REVUE DE STATISTIQUE APPLIQUÉE

J. R. BARRA

Contrôle statistique des suites de nombres « au hasard »

Revue de statistique appliquée, tome 19, n° 3 (1971), p. 19-26

<http://www.numdam.org/item?id=RSA_1971__19_3_19_0>

CONTROLE STATISTIQUE DES SUITES DE NOMBRES « AU HASARD »

J. R. BARRA
Faculté des Sciences de Grenoble

En analyse Numérique, en Simulation, en Statistique, etc..., on a parfois besoin de "nombres au hasard" ; il est toujours nécessaire de tester statistiquement les suites que l'on utilise. On a déjà indiqué dans [1] un test simple et commode d'emploi, mais il est facile de comprendre que pour une analyse poussée il faut utiliser un ensemble de tests et les combiner. On étudie ci-après quelques problèmes apparaissant quand on veut construire un tel système de tests en s'appuyant sur les fréquences d'apparition d'un certain nombre de "figures".

On désigne par $\mathbb{N}$ l'ensemble des entiers positifs non nuls et par card H le nombre d'élément de l'ensemble fini H.

I - TESTS DE FREQUENCES

On suppose que tous les éléments de la suite dont on étudie le caractère aléatoire appartiennent à un même ensemble fini E ; dans la pratique E sera le plus souvent un ensemble de lettres, de chiffres, de symboles "machine"...

DEFINITION 1 - "Soit E un ensemble fini, pour tout entier positif r, on appelle figure d'ordre r une partie non vide Φ de E^r telle qu'il n'existe pas de partie Ψ de E^{r-1} telle que :

$$\Phi = E \times \Psi \quad \text{ou} \quad \Phi = \Psi \times E$$

en particulier une figure d'ordre 1 est une partie de E distincte de E".

DEFINITION 2 - "On appelle n-fréquence de la figure Φ d'ordre r, dans la suite $s \in E^{\mathbb{N}}$, l'entier $\nu_{\Phi}^{n}(s)$ défini par :

$$\nu_{\Phi}^{n}(s) = \sum_{i=r}^{i=r+n} 1_{\Phi}^{i}(s)$$

où $1_{\Phi}^{i}(s)$ est l'indicatrice de la partie de $E^{\mathbb{N}}$ égale à $E^{i-r} \times \Phi \times E^{\mathbb{N}}$".

En termes concrets, $\nu_\Phi^n(s)$ est tout simplement le nombre de fois ou on a observé la figure Φ dans une partie de longueur n de la suite s, sachant que l'on a commencé le comptage dès que possible c'est-à-dire au rang r et que l'on a effectué en tout n comptages consécutifs.

L'application $s \to \nu_\Phi^n(s)$ est une statistique ([4]) sur la structure statistique $[E, \mathcal{A}, \mathcal{L}]$ où $\mathcal{L}$ est la famille des lois de probabilité sur E^N muni de sa tribu $\mathcal{A}$ usuelle ; en particulier soit U la loi sur E^N définie par l'indépendance et l'équirépartition des composantes de s, on a :

$$E_U(\nu_\Phi^n) = n \frac{\text{card}\,\Phi}{(\text{card}\ E)^r}$$

Nous considérons dans la suite que contrôler le caractère aléatoire d'une suite s c'est tester l'hypothèse U ; et on étudie ci-après des tests construits à partir des fréquences d'un certain nombre de figures.

On désignera par $\rho(X, Y)$ la covariance $E(XY) - E(X)\,E(Y)$ de deux variables aléatoires X et Y.

DEFINITION 3 - "Soient Φ et Φ' deux figures respectivement d'ordre r et r', on appelle covariance de ces deux figures, le réel $\lambda(\Phi, \Phi')$ défini par :

$$\lambda(\Phi, \Phi') = \sum_{j=1}^{r+r'-1} \rho\left(1_\Phi^{r+r'-1},\ 1_\Phi^{r'+j-1}\right)$$

quand E^N est muni de la loi U".

DEFINITION 4 - "Soit $\mathcal{F}$ une famille de figures, on appelle matrice de covariance de $\mathcal{F}$ la matrice $\Lambda_\mathcal{F}$ de terme général $\lambda(\Phi, \Phi')$, $(\Phi, \Phi' \in \mathcal{F})$; si $\Lambda_\mathcal{F}$ est non singulière, en désignant par $\mu(\Phi, \Phi')$ le terme général de son inverse, on appelle indice d'aléa défini par $\mathcal{F}$ toute statistique $I_\mathcal{F}^n(s)$ définie par :

$$I_\mathcal{F}(s) = \frac{1}{n} \sum_{\Phi, \Phi' \in \mathcal{F}} \mu(\Phi, \Phi').\ [\nu_\Phi^n(s) - E_U(\nu_\Phi^n)]\ [\nu_{\Phi'}^n(s) - E_U(\nu_{\Phi'}^n)]$$

où n est un entier positif".

THEOREME - "Soit $\mathcal{F}$ une famille de k figures, de matrice de covariance non singulière ; si E^N est muni de la loi U, la loi limite de $I_\mathcal{F}^n$ quand n tend vers l'infini est la loi du Khi-deux à k degrés de liberté".

Il est clair que ce théorème est une conséquence du lemme suivant :

LEMME - "Dans les conditions du théorème, la loi limite quand $n \longrightarrow \infty$ du système des variables aléatoires :

$$\Delta_\Phi^n = \frac{\nu_\Phi^n - E_U(\nu_\Phi^n)}{\sqrt{n}},\ \Phi \in \mathcal{F}$$

est la loi $N(0, \Lambda_\mathcal{F})$".

En effet, soit r l'ordre maximum des figures de la famille $\mathcal{F}$; les composantes $x_1, \ldots, x_n, \ldots$ de S sont des variables aléatoires indépendantes uniformément réparties sur E et donc les vecteurs aléatoires :

$$Z_k = (x_{k-r+1}, \ldots, x_{k-1}, x_k) \quad k = r, \ldots$$

évoluent en chaîne de Markov finie et régulière. D'après le théorème Central Limite ([2] p. 401), les n-fréquences des figures définies par les points de E^r ont une loi limite gaussienne et il en est de même du système des $\nu_\Phi^n (\Phi \in \mathcal{F})$ qui en sont une combinaison linéaire.

Enfin, un calcul élémentaire montre que :

$$\frac{1}{n} \rho(\nu_\Phi^n, \nu_{\Phi'}^n) \xrightarrow[n \to \infty]{} \lambda \ (\Phi, \Phi')$$

et la démonstration est achevée.

L'utilisation pratique de ce résultat est simple :

a) On fait choix d'une famille de figures $\mathcal{F}$.

b) On calcule le cardinal de chaque figure et leurs covariances deux à deux : ce sont des calculs de dénombrement et la formule de la définition 3 se prête bien à un calcul sur ordinateur si c'est nécessaire.

c) On inverse la matrice $\Lambda_{\mathcal{F}}$; cette obligation limite souvent en pratique le nombre de figures considérées.

d) On accepte l'hypothèse U, c'est-à-dire le caractère aléatoire de la suite testée si :

$$I_{\mathcal{F}}^n(s) \leqslant Q_k^{-1} (1 - \alpha)$$

où $Q_k(x)$ est la fonction de répartition de la loi du Khi-deux à k degrés de liberté et α le niveau de signification choisi.

Remarque - Il est clair que les familles de figures dont la matrice de covariance est diagonale sont particulièrement intéressantes en pratique car elles évitent les calculs b) et c) ci-dessus ; dans la suite nous déterminerons de telles familles quand elles existent. On constate sans difficulté que de telles familles de figures, même en se limitant à des figures d'ordre 1, n'existent pas toujours, par contre nous verrons au paragraphe 3 que si le cardinal de E est une puissance de 2, on sait construire de telles familles de figures.

II - TRANSFORMATIONS LINEAIRES DE LA SUITE DES OBSERVATIONS

LEMME - "Soient E un anneau commutatif fini ayant un élément unité et muni de la loi de probabilité U_o uniforme, n un entier positif et $\{Y_k, k \in K\}$ une famille finie d'applications linéaires de E^n dans E, telles qu'il existe une famille $\{\xi_k, k \in K\}$ d'éléments de E^n tels que :

$$Y_k (\xi_{k'}) = \begin{cases} 0 & \text{si } k \neq k' \\ 1 & \text{si } k = k' \end{cases} ;$$

les variables aléatoires définies sur $[E, \mathcal{P}(E), U_o]^n$ par les applications Y_k ($k \in K$) sont indépendantes et de loi U_e.

La linéarité des applications Y_k signifie que pour tout k de K on a :

$$\forall (x_1, \ldots, x_n), (y_1, \ldots, y_n) \in E^n \qquad (1)$$

$$Y_k(x_1 + y_1, \ldots, x_n + y_n) = Y_k(x_1, \ldots, x_n) + Y_k(y_1, \ldots, y_n)$$

$$\forall \alpha \in E, \forall (x_1, \ldots, x_n) \in E^n \qquad (2)$$

$$Y_k(\alpha x_1, \ldots, \alpha x_n) = \alpha Y_k(x_1, \ldots x_n).$$

D'autre part, la condition de l'énoncé entraîne que l'application Y de E^n dans E^K définie par la famille $\{Y_k, k \in K\}$ est surjective ; en effet :

$$\forall \{y_k, k \in K\} \in E^K, \quad Y_k\left(\sum_{i \in K} y_i \xi_i\right) = y_k \quad \forall k \in K,$$

où l'addition et la multiplication dans E^n sont celles apparaissant dans (1) et (2).

Désignons donc par n_y le nombre de solutions du système d'équations :

$$Y_k(u) = 0 : \forall k \in K,$$

il vient :

$$\forall \{y_k, k \in K\} \in E^K \quad P\{Y_k = y_k, \forall k \in K\} = \frac{n_y}{(\text{card } E)^n}$$

et la démonstration est achevée.

THEOREME 1 - "Supposons que E soit muni de la loi de probabilité uniforme U_o et d'une structure d'anneau commutatif avec élément unité ; soient r un entier supérieur à 1 et $\alpha_1, \ldots, \alpha_{r-1}$ des éléments donnés de E. Les variables aléatoires définies sur $[E, \mathcal{P}(E), U_o]^N$ par :

$$n = r, \ldots \quad Z_n = x_n + \alpha_{r-1} x_{n-1} + \ldots \alpha_1 x_{n-r+1}$$

où $(x_1, \ldots, x_n, \ldots) \in E^N$, sont indépendantes et de loi U_o".

Il est d'abord facile de construire par récurrence une suite $s = (s_1, \ldots, s_n, \ldots)$ dont les $r-1$ premiers termes sont nuls et telle que :

$$Z_r(s) = 1, \quad Z_n(s) = 0 \quad \forall n > r ;$$

pour tout entier positif j, la suite S^j égale à la suite s précédée de j zéros, est telle que :

$$Z_{r+j}(S^j) = 1, \quad Z_n(S^j) = 0 \quad \forall n \neq r+j.$$

Pour tout entier n supérieur à r, on peut appliquer le lemme 1 aux variables aléatoires $Z_r, \ldots, Z_n$, en prenant pour éléments ξ_{r+j} de E^n respectivement les n premiers termes des suites S^j ; la démonstration est alors achevée.

Remarque - Les suites obtenues par deux transformations linéaires différentes, ne sont pas indépendantes entre elles, mais la covariance entre certaines figures définies par leur intermédiaire peut être calculée assez aisément.

THEOREME 2 - "Supposons que E soit muni d'une structure d'anneau commutatif avec élément unité ; soient r un entier supérieur à 1 et $(\alpha_o, \beta_o, \alpha_1, \beta_1, \ldots, \alpha_{r-1}, \beta_{r-1})$ des éléments de E. En désignant par $(x_1, \ldots x_r)$ un point de E^r, la covariance entre les figures ;

$$\Phi = \{\alpha_1 x_1 + \ldots + \alpha_{r-1} x_{r-1} + x_r = \alpha_o\}$$

$$\Phi' = \{\beta_1 x_1 + \ldots + \beta_{r-1} x_{r-1} + x_r = \beta_o\}$$

est égale à :

$$\frac{\text{card } (\Phi \cap \Phi')}{(\text{card } E)^r} - \frac{1}{(\text{card } E)^2};$$

en particulier si l'une des différences $\alpha_1 - \beta_1, \ldots, \alpha_{r-1} - \beta_{r-1}$ est inversible dans E cette covariance est nulle".

