REVUE DE STATISTIQUE APPLIQUÉE

M. D. INDJOUDJIAN

Estimation de certaines probabilités

Revue de statistique appliquée, tome 13, nº 4 (1965), p. 21-31
<http://www.numdam.org/item?id=RSA_1965__13_4_21_0>

ESTIMATION DE CERTAINES PROBABILITÉS

M. D. INDJOUDJIAN

Professeur de Statistique Mathématique
à l'Institut de Statistique de l'Université de Paris

INTRODUCTION

Les problèmes classiques d'estimation portent sur l'estimation du ou des paramètres inconnus intervenant dans une loi de probabilité de type connu.

Exemples : Estimation de la probabilité d'un succès individuel dans le cas d'une loi binomiale, du paramètre d'une loi de Poisson, de la moyenne ou de la variance d'une loi de Laplace-Gauss, etc.

Or, dans les applications, en est souvent intéressé non par une estimation, disons sans distorsion et si possible de variance minimale, d'un tel paramètre mais par celle - avec les mêmes qualités - de la probabilité $P = \Pr\{a \leqslant X \leqslant b\}$ que la variable aléatoire considérée X soit comprise entre deux liùites données a et b. Certes P est une fonction du ou des paramètres dont on suppose connus des estimateurs sans distorsion, mais la fonction de ce ou de ces estimateurs, si elle fournit un estimateur de P, n'en fournit pas, en général, un estimateur sans distorsion.

La présente note a pour but d'attirer l'attention sur une méthode générale [1] fournissant dans des cas étendus un estimateur sans distorsion de variance minimale de P.

Le paragraphe 1 sera consacré à un bref exposé de cette méthode. Le paragraphe 2 à l'énoncé de résultats déjà en partie publiés par divers auteurs, et qui, ici complétés, nous paraissent avoir une réelle utilité pratique. Dans le paragraphe 3, il sera vérifié sur quelques uns des exemples du paragraphe 2 que les estimations sont bien sans distorsion, les autres vérifications étant laissées au lecteur ou résultant du paragraphe 4 où, sur trois cas pris comme exemples, la construction desdits estimateurs sera effectuée selon la méthode du paragraphe 1. Enfin, une annexe sur la loi de l'alternative généralisée répétée établira d'une manière qui parait nouvelle une formule générale commode.

1. Soit ν le nombre des observations, parmi les observations indépendantes $x_1, x_2, \ldots, x_n$ sur une population-mère donnée, qui se trouvent dans l'intervalle donné [a, b].

Si $$P = \Pr\{a \leqslant X \leqslant b\}, \quad E\{\nu\} = nP \text{ ou } E\left\{\frac{\nu}{n}\right\} = P.$$

S'il existe un estimateur exhaustif t du paramètre θ intervenant dans la loi de probabilité de X, il résulte des résultats de Blackwell [3], Rao [6] et Lehmann et Scheffé [5] que

$$P^* = E\left\{\frac{\nu}{n} \mid t\right\} = \frac{1}{n} E\{\nu \mid t\}$$

est un estimateur sans distorsion de variance minimale de P.

Or $E\{\nu|t\} = n \Pr\{a \leqslant x_j \leqslant b \mid t\}$ quel que soit j pris de 1 à n, donc l'estimateur cherché est

$$\underline{P^* = \Pr\{a \leqslant x_j \leqslant b \mid t\}} \qquad (1)$$

Bien entendu, j devra être choisi indépendamment des résultats mêmes des observations, disons $j = 1$.

2. Résultats (estimateurs sans distorsion de variance minimale)

a) Loi binomiale (ou de l'alternative répétée).

Un estimateur de $P = \Pr\{r \text{ succès en } n \text{ épreuves}\} = C_n^r p^r q^{n-r}$ est, si R désigne le nombre observé de succès en N épreuves :

$$\underline{P = \frac{C_n^r}{N^{[n]}} R^{[r]} (N - R)^{[n-r]} = \frac{C_R^r C_{N-R}^{n-r}}{C_N^n}} \qquad (2)$$

où la notation du crochet en exposant est celle définie à l'annexe.

