MATHÉMATIQUES ET SCIENCES HUMAINES

Problèmes d'enseignement. Applications pratiques de lois de probabilité

Mathématiques et sciences humaines, tome 25 (1969), p. 35-40

<http://www.numdam.org/item?id=MSH_1969__25__35_0>

© Centre d'analyse et de mathématiques sociales de l'EHESS, 1969, tous droits réservés.

L'accès aux archives de la revue « Mathématiques et sciences humaines » (http://msh.revues.org/) 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/

M.S.H. (7e année, no 25, 1969, p. 35-40)

PROBLÈMES D'ENSEIGNEMENT

APPLICATIONS PRATIQUES DE LOIS DE PROBABILITÉ

par

B. LECLERC

LOIS DE PARETO *LINGUISTIQUE*

MANDELBROT Benoît. — Information theory and Psycholinguistics : A theory of word frequencies. *Readings in Math. Social Science*, 1966, Paul F. Lazarsfeld & Neil W. Henry, Editors, Chicago, Science Research Associates.

Modèles.

A partir d'un long échantillon d'un texte provenant d'un individu donné, on range les mots apparus dans le texte par ordre de fréquence croissante.

Si $i\,(r, k)$ est le nombre d'apparitions du r-e mot, le texte étant de k mots, on obtient en première approximation :

$$\frac{i\,(r, k)}{k} = \frac{1}{10\,r} \qquad \text{(Loi de Zipf),}$$

ce qui donne une droite sur le papier double logarithmique.

Certains auteurs pensent que cette loi est vérifiée quels que soient l'auteur du texte et la langue dans laquelle il s'exprime. L'auteur remarque qu'elle ne semble pas être habituellement suivie pour les mots les plus fréquents et que la pente de la droite dépend de la « richesse de vocabulaire» du sujet. Il propose un modèle plus général :

$$i\,(r, k) = p\,k\,(r + V)^{-B}$$

où p, V, B sont des paramètres dépendant du sujet. Le modèle de Zipf correspond à $p = 1/10$, $V = 0$, $B = 1$.

Estimation.

B est la pente de la droite du tracé logarithmique. L'auteur ne précise pas ici l'estimation des autres paramètres.

Discussion.

L'auteur considère que, les statistiques en Sciences Sociales étant ce qu'elles sont, la seconde loi est parmi les résultats les mieux établis en ce domaine. En fait, c'est l'une des rares lois qui sont constamment confirmées par l'expérience.

Quelques données sont jointes à l'article sous forme de courbes : textes en allemand, en norvégien et en anglais.

RÉFÉRENCES BIBLIOGRAPHIQUES

ESTOUP Jean-Baptiste. — *Les gammes sténographiques*. Paris, Institut Sténographique, 1916.

MANDELBROT Benoît. — Adaptation d'un message à la ligne de transmission. *Compte rendu des séances hebdomadaires de l'Acad. des Sc. de Paris*, 232, 1951, 1638-40.

MANDELBROT Benoît. — On Recurrent noise limiting coding. In Information Networks, *The Brooklyn Polytechnic Institute Symposium*, 1954, 205-21.

MANDELBROT Benoît. — Linguistique statistique macroscopique. In L. Apostel, B. Mandelbrot & A. Morf. Logique, *Langage et Théorie de l'Information*. Paris, P.U.F.

MANDELBROT Benoît. — On the theory of word frequencies and on related Markovian Models of discours. In Vol. XII, Structure of Language and its Math. Aspects, Proc. of Symposia on Applied Math. Ed. Roman Jakobson. Providence R.I. *American Math. Soc.* 1961.

MANDELBROT Benoît. — New Methods in Statistical Economics. *Journal of Political Economy*, 71, 1963, 421-40, in Bulletin of the International Statistical Institute.

MANDELBROT Benoît. — La Théorie de l'Information est-elle *encore* utile ? in *Le Concept d'Information dans la Science Contemporaine*, Cahiers de Royaumont. Paris, Gauthier-Villars et Éd. de Minuit. 1965, 78-98.

MARKOV A. — Essai d'une recherche statistique sur le texte du roman « Eugène Onéguine». *Bulletin de l'Acad. Impériale des Sc. de St Pétersbourg*. VII. 1913.

