ANNALES DE L'I. H. P., SECTION B

ALAIN ROUAULT

Lois de Zipf et sources markoviennes

Annales de l'I. H. P., section B, tome 14, n° 2 (1978), p. 169-188

<http://www.numdam.org/item?id=AIHPB_1978__14_2_169_0>

Ann. Inst. Henri Poincaré,
Vol. XIV, n° 2, 1978, p. 169-188.

Section B :
Calcul des Probabilités et Statistique.

Lois de Zipf et sources markoviennes

par

Alain ROUAULT

Département de Mathématiques, Université Paris-Sud,
Centre d'Orsay, 91405 Orsay

RÉSUMÉ. — On étudie ici les lois de Zipf — d'origine linguistique — dans le cadre d'un n-échantillon d'une v. a. dénombrable. Puis on montre que la suite des valeurs prises par une chaîne de Markov finie, avec un état ne se succédant pas à lui-même (blanc) peut être considérée comme un échantillon d'un ensemble de mots, vérifiant les lois de Zipf. On montre aussi un résultat asymptotique sur une v. a. Z_n^{m*}, nombre de mots — parmi les n premiers — ayant été utilisés au moins m fois.

SUMMARY. — We study here Zipf's laws (of linguistic origine) in the context of a n-sample of a denumerable random value. Then we prove that the sequence of values taken by a finite Markov chain with a not-self-following state (blank) may be considered as a sample from a set of words, fulfilling Zipf's laws. We also prove an asymptotical result on Z_n^{m*}, number of words (among the n first) used more than m-times.

I. INTRODUCTION

Le nom de Zipf est connu des linguistes, des démographes, des biologistes et une littérature mathématique substantielle a été consacrée à ses lois. Nous emploierons la terminologie linguistique pour introduire le problème. Il s'agit d'étudier le rapport entre les mots potentiels d'un texte (le vocabulaire de l'auteur) et les mots rencontrés effectivement au cours de sa lecture, les répétitions...

La première loi concerne l'étude, lorsque la lecture avance, du rapport

du nombre de mots rencontrés s fois, au nombre total de mots différents rencontrés $s = 1, 2, \ldots$

La seconde concerne la relation entre la fréquence d'un mot, et la place qu'il occupe dans le rangement des mots par ordre de fréquences décroissantes.

Enfin une autre loi (sans nom) dit que pour n grand, le nombre de mots différents dans un texte de longueur n, considéré comme aléatoire a approximativement une distribution de Gauss.

Selon un premier point de vue [7] [8] un auteur écrit un texte en effectuant des tirages successifs de mots, avec remplacement, dans un dictionnaire (fini ou non), muni d'une loi de probabilité.

Selon un second [3] [4] [5], on opère uniformément dans un ensemble fini N divisé en M classes et on fait croître N et M.

Nous adoptons le premier. McNeil [8] suppose le dictionnaire fini et estime sa taille. Mandelbrot [7] fait des approximations qui prennent tout leur sens quand on les rapproche des résultats de Karlin [6]. Dans le III nous remarquons ainsi que la première loi de Zipf est une conséquence immédiate de ces résultats (sous une condition simple — grosso modo $p_k \simeq k^{-\frac{1}{\gamma}}$, $0 < \gamma < 1$ pour la probabilité). La seconde nous dira seulement que la k-ième fréquence (par ordre décroissant) converge p. s. vers la k-ième probabilité. Ceci est une simple conséquence de la loi forte des grands nombres.

Dans le IV, on part d'un alphabet fini, comprenant un élément « blanc » ou « espace » et on suppose que le texte est composé, lettre après lettre, selon un procédé markovien. Ceci implique pour les mots formés, un tirage avec remplacement et nous montrons que la probabilité correspondante vérifie la condition ci-dessus, le γ étant le paramètre malthusien d'un certain processus de ramification.

Dans V, nous démontrons une convergence de Z_n^{m*} (nombre de mots, parmi les n premiers, rencontrés au moins m fois) dans le cas où m et n tendent vers l'infini, le quotient $\frac{m}{n}$ restant entre certaines limites.

II. NOTATIONS

Considérons un ensemble dénombrable assimilé à $\mathbb{N}^*$ muni d'une probabilité $p = (p_n)_{n \geq 1}$. Pour tout x tel que $0 < x < 1$, définissons

$$\alpha(x) = \text{card } \{ k : p_k > x \} \tag{2.1}$$

CONDITION 1. — $\forall n$, $p_n > 0$ et $\exists 0 < \gamma < 1$ et $L\,[0, 1] \to \mathbb{R}^+$ à variation lente au voisinage de 0, tels que

$$\alpha(x) = x^{-\gamma} L(x) \tag{2.2}$$

Soit $(X_n)_{n \geq 1}$ une suite de v. a. à valeurs dans $\mathbb{N}^*$, indépendantes et de même loi p. Si δ_n, $n = 1, 2, \ldots$ est la n-ième fonction de Kronecker, notons :

$$X_n^k = \sum_{m=1}^{n} \delta_k(X_m) \qquad k = 1, 2, \ldots$$

$$Z_n^r = \sum_{k \geq 1} \delta_r(X_n^k) \qquad r = 1, 2, \ldots \quad \text{(nombre de valeurs prises } r \text{ fois)}$$

$$Z_n^{r*} = \sum_{s \geq r} Z_n^s, \qquad Z_n^* = Z_n^{1*} \quad \text{(nombre de valeurs prises)} \tag{2.3}$$

On notera $X_n^{(1)} \geq X_n^{(2)} \geq \ldots \geq X_n^{(Z_n^*)} > 0$ le réarrangement par ordre décroissant des Z_n^* termes non nuls de la suite $(X_n^k)_{k \geq 1}$ et $p_{(1)} \geq \ldots \geq p_{(r)} \geq \ldots$ celui correspondant aux p.

III. LOIS DE ZIPF

PROPOSITION 1 (1re loi). — Sous la condition 1, quand n tend vers l'infini, on a

$$\forall s \geq 1, \qquad \lim \text{p. s.} \frac{Z_n^s}{Z_n^*} = \frac{\gamma}{\Gamma(1-\gamma)} \frac{\Gamma(s-\gamma)}{\Gamma(s+1)} \tag{3.1}$$

Démonstration. — Karlin a démontré [6] :

$$\lim_{n \to +\infty} \text{p. s.} \frac{Z_n^*}{EZ_n^*} = 1, \qquad \forall s \geq 1 \quad \lim_{n \to +\infty} \text{p. s.} \frac{Z_n^s}{EZ_n^s} = 1 \qquad \text{sans condition}$$

$$EZ_n^* \underset{n \to +\infty}{\sim} \Gamma(1-\gamma) n^\gamma L\left(\frac{1}{n}\right), \qquad EZ_n^s \underset{n \to +\infty}{\sim} \gamma \frac{\Gamma(s-\gamma)}{\Gamma(s+1)} n^\gamma L\left(\frac{1}{n}\right)$$

sous la condition 1

d'où le résultat. Quand s tend vers l'infini $\dfrac{\gamma}{\Gamma(1-\gamma)} \dfrac{\Gamma(s-\gamma)}{\Gamma(s+1)} \sim \dfrac{\gamma}{\Gamma(1-\gamma)} \dfrac{1}{s^{1+\gamma}}$ qui est de la forme proposée par différents auteurs [3] [4] [5] comme caractérisant la loi de Zipf.

