REVUE DE STATISTIQUE APPLIQUÉE

D. CHESSEL
C. GAUTIER

Des statistiques non paramétriques pour l'analyse des données binaires

Revue de statistique appliquée, tome 25, n° 1 (1977), p. 57-73
<http://www.numdam.org/item?id=RSA_1977__25_1_57_0>

DES STATISTIQUES NON PARAMÉTRIQUES POUR L'ANALYSE DES DONNÉES BINAIRES [1]

D. CHESSEL - C. GAUTIER
Laboratoire de Biométrie
Université Lyon I 69621 Villeurbanne (France)

I – INTRODUCTION.

Divers problèmes de traitement des données biologiques, rencontrés dans l'analyse de la dispersion spatiale des populations animales ou végétales d'une part (CHESSEL 1975), dans l'étude des structures des protéines d'autre part (GAUTIER, 1976, GRANTHAM 1975), nous ont amené à employer les modèles non paramétriques d'équiprobabilité dans les espaces

– des N^R applications des $\{1, \ldots, R\}$ dans $\{1, \ldots, N\}$

– des $N!$ permutations des $\{y_1, \ldots, y_N\}$

– des $\binom{N}{M}$ parties à M éléments de $\{1, 2, \ldots, N\}$

– des $\dfrac{N!}{z_1! \ldots z_m!}$ mots du type $1^{z_1} \ldots m^{z_m}$.

Conduits à définir des statistiques adaptées aux questions posées dans les plans d'expériences au laboratoire ou dans la nature, nous avons réuni ici les démonstrations correspondant aux propositions faîtes dans des publications biologiques. L'ensemble s'appuie sur une des techniques fondamentales des probabilités combinatoires : les fonctions génératrices écrites à partir des énumérateurs.

Un premier paragraphe porte sur le modèle classique d'occupation des N^R applications de $\{1, \ldots, R\}$ dans $\{1, \ldots, N\}$ pour lequel on peut employer le langage des attributions de R boules dans N boîtes ou celui des R-échantillons avec remise d'une urne à N boules de couleurs distinctes. Nous rappelons les résultats connus sur les v.a. de ce modèle et une démonstration de RIORDAN contenant une indication que nous avons systématiquement utilisée, sur l'emploi des f.g. (fonctions génératrices) de f.g.p. (fonctions génératrices de probabilités). Nous proposons une solution au problème de la distinction entre contagion vraie et contagion fausse posé, il y a fort longtemps, par FELLER (1943).

---

(1) Article remis en juin 1976.

Mots-clés : **Probabilités combinatoires. Problèmes d'occupation. Loi multinomiale. Loi hypergéométrique.**

Une deuxième partie étend les résultats précédents et la technique repérée au modèle multi-hyper géométrique de l'équiprobabilité des $\binom{N}{M}$ choix de M positions dans N possibles, ces N positions possibles formant B blocs de K positions. On obtient ainsi diverses statistiques, descriptives des structures d'échantillons de variables binaires, définies dans le cadre d'une analyse non paramétrique.

Les notations suivantes sont utilisées :

$$(N)_k = N(N-1)\ldots(N-k+1)$$

Si Y est une v.a., dont la loi dépend des paramètres $\mu_1, \ldots, \mu_\varrho$ sa f.g.p. est :

$$G_Y(s\ ;\mu_1,\ldots,\mu_\varrho) = \sum_k s^k P(Y=k\ ;\mu_1,\ldots,\mu_\varrho)$$

Ses moments factoriels sont :

$$E((Y)_k) = F_k(Y\ ;\mu_1,\ldots,\mu_\varrho)$$

La f.g. des probabilités conjointes de $Y_1, \ldots, Y_u$ (f.g.p.c.) est :

$$G_{Y_1,\ldots,Y_u}(s_1,\ldots,s_u\ ;\mu_1,\ldots,\mu_\varrho) = \sum_{k_1,\ldots,k_u} s_1^{k_1}\ldots s_u^{k_u}$$
$$P(Y_1=k_1,\ldots,Y_u=k_u\ ;\mu_1,\ldots,\mu_\varrho)$$

II – MODELE CLASSIQUE D'OCCUPATION.

Il est défini par l'équiprobabilité des N^R attributions de R boules distinctes dans N cases distinctes.

Les variables aléatoires définies sur cet espace sont :

* Y_k est le nombre d'objets attribués à la k^{ieme} case. Y_k suit une loi binomiale de paramètres R et 1/N.

* $S = (Y_1, \ldots, Y_N)$. On peut appeler *répartition* une valeur observée de S. S suit une loi multinomiale de paramètres R, 1/N, . . . , 1/N.

* $T = \sum_{i=1}^{N} Y_i^2$ est une statistique de dispersion. COX et LEWIS (1969, ch. 9, p. 225) rappelle la moyenne et la variance de T

$$E(T) = \frac{R(R+N-1)}{N} \qquad V(T) = \frac{2R(R-1)(N-1)}{N^2}$$

* $I_D = \dfrac{\sum_{i=1}^{N}\left(Y_i - \frac{R}{N}\right)^2}{(N-1)\frac{R}{N}}$ est l'indice de dispersion. HOEL (1943) a pré-

cisé les conditions d'approximations de $(N-1)I_D$ par un $\chi^2(N-1)$. On peut consulter la bibliographie sur le sujet dans STITELER et PATIL (1969).

* X est le nombre de cases occupées. X suit la loi de Stevens-Craig

$$P(X = j) = \frac{1}{N^R} S(R, j) (N)_j$$

où S(R, j) est le nombre de Stirling de 2e ordre (cf JOHNSON et KOTZ (1969 ch. 10 sec. 5), PATIL et JOSHI (1968 p. 49)). Une approximation est définie par FELLER (1957, ch. IV, sec. 2).

