ANALYSES BIBLIOGRAPHIQUES

LINDSEY J. K.
Introductory Statistics : A Modelling Approach.
ISBN 0-19-852345-9 ; cartonné ou non ; 23,2 x 15,5 x 1,5 cm ; XI + 214 pages ; Oxford University Press, Walton Street, OX2 6DP, 1995.

A la lecture de ce manuel, on suit aisément le cheminement de l'auteur qui aborde les statistiques par le biais de la notion intuitive de probabilité. En guise d'introduction, on trouve des exemples de données simples ; ces données peuvent logiquement se résumer sous la forme de tableaux de contingence ; au niveau de ces tableaux, on peut facilement tester l'indépendance d'une variable réponse par rapport à une ou plusieurs variables explicatives ; la relation entre variables peut le plus souvent s'écrire sous forme d'une équation ; et ceci nous amène tout naturellement au principe de l'élaboration de modèles décrivant les relations entre variables étudiées. Ainsi, au fil du livre, les non-mathématiciens vont peu à peu parvenir à « mettre un visage » sur un certain nombre de termes fréquemment trouvés dans le chapitre Matériel & Méthodes de publications récentes dans leur discipline ou encore dans le menu de leur programme de statistique préféré (régression logistique, modèles log-linéaires, modèles linéaires généralisés,...) ; on rappelle aussi, sous un autre jour, dans le contexte de la modélisation, des tests connus comme la régression linéaire simple, le t de Student ou l'ANOVA.

De nombreux exercices permettent, à l'aide d'une machine à calculer, d'ajuster soi-même des modèles simples à des ensembles de données présentés dans le texte.

Au fur et à mesure des chapitres, un certain nombre d'exemples précédemment donnés sont réutilisés afin de compléter et d'affiner les analyses effectuées de même que les notions qu'ils permettent d'introduire. Ainsi, partant du plus simple vers le plus compliqué, on se familiarise avec la terminologie et les principes de base des modèles. Bien que s'adressant initialement à des étudiants, ce manuel permettra au chercheur d'une discipline non mathématique de mieux assimiler les traitements appliqués par certains de ses collègues et, parallèlement, d'entrevoir avec intérêt les applications potentielles de la modélisation dans le cadre de sa recherche personnelle.

La **préface** nous explique entre autre que le fil conducteur du manuel est la construction de modèles visant à expliquer la structure des données récoltées : un modèle décrit comment une distribution de probabilité change de forme dans les différents sous-groupes d'une population. Cette démarche de modélisation est d'application dans un ensemble très vaste de domaines, l'interprétation concrète des résultats restant toujours la finalité du traitement.

Le **premier chapitre** établit les concepts de base. On y trouve des définitions simples et pratiques de termes tels que variable, probabilité ou échantillon. L'importance de la présentation des données sous la forme de tableaux de contingence et de graphes (histogrammes) est soulignée. Deux importantes lois des probabilités (Loi de produit et Loi d'addition) sont formulées et la notion de probabilité conditionnelle à la base de la construction des modèles statistiques est introduite. La distribution multinomiale est définie, ainsi que la fonction de densité de probabilité. La distinction est faite entre variable(s) réponse(s) et variable(s) explicative(s). L'importance de l'établissement d'un protocole de prise de données et du plan de recherche est évoquée.

Dans le **chapitre deux**, on étudie les modifications de la distribution multinomiale au niveau de sous-groupes d'une population par l'utilisation de modèles appelés logistiques. Ces modèles pour données discrètes sont applicables à n'importe quelle réponse ; néanmoins, ce type de modélisation requiert un nombre suffisant d'observations ainsi que le calcul d'un nombre de paramètres relativement élevé.

On voit, sur base de calculs de probabilités, comment mesurer le degré d'indépendance des variables dans un tableau de contingence grâce à la notion de cote (*odds* en anglais, par similarité avec la terminologie des courses de chevaux) utilisée sous la forme du rapport de cote ou rapport du produit croisé.

