MATHÉMATIQUES ET SCIENCES HUMAINES

Problèmes d'enseignement

Mathématiques et sciences humaines, tome 32 (1970), p. 75-86

<http://www.numdam.org/item?id=MSH_1970__32__75_0>

M.S.H. (8e année, no 32, 1970, p. 75-82)

APPLICATIONS PRATIQUES DES LOIS DE PROBABILITÉ (9)

par

B. LECLERC

LOI DE PASCAL *PSYCHOLOGIE*

LOI DE FISHER

RIDDEL W. J. B., "The relation between the number of speakers and the number of contributions to the *Transactions of the ophtalmological society of the United Kingdom* between 1881 and 1890", *Annals of Eugenics*, 12, pp. 274-279, 1945.

Cet article est similaire à celui de Williams (fiche dans *Math. Sci. hum.*, no 28), sur les publications de biologistes.

Modèles

Deux modèles sont confrontés pour l'ajustement de la distribution du nombre de publications par chercheur.

Loi géométrique, ou loi de Pascal: le nombre de n_k de chercheurs ayant publié k articles est approximativement:

$$n_k = n_1 x^{k-1} \qquad k \geqslant 1, \qquad 0 \leqslant x \leqslant 1.$$

Loi logarithmique de Fisher:

$$n_k = n_1 \frac{x^{k-1}}{k} \qquad k \geqslant 1, \qquad 0 \leqslant x \leqslant 1.$$

Estimation

Si S est le nombre d'auteurs et N celui des publications, on a pour la loi de Pascal:

$$n_1 = \frac{S^2}{N} \qquad \text{et} \qquad x = \frac{N - S}{N}$$

pour la loi de Fisher:

$$S = \frac{n_1}{x}(-\log(1-x)) \qquad \text{et} \qquad N = \frac{n_1}{1-x}$$

d'où le calcul de n_1 et x.

Test du χ^2, utilisé uniquement pour comparer entre eux les ajustements aux deux lois. Le niveau de signification n'est pas pris en compte.

Application à 760 publications par 138 auteurs citées dans l'index décennal des *Transactions of the Ophtalmological Society of the United Kingdom* entre 1881 et 1890 (volumes 1 à 10). L'auteur étudie la distribution globale, puis celle plus restreinte portant sur les 64 membres fondateurs survivants en 1890, enfin la distribution des autres chercheurs. Tous les tableaux numériques sont inclus dans l'article.

Pour la distribution globale, la loi logarithmique s'ajuste bien, mais non la loi géométrique. Pour les membres fondateurs, l'ajustement géométrique est le meilleur ($\chi^2 = 2{,}33$ pour 4 degrés de liberté, au lieu de 7,19). Pour la distribution « résiduelle », l'ajustement logarithmique est supérieur ($\chi^2 = 8{,}78$ pour 3 degrés de liberté, au lieu de 25,80).

BIBLIOGRAPHIE

DUFRESNOY J., " The publishing behaviour of biologists ", *Quart. Rev. Biol.*, 13, p. 207, 1938.

FISHER R. A., CORBETT, A. S. et WILLIAMS, C. B., " The relation between the number of species and the number of individuals in a random sample of an animal population ", *J. animal Ecol.*, 12, 12, p. 42, 1943.

WILLIAMS, C. B., " The number of publications written by biologists ", *Annals of Eugenics*, 12, pp. 143-146, 1945.

LOI DE FISHER — *RECHERCHE OPÉRATIONNELLE* — *ÉCOLOGIE*

HARRISON J. L., "Stored products and the insects infesting them as examples of the logarithmic series", *Annals of Eugenics*, 12, pp. 280-282, 1945.

Cet article fait suite au précédent et reprend le modèle de la loi logarithmique de Fisher pour une autre application. La méthode d'ajustement et les références bibliographiques sont communes. Le *test* du χ^2 est effectué ici pour mesurer la qualité des ajustements.

