PUBLICATIONS MATHÉMATIQUES ET INFORMATIQUES DE RENNES

MICHELINE PETRUSZEWYCZ

Chaînes de Markov discrètes dans le domaine linguistique : l'article de 1913 de A.A. Markov

Publications des séminaires de mathématiques et informatique de Rennes, 1981, fascicule 2
« Séminaire d'histoire des mathématiques au XXe siècle », , exp. n° 2, p. 1-12

<http://www.numdam.org/item?id=PSMIR_1981___2_A3_0>

CHAINES DE MARKOV DISCRETES DANS LE DOMAINE LINGUISTIQUE : l'article de 1913 de A.A. MARKOV

Micheline PETRUSZEWYCZ *

Je présenterai d'abord un bref résumé biographique. A.A. Markov est né le 2 juin 1856 à Riazan où son père était un petit fonctionnaire des Eaux et Forêts. Il arrive adolescent à St-Petersbourg et y fait des études quelconques sauf en mathématiques où son génie se révélera précocement. Etudiant il suit les cours des mathématiciens russes Zockin, Zolotariev et surtout Čebyšev dont il sera le véritable disciple et, disons-le, le disciple préféré. Ayant soutenu ses thèses successives, il est nommé en 1886 professeur à l'Université et stagiaire à l'Académie des Sciences. A partir du retrait de Čebyšev en 1883 il a pris en charge le cours de calcul des probabilités qu'il assumera même au delà de la retraite jusqu'à sa mort à Petrograd le 20 juillet 1922. Markov est l'un des plus éminents représentants de l'Ecole dite de St Petersbourg ou de Čebyšev car c'est à la pleïade de mathématiciens formés par ce grand maître que l'on donne ce nom. J'ai dit disciple préféré et c'est justice car tout en gardant la grande indépendance de jugement qui le caractérise Markov s'est employé à n'utiliser constamment que les méthodes mathématiques introduites par Čebyšev dans le domaine du calcul des probabilités, en particulier la méthode des fractions continues et la méthode des moments, à condition cependant de rendre l'application de celle-ci plus générale par l'introduction, due à Markov, de la notion de variable aléatoire tronquée.

Je vais aujourd'hui vous présenter quelques textes de Markov : celui où apparaît le concept de chaîne puis deux articles énonçant une propriété des chaînes. Enfin je parlerai d'une façon plus détaillée du seul exemple d'application donné par Markov lui-même.

---

* Centre de Mathématique Sociale, Ecole des Hautes Etudes en Sciences Sociales, Paris

I - LES ARTICLES THEORIQUES

a) L'article de Kazan (1907) appelé ainsi car son noyau initial a paru dans les Isvestyah de l'Université de Kazan. Je me réfère au texte publié dans les Oeuvres choisies et qui porte deux dates : 16 janvier 1907 et 25 mars 1907. Le texte a donc été remanié et ce remaniement semble postérieur à la lecture devant l'Académie du prochain texte dont je parlerai ensuite, lecture faite le 14 février 1907. Dans cette communication parue au Bulletin Impérial ne figure pas le mot chaîne (цепь = cepь) alors que le concept est nommé et défini à la fin de l'article de Kazan. Le mot a donc été adopté par Markov entre le 14 février et le 25 mars. Markov part, comme très souvent, d'un résultat de Čebyšev qui avait étendu la loi des grands nombres au cas de variables aléatoires indépendantes ayant des variances uniformément bornées, $D(x) \leq C$. Ce ne sont pas les deux premières parties de l'article qui nous intéressent mais seulement la troisième. Au lieu d'étudier le nombre d'apparitions d'un certain événement, dit-il, on peut examiner la somme des grandeurs *liées en chaine*. Et voici sa définition :

> "Quand une grandeur reçoit une valeur déterminée celles qui la suivent "deviennent indépendantes de celles qui la précédaient. Nous poserons "$x_1, x_2, \ldots, x_k, x_{k+1}, \ldots$ série illimitée de grandeurs liées de telle "façon que x_{k+1}, pour tout k, ne dépend pas de $x_1, x_2, \ldots, x_{k-1}$ quand "on connaît la valeur de x_k".