Exemple : Pour $n = 3$, $r = 2$; $N = 10$, $R = 3$, $P^* = \frac{3.3.2.7}{10.9.8} = 0,175$, tandis que l'estimation (non sans distorsion) $P' = 3p'^2 q'$ obtenue à partir de l'estimation sans distorsion de p, soit $p' = \frac{R}{N} = 0,3$ est

$$P' = \frac{3.3^2.7}{10^3} = 0,189.$$

b) Loi de l'alternative répétée généralisée.

Un estimateur de $P = \Pr\{n_1$ succès de la première catégorie ; ..., n_i de la $i^{\text{ème}}$, ..., n_k de la $k^{\text{ième}}\} = n! \prod_{i=1}^{k} \frac{p_i^{n_i}}{n_i!}$,

si en N épreuves on observe N_1 succès de la première catégorie, etc. est

$$\underline{P^* = \frac{n!}{N^{[n]}} \prod_{i=1}^{k} \frac{N_i^{[n_i]}}{n_i!} = \frac{\prod_{i=1}^{k} C_{N_i}^{n_i}}{C_N^n}} \qquad (3)$$

c) Loi de Poisson

Un estimateur de $P_k = \Pr\{X = k\} = e^{-\lambda}\frac{\lambda^k}{k!}$,

si S désigne la somme des observations $x_1, \ldots, x_n$, est

$$P_k^* = \frac{S^{[k]}}{k!\ n^k}\left(1 - \frac{1}{n}\right)^{S-k} = C_S^k \left(\frac{1}{n}\right)^k \left(1 - \frac{1}{n}\right)^{S-k} \tag{4}$$

En particulier l'estimateur de $P_o = Pr\{X = 0\} = e^{-\lambda}$ est

$$P_o^* = \left(1 - \frac{1}{n}\right)^S \tag{4'}$$

Remarque : La simplicité des formules (1) à (4) est remarquable, alors même qu'elles sont à préférer aux formules qui, déduites des formules classiques d'estimation des paramètres, ne donneraient pas de P une estimation sans distorsion. L'analogie formelle de ces expressions avec celle de probabilités hypergéométriques ou binomiale est à noter.

d) Loi exponentielle de densité $\frac{1}{\theta}\ e^{-\frac{x}{\theta}}$ pour $x > 0$, nulle pour $x < 0$ étant supposé inconnu, un estimateur de la probabilité de "survie" au-delà du "temps" $\xi > 0$ donné $P = Pr\{X > \xi\}$ est

$$P^* = \begin{cases} 0 & \text{si } S < \xi \\ \left(1 - \frac{\xi}{S}\right)^{n-1} & \text{si } S > \xi \end{cases} \tag{5}$$

S désignant la somme des n observations $x_1, \ldots, x_n$.

Un estimateur de $Pr\{a \leqslant X \leqslant b\}$ est égal à 0 si $S < a$, à

$\left(1 - \frac{a}{S}\right)^{n-1}$ si $a \leqslant S \leqslant b$ et à $\left(1 - \frac{a}{S}\right)^{n-1} - \left(1 - \frac{b}{S}\right)^{n-1}$ si $S > b$

e) Loi exponentielle de densité $\frac{1}{\theta}\ e^{-\frac{x-\mu}{\theta}}$ pour $x > \mu$, nulle pour $x < 0$. θ et μ étant supposés inconnus, un estimateur de la probabilité de survie au-delà de ξ, peut s'obtenir (4) à partir du fait que

$$(x_1^*,\ \overline{x} - x_1^* = n\,S - x_1^*)$$

- où x_1^* est la plus petite des n observations $x_1, \ldots, x_n$ de somme S - constitue un estimateur exhaustif de (μ, θ).

