier tableau, qui lui n'est acceptable qu'au seuil de 99 %.

Ces calculs ont été menés avec une valeur de ϑ fixe.

Voici deux exemples avec ϑ variable :

N	Δ	ϛ	r	Intervalle de confiance de r au seuil de 95 %
1 000	100	0	0,0459	0,0666
	150	0	0,0544	0,0672
	200	0	0,0553	0,0693
	250	0	- 0,0479	0,0715
	300	0	- 0,0118	0,0740
	350	0	- 0,0081	0,0768
	400	0	0,0750	0,0799
	450	0	0,0225	0,0834
	500	0	- 0,0854	0,0894
1 000	100	0	0,0025	0,0666
	150	0	- 0,0105	0,0672
	200	0	0,0524	0,0693
	250	0	0,0168	0,0715
	300	0	0,0016	0,0740
	350	0	- 0,0022	0,0768
	400	0	- 0,0368	0,0799
	450	0	- 0,0358	0,0834
	500	0	0,0963	0,0894

Sur ces deux exemples on constate que le procédé donne des résultats du même ordre que ceux que nous avions obtenus avec une valeur de $\mathcal{E}$ constante, puisqu'on trouve ici encore une valeur, la dernière du deuxième tableau, qui est inacceptable au seuil de 95 %.

N	Δ	$\mathcal{E}$	r	Intervalle de confiance de r au seuil de 95 %
200	10	0	0,0224	0,145
	15	0	- 0,0539	0,149
	20	0	0,0120	0,152
	25	0	- 0,0202	0,155
	30	0	- 0,0265	0,158
	35	0	- 0,1457	0,161
	40	0	- 0,0668	0,164
	45	0	0,0540	0,167
	50	0	0,0545	0,170
	55	0	0,0857	0,173
	60	0	0,1055	0,176
	65	0	- 0,1541	0,179
	70	0	- 0,0428	0,182
	75	0	- 0,0317	0,185
	80	0	0,0479	0,188
	85	0	0,0425	0,191
	90	0	0,0412	0,194
	95	0	0,0022	0,197
	100	0	0,0283	0,200

N	Δ	ϑ	r	Intervalle de confiance de r au seuil de 95 %
200	10	0	0,0058	0,145
	15	0	- 0,1523	0,149
	20	0	- 0,0163	0,152
	25	0	0,0347	0,155
	30	0	0,0472	0,158
	35	0	- 0,0016	0,161
	40	0	- 0,0708	0,164
	45	0	- 0,1568	0,167
	50	0	- 0,0194	0,170
	55	0	- 0,0723	0,174
	60	0	- 0,0787	0,176
	65	0	- 0,1304	0,179
	70	0	- 0,0161	0,182
	75	0	0,0110	0,185
	80	0	- 0,0137	0,188
	85	0	- 0,0212	0,191
	90	0	0,0830	0,194
	95	0	0,1324	0,197
	100	0	0,0164	0,200

N	Δ	S	r	Intervalle de confiance de r au seuil de 95 %
100	10	0	0,0954	0,211
	15	0	0,0080	0,217
	20	0	0,0060	0,224
	25	0	- 0,0153	0,231
	30	0	0,0057	0,239
	35	0	- 0,1451	0,248
	40	0	- 0,0001	0,258
	45	0	0,2907	0,269
	50	0	0,0233	0,283

N	Δ	S	r	Intervalle de confiance de r au seuil de 95 %
100	10	0	0,0441	0,211
	15	0	- 0,0042	0,217
	20	0	0,0757	0,224
	25	0	- 0,0559	0,231
	30	0	0,0087	0,239
	35	0	0,0357	0,248
	40	0	0,2568	0,258
	45	0	0,0124	0,269
	50	0	- 0,3138	0,283

N	Δ	ρ	r	Intervalle de confiance de r au seuil de 95 %
100	3	0	0,0114	0,203
	7	0	- 0,0402	0,207
	11	0	- 0,0893	0,212
	15	0	- 0,2087	0,217
	19	0	0,2111	0,223
	23	0	0,0216	0,228
	27	0	- 0,1226	0,234
	31	0	0,0904	0,241
	35	0	0,0056	0,248
	39	0	- 0,0965	0,256
	43	0	0,0522	0,265
	47	0	- 0,1501	0,275

N	Δ	S	r	Intervalle de confiance de r au seuil de 95 %
100	3	0	0,0114	0,203
	6	0	- 0,0654	0,206
	9	0	0,0643	0,210
	12	0	0,0968	0,213
	15	0	- 0,2087	0,217
	18	0	- 0,0620	0,221
	21	0	0,0670	0,225
	24	0	- 0,1345	0,230
	27	0	- 0,1226	0,234
	30	0	0,1056	0,239
	33	0	0,0178	0,244
	36	0	- 0,2333	0,250
	39	0	- 0,0965	0,256
	42	0	- 0,0002	0,261
	45	0	- 0,1479	0,270
	48	0	- 0,0982	0,278

A moins une simulation très puussée, on a rarement à utiliser des listes de 1 000 nombres, c'est pourquoi nous

avons donné davantage d'exemples portant sur des listes de 100 ou 200 nombres.