Avec les notations du théorème 1, notons de plus :

$$Z'_n = x_n + \beta_{r-1} x_{n-1} + \ldots + \beta_1 x_{n-r+1}$$

toutes les variables aléatoires Z_n et Z'_n sont de loi U_o, mais de plus, si $n \neq n'$, on constate aisément par application du lemme, que les variables aléatoires Z_n et Z'_n, sont indépendantes. Dans la formule de la définition 3 paragraphe 1, seul est donc non nul le terme correspondant à j = r et la première partie du théorème est établie. De plus si l'une des différences $\alpha_1 - \beta_1, \ldots, \alpha_{r-1} - \beta_{r-1}$, est inversible, le lemme permet de montrer que Z_n et Z'_n sont indépendantes et le dernier terme de la covariance disparaît lui aussi.

THEOREME 3 - "Supposons que E soit muni d'une structure de corps commutatif et soient $a^i = \left(a^i_1, \ldots, a^i_{r-1}\right)$ (i = 1 ... k) k éléments distincts de de E^{r-1}; en notant :

$$\forall\, a \in E,\ i = 1 \ldots k\ ,\quad \Phi^i_a = \left\{x_1 a^i_1 + \ldots + x_{r-1} a^i_{r-1} + x_r = a\right\}$$

la variable aléatoire :

$$Q = \frac{1}{n \text{ card } E} \sum_{\substack{a \in E \\ i=1 \ldots k}} (\nu^n_{\Phi^i_a} \text{ card } E - n)^2$$

a pour loi limite, quand $n \longrightarrow \infty$, la loi du Khi-deux à k(card (E) -1) degrés de liberté".

En particulier, en utilisant tous les points de E^{r-1}, on obtient un indice d'aléa d'ordre r, dont la loi est asymptotiquement la loi du Khi-deux à $(\text{card } E)^{r-1}$ (card E - 1) degrés de liberté.

Démonstration - D'après le théorème 1 et le théorème de Karl Pearson, pour tout a_i, la variable aléatoire

$$Q_i = \frac{1}{n} \sum_{a \in E} (\nu^n_{\Phi^i_a} \text{ card } E - n)^2$$

a pour loi limite, quand $n \longrightarrow \infty$, la loi du χ^2 à (card E $-$ 1) degrés de liberté.

D'autre part, d'après le lemme 1 paragraphe 1 et la deuxième partie du théorème 2 ci-dessus, ces variables aléatoires $Q_i \{i = 1 \ldots\}$ sont aysmptotiquement indépendantes et le théorème est donc établi.

Remarque - Ce théorème 3 qui généralise, dans un certain sens, le théorème de K. Pearson, montre avec le théorème 2 que les "hyperplans" de E^r définis par une condition linéaire :

$$\alpha_1 x_1 + \ldots + \alpha_{r-1} x_{r-1} + x_r = \alpha_o$$

définissent des figures dont le calcul de la covariance est relativement simple. De plus, si les conditions du théorème 3 sont remplies, on obtient directement l'indice d'aléa associé à une famille de telles figures. On a ainsi résolu presque complètement le problème posé par la remarque du paragraphe 1 ; c'est seulement si le cardinal de E est une puissance de 2 que l'on peut résoudre entièrement ce problème et trouver même une solution d'une remarquable simplicité, comme on va le voir au paragraphe 3.

III - FIGURES FONDAMENTALES (card E = 2^k)

Dans tout ce paragraphe on suppose donc que card E = 2^k et on représente un élément x de E par un vecteur colonne à composantes bivalentes :

$$x \in E \Longrightarrow x = \begin{pmatrix} x^1 \\ \vdots \\ x^k \end{pmatrix} x^j = 0 \text{ ou } 1, \quad j = 1 \ldots k.$$

DEFINITION 1 - "Soit I une famille de couples d'entiers (i, j), ($1 < i < r$, $1 < j < k$) contenant au moins un couple pour lequel i = 1 et au moins un couple pour lequel j = r, alors la partie de E^r définie par :

$$(x_1, \ldots, x_r) \in E^r \quad \sum_{(i,j)\in I} x_i^j = 0 \qquad \text{(modulo 2)}$$

est une figure d'ordre r, appelée figure fondamentale définie par I et notée F_I".