C'est

$$P^* = \begin{cases} 1 & \text{si } \xi < x_1^* \\ \left(1 - \frac{1}{n}\right)\left(1 - \frac{\xi - x_1^*}{S - nx_1^*}\right)^{n-2} & \text{si } x_1^* \leqslant \xi \leqslant S - (n-1)\,x_1^* \\ 0 & \text{si } \xi > S - (n-1)\,x_1^* \end{cases} \tag{6}$$

f) Loi de Laplace-Gauss de variance σ^2 connue et de moyenne m inconnue.

Un estimateur de $P = Pr\{a < X < b\}$ est, en désignant par G(u) la fonction de répartition de la loi de Laplace-Gauss réduite :

$$P^* = G\left(\frac{b-\overline{x}}{\sigma}\sqrt{\frac{n}{n-1}}\right) - G\left(\frac{a-\overline{x}}{\sigma}\sqrt{\frac{n}{n-1}}\right) \tag{7}$$

g) Loi de Laplace-Gauss de moyenne m et de variance σ^2 inconnues. Un estimateur de $P = Pr\{a < X < b\}$ est

$$P^* = \int_{\mathcal{J}} \frac{(1-u^2)^{\frac{n-4}{2}}}{B\left(\frac{1}{2}, \frac{n}{2}-1\right)} du \tag{8}$$

étendue à l'intervalle $\mathcal{J} = [-1, +1] \cap [u_a, u_b] = [\alpha', \beta']$ où

$$u_a = \frac{a-\overline{x}}{s\sqrt{n-1}} \quad \text{et} \quad u_b = \frac{b-\overline{x}}{s\sqrt{n-1}}$$

Le calcul de P^* peut s'effectuer, comme le montre le changement de variable $v = \frac{1+u}{2}$, par recours aux tables de Pearson de la fonction eulérienne incomplète de première espèce $I_p\left(\frac{n}{2}-1, \frac{n}{2}-1\right)$, en remarquant que

$$P^* = I_\alpha - I_\beta \tag{9}$$

où

$$\alpha = \frac{1+\alpha'}{2} \quad \text{et} \quad \beta = \frac{1+\beta'}{2}$$

3. Vérifions dans les cas b, c et f pris à titre d'exemples que les estimateurs indiqués sont sans distorsion. La construction de tels estimateurs (sans distorsion et de variance minimale) sera exposée au § 4 à partir des principes donnés en 1, dans les cas a, d et g.

Cas b. D'après la formule fondamentale (5) de l'annexe appliquée à N épreuves - dont on constate qu'elles se traduisent par N_i succès de la ième catégorie (i = 1 à k) - et à des exposants pris égaux aux nombres n_i qui sont ici donnés :

$$E\left\{\prod_{i=1}^{k} N_i^{[n_i]}\right\} = N^{[n]} \prod_{i=1}^{k} p_i^{n_i}$$

égalité équivalente à celle qu'il s'agit d'établir : $E\{P^*\} = P$.

Cas c. La somme S obéit à une loi de Poisson de paramètre $n\lambda$. Donc

$$E\{P_k^*\} = \sum_{S=0}^{+\infty} \frac{S^{[k]}}{k!\, n^k}\left(1-\frac{1}{n}\right)^{S-k} e^{-n\lambda}\frac{(n\lambda)^S}{S!}$$

$$= e^{-\lambda}\frac{\lambda^k}{k!}\sum_{S=k}^{+\infty}\frac{S^{[k]}}{n^k}\frac{(n-1)^{S-k}}{n^{S-k}} e^{-(n-1)\lambda} n^S \frac{\lambda^{S-k}}{S!}$$

$$= P_k \sum_{S-k=0}^{+\infty} e^{-(n-1)\lambda}\frac{[(n-1)\lambda]^{S-k}}{(S-k)!} = P_k$$

Remarque : Considérons en particulier l'efficacité de l'estimateur P_o^* de $P_o = e^{-\lambda}$ par rapport à l'estimateur - également sans distorsion - qu'est $P_o' = \frac{n_o}{n}$ où n_o est le nombre des observations $x_1, \ldots, x_n$ égales à zéro.