Nous avons beaucoup insisté sur les résultats obtenus car ces calculs d'auto-corrélation semblent être, en définitive, les plus significatifs en ce qui concerne la validité des listes obtenues. On constate d'ailleurs que si les résultats sont très bons au seuil de 95 %, on ne relève aucune contradiction au seuil de 99 %.

2. Tests de fréquence.

Nous avons testé la fréquence d'apparition de chacun des dix chiffres dans chacune des dix colonnes qui composent les nombres pseudo-aléatoires obtenus. Nous avons toujours travaillé sur des listes de 1 000 nombres. La fréquence théorique d'apparition de chacun des dix chiffres est donc de 100.

Voici quelques résultats obtenus :

- Avec δ constant.

fréquence du \ colonnes	1	2	3	4	5	6	7	8	9	10
0	100	106	102	117	91	104	100	100	0	0
1	100	101	105	97	101	96	100	100	0	250
2	96	97	83	103	96	104	100	100	250	0
3	118	100	103	108	96	96	100	100	0	250

colonnes / fréquence du	1	2	3	4	5	6	7	8	9	10
4	107	111	109	104	110	104	100	100	0	0
5	91	104	97	81	105	96	100	100	500	0
6	101	84	96	84	103	104	100	100	0	0
7	92	99	102	111	94	96	100	100	0	250
8	99	105	99	107	100	104	100	100	250	0
9	96	93	104	88	104	96	100	100	0	250
Ecart maximum	18	16	17	19	10	x	x	x	x	x

Nous avons relevé dans la dernière ligne, l'écart maximum pour chaque colonne, entre les fréquences observées et la fréquence théorique. On voit très nettement apparaître sur cet exemple le phénomène de périodicité sur les derniers chiffres. En effet les cinq dernières colonnes sont inutilisables, et il est très probable que l'on pourrait mettre en évidence une corrélation interne entre les chiffres des quatrième et cinquième colonne.

colonnes / fréquences du	1	2	3	4	5	6	7	8	9	10
0	120	104	103	96	108	103	94	121	77	93
1	105	104	109	97	96	107	108	95	104	3
2	96	92	94	89	113	92	110	107	79	471
3	102	97	105	109	84	108	104	95	103	2
4	94	107	111	81	92	105	88	95	78	190
5	93	95	98	95	84	79	97	96	106	2
6	99	105	112	98	88	97	107	98	75	141
7	106	95	85	104	106	101	94	99	196	2
8	92	108	99	121	105	98	88	103	75	94
9	93	95	84	110	124	110	110	91	107	2
Ecart maximum	20	8	16	21	24	21	12	21	x	x

Ce deuxième exemple a été effectué sur une liste obtenue avec ⊖ variable, par le premier des deux procédés signalés. On peut remarquer deux choses sur ce tableau. D'abord, les écarts maximum relevés sont très élevés (20, 21, 24), ensuite un mauvais choix des données a provoqué une mauvaise répartition dans la dernière colonne où l'on constate un déficit très net des chiffres impairs ; cette mauvaise répartition se répercute sur la neuvième colonne où les chiffres pairs cette fois, sont sensiblement en infériorité.

C'est à dessein que nous avons choisi un exemple défavorable, car on peut constater qu'il reste tout de même huit colonnes d'utilisables. On le contrôle par un test du χ^2.

En effet, à neuf degrés de liberté,

Prob $\left[\chi^2 < \chi^2_0\right]$ = 95 %, donne χ^2_0 = 16,92.

1ère colonne χ^2 = 6,8
2ème colonne χ^2 = 3,18
3ème colonne χ^2 = 9,01
4ème colonne χ^2 =10,76
5ème colonne χ^2 =15,19
6ème colonne χ^2 = 7,66
7ème colonne χ^2 = 6,99
8ème colonne χ^2 = 6,74

colonnes / fréquences du	1	2	3	4	5	6	7	8	9	10
0	94	107	118	88	82	83	117	98	74	0
1	97	112	93	103	114	102	107	100	123	125
2	96	84	114	109	102	96	107	93	74	0
3	105	106	101	103	100	101	102	100	125	209
4	95	100	85	97	106	96	86	96	74	0
5	96	96	79	97	107	105	94	99	126	248
6	114	115	100	101	98	94	92	103	78	0
7	90	94	104	84	86	113	92	112	125	167
8	101	98	99	113	103	104	96	97	75	0
9	112	88	107	105	102	106	107	102	126	251
Ecarts maximum	14	16	21	16	18	17	17	12	x	x

Dans cet exemple, on a utilisé, pour la formation des nombres, une valeur de Θ variable, selon le second procédé indiqué.