* Z_j est le nombre de cases contenant j objets. RIORDAN (1968, ch. 5 ex. 3) donne les moments factoriels de Z_j, de la manière suivante :

Appelons $C_1, \ldots, C_N$ les cases. L'écriture formelle

$$\prod_{i=1}^{N} \left(1 + C_i t + \ldots + C_i^k \frac{t^k}{k!} + \ldots + s \frac{C_i^j}{j!} t^j + \ldots\right)$$

$$= \sum_{R \geq 0} \frac{t^R}{R!} \sum_{k \geq 0} s^k \sum_{\substack{\Sigma \lambda_i = R \\ \lambda_{i_1} = \ldots = \lambda_{i_k} = j}} \frac{R!}{\lambda_1! \ldots \lambda_n!} C_1^{\lambda_1} \ldots C_n^{\lambda_n} \tag{1}$$

permet d'avoir les quantités $C_1^{\lambda_1} \ldots C_n^{\lambda_n}$ multipliées par $\frac{R!}{\lambda_1! \ldots \lambda_n!}$, c'est-à-dire exactement chaque répartition (la case C_i contient λ_i objets) où k cases contiennent j objets multipliée par le nombre de possibilités correspondantes. Si on fait $C_1 = \ldots = C_N = 1$ la somme finale de (1) est le nombre de façons de distribuer R objets dans N cases avec exactement k cases contenant j objets. En multipliant et divisant par N^R on passe aux probabilités

$$\left(1 + t + \ldots + \frac{t^k}{k!} + \ldots + s\frac{t^j}{j!} + \ldots\right)^N = \sum_{R \geq 0} \frac{N^R t^R}{R!} \sum_{k \geq 0} s^k P(Z_j = k ; R, N) \tag{2}$$

et aux f.g.p.

$$\left(e^t + (s-1)\frac{t^j}{j!}\right)^N = \sum_{R \geq 0} \frac{(Nt)^R}{R!} G_{Z_j}(s ; R, N) \tag{3}$$

soit

$$\left(e^{t/N} + (s-1)\frac{\left(\frac{t}{N}\right)^j}{j!}\right)^N = \sum_{R \geq 0} \frac{t^R}{R!} G_{Z_j}(s ; R, N) \tag{4}$$

On obtient ainsi une f.g. de f.g.p.. On sait que

$$G_{Z_j}(s + 1 ; R, N)$$

est la f.g. des quantités $\frac{F_k(Z_j ; R, N)}{k!}$

En développant en t et s après avoir substitué s + 1 à s dans (4) on obtient

$$F_k(Z_j\ ;\ R\ ,\ N) = k!\ \binom{N}{j} \frac{R!}{(R-kj)!\ (k!)^j}\ \frac{(N-j)^{R-kj}}{N^R}$$

La formule d'inversion entre moments factoriels et probabilités donne la loi des Z_j et le théorème de Fréchet (DAVID et BARTON, 1962, p. 223) définit la précision numérique obtenue par cette méthode.

Un cas particulier important du modèle est constitué par N = 2. On définit dans ce cas le test des signes et le test de Mac Nemar (cf. CONOVER 1971 p. 121-130). Notons que R objets sont distribués au hasard entre 2 cases. U et V sont les effectifs respectifs de chacune des cases. Le test des signes s'appuie sur la distribution binomiale (R, 1/2) de U et le test de Mac Nemar s'appuie sur l'approximation de $\frac{(U-V)^2}{U+V}$ par un $\chi^2(1)$.

La situation expérimentale où l'on a un certain nombre de couples d'observation U_i, V_i n'est pas rare : individus mâle ou femelle par unité d'échantillonnage, effectifs a_{ij} et a_{ji} d'une table de contingence carrée, effectifs par couple d'échantillons élémentaires. Un test du modèle binomial (sex-ratio égal à 1/2 et indépendance des individus, symétrie de la table de contingence, dispersion localement poisonnienne) est constitué par l'application du théorème central limite aux v.a.

$$D = |U - V|\ .$$

Il a été proposé indépendamment par CHESSEL et DE BELAIR (1973) et MEAD (1974). La statistique ainsi obtenue peut s'appeler indice de contagion vraie par référence au problème soulevé par FELLER (1943). Il est en effet connu que nombre de distributions discrètes contagieuses sont soit des mélanges de loi de Poisson (contagion fausse : par exemple les accidents d'un conducteur sont poissoniens mais le taux varie d'un conducteur à l'autre) soit générées par des modèles de distributions corrélées (contagion vraie : par exemple les taux d'accidents par conducteurs sont égaux mais l'occurrence d'un accident augmente la probabilité d'un nouvel accident pour le même conducteur).

L'examen de couples d'unité de temps pour un ensemble de conducteurs permettrait de choisir entre les deux modèles.

Techniquement, abordons le problème plus largement. Soit S et E le nombre respectif des succès (probabilité p) et des échecs (probabilité q = 1 − p) dans un échantillon de R essais indépendants. Calculons moyenne et variance de |S − E|. Le cas particulier p = q = 1/2 définit alors l'indice de contagion vraie.

– Notons d'abord que

$$E(D^2\ ;\ R\ ,\ p) = 4Rpq + R^2(1-4pq)$$

car $D^2 = (2S - R)^2$. Pour p = 1/2 on a $E(D^2\ ;\ R) = R$. L'essentiel du problème, est donc le calcul de la moyenne.

– *LEMME PRELIMINAIRE.* Si X et Y sont deux v.a. la f.g.p. de $D = |X - Y|$ vérifie :

$$G_D(t) + G_D\left(\frac{1}{t}\right) = G_{X,Y}\left(t, \frac{1}{t}\right) + G_{X,Y}\left(\frac{1}{t}, t\right)$$

La quantité $G'_D(t)$ est alors la partie entière de la série

$$\left[G_{X,Y}\left(t, \frac{1}{t}\right) + G_{X,Y}\left(\frac{1}{t}, t\right)\right]'.$$

Ceci permet le calcul de $E(D) = G'_D(1)$.

La moyenne de D peut donc s'obtenir par l'étude de la fonction

$$\frac{d}{dt}\left[G_{S,E}\left(t, \frac{1}{t}\ ; R, p\right) + G_{S,E}\left(\frac{1}{t}, t\ ; R, p\right)\right]$$

Pour alléger les notations nous écrirons dans cette démonstration

$$G_{S,E}(s_1, s_2\ ; R, p)$$

sous la forme $\phi_R(s_1, s_2) = \Sigma\, a_{ij;R}\, s_1^i s_2^j$; ce qui est possible p restant invariant dans toute la démonstration.

S étant le nombre de succès d'une binomiale on a :

$$\phi_R(s_1, s_2) = (ps_1 + qs_2)^R$$

d'où :

$$\frac{d}{dt}\left[\phi_R\left(t, \frac{1}{t}\right) + \phi_R\left(\frac{1}{t}, t\right)\right] = R\left(p - \frac{q}{t^2}\right)\left(pt + q\frac{1}{t}\right)^{R-1}$$

$$+ R\left(q - p\frac{1}{t^2}\right)\left(p\frac{1}{t} + qt\right)^{R-1} = R\left(p - \frac{q}{t^2}\right)\phi_{R-1}\left(t, \frac{1}{t}\right)$$

$$+ R\left(q - \frac{p}{t^2}\right)\phi_{R-1}\left(\frac{1}{t}, t\right) = R\sum_{k=-\infty}^{\infty} t^k\left[p\sum_{i-j=k} a_{ij,R-1} - q\sum_{i-j=k+2} a_{ij,R-1}\right.$$

$$\left. + q\sum_{j-i=k} a_{ij,R-1} - p\sum_{j-i=k+2} a_{ij,R-1}\right]$$

Pour t = 1 le lemme préliminaire permet d'écrire

$$E(D_R) = R\sum_{k\geqslant 0}\left[p\left(\sum_{i-j=k} a_{ij;R-1} - \sum_{j-i=k} a_{ij,R-1}\right)\right.$$

$$\left. - q\left(\sum_{i-j=k} a_{ij,R-1} - \sum_{j-i=k} a_{ij,R-1}\right)\right] + R\left[q\sum_{\substack{i-j=0\\ i-j=1}} a_{ij,R-1} + p\sum_{\substack{j-i=0\\ j-i=1}} a_{ij,R-1}\right]$$

On note T_R et V_R les deux termes de cette somme.

V_R se calcule très simplement suivant la parité de R :

$$V_{2L} = 2L\,[q\,a_{L,L-1;2L-1} + p\,a_{L-1,L;2L-1}] = 2L\binom{2L-1}{L} q^{L-1}p^{L-1}$$

$$V_{2L+1} = (2L+1)\,a_{L,L;2L} = (2L+1)\binom{2L}{L} p^L q^L$$

Si on note $P_R(A)$ la probabilité de l'évènement A s'il y a R tirages, on peut écrire T_R sous la forme suivante :

$$T_R = R(p-q)\,[P_{R-1}(S \geqslant E) - P_{R-1}(E \geqslant S)]$$

On peut calculer $P_R = P_R(S \geqslant E)$ par récurrence :

$$P_{2L+1} = p\,P_{2L} + q\left(P_{2L} - \binom{2L}{L} p^L q^L\right) = P_{2L} - \binom{2L}{L} p^L q^{L+1}$$

$$P_{2L+2} = p\left(P_{2L+1} + \binom{2L+1}{L} p^L q^{L+1}\right) + q\,P_{2L+1}$$

$$= P_{2L+1} + \binom{2L+1}{L} p^{L+1} q^{L+1}$$

On obtient alors T_R par :

$$T_{2L} = 2L(p-q)\,(2P_{2L-1} - 1)$$

$$T_{2L+1} = (2L+1)\,(p-q)\left(2P_{2L} - 1 + \binom{2L}{L-1} p^L q^L\right)$$

Dans le cas particulier important $p = q = \frac{1}{2}$ T_R est nul et on a donc

$$E(D_{2L}) = 2L\binom{2L-1}{L}\left(\frac{1}{2}\right)^{2L-1}$$

$$E(D_{2L+1}) = (2L+1)\binom{2L}{L}\left(\frac{1}{2}\right)^{2L}$$

Résultat que MEAD (1974) écrit sous la forme $\prod_{j=2}^{(n+1)/2}\left(\frac{2j-1}{2j-2}\right)$. On remarque aisément que $E(D_{2L}) = E(D_{2L-1})$.

Le calcul de $E(D_R)$ étant compliqué pour N grand une forme limite peut être utile. L'utilisation des formules de Stirling donne :

$$L \to \infty \Rightarrow E(D_{2L+1}) \sim \frac{2L+1}{\sqrt{\pi L}}$$

L'approximation ainsi obtenue semble assez rapidement valable puisque pour $L = 10$ les valeurs exactes et approchées sont respectivement 3,700 et 3,746.