L'auteur, au cours d'une tournée d'inspection de dépôts alimentaires au Moyen-Orient, a recensé les types d'aliments et les espèces d'insectes représentés. Il ajuste la loi logarithmique à la distribution des nombres d'apparition des espèces d'insectes, puis des sortes d'aliments. Pour la première distribution, l'ajustement est excellent, la probabilité pour une variable du χ^2 de dépasser la quantité-test est supérieur à 0,99. Pour la seconde, il est acceptable, avec une probabilité dépassant 0,3. Un tableau fournit les données détaillées.

GOOD I. J., "The population frequencies of species and the estimation of population parameters", *Econometrica*, 40, 1953, pp. 237-264.

Un échantillon au hasard de taille N est extrait d'une population supposée infinie d'animaux d'espèces diverses. Soit n_r le nombre d'espèces représentées r fois exactement dans l'échantillon.

Dans les premiers paragraphes de l'article, l'auteur suggère une méthode pour estimer certaines caractéristiques de la population, notamment:

— la fréquence de chaque espèce,

— la proportion de la population représentée dans l'échantillon, c'est-à-dire appartenant à l'une des S espèces présentes dans l'échantillon,

— des paramètres pour mesurer l'hétérogénéité de la population, en particulier son entropie.

L'auteur s'intéresse ensuite au problème du « lissage » des données: remplacer les n_r par des nombres n'_r voisins mais présentant une certaine régularité.

Une première méthode est graphique et manuelle. On travaille en fait sur les $\sqrt{n_r}$. Sous certaines hypothèses, en effet (travaux de Bartlett (1966), Anscombe (1948)), la variance de $\sqrt{n_r}$ est voisine de $\frac{1}{2}$, indépendamment de r.

On peut aussi faire des hypothèses sur la distribution des fréquences des espèces de la population. Soit $f(p)\,dp$ le nombre d'espèces dont la fréquence est comprise entre p et $p + dp$. On en déduit $E(n_r)$, espérance mathématique de n_r.

$$E(n_r) = \binom{N}{r} \int_0^1 p^r (1-p)^{N-r} f(p)\,dp$$

$$\simeq \frac{1}{r!} \int_0^1 (pN)^r e^{-pN} f(p)\,dp$$

la formule approchée étant utilisable pour r^2 petit devant N.

L'auteur propose sept modèles théoriques de cette forme.

Moins précisément, on peut se contenter d'hypothèses sur les nombres $E(n_r)$. Trois formules sont suggérées:

$$E(n_r) = \frac{\lambda}{r^\zeta} \qquad \zeta > 1$$

$$E(n_r) = \frac{\lambda x^r}{r^\xi} \qquad 0 < x < 1 \qquad \zeta \geqslant 1$$

$$E(n_r) = \frac{\lambda x^r}{r(r+1)} \qquad 0 < x < 1$$

Les n'_r sont alors ces nombres $E(n_r)$. Notons que pour les trois formules ci-dessus, il s'agit d'ajuster les n_r à une distribution parétienne au sens large.

L'auteur donne quelques exemples de tels ajustements:

1) Captures de macrolépidoptères dans un piège lumineux, données de Williams (1943).

$$N = 15\,609 \qquad S = 240.$$

Après avoir utilisé trois méthodes graphiques manuelles, l'auteur reprend le modèle théorique de Fisher pour ces mêmes données pour lequel on a:

$$f(p) = \frac{\beta\, e^{-\beta}\, p}{p}$$

d'où:

$$E(n_r) = \frac{\beta\, x^r}{r}$$

Fisher a trouvé:

$$\beta \simeq 40{,}2 \qquad \text{et} \qquad x \simeq 0{,}9974.$$

La méthode d'ajustement n'est pas précisée ici. L'auteur considère que la quantité:

$$\chi^2 = \sum_{t=1}^{r} \frac{(n_t - E(n_t))^2}{V(n_t)}$$

suit approximativement une loi du χ^2 à r degrés de liberté.

Il s'en sert pour tester l'ajustement. Le résultat n'est pas précisé, mais l'ajustement est estimé satisfaisant, meilleur que ceux par la méthode graphique.

2) Statistique d'Eldridge (1911) sur les mots à désinence des journaux américains.

$$N = 43\,989, \qquad S = 6\,001$$

Un ajustement théorique remarquablement bon pour des valeurs de r pas trop élevées est donné par:

$$n_r \simeq \frac{s}{r(r+1)}$$

Pour obtenir un bon ajustement, on prend:

$$n_r \simeq \frac{\lambda\, x^r}{r(r+1)}$$

A partir de:

$$N = \lambda \sum_{r=1}^{\infty} \frac{x^r}{r+1} = -\frac{\lambda}{x}(x + \log(1-x))$$

$$S = \lambda \sum_{r=1}^{\infty} \frac{x^r}{r\,(r+1)} = \frac{\lambda}{x}\,(x + (1-x)\log(1-x))$$

on a:

$$x = 1 - e^{-y},$$

où y est obtenu itérativement:

$$y = \lim_{n \to \infty} y_n$$

$$y_{n+1}^{-1} = (1 - e^{-y_n})^{-1} - \left(1 + \frac{S}{N}\right)^{-1}$$

et:

$$\lambda = \frac{x\ N}{y - x}.$$

Ici:

$$\lambda \simeq 6\,017{,}4 \qquad \text{et} \qquad x \simeq 0{,}999667$$

3) Échantillon de substantifs dans l'essai de Macoulay sur Bacon (données de Yule (1944)) :

$$N = 8\,045, \qquad S = 2\,048.$$

Le modèle théorique est le même que le précédent.

On trouve:

$$\lambda \simeq 2\,138{,}90, \qquad x \simeq 0{,}991074.$$

Un test du χ^2 classique indique un ajustement suffisamment bon.

4) Ouvertures au jeu d'échecs (3 premiers échanges), publiées dans le *British Chess Magazine*, 1951.

$$N = 385, \qquad S = 174.$$

Avec l'hypothèse théorique:

$$f(p) = \frac{e^{-\beta p}}{k\,p^{-2}} \qquad p > p_0$$

$$f(p) = 0 \qquad p < p_0$$

On a:

$$E(n_r) = \frac{\lambda\,x^r}{r\,(r-1)}, \quad r \geqslant 2, \quad r^2 \text{ petit devant } N.$$

Ce modèle est donc voisin du précédent. On trouve:

$$x \simeq 0{,}99473 \qquad \lambda \simeq 49{,}635$$

d'où:

$$\beta \simeq 2{,}040 \qquad p_0 \simeq \frac{1}{8\,846}.$$

BIBLIOGRAPHIE

ANSCOMBE F. J., "The transformation of Poisson, binomial and negative binomial data", *Biometrika* 35, pp. 246-254, 1948.

ANSCOMBE F. J., "Sampling theory of the negative binomial and logarithmic series distribution", *Biometrika*, 37, pp. 358-382, 1950.

CHAMBERS E. G. et YULE G. U., "Theory and observation in the investigation of accident causation", *J. r. statist. Soc.*, Suppl. 7, pp. 89-109, 1948.

CORBETT A. S., FISHER R. A. et WILLIAMS C. B., "The relation between the number of species and the number of individuals in a random sample of an animal population", *J. animal Ecology*, 12 pp. 42-58, 1943.

ELDRIDGE R. C., *Six thousand common English words* Buffalo, the Clements Press, 1911.

GOODMAN L. A., "On the estimation of the number of classes in a population", *Ann. math. statist.*, 20, pp. 572-579, 1949.

GREENWOOD M. et YULE G. U., "An inquiry into the nature of frequency distributions of multiple happenings", *J. r. statist. Soc.*, 83, p. 255, 1920.

HARDY G. H., *Divergent series*, Clarendon Press, Oxford, 1949.

JEFFREYS H., *Theory of probability*, Clarendon Press, Oxford, 1948.

JEFFREYS H. et JEFFREYS B. S., *Methods of mathematical Physics*, Cambridge University Press, 1946.

NEWBOLD E. M., "Practical application of the statistics of repeated events, particularly to industrial accident", *J. r. statist. Soc.*, 90, pp. 487-547, 1927.

PRESTON F. W., "The commonness and rarity of species", *Ecology*, 29, pp. 254-283, 1948.