Markov va considérer les espérances mathématiques de ces grandeurs et en particulier l'espérance mathématique du carré :
$(x_1 - a_1$ (a_1 est l'esp. math. de x_1) $+ x_2 - a_2 + \ldots + x_n - a_n)^2$ puis notant z_k la différence $x_k - a_k$ et reprenant le cheminement de la première partie de l'article il établit que :

$$e.m\ z_k(z_1 + z_2 + \ldots + z_{k-1}) < D\,(H + H^2 + \ldots + H^{k-1})$$

et $e.m\ (x_1 - a_1 + x_2 - a_2 + \ldots + x_n - a_n)^2 < G\,n$,

D, G et H étant des nombres constants.

D'autre part en effectuant $e.m\ (x_1 - a_1 + x_2 - a_2 + \ldots + x_n - a_n)^2$

où $a = \lim a_{k+i}$ avec $i = \infty$ $\quad - \quad e.m\ (x_1 + x_2 + \ldots + x_n - n\,a)^2$

il obtient l'expression : $(a_1 - a + a_2 - a + \ldots + a_n - a)^2$

qui garde une valeur finie pour un accroissement sans limite de n.

Alors sous cette condition : $e.m\ \left(\frac{x_1 + x_2 + \ldots x_n}{n} - a\right)^2 \to 0$

Je reprends alors les termes mêmes de Markov :

"Mais de là aussitôt résulte la loi des grands nombres ; pour aussi petits
"que soient les nombres positifs ε et η, la probabilité de la réalisation
"des inégalités

$$-\varepsilon < \frac{x_1 + x_2 + \ldots + x_n}{n} - a < +\varepsilon$$

"sera plus grande que $1 - \eta$, pour toutes les valeurs suffisamment grandes
"de n.
"Donc l'indépendance (nezabucumbocmь = nezavisimost') des grandeurs ne
"constitue pas une condition nécessaire pour l'application de la loi des
"grands nombres".

Je reviens sur la définition de la *chaîne* : dans cet article où le concept est défini et nommé pour la première fois Markov dit textuellement qu'il va examiner
la somme des grandeurs *liées* [= cbezaвиmыx = svjazannyx, ce terme ayant en russe le même champ sémantique que le français lier mais ayant aussi le sens de tricoter. Il s'agit vraiment d'attacher par un noeud] en chaîne ;

"ainsi quand l'une reçoit une valeur déterminée, celles qui la suivent
"deviennent indépendantes de celles qui la précédaient. Soit
"$x_1, x_2, \ldots, x_k, x_{k+1}$, série illimitée de grandeurs liées de sorte que
"x_{k+1}, pour tout k ne dépend pas de $x_1, x_2, \ldots, x_{k-1}$ quand on connaît
"la valeur de x_k".

Devant un public dont une partie s'intéresse à la pédagogie des mathématiques et pour rendre hommage au grand pédagogue que fut Markov je vais en parallèle vous énoncer la définition qu'il a donnée dans son manuel, en me référant à l'édition posthume de 1924.

Il parle là aussi d'une série illimitée d'épreuves successives numérotées : 1, 2, 3, ..., k, $k + 1$ et il poursuit :

"Nos épreuves seront liées relativement à un certain événement E en chaîne
"simple qui se divise en deux parties : une partie des épreuves seulement
"entraîne la réalisation de E, l'autre pas ; les épreuves suivantes après
"celles-ci deviennent indépendantes de celles qui les précèdent... Le
"nombre p_1 note la probabilité de l'événement E pour la $k + 1$ ième épreuve
"s'il est donné que E se réalise au cours de la k ième épreuve, les ré-
"sultats des épreuves les suivant restant indéterminés. Le nombre p_2 note
"aussi la probabilité de l'événement E pour la $k + 1$ ième épreuve alors
"que les résultats des épreuves numérotées $k + 1$, $k + 2$, $k + 3$ restent

"indéterminés mais seulement dans les cas où la k-ième épreuve a amené non "pas la réalisation de l'événement E mais au contraire celle de l'événement "F. Conformément à l'explication donnée ci-dessus, en rapport avec le lien "des épreuves en chaîne simple, les probabilités indiquées de l'événement "E pour la k + 1 ième épreuve sont formées entièrement indépendamment des "résultats des k - 1 premières épreuves... Pour donner à notre conclusion "une entière précision il convient d'introduire encore le nombre p' repré- "sentant la probabilité de l'événement E pour la première épreuve... toute- "fois ce dernier nombre dans nos conclusions finales ne joue aucun rôle.

J'espère que la différence de ton reste perceptible malgré l'écran de la traduction.

b) Article de 1907. Recherches sur un cas remarquable d'épreuves dépendantes.