Remarquons que si $s = 1$, on a $\lim \text{p. s.} \dfrac{Z_n^1}{Z_n^*} = \gamma$, d'où une possibilité d'estimer γ.

PROPOSITION 2 (2e loi). — Quand n tend vers l'infini, on a

$$\forall r \geqslant 1 \qquad \lim \text{p. s.} \frac{X_n^{(r)}}{n} = p_{(r)} \tag{3.2}$$

Démonstration. — C'est une conséquence immédiate de la loi forte des grands nombres. Nous supposerons pour simplifier que $\forall n \; p_n \geqslant p_{n+1}$. Soit r fixé et

$$a = \alpha(p_r) \qquad b = (p_r^-) \qquad a < r \leqslant b \tag{3.3}$$

soit R tel que

$$\sum_{k=R+1}^{+\infty} p_k < p_{b+1} \tag{3.4}$$

soit $j_1 < j_2 < \ldots < j_n < \ldots$ l'image de la fonction α et

$$0 < \varepsilon < \frac{1}{2} \min (p_{j_s} - p_{j_{s+1}}) \qquad s = 1, 2, \ldots, \alpha(p_R^-) \tag{3.5}$$

Alors pour $s = 1, 2, \ldots, \alpha(p_R^-)$ les intervalles $[p_{j_s} - \varepsilon, p_{j_s} + \varepsilon]$ ont deux à deux une intersection vide. Or

$$\text{p. s. } \exists N_1 \text{ tel que } n > N_1 \Rightarrow \forall k \leqslant R \qquad \left| \frac{X_n^k}{n} - p_k \right| < \varepsilon$$

$$\text{p. s. } \exists N_2 \text{ tel que } n > N_2 \Rightarrow \left| \frac{\sum_{k=R+1}^{+\infty} X_n^k}{n} - \sum_{k=R+1}^{+\infty} p_k \right| < \varepsilon$$

donc

$$\text{p. s. } \exists N \text{ tel que } n > N \Rightarrow \forall k \leqslant b \qquad \left| \frac{X_n^k}{n} - p_k \right| < \varepsilon$$

$$\forall b < k \leqslant R \qquad \frac{X_n^k}{n} < p_k + \varepsilon \leqslant p_{b+1} + \varepsilon$$

$$\forall k \geqslant R + 1 \qquad \frac{X_n^k}{n} \leqslant \frac{1}{n} \sum_{R+1}^{+\infty} X_n^j$$

$$< \left(\sum_{R+1}^{+\infty} p_j \right) + \varepsilon < p_{b+1} + \varepsilon$$

Par suite, p. s. il existe N, tel que $n > N$ entraîne :
il y a exactement

j_1 fréquences dans l'intervalle $[p_{j_1} - \varepsilon, p_{j_1} + \varepsilon]$
$j_2 - j_1$ » » $[p_{j_2} - \varepsilon, p_{j_2} + \varepsilon]$
.....
$b - a$ » » $[p_b - \varepsilon, p_b + \varepsilon] = [p_r - \varepsilon, p_r + \varepsilon]$

et donc

$$\text{p. s. } \forall \varepsilon > 0 \quad \exists N : \qquad n > N \Rightarrow \left| \frac{X_n^{(r)}}{n} - p_r \right| < \varepsilon .$$

D'où la proposition.

On remarque que de la majoration exponentielle classique (à r fixé) de $P\left(\left| \frac{X_n^r}{n} - p_r \right| > \varepsilon \right)$, on peut déduire une majoration exponentielle pour $P\left(\left| \frac{X_n^{(r)}}{n} - p_{(r)} \right| > \varepsilon \right)$.

IV. SOURCES MARKOVIENNES

1) Définitions

a) On appelle source markovienne une chaîne de Markov définie sur un espace $\bar{E} = \{ e_1, \ldots, e_d, \sigma \}$ $(d \geqslant 1)$ ayant pour probabilité initiale la mesure de Dirac en σ et dont tous les éléments de la matrice de transition sont strictement positifs, sauf celui correspondant à un passage de σ à σ.

$\Omega = \bar{E}^{\mathbb{N}}$ $X = (X_k)_{k \in \mathbb{N}}$ les coordonnées $P(X_0 = \sigma) = 1$

$$A = \begin{bmatrix} p_{11} & \cdots & p_{1d} & p_{1\sigma} \\ & & & \\ p_{d1} & \cdots & p_{dd} & p_{d\sigma} \\ p_{\sigma 1} & \cdots & p_{\sigma d} & 0 \end{bmatrix} \tag{4.1}$$

$$T_1 = \inf \{ n \geqslant 1, \ X_n = \sigma \} \qquad T_{k+1} = \inf \{ n > T_k, \ X_n = \sigma \} \qquad k \geqslant 1$$

$$E = \{ e_1, \ldots, e_d \} \qquad E^* = \bigcup_{n=1}^{+\infty} E^n$$

b) Un élément de E^n s'appelle un mot de longueur n et sera noté $x = x_1 x_2 \ldots x_n$ avec, pour tout i, $x_i \in E$.

$$\forall k \quad M_k \ \Omega \to E^* \qquad \begin{matrix} M_1(\omega) = X_1(\omega) X_2(\omega) \ldots X_{T_1(\omega)-1}(\omega) \\ k > 1 \qquad M_k(\omega) = X_{T_{k-1}(\omega)+1}(\omega) \ldots X_{T_k(\omega)-1}(\omega) \end{matrix} \tag{4.2}$$

2) Échantillonnage

PROPOSITION 3. — Les M_k sont indépendants et de même loi (notée μ).

Ceci résulte du caractère markovien et irréductible de la chaîne. Les n premiers mots (aléatoires) sont donc n tirages indépendants dans l'ensemble E* suivant la loi μ. On est alors conduit à se demander si la loi μ vérifie la condition 1 (cf. II) c'est-à-dire à étudier :

$$0 < t < 1 \qquad \alpha(t) = \text{card } \{ x \in E^* ; \mu(x) > t \}$$

La réponse est donnée par le

THÉORÈME 4. — Quand t tend vers zéro, il existe $0 < \gamma < 1$ tel que $t^\gamma \alpha(t)$ ait une limite stricte positive.

