MATHÉMATIQUES ET SCIENCES HUMAINES

Problèmes d'enseignement

Mathématiques et sciences humaines, tome 28 (1969), p. 59-65

<http://www.numdam.org/item?id=MSH_1969__28__59_0>

M.S.H. (7e année, no 28, 1969, p. 59-64)

PROBLÈMES D'ENSEIGNEMENT

APPLICATIONS PRATIQUES DES LOIS DE PROBABILITÉ (5)

par

B. LECLERC

LOI DE PASCAL *PSYCHOLOGIE*

LOI DE FISHER

C. B. WILLIAMS, « The number of publications written by biologists », *Annals of Eugenics*, 12, 1964, p. 143-146.

Modèles.

Deux modèles sont ajustés à des données portant sur le nombre de publications de chercheurs en biologie.

Loi géométrique (ou loi de Pascal, pour divers auteurs) : le nombre n_k de chercheurs ayant publié k articles est approximativement donné par :

$$n_k = n_1 x^{k-1} \qquad k \geqslant 1$$

où x est une constante ($0 < x < 1$).

Loi logarithmique de Fisher :

$$n_k = n_1 \frac{x^{k-1}}{k} \qquad \begin{matrix} k \geqslant 1 \\ (0 < x < 1) \end{matrix}$$

Estimation.

Si N est le nombre total de publications et S le nombre d'auteurs, on a pour la série géométrique :

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

et pour la série logarithmique :

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

à partir de quoi l'on peut calculer n_1 et x, S et N étant connus. Dans les deux cas, la distribution ajustée est donc calculée à partir de N et S.

Application.

L'auteur reprend d'abord des données auxquelles J. Dufresnoy avait ajusté une loi géométrique : le recensement, dans *Review of applied Mycology* (1935) de 2 229 publications par 1 527 chercheurs. L'auteur ajuste ces données à une loi géométrique (différente de celle de Dufresnoy, qu'il précise) et à une loi logarithmique. Sans faire de test, il fournit les tableaux numériques et les tracés sur papier logarithmique. L'ajustement à la loi logarithmique est nettement le meilleur.

Ce travail est ensuite repris pour deux autres données : 656 publications par 411 auteurs, puis 2 375 publications par 1 534 auteurs, recensés dans *Review of applied Entomology*, respectivement en 1913 et en 1936. Les conclusions sont les mêmes que précédemment. L'auteur pense que la loi logarithmique rend mieux compte que la loi géométrique de l'effet suivant : un chercheur qui a publié un grand nombre d'articles en fera plus facilement un supplémentaire qu'un chercheur qui n'en a publié qu'un.

RÉFÉRENCES BIBLIOGRAPHIQUES

DUFRESNOY J., « The publishing behavior of biologists », *Quart. Rev. Biol.*, *13*, 1938, p. 207.

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

LOI DE PARETO *ÉCONOMIE*

D. G. CHAMPERNOWNE, « The graduation of Income distributions », *Econometrica*, vol. 20, n° 4, octobre 1952, p. 591-615.

Modèle.

Dans un article paru en 1937, l'auteur avait proposé l'ajustement de la distribution des revenus à la courbe :

$$\varphi(x) = \frac{n}{\operatorname{ch} \alpha (x - x_0) + \lambda}$$

avec quatre paramètres n, α, x_0, λ ($\lambda > -1$), et où x est le logarithme népérien des revenus.

Ceci conduit, si $F(t)$ est le nombre de personnes ayant un revenu supérieur ou égal à t, aux formules :

$$F(t) = \frac{N}{\Theta} \operatorname{tg}^{-1} \left(\frac{\sin \Theta}{\cos \Theta + \left(\frac{t}{t_0} \right)^{\alpha}} \right)$$

si $\lambda < 1$

où Θ est exprimé en degrés et $\cos \Theta = \lambda$

$$F(t) = \frac{N t_0^\alpha}{t_0^\alpha + t^\alpha}$$

si $\lambda = 1$

$$F(t) = \frac{N}{2\eta} \log_{10} \left(\frac{t^\alpha + 10^\eta t_0^\alpha}{t^\alpha + 10^{-\eta} t_0^\alpha} \right)$$

avec $\text{ch}\, \eta = \lambda$ si $\lambda > 1$

où $N = F(0)$ est le nombre de gens ayant un revenu, t_0 est la médiane de la distribution des revenus, α est le coefficient de Pareto.

L'auteur précise que, généralement, on est dans le cas $-1 < \lambda < 1$.

Estimation.

Sept méthodes sont proposées par l'auteur pour l'estimation de t_0 et Θ. Le tracé sur papier logarithmique donne pour les grandes valeurs de t :

$$y = \beta - \alpha \log t$$

d'où l'estimation de α et β (méthode non précisée).

Si N est connu :

1. t_0 est obtenu par identification des médianes empirique et théorique, Θ s'obtient en fonction de t_0 et β.

2) Identification des moyennes empirique et théorique.

3) Identification des modes.

Si N est inconnu :

4) Identification de $g_{\max'}$, plus grand revenu total de la population d'une classe de revenus avec $t_0 \dfrac{dF}{dt}(t_0)$ et du niveau de la classe correspondante avec t_0.