La répartition dans la dernière colonne est encore plus mauvaise que dans l'exemple précédent (on peut palier très facilement cet inconvénient), cependant huit colonnes demeurent utilisables et les écarts maximum observés sont en baisse.

χ^2_0 limite, au seuil de 95 % valant toujours 16,92, on trouve :

1ère colonne	χ^2	= 7,64
2ème colonne	χ^2	= 9,10
3ème colonne	χ^2	= 13,12
4ème colonne	χ^2	= 7,11
5ème colonne	χ^2	= 5,62
6ème colonne	χ^2	= 5,18
7ème colonne	χ^2	= 7,36
8ème colonne	χ^2	= 2,36

3. Tests graphiques.

- Nous avons donné dans ce type de tests la répartition de 150 nombres pseudo-aléatoires consécutifs, tirés d'une liste de 1 000 nombres.

En effectuant un rapide groupement en classes, on peut effectuer un test du χ^2.

Pour le premier exemple donné, on trouve $\chi^2 = 5,5$

Pour le deuxième exemple, on trouve $\chi^2 = 2,90$

Alors qu'au seuil de 95 % χ^2 limite = 16,92.

V - Conclusion -

On obtient donc des suites de nombres pseudo-aléatoires satisfaisantes dans les conditions suivantes.

1. Procédé de génération.

$$x_{n+1} = N\, x_n + \Theta - \left[N\, x_n + \Theta\right]$$

$[\ ]$ signifie "partie entière".

Il est indispensable, dans ce mode de génération, de faire varier Θ de façon aléatoire au cours des calculs. Cette variation peut se faire simplement en prenant pour nouvelle valeur de Θ un des nombres aléatoires déjà formés, soit une combinaison de plusieurs de ces nombres soit encore la partie entière de cette combinaison. Il semble que l'on obtienne un bon résultat en modifiant Θ après avoir calculé quatre nombres de la suite.

2. Constantes.

- On peut très bien prendre $\Theta = 0$ au départ.
- Nous n'avons naturellement pas essayé beaucoup de valeurs de N. Des puissances impaires de 3 ou de 7 donnent de bons résultats. Nous recommandons pour notre part $N = 7^9$.
- La valeur de x_0 est sans influence sensible sur le déroulement. Cependant on démontre facilement que le résultat est symétrique en x_0 par rapport à la valeur 0,5. On peut donc se contenter des valeurs comprises entre 0 et 0,5.

Enfin, signalons que sur calculateur I B M 650, on calcule environ 700 nombres, par ce procédé de génération, en une minute environ.

VI - Application :

- Résolution d'une équation de Fredholm par une méthode de Monte-Carlo.

Nous nous contenterons d'un bref exposé théorique, pour plus de détails se reporter à l'article de M.M. GATESOUPE et GUILLOUD dans le vol. X fascicule 4 1961, des Publications de l'Institut de Statistiques de l'Université de Paris.

Soit à résoudre l'équation :

$$f(x) = g(x) + \int_0^1 H(xy)\, f(y)\, dy$$

dans laquelle g(x) est une fonction donnée sur l'intervalle fermé $[0,1]$, H (xy) est une fonction donnée de deux variables continue sur le pavé $[0,1] \times [0,1]$, f (x) est une fonction inconnue, continue sur $[0,1]$.

On sait que f (x) s'obtient comme limite d'une suite de termes f_n (x) calculés par itérations.

Calcul de f_n (x) :

On remplace H (xy) par un opérateur K (xy) et g (x) par g (x) de telle façon que :

$$f_n = K^{n+1} g_1$$

K (xy) est lui-même décomposé sous la forme :

$$K(xy) = z(xy) \quad (xy)$$

où (xy) est une densité de probabilité de passage, relative à une suite de variables aléatoires en chaîne $A_0\ A_1 \ldots\ldots A_n \ldots\ldots$, et z (xy) une fonction déterminée par la donnée de K (xy) et de (xy).