On trouvera des exemples d'applications numériques de l'indice de contagion vraie dans JARRY (1974 : dispersion de la ponte d'un insecte), PRODON (1976 : choix multiple de milieux par des larves aquatiques), DEBOUZIE et COLL. (1975 : échantillonnage systématique par unité de plantes steppiques), MOUEZA (1976), LEGAY et CHESSEL (1976 : sex-ratio par unité), GAUTIER (1976 : symétrie des tables de contingence).

Nous allons maintenant étendre l'ensemble des résultats précédents au modèle hypergéométrique.

III – LE MODELE HYPERGEOMETRIQUE.

L'espace de probabilité constitué des $\binom{N}{M}$ parties à M éléments de $\{1, 2, \ldots, N\}$ muni de la probabilité uniforme est l'un des plus étudiés en statistique non paramétrique. On en trouve une étude très complète dans DAVID et BARTON (1962). Les statistiques s'organisent en trois groupes fondés respectivement sur des sommes de rangs, les suites et l'utilisation d'une coupure.

* Les premières sont utilisées en particulier pour la comparaison des échantillons soit relativement à leur moyenne (WILCOXON) soit à leur variance (ANSARI-BRADLEY, SIEGEL-TUKEY). Une revue de ces tests peut être faite dans HOLLANDER et WOLFE (1973 page 68 et suivantes) et dans CONOVER (1971 page 223 et suivantes).

* DAVID et BARTON (1962 page 188) font une étude des statistiques liées à une coupure : nombre d'éléments choisis avant la coupure, différence entre la somme des rangs des éléments précédant et suivant la coupure, cette même statistique conditionnée par le nombre d'éléments choisis avant la coupure.

* Un dernier test peut être rattaché à ce modèle : le test de comparaison de deux échantillons par la statistique de Kolmogorov-Smirnov (par exemple HOLLANDER et WOLFE 1973 page 219).

Notre souci de recherche de structure introduit de manière naturelle des statistiques liées à une partition de $\{1, 2, \ldots, N\}$ en blocs. Or malgré la grande diversité des études réalisées sur ce modèle il ne nous semble pas que de telles statistiques aient été étudiées bien qu'on puisse rapprocher ce problème de l'un des suivants :

* Distribution de R boules dans N cases avec limitation du nombre d'objets par case (RIORDAN 1958, chapitre 5 exercice 6).

* Distribution de R objets identiques dans N cases distinctes. L'espace compte alors $\binom{R+N-1}{R}$ éléments (RIORDAN 1958, page 92).

* Distribution de R objets distincts dans N cases identiques. L'espace contient $S(R, 1) + \ldots + S(R, N)$ éléments (KAUFMANN, 1968, p. 156).

* Distribution d'objets de couleurs différentes dans N boîtes et comptage des couleurs représentées par boîtes (DAVID et BARTON, 1959).

Notons enfin, pour situer le problème abordé, qu'il s'apparente aux résultats rapportés par CLIFF et ORD (1973) sur les mesures d'autocorrélation spatiale dans le cadre de données binaires et du modèle "Non free sampling".

De même que dans le modèle précédent, les démonstrations reposeront sur l'utilisation des f.g.p.. On notera $\mathcal{H}(N, M)$ l'espace considéré. Un énumérateur de $\mathcal{H}(N, M)$ est :

$$(1 + xt_1) \ldots (1 + xt_N) = \sum_M x^M \sum_{(i_1,\ldots,i_M)\in \mathcal{H}(N,M)} t_{i_1} \ldots t_{i_M}$$

III.1. Choix dans un bloc.

Soit K la taille du bloc, l'énumérateur des choix avec j positions choisies dans ce bloc est :

$$(1 + xta_1) \ldots (1 + xta_K)(1 + xa_{K+1}) \ldots (1 + xa_N)$$

$$= \Sigma x^M \Sigma t^j \sum_{\substack{(i_1,\ldots,i_j)\in \mathcal{H}(K,j) \\ (k_1-K,\ldots,k_{M-j}-K)\in \mathcal{H}(N-K,M-j)}} a_{i_1} \ldots a_{i_j}\ a_{k_1} \ldots a_{k_{M-j}}$$

$$(1 + xt)^K (1 + x)^{N-K} = \Sigma x^M \binom{N}{M} G_Y(t ; N, M, K) \qquad (5)$$

en notant Y la variable aléatoire : nombre de positions choisies dans le bloc. L'identification des coefficients des termes en x^M puis en t^j conduit à la relation

$$P(Y = j) = \frac{\binom{K}{j}\binom{N-K}{M-j}}{\binom{N}{M}}$$

On retrouve ainsi un résultat classique : Y suit une loi hypergéométrique. L'utilisation de la fonction génératrice des f.g.p. permet d'autre part de retrouver rapidement les moments factoriels de Y :

On dérive (5) par rapport à t

$$Kx(1 + xt)^{K-1} (1 + x)^{N-K} = \Sigma x^M \binom{N}{M} G'_Y(t ; N, M, K)$$

$$Kx \Sigma x^M \binom{N-1}{M} G_Y(t ; N-1, M, K-1) = \Sigma x^M \binom{N}{M} G'_Y(t ; N, M, K)$$

$$G'_Y(t\,;N,M,K) = \frac{KM}{N} G_Y(t\,;N-1,M-1,K-1)$$

de manière plus générale

$$\frac{d^k G_Y(t\,;N,M,K)}{dt^k} = \frac{(K)_k (M)_k}{(N)_k} G_Y(t\,;N-k,M-k,K-k)$$

En donnant à t la valeur 1 on obtient

$$F_k(Y\,;N,M,K) = \frac{(K)_k (M)_k}{(N)_k}$$

On peut remarquer que cette méthode pour obtenir les moments factoriels d'une hypergéométrique n'est peut être pas la plus rapide (par exemple utilisation du théorème sur la somme de variables caractéristiques cité dans la suite) mais elle se généralise au cas multi-hypergéométrique étudié au paragraphe suivant.