Conséquences. — Une source markovienne vérifie les lois de Zipf (III) et aussi possède la propriété démontrée par Karlin [6] : Z_n^* convenablement normalisée converge faiblement vers une loi de Gauss. Notre modèle, s'il ne décrit que grossièrement la réalité, se trouve donc néanmoins en accord avec certaines tendances décelées depuis longtemps dans les données empiriques.

Démonstration du théorème 4. —

$$\alpha(t) = \sum_{n=1}^{+\infty} \text{card } \{ x \in E^n ; -\log \mu(x) < -\log \mu(t) \} = \sum_{n=1}^{+\infty} \alpha_n(t) \quad (4.3)$$

or si

$$x \in E^n, \quad n \geqslant 2, \quad x = x_1 x_2 \ldots x_n \qquad \mu(x) = p_{\sigma,x_1} p_{x_1,x_2} \ldots p_{x_n,\sigma} \quad (4.4)$$

$$-\log \mu(x) = -\log p_{\sigma,x_1} - \log p_{x_n,\sigma} + \sum_{k=1}^{n-1} -\log p_{x_k,x_{k+1}} \quad (4.5)$$

Posons

$$u_k = -\log p_{x_k,x_{k+1}} \qquad k = 1, 2, \ldots, n-1 \quad (4.6)$$

Les u_k appartiennent à $F = \{ f_{i,j} = -\log p_{i,j} \; 1 \leqslant i, j \leqslant d \}$ F a d^2 éléments ; mais si u_k est donné, u_{k+1} ne peut prendre que d valeurs, celles correspondant aux transitions à partir de x_{k+1} si u_1 est donné, il y a exactement d^{n-1} $n-1$-uplets possibles $(u_2, \ldots, u_n)$. Nous devons faire du dénombrement sur ces n-uplets. On va donc passer par un calcul sur un modèle

probabiliste sur F assurant à tous les $n-1$-uplets ci-dessus la même probabilité.

On prend $F^{\mathbb{N}}$, $(U_k)_{k\in\mathbb{N}^*}$ les coordonnées. Si (i, j) désigne $f_{i,j}$ la matrice de transition sera :

$$\Pi=(\pi_{a,b})_{a,b\in F}=\frac{1}{d}\begin{array}{c}\begin{matrix}(1,1)\ldots(1,d)\ldots\ldots\ldots\ldots(d,1)\ldots(d,d)\end{matrix}\\ \begin{bmatrix}1\ldots\ldots 1\ 0\ldots\ldots\ldots\ldots 0\ldots\ldots 0\\ 0\ldots\ldots 0\ 1\ldots 1\ 0\ldots\ldots 0\ldots\ldots 0\\ \vdots\ \ \vdots\ \ \vdots\ \ \vdots\ \ \vdots\\ 0\ldots\ldots 0\ldots\ldots\ldots\ldots 0\ 1\ldots\ldots 1\\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\\ 1\ldots\ldots 1\ 0\ldots\ldots\ldots\ldots\ldots\ldots 0\\ 0\ldots\ldots 0\ 1\ldots\ldots 1\ 0\ldots\ldots 0\ldots\ldots 0\\ \vdots\ \ \vdots\ \ \vdots\ \ \vdots\ \ \vdots\\ 0\ldots\ldots\ldots\ldots\ldots\ldots\ldots 0\ 1\ldots\ldots 1\end{bmatrix}\end{array}\begin{matrix}(1,1)\\ \\ \\ (1,d)\\ \\ (d,1)\\ \\ \\ (d,d)\end{matrix}\qquad(4.7)$$

c'est-à-dire $\pi_{(i,j),(k,m)}=\dfrac{1}{d}$ si $j=k$, 0 sinon.

On notera Q^a la probabilité correspondant à l'état initial a, $a\in F$.

$$\alpha_{n+1}(t)=\sum_{\substack{j,k\leqslant d\\ m,q\leqslant d}}\text{card}\,\{(u_1,\ldots,u_n);\ u_1=f_{j,m},\ u_n=f_{q,k}\quad\text{et}\quad u_1+\ldots+u_n<\log(t^{-1}p_{k,\sigma}p_{\sigma,j})\}\qquad(4.8)$$

c'est-à-dire

$$\alpha_{n+1}(t)=\sum_{j,k,m,q\leqslant d}d^{n-1}Q^{f_{j,m}}(U_n=f_{q,k}\text{ et }U_1+\ldots+U_n<\log(t^{-1}p_{k,\sigma}p_{\sigma,j}))\qquad(4.9)$$

Soit $S_n=\sum_1^n U_i$ et si $a, b\in F$

$$G_{a,b}(\theta)=\sum_{k=1}^{+\infty}Q^a(S_k<\theta,\ X_k=b)d^{k-1}\qquad(4.10)$$

alors

$$\sum_{n=1}^{+\infty}\alpha_{n+1}(t)=\sum_{a,b\in F}G_{a,b}(\log(t^{-1}\varphi(a,b)))\qquad(4.11)$$

où $\varphi(a, b)$ est défini ainsi : si

$$a = f_{j,m} \quad b = f_{q,k} \qquad \varphi(a, b) = p_{k\sigma}p_{\sigma j}. \tag{4.12}$$

Il nous faut donc le comportement asymptotique de $G_{a,b}(\theta)$ quand θ tend vers l'infini. On a les relations

$$\begin{aligned} a \neq b, \qquad & G_{a,b}(\theta) = d\, 1_{\theta > a} \sum_{c \in F} \pi_{a,c} G_{c,b}(t - a) \\ & G_{a,a}(\theta) = 1_{\theta < a} + d\, 1_{\theta > a} \sum_{c \in F} \pi_{a,c} G_{c,a}(t - a) \end{aligned} \tag{4.13}$$

c'est-à-dire, en passant aux transformées de Laplace :

$$\Phi_{a,b}(\lambda) = \int_0^{+\infty} e^{-\lambda x} G_{a,b}(x)dx \qquad \lambda \in \mathbb{C},\ \mathrm{Re}\ \lambda > 0$$

$$\Phi(\lambda) = (\Phi_{a,b}(\lambda))_{a,b \in F}, \qquad H(\lambda) = (h_{a,b}(\lambda))_{a,b \in F} \qquad h_{a,b}(\lambda) = d\, \pi_{a,b} e^{-\lambda\omega} \tag{4.14}$$

$D(\lambda)$ matrice diagonale $d_a(\lambda) = \dfrac{e^{-\lambda a}}{\lambda}$, I matrice identité dans F (4.13) se transforme en

$$\Phi(\lambda)(I - H(\lambda)) = D(\lambda) \tag{4.15}$$

On aperçoit alors une analogie avec des processus de ramification : considérons une population ayant d^2 « types », dont chaque élément produit à la fin de sa vie d'enfants, tous de même type, la transition entre le type du géniteur et celui de ses enfants étant régie par la matrice Π (4.7). On suppose que la durée de vie d'un individu ne dépend que de son type. Alors les $G_{a,b}$ interviennent dans le calcul du nombre moyen d'individus en vie à l'instant t (cf. [9] et [1]).