Dans cet article Markov revient à l'examen du nombre d'apparitions d'un événement E au cours d'un nombre connu n d'épreuves successives et il précise les conditions de la chaîne et la probabilité initiale.

1. - la probabilité de l'événement E a une seule et même valeur p alors que les résultats des épreuves restent indéterminés. (Cette condition a dans certains autres textes été modifiée, à la suite de la critique de Ljapounov, mais je m'y tiendrai car c'est elle qui est posée dans l'article d'application sur la succession des voyelles et des consonnes).

2. - la probabilité de l'événement E présente une deuxième valeur dénotée p_1 si les résultats des épreuves suivantes demeurent indéterminés mais que l'épreuve immédiatement précédente a amené la réalisation de l'événement E.

3. - enfin la probabilité de l'événement E a chaque épreuve présente une troisième valeur p_2 si les résultats des épreuves suivantes restent indéterminés mais que l'épreuve immédiatement précédente n'a pas amené la réalisation de l'événement E.

Markov introduit alors l'événement F complémentaire et la symétrie en p et q de ses formules. Des trois nombres p, p_1 et p_2, dit-il, deux seulement peuvent être pris arbitrairement car ils sont liés par la relation $p = pp_1 + qp_2$ déduite de la considération des épreuves voisines de l'épreuve considérée. Puis il introduit le paramètre de *chaîne simple* $\delta = p_1 - p_2$ et en déduit ses quatre formules fondamentales :

$$p_1 = p + \delta q \qquad p_2 = p - \delta p$$
$$+ \quad q_1 = q - \delta q \qquad q_2 = q + \delta p$$
$$1$$

Il va considérer la fonction génératrice déterminant la probabilité que l'événement E, lors de n épreuves, apparaisse un nombre déterminé de fois m au cours

des $\underline{k}$ premières épreuves. Après introduction d'un nombre arbitraire ξ il écrit :

$$\omega_k = \sum_m P_{m,k}\, \xi^m \qquad o \leq m \leq k$$

puis en introduisant à nouveau un nombre arbitraire t

$$\Omega(\xi,t) = \omega_o + \omega_1 t + \omega_2 t^2 + \ldots$$

$$= \frac{1 - \delta(q\,\xi + p)t}{1 - \{p\,\xi + q + \delta(q\,\xi + p)t + \delta\,\xi\,t^2\}}$$

expression qui lui permet d'étudier les moments factoriels. Mais en fait, il se ramène à l'étude de certaines fractions développées en séries convergentes et ordonnées par degrés croissants de δ et indépendantes de n. Puis il réussit à exprimer l'espérance mathématique de $(m - pn)^k$ sous la forme du polynôme :

$$R_k^{(k)}\, n^k + R_{(k-1)}^{(k)}\, n^{(k-1)} + \ldots + R_i^k\, n^i + \ldots R_o^k$$

dont les coefficients sont des fonctions entières de $\underline{p}$, $\underline{q}$ et $\underline{\delta}$.
Il poursuit par l'étude de la parité de $\underline{k}$ au moyen des deux nombres pair et impair 2ℓ et $2\ell - 1$ et en effectuant un passage à la limite il écrit :

$$R_\ell^{2\ell} = (a_o + a_1 + a_2 + \ldots + a_\ell)\, p^\ell q^\ell$$

invariable quand on permute $\underline{p}$ et $\underline{q}$. Il montre alors que $a_0\ a_1 \ldots$ sont indépendants de $\underline{p}$ et $\underline{q}$ et que leur somme a une limite quand $\underline{n}$ augmente.

Il peut alors énoncer le théorème limite :

"La probabilité des inégalités

$$np + t_1 \sqrt{2pq\,\frac{1+\delta}{1-\delta}\,n} < m < np + t_2 \sqrt{2pq\,\frac{1+\delta}{1-\delta}\,n}$$

"converge à la limite vers :

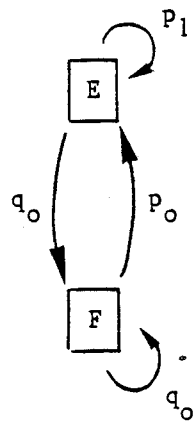
$$\frac{1}{\sqrt{\pi}} \int_{t_1}^{t_2} e^{-t^2}\, dt$$

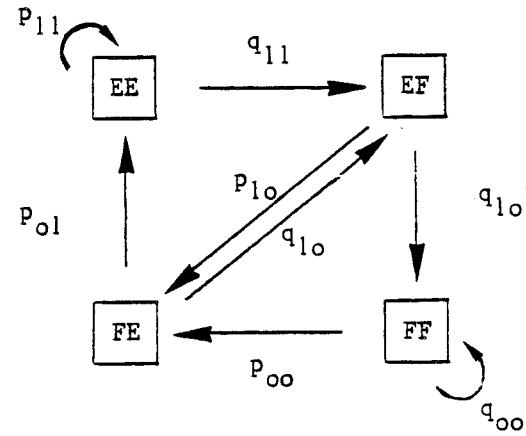
"quand $\underline{n}$ croît sans limite, p, q, δ, t_1 et t_2 restant constants."