III.2. Choix dans une partition.

L'ensemble des N positions est divisé en B blocs de tailles respectives $K_1, K_2, \ldots, K_B$. On note $Y_1, \ldots, Y_B$ les nombres de positions choisies dans chaque bloc. Par la même méthode qu'au paragraphe précédent on obtient :

$$(1 + xt_1)^{K_1} \ldots (1 + xt_B)^{K_B} = \Sigma\, x^M \binom{N}{M} G_{Y_1,\ldots,Y_B}(t_1, \ldots, t_B\,;K_1, \ldots K_B, M)$$

Ce qui permet de connaître la loi conjointe des Y_i :

$$P(Y_1 = y_1, \ldots, Y_B = y_B) = \frac{\binom{K_1}{y_1} \ldots \binom{K_B}{y_B}}{\binom{N}{M}}$$

On retrouve ainsi que cette loi est une multi-hypergéométrique. En utilisant la méthode utilisée pour calculer $F_k(Y; N, M, K)$ au paragraphe précédent, appliquée à

$$\frac{\partial^{\Sigma j_i} G_{Y_1 \ldots Y_B}(t_1, \ldots, t_B\,;K_1, \ldots, K_B, M)}{\partial^{j_1} t_1 \ldots \partial^{j_B} t_B}$$

on obtient les moments factoriels conjoints :

$$E((Y_1)_{j_1} \ldots (Y_B)_{j_B}) = (K_1)_{j_1} \ldots (K_B)_{j_B} \frac{(M)_{\Sigma j_i}}{(N)_{\Sigma j_i}} \qquad (6)$$

III.3. Indice de dispersion T.

L'ensemble des N positions est divisé en B blocs de tailles K ; les notations du paragraphe précédent sont conservées.

$$T = \sum_{i=1}^{B} Y_i^2$$

Les moments factoriels déterminés par l'égalité (6) permettent le calcul de $E(Y_i^4)$, $E(Y_i^2 Y_j^2)$ et donc de la moyenne et de la variance de T :

$$E(T) = M((K-1)(M-1)/(N-1) + 1)$$

$$V(T) = \frac{2M(M-1)(K-1)(N-K)(N-M)(N-M-1)}{(N-1)^2(N-2)(N-3)}$$

CHESSEL et CROZE (1976) donnent des exemples d'utilisation de T en écologie, et montrent par simulation que l'approximation normale est valable pour la queue droite de la distribution dans un large domaine. On observe d'autre part que le maximum de $\frac{T - E(T)}{\sqrt{V(T)}}$ pour des tailles de blocs en progression géométrique de raison 2 a un seuil, au niveau 0.05, d'environ 3. En particulier ce seuil ne semble pas dépendre de N ou de M.

III.4. Nombre de blocs vides.

L'ensemble des N positions est divisé en B blocs de taille K. On note V_K le nombre de blocs vides.

Nous allons utiliser un théorème sur les sommes de variables caractéristiques dont l'énoncé suit (DAVID et BARTON 1962, p. 70-71).

THEOREME. *Soit* N *variables caractéristiques* α_i *des évènements* E_i ; *on note* S *la somme des* α_i. *Si la loi conjointe des* α_i *est symétrique on a* :

$$F_\ell(S) = (N)_\ell \; P(\alpha_1 \alpha_2 \ldots \alpha_\ell = 1).$$

L'application du théorème à V_k est immédiate et on obtient

$$F_\ell(V_K) = (B)_\ell \frac{(N-M)_{K\ell}}{(N)_{K\ell}}$$

On en déduit moyenne et variance de V_k :

$$E(V_K) = B \frac{(N-M)_K}{(N)_K}$$

$$V(V_K) = (B)_2 \frac{(N-M)_{2K}}{(N)_{2K}} + E(V_K) - (E(V_K))^2.$$

Cette méthode qui fournit très rapidement les moments de V_K a le désavantage de ne pas donner la loi autrement qu'à travers les formules assez compliquées d'inversion. Aussi allons-nous donner une courte approche utilisant les fonctions génératrices.

On a :

$$\left(t + \binom{K}{1} u + \binom{K}{2} u^2 + \ldots + \binom{K}{K} u^K\right)^B = \Sigma\, u^M \binom{N}{M} G_{V_K}(t\,;\, N,\, M)$$

On peut alors retrouver les moments factoriels par dérivations successives.

D'autre part en développant la fonction génératrice de f.g.p. écrite ci-dessus et après quelques calculs simples on obtient

$$P(Y_K = j\,;\, N, K, M) = \sum_{h=j}^{B} (-1)^{h-j} \frac{(B)_h}{j!\,(h-j)!} \frac{(N-M)_{Kh}}{(N)_{Kh}}$$

ce qui correspond aux formules d'inversions mentionnées ci-dessus mais on peut également en déduire les relations de récurrence suivantes :

$$j!\; P(V_K = j\,;\, N, K, M) = \frac{(N-M)_{Kj}}{(N)_{Kj}} (B)_j \quad P(V_K = 0\,;\, N - Kj, K, M)$$

$$P(V_K = 0\,;\, N, K, M) = \sum_{j=1}^{K} \frac{\binom{K}{j}\binom{N-K}{M-j}}{\binom{N}{M}} \quad P(V_K = 0\,;\, N-K, K, M-j)$$

III.5. L'indice d'autocorrélation H.