Le résultat final est que :

1) $\exists f, g\ 0 < g < f$ tels que $\forall a, b$

$$G_{a,b}(\theta) = k_{a,b} e^{f\theta} + O(e^{g\theta}) \tag{4.16}$$

où $k_{a,b}$ ne dépend pas de θ

2) f est racine de l'équation $\rho(\lambda) = 1$ où $\rho(\lambda)$ est la « plus grande valeur propre (*) de $H(\lambda)$. La fonction ρ est décroissante et continue. (4.17)

Il suffira alors de montrer que $\rho(0) > 1$ et $\rho(1) < 1$ pour pouvoir conclure :

(*) Racine de Frobenius de $H(\lambda)$.

$\alpha(t)$ est somme d'un nombre fixe de termes dont la partie principale, quand t tend vers 0 est en $ke^{f(-\log t)} = kt^f$ avec $0 < f < 1$. f est le γ cherché.

Or soit μ une valeur propre de $H(\lambda)$ et $(w_{i,j})$ un vecteur propre correspondant à μ.

$H(\lambda)w = \mu w$ c'est-à-dire $\sum_{b\in F} h_{a,b}(\lambda)w_b = \mu w_a$ ou encore si $a = (i, j)$

$$\sum_{l=1}^{d} (p_{i,j})^{\lambda} w_{j,l} = \mu w_{i,j} \tag{4.18}$$

d'où

$$\mu w_{i,j} = (p_{i,j})^{\lambda} \sum_{l} w_{j,l} \quad \text{et} \quad \sum_{j=1}^{d} (p_{i,j})^{\lambda} \left(\sum_{l} w_{j,l} \right) = \mu \left(\sum_{j} w_{i,j} \right) \tag{4.19}$$

donc μ est valeur propre de la matrice $C(\lambda) = (c_{i,j}(\lambda))_{i,j\leq d} c_{i,j} = (p_{i,j})^{\lambda}$ et

$$v = (v_i)_{1\leq i\leq d} \qquad v_i = \sum_{j=1}^{d} w_{i,j} \tag{4.20}$$

un vecteur propre de $C(\lambda)$ correspondant à μ; de plus

$$w_{i,j} = \frac{1}{\mu} (p_{i,j})^{\lambda} v_j \qquad \text{si} \quad \mu \neq 0 \tag{4.21}$$

et

$$\forall j \quad \sum_{l} w_{j,l} = 0$$

définit le sous-espace propre correspondant à la valeur propre 0 de $H(\lambda)$.

Donc 0 est toujours valeur propre de $H(\lambda)$ et les autres sont les valeurs propres non nulles de $C(\lambda)$.

Pour $\lambda = 0$

$$C(0) = \begin{pmatrix} 1 \cdots\cdots 1 \\ \vdots \quad\quad \vdots \\ 1 \cdots\cdots 1 \\ \vdots \quad\quad \vdots \\ 1 \cdots\cdots 1 \end{pmatrix} \tag{4.22}$$

donc la plus grande valeur propre est d. Donc $\rho(0) > 1$.

Pour $\lambda = 1$

$$C(1) = (p_{i,j})_{1\leq i,j\leq d} \tag{4.23}$$

or d'après la définition (IV.1) pour tout i, $\sum_{j=1}^{d} p_{i,j} < 1$ si μ est une valeur propre positive et v un vecteur propre de C(1) correspondant à μ avec par exemple $|v_1| = \max_{1 \leqslant j \leqslant d} |v_j|$

$$\mu |v_1| = |\Sigma p_{1j} v_j| \leqslant \sum_{1}^{d} p_{1j} |v_j| \leqslant |v_1| \sum_{1}^{d} p_{1j} < |v_1| \qquad (4.24)$$

d'où $\rho(1) < 1$.

Ceci termine la démonstration du théorème 4.

V. ÉTUDE DE Z_n^{m*}

Reprenons les notations de II. Il est facile de voir l'équivalence :

$$\{ Z_n^{m*} < k \} = \{ X_n^{(k)} < m \} \qquad (5.1)$$

et donc la proposition 2 entraîne :

COROLLAIRE 5. — $\forall 0 < f < 1$ telle que $\forall k$, $f \neq p_k$

$$\text{p. s. } Z_n^{(fn)*} \xrightarrow[n \to +\infty]{} \alpha(f) \qquad (5.2)$$

Ceci nous amène à considérer les Z_n^{m*} d'un peu plus près. Karlin [6] a montré que pour m fixé, n tendant vers l'infini $\dfrac{Z_n^{m*}}{EZ_n^{m*}} \to 1$ p. s. et a donné un équivalent de EZ_n^{m*}. Le corollaire 5 ci-dessus nous suggère d'étudier le cas n, m tendant simultanément vers l'infini. Cependant si $\dfrac{m}{n}$ passe par une valeur p_k il y aura problème d'où :

THÉORÈME 6. — Soient l et L, deux réels fixés tels que $0 < l < L < 1$ et $[l, L]$ ne contienne pas de p_k ($k \in \mathbb{N}^*$). Si on pose

$$Z_n^{m*} = \alpha\left(\frac{m}{n}\right) + \Delta_{m,n} \qquad (5.3)$$

il existe K et δ positifs tels que

$$\forall m, n \qquad l < \frac{m}{n} < L \Rightarrow E(\Delta_{m,n}^2) \leqslant K e^{-\delta n} \qquad (5.4)$$

Démonstration du théorème 6. — Posons

$$A_{s,k} = \{ X_n^k = s \} \tag{5.5}$$

$$Z_n^{m*} = \sum_{s \geqslant m | k \geqslant 1} 1_{A_{s,k}}, \qquad EZ_n^{m*} = \sum_{\substack{s \geqslant m \\ k \geqslant 1}} P(A_{s,k}) \tag{5.6}$$

$$(Z_n^{m*})^2 = Z_n^{m*} + \sum_{\substack{s,t \geqslant m \\ k,i \geqslant 1 \\ (s,k) \neq (t,i)}} 1_{A_{s,k}} 1_{A_{t,i}} \tag{5.7}$$

Or si $s \neq t$

$$A_{s,k} A_{t,k} = \emptyset, \qquad (Z_n^{m*})^2 - Z_n^{m*} = \sum_{\substack{k \neq i \\ s,t \geqslant m}} 1_{A_{s,k}} 1_{A_{t,i}} \tag{5.8}$$

d'où

$$E(Z_n^{m*})^2 - EZ_n^{m*} = \sum_{\substack{k \neq i \\ s,t \geqslant m}} P(A_{s,k} A_{t,i}) \tag{5.9}$$

or

$$P(A_{s,k}) = \binom{n}{s} (p_k)^s (1 - p_k)^{n-s}, \qquad s \leqslant n \tag{5.10}$$

$$P(A_{s,k} A_{t,i}) = \binom{n}{s} \binom{n-s}{i} (p_k)^s (p_i)^t (1 - p_k - p_i)^{n-s-t} \tag{5.11}$$

si $k \neq i$ et $s + t \leqslant n$.

Il faut dissocier le cas $2m > n$ et $2m \leqslant n$. On se placera désormais dans le cas $2m \leqslant n$.

Sur le triangle $0 \leqslant x \leqslant 1$, $0 \leqslant y \leqslant 1$, $x + y \leqslant 1$, définissons

$$\Phi(x, y) = \sum_{\substack{s,t \geqslant m \\ s+t \leqslant n}} \binom{n}{s} \binom{n-s}{t} x^s y^t (1 - x - y)^{n-s-t} \tag{5.12}$$

alors

$$E(Z_n^{m*})^2 - EZ_n^{m*} = \sum_{k \neq i} \Phi(p_k, p_i) \tag{5.13}$$

si tous les p_k sont inférieurs ou égaux à $\frac{1}{2}$, on a alors

$$E(Z_n^{m*})^2 - EZ_n^{m*} = \sum_{k,i \geqslant 1} \Phi(p_k, p_i) - \sum_{k \geqslant 1} \Phi(p_k, p_k) \tag{5.14}$$

De même si on pose

$$\psi(x) = \sum_{s=m}^{n} \frac{n}{s} x^s(1-x)^{n-s} \qquad x \in [0, 1] \tag{5.15}$$

$$EZ_n^{m*} = \sum_{k \geq 1} \psi(p_k) \tag{5.16}$$

Interprétation des fonctions Φ *et* ψ. — Soient $0 \leqslant x,\ y \leqslant 1\ x + y \leqslant 1$

$$I_1 = [0, x], \quad I_2 =]x, 1 - y[, \quad I_3 = [1 - y, 1]$$

et considérons n tirages au sort sur $[0, 1]$, indépendants et suivant la loi uniforme, $(U_1, \ldots, U_n)$ et soit $U_n^{(1)} \leqslant U_n^{(2)} \leqslant \ldots \leqslant U_n^{(n)}$ le n-uplet ordonné. (5.17)

Alors $\Phi(x, y) = P$(au moins m points dans I_1, au moins m points dans I_3)

$$= P(U_n^{(m)} < x,\ U_n^{(n-m+1)} > 1 - y) \tag{5.18}$$

$(U_n^{(m)},\ 1 - U_n^{(n-m+1)})$ a pour densité (cf. Wiks [*10*])

$$D_{m,m;n-2m+1}(\xi, \eta) = \frac{n!}{[(m-1)!]^2(n-2m)!} \xi^{m-1}\eta^{m-1}(1 - \xi - \eta)^{n-2m},$$
$$0 < \xi, \eta < 1, \quad \xi + \eta < 1 \tag{5.19}$$

De même $\psi(x) = P(U_n^{(m)} < x)$ or $U_n^{(m)}$ a pour densité

$$\beta_{m,n-m+1}(\xi) = \frac{n!}{(m-1)!(n-m)!} \xi^{m-1}(1 - \xi)^{n-m}, \qquad 0 < \xi < 1 \tag{5.20}$$

d'où

$$\Phi(x, y) = \int_0^x \int_0^y D_{m,m,n-2m+1}(\xi, \eta) d\xi d\eta$$
$$\psi(x) = \int_0^x \beta_{m,n-m+1}(\xi) d\xi \tag{5.21}$$

Mais, en intégrant par parties :

$$\sum_{k,i} \Phi(p_k, p_i) = \iint \alpha(dx)\alpha(dy)\Phi(x, y)$$
$$= \iint_{\substack{0 \leqslant x \leqslant 1 \\ 0 \leqslant y \leqslant 1 \\ x - y \leqslant 1}} \alpha(x)\alpha(y) D_{m,m;n-2m+1}(x, y) dx dy \tag{5.22}$$

$$\sum_k \Phi(p_k, p_k) = -\int_0^1 \alpha(dx)\Phi(x, x) = \int_0^1 \alpha(x)\left[\frac{\partial \Phi}{\partial x}(x, x) + \frac{\partial \Phi}{\partial y}(x, x)\right] dx$$
$$= 2\int_0^1 \alpha(x)\left[\int_0^x D_{m,m;n-2m+1}(x, \eta) d\eta\right] dx \tag{5.23}$$

De même

$$\sum_k \psi(p_k) = -\int_0^1 \alpha(dx)\psi(x) = \int_0^1 \alpha(x)\beta_{m,n-m+1}(x)dx \qquad (5.24)$$

Or il est bien connu que si G et G' sont deux v. a. indépendantes, de lois gamma de paramètres λ et λ', G + G' a une loi gamma de paramètre $\lambda + \lambda'$ et $\dfrac{G}{G + G'}$ une loi bêta de paramètres λ et λ'.

Le résultat analogue à 3 dimensions est le suivant : si G_1, G_2, G_3 sont des v. a. indépendantes, de lois gamma de paramètres λ_1, λ_2, λ_3 et si on pose

$$H_1 = \frac{G_1}{G_1 + G_2 + G_3}, \qquad H_2 = \frac{G_2}{G_1 + G_2 + G_3},$$

(H_1, H_2) a une loi $D_{\lambda_1,\lambda_2,\lambda_3}$ (et H_1 une loi $\beta_{\lambda_1,\lambda_2+\lambda_3}$, H_2 une loi $\beta_{\lambda_2,\lambda_1+\lambda_3}$).