USPENSKY J. V., *Introduction to mathematical probability*, McGraw Hill, New York, 1937.

WHITTAKER E. T. et ROBINSON G., *The calculus of observations*, Blackie, London and Glasgow, 1944.

YULE G. U., *Statistical study of litterary vocabulary*, Cambridge University Press, 1944.

ZIPF G. K., *Human behaviour and the principle of least effort*, Addison Wesley, Reading, 1949.

LESAVRE J., "Contrôle de l'homogénéité d'un mélange de solides", *Rev. Statist. appl.*, VIII, n° 3, pp. 75-85, 1960.

Il s'agit de contrôler l'homogénéité de mélanges vitrifiables (sable, calcaire, dolomie, carbonates) avant leur introduction dans les fours de fusion. La méthode employée est celle des traceurs colorés: des grains de calcaires légèrement colorés en vert sont introduits en quantité convenable dans le mélange. On retire ensuite une trentaine d'échantillons de l'ordre du litre et l'on étudie la distributioo du nombre de grains colorés par échantillon.

Modèle. — Dans l'hypothèse d'homogénéité du mélange, la distribution est poissonnienne.

Estimation. — Identification des moyennes observées et théoriques.

Test. — La quantité:

$$\frac{\sum_i (x_i - \bar{x})^2}{\bar{x}}$$

où les x_i sont les nombres de grains observés,

$\bar{x}$ est la moyenne observée

suit, si l'hypothèse est vérifiée, la loi du χ^2 à $k - 1$ degrés de liberté. L'auteur fait un test bilatéral à 5 %, remarquant que l'on réputera anormaux 5% des résultats normaux.

Application à six semaines de 30 mesures provenant d'une usine allemande. Pour des raisons techniques, les poids des prélèvements ne sont pas rigoureusement constants. Ceci a été corrigé en rapportant le nombre des grains au poids moyen des prélèvements. Les données sont fournies dans l'article.

Un seul des six tests est significatif. Pour les autres, la valeur du χ^2 est proche de son espérance.

L'auteur teste ensuite le caractère aléatoire des séries chronologiques de trente prélèvements successifs, par trois méthodes:

— Dénombrement des maxima et minima;

— Longueur des phases monotones;

— Nombres de croisements avec la moyenne.

Un seul mélange peut être suspect. On observe aussi une certaine différence entre les trois premiers et les trois derniers mélanges. Tout ceci n'est jamais significatif: l'auteur a appris par la suite qu'il avait plu le jour des trois premiers mélanges. D'autres études ont permis de mettre en évidence le rôle dominant du facteur humidité du sable.

Références bibliographiques. — 12 titres se rapportant presque tous au problème technique particulier étudié.

VINING R., " Delimitation of economic areas: statistical conceptions in the study of the spatial structure of an economic system ", *J. am. statist. Ass.*, 48, n_y 1, pp. 44-64, 1953.

A la fin de cet article où l'auteur présente les traits généraux du travail effectué par le Bureau of Census pour regrouper les comtés des États-Unis en 501 nouvelles régions dénommées States Economic Areas, l'auteur s'intéresse à la distribution des éloignements de destinations des marchandises depuis leur centre d'origine.

Modèle et ajustement. Des distributions de longueurs de transports sont ajustées à la loi log-normale, de forme analytique non précisée dans l'article. Cet ajustement est fait par tracé de droites de Henry sur papier gausso-logarithmique.

Applications. (Une partie des données est présentée sous forme de figures):

Distances moyennes de 1 252 groupes de marchandises transportées par rail aux États-Unis pendant les premiers trimestres 1947, 1948, 1949 et 1950 (quatre distributions s'ajustant à la même loi).

Distances par wagon de marchandise, 1949 et premier trimestre 1947 (deux distributions ajustées à la même loi).

Distances parcourues par des charges de wagons originaires d'Alabama, puis de Virginie, 1949.

Distances parcourues par des charges de wagons en 1948 et au troisième trimestre 1947, États-Unis, puis Alabama seul.