5) Identification de $g_{\max}$ comme précédemment et de T, revenu total de la population étudiée.

6) Identification de $F(t_1)$, t_1 convenablement choisi et de T.

7) Identification de $F(t_1)$ et de $F(t_2)$, t_1 et t_2 convenablement choisis.

L'auteur donne toutes les formules nécessaires.

Applications et discussion.

Revenus annuels nets aux États-Unis, 1918 (en dollars) : tableau et courbe donnés par l'auteur.

Application des méthodes 1) à 6) :

$$N = 37\,569\,000 \qquad \alpha = 1{,}59739 \qquad \beta = 11{,}85104$$

L'estimation de t_0 varie entre 1 068,6 et 1 201 selon la méthode employée.

Celle de Θ oscille autour de 139°.

Si N est supposé inconnu, les méthodes 4), 5), 6) donnent respectivement :

$$N = 36\,370\,000 \qquad N = 34\,881\,000 \qquad N = 35\,260\,000$$

Application de la méthode 3) aux données japonaises de Hayakawa (voir fiche correspondante, dans *M.S.H.*, n° 26). On a N = 13 941 085 et l'on obtient :

$$\alpha = 1{,}55837 \qquad \beta = 10{,}84448 \qquad t_0 = 1\,140 \qquad \Theta = 142{,}3°.$$

L'auteur présente dans un tableau sa courbe théorique et celles de Hayakawa et de H. T. Davis. Il ne redonne pas les chiffres empiriques de Hayakawa.

Application de la méthode 7 aux revenus des citadins norvégiens en 1930 (en milliers de couronnes). On trouve :

$$\alpha = 2 \quad \text{et} \quad \beta = 5{,}9 \qquad \Theta = 117{,}8° \qquad t_0 = 3{,}35.$$

L'auteur donne dans un tableau les valeurs de t_0 et α pour 21 distributions dont la première citée ici. Il signale que son modèle peut fréquemment être simplifié en la forme :

$$F(t) = \frac{N}{90} \operatorname{tg}^{-1} \left(\frac{t_0}{t}\right)^{\alpha}.$$

Il donne les ajustements de la distribution des revenus en Bohême (1933) avec ses deux modèles, et avec les modèles à quatre paramètres de Pareto, Davis et Gibrat. Il ne donne pas les formules de bases de ceux-ci, mais les formules avec les valeurs numériques des paramètres calculées.

Pareto :

$$F(t) = 124\,500\,000\,(t + 8)^{-2{,}1}$$

Gibrat :

$$-\frac{dF}{dt} 455\,000 \exp\left(-2{,}509 \log \left(\frac{t-3}{9}\right)^2\right)$$

Davis :

$$-\frac{dF}{dt} = \frac{1\,907\,200\,000}{(t-2)^{3{,}94}\left(\exp \frac{2{,}31219}{t-2} - 1\right)}$$

Enfin l'auteur termine en donnant encore quelques exemples avec les tableaux détaillés :

Royaume-Uni : 1938-1939.

U.S.A. : 1947.

Il n'émet pas de jugements sur la qualité des ajustements.

RÉFÉRENCES BIBLIOGRAPHIQUES.

CHAMPERNOWNE D G., « The theory of Income Distribution », *Econometrica*, vol. 15, p. 379-381.

DAVIS H. T., « The analysis of Economic Time-Series », *Cowles Commission monograph*, n° 6, Evanston, Ill., Principia Press, 1941.

GIBRAT R., *Les inégalités économiques*, Paris, Recueil Sirey, 1931.

HAYAKAWA M., « The application of Pareto's Law of Income to Japanese Data », *Econometrica*, vol. 19, p. 174-183.

PARETO V., *Cours d'économie politique*, vol. 2, Paris, F. Pichon, 1897.

LOI NORMALE *DURÉES DE VIE*

J. P. GIVRY, I. « Gestion d'un parc aléatoire ». — II. « Estimation de la loi de probabilité des durées de vie », *Revue de Statistique Appliquée*, vol. 13, n° 1, p. 5-47, et n° 2, p. 5-28.

Modèle.

La distribution des durées de vie de cathodes de cellules d'électrolyse pour la production de l'aluminium s'est révélée être sensiblement gaussienne. Cette normalité a été testée sur une quarantaine de lots de 20 à 150 cathodes chacun.

Estimation.

Le problème est d'estimer la moyenne m et l'écart-type σ de la distribution sans attendre la mort de toutes les cathodes, c'est-à-dire à partir d'un échantillon tronqué. Toutes les cathodes ne sont pas entrées en service au même moment.

Une estimation de $F(x)$, probabilité pour une cathode d'avoir une vie de durée supérieure à x est donnée à chaque instant par la fraction des cathodes qui ont dépassé en service l'âge x parmi celles qui ont une ancienneté supérieure à x. Le tracé d'une droite de Henry donne une estimation ponctuelle du couple (m, σ).