Cet article est accessible en français car Markov en a donné une réédition légèrement modifiée dans Acta Mathematica, 1910, 33, 87-104.

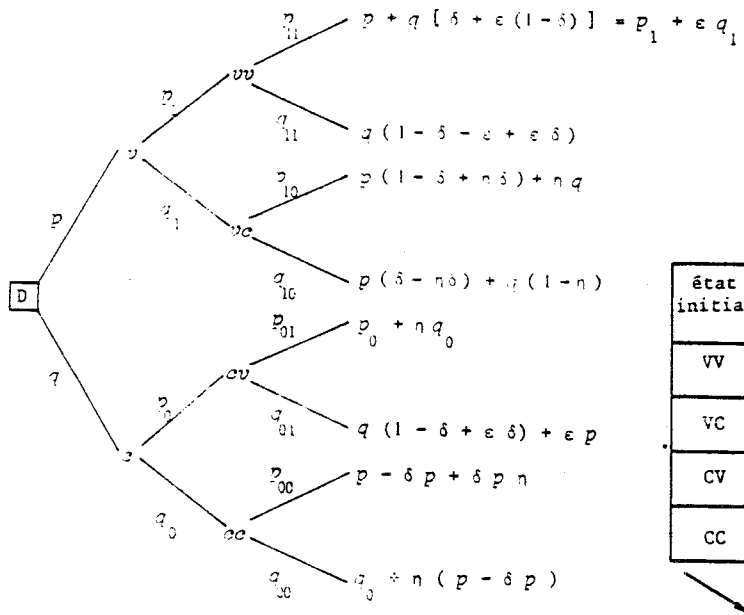
c) Article de 1911. Sur un cas d'épreuves liées en chaîne multiple.

On passe de la chaîne simple à la chaîne multiple (простой: prostoj / сложной = složnoj) quand au lieu de considérer à l'étape k le résultat de l'étape k-1 seulement on regarde aussi le résultat de l'étape k-2. Il est à mon sens plus éclairant d'en donner des *représentations* ce que ne fait jamais Markov.





On peut aussi représenter la chaîne par l'arbre des mots ou états du système. On peut aussi donner la matrice des probabilités de passage qui est une matrice stochastique les $p_{ij} \geq 0$ et les sommes en ligne sont toutes égales à l'unité.



L'arbre des états successifs de la chaîne complexe.

état initial					
VV	p_{11}	q_{11}	0	0	
VC	0	0	p_{10}	q_{10}	
CV	p_{01}	q_{01}	0	0	
CC	0	0	p_{00}	q_{00}	
	VV	VC	CV	CC	état final

Matrice des probabilités de passage (1)

(1) Nous avons remplacé ici les événements désignés dans le texte de Markov par les lettres E et F par les initiales des événements du phénomène étudié : apparition dans un texte d'un graphème Vocalique et d'un graphème Consonantique.

Aux équations de base de la chaîne simple il convient alors d'ajouter celles liant les nouvelles probabilités de transition :

$$p_1 = p_1\, p_{11} + q_1\, p_{01} \quad \text{et} \quad p_o = p_o\, p_{1o} + q_o\, p_{oo}$$

Markov introduit alors les paramètres de chaîne multiple

$$\varepsilon = \frac{p_{11} - p_1}{q_1} \qquad \text{et} \qquad \eta = \frac{q_{oo} - q_o}{p_o}$$

et donne une expression qui est assez complexe de la dispersion. Après diverses considérations sur les cas possbiles concernant les paramètres il montre que dans le cas de chaîne multiple et si la condition $\varepsilon = -\delta$ se réalise alors l'écart probable du nombre d'apparition de $\underline{m}$ égal à $\sqrt{2\ pq\ \frac{1+\delta}{1-\delta}\ n}$ dans le cas de chaîne simple est égal à $\sqrt{2\ pqn}$ c'est-à-dire celui d'une distribution normale (le 2 vient de l'expression analytique utilisée alors $^1/_{\sqrt{\pi}} \int e^{-x^2}\, dx$ si on fait $x = \frac{u}{\sqrt{2}}$ on a aujourd'hui $^1/_{\sqrt{2\pi}} \int e^{-1/2u^2}\, du$).