Comparaison du pourcentage de wagons originaires, puis destinataires de la Virginie parcourant une distance donnée avec les nombres théoriques correspondants à une loi log-normale, 1948.

Sources. I C C Carload Waybill Analyses, divers numéros entre 1947 et 1950.

M.S.H. (8e année, no 32, 1970, p. 83-86)

QUELQUES REMARQUES SUR L'ENSEIGNEMENT DES MATHÉMATIQUES EN GÉOGRAPHIE

par

J. ZEITOUN *

Les remarques qui suivent portent essentiellement sur une expérience de deux années d'enseignement et sur quelques rencontres avec des géographes, qu'ils soient enseignants ou pratiquants.

La première constatation est dans la diversité, relative, des techniques mathématiques et de l'informatique, nécessaires à la pratique du géographe.

Qu'il s'agisse de géographie humaine ou physique, l'étudiant, surtout lorsqu'il n'est pas encore spécialisé, devrait normalement posséder quelques éléments d'analyses, d'algèbre linéaire, de statistique inductive et descriptive et d'algèbre et combinatoire, et même de géométrie analytique et différentielle. En bref, un peu de tout, sans oublier des notions de topologie générale.

Les besoins ainsi exprimés ne peuvent être satisfaits convenablement, semble-t-il, en deux années de D.U.E.L. De plus, se pose en permanence la question du choix du mode d'enseignement: la tendance recette ou la tendance compréhensive, plus approfondie, la seconde nécessitant des choix en fonction d'une demande précise des géographes.

Une seconde remarque concerne l'esprit dans lequel l'étudiant géographe perçoit les mathématiques. D'une part, et bien que la géographie enseignée soit essentiellement « littéraire », le souci existe chez l'étudiant de comprendre les textes (articles ou ouvrages), anglo-saxons le plus souvent, qui font appel à des techniques mathématiques. D'autre part, l'étudiant supporte, en général, assez mal d'apprendre des mathématiques, séparément des problèmes concrets de géographie auxquels elles pourraient s'appliquer. La question la plus difficile à éviter, semble-t-il, sur le plan pédagogique, est: « à quoi cela sert-il ? »

Enfin, une dernière remarque porte sur la situation de la géographie, en France, au moins, selon le point de vue partiel que nous avons pu acquérir jusque là.

La plupart des enseignants géographes sont effectivement qualifiés de « littéraires », *i.e.*, ne pratiquant pas les techniques mathématiques. De plus, la géographie, quand il ne s'agit pas de cartographie, ou de géographie physique en général, aborde des problèmes qui figurent aussi dans l'environnement de l'économiste, du planificateur, du sociologue.

Du fait du flou de ces limites, on est conduit, peut-être à tort, à aborder des techniques mathétiques pratiquées en économie, en aménagement du territoire, etc. Par ailleurs, la géographie dans les

* U.E.R. de Mathématique, Logique formelle et informatique, Paris (5e).

pays anglo-saxons, qui semble consommer beaucoup de mathématiques est pratiquée par des gens ayant une formation technique sensiblement différente de celle de l'étudiant français.

Ces trois remarques générales étant faites, il convient de souligner encore que les ouvrages et articles semblent difficilement disponibles et que les étudiants rencontrent quelques difficultés linguistiques.

Nous avons avec M. B. Marchand, tenté une expérience, avec un groupe d'une vingtaine d'étudiants, à l'occasion d'un certificat appelé malheureusement « géographie quantitative ». En effet, il ne s'agit pas de techniques mathématiques surajoutées, permettant de mesurer, mais dans bien des cas, et pour des raisons données plus haut, d'un état d'esprit et d'un vocabulaire condensé et clair dans l'approche et le traitement de certains cas. (Certains appellent cette géographie, « la nouvelle géographie ».) Un enseignement d'algèbre linéaire et de statistiques et un enseignement d'informatique, venaient étayer les études de cas exposées par M. Marchand. Mais il y avait forcément quelques distorsions entre les différents cours, compte tenu des délais nécessaires pour l'explication, la pratique et l'assimilation des notions enseignées.