Introduisons alors la variable aléatoire :

$$\zeta = z(A_0, A_1) \ldots\ldots z_n(A_n, A_{n+1})\, g_1(A_{n+1})$$

on montre alors que :

$$E\left(Z / A_0 = x_0\right) = f_n(x_0) \quad \text{avec} \quad 0 \quad x_0 \quad 1$$

Fixons alors :

$$(xy) = (1-S) \text{ sur } [0,1] \times [0,1]$$
$$(xy) = S \quad (y-1-x) \text{ sur } [0,1] \times [0,1]$$
$$(xy) = 0 \text{ sur } [1,2] \times [0,1]$$
$$(xy) = (y-x) \text{ sur } [1,2] \times [1,2]$$

et

$$z(xy) = \frac{1}{1-S} \; H(xy) \text{ sur } [0,1] \times [0,1]$$

$$z(xy) = \frac{1}{S} \quad \text{pour } y - 1 - x = 0$$

$$z(xy) = 1 \text{ ailleurs}$$

où S représente un nombre compris entre 0 et 1 et δ la fonction de Dirac.

Le tirage de la chaîne A_0, A_1 ..... A ..... se fait alors de la façon suivante :

On compare un nombre R_1 uniformément réparti sur (0,1) au nombre S. Si $R_1 \leqslant S$ la partie est terminée et on a :

$$A_1 = A_2 = \ldots\ldots\ldots = A_0 + 1$$

Si $R_1 > S$ on tire un deuxième nombre au hasard R_2 et on fait

$$A_1 = R_2$$

On prend alors un nouveau nombre au hasard R_1 et le cycle recommence.

Nous avons appliqué ce procédé à la résolution de l'équation :

$$f(x) = 2x + \int_0^1 H(xy)\, f(y)\, dy$$

où $H(xy) = \begin{cases} 2y(1-x) & \text{si } y < x \\ 2x(1-y) & \text{si } y \geqslant x \end{cases}$

la solution formelle de l'équation est :

$$f(x) = 2\,\frac{\sin\sqrt{2}\,x}{\sin\sqrt{2}}$$

a) convergence.