MILLER G. A. & CHOMSKY A. Noam. — « Finitary Models of Language Users», in Vol. II. *Handbook of Math. Psychology*, Eds R. R. Bush, E. Galanter and R. D. Luce. New York. Wiley, 1963.

SHANNON C. A. — Math. Theory of Communication. *Bell System Technical Journal*, 28, 1948, 379-423, 623-56.

ZIPF G. K. — *Human behavior and the Principle of Least Effort*. Reading, Mass. : Addison-Wesley, 1949.

LOI DE YULE *PSYCHOLOGIE*

HORVATH William J. — A Stochastic Model for Word Association Tests. *Psychological Review*, 1963, Vol. 70, n° 4, 361-64.

Le modèle appliqué originellement par Yule (1924) à la distribution des genres classés par le nombre de leurs espèces, en biologie, a été repris par Simon (1955) pour la distribution de fréquence des mots dans les textes. Simon le fonde sur deux hypothèses :

1) Soit un texte T de k mots. La probabilité que le $(k+1)$ —e mot ait apparu i fois exactement dans T est proportionnelle au nombre de mots ayant apparu i fois exactement dans T et à i ;

2) La probabilité α que le $(k+1)$ —e mot soit nouveau est constante par rapport à k.

D'où la distribution, f (i) étant le nombre de mots utilisés exactement i fois :

$$f(i) = f^*(1)\, B(i, \rho + 1) \qquad (1)$$

Cette loi est parétienne au sens large : $\dfrac{f(i)}{i^{-\alpha}} \to \text{Cte}$, où $\alpha \geqslant 1$ est une constante, pour $i \to +\infty$.

$f^*(1) = \dfrac{n_k}{2 - \alpha}$ est le nombre de mots n'apparaissant qu'une fois dans un texte de k mots, n_k étant le nombre de mots différents dans le texte, et $\dfrac{\alpha = n_k}{k}$.

B est la fonction β et $\rho = \dfrac{1}{1 - \alpha}$.

Pour α petit et $\rho \simeq 1$, (1) se simplifie en $f(i) = \dfrac{n_k}{i(i+1)}$

et pour i assez grand $f(i) = \dfrac{n_k}{i^2}$.

Estimation.

Approximation de α par la donnée empirique. D'où calcul de f* (1), puis des f (i) pour obtenir la distribution théorique correspondant à la valeur de α estimée.

Test du χ^2.

Test du χ^2 dans l'application exposée par l'auteur. Il s'agit ici du test de Kent-Rosanoff d'association de mots. L'auteur présente des résultats pour les quatre mots-tests : table, Music, Blossom et High.

Exemple. 1 009 personnes ont répondu au mot table, donnant 47 mots-réponses différents. Le résultat peut être résumé dans le tableau suivant :

Nombre i d'occurences d'un mot réponse particulier	Nombre de mots répondus i fois	Valeurs théoriques de la loi de Yule après estimation de α
1	28	24
2	7	8
3	3	4
4 ou plus	9	11

$\chi^2 = 1{,}41$ ayant une probabilité 0,7 d'être dépassée. Les probabilités analogues pour les trois autres mots sont 0,4 ; 0,7 et 0,6.

RÉFÉRENCES BIBLIOGRAPHIQUES

Russel W. A. & Jenkins J. J. — The complete Minnesota norms for responses to 100 words for the Kent-Rosanoff word association test. *Technical Report*, n° 2, 1954, Univ. of Minnesota, Department of Psychology.

Simon H. A. — On a class of skew distributions functions. *Biometrika*, 1955, 42, 425.

Skinner B. F. — The distribution of associated words. *Psychol. Rec.*, 1937, 1, 71-76.

White H. — Chance models of systems of casual groups. *Sociometry*, 1962, 25, 153.

Yule G. U. — A mathematical theory of evolution based on the conclusions of Dr. J. C. Willis F.R.S. *Phil. Trans.* B., 1924, 213, 21.

LOI BINOMIALE NÉGATIVE — *DISTRIBUTION DES ACCIDENTS*

Sichel Herbert S. — The estimation of the parameters of a negative binomial distribution with special reference to psychological data. *Psychometrika*, Vol. 16, n° 1, March 1951, 107-127.

Modèle.

On considère une population d'individus susceptibles d'avoir des accidents. Pour un individu déterminé, la probabilité d'avoir r accidents en une unité de temps, est supposée suivre la loi de Poisson.

$$P_\lambda (r) = \frac{e^{-\lambda}\, \lambda^r}{r\,!}.$$

Le paramètre réel positif λ représente la propension aux accidents de cet individu. Dans la population entière, on suppose que λ est distribué suivant une loi γ de loi élémentaire :

$$\mathrm{d}\,\mathrm{F}(\lambda) = \frac{c^p}{\Gamma(p)}\, e^{-c\lambda}\, \lambda^{p-1}\, \mathrm{d}\,\lambda$$

où p et c sont des paramètres réels positifs.

La distribution du nombre d'accidents par individu et par unité de temps est donc, $\forall\ r$ entier positif ou nul :

$$f(r) = \int_{\lambda=0}^{\infty} P_\lambda(r)\, \mathrm{d}\,\mathrm{F}(\lambda) = \frac{\Gamma(r+p)}{\Gamma(p)\,\Gamma(r+1)} \left(\frac{c}{c+1}\right)^p \left(\frac{1}{c+1}\right)^r$$

r suit la loi binômiale négative, ou loi de Greenwood-Yule.

Estimation.

Le but de cet article est de confronter deux méthodes d'estimation : la méthode des moments, habituellement employée, et la méthode du maximum de vraisemblance. L'auteur se propose de montrer que le manque d'efficacité de la première peut être à l'origine de refus non justifiés du modèle.

L'auteur cherche à estimer p et $m = \frac{p}{c}$.

La méthode des moments donne les estimateurs $\overline{m}$ et $\overline{p}$:

$$\overline{m} = m' \qquad \overline{p} = \frac{m'^2}{s^2 - m'}$$

où m' et s^2 sont la moyenne et la variance empirique.

La méthode du maximum de vraisemblance donne les estimateurs $\hat{m} = \overline{m}$ et $\hat{p}$ solution d'une équation assez compliquée où intervient la fonction digramma. Des tables sont jointes à l'article pour faciliter l'évaluation de $\hat{p}$.

Test.

L'ajustement est mesuré par la distance du χ^2. Le calcul des quantités-tests est donné de façon précise. Les probabilités pour des variables du χ^2 d'excéder leur valeur sont indiquées.

Applications.

Données numériques dont les origines ne sont pas précisées.

1) Nombre d'absences d'ouvriers d'une aciérie durant une période de six mois. L'ajustement de la méthode des moments donne une quantité-test ayant la probabilité 0,03 d'être dépassée. Le modèle doit être rejeté, alors qu'il est acceptable pour l'ajustement du maximum de vraisemblance, qui donne une quantité-test ayant une probabilité 0,18 d'être dépassée.

2) et 3) Distribution des accidents mineurs de deux groupes d'ouvriers.

4) Distribution du nombre d'erreurs dans un test de coordination des deux mains.

Dans ces exemples, le modèle est toujours acceptable, mais l'ajustement du maximum de vraisemblance donne chaque fois une quantité-test nettement plus faible, et peut donc être considérée comme meilleur.

RÉFÉRENCES BIBLIOGRAPHIQUES

FISHER R. A. — On the mathematical foundations of theoretical statistics. *Philosophical transactions*, 1921, A 222, 309.

FISHER R. A. — *Statistical methods for research workers*. Edinburgh, Oliver & Boyd, 1948.

FISHER R. A. — The negative binomial distribution. *Annals of Eugenics*, 1941, 11, 179.

GREENWOOD M. & Yule G. U. — An inquiry into the nature of frequency distributions of multiple happenings, *J. Roy. Stat. Soc.*, 1920, 83, 255.

KENDALL M. G. — *The advanced theory of statistics*, London, Griffiths, 1945.

HALDANE J. B. S. — The fitting of binomial distributions, *Annals of Eugenics*, 1941, 11, 179.

MARITZ J. S. — On the validity of inferences drawn from the fitting of Poisson and negative binomial distributions to observed accident data, *Psychological Bulletin*, 1950, 47, 434.