Nous avons pu remarquer que les titres ci-dessous correspondent globalement aux techniques utilisées ou recommandées dans les articles et ouvrages de géographie quantitative, et par conséquent qu'ils pourraient figurer dans un programme de mathématiques destiné à l'étudiant en géographie, pendant ses quatre années de maîtrise.

1. *Relations, graphes, combinatoire*

Typologie des relations (préordre, équivalence, ordre), matrice booléenne associée, graphe, partitions.

Éléments de vocabulaire de la théorie des graphes et quelques définitions.

Chemins hamiltoniens, etc.

r-combinaisons et r-permutations de n objets avec ou sans répétitions permises.

Dénombrement de partitions.

2. *Espaces vectoriels, calcul matriciel*

Espaces vectoriels; applications linéaires; matrice d'un opérateur.

Notions sur les valeurs propres et vecteurs propres et changement de base.

Calcul matriciel.

Espace métrique, norme, distance euclidienne.

Géométrie des masses; barycentres; produit vectoriel, produit mixte.

3. *Topologie, fonctions, convergence*

Notions élémentaires de topologies, espace métrique.

Fonction réelle d'une variable réelle; graphe; monotonie.

Limite, continuité.

Suites, séries entières, critères de convergence.

Étude de quelques suites et séries remarquables.

4. *Intégration, différentiation*

Notion de mesure positive (Lebesgue); intégrale d'une fonction sur N et sur R; calcul intégral usuel.

Usage d'une table d'intégrales; formule de la moyenne.

Application linéaire tangente, fonction dérivée, dérivée partielle d'une fonction de R dans R; calcul de dérivées usuelles.

Équations différentielles du 1er et 2e ordre; quelques résolutions simples usuelles.

5. *Probabilités et statistique inductive*

Notion de loi de probabilité (binomiale, normale, Poisson, etc.).

Espérance mathématique, variance, écart-type.

Inégalité de Bienaymé-Tchebicheff.

Probabilités conditionnelles; formule de Bayes.

Principe des tests.

Comparaison de moyennes et pourcentages entre elles, respectivement, ou à des valeurs théoriques; tests de Student, Fischer.

Test du chi-deux, entropie (à partir du développement du χ^2).

Analyse de variance (2 et 3 facteurs).

Notion sur les processus markoviens finis.

Présentation de quelques tests non paramétriques (Mann et Witney, Run Test).

6. *Statistique descriptive.*

Éléments de calcul numérique; pratique de la règle, tracé de courbes, échelles.

Mise en forme des données; histogramme, diagrammes en bâtons, fréquences cumulées.

Calcul des moyennes et d'écarts-types, etc.

Notion de corrélation; (test associé) corrélation et régression; droites de régression; changements d'échelles; corrélation multiple.

Analyse en composantes principales; présentation d'autres analyses factorielles.

7. *Géométrie analytique (optionnel).*

Géométrie cartésienne; les différents repères; représentations cartographiques, droites et plans, sphères; équation d'une surface.

Calcul vectoriel; étude des courbes (en coordonnées cartésiennes ou polaires, avec paramètres).

Droites, plan tangent; plan osculateur; courbure; courbures principales, courbes sur une surface; présentation de quelques lignes remarquables (géodésiques, courbes de niveau à partir d'un potentiel).

Potentiel, gradient, champ.

8. *Éléments de mécanique (optionnel).*

Éléments de cinématique; vitesse, accélération; principe fondamental de la dynamique.

Moment d'inertie; notion d'équilibre statique (torseur).

9. *Étude de quelques exemples*

Quelques exemples d'optimisation, à l'occasion de présentation de modèles linéaires; principe de dualité.

Étude de quelques modèles stochastiques et de simulation.

Quant au premier cycle, et d'après les remarques faites lors d'un stage (5-9 octobre 1970, Centre de Mathématique Sociale, Maison des Sciences de l'Homme) portant sur la géographie et les mathématiques, il peut être efficacement soutenu par le cursus suivant:

Algèbre et combinatoire; algèbre linéaire (1, 2).

Topologie générale; éléments d'analyse (3, 4).

Probabilité, statistiques (5, 6).