D'une part

$$V\{P_o'\} = \frac{P_o(1 - P_o)}{n} = \frac{e^{-\lambda}(1 - e^{-\lambda})}{n}$$

D'autre part

$$E\{P_o^{*2}\} = \sum_{S=0}^{+\infty} \left(1 - \frac{1}{n}\right)^{2S} e^{-n\lambda} \frac{(n\lambda)^S}{S!} = e^{-n\lambda} \sum_{S=0}^{\infty} \frac{\left[n\left(1 - \frac{1}{n}\right)^2 \lambda\right]^S}{S!}$$

$$= e^{-n\lambda} \sum_{S=0}^{\infty} \frac{\left[\left(n - 2 + \frac{1}{n}\right)\lambda\right]^S}{S!} = e^{-n\lambda + \left(n - 2 + \frac{1}{n}\right)\lambda} = e^{-2\lambda + \frac{\lambda}{n}}$$

d'où $$V\{P_o^*\} = e^{-2\lambda + \frac{\lambda}{n}} - e^{-2\lambda} = e^{-2\lambda}\left(e^{\frac{\lambda}{n}} - 1\right)$$

L'efficacité relative $\mathcal{E}$ de P_o' par rapport à P_o^* est donc

$$\mathcal{E} = \frac{e^{-2\lambda}\left(e^{\frac{\lambda}{n}} - 1\right)}{e^{-\lambda}(1 - e^{-\lambda})} n = \frac{e^{\frac{\lambda}{n}} - 1}{e^{\lambda} - 1} n$$

inférieure à un et tendant vers $\frac{\lambda}{e^{\lambda} - 1}$ lorsque $n \longrightarrow +\infty$ pour λ fixe, limite d'autant plus petite que λ est plus grand.

Cas f. En désignant par $g(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}}$ la densité de probabilité de la loi de Laplace-Gauss réduite, $P^* = \int_\alpha^\beta g(u)\, du$ o $\alpha = \frac{a - \bar{x}}{\sigma} \sqrt{\frac{n}{n - 1}}$ et $\beta = \frac{b - \bar{x}}{\sigma} \sqrt{\frac{n}{n - 1}}$.

Or, P^* est fonction de la variable aléatoire x telle que $y = \frac{\bar{x} - m}{\sigma}\sqrt{n}$ obéit à une loi de Laplace-Gauss réduite.

Donc $E\{P^*\} = \int_{-\infty}^{+\infty} P^* g(y)\, dy$ où, dans l'expression de P^*, $\bar{x}$ est supposée exprimée en fonction de y. Il résulte, en remplaçant u par z au moyen du changement de variable

$$z\sqrt{n} = u\sqrt{n - 1} + y, \text{ que}$$

$$E\{P^*\} = \int_{-\infty}^{+\infty} g(y)\, dy \int_{\frac{a-m}{\sigma}}^{\frac{b-m}{\sigma}} g\left(\frac{z\sqrt{n} - y}{\sqrt{n - 1}}\right) \sqrt{\frac{n}{n - 1}}\, dz$$

$$= \frac{1}{2\pi} \sqrt{\frac{n}{n - 1}} \int_{-\infty}^{+\infty} dy \int_{\frac{a-m}{\sigma}}^{\frac{b-m}{\sigma}} e^{-\frac{Q}{2(n-1)}}\, dz$$

où $Q = (n-1)y^2 + (nz^2 - 2\sqrt{n}\,yz + y^2) = ny^2 - 2\sqrt{n}\,yz + nz^2$

$= (n-1)z^2 + (ny^2 - 2\sqrt{n}\,yz + z^2)$

Par suite en intervertissant les signes sommes :

$$E\{P^*\} = \int_{\frac{a-m}{\sigma}}^{\frac{b-m}{\sigma}} g(z)\,dz \int_{-\infty}^{+\infty} g\left(\frac{y\sqrt{n} - z}{\sqrt{n-1}}\right) d\left(y\sqrt{\frac{n}{n-1}}\right)$$

La seconde intégrale étant égale à l'unité, $E\{P^*\} = \int_{\frac{a-m}{\sigma}}^{\frac{b-m}{\sigma}} g(z)\,dz$, ce qui est bien la probabilité P que la variable aléatoire obéissant à une loi de Laplace-Gauss de moyenne m et de variance σ^2 soit comprise entre les deux limites données a et b.