Les modèles linéaires sont introduits par leur champs d'utilisation, leur formulation en équations et la résolution de ces équations pour trouver les paramètres inconnus qui décriront la variation d'une variable réponse en fonction d'une ou de plusieurs variable(s) explicative(s).

L'introduction du rapport de cote dans le modèle linéaire simple permet d'éviter certains problèmes inhérents aux probabilités ; on aborde alors le modèle logistique qui se résout en appliquant certaines contraintes. Peu à peu, on envisage des cas de plus en plus complexes où la variable explicative n'est plus binaire mais peut prendre plusieurs valeurs, puis lorsque plusieurs variables explicatives interviennent et interagissent pour produire des effets différents. Le passage à une variable quantitative et non plus qualitative permet de simplifier les équations au niveau du modèle de régression logistique. A nouveau, on envisage la cas d'une variable réponse binaire, puis polytomique (peut prendre plus de deux valeurs). Lorsque, parmi les différentes variables étudiées, on ne peut clairement faire la distinction entre la réponse et les variables explicatives, mais qu'on s'intéresse plutôt à la distribution commune de ces variables, on applique un modèle log-linéaire pour données discrètes. Et, d'une manière similaire à la régression logistique, on peut étudier des variables explicatives non plus qualitatives mais quantitatives en développant un modèle de régression log-linéaire. En ce qui concerne les réponses de type ordinal, il est possible de respecter la notion d'ordre des catégories de la réponse en développant des modèles adaptés (modèle avec continuation du rapport ou modèle avec cotes proportionnelles).

Le **troisième chapitre** introduit la notion d'inférence : dans quelle mesure les relations entre variables étudiées au niveau d'un échantillon

s'appliquent-elles à la population ? Dans un modèle, on peut fixer les valeurs des paramètres des équations et calculer les probabilités d'obtenir différents échantillons. Inversement, il est possible de retrouver la valeur des paramètres qui rendent l'échantillon le plus probable, c'est l'estimation correspondant à la vraisemblance maximale. On définit, à partir de cette valeur, la déviance qui constitue une mesure de la différence entre un modèle élaboré et le meilleur modèle possible.

En fait, pour décrire des données, il existe un grand nombre de modèles possibles et plus le nombre de paramètres intervenant dans un modèle est élevé, plus il sera fidèle à la structure observée des données. En comparant deux modèles, on prend en compte la différence entre le nombre de paramètres intervenant dans chacun des modèles, c'est-à-dire le nombre de degrés de liberté.

Paradoxalement, le fait de bien représenter l'échantillon ne signifie pas forcément qu'un modèle représente bien la population : plus un modèle est complexe (plus il fait intervenir de paramètres) moins il est facile à interpréter ; ensuite, un modèle trop proche des données d'un échantillon aura moins de chance de correctement décrire un autre échantillon de la même population. La modélisation est donc question de compromis entre l'adéquation et la simplicité du modèle.

A ce niveau, sont introduits les tests de signification fixant des seuils de probabilité au-delà desquels un modèle est rejeté. Le choix entre différents modèles établis pour les mêmes données peut aussi se faire grâce au critère d'information d'Akaike. Le modèle présentant le plus petit AIC est le meilleur.

Les notions de modèle saturé et de résidus sont encore abordées, ainsi que la méthode de calcul de la taille de l'échantillon nécessaire à la détection des différences de distribution entre sous-groupes.

Les modèles vus précédemment étaient basés sur une distribution multinomiale et bien adaptés aux réponses nominales, sans condition quant à la forme de leur distribution. Dans le **chapitre quatre**, on aborde les données quantitatives telles que l'on peut contraindre leur distribution à prendre une forme théorique. On considère que la population étudiée est homogène et que la même distribution s'applique à tous les individus. Dans la plupart des cas, on simplifie encore en considérant que seule change dans les sous-populations la position de la distribution, déterminée par le paramètre de moyenne. Les avantages d'imposer une distribution théorique aux données sont de simplifier les équations du modèle, de lisser l'histogramme rendant la courbe plus facile à interpréter, d'augmenter la probabilité que le modèle soit applicable dans d'autres études similaires et d'obtenir des informations sur le mécanisme par lequel les données sont générées. Le prix à payer pour ces avantages est de se plier à un certain nombre de conditions d'application sur la forme de la fonction de densité.