II - L'ARTICLE DE 1913 : un exemple d'étude statistique sur le texte d'Evgenij Onegin illustrant la liaison des épreuves en chaîne.

Lors de la date de première parution de ce texte le modèle mathématique de chaîne était déjà raffiné mais il est certain, puisqu'il l'a dit expressement, que Markov a voulu donner un exemple *simple*. Il n'a pas donné, à ma connaissance, d'explication sur l'extraordinaire choix qu'il a fait : la succession des voyelles et des consonnes dans une langue naturelle. Qu'il y ait pensé en 1913 fait problème ; qu'il ait "récidivé" en 1922 quand il l'insère, avec un deuxième exemple dans son manuel constitue un autre problème. En effet, à cette date, à mon sens, il ne pouvait plus ignorer quelques autres des innombrables domaines d'application de sa théorie.

Markov considère les 20.000 premières lettres du roman en vers de Puškin Evgenij Onegin. Cette suite fournit 20.000, exactement 19.999 épreuves liées chacune voyant apparaître soit une voyelle, soit une consonne. Il admet l'existence d'une probabilité inconnue constante $\underline{p}$ de l'apparition de l'événement qu'il privilégie, l'apparition de V, et va en chercher une estimation.

De même il estime au moyen du nombre d'apparition des événements correspondants les probabilités p_1 = V apparaît après V : $\frac{\text{nbre de VV}}{\text{nbre de V}} \simeq$; p_o = V apparaît après C : $\frac{\text{nbre de CV}}{\text{nbre de C}}$, etc....

Pour cela il fait le relevé des lettres consécutives selon le schéma distribué de tableau élémentaire et il va montrer après les manipulations dont je vais vous donner le détail que la variable = nombre de V dans 100 lettres est

TABLEAU I. Les deux premiers tableaux élémentaires du texte d'Evgenii Onegin.

м	о	й	д	ы	п	я	с	а	м	5
ы	х	ч	е	с	т	н	ы	х	п	3
р	а	в	и	л	к	о	г	д	а	4
н	е	э	ш	у	т	к	у	з	а	4
к	е	м	о	г	о	н	у	в	а	5
ж	а	т	с	е	б	я	з	а	с	4
т	а	в	и	л	и	л	у	ч	ш	4
е	в	ы	д	у	м	а	т	н	е	5
м	о	г	е	г	о	п	р	и	м	4
е	р	д	р	у	г	и	м	н	а	4
3	7	2	5	5	3	5	4	3	5	42

(1,6) : 3 + 3 = 6
(2,7) : 7 + 5 = 12
(3,8) : 2 + 4 = 6
(4,9) : 5 + 3 = 8
(5,10): 5 + 5 = 10
42

у	к	а	н	о	б	о	ж	е	м	5
о	й	к	а	к	а	я	с	к	у	6
к	а	с	б	о	л	н	ы	м	с	3
и	д	е	т	и	д	е	н	и	н	5
о	ч	н	е	о	т	х	о	д	я	5
н	и	ш	а	г	у	п	р	о	ч	4
к	а	к	о	е	н	и	з	к	о	5
е	к	о	в	а	р	с	т	в	о	4
п	о	л	у	ж	и	в	о	г	о	5
з	а	б	а	в	л	я	т	е	м	4
5	6	3	6	6	3	5	3	4	5	46

(1,6) : 5 + 3 = 8
(2,7) : 6 + 5 = 11
(3,8) : 3 + 3 = 6
(4,9) : 6 + 4 = 10
(5,10): 6 + 5 = 11
46

Les lettres soulignées sont des "voyelles".

TABLEAU II. Les quarante tableaux de Markov constitués sur les 20.000 premières lettres du texte (reproduit d'après les pp. 563-569 de [4] (les vingt premiers).

