

JOURNAL DE LA SOCIÉTÉ STATISTIQUE DE PARIS

J. DUFRÉNOY

1868-1968 : centenaire du mémoire de Bienaymé « Sur la probabilité des erreurs d'après la théorie des moindres carrés »

Journal de la société statistique de Paris, tome 109 (1968), p. 128-130

http://www.numdam.org/item?id=JSFS_1968__109__128_0

© Société de statistique de Paris, 1968, tous droits réservés.

L'accès aux archives de la revue « Journal de la société statistique de Paris » (<http://publications-sfds.math.cnrs.fr/index.php/J-SFdS>) implique l'accord avec les conditions générales d'utilisation (<http://www.numdam.org/conditions>). Toute utilisation commerciale ou impression systématique est constitutive d'une infraction pénale. Toute copie ou impression de ce fichier doit contenir la présente mention de copyright.

NUMDAM

Article numérisé dans le cadre du programme
Numérisation de documents anciens mathématiques
<http://www.numdam.org/>

1868-1968 : CENTENAIRE DU MÉMOIRE DE BIENAYMÉ « SUR LA PROBABILITÉ DES ERREURS D'APRÈS LA THÉORIE DES MOINDRES CARRÉS »

Mém. Savants Étrangers, Acad. Sci. Paris, 15 : 615-663

Dans un article, consacré aux « Précurseurs du χ^2 de Pearson » (*The Australian Jour. of Statistics*, 8, 3, nov. 1966, pp. 117-126), H. O. Lancaster rappelle les études effectuées dès 1838, par Bienaymé, selon Jakob Bernoulli, de Moivre, Bayes, Poisson et enfin Laplace, au sujet de la relation entre valeur observée (y_o) et valeur correspondantes calculée (y_c) d'après un certain modèle mathématique de la distribution des valeurs de la variable Y.

Soient p_i (avec $i = 1, 2, 3$) les probabilités (inconnues *a priori*) qu'une valeur observée (y_o) appartienne à l'une de trois classes caractérisées chacune par la fréquence N_1 , N_2 ou N_3 telles que $N_1 + N_2 + N_3 = N$ avec $p_1 + p_2 + p_3 = 0$.

Utilisant N valeurs observées dont N_i appartiennent à la classe i , Bienaymé, développant les travaux de Laplace, arrive par sommation pour toutes les classes, à l'expression $S \frac{(y_o - y_c)^2}{y_c}$ qui correspond à la formule du χ^2 de Pearson; en 1852, Bienaymé complétait ces premiers résultats dans un mémoire publié par le *Journal Math. Pures et Appl.*, 17 : 37 + 78 et qui fut réimprimé en 1868 dans les *Mémoires de l'Académie des Sciences* ⁽¹⁾.

En 1844, Ellis utilisa les transformations de Laplace et de Fourier pour déterminer la distribution de la somme de n valeurs aléatoires indépendantes, notamment au sujet du problème proposé par Isaac Newton, quant à la durée du règne de chaque roi; Ellis parvenait ainsi à la distribution de χ^2 en tant que manifestation de processus stochastique.

En 1865, Mendel présentait devant la Société des Sciences naturelles de Brunn son mémoire sur les hybrides des plantes. Reprenant, sans s'en douter, le concept corpusculaire de l'hérédité, proposé un siècle plus tôt par Maupertuis dans la *Vénus physique* (1746), Mendel interprète les résultats de ses croisements entre deux types de pois, différents par un caractère, en fonction de l'alternative : présence (A) ou absence (a) dans chacun des deux géniteurs; d'où possibilité pour un descendant de seconde génération d'appartenir à l'une de 3 classes AA, Aa, ou aa. Si les caractères « phénotypiques » permettent sans ambiguïté l'attribution de chaque descendant à l'une des 3 classes (Aa étant équivalent à aA) les fréquences observées dans chaque classe (N_1 , N_2 , N_3) pour un total de N descendants avec $N = N_1 + N_2 + N_3$ correspondent à celles que permettent de calculer les probabilités : 1/4 pour AA, 1/2 pour Aa ou aA et 1/4 pour aa; si, au lieu d'une seule alternative A ou a, interviennent, indépendamment, n alternatives, le nombre des classes « phénotypiques » devient 3^n , et l'analyse combinatoire permet de calculer, pour chacune de ces classes, la fréquence théorique y_c à comparer à la fréquence observée, y_o .

Dès cette époque, Mendel aurait pu trouver dans les publications de Bienaymé la méthode permettant d'évaluer statistiquement la signification de l'écart entre fréquences observée (y_o) et calculée (y_c).

En 1875, Francis Galton, redécouvrant le concept de l'hérédité corpusculaire de Maupertuis et de Mendel, et considérant chaque « corpuscule » ou unité structurale comme une molécule organique, parvenait théoriquement à la proportion 1 : 2 : 1, soit $1/4 + 1/2 + 1/4$ pour la distribution des fréquences dans les « classes » de descendants de parents différant quant à une paire de caractères (A ou a); F. Galton remarquait d'ailleurs que les lignes successives du triangle de Pascal

1. BIENAYMÉ, Sur la probabilité des résultats moyens d'observations : démonstration directe de la règle de Laplace, *Mém. Sav. Etr. Acad. Sci. Paris*, 5 : 513-558.

indiquent les fréquences dans chaque classe, à mesure que la situation devient de plus en plus complexe; par exemple, sur la 5^e ligne, les chiffres 1, 4, 6, 4, 1 correspondent aux fréquences dans chacune des 4 classes auxquelles peuvent être attribuées l'un des descendants de deuxième génération d'un croisement entre géniteurs mettant en jeu ces deux alternatives: A ou a et B ou b, avec équiprobabilité (1).