Les distributions de probabilité ou fonctions de densité décrites dans ce chapitre sont triées sur base de leur cadre d'applications (certaines conviennent bien à des comptages, d'autres à des mesures, d'autres à des durées) ; elles sont illustrées graphiquement et on trouve la formule permettant de calculer leur moyenne et leur variance ainsi que leur adéquation maximale (plus la déviance

est élevée, plus la distribution de la fonction de densité testée est éloignée de la distribution multinomiale correspondant parfaitement aux données). On peut également utiliser l'AIC : la distribution testée est plausible lorsque l'AIC de cette distribution est plus petit que celui de la distribution multinomiale.

Au **chapitre cinq**, les particularités de la distribution normale sont mises en valeur pour simplifier encore les modèles et étendre leur champs d'action ; par contre, les conditions d'application sont plus strictes (notamment une variabilité constante et une courbe de réponse symétrique en cloche dans tous les segments de la population). L'équation du modèle de régression linéaire simple — le plus connu des modèles linéaires généralisés — est analysée ; ses deux paramètres, l'interception et la pente, peuvent être estimés même avec un très petit nombre d'observations.

Les analyses de variance ou ANOVA sont également présentées comme des modèles similaires à ceux évoqués au chapitre deux mais sous l'hypothèse d'une distribution normale des données et d'une variance identique dans tous les sous-groupes formés sur base d'une variable explicative nominale. La généralisation du modèle linéaire à deux variables explicatives est directe. On étudie l'effet des deux variables séparément puis de leur interaction. L'analyse de covariance vient ensuite, où l'on combine régression linéaire et analyse de variance.

Si deux variables réponses continues ont une distribution commune de type normal, elles peuvent avoir des moyennes et des variances différentes : le coefficient de corrélation qui décrit la dépendance entre ces deux réponses est positif (négatif) lorsqu'une des variables étant supérieure à sa moyenne, la plus forte probabilté est que l'autre variable soit également supérieure (inférieure) à sa moyenne ; on rappelle aussi que dans le cas de variables formant une distribution normale, un coefficient de corrélation nul indique l'indépendance des variables.

Comme pour les modèles logistique et log-linéaire, on peut calculer la taille de l'échantillon minimum nécessaire pour déceler des différences entre distributions de la réponse dans des sous-groupes formés sur base de variables explicatives. La taille de l'échantillon est directement proportionnelle à la variance dans le modèle et inversement proportionnelle au carré de la différence entre les moyennes.

Dans le **sixième chapitre**, on aborde la modélisation de données qui ne sont plus indépendantes les unes des autres. On parle de mesures répétées lorsque la variable réponse est mesurée plusieurs fois pour chaque individu. Un processus ponctuel est une série de un et de zéro codant l'occurrence ou l'absence d'occurrence d'un événement à chaque instant (à chaque point). Si on considère que chaque réponse ne dépend que de celle qui la précède, on parle de chaîne de Markov. Une approche possible pour modéliser ce genre de données est proposée. On définit également la fragilité : le fait que certains individus, toutes variables explicatives étant identiques, vont répondre différemment que d'autres.

En outre, un modèle semblable aux chaînes de Markov peut être construit lorsqu'une variable réponse continue et distribuée normalement est observée dans le temps : c'est l'autorégression ; le paramètre de pente est appelé autocorrélation et donne la corrélation entre réponses consécutives.

Quand, au lieu de s'étaler dans le temps, plusieurs observations d'un individu sont faites à peu près simultanément, on peut créer un modèle d'effets aléatoires : pour une distribution normale, on utilise une analyse de variance avec plusieurs observations de la réponse pour chaque individu et une variable explicative nominale distinguant entre ces réponses.

Pour terminer, les tables de survie sont abordées ; l'idée de base d'une table de survie est le suivi d'individus dans le temps et le relevé, au cours d'une série de périodes, du nombre d'entre eux qui subissent certains événements. Au cours de chaque période, un individu peut subir un événement ou pas, ce qui peut également se modéliser.