4. Construisons, à titre d'exemples, dans les cas a, d et g les estimateurs sans distorsion de variance minimale au moyen de la méthode indiquée en 1.

Cas a. Etant donné N épreuves, le nombre R des succès est un estimateur exhaustif de p, donc r désignant le nombre de succès de n épreuves prises parmi les N, l'estimateur cherché de la probabilité P de r succès en n épreuves est :

$$P^* = \Pr\{r \mid R\} = \frac{\Pr\{r, R\}}{\Pr\{R\}} = \frac{\Pr\{R \mid r\}\Pr\{r\}}{\Pr\{R\}}$$

$$= \frac{C_{N-n}^{R-r}\, p^{R-r}\, q^{N-n-R+r} \cdot C_n^r\, p^r\, q^{n-r}}{C_N^R\, P^R\, Q^{N-R}}$$

$= \dfrac{C_{N-n}^{R-r}\, C_n^r}{C_N^R}$, quantité égale à celles de la formule (2).

Cas d. $S = \sum_{i=1}^{n} x_i$ est un estimateur exhaustif de θ, donc

$$P^* = \Pr\{x_1 > \xi \mid S\}$$

est l'estimateur cherché de $P = \Pr\{X > \xi\}$. Or, en désignant par $\delta(U)$ la densité de probabilité d'une variable aléatoire U :

$$\delta(x_1 \mid S) = \frac{\delta(S \mid x_1)\,\delta(x_1)}{\delta(S)}$$

Mais $\delta(x_1) = \dfrac{1}{\theta} e^{-\frac{x_1}{\theta}}$, $\delta(S) = \dfrac{1}{\theta^n} e^{-\frac{S}{\theta}} \dfrac{S^{n-1}}{(n-1)!}$

et $\delta(S \mid x_1) = \dfrac{1}{\theta^{n-1}} e^{-\frac{S-x_1}{\theta}} \dfrac{(S-x_1)^{n-2}}{(n-2)!}$ pour $S > x_1$ et 0 pour $S < x_1$

Donc, pour $S > x_1$, $\delta(x_1 \mid S) = (n-1)\dfrac{(S-x_1)^{n-2}}{S^{n-1}}$, qui est la dérivée par rapport à x_1 de

$$- \frac{(S - x_1)^{n-1}}{S^{n-1}}$$

Il en résulte bien pour $P^* = Pr\{x_1 > \xi | s\}$ l'expression indiquée en 2.d :

$$P^* = \begin{cases} \left(1 - \frac{\xi}{S}\right)^{n-1} & \text{si } S > \xi \\ 0 & \text{si } S \leqslant \xi \end{cases}$$

Cas g. L'estimateur cherché est égal à la probabilité P^* que l'observation x_1 est comprise entre a et b,

$$\bar{x} = \frac{x_1 + \ldots + x_n}{n} \quad \text{et} \quad ns^2 = \sum_{i=1}^{n} (x_i - \bar{x})^2$$

étant donnés.

Or $y = \frac{\bar{x} - m}{\sigma} \sqrt{n}$ et $z = \frac{ns^2}{\sigma^2}$ suivent indépendamment une loi de Laplace Gauss réduite et une loi de χ^2 à $(n - 1)$ degrés de liberté respectivement. La probabilité élémentaire de y et z est donc

$$\frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} dy \cdot \frac{1}{2^{\frac{n-1}{2}} \Gamma\left(\frac{n-1}{2}\right)} z^{\frac{n-3}{2}} e^{-\frac{z}{2}} dz$$