Il existe exactement $2^{rk} - 2^{(r-1)k}$ figures fondamentales d'ordre au plus égal à r.

THEOREME 1 - "Soit Φ une figure fondamentale, on a :

$$E_U(\nu_\Phi^n) = \frac{n}{2}, \quad \lambda(\Phi, \Phi) = \frac{n}{4}$$

de plus si Φ' est une autre figure fondamentale, on a :

$$\lambda(\Phi, \Phi') = 0".$$

Ce théorème résulte directement de ce que, si E est muni de la loi uniforme U_o, les variables aléatoires $1_\Phi^{i'}$ et $1_{\Phi'}^{i'}$ sont indépendantes si $i \neq i'$ ou $\Phi \neq \Phi'$. Or ce dernier point résulte lui-même du lemme suivant, cas particulier du lemme du paragraphe 2, et que l'on peut d'ailleurs vérifier directement :

LEMME - "Soient $\mathcal{J}$ un ensemble fini et X_i, $(i \in \mathcal{J})$ des variables aléatoires de Bernouilli indépendantes de loi B :

$$P(X_i = 0) = P(X_i = 1) = \frac{1}{2} \quad \forall\ i \in \mathcal{J} ;$$

alors quelles que soient les deux parties I et J de $\mathcal{J}$, non vides et distinctes, les deux variables aléatoires

$$U = \sum_{i\in I} X_i \qquad V = \sum_{i\in J} X_i \qquad \text{(modulo 2)}$$

sont indépendantes et de loi B".

Remarques

1/ En particularisant le théorème 1 paragraphe 2, on constate aisément que si Φ est une figure fondamentale, ν_Φ^n a pour loi de probabilité la loi binomiale $\mathcal{B}\left(n, \frac{1}{2}, \frac{1}{2}\right)$.

2/ Les figures fondamentales sont d'un usage pratique très commode, puisque l'indice d'aléa associé à une famille $\mathcal{F}$ de ces figures est égal à :

$$\frac{4}{n} \sum_{\Phi\in\mathcal{F}} \left[\nu_\Phi^n - \frac{n}{2}\right]^2 .$$

De plus, on peut faire une étude de plus en plus fine de l'aléa en augmentant l'ordre des figures fondamentales utilisées.

3/ Enfin la famille des figures fondamentales est minimum au sens suivant ; la n-fréquence de toute figure d'ordre inférieur ou égal à r est, à quelques unités près, combinaison linéaire de n-fréquences des figures fondamentales d'ordre au plus égal à r.

On trouvera dans [3] un exemple de programmation sur ordinateur de l'indice d'aléa défini par une famille de figures fondamentales. On a appliqué cette méthode aux 50 000 premiers bits de la représentation binaire de π, en prenant les 16 figures fondamentales d'ordre au plus égal à 5.

On représente ci-après une figure fondamentale Φ par l'entier dont l'écriture binaire représente la partie I définissant Φ (k est ici égal à 1) ; on a observé :

Figure	Fréquence	Figure	Fréquence
16	25 799	24	25 227
17	25 233	25	25 131
18	24 961	26	25 389
19	25 031	27	25 069
20	25 091	28	25 211
21	25 225	29	25 223
22	25 105	30	25 085
23	25 115	31	24 989

l'indice d'aléa aisément calculé est alors supérieur à 34,267 qui est le seuil à 5 %.

BIBLIOGRAPHIE

[1] J.R. BARRA - Contrôle statistique d'une suite de digits aléatoires Revue de statistique appliquée. Vol XV N° 3 (1967).

[2] HENNEQUIN et TORTRAT - Théorie des probabilités et quelques applications - Masson 1965.

[3] HISLEUR - Exemples d'utilisation d'un calculateur en statistique - Thèse Grenoble 1969.

[4] J.R. BARRA - Notions fondamentales de statistique mathématique - DUNOD 1971.