L'ensemble des N positions est divisé en N/2 blocs de deux positions. H est le nombre de ces blocs dont une, et une seule, des deux positions est choisie. Il est clair qu'une autocorrélation positive diminue H et que H est grand en cas d'autocorrélation négative. CHESSEL (1975) donne un exemple d'utilisation de H.

La loi de H peut être obtenue à partir de l'énumérateur

$$\prod_{i=1}^{N/2} (1 + 2t_i s x + t_i^2 s^2) = \Sigma\, s^M \Sigma x^j \Sigma 2^j t_{i_1} \ldots t_{i_j} t_{k_1}^2 \ldots t_{\frac{k_{M-j}}{2}}^2$$

qui conduit à la fonction génératrice

$$(1 + 2sx + s^2)^{N/2} = \Sigma\, s^M \binom{N}{M} G_H(x\,;\, N, M)$$

En dérivant par rapport à x on obtient

$$N \Sigma\, s^M \binom{N-2}{M-1} G_H(x\,;\, N-2, M-1) = \Sigma\, s^M \binom{N}{M} G'_H(x\,;\, N, M)$$

$$M(N-M)\, G_H(x\,;\, N-2, M-1) = G'_H(x\,;\, N, M)\,(N-1)$$

On peut alors calculer moyenne et variance de H

$$E(H) = \frac{M(N - M)}{N - 1}$$

$$V(H) = 2\,\frac{M(N - M)\,(M - 1)\,(N - M - 1)}{(N - 1)^2\,(N - 3)}$$

De manière plus générale on peut obtenir les moments factoriels soit par dérivations successives soit en appliquant le théorème du paragraphe précédent :

$$F_k(H) = \frac{(M)_k\,(N - M)_k}{(N)_{2k}}\; 2^k \left(\frac{N}{2}\right)_k$$

D'autre part en développant l'expression de la fonction génératrice on obtient sans difficulté la loi de H :

$$P(H = j\ ;\ N\ ,\ M) = \binom{N/2}{\frac{M + j}{2}} \binom{\frac{M + j}{2}}{j} \frac{2^j}{\binom{N}{M}}$$

Une tabulation très facile de H est réalisable grâce aux formules de récurrence suivantes :

$$P(H = 0\ ;\ N,\ 0) = 1 \qquad P(H = 0\ ;\ N,\ M) = \frac{M - 1}{N - M + 1}\,P(H = 0\ ;\ N,\ M - 2)$$

$$P(H = 1\ ;\ N,\ 1) = 1 \qquad P(H = 1\ ;\ N,\ M) = \frac{M}{N - M + 2}\,P(H = 1\ ;\ N,\ M - 2)$$

$$P(H = j + 2\ ;\ N,\ M) = \frac{(M - j)\,(N - M - j)}{(j + 1)\,(j + 2)}\,P(H = j\ ;\ N,\ M)$$

III.6. Indice de contagion vraie.

L'indice de contagion vraie développé au paragraphe I est adapté aux tests portant sur les lois binomiales. Le très mauvais ajustement obtenu par MEAD (1974) lors d'une simulation conduisant à une comparaison de fréquence est une des justifications au développement d'un indice de contagion vraie dans le cadre du modèle hypergéométrique. Ce modèle répond en effet, non seulement aux problèmes de contagion dans des relevés en présences-absences, mais aussi à celui de la comparaison de fréquences. On trouvera un exemple d'utilisation de cet indice dans CHESSEL et DEBOUZIE (1974).

Les N positions étant groupées en deux blocs de tailles respectives K et $N - K$ on étudie la variable aléatoire $D_{N,K,M} = |Y_1 - Y_2|$ où Y_i est le nombre de positions choisies dans le i^e bloc.

$E(D^2_{N,K,M})$, comme dans le premier modèle est d'un calcul simple ; par contre, le calcul de la moyenne a requis l'utilisation de f.g.

$$E(D^2_{N,K,M}) = E((Y_1 - Y_2)^2) = E((2Y_1 - M)^2)$$

$$= \frac{4K(K-1)M(M-1)}{N(N-1)} + \frac{4KM(1-M)}{N} + M^2$$

De même que pour le modèle d'occupation nous poserons

$$\phi_{N,K,M}(s_1, s_2) = G_{Y_1Y_2}(s_1, s_2 ; N, K, M)$$

Nous poserons de plus

$$F(t) = \frac{d}{dt}\left[\phi_{N,K,M}\left(t, \frac{1}{t}\right) + \phi_{N,K,M}\left(\frac{1}{t}, t\right)\right]$$

$$\phi_{N,K,M}(s_1, s_2) = \Sigma\, a_{ij;N,K,M}\, s_1^i s_2^j$$

Nous avons précédemment (III.1) démontré la relation :

$$(1 + sx)^K (1 + tx)^{N-K} = \Sigma\, x^M \binom{N}{M} \phi_{N,K,M}(s, t)$$

On en déduit que $\binom{N}{M} F(t)$ est le coefficient de x^M dans l'expression ci-dessous :

$$\frac{d}{dt}\left[(1 + tx)^K \left(1 + \frac{x}{t}\right)^{N-K} + \left(1 + \frac{x}{t}\right)^K (1 + xt)^{N-K}\right]$$

$$= Kx(1 + tx)^{K-1}\left(1 + \frac{x}{t}\right)^{N-K} - \frac{N-K}{t^2}\, x(1 + tx)^K \left(1 + \frac{x}{t}\right)^{N-K-1}$$