En 1897 Sir Francis Galton écrivait (*Proc. Roy. Soc. London*, 61 : 401-13) : « Les deux parents à eux deux contribuent en moyenne la moitié (soit 0,5) de l'héritage total des descendants; les quatre grands-parents un quart (ou (0,5)²); les huit grands-grands parents, un huitième (ou (0,5)³). Cette loi statistique n'exprime rien qui ne soit conforme à l'opinion que la ligne de descendance procède de germe à germe et non de personne à personne... La loi rend compte de la subdivision binaire des cellules germinales, ainsi que de l'extrusion concomitante, donc perte, d'une moitié des contributions de chaque parent à la cellule germinale d'un descendant... »

Selon les conceptions cytogénétiques, chaque alternative, telle que A ou a, est évoquée par la présence en un certain *locus* d'un chromosome, du facteur A chez un parent, et au *locus* correspondant du chromosome homologue de l'autre parent, de l'allèle a.

Chaque parent, possède, dans chaque cellule non sexuelle, une paire de ces chromosomes : chez le parent homozygote, chacun de ces deux chromosomes possède le facteur A (homozygote dominant) de génotype AA ou le facteur a (homozygote récessif) de génotype aa; un individu hétérozygote possède le facteur A sur l'un des chromosomes de la paire, l'allèle a au *locus* correspondant de l'autre chromosome; son génotype est Aa.

Une cellule sexuelle ne recevant que l'un des chromosomes de la paire, recevra évidemment un A chez l'homozygote dominant, un a chez l'homozygote récessif, mais soit A soit a; avec équiprobabilité, chez l'hétérozygote.

Dans une population soumise à autofécondation (2), l'équilibre génotypique des proportions de AA, Aa, aa peut se définir par l'égalité $(P, R, Q) = p^2 + pqF, 2pq(1 - F), q^2 + PQF$ caractérisée par la valeur de F dans l'égalité, lorsque $p = P + 1/2 R, q = 1 - p = Q + 1/2 R$.

Un récent traité sur les fondements mathématiques de la Génétique (3) fait état du triangle de Pascal, pages 39-40; Pascal étant d'ailleurs l'un des trois auteurs français cités, les pages 45 et 46 étant consacrées aux séries de Poisson.

Les tests de χ^2 font l'objet des pages 73-105, les pages suivantes (106 sqq.) étant consacrées à R. A. Fisher; R. A. Fisher, ayant utilisé le test de χ^2 pour évaluer statistiquement la signification des écarts entre fréquences observées par Mendel dans ses différentes classes d'hybrides de deuxième génération et les fréquences correspondantes calculées selon les Lois de Mendel, arriva à cette conclusion que pour les premiers croisements effectués, les valeurs de χ^2 sont très probables; et que les résultats des croisements ultérieurs s'accordent trop bien avec ceux que permettaient de prévoir la théorie, les valeurs de χ^2 devenant inférieures à celles qui auraient le plus de chances de se manifester du seul fait des erreurs aléatoires... R. A. Fisher suspectait donc les résultats d'avoir été manipulés en vue d'un accord trop parfait avec les prévisions; il paraît plus logique de penser qu'à mesure que Mendel disposait d'une plus grande masse de résultats expérimentaux s'accordant avec ses prévisions, il employait de plus en plus une méthode d'analyse séquentielle, interrompant les dénombrements dès que les fréquences dénombrées lui paraissaient suffisantes pour conclure à la vérification de la théorie.

Après le critérium de χ^2 , le traité étudie longuement le critérium de Fisher et expose les méthodes de calcul du Maximum de vraisemblance.

Nous avons été fort surpris de ne trouver, dans ce traité de 464 pages, qu'une seule page traitant de l'Information (p. 111); la formule du critérium d'Information est donnée sous une forme qui ne sera intelligible que pour des initiés.

1. R. C. OLBY, Francis Galton's derivation of Mendelian ratios in 1875, *Heredity*, 20, 4, 636-638, 1965.
2. S. K. JAIN et P. L. WORKMAN, Generalized F. Statistics and the Theory of Inbreeding and selection, *Nature*, 214, may 13, 67, pp. 674-000.
3. Genetik Grundlagen, Ergebnisse und Probleme in Einzeldarstellungen. Beitrag 5, Mathematische Grundlagen der Genetik von Erna Weber, G. Fischer Verlag, Berlin, 1967.



En conclusion, une brève étude historique des « bases mathématiques de la Génétique » semble justifier la prise en considération de l'article de H. O. Lancaster, rappelant le rôle du mathématicien français Bienaymé comme fondateur des théories mathématiques qui ont au cours de ce siècle joué ou permis le développement des théories statistiques de l'Hérédité, en réhabilitant d'ailleurs les théories émises par Maupertuis au milieu du xviii^e siècle (¹, ²).

J. DUFRÉNOY