En posant $y' = \frac{\bar{x}' - m}{\sigma} \sqrt{n-1}$ et $z' = \frac{(n-1)s'^2}{\sigma^2}$, avec $\bar{x}' = \frac{x_2 + \ldots + x_n}{n-1}$ et $(n-1)s'^2 = \sum_{i=2}^{n} (x_i - \bar{x}')^2$, on en déduit aisément la probabilité élémentaire de y' et z'. Or $(n - 1)\bar{x}' = n\bar{x} - x_1$, d'où, en désignant $\frac{x_1 - m}{\sigma}$ par ξ, $y'\sqrt{n-1} = y\sqrt{n} - \xi$ et, puisque

$$(n-1)s'^2 = ns^2 - \frac{n-1}{n} (\bar{x}' - x_1)^2 , \quad z'\sqrt{n-1} = \sqrt{n-1} - (y - \xi\sqrt{n})^2$$

Le jacobien $\frac{d(y', z')}{d(y, z)}$ étant égal à $\sqrt{\frac{n}{n-1}}$, on déduit de la probabilité élémentaire de y' et z' celle de y et z, connaissant ξ :

$$\frac{1}{\sqrt{2\pi}\, 2^{\frac{n-2}{2}} \Gamma\left(\frac{n-2}{2}\right)} \sqrt{\frac{n}{n-1}}\, e^{-\frac{z}{2}} e^{-\frac{y^2 - \xi^2}{2}} \left[z - \frac{1}{n-1}(y - \xi\sqrt{n})^2\right]^{\frac{n-4}{2}} dy\, dz$$

Or, avec une notation déjà définie, $\delta(\xi | y, z) = \frac{\delta(y, z | \xi)\, \delta | \xi)}{\delta(y, z)}$. D'où l'on déduit par un calcul facile la probabilité élémentaire de ξ connaissant y et z ;

$$\frac{1}{B\left(\frac{1}{2}, \frac{n}{2} - 1\right)} \sqrt{\frac{n}{n-1}} \left[z - \frac{n}{n-1}\left(\frac{y}{\sqrt{n}} - \xi\right)^2\right]^{\frac{n-4}{2}} \frac{d\xi}{z^{\frac{n-3}{2}}}$$

et en remarquant que $\frac{y}{\sqrt{n}} - \xi = \frac{\bar{x} - x_1}{s\sqrt{n}}\sqrt{z}$ et en posant $u = \frac{x_1 - \bar{x}}{s\sqrt{n-1}}$, la probabilité élémentaire de u :

$$\frac{1}{B\left(\frac{1}{2}, \frac{n}{2} - 1\right)} (1 - u^2)^{\frac{n-4}{2}} du,$$

l'intégrale n'étant à prendre que pour les valeurs de u inférieures à un en valeur absolue, car $(n - 1) s^2 - (\bar{x} - x_1)^2 = \frac{(n-1)^2}{n} s'^2$ est une quantité toujours positive.

Ainsi se trouvent établis les résultats du § 2. g.

Remarque : Etant donné la forme de la probabilité élémentaire de u, on peut préférer à la formule (9) du paragraphe 2.g, le calcul de l'intégrale P^* par une méthode trigonométrique élémentaire, le changement de variable $u = \sin\varphi$ ramenant à l'intégrale $\int \cos^{n-3}\varphi\, d\varphi$.

BIBLIOGRAPHIE

[1] BARTON D.E. (1961) - Unbiased estimation of a set of probabilities. Biometrika *48*. 227-9

[2] BASU A.P. (1964) - Estimates of reliability for some distributions udeful in life testing. Technometrics *6*, 215-9.

[3] BLACKWELL D. (1947) - Conditional expectation and unbiased sequential estimation. Ann. Math. Stat. *18*, 105-10.

[4] LAURENT A.G. (1963) - Conditional distribution of order statistics and distribution of the reduced i^{th} order statistic of the exponentiel model. Ann. Math. Stat. *34*, 652-7

[5] LEHMANN E.L. and SCHEFFE H. (1950) - Completeness, similar regions and unbiased estimates. Sankhya *10*, 305-40.