$$- \frac{Kx}{t^2}\left(1 + \frac{x}{t}\right)^{K-1}(1 + tx)^{N-K} + (N - K)\, x\left(1 + \frac{x}{t}\right)^K (1 + tx)^{N-K-1}.$$

$$\binom{N}{M} F(t) = K\binom{N-1}{M-1} \underset{N-1,K-1,M-1}{\phi}\left(t, \frac{1}{t}\right) - \frac{N-K}{t^2}\binom{N-1}{M-1} \underset{N-1,K-1,M-1}{\phi}\left(t, \frac{1}{t}\right)$$

$$- \frac{K}{t^2}\binom{N-1}{M-1} \underset{N-1,K-1,M-1}{\phi}\left(\frac{1}{t}, t\right) + (N - K)\binom{N-1}{M-1} \underset{N-1,K,M-1}{\phi}\left(\frac{1}{t}, t\right).$$

En utilisant le lemme préliminaire énoncé en II on obtient :

$$\frac{N}{M} E(D_{N,K,M}) = \sum_{k \geq 0}' \left[K \sum_{i-j=k} a_{ij;N-1,K-1,M-1} - (N - K) \sum_{i-j=k+2} a_{ij;N-1,K,M-1}\right.$$

$$\left. - K \sum_{j-i=k+2} a_{ij;N-1,K-1,M-1} + (N - K) \sum_{j-i=k} a_{ij;N-1,K,M-1}\right]$$

$$= K\left[P_{N-1,K-1,M-1}(Y_1 \geqslant Y_2) - P_{N-1,K-1,M-1}(Y_2 \geqslant Y_1)\right] - (N-K)\left[P_{N-1,K,M-1}(Y_1 \geqslant Y_2) - P_{N-1,K,M-1}(Y_2 \geqslant Y_1)\right]$$

$$+ (N-K) \sum_{\substack{i-j=0 \\ i-j=1}} a_{ij;N-1,K,M-1} + K \sum_{\substack{j-i=0 \\ j-i=1}} a_{ij;N-1,K-1,M-1}.$$

On note $P_{N,K,M} = P_{N,K,M}(Y_1 \geqslant Y_2)$

On a alors suivant la parité de M les relations :

a) $M = 2L$

$$P_{N,K,M} = P_{N,K-1,M} + \frac{\binom{K-1}{L-1}\binom{N-K}{L}}{\binom{N}{M}}$$

b) $M = 2L + 1$

$$P_{N,K,M} = P_{N,K-1,M} + \frac{\binom{K-1}{L}\binom{N-K}{L}}{\binom{N}{M}}$$

Suivant la parité de M l'expression de $E(D_{N,K,M})$ peut être simplifiée :

a) $M = 2L$

$$\frac{N}{M} E(D_{N,K,M}) = K\left[2P_{N-1,K-1,M-1} - 1 - P_{N-1,K-1,M-1}(Y_1 = Y_2)\right]$$

$$- (N-K)\left[2P_{N-1,K,M-1} - 1 - P_{N-1,K,M-1}(Y_1 = Y_2)\right]$$

$$+ (N-K)\, a_{L,L-1;N-1,K,M-1} + K\, a_{L-1;N-1,K-1,M-1}$$

$$= (2K-N)\left(2P_{N-1,K-1,M-1} - 1\right) + 2(N-K)\frac{\binom{K-1}{L-1}\binom{N-1-K}{L-1}}{\binom{N-1}{2L-1}}\left(\frac{K}{L} - 1\right)$$

b) $M = 2L + 1$

$$\frac{N}{M} E(D_{N,K,M}) = (2K-N)\left(2P_{N-1,K-1,M-1} - 1\right) + 2(N-K)\frac{\binom{K-1}{L-1}\binom{N-1-K}{L}}{\binom{N-1}{2L}}\left(\frac{K}{L} - 1\right)$$

Du point de vue calcul numérique on peut remarquer qu'ils se résument au calcul du deuxième terme dans deux cas particuliers :

- $K = \frac{N}{2}$ qui est étudié plus en détail ci-dessous
- $K < L$ ou $N - K < L$ (Le nombre de positions du plus petit des deux blocs est inférieur à la moitié du nombre de positions choisies).

Dans les autres cas le calcul par récurrence sur $P_{N,K,M}$ est d'autant plus simple que K (ou N – K) est voisin de L. Il est de toute manière plus simple que par intégration directe de la loi de $|Y_1 - Y_2|$.

Dans le cas particulier N = 2K les formules deviennent :

a) M = 2L

$$E(D_{N,K,M}) = \frac{M\binom{K-1}{L-1}}{\binom{N-1}{M-1}}\left(\frac{K}{L} - 1\right)$$

b) M = 2L + 1

$$E(D_{N,K,M}) = \frac{M\binom{K-1}{L-1}\binom{K-1}{L}}{\binom{N-1}{M-1}}\left(\frac{K}{L} - 1\right)$$

CONCLUSION.

Nous avons étudié diverses variables aléatoires définies par le modèle de l'équiprobabilité des choix de M positions sur N possibles formant B blocs de K cases.

A l'origine il s'agissait de recherche de structures sur les échantillons systématiques de variables binaires. Ainsi le coefficient d'autocorrélation étend le test des suites à une grille de points, l'indice de dispersion non paramétrique repère une éventuelle échelle privilégiée d'hétérogénéité, l'indice de contagion vraie mesure cette hétérogénéité entre deux blocs de comptages, la variable nombre de blocs vides repère les "trous" dans un processus 0 – 1, etc. . .