La loi $D_{m,m;n-2m+1}$ (5.19) peut alors être considérée comme la loi d'un couple de v. a. $W_{m,n}$, $W'_{m,n}$ obtenu de la façon suivante : soient S_m, S'_m, S''_{n-2m+1} 3 v. a. indépendantes, de loi gamma de paramètre m (resp. m, $n - 2m + 1$) et

$$W_{m,n} = \frac{S_m}{S_m + S'_m + S''_{n-2m+1}} \quad W'_{m,n} = \frac{S'_m}{S_m + S'_m + S''_{n-2m+1}} \qquad (5.25)$$

Les formules (5.14) à (5.25) entraînent, si E désigne à la fois l'espérance dans l'espace où ont été définies les v. a. de (5.25) et l'espérance dans l'espace où ont été définies les (Z_n^m) :

$$\begin{aligned} EZ_n^{m*} &= E\alpha(W_{m,n}) \\ E\left[Z_n^{m*} - \alpha\left(\frac{m}{n}\right)\right]^2 &= E\left[\alpha(W_{m,n}) - \alpha\left(\frac{m}{n}\right)\right]\left[\alpha(W'_{m,n}) - \alpha\left(\frac{m}{n}\right)\right] + E\alpha(W_{m,n}) \\ &\quad - 2E[\alpha(W_{m,n})1_{W'_{m,n} < W_{m,n}}] \end{aligned} \qquad (5.26)$$

L'idée de la suite de la démonstration est la suivante : d'après la loi forte des grands nombres

$$\lim_{m\to+\infty} \text{p. s.} \frac{S_m}{m} = 1 \qquad \lim_{m\to+\infty} \text{p. s.} \frac{S'_m}{m} = 1 \qquad \lim_{n-2m+1\to+\infty} \text{p. s.} \frac{S''_{n-2m+1}}{n-2m+1} = 1$$

donc si $\dfrac{m}{n}$ a une limite f $\lim\limits_{m,n\to+\infty}$ p. s. $(W_{m,n}, W'_{m,n}) = (f, f)$.

α est une fonction en escalier, décroissante, tendant vers l'infini quand son argument tend vers zéro, intégrable :

$$\int_0^1 \alpha(x)dx = \sum_1^{+\infty} p_k = 1$$

Nous allons avoir besoin d'une majoration pour $P\left|W_{m,n} - \frac{m}{n}\right| > \varepsilon$ que nous démontrerons ultérieurement.

LEMME 7. — Soit $W_{m,n}$ une v. a. ayant une loi bêta de paramètres m et $n - m + 1$. Alors :

$$\forall 0 < \varepsilon < 1 \quad \exists N : \quad n \geqslant N \Rightarrow \forall m \leqslant \frac{n}{2}$$

$$P\left(\left|W_{m,n} - \frac{m}{n}\right| > \varepsilon\right) \leqslant 6e^{-\frac{m\varepsilon^2}{128}} \tag{5.27}$$

D'après l'hypothèse du théorème 6, il existe $i \in \mathbb{N}$, $l, L \in [0, 1]$, $N' \in \mathbb{N}$ tels que

$$n, m \geqslant N' \Rightarrow p_{i+1} < l < \frac{m}{n} < L < p_i \tag{5.28}$$

(on suppose que les p_k sont décroissants pour simplifier ; le cas $l > p_1$ sera vu plus tard). Or

$$\begin{aligned} &\alpha(W) = \alpha\left(\frac{m}{n}\right) && \text{si} \quad p_{i+1} \leqslant W < p_i \\ &\alpha(W) < \alpha\left(\frac{m}{n}\right) && \text{si} \quad W \geqslant p_i \\ &\alpha(W) > \alpha\left(\frac{m}{n}\right) && \text{si} \quad W < p_{i+1} \end{aligned} \tag{5.29}$$

(5.26) entraîne donc

$$\begin{aligned} E\left[Z_n^{m*} - \alpha\left(\frac{m}{n}\right)\right]^2 \leqslant\ & E\alpha(W_{m,n})\alpha(W'_{m,n})1_{W_{m,n} < p_{i+1}} 1_{W'_{m,n} < p_{i+1}} \\ & + \alpha\left(\frac{m}{n}\right)^2 P(W_{m,n} \geqslant p_i) \\ & + 2\alpha\left(\frac{m}{n}\right) P(W_{m,n} \geqslant p_i) \\ & + E\alpha(W_{m,n}) 1_{W_{m,n} < p_{i+1}} \end{aligned} \tag{5.30}$$

Or si $n, m > N'$, $\alpha\left(\frac{m}{n}\right) = i$ fixe

$$P(W_{m,n} \geqslant p_i) \leqslant P\left(\left|W_{m,n} - \frac{m}{n}\right| > p_i - \frac{m}{n}\right) \leqslant P\left(\left|W_{m,n} - \frac{m}{n}\right| > p_i - L\right)$$

donc d'après (5.27)

$$\begin{aligned} n, m \underset{N'}{\geqslant} N \Rightarrow P(|\, W_{m,n} \geqslant p_i) &\leqslant 8e^{-\frac{m(p_i - L)^2}{128}} \\ &\leqslant 8e^{-\frac{l(p_i - L)^2}{128}} \end{aligned} \tag{5.31}$$

qui est une majoration du type demandé (5.4).

D'après (5.26) il reste donc à trouver des majorations exponentielles

$$\begin{aligned} E_1 &= E\alpha(W_{m,n})\alpha(W'_{m,n})1_{W_{m,n}<p_{i+1}}1_{W'_{m,n}<p_{i+1}} \\ E_2 &= E\alpha(W_{m,n})1_{W_{m,n}<p_{i+1}} \end{aligned} \tag{5.32}$$

$$E_1 = \int_0^{p_{i+1}}\int_0^{p_{i+1}} \alpha(x)\alpha(y)D_{m,m;n-2m+1}(x, y)dxdy$$

soit b quelconque, inférieur à p_{i+1}

$$E_1 \leqslant \int_0^b\int_0^b \alpha(x)\alpha(y)D_{m,m;n-2m+1}(x, y)dxdy + \alpha(b)^2P(W_{m,n} < p_{i+1}) + 2\alpha(b)\int_0^b \alpha(x)\beta_{m,n-m+1}(x)dx \tag{5.33}$$

De même

$$E_2 \leqslant \int_0^b \alpha(x)\beta_{m,n-m+1}(x)dx + \alpha(b)P(W_{m,n} < p_{i+1}) \tag{5.34}$$

On majore $P(W_{m,n} < p_{i+1})$ de la même manière que $P(W \geqslant p_i)$, à l'aide de (5.27).