8 8 11 11 13 49	16 11 9 8 7 51	14 12 7 3 6 42	5 11 10 6 10 42
12 11 7 7 5 42	4 8 9 11 10 42	5 5 11 9 11 41	12 8 8 11 7 46
6 6 6 7 13 38	9 9 5 7 10 44	8 10 5 10 7 41	7 7 12 10 9 45
8 10 11 5 4 42	12 9 6 10 7 44	11 11 8 3 10 43	8 12 7 9 9 45
10 11 5 10 8 44	3 8 10 8 9 38	4 4 11 14 8 41	12 8 10 9 8 47
42 46 40 44 43 15	44 45 43 44 43 19	42 42 43 39 42 8	44 46 47 45 43 25

10 6 6 6 7 35	8 7 8 7 10 40	11 11 8 7 7 44	11 10 10 12 6 49
9 12 15 6 9 51	10 9 9 8 8 44	9 8 10 11 11 47	4 4 9 7 9 33
9 3 6 10 9 37	8 9 8 8 8 41	12 9 9 5 6 41	11 13 6 3 10 49
9 11 8 5 6 39	10 6 13 6 12 47	10 8 6 11 11 46	6 7 11 8 6 38
9 10 10 10 9 48	8 12 5 13 6 44	7 6 8 9 8 38	8 6 10 7 12 43
46 42 45 37 40 10	44 43 43 42 44 16	49 40 41 43 43 16	40 40 46 43 43 12

12 9 8 10 10 49	8 9 9 5 8 39	7 7 7 7 9 37	12 7 7 6 8 40
3 10 12 9 10 44	7 9 9 11 7 43	9 13 6 8 4 40	6 8 7 10 8 39
11 11 6 11 10 49	10 6 6 9 9 40	9 7 11 12 14 53	9 10 10 8 7 44
10 8 11 6 7 42	7 8 15 6 9 45	7 11 8 9 7 42	9 5 6 7 7 34
6 8 7 9 6 36	11 7 6 11 10 45	8 10 10 11 9 48	7 11 9 13 7 47
42 46 44 45 43 20	43 39 45 42 43 12	40 48 42 47 43 20	43 41 39 44 37 4

7 4 11 5 7 34	5 5 7 5 9 31	8 6 5 14 11 44	10 9 13 6 12 50
11 14 9 11 9 54	12 6 10 10 8 46	8 12 10 7 4 41	8 8 8 9 5 38
7 6 9 8 9 39	8 14 11 11 10 54	8 10 9 8 14 49	10 10 8 9 10 47
10 9 8 10 5 42	4 3 9 5 9 30	9 5 9 9 6 38	7 9 10 7 10 43
11 10 8 9 11 49	13 14 9 11 7 54	8 13 11 5 10 47	9 8 3 11 7 38
46 43 45 43 41 18	42 42 46 42 43 15	41 46 44 43 45 19	44 44 42 42 44 16

4 11 10 12 5 42	5 11 10 6 5 37	4 4 10 11 5 34	13 11 13 10 10 57
14 9 8 7 14 52	8 9 8 10 10 45	6 12 9 8 10 45	7 10 9 6 2 34
4 8 9 8 4 33	8 8 6 9 9 40	13 4 10 8 6 41	8 8 7 8 12 43
8 14 11 12 6 51	10 6 9 7 6 38	7 10 7 12 11 47	9 11 9 10 6 45
11 6 7 4 14 42	11 9 8 10 12 50	9 13 8 1 8 39	6 3 7 9 9 34
41 48 45 43 43 20	42 43 41 42 42 10	39 43 44 40 40 6	43 43 45 43 39 13

(à suivre)

selon comme il la considère une variable indépendante ou plus ou moins liée.

Il constitue pour cela quarante tableaux groupant cinq centaines de lettres ; dans ceux-ci chaque colonne représente une centaine, le premier chiffre représentant la somme des V contenues dans les colonnes 1 et 6 du premier tableau élémentaire ; le deuxième chiffre représente la somme des V des colonnes 2 et 7 et ainsi de suite.

Ceci fait il dresse la statistique des totaux en colonnes ou marges horizontales des 40 tableaux et il calcule la moyenne qui lui permet d'estimer $\hat{p} = 0,4319 \simeq 432$. Il calcule ensuite la somme des carrés des écarts de ces 200 nombres à 43,2 et obtient 1022,8 qui divisé par 200 → variance $\simeq 5,114$. Il présente ensuite deux argumentations pour montrer qu'alors le nombre de V par 100 lettres est une variable indépendante dont on peut rendre compte par une loi normale de mêmes paramètres.

