

JSFS

Comptes rendus de lecture

Journal de la société française de statistique, tome 142, n° 2 (2001),
p. 81-89

http://www.numdam.org/item?id=JSFS_2001__142_2_81_0

© Société française de statistique, 2001, tous droits réservés.

L'accès aux archives de la revue « Journal de la société française de statistique » (<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/>

COMPTES RENDUS DE LECTURE

Théorie des sondages – Échantillonnage et estimation en populations finies Cours et exercices avec solutions

Yves TILLÉ

284 pages, Dunod 2001

ISBN 2-10-005484-8

Plan de l'ouvrage : 1) Introduction. 2) Une histoire des idées en théorie des sondages. 3) Les fondements de la théorie des sondages. 4) Plans simples. 5) Plans à probabilités inégales. 6) Échantillonnage par scission et estimation de variance. 7) Stratification. 8) Plans équilibrés. 9) Plans par grappe, à plusieurs degrés et deux phases. 10) Estimation avec informations auxiliaires et plans simples. 11) Estimation avec informations auxiliaires et plans complexes. 12) Estimation de variance par linéarisation. 13) Traitement des non-réponses. Quelques solutions. Bibliographie. Index

Schématiquement, on distingue deux approches en théorie des sondages :

- (1) l'approche basée sur le plan de sondage, où l'aléatoire vient du mécanisme de tirage des individus dans la population finie qui nous intéresse et
- (2) l'approche basée sur le modèle, où la population finie est considérée comme un échantillon obtenu par tirage simple dans une population infinie.

Il existe aussi des points de vue intermédiaires, comme l'approche « assistée par un modèle ».

Ce livre est basé sur la première approche. Ce choix a le mérite de la clarté car il met en lumière la spécificité de l'aