MATHÉMATIQUES ET SCIENCES HUMAINES

J. FAUCOUNAU

Note sur une loi de probabilité discrète méconnue

Mathématiques et sciences humaines, tome 52 (1975), p. 55-68

<http://www.numdam.org/item?id=MSH_1975__52__55_0>

Math. Sci. hum. (13^{e} année, n°52, *1975*, p.55-68)

NOTE SUR UNE LOI DE PROBABILITE DISCRETE MECONNUE

J. FAUCOUNAU

Les ouvrages sur le Calcul des Probabilités mentionnent généralement, au titre des lois de répartition d'une variable à valeurs discrètes, les lois de répartition usuelles : binomiale, géométrique, hypergéométrique et loi de Poisson. Notre intention est d'attirer ici l'attention sur une loi très simple, généralement méconnue (1), mais qui mériterait, à notre avis, de figurer en bonne place parmi les lois de probabilité discrètes citées ci-dessus. Cette loi, que nous désignerons du nom de "loi k-dimensionnelle d'une variable multinomiale dégénérée", présente un intérêt pédagogique non négligeable car elle introduit de façon naturelle les nombres de Stirling de deuxième espèce dont on sait le rôle important dans le calcul des différences finies.

Le problème auquel elle se rapporte remonte à de Moivre qui l'énonça pour la première fois sous la forme d'un problème de dés : Si on lance un dé parfaitement symétrique, possédant n faces distinctes (numérotées par exemple de 1 à n) , quelle est la probabilité de voir apparaître exactement k de ces n faces en m coups ? De Moivre en donna la solution en utilisant ce que nous désignons usuellement aujourd'hui du nom de "formule de Boole" relative à la composition des probabilités :

---

(1) ce qui lui a valu d'être "redécouverte" à plusieurs reprises, même récemment (Voir par exemple C.Craig [1] p 173)

$$P_r\{E_1 \cup E_2 \cup \ldots E_n\} = \sum_{i=1}^{n} P_r\{E_i\} - \sum_{i<j} \sum P_r\{E_i \cap E_j\} + \sum_{i<j<k} \sum \sum P_r\{E_i \cap E_j \cap E_k\} - \ldots\ldots$$

Choisissant comme événement E_i la non-apparition de la face i (événement dont la probabilité est de (1-1/n) lors de chaque lancer), il trouva ainsi la formule :

$$P = \sum_{i=1}^{k} C_k^i (-1)^i (1-i/n)^m$$

où P est la probabilité de voir apparaître (au moins une fois) k faces déterminées.

Laplace [8] reprit et développa le même problème sous la forme de la "loterie lorraine" étudiée par Trembley [12] et Euler [3]. Dans cette loterie, n billets sont distribués et tirés au sort avec remise , d'où il résulte un nombre k de billets gagnants (compris entre 1 et n). Enoncé sous cette forme, le problème revient à chercher la probabilité pour que m boules tombant au hasard dans n boîtes identiques, k de ces boîtes soient occupées. Cette remarque explique l'origine des termes d'"occupancy" et de "besetzung" donnés respectivement au problème par les auteurs anglo-saxons et allemands. Plus près de nous, différents auteurs parmi lesquels nous citerons W.L. Stevens [11],C.C. Craig [1] , Ch. Jordan [7] (p 178) , F.N. David [2] et N.L. Johnson [6] ont traité de cette même question.

Nous l'exposerons pour notre part comme suit :

Soit une urne contenant N boules dont une proportion
p_a est constituée de boules de couleur a
p_b -"- -"- b
.....
p_z -"- -"- z
le nombre de couleurs différentes étant n au total.

On prélève dans cette urne par tirages successifs avec remise un échantillon de m boules.

Soit X_1 , X_2 , ... X_n les nombres de boules de couleur

a,b,...z figurant dans l'échantillon ($X_i \geq 0$).

Dans la loi multinomiale , on étudie la répartition de la variable aléatoire $\vec{X}$ à n composantes discrètes $X_1, X_2..X_n$. En d'autres termes, on cherche la probabilité pour que $\vec{X}$ soit égal à $\vec{x}$ donné , $\vec{x}$ étant un vecteur d'un espace R_n à n dimensions dont les coordonnées sont liées par la relation :

$$(1)\quad x_1 + x_2 + ... + x_n = m$$

les x_i étant des entiers non négatifs (Dans l'espace R_n , l'extrémité du vecteur $\vec{x}$ se déplace dans l'hyperplan $\sum x_i = m$).

On montre facilement que :

$$(2)\quad P_r\{X_1=x_1 \ , \ X_2=x_2, \ ...X_n=x_n\} = \frac{m!}{x_1!x_2!..x_n!}\ p_a^{x_1}p_b^{x_2}..p_z^{x_n}$$

La loi k-dimensionnelle de la variable multinomiale s'introduit lorsque l'on s'intéresse non aux composantes X_i elles-mêmes, mais au nombre K de composantes non-nulles de $\vec{X}$.

En d'autres termes, on cherche la probabilité pour que l'échantillon comporte k ensembles non vides de boules de couleurs différentes (k étant un entier tel que $1 \leq k \leq m$ ou n suivant que m est inférieur ou supérieur à n).

On notera P(m,k) la probabilité que K = k pour un échantillon de taille m .

Nous examinerons tout d'abord quelques cas particuliers:

Si on tire une seule boule (m = 1) , cette boule sera obligatoirement de l'une des couleurs a,b,...z. Par composition des probabilités, on a par conséquent :

$$(3)\quad P(1,1) = p_a + p_b + ... + p_z = 1$$

Si on tire deux boules (m = 2) , on se trouvera en présence de deux possibilités :

a)- les deux boules sont de la même couleur. Alors :

$$(4)\quad P(2,1) = p_a^2 + p_b^2 + \ldots + p_z^2 = \sum p_i^2$$

b)- les deux boules sont de couleurs différentes :

$$(5)\quad P(2,2) = p_a p_b + p_a p_c + \ldots + p_y p_z = \sum p_i p_j \qquad i \neq j$$

P(2,1) et P(2,2) correspondent donc respectivement à la somme des termes en p_i^2 et en $p_i p_j$ du développement de $(\sum p_i)^2$.

La formule se généralise aisément pour m quelconque et on a :

(6) P(m,k) = Somme des termes en $p_a^{x_1} p_b^{x_2} \ldots p_z^{x_n}$ du développement de $(\sum p_i)^m$

pour lesquels chaque ensemble $x_1, x_2, \ldots x_n$ est une racine en entiers non négatifs de l'équation

$$(1)\quad x_1 + x_2 + \ldots + x_n = m$$

le nombre de termes <u>non nuls</u> de cet ensemble étant égal à k.

La formule (6) donne théoriquement le moyen de calculer P(m,k) en fonction des p_i.

Si on note $\sigma_1, \sigma_2, \ldots, \sigma_n$ les fonctions symétriques fondamentales des p_i , soit :

$$\sigma_1 = p_a + p_b + \ldots + p_z$$
$$\sigma_2 = p_a p_b + p_a p_c + \ldots + p_y p_z$$
$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$
$$\sigma_n = p_a p_b p_c \ldots p_z$$

on pourra exprimer P(m,k) en fonction des σ_i . Par exemple , pour m = 4 et n $\geqslant$4 , on trouve :

$$P(4,1) = 1 - 4\,\sigma_2 + 4\,\sigma_3 - 4\,\sigma_4 + 2\,\sigma_2^2$$

$$P(4,2) = 4\sigma_2 - 16\sigma_3 + 28\sigma_4 - 2\sigma_2^2$$
$$P(4,3) = 12\sigma_3 - 48\sigma_4$$
$$P(4,4) = 24\sigma_4$$

Pour $n < 4$ (par exemple $n = 3$) les mêmes formules sont valables à condition d'annuler les termes en σ de rang supérieur à n (σ_4 dans l'exemple cité).

En pratique, le calcul se complique très vite dès que la valeur de m dépasse quelques unités.

Nous étudierons maintenant le cas particulier, qui est l'objet de la présente note, où la proportion de boules d'une couleur est la même quelle que soit la couleur considérée (loi multinomiale "dégénérée") :

$$p_a = p_b = ... = p_z = 1/n$$

La probabilité P(m,k) pour qu'un tirage de m boules fournisse k ensembles de couleurs différentes se calcule alors aisément par un raisonnement de récurrence :

Pour qu'un tirage de m boules fasse apparaître k ensembles de couleurs différentes, il faut et il suffit :
- soit que le tirage (m-1) ait fait apparaître k ensembles de couleurs différentes et que le $m^{ième}$ tirage fasse apparaître une boule de l'une des couleurs déjà tirées
- soit que le tirage (m-1) ait fait apparaître (k-1) ensembles de couleurs différentes et que le $m^{ième}$ tirage fasse apparaître une boule d'une couleur nouvelle.

Par composition des probabilités, on a donc :

$$(7)\quad P(m,k) = P(m-1,k)\frac{k}{n} + P(m-1,k-1)\frac{n-k+1}{n}$$

Si nous posons :

$$(8)\quad P(m,k) = S_m^k \frac{n(n-1)....(n-k+1)}{n^m}$$

(8 bis) $= S_m^k \frac{n!}{(n-k)!} \cdot \frac{1}{n^m}$

il est facile de voir que la relation (7) entraine pour les coefficients S_m^k la relation de récurrence :

(9) $S_m^k = k\, S_{m-1}^k + S_{m-1}^{k-1}$, avec

(10) $S_1^1 = S_2^2 = \ldots = S_m^m = 1$

Cette relation est l'une des définitions relatives aux "nombres de Stirling de deuxième espèce", nombres qui s'introduisent en Mathématiques dans diverses questions, dont en particulier le Calcul des différences finies .(1)

Il est possible de retrouver la formule (8) en utilisant une deuxième définition des nombres de Stirling de deuxième espèce : On peut en effet définir S_m^k comme le nombre de façons de ranger m objets distincts dans k cases indiscernables , chaque case contenant au moins un objet ($k \leqslant m$).

Il est facile de voir que cette définition correspond à la précédente : Considérons en effet les deux ensembles de rangements suivants :

(A):ensemble des rangements de (m-1) objets dans k cases (aucune case vide)

(B):ensemble des rangements de (m-1) objets dans (k-1) cases (-id-)

Ajoutons un $m^{ième}$ objet:

Il y a k possibilités de l'ajouter à chaque rangement de (A) puisque on peut l'ajouter dans l'une quelconque des k cases

Il n'y a par contre qu'une seule façon de l'ajouter à l'un des rangements (B) de manière à créer une $k^{ième}$ case : C'est de mettre le $m^{ième}$ objet seul dans la $k^{ième}$ case (rappelons que les cases sont indiscernables, c'est-à-dire que l'

---

(1) Voir page suivante

ordre dans lequel on les considère est sans importance).

On obtient donc la relation de récurrence cherchée :

(9) $S_m^k = k\, S_{m-1}^k + S_{m-1}^{k-1}$, avec comme il est facile de le voir :

(10) $S_1^1 = S_2^2 = \ldots = S_m^m = 1$

Schématisons alors un tirage m par une correspondance telle que :

1 2 3 4 5 .....
a b a a c .....

signifiant qu'une boule de couleur a est sortie aux premier, troisième et quatrième tirages, une boule de couleur b au second, etc...

Les événements constitués par les suites ordonnées telles que abaac... sont les événements équiprobables. Leur nombre total est n^m puisqu'à chaque tirage élémentaire il existe n possibilités d'amener une couleur.

Soit k le nombre de couleurs distinctes obtenues parmi les n possibles. Créons k cases et mettons dans ces cases les numéros repères des tirages : Exemple : La case a contiendra 1,3,4,... , la case b : 2,..., la case c : 5,..., etc.... Il y a S_m^k façons de le faire si on ne discerne pas les cases et $k!S_m^k$ façons si on les distingue. Le nombre de cas favorables c'est-à-dire le nombre de façons d'obtenir k couleurs distinctes lors du tirage m sera donc :

$k!\, S_m^k\, C_n^k$, d'où la probabilité cherchée :

---

(1) (renvoi page précédente) Utilisant les symboles usuels dans cette branche des Mathématiques, certains auteurs emploient la notation $\Delta^k 0^m$ qui exprime, par définition, le résultat obtenu lorsque l'on attribue à x la valeur zéro dans la différence $k^{ième}$ de x^m. On a ainsi : $k!\, S_m^k = \Delta^k 0^m$.

$$(8\ \text{ter})\quad P(m,k) = \frac{k!\ S_m^k\ C_n^k}{n^m}$$

En remplaçant C_n^k par sa valeur $\frac{n!}{k!(n-k)!}$, on retrouve les formules (8) et (8 bis).

Il existe des tables relatives aux nombres de Stirling de deuxième espèce pour m = 1 à m = 25 , ce qui permet pour les faibles valeurs de m , de calculer sans difficulté les P(m,k) à partir de la formule (8 bis).

Pour les grandes valeurs de m , on utilisera la formule d'approximation établie par L. Moser et M. Wyman [10] dont le premier terme est :

$$S_m^k \sim \frac{m!(\exp(r) - 1)^k}{2r^m\ k!\ (\pi krh)^{1/2}}$$

où r est solution de :

$$r(1 - e^{-r})^{-1} = m/k$$

et $h = e^r.(e^r-1-r)/2(e^r-1)^2$

Les nombres de Stirling de deuxième espèce possèdent un certain nombre de propriétés remarquables sur lesquelles nous n'insisterons pas, nous contentant de rappeler les principales:

a)- Pour m fixe, on a évidemment puisque les P(m,k) sont des probabilités :

$$\sum_{k=1}^{m} P(m,k) = 1$$

(Cette formule est valable même si $n < m$. Dans ce dernier cas, P(m,k) = 0 pour tout $k > n$ du fait de la présence d'un facteur nul au numérateur).

$$\text{soit :}\ (11)\quad n^m = \sum_{k=1}^{m} S_m^k\ n(n-1)\ldots(n-k+1)$$

b)- Soit p(x) un polynôme de degré m . On peut l'écrire sous la forme :

$$(12)\quad p(x) = a_0 + a_1(x-x_0) + a_2(x-x_0)(x-x_1)+\ldots+a_m(x-x_0)\ldots(x-x_{m-1})$$

Faisons successivement $x = x_0$, $x = x_1$, $x = x_2$, etc... Il vient :

$$
\begin{aligned}
& p(x_0) = a_0 \\
(13)\quad & p(x_1) = a_0 + a_1(x-x_0) \\
& p(x_2) = a_0 + a_1(x_2-x_0) + a_2(x_2-x_0)(x_2-x_1) \\
& \cdots\cdots\cdots\cdots\cdots\cdots
\end{aligned}
$$

Les relations (13) permettent de calculer de proche en proche les coefficients a_0, a_1, a_2, ..etc.. en fonction de $p(x_0)$, $p(x_1)$, etc...

Choisissons maintenant comme polynôme $p(x) = x^m$ et assignons aux variables x_i les valeurs respectives :

$$x_0 = 0 \ , \ x_1 = 1 \ , \ \ldots\ , \ x_{m-1} = m-1$$

Nous voyons que, d'après l'identité (11) , les coefficients a_0, a_1, ..., a_m s'identifient aux nombres de Stirling de deuxième espèce, l'expression (12) devenant :

$$(11\text{ bis})\quad x^m = S_m^1\, x + S_m^2\, x(x-1) + \ldots + S_m^m\, x(x-1)..(x-m+1)$$

soit : $a_i = S_m^i$

Les formules (13) permettent donc de calculer les valeurs des S_m^k. Il vient en particulier :

$$(14)\quad S_m^1 = 1 \ , \ S_m^2 = 2^{m-1} - 1 \ , \ S_m^3 = \frac{3^m}{6} - \frac{2^m-1}{2} \ , \ \text{etc...}$$

La formule générale est facile à obtenir grâce aux identités (13) écrites sous la forme :

$$
\begin{aligned}
1^m &= S_m^1 \\
2^m &= \frac{2!}{1!} S_m^1 + 2! S_m^2 \\
3^m &= \frac{3!}{2!} S_m^1 + \frac{3!}{1!} S_m^2 + 3! S_m^3 \\
&\cdots\cdots\cdots\cdots\cdots\cdots \\
k^m &= \frac{k!}{(k-1)!} S_m^1 + \frac{k!}{(k-2)!} S_m^2 + \ldots\ldots + k! S_m^k
\end{aligned}
$$

Multiplions la première ligne par C_k^1 , la seconde par $-C_k^2$, la troisième par C_k^3 , etc... et additionnons membre à membre. Le premier membre donne la somme :

$$1^m C_k^1 - 2^m C_k^2 + \ldots\ldots + (-1)^{k+1}\, k^m C_k^k$$

Au second membre, grâce à une propriété bien connue des C_k^j , tous les coefficients des S_m^j sont nuls, à l'exception du dernier : S_m^k . On a donc :

$$(16) \quad S_m^k = \frac{1}{k!} \sum_{j=1}^{k} (-1)^{k-j}\, j^m\, C_k^j$$

Revenons maintenant à la loi k-dimensionnelle d'une variable multinomiale dégénérée, représentée par la loi de distribution (8).

On voit facilement que cette loi de distribution revêt des formes très différentes suivant que m est petit devant n , équivalent à n ou très grand devant n .

Dans le premier et le dernier cas, les courbes enveloppes joignant les sommets du spectre des probabilités P(m,k) ont la forme de "dents de scie".

Lorsque m se rapproche de n , la courbe se symétrise et prend une forme "en cloche", la symétrie maximum ayant lieu pour m = n .

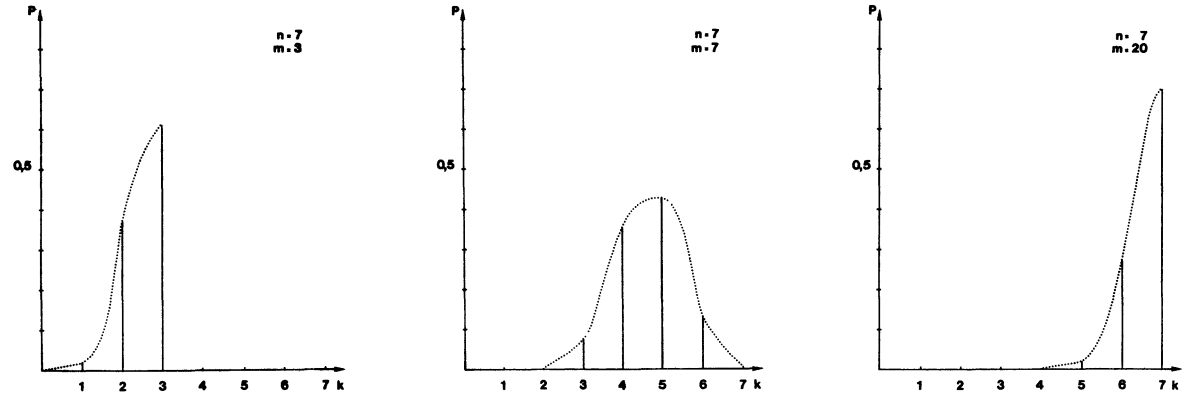

Figure 1 : Diverses formes de la loi de distribution

Pour déterminer les moments de la loi de répartition de la variable aléatoire k , le plus aisé est de partir des mo-

ments factoriels qui ont une expression très simple (voir la formule (19) ci-dessous).

Nous utiliserons pour alléger l'écriture la notation commode introduite par N. Vandermonde :

$$x^{(m)} = x(x-1)\ldots(x-m+1)$$

qui conduit à la formule connue, parallèle à la formule de dérivation d'une puissance :

$$\Delta x^{(m)} = m x^{(m-1)}$$

où Δ est l'opérateur différence :

$$\Delta f(x) = f(x+1) - f(x)$$

On a alors :

$$(8\text{ ter}) \quad P(m,k) = S_m^k \frac{n^{(k)}}{n^m} \qquad \text{et :}$$

$$(11\text{ ter}) \quad x^m = \sum_{k=1}^{m} S_m^k x^{(k)}$$

En appliquant l'opérateur Δ à x^m, il vient successivement :

$$\Delta(x^m) = \sum_{k=1}^{m} k\, S_m^k\, x^{(k-1)}$$

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

$$\Delta^s(x^m) = \sum_{k=1}^{m} k^{(s)}\, S_m^k\, x^{(k-s)}$$

En remplaçant x par (n-s), on obtient la relation :

$$(18) \quad \Delta^s (n-s)^m = \sum_{k=1}^{m} k^{(s)}\, S_m^k\, (n-s)^{(k-s)}$$

Le $s^{\text{ième}}$ moment factoriel de la distribution de la variable aléatoire k est :

$$\mu(s) = E\{k^{(s)}\} = \sum_{k=1}^{m} S_m^k \frac{n^{(k)}}{n^m} k^{(s)}$$

(si $m > n$, les termes dont l'indice est plus grand que n sont nuls).