Première argumentation : l'écart possible ω. Il est défini par le fait que la variable a une probabilité 1/2 de se situer entre $F(2+\omega)$ et $F(2-\omega)$ ω ayant pour la loi normale la valeur $0,6745\sigma$, $\sigma = \sqrt{5,11}$ → $\omega \approx 1,5$ et effectivement entre 43,2 + 1,5 = 44,7 et 43,2 - 1,5 = 41,7 on trouve la moitié à peu près des observations

31 fois 42 , 43 fois 43 et 29 fois 44 = 103 valeurs sur 200.

Deuxième argumentation est basée sur la considération de la valeur de la somme des carrés des écarts à la moyenne. Plus la liaison entre les épreuves est grande plus cette somme est petite ; elle est même nulle quand la liaison est parfaitement déterminée. Si on considère des séquences de n épreuves le nombre d'apparitions de V varie selon le longueur de la séquence et on peut calculer le nombre moyen d'apparitions, puis les écarts à ce nombre moyen. Si le processus fait apparaître systématiquement le même événement le nombre de réalisations de V dépend de la longueur de la séquence : on observe toujours C ou toujours V ou une parfaite alternance VC VC VC ; si on considère plusieurs séries d'épreuves le nombre de V reste constant il n'y a pas d'écarts au nombre moyen. Markov groupe ensuite par deux, puis quatre et enfin cinq les valeurs marginales horizontales. Il calcule les moyennes de ces groupements respectifs

86,4 172,8 216

et les sommes des carrés des écarts

827,6 975,2 1004

et dit-il : "ces trois nombres ne se distinguent pas beaucoup de 1022,8" relatif aux 200 nombres élémentaires qui sont les résultats d'épreuves indépendantes et donc qu'on les groupe par paquets de 2, 4 ou 5 ne fait pas grande différence.

Markov dénombre alors dans le texte, et donc dans le cas de chaîne les configurations relevant de la chaîne simple c'est-à-dire les configurations de 2 lettres. On peut le faire rapidement en dressant la table de contingence. Je ne traiterai pas aujourd'hui du maniement de ces tableaux sous-tendus par des propriétés combinatoires*. Il peut alors calculer les valeurs estimées

1104 VV	7534 VC	8638
7534 CV	3827 CC	11361
8638	11361	19999

$$p_1 = \frac{1104}{8638} \simeq 0{,}128$$

$$p_2 = \frac{CV}{C} \simeq 0{,}663$$

Il en tire $\delta = p_1 - p_2 \simeq -0{,}535$ et pour M = (coefficient de dispersion) $\frac{1+\delta}{1-\delta} \simeq 0{,}3$.

Et il commente : "il est vrai que ce nombre n'est pas égal à

$$\frac{\text{variance observée}}{\text{var.théor.loi de même } \bar{p} \text{ et } q} = \frac{5{,}114}{0{,}432 \times 0{,}568} = \frac{5{,}114}{24{,}537} \simeq 0{,}208$$

, mais il en est plus proche que de 1 qui lui correspondrait si les épreuves étaient indépendantes.

Il adopte ensuite le modèle de chaîne multiple et dénombre les configurations de trois lettres ce qui lui permet d'estimer les probabilités de transition. Il obtient

$$p_{11} = \frac{VVV}{VV} \simeq 0{,}104 \qquad q_{ooo} = \frac{CCC}{CV} \simeq 0{,}132$$

puis de calculer les paramètres de chaîne multiple $\varepsilon \simeq 0{,}027$ $\eta = -0{,}309$. Il peut alors calculer le coefficient de dispersion C = 0,195 qui approche au mieux la valeur 0,208 ce qui montre que la chaîne d'ordre 2 (2 transitions, 3 lettres) constitue une meilleure représentation de l'effet de chaîne que le modèle de chaîne simple.

Il revient alors aux quarante tableaux de cinq centaines et considère les marges verticales. Il dresse la statistique de ces totaux et dit-il

"chacun de ces totaux est la somme de *grandeurs presqu'indépendantes*".
"En effet les cinq termes de la 1ère ligne sont obtenus par addition des "colonnes 1 et 6 des 5 premiers tableaux élémentaires ; les cinq termes de "la 2e ligne otenus par l'addition des colonnes 2 et 7 des 5 premiers tableaux "élémentaires etc."Mais *ces termes sont liés par cinq*",

en effet dans chaque groupe les lettres de la première centaine (1ère ligne) sont contigües à celle de la 2e centaine (2e ligne) ; les lettres de la 2e centaine sont contigües à celles de la 1ère et de la 3e centaines et ainsi de suite.