[6] RAO C.R. (1949) - Sufficient estimates and minimum variance estimates. Proc. Cambr. Phil. Soc. *45*, 213-18.

ANNEXE SUR LA LOI DE L'ALTERNATIVE GENERALISEE REPETEE

1. Cette loi généralise la loi de Bernoulli ou loi binomiale. Si une épreuve se traduit par un événement parmi k s'excluant mutuellement et de probabilités respectives $p_i (i = 1, 2, \ldots, k)$ avec $\sum_{i=1}^{k} p_i = 1$; la probabilité que, dans un ensemble de n épreuves indépendantes, le premier évènement se produise n_1 fois,..., le $i^{ème}$ n_i fois, et le $k^{ième}$ n_k fois est :

$$\frac{n!}{n_1!\dots n_i!\dots n_k!}\, p_1^{n_1} \dots p_i^{n_i} \dots p_k^{n_k} = n! \prod_{i=1}^{k} \frac{p_i^{n_i}}{n_i!} \qquad (1)$$

2. Nous nous proposons d'établir une formule générale donnant les moments factoriels d'ordres quelconques relatifs à $n_1, \dots, n_i, \dots, n_k$. Utilisant la notation $\nu^{[\alpha]} = \nu(\nu - 1) \dots (\nu - \alpha + 1)$ pour α entier positif et $\nu^{[0]} = 1$, nous cherchons donc l'expression de l'espérance mathématique :

$$E \left\{ \prod_{i=1}^{k} n_i^{[\alpha_i]} \right\} \qquad (2)$$

où les α_i sont des entiers donnés positifs ou nuls, leur somme étant désignée par α.

3. A cette fin, considérons la somme $S = p_1 + \dots + p_i + \dots + p_k$ où nous regardons, provisoirement, les p_i comme analytiquement indépendants.

$$\frac{\partial^{\alpha_1} (S^n)}{\partial_{p_1}^{\alpha_1}} = n(n - 1) \dots (n - \alpha_1 + 1)\, S^{n-\alpha_1}$$

et plus généralement,

$$\frac{\partial^{\alpha} (S^n)}{\partial p_1^{\alpha_1} \dots \partial p_i^{\alpha_i} \dots \partial p_k^{\alpha_k}} = n^{[\alpha]}\, S^{n-\alpha} \qquad (3)$$

Mais, d'autre part S^n pouvant être obtenu par la sommation (notée $\sum$) des termes de la forme (1) correspondant à tous les ensembles d'entiers positifs ou nuls $(n_1, \dots, n_i, \dots, p_k)$ de somme égale à n, le premier membre de (3) est aussi égal à

$$n! \sum \prod_{i=1}^{k} \frac{n^{[\alpha_i]}}{n_i!} p_i^{n_i-\alpha_i} \qquad (4)$$

En multipliant le second membre de (3) et cette expression (4) par $\prod_{i=1}^{k} p_i^{\alpha_i}$ et en égalant les expressions résultantes, on obtient :

$$\sum n! \prod_{i=1}^{k} \frac{p_i^{n_i}}{n_i!} n_i^{[\alpha_i]} = n^{[\alpha]}\, S^{n-\alpha} \prod_{i=1}^{k} p_i^{\alpha_i}$$

c'est-à-dire, en faisant S = 1 :

$$\boxed{E \left\{ \prod_{i=1}^{k} n_i^{[\alpha_i]} \right\} = n^{[\alpha]} \prod_{i=1}^{k} p_i^{\alpha_i}} \qquad (5)$$

où $$\alpha = \sum_{i=1}^{k} \alpha_i$$

ce qui est la formule générale annoncée.