En conclusion, l'auteur insiste sur le fait que ce livre ne donne qu'un simple aperçu des très nombreux types de modèles existants et en développement. Il nous rappelle que la statistique est un outil qui devrait aider les chercheurs de nombreux domaines à la conception du plan de recherche, à l'enregistrement clair et efficace des données, à la détection des erreurs, des valeurs aberrantes ou des résultats inattendus, à la description des relations entre variables étudiées, à suggérer les mécanismes par lesquels les données ont pu être produites ainsi qu'à l'interprétation et à la communication des résultats.

Ce livre a notamment été écrit dans l'idée de faciliter le dialogue entre chercheurs de différentes disciplines et les statisticiens. En outre, il peut nous aider à mieux utiliser les analyses statistiques générées par ordinateur en s'assurant de la bonne compréhension des résultats fournis.

Enfin, la modélisation d'un ensemble de données est résumée de la manière suivante :

— faire un choix raisonnable quant à la (ou les) distribution(s) possible(s) pour la variable réponse étudiée ;
— prévoir quel type de dépendance il peut y avoir entre les réponses ;
— sélectionner les variables explicatives appropriées et vérifier si elles expliquent effectivement les changements dans la distribution entre sous-groupes de la population ;
— examiner l'adéquation du modèle choisit, souvent grâce à l'utilisation de graphes et des résidus qui indiquent des aspects inattendus des données qui doivent être pris en compte ;
— si le premier modèle s'avère peu satisfaisant, répéter l'opération avec d'autres plus appropriés ;
— étudier ce que le modèle finalement sélectionné nous apprend au sujet des mécanismes qui ont généré les données ;
— éventuellement, essayer des modèles alternatifs pour tenter d'en apprendre plus sur la structure des données, les modèles n'étant jamais que des approximations des données récoltées.

A. CAUDRON

MOURIER H. & J. D'AGUILAR
250 animaux et insectes, hôtes cachés de nos maisons, 222 pp.
Delachaux et Niestlé, 1996.

Voici plusieurs années que le « Guide des petits animaux sauvages de nos maisons et jardins » de H. Mourier et O. Winding, édité chez Delachaux et Niestlé en 1979, était épuisé. C'était un guide de terrain de grande vulgarisation unique en son genre, utile pour identifier les animaux commensaux des maisons, leurs traces et leurs dégâts, et donnant quelques conseils préventifs et curatifs. Le nouveau guide illustré paru sur le sujet chez les mêmes éditeurs n'en est pas une simple réédition, même si le texte en est très largement inspiré et que de nombreuses photographies et dessins sont communs aux deux ouvrages. Il est dès lors fort dommage que les principaux défauts de l'ancien guide n'aient pas été corrigés dans le second.

Après 21 pages de jolis dessins en couleurs des arthropodes les plus fréquents dans les maisons (certains auraient pu être améliorés et rendus plus conformes à la réalité ; voyez pour comparaison le guide des insectes de M. Chinery paru en 1995 chez Arthaud), le livre présente les commensaux de nos habitations de façon thématique : les arachnides et insectes piqueurs ou suceurs de l'Homme, les ravageurs des aliments, ceux des textiles, ceux du bois, les parasites des plantes d'appartement — revue très succincte —, quelques animaux nidifiant dans les maçonneries et les matériaux d'isolation, les habitants des toits de chaume, les insectes et rongeurs s'attaquant aux métaux, les « occupants clandestins », dont plusieurs espèces auraient pu être abordées dans les chapitres précédents —pourquoi, en effet, classer ici les abeilles maçonnes si fréquentes dans les vieux murs mal jointoyés ? —, et les « hôtes occasionnels » — encore une fois, pourquoi classer ici la scutigère, strictement inféodée aux habitations en région méditerranéenne ? (p. 192). On constate que le thème choisi pour aborder une espèce n'est pas toujours des plus judicieux ; dans certains cas, une espèce est citée dans différents chapitres thématiques.