CHAMBERS E. G. & Yule G. U. — Theory and observations in the investigation of accident causations, *Supp. J. Roy. Stat. Soc.*, 1941, 7, n° 2, 09.

LOI DE POISSON
LOIS BINOMIALES
LOI HYPERGÉOMÉTRIQUE

DISTRIBUTION DES ACCIDENTS
DISTRIBUTIONS DISCRÈTES

GULDBERG Alf. — Fonctions de fréquences discontinues et séries statistiques. *Annales de l'Institut Henri Poincaré*, fasc. 3, Vol. 3, 1933, 229-278.

L'article porte sur les méthodes pour approximer une série statistique à valeurs entières par une fonction de fréquence (loi de probabilité).

L'auteur résume les méthodes de K. Pearson d'une part, et de Gram-Charlier d'autre part, ainsi que les critiques qui en ont été faites.

Modèles.

Il donne ensuite une méthode d'ajustement pour quelques lois discrètes. En désignant la fréquence théorique de x entier par $f(x)$.

Loi de Poisson de paramètre λ :

$$f(x) = \frac{e^{-\lambda}\,\lambda^x}{x\,!}$$

Loi binômiale de paramètres k et p, k entier, $p \in\,]\,0,1\,[$:

$$f(x) = C_k^x\, p^x\, (1 - p)^{k-x}$$

Loi de Pascal de paramètre k entier (ou loi binômiale négative) :

$$f(x) = C_{k+x}^k\, p^k\, q^x$$

Loi hypergéométrique :

$$f(x) = \frac{C_k^x\, C_h^{m-x}}{C_{k+h}^m}$$

Ajustement.

Pour chacune de ces lois, l'auteur établit une formule liant $\dfrac{f(x+1)}{f(x)}$ aux premiers moments, et en tire un critère pour déterminer si une série statistique peut être ajustée à la loi considérée.

Loi de Poisson :

$$\frac{f(x+1)}{f(x)}(x+1) = \lambda = m.$$

D'où :

$$\alpha(x) = \frac{f(x+1)}{f(x)} \frac{(x+1)}{m} = 1.$$

Loi binômiale et loi de Pascal :

$$\frac{f(x+1)}{f(x)}(x+1) + \frac{m-\sigma^2}{\sigma^2} x = \frac{m^2}{\sigma^2}$$

D'où :

$$\alpha(x) = \frac{\frac{f(x+1)}{f(x)}(x+1) + \frac{m-\sigma^2}{\sigma^2}}{\frac{m^2}{\sigma^2}} = 1,$$

avec pour la loi binômiale : $m > \sigma^2$,

pour la loi de Pascal : $m < \sigma^2$,

et pour la loi de Poisson: $m = \sigma^2$-.

Pour la loi hypergéométrique, on aboutit à une formule non simple liant $f(x+1)$, $f(x)$, x, et les trois premiers cumulants.

En remplaçant dans ces formules fréquences et moments théoriques par leurs homologues empiriques, on constate si la série statistique donnée se rapproche de la distribution théorique considérée. Les paramètres s'estiment alors par identification des premiers moments, pris en nombre suffisant.

Nombreux exemples.

Nombre de soldats tués par coups de pieds de cheval dans l'armée prussienne entre 1875 et 1894, par an et par corps d'armée (loi de Poisson).

Nombre de glandes de Muller de la patte droite antérieure de 2 000 truies (loi de Poisson non satisfaisante — loi binômiale).

Nombre de suicides par an, des personnes assurées auprès d'une société d'assurance sur la vie en Norvège, de 1900 à 1929 (loi de Poisson).

Nombre de particules α rayonnées par le polonium dans un temps donné (loi de Poisson).

Nombre de zéros dans les colonnes du livre *Thesaurus logarithmorum*, de Véga (loi binômiale).

Nombre par an de suicides d'enfants en Prusse, pend ant la période 1869-1893 (loi de Pascal — loi de Poisson (approximation de Pascal meilleure, par l'auteur)).

Nombre de fois où l'avant-main dans le jeu de whist a reçu x atouts dans 25 000 donnes (K. Pearson) (loi hypergéométrique).