pour $x = 0{,}125$, $f(0{,}125) = 0{,}356071$

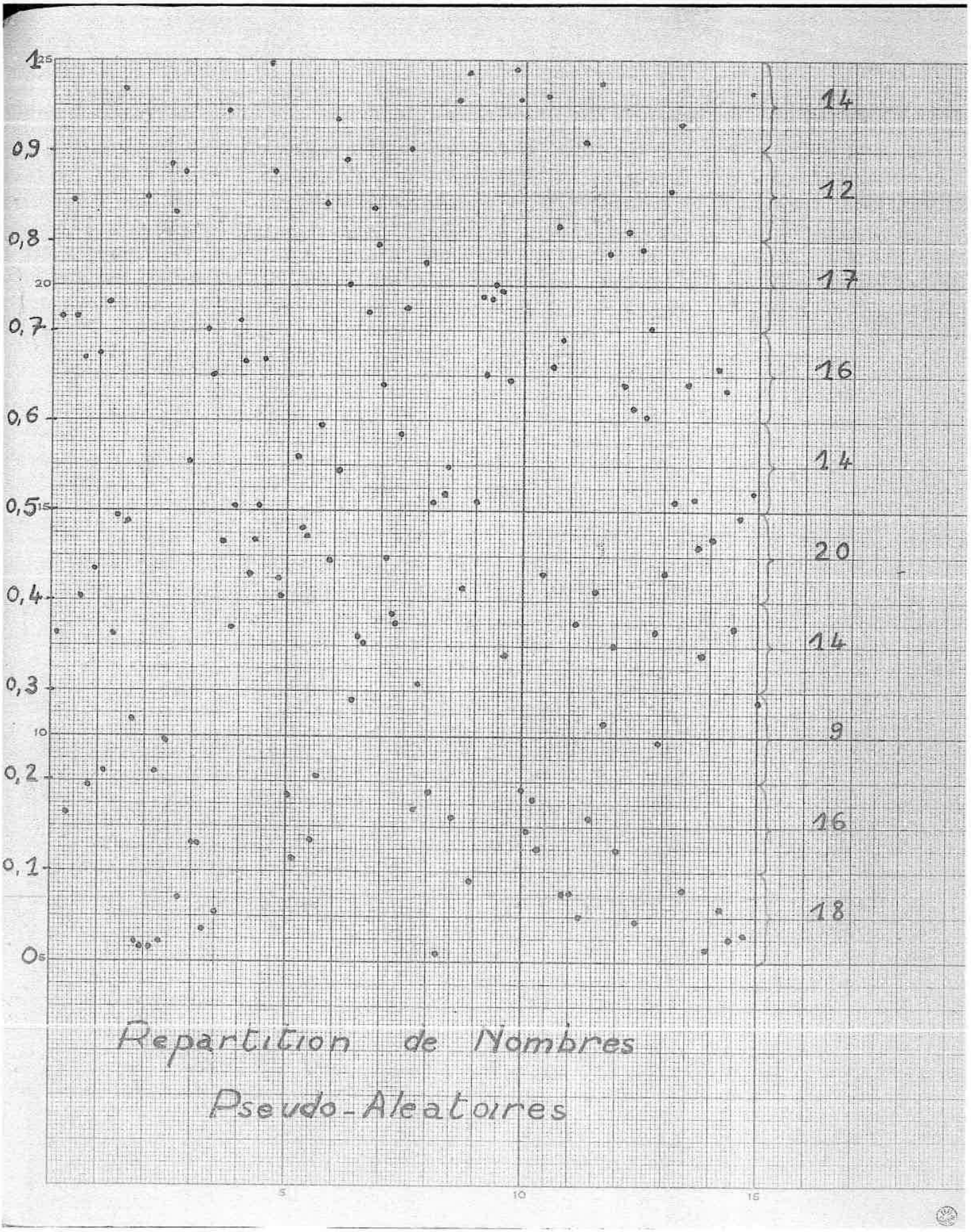
Nous avons calculé $f_{10}(0{,}125)$

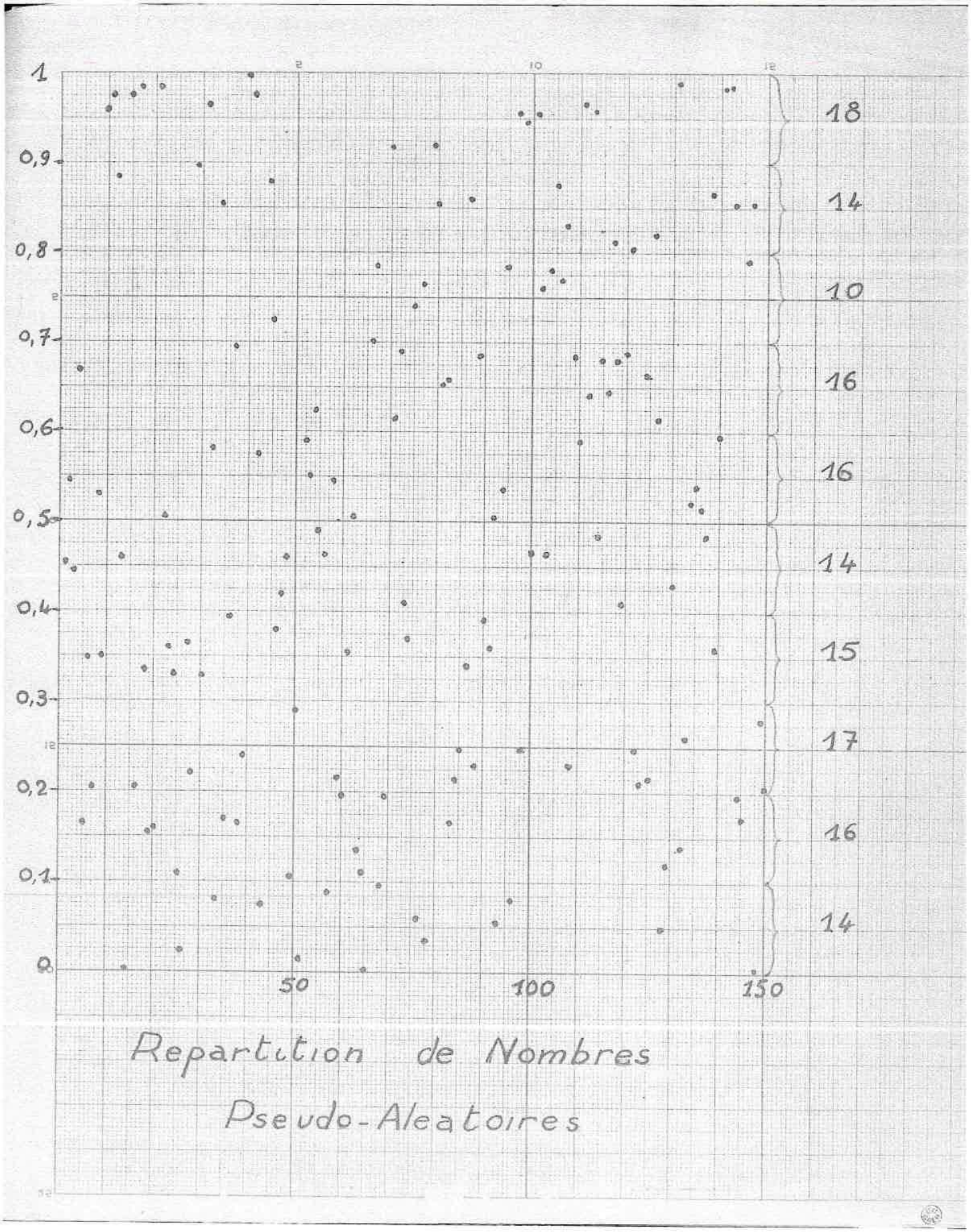
On suivra sur la courbe 1 la convergence de $f_{10}(0{,}125)$ vers $f(0{,}125)$ en fonction du nombre d'épreuves effectuées (Nombre de valeurs de Z obtenues). Nous nous sommes arrêtés après 27 000 épreuves. On constate que la convergence est satisfaisante.

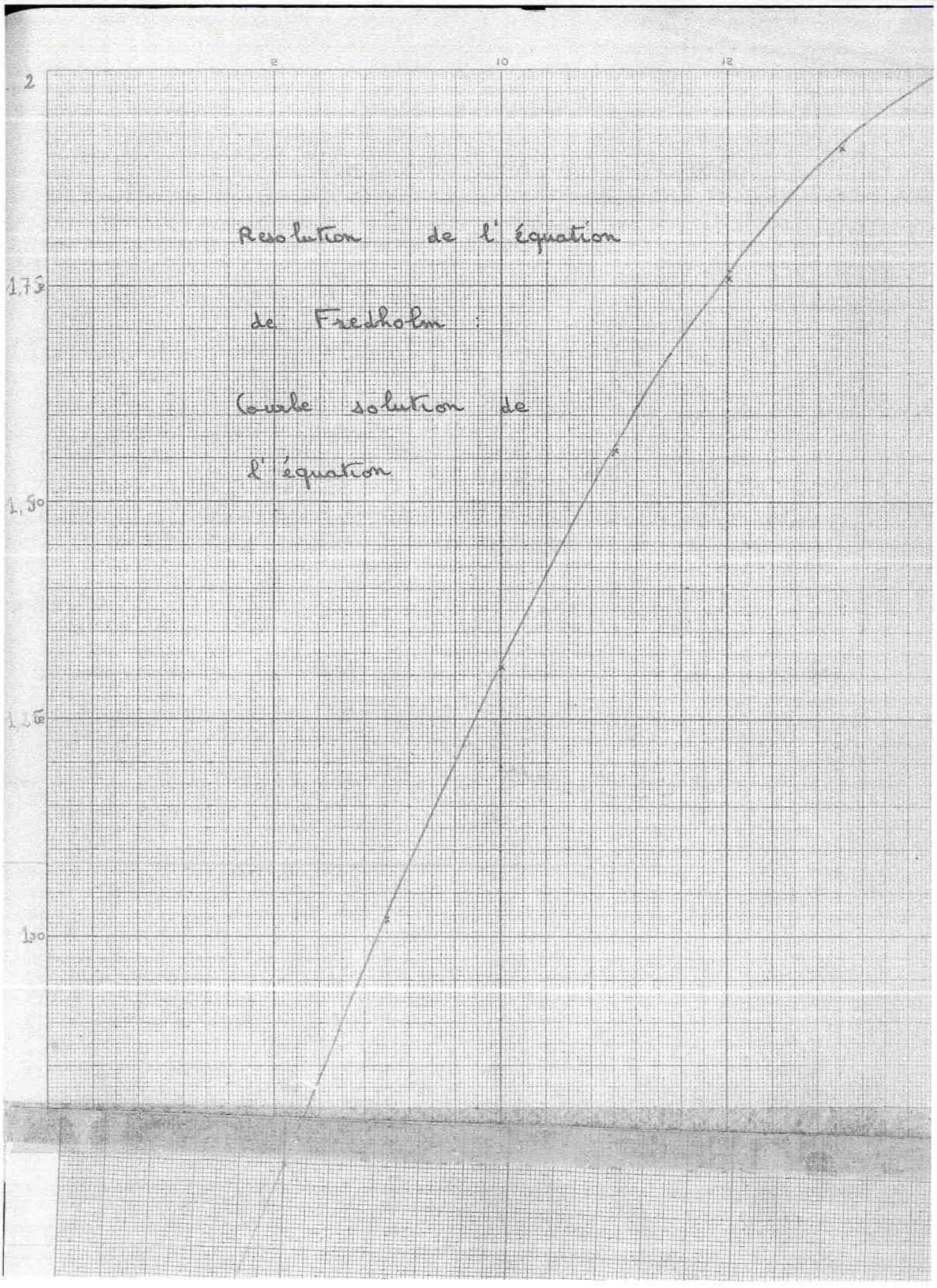
b) détermination de la courbe $f(x)$.

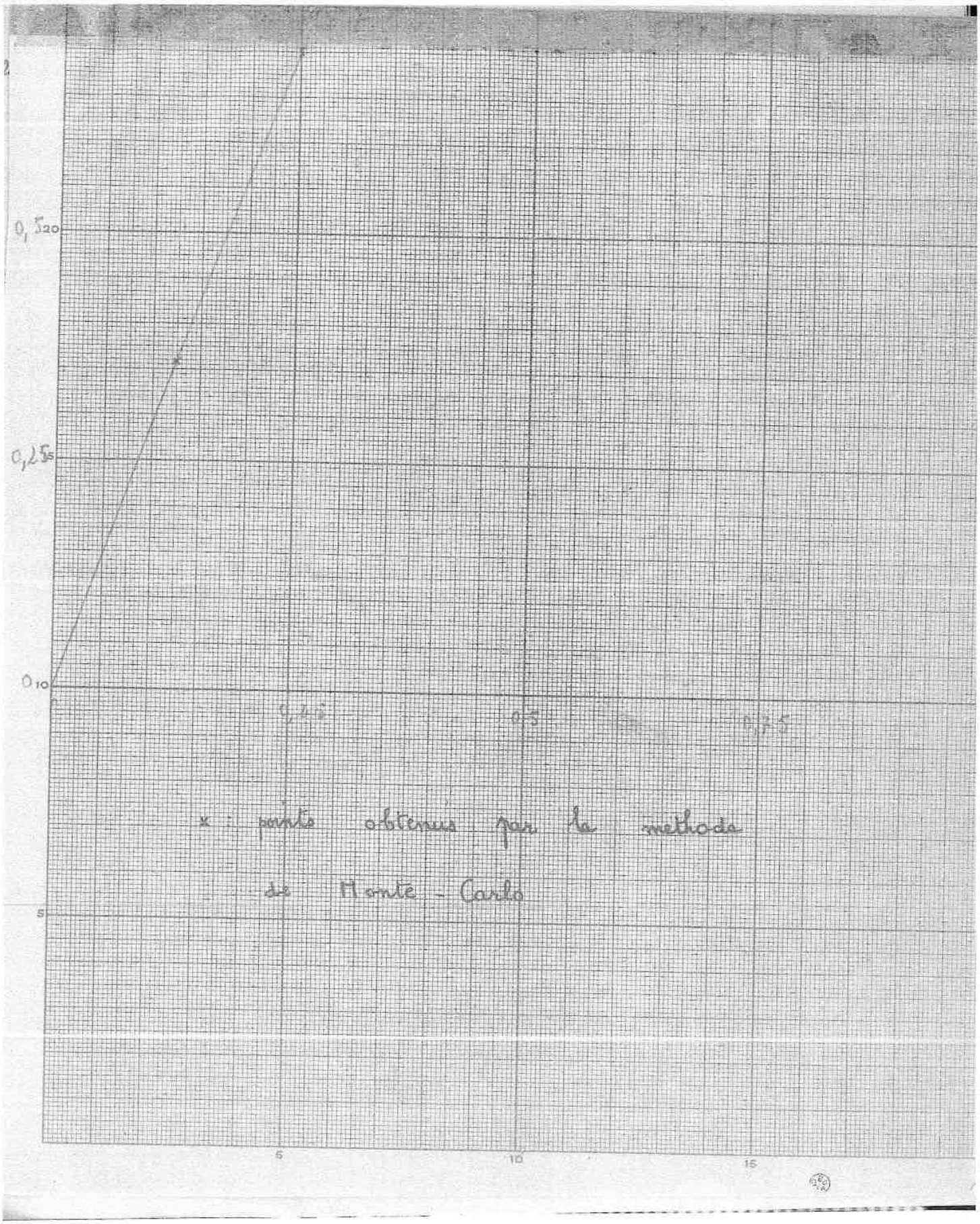
Nous avons représenté la courbe $f(x)$ sur $[0,1]$. Pour chaque point nous avons utilisé un résultat obtenu après 15 000 épreuves. La courbe 2 représente la fonction théorique et la fonction calculée.