On regrettera la maigreur des deux chapitres suivants, l'un traitant des excréments (3 pages), l'autre des empreintes (1/2 page), sans préciser l'échelle ni la taille réelle des quelques illustrations ; nulle indication ne précise s'il s'agit d'une empreinte de patte antérieure ou postérieure. Une page est consacrée aux odeurs. La page abordant les bruits n'évoque nullement les sifflements caractéristiques des lérots, si fréquents la nuit dans les maisons qui en abritent.

Enfin, en plus des conseils préventifs et curatifs évoqués dans chaque chapitre thématique, les auteurs consacrent un chapitre aux diverses méthodes de lutte, heureusement sans se limiter à l'emploi de biocides et en évoquant tout juste leurs problèmes de toxicité et d'accoutumance. Quelques lignes pleines de poncifs abordent de manière superficielle et incomplète les pesticides les plus utilisés, et citent le DDT, pourtant interdit depuis des années ! S'il est vrai que les enfants pouilleux étaient saupoudrés de DDT jusque dans les années '60, son absence de biodégradabilité et sa capacité de bioamplification dans les chaînes trophiques ont conduit à son interdiction dans la plupart des pays industrialisés au cours des années '70.

De façon générale, cet ouvrage est fort bien venu pour répondre aux nombreuses demandes de renseignements à propos de « petites bêtes » trouvées dans les maisons, comme l'attestent les sollicitations parfois affolées reçues au Musée de Zoologie de l'Université de Liège. Rappelons que l'ouvrage aborde très utilement non seulement les animaux — larves et imagos —, mais aussi les traces laissées et les dégâts causés par les commensaux. Mais indépendamment des remarques énoncées ci-dessus, plusieurs griefs doivent être néanmoins soulevés.

Le principal manquement est l'absence de clés d'identification et/ou de bons caractères diagnostics spécifiques pour les groupes où de nombreuses espèces voisines se côtoient. Le lecteur sera bien en peine de savoir quel texte lire lorsqu'il a une petite «mite» dans un pot. On trouve ici et là des illustrations comparatives peu sérieuses sensées permettre une identification spécifique de plusieurs espèces proches. Bien malin, par exemple, celui qui identifiera les diverses espèces de puces à partir des dessins succincts de la p. 47. Qui peut se targuer de reconnaître spécifiquement un bourdon (p. 33) ou un moustique (p. 34) à partir d'un simple dessin général en couleurs ? Par ailleurs, on cherchera vainement une illustration de nombreuses espèces citées, tels la musaraigne musette ou le campagnol des neiges.

Par ailleurs, plusieurs erreurs ou omissions révèlent un manque de connaissances ou d'observation. Les réduves observés dans les maisons ne sont nullement des « égarés » (p. 52), car, la nuit, les larves et imagos lucifuges chassent sur les murs. Le dessin illustrant la position des moustiques *Anopheles* — qui permet de les distinguer des *Culex* — est erroné : les pattes métathoraciques ne reposent pas mais sont très clairement relevées. Les auteurs n'expliquent pas, p. 197, la raison de la curieuse attitude que prend *Ocypus olens* lorsqu'il est dérangé : il présente, abdomen relevé, une paire de glandes abdominales évaginées dont l'odeur nauséabonde est répulsive. Pourquoi citer la musaraigne musette comme « occupant clandestin » et lérot, mulots et campagnol des neiges comme « hôtes d'hiver », mais omettre le campagnol roussâtre qui n'hésite pas à s'introduire dans les habitations en hiver — et pourtant répandu au Danemark dont les auteurs sont originaires. Il est curieux d'oublier les célèbres dégâts aux câbles de moteurs de voitures occasionnés en Allemagne par les fouines, et d'affirmer p. 182 que la fouine ne ronge pas ! On ne peut, en aucune façon, parler de la fusion d'une tête et d'un thorax chez les araignées, les scorpions et les acariens (p. 12) puisqu'il n'y a jamais eu une telle différenciation dans l'évolution des arachnides, comme cela s'est fait chez les insectes. Les auteurs citent trois espèces ubiquistes de syrphes p. 198, mais ignorent *Myathropa florea*, espèce qui s'introduit le plus communément dans les maisons. La tégénaire — comme s'il n'y avait qu'une espèce — présentée comme espèce commensale est *Tegenaria domestica*, alors que celle-ci est peu fréquente dans les maisons. Par contre, on ne parle pas de *Tegenaria atrica* si fréquente, dont on présente p. 165 une photographie erronément légendée. Plus grave consiste à illustrer les propos sur *Zygiella x-notata* p. 166 par une photographie mal légendée d'*Araneus diadematus* sur sa toile, dont la structure ne correspond pas au dessin au trait de la page en regard ! Que dire de telles confusions ?