Sur $[0, b]^2$, pour $b < \dfrac{m-1}{n-2}$

$$\begin{aligned} D_{m,m;n-2m+1}(x, y) &\leqslant \frac{n!}{[(m-1)!]^2(n-2m)!}b^{2m-2}(1-2b)^{n-2m} \\ &= \frac{m^2}{b^2}\underbrace{\frac{n!}{(m!)^2(n-2m)!}b^{2m}(1-2b)^{n-2m}}_{E_3} \end{aligned} \tag{5.35}$$

Et pour $b < \dfrac{m-1}{n-1}$ $0 < x < b$

$$\beta_{m,n-m+1}(x) < \beta_{m,n-m+1}(b) = \frac{m}{b}\underbrace{\frac{n!}{m!(n-m)!}b^m(1-b)^{n-m}}_{E_4} \tag{5.36}$$

Mais

$$\begin{aligned} \log E_3 &= \log(n!) - 2\log(m!) - \log(n-2m)! \\ &\quad + 2m\log b + (n-2m)\log(1-2b) \\ \log E_4 &= \log(n!) - \log(m!) - \log[(n-m)!] \\ &\quad + m\log b + (n-m)\log(1-b) \end{aligned} \tag{5.37}$$

or d'après la formule de Stirling

$$\log k! = k\log k - k + \frac{1}{2}\log k + \frac{1}{2}\log 2\pi + \frac{\theta_k}{12k} \qquad 0 < \theta_k < 1$$

d'où

$$\log E_3 = -2m \log \frac{m}{n} + 2m \log b - (n-2m) \log \left(1 - \frac{2m}{n}\right)$$

$$+ (n-2m) \log (1-2b) + \frac{1}{2} \log n - \log m - \frac{1}{2} \log (n-2m) - \log 2\pi$$

$$+ \frac{1}{12}\left[\frac{\theta_n}{n}, \frac{\theta_m}{m} - \frac{\theta_{n-2m}}{n-2m}\right] \tag{5.38}$$

de même

$$\log E_4 = -m \log \frac{m}{n} - (n-m) \log \left(1 - \frac{m}{n}\right) + m \log b + (n-m) \log (1-b)$$

$$+ \frac{1}{2} \log n - \frac{1}{2} \log m - \frac{1}{2} \log n - 2m - \frac{1}{2} \log 2\pi$$

$$+ \frac{1}{12}\left[\frac{\theta_n}{n} - \frac{\theta_m}{m} - \frac{\theta_{n-m}}{n-m}\right] \tag{5.39}$$

Posant pour $0 < x < \frac{1}{2}$

$$f(x) = -2x \log \frac{x}{b} - (1 - 2x) \log \frac{1 - 2x}{1 - 2b}$$

$$g(x) = -x \log \frac{x}{b} - (1 - x) \log \frac{1 - x}{1 - b} \tag{5.40}$$

on a

$$\log E_3 \leqslant nf\left(\frac{m}{n}\right) + \frac{1}{2} \log n + \frac{1}{12n}$$

$$\log E_4 \leqslant ng\left(\frac{m}{n}\right) + \frac{1}{2} \log n + \frac{1}{12n} \tag{5.41}$$

Or f a un maximum égal à 0 pour $x = b$, et après f décroît.
g a un maximum égal à 0 pour $x = b$, et après g décroît.

D'après (5.28), si on prend $b = p_{i+1}$, on a pour $\begin{matrix} n \\ m \end{matrix} > \begin{matrix} \mathrm{N} \\ \mathrm{N}' \end{matrix}$

$$\log E_3 \leqslant nf(l) + \frac{1}{2} \log n + \frac{1}{12n} \quad \text{avec} \quad f(l) < 0$$

$$\log E_4 \leqslant ng(l) + \frac{1}{2} \log n + \frac{1}{12n} \quad \text{avec} \quad g(l) < 0 \tag{5.42}$$

Les formules (5.33) à (5.42), jointes à $\int_0^1 \alpha(x)dx = 1$ permettent alors d'aboutir à l'inégalité (5.4) cherchée.

Ceci dit, nous avons au cours de la démonstration, fait les restrictions $2m \leqslant n$, $l \leqslant p_1$ et $p_{(1)} < \frac{1}{2}$. Mais si $2m > n$, $Z_n^{m*} = 0$ ou 1, et donc $EZ_n^{m*} = E(Z_n^{m*})^2$ de même $\alpha\left(\frac{m}{n}\right) = 0$ ou 1.

Alors

$$E\left[Z_n^{m*} - \alpha\left(\frac{m}{n}\right)\right] = \left|E\left(\alpha(W_{m,n}) - \alpha\left(\frac{m}{n}\right)\right)\right|$$

On utilise les mêmes méthodes, on modifie le lemme 7 en changeant m et $n - m + 1$.

On étudie alors les différents cas de figure correspondant à la position respective de

$$p_1,\ p_2,\ l,\ L,\ \frac{1}{2}.$$

On aboutit à un résultat identique à celui obtenu précédemment.

Démonstration du lemme 7. — On cherche une majoration de $P\left(\left|W_{m,n} - \frac{m}{n}\right| > \varepsilon\right)$. Or $EW_{m,n} = \frac{m}{n+1}$, donc si $n > \frac{1}{\varepsilon}$ et $m \leqslant \frac{n}{2}$

$$\left|\frac{m}{n+1} - \frac{m}{n}\right| < \frac{\varepsilon}{2} \quad \text{et donc} \quad P\left(\left|W_{m,n} - \frac{m}{n}\right| > \varepsilon\right) \leqslant P(A_{m,n}) \tag{5.43}$$

où on a posé

$$A_{m,n} = \left\{\left|W_{m,n} - \frac{m}{n+1}\right| > \frac{\varepsilon}{2}\right\} \tag{5.44}$$

Posons encore, avec les notations (5.25)

$$B_{m,n} = \left\{S_m - m > -\frac{m}{2}\right\}$$

$$C_{m,n} = \left\{S'_m + S''_{n-2m+1} - (n - m + 1) > -\frac{n - m + 1}{2}\right\} \tag{5.45}$$

alors

$$P(A_{m,n}) \leqslant P(A_{m,n}B_{m,n}C_{m,n}) + P(\bar{B}_{m,n}) + P(\bar{C}_{m,n}) \tag{5.46}$$

Mais

$$A_{m,n} = \left\{\left|\frac{S_m}{S_m + S'_m + S''_{n-2m+1}} - \frac{m}{n+1}\right| > \frac{\varepsilon}{2}\right\}$$

donc

$$A_{m,n}B_{m,n}C_{m,n} \subset \left\{ \left| S_m\left(1 - \frac{m}{n+1}\right) - (S'_m + S''_{n-2m+1})\frac{m}{n+1} \right| > \frac{(n+1)\varepsilon}{4} \right\} \quad (5.47)$$

d'où

$$P(A_{m,n}B_{m,n}C_{m,n}) \leqslant P\left(|S_m - m| > \frac{m\varepsilon}{4}\right) + P\left(|S'_m + S''_{n-2m+1} - (n-m+1)| > \frac{(n-m+1)\varepsilon}{4}\right) \quad (5.48)$$

car si

$$|S_m - m| \leqslant \frac{m\varepsilon}{4} \quad \text{et} \quad |S'_m + S''_{n-2m+1} - (m-n+1)| \leqslant \frac{n-m+1}{4}\varepsilon$$

alors

$$\left(1 - \frac{m}{n+1}\right)|S_m - m| \leqslant \frac{m\varepsilon}{4}$$

$$\frac{m}{n+1}|S'_m + S''_{n-2m+1} - (m-n+1)| \leqslant \frac{n-m+1}{4}\varepsilon$$

et donc

$$\left| S_m\left(1 - \frac{m}{n+1}\right) - (S'_m + S''_{n-2m+1})\frac{m}{n+1} \right| \leqslant \frac{m\varepsilon}{4} + \frac{n-m+1}{4}\varepsilon = \frac{n+1}{4}$$