* GUILBAUD G.Th., Note sur les comptabilités markoviennes, Math. Sci. hum n°66, 1979, p.99-112.

S'il considère les 200 résultats, leur moyenne est évidemment égale à celle de la distribution en colonnes : 43,19 mais quand il calcule la somme des carrés des écarts celles-ci est de 5 788,2 et la variance $\frac{5\ 788,2}{200} \simeq 28,944$.

Si on fait $\frac{\text{variance observée}}{\text{var. théor. de même paramètres}} = \frac{5\ 788,2/200}{0,432 \times 0,568} \simeq 1,18$ ce qui traduit le caractère non lié des totaux marginaux. Maintenant on somme à nouveau par 2, 4 ou 5 ce faisant on restitue une *certaine contigüité* et donc l'effet de chaîne : et que constate-t-on ? les moyennes sont :

86,4 172,8 216

alors que les sommes des carrés des écarts sont :

3 551,6 3 089,2 1 004, et 1 004 est presque 6 fois plus petit que 5 788,2, disparité qui traduit le caractère lié des épreuves quand elles sont sommées c'est-à-dire quand on a restitué la chaîne.

Je viens de raconter la naissance et les tout premiers pas du premier modèle de probabilités en chaîne tels que Markov les a écrits dans les années 1907-1913. Si la théorie a connu le développement extraordinaire que l'on sait, si les applications ont touché de nombreux domaines il est très remarquable que le seul article d'application écrit par Markov lui-même n'ait pas suscité d'autres recherches dans le domaine littéraire. Je pense que c'est à tort et j'ai montré qu'on pouvait utiliser descriptivement les valeurs des paramètres de chaîne de Markov pour l'étude de textes littéraires mais ce n'est pas le lieu d'en parler ici.

BIBLIOGRAPHIE : OEUVRES DE MARKOV.

- Пример статистического исследования над текстом "Евгения Онегина" иллюстрирующий связь испитаний в цепь.

  Primer statisticeskogo issledovanija nad tekstom "Evgenija Onegina" illjustrirujuscii svjaz' ispytanii v cep'.

  Un exemple de recherche statistique sur le texte d' "Eugène Onéguine" illustrant la liaison des épreuves en chaîne, Bull. Acad. Imp. Sc., 6e série, t.7, (1913), 153-162.

- Исследование замечательного случая зависимых испытаний.

  Issledovanie zamečatel'nogo slučaja zavisimyh ispytanij.

  Recherches sur un cas remarquable d'épreuves dépendantes, Bull. Acad. Imp. Sc., 6e série, t.1, (1907), 61-80.

- Об одном случае испитаний, связанных в сложную цепь.

  Ob odnom slučae ispytanij, svjazannyh v složnuju cep'.

  Sur un cas d'épreuves liées en chaîne multiple. Bull. Acad. Imp. Sc., 6e série, t.5, (1911), 171-186.

- Исчисление Вероятностей, переработанное автором, т.е. посмертное издание. С портретом автора и биографическим очерком проф. А.С. Безиковича, Москва, Государственное издательство, 1924.

  Isčislenie Verojatnostei, Pererabotannoe avtorom, 4-e posmertnoe izdanie, s portretom avtora i biografičeskim očerkom Prof. A.S. Bezikoviča, Gocudarstvennoe Izdatel'stvo, Moskva, 1924.

  Calcul des Probabilités, revu par l'auteur, 4e édition posthume, avec un portrait de l'auteur et un aperçu biographique par le Professeur A.S. Bezikovič, Moscou, Editions de l'Etat, 1924.

- Избранные Труды, Теория чисел, Теория вероятостей, Редакция профессора Ю.В. Линника. Комментарии Н.В.Линника, Н.Л.Сапогова, О.В.Сарматова и В.Н. Тимофеева. Издательство Академии Наук, СССР, 1951.

  Izbrannye Trudy, Teorija čisel,Teorija verojatnostei, Redakcija professora Ju. V. Linnika, Kommentarii Ju. V. Linnika, N.L. Sapoga , O.V. Sarmatova i V.N. Timofeeva, Izdatel'stbo Akademii Nauk SSSR, 1952.

  Oeuvres choisies, Théorie des nombres, Théorie des Probabilités, Rédacteur : Professeur Ju.V. Linnik, Commentaires de : Ju.V. Linnik, N.L. Sapogob, O.V. Sarmatov et V.N. Timofeev, Editions de l'Académie des Sciences, URSS, 1952.