Notons cependant que leur emploi est d'intérêt plus large puisqu'il recouvre les analyses non paramétriques de comparaison de fréquences. En effet si $Y_1, \ldots, Y_B$ sont des variables binomiales, de paramètres respectifs $(K_1, p_1) \ldots (K_B, p_B)$, indépendantes, et si les p_i sont égaux, la distribution conditionnelle de $(Y_1, \ldots, Y_B)$ sachant $\sum_{i=1}^{B} Y_i = M$, est multi hypergéométrique. Dans toutes les conditions numériques où l'approximation par la loi multinomiale est peu sûre ou franchement mauvaise (les simulations de MEAD

(1974) sur l'indice de contagion vraie sont un exemple très explicite d'une telle rencontre) on pourra donc préférer les résultats exposés ici aux tests classiques de l'indice de dispersion, de Mac-Nemar, du χ^2 de la table de contingence 2 x n, etc. . .

Remarquons encore combien la technique des f.g. de f.g.p. se montre productive. Elle permet d'éviter, pour un groupe de problèmes, les raisonnements combinatoires. Une généralisation est en cours sur les processus qualitatifs à plus de deux états. L'analyse mathématique des protéines (suites d'acides aminés) et l'étude des structures spatiales des variables écologiques (échantillonnage systématique d'une variable qualitative) sollicitent une telle extension.

BIBLIOGRAPHIE.

CHESSEL D. (1975) – *Mesure de dispersion spatiale et méthodes d'échantillonnage.* D.G.R.S.T., Comité Equilibre et lutte biologique. Note Technique n° 3 – Ronéo 29 p.

CHESSEL D., DE BELAIR G. (1973) – *Mesure de la contagion vraie en échantillonnage par carrés dans l'analyse des populations végétales.* C.R.A.S. (D), 277, 1483-1486.

CHESSEL D., DEBOUZIE D. (1974) – *Mesure de la dispersion spatiale des végétaux en échantillonnage systématique par présence-absence.* C.R.A.S. (D), 278, 2027-2030.

CHESSEL D., CROZE J.P. (1976) – *Un indice de dispersion pour les mesures de présence-absence. Application à la répartition des animaux ou des plantes* (proposé pour publication).

CLIFF A.D., ORD J.K. (1973) – *Spatial autocorrelation.* Pion. Londres.

CONOVER W.J. (1971) – *Pratical non parametric statistics.* Wiley. Londres.

COX D.R., LEWIS P.A.W. (1966) – *The statistical analysis of series of events.* Methuen, Londres.

DAVID F.N., BARTON D.E. (1962) – *Combinatorial chance.* Griffin, Londres.

DAVID F.N., BARTON D.E. (1959) – *The dispersion of a number of species.* J. Roy. Statist. Soc. (B) 21, 190-194.

DEBOUZIE D. et Coll. (1975) – *Introduction à l'étude de la structure horizontale en milieu steppique. II Le traitement statistique des lignes de placettes contigües.* Oecol. Plant. 10 (3), 211-231.

FELLER W. (1943) – *On a general class of "contagious" distributions.* Ann. Math. Statist. 14, 389-400.

FELLER W. (1957) – *An introduction to probability theory and its applications.* Vol. I, 2nd edition. Wiley, Londres.

GAUTIER C. (1975) – *Les taux de mutation orientés : une mesure de l'évolution des protéines.* Année Biologique, sous presse.

GRANTHAM R. (1975) – *Le Code Génétique : Contribution à l'étude de son origine, sa structure et son fonctionnement. Vers un modèle moléculaire de l'évolution.* Thèse d'Etat, Lyon.

HOEL P.G. (1943) – *On indices of dispersion.* Ann. Math. Statist. 14, 155-62.

HOLLANDER M., WOLFE D.G. – *Nonparametric statistical methods.* Wiley, Londres.

JARRY M. (1971) – *Etude de la dispersion de la ponte chez Acanthoscelides obtectus SAY (Coléoptère-Bruchidae) dans un espace limité.* Thèse de troisième cycle, Lyon.

JOHNSON N.L., KOTZ S. (1969) – *Discrete distributions.* Mifflin, Boston.

KAUFMANN A. (1958) – *Introduction à la combinatorique en vue des applications,* Dunod, Paris.

LEGAY J.M., CHESSEL D. (1976) – *Description et analyse de la répartition des insectes dans une population végétale. Cas du doryphore sur pomme de terre* (proposé pour publication).

MEAD R. (1974) – *A test for spatial pattern at several scales using data from a grid of contiguous quadrats.* Biometrics, 30, 295-307.

MOUEZA M. (1975) – *Contribution à l'étude de la biologie de Donax trunculus L. Mollusque lamellibranche dans l'Algérois.* Thèse d'Etat, Marseille.

PRODON R. (1976) – *Le substrat facteur écologique et éthologique de la vie aquatique : observation et expériences sur des larves de Micropterna testacea et Cordulegaster annulatus.* Thèse de 3e cycle, Lyon.

PATIL G.C., JOSHI S.W. (1962) – *A dictionnary and bibliography of discrete distribution* Olivier-Boyd. Edinburgh.

RIORDAN J. (1958) – *An introduction to combinatorial analysis.* Wiley, Londres.

STITELER W.M., PATIL G.P. (1969) – *Variance to man ratio and Morosita's index as measures of spatial patterns in ecological populations.* Preprint n° 8, Department of statistics, Pennsylvania State University.

WALTER S.D. (1974) – *On the detection of household aggregation of disease.* Biometrics 30, 525-538.

REMARQUE : Le résultat III-3 et son extension aux blocs inégaux a été publié par WALTER (1974) dans le cadre d'applications à l'épidémiologie.