Multiplions et divisons par $n^{(s)}$. En remarquant que $\frac{n^{(k)}}{n^{(s)}} = (n-s)^{(k-s)}$, il vient :

$$(s) = \frac{n^{(s)}}{n^m} \sum_{k=1}^{m} S_m^k \, (n-s)^{(k-s)} \, {}_k(s)$$

soit en tenant compte de la relation (18) :

$$(19) \qquad \mu_{(s)} = \frac{n^{(s)}}{n^m} \, \Delta^s (n-s)^m$$

L'<u>espérance mathématique</u> de k s'obtient en faisant s=1 dans la formule ci-dessus. D'où :

$$(20) \quad E\{k\} = \frac{1}{n^{m-1}} \, \Delta (n-1)^m = \frac{n^m - (n-1)^m}{n^{m-1}}$$

La <u>variance</u> s'obtient en faisant s=2 et en tenant compte de la relation entre variance et moments factoriels :

$$\text{var}\{k\} = \mu_{(2)} - \mu_{(1)}^2 + \mu_{(1)}$$

Il vient ainsi :

$$(21) \quad \text{var}\{k\} = n(1-1/n)^m + n(n-1)(1-2/n)^m - n^2(1-1/n)^{2m}$$

<u>Cas particuliers</u> :

a)- Si n est fixe et $m \to \infty$

$E\{k\} = n - (1-1/n)^m . n$ n , le terme $(1-1/n)^m < 1$

tendant vers zéro pour $m \to \infty$

$$\text{var}\{k\} = n(1-1/n)^m + (1-2/n)^m . n(n-1) - n^2(1-1/n)^{2m}$$
$$\to 0$$

Le spectre des probabilités tend, comme il était intuitivement évident, à se réduire à la seule composante P(m,n) qui tend vers 1.

b)- Si m est fixe et $n \to \infty$

$E\{k\} = n[1 - (1-1/n)^m] \to m$

$\text{var}\{k\} \to 0$

Le spectre des probabilités tend à se réduire à la seule composante P(m,m) qui tend vers 1.

c)- Si m et n tendent vers l'infini, le rapport $m/n = \alpha$ restant constant :

$$E\{k\} = n\,[\,1 - (1-1/n)^n\,] \quad \to \quad n(1 - e^{-\alpha})$$

$$\mathrm{var}\{k\} = n(1-1/n)^n + n(n-1)(1-2/n)^n - n^2(1-1/n)^{2n}$$

$$\to \quad n[\,e^{-\alpha} - (\alpha + 1)\,e^{-2\alpha}]$$

Nous terminerons en signalant que la loi k-dimensionnelle s'introduit de façon naturelle dans un certain nombre de questions en numismatique, lexicologie, botanique, biologie, etc... Nous citerons à titre d'exemple son application dans le recensement d'une population d'insectes volants (papillons). Pour estimer le nombre de papillons vivant dans une aire donnée, on fait m captures au hasard (en prenant soin que chaque papillon ait la même chance d'être attrapé à chaque capture), on marque l'insecte capturé et on le relâche aussitôt. Un papillon peut ainsi être attrapé zéro, une, deux, trois, etc....fois. De la statistique obtenue, on peut déduire une estimation du nombre total de papillons dans l'aire considérée.(Pour plus de détail voir C.C. Craig [1]et I.J. Good [4] et [5]). On remarquera que le problème revient à affecter une case à chaque papillon et à laisser tomber au hasard m boules (chaque boule correspondant à une capture) dans ces cases.

BIBLIOGRAPHIE

[1] - CRAIG C.C.,"On the utilization of marked specimens in estimating populations of flying insects", Biometrika, 40 (1953), 170-176

[2] - DAVID F.N. & BARTON D.E., Combinatorial Chance, London, Charles Griffin & Company Limited, 1962,243-286

[3] - EULER L.,Opuscula Analytica , vol II (1785), 331-346

[4] - GOOD I.J.,"The population frequencies of species.." , Biometrica, 40, (1953), 237-264

[5] - GOOD I.J. & TOULMIN G.H.,"The number of new species...."

Biometrica, 43, (1956), 45-63

[6]- JOHNSON N.L. & KLOTZ S., Discrete Distributions ,Boston Houghton Mifflin Co , (1969), 251-253

[7]- JORDAN Ch., Calculus of Finite Differences, New-York , Chelsea Publishing Co , (1965), 168-179

[8]- LAPLACE P.S., Théorie Analytique des Probabilités,Paris, Courcier, (1812), 192-201

[9]- De MOIVRE A. , Doctrine of Chances , London,(1718) , Problème XXXIX

[10]- MOSER L. & WYMAN M. , Duke Mathematical Journal ,(1958) 29-43

[11]- STEVENS W.L. , "Significance of grouping",Annals of Eugenics, 8, (1937), 57-69

[12]- TREMBLEY ,"Recherches sur une question relative au calcul des probabilités", Mémoires de l'Accad...Berlin (1794/1795), -9-108