4. Voici quelques formules particulières en résultant :

a) Un seul des α_i est égal à 1, les autres sont nuls :

$$E\{n_i\} = np_i$$

formule classique de la moyenne théorique.

b) Un seul des α_i est égal à 2, les autres sont nuls :

$$E\{n_i^{[2]}\} = n^{[2]}\ p_i^2, \quad \text{ou } E\{n_i(n_i - 1)\} = n(n-1)\ p_i^2$$

Donc

$$V\{n_i\} = E\{n_i^2\} - [E\{n_i\}]^2 = n(n-1)\ p_i^2 + np_i - n^2\ p_i^2 = np_i q_i$$

formule classique de la variance, où $q_i = 1 - p_i$.

c) Seuls α_i et α_j sont non nuls et égaux à un :

$$E\{n_i\ n_j\} = n^{[2]}\ p_i\ p_j = n(n-1)\ p_i\ p_j$$

d'où $\text{cov}\{n_i, n_j\} = E\{n_i\ n_j\} - E\{n_i\}\ E\{n_j\} = n(n-1)\ p_i\ p_j - n^2 p_i p_j = -\ np_i\ p_j$

et
$$\rho\{n_i, n_j\} = \frac{-\ n\ p_i\ p_j}{\sqrt{np_i\ q_i\ .\ np_j\ q_j}} = -\sqrt{\frac{p_i\ p_j}{q_i\ q_j}}$$

d) Application au calcul de la moyenne et de la variance du χ^2.

$$\chi^2 = \sum_{i=1}^{k} \frac{(n_i - np_i)^2}{np_i} = \sum_{i=1}^{k} \frac{n_i^2}{np_i} - n, \quad \text{donc } E\{\chi^2\} = \frac{1}{n} \sum_{i=1}^{k} \frac{n(n-1)\ p_i^2 + np_i}{p_i} - n$$

$$= (n-1) \sum_{i=1}^{k} p_i + k - n = k - 1.$$

Le fait que l'espérance mathématique du χ^2 est égale à $k - 1$, k étant le nombre des catégories, est vrai exactement : il ne résulte d'aucune approximation.

Il n'en est pas de même en ce qui concerne le résultat classique sur la variance : $V\{\chi^2\}$ n'est égale à $2(k-1)$ qu'au prix d'une approximation, comme le montre la formule exacte que nous allons établir :

Puisque

$$V\{\chi^2\} = E\{[\chi^2 - (k-1)]^2\} = E\{\chi^4\} - (k-1)^2$$

calculons

$$\chi^4 = \sum \frac{n_i^4}{n^2\ p_i^2} + 2 \sum_{i} \mathop{\sum}_{<\ j} \frac{n_i^2 n_j^2}{n^2\ p_i\ p_j} - 2n \sum \frac{n_i^2}{np_i} + n^2$$

Le calcul de $E\{\chi^4\}$ est simple, en utilisant les deux formules suivantes qui résultent de cas particuliers de (5) en remarquant que

$$n_i^4 = n_i^{[4]} + 6\ n_i^{[3]} + 7\ n_i^{[2]} + n_i^{[1]}$$

et que

$$n_i^2 n_j^2 = n_i(n_i - 1)\, n_j(n_j - 1) + n_i(n_i - 1)\, n_j + n_i n_j(n_j - 1) + n_i n_j ,$$

$$E\{n_i^4\} = n(n-1)(n-2)(n-3)\, p_i^4 + 6\, n(n-1(n-2)\, p_i^3 + 7\, n(n-1)\, p_i^2 + np_i$$

$$E\{n_i^2 n_j^2\} = n(n-1)(n-2)(n-3)\, p_i^2 p_j^2 + n(n-1)(n-2)\, p_i^2 p_j + n(n-1)(n-2)\, p_i p_j^2 + \\ + n(n-1)\, p_i p_j .$$

On en déduit aisément que

$$V\{\chi^2\} = 2(k-1) + \frac{1}{n}\left[\sum_{i=1}^{k} \frac{1}{p_i} - (k^2 + 2k - 2)\right]$$

et qu'aucune des espérances mathématiques np_i ne doit être trop petite pour que l'on puisse assimiler $V\{\chi^2\}$ à $2(n-1)$ et la loi du χ^2 à une loi de Pearson à $k-1$ degrés de liberté.