Les $F(x)$ sont également estimés par intervalle : on détermine pour chacun d'eux un intervalle de confiance à 95%. Un point d'un tel intervalle représente un couple $(x, F(x))$ admissible. On écrit :

$$F(x) = G\left(\frac{x - m}{s}\right) = G(r)$$

où G est la fonction de répartition de la loi normale $N(0, 1)$. r est lu dans les tables.

La droite $m = x + rs$ du plan (m, s) est admissible. Le domaine des points admissibles se ferme et se resserre lorsque le nombre de décès augmente. On obtient ainsi une estimation par domaine du couple (m, s).

Tests.

Vérification de la linéarité du tracé partiel sur le graphique de Henry.

Test du χ^2 au seuil 5%, avec regroupement des classes d'observations, pour obtenir des nombres théoriques au moins égaux à 5.

RÉFÉRENCE BIBLIOGRAPHIQUE

GUPTA A. K., « Estimation of the mean and standard deviation of a normal population from a censored sample » (lieu de publication non précisé).

M. L. DUFRESNOY, « Statistique linguistique appliquée aux *Lettres Persanes* », *Journal de la Statistique de Paris*, 2e trimestre 1966, p. 130-134.

L'auteur s'intéresse à la distribution des longueurs de phrases (nombre de mots par phrase) dans les *Lettres Persanes* de Montesquieu.

Modèles.

Sans préciser la forme analytique de la loi, l'auteur opte pour la distribution log-normale comme représentant les résultats de la fragmentation « au hasard » d'une quantité initialement indivise de matériaux fragmentables.
Estimation et *test* non précisés.

Applications.

Cinq passages des *Lettres Persanes*, pour la plupart des pastiches d'auteurs contemporains (tels que Fénelon, La Bruyère) sont étudiés. Trois tracés sur papier gausso-logarithmique sont donnés. L'un d'eux notamment, permet de comparer « Les Troglodytes » (Lettres XI à XIV) à « La Bétique » (extrait de *Télémaque* de Fénelon). Pour celle-ci et pour les extraits des *Lettres Persanes*, l'auteur détermine les quantiles correspondant aux proportions 16%, 50% et 84%, comme valeurs caractéristiques de centralité et de dispersion.

RÉFÉRENCES BIBLIOGRAPHIQUES. Une vingtaine de titres, dont :

YULE G. U., « On sentence length as a statistical characteristic of style in prose », *Biometrika 30*, III et IV, 1939.

DUFRESNOY M. L., « Analyse statistique du langage », *J. Soc. Stat. 97*, 1946, p. 208-218.

HERDAN G., *Language as choice and chance*, Groningen, 1956.

WAKE, « Sentence length distributions », *J. Roy. Stat. Soc. 120*, 1957, p. 331-346.

GUIRAUD P., *Problèmes et méthodes de la statistique linguistique*, Paris, 1960.

HERDAN G., *Quantitative linguistics*, London, 1964.

PAPP F., « Mathematical linguistics in the Soviet Union », *Acta linguistica Acad. Sci. Hungar., 14*, 1964, p. 119-137.

(A suivre.)

M.S.H. (7e année, n° 28, 1969, p. 65)

CONTE POUR COMPTER

par

P. JULLIEN

Le Conseil est réuni. Il s'agit ce soir, de constituer l'équipe qui défendra les couleurs du village aux jeux intercommunaux. Ce n'est pas facile de désigner $(p + 1)$ joueurs, dont l'un sera le roi, les autres les pions, parmi ces n jeunes gens, si dynamiques, tous également capables d'être joueurs, et même d'être roi.

Hubert dit : « Choisissons d'abord les joueurs, puis le roi parmi les joueurs. — Non ! crie Isidore, choisissons d'abord les pions, puis le roi parmi les restants. — Je propose mieux, intervient Jules, choisissons d'abord le roi, puis les pions parmi les restants. » Un débat passionné s'instaure sur les avantages comparés de ces diverses solutions.

Dans son coin, Alfred, que ces joutes oratoires n'intéressent pas, constate que chaque méthode proposée permet de calculer le nombre d'équipes possibles. Manifestement suivant Hubert, ce nombre est :

$$C_n^{p+1} \times (p + 1),$$

suivant Isidore :

$$C_n^p \times (n - p)$$

et suivant Jules :

$$n \times C_{n-1}^p.$$

Ainsi poursuivant son idée, Alfred griffonne la double égalité :

$$(p + 1)\, C_n^{p+1} = (n - p)\, C_n^p = n\, C_{n-1}^p.$$

De la première en fixant n, il fait une récurrence sur p, à partir de $p = 0$, car il sait que $C_n^0 = 1$. Il retrouve avec satisfaction :

$$C_n^k = \frac{n\,(n-1) \ldots (n+1-k)}{1 \times 2 \times \ldots \times k}$$

De la deuxième en fixant p, il fait une récurrence sur n, à partir de $n = p + 1$, car il sait que $C_p^p = 1$. Il trouve, (le saviez-vous ?) :

$$C_{p+k}^p = \frac{(p+1)\,(p+2) \ldots (p+k).}{1 \times 2 \times \ldots \times k}$$

C'est bien beau les mathématiques, mais le Conseil est terminé et Alfred n'est pas capable de dire, quelle a été l'équipe désignée...