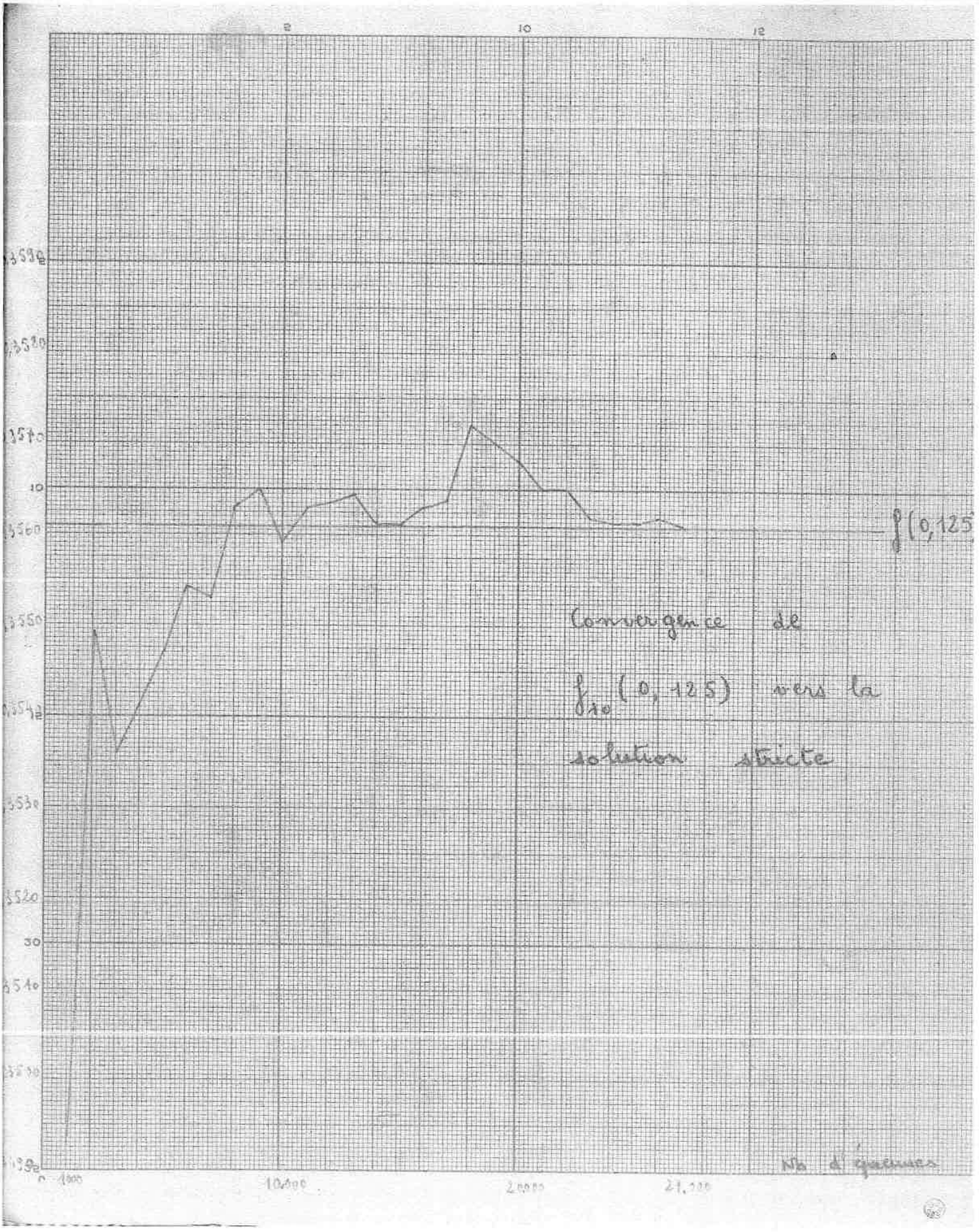
Signalons pour terminer que sur I.B.M. 650 il faut 5 minutes pour calculer 1 000 épreuves.


1
0,9
0,8
0,7
0,6
0,5
0,4
0,3
0,2
0,1
0
14
12
17
16
14
20
14
9
16
18
Repartition de Nombres
Pseudo-Aleatoires


1
0,9
0,8
0,7
0,6
0,5
0,4
0,3
0,2
0,1
0
50
100
150
18
14
10
16
16
14
15
17
16
14
Repartition de Nombres
Pseudo-Aleatoires


8
10
12
2
1,75
1,50
1,25
1,00
Résolution de l'équation
de Fredholm :
Courbe solution de
l'équation


0,5
0,25
0,10
5
10
15
x : points obtenus par la methode de Monte - Carlo


8
10
12
3590
3580
3570
10
3560
3550
3540
3530
3520
30
3510
3500
3490
1000
10000
20000
21.000
Nb d'itérations
$f(0,125$
Convergence de
$f_{10}(0, 12.5)$ vers la
solution stricte

Vu et Approuvé

Nancy, le

Le Doyen,

M. ROUBAULT

Vu et Permis d'imprimer

Nancy, le

le Recteur :

Président du Conseil de
l'Université.

P. IMBS.