Nous avons alors besoin d'un lemme sur les v. a. de lois gamma :

LEMME 8. — Si $0 < \sigma < \mu$ et si ξ est une v. a. gamma de paramètre μ alors

$$P(|\xi - \mu| \geqslant \sigma) \leqslant 2e^{-\frac{\sigma^2}{8\mu}} \quad (5.49)$$

Démonstration du lemme 8. — Si on pose $M(\varepsilon) = Ee^{\varepsilon(\xi-\mu)}$ défini pour $\varepsilon < 1$, on a

$$M(\varepsilon) = \frac{e^{-\mu\varepsilon}}{(1-\varepsilon)^{\mu}} \quad (5.50)$$

si $\varepsilon > 0$, en appliquant l'inégalité de Markov, on a :

$$\forall t > 0 \qquad P\left(\xi - \mu > \frac{1}{\varepsilon}(t + \log M(\varepsilon))\right) \leqslant e^{-t} \quad (5.51)$$

$$P\left(\xi - \mu < -\frac{1}{\varepsilon}(t + \log M(-\varepsilon))\right) \leqslant e^{-t} \quad (5.51)$$

or si $0 < \varepsilon < \frac{1}{2}$ on a

$$\log M(\varepsilon) \leqslant \frac{3\mu\varepsilon^2}{2} \quad (5.52)$$

donc

$$\forall t > 0 \qquad P\left(\xi - \mu > \frac{1}{\varepsilon}\left(t + \frac{3\mu\varepsilon^2}{2}\right)\right) \leqslant e^{-t} \tag{5.53}$$

Posant $t = \dfrac{\sigma^2}{8\mu}$ et $\varepsilon = \dfrac{\sigma}{2\mu}$ on trouve, si $\sigma < \mu$

$$P(\xi - \mu > \sigma) \leqslant e^{-\frac{\sigma^2}{8\mu}} \tag{5.54}$$

si $\varepsilon > 0$ on a

$$0 \geqslant -\log M(-\varepsilon) \geqslant -\frac{\mu\varepsilon^2}{2} \tag{5.55}$$

donc

$$\forall t > 0 \qquad P\left(\xi - \mu \leqslant -\frac{1}{\varepsilon}\left(t + \frac{\mu\varepsilon^2}{2}\right)\right) \leqslant e^{-t} \tag{5.56}$$

Posant $t = \dfrac{\sigma^2}{2\mu}$, $\varepsilon = \dfrac{\sigma}{\mu}$ il vient

$$P(\xi - \mu < -\sigma) \leqslant e^{-\frac{\sigma^2}{2\mu}} \tag{5.57}$$

Remarquons que si $\sigma > \mu$ cette dernière probabilité est nulle. Le lemme 8 est donc démontré en réunissant (5.54) et (5.57).

Pour terminer la démonstration du lemme 7, il suffit alors d'appliquer ce qui vient d'être fait à (5.46) et (5.48). En effet :

$$P(\bar{B}_{m,n}) \leqslant e^{-\frac{m}{8}}, \qquad P(\bar{C}_{m,n}) \leqslant e^{-\frac{n-m+1}{8}} \qquad \text{d'après (5.57)}$$

Mais $m < n - m + 1$, d'où

$$P(\bar{B}_{m,n}) + P(\bar{C}_{m,n}) \leqslant 2e^{-\frac{m}{8}} \tag{5.58}$$

Enfin

$$P\left(|S_m - m| > \frac{m\varepsilon}{4}\right) \leqslant 2e^{-\frac{m\varepsilon^2}{128}} \tag{5.59}$$

$$P\left(|S'_m + S''_{n-2m+1} - (n - m + 1)| > \frac{(n - m + 1)\varepsilon}{4}\right)$$

$$\leqslant 2e^{-\frac{(n-m+1)\varepsilon^2}{128}} \leqslant 2e^{-\frac{m\varepsilon^2}{128}} \tag{5.60}$$

(5.58), (5.59) et (5.60) donnent

$$P(A_{m,n}) \leqslant 6e^{-\frac{m\varepsilon^2}{128}} \tag{5.61}$$

Remarque. — Il ne nous a pas été possible de déduire du théorème 6 des conséquences sur les $X_n^{(k)}$.

2) On peut, avec le théorème 6, estimer la fonction α sur les rationnels et donc, d'une certaine manière, estimer les p_k.

BIBLIOGRAPHIE

[1] K. B. ATHREYA, P. E. NEY, *Branching processes*, Springer-Verlag, 1972.
[2] B. BRAINERD, On the relation between types and tokens in a litterary text. *J. Appl. Prob.*, t. **9**, 1972, p. 507-518.
[3] B. M. HILL, The rank-frequency form of Zipf's law. *J. Amer. Stat. Assoc.*, vol. 69, no. 348, 1974, p. 1017-1026.
[4] Hill-Woodroofe stronger forms of Zipf's law. *J. Amer. Stat. Assoc.*, vol. 70, no. 349, 1975, p. 212-219.
[5] B. M. HILL, M. WOODROOFE, On Zipf's law. *J. of Appl. Prob.*, t. **12**, 1975, p. 425-434.
[6] S. KARLIN, Central limit theorems for certain infinite urn schemes. *J. of Math. and Mech.*, vol. 17, no 4, 1967, p. 373-401.
[7] B. MANDELBROT, On the theory of word frequencies and on related markovian models of discourse. *Symposium an Applied Mathem.*, vol. XII, 1962.
[8] D. N. MCNEIL, Estimating an author's vocabulary. *J. of Amer. Stat. Assoc.*, vol. 68, no. 341, 1973, p. 92-96.
[9] C. J. MODE, *Multitype branching processes.* American Elsevier, New York, 1971.
[10] S. S. WILKS, *Mathematical Statistics.* Wiley, New York, London, 1962.

(Manuscrit reçu le 8 décembre 1977)

* Épreuves non corrigées par l'auteur.