Ensuite, l'auteur s'appuie sur certaines conceptions taxonomiques ou nomenclaturales désuètes. Il y a belle lurette que le règne animal n'est plus divisé en « 14 phylums », mais qu'il en comporte une trentaine. Pas étonnant si les auteurs considèrent encore « l'embranchement des vers » ! Il y a longtemps aussi que *Gryllus domesticus* est unanimement classé dans le genre *Acheta* (p. 20) et que *Staphylinus olens* (p. 26) appartient au genre *Ocypus*, genre qu'il rejoint à la p. 197. Les pucerons et les cochenilles appartiennent à l'ordre des homoptères et non des hémiptères. Pourtant l'ordre des hétéroptères est cité p. 152 à propos du réduve.

La relecture a été bâclée quant aux termes scientifiques et des fautes de frappe apparaissent dans la nomenclature. L'annélide *Allolobophora caliginosa* est par exemple mal orthographié p. 16.

La précision de présentation laisse parfois à désirer. Pourquoi dessiner la plupart des espèces en position naturelle mais représenter certaines, tel le frelon p. 32, avec les pattes et les antennes rabattues comme sur un cadavre ? Les échelles des dessins sont parfois fantaisistes : pour exemple, *Chelifer cancroides*, d'une longueur réelle de 3,5 mm, est représenté p. 17 par un dessin de 15 mm affublé d'un rapport x 10 ! Dans le cas où existe un important dimorphisme sexuel, aucune indication ne précise le sexe : la femelle d'*Urocerus gigas* (p. 32) diffère pourtant du mâle par sa coloration et sa taille supérieure, et bien sûr par son oviscapte.

Enfin de nombreuses erreurs proviennent sans doute d'une traduction rapide et peu scrupuleuse du texte original. Pour exemple, la larve d'*Ixodes ricinus* ne se nourrit pas sur les souris (p. 41), mais bien sur tous les petits rongeurs — mulots et campagnols principalement —, pas plus que le faucon crécerelle ne se nourrit de souris (p. 180) — mais bien de campagnols. On s'étonnera que la tique du pigeon s'attaque éventuellement « à l'homme », mais épargne la femme... Comment se fait-il que « des réduves **mordent** des personnes » alors que ces insectes « possèdent un rostre très court et trapu » qui leur permet de « **sucer** le sang et les humeurs » ? En réalité les réduves **piquent** à l'aide d'un très long rostre.

Ce livre pêche donc à bien des égards et mériterait une édition revue, corrigée, augmentée et plus terre à terre pour mériter le titre générique de « Guide pratique du Naturaliste ».

De façon générale, et pour conclure, l'ouvrage est entaché d'une connotation très péjorative alors qu'il devrait être rassurant pour bien des espèces commensales qui font peur au commun des mortels. Pourquoi parler des espèces hématophages de vampires d'Amérique tropicale plutôt que d'insister sur la dramatique raréfaction de nos chauves-souris insectivores ? Est-il vraiment utile d'affirmer que « il convient de bâtir les immeubles de façon que les oiseaux ne trouvent pas d'emplacement pour nicher ; de plus les trous de mur doivent être bouchés et les ouvertures fixes équipées de grillage ». Adieu donc aux hirondelles et martinets, rougequeues, effraies, pipistrelles et abeilles solitaires ?...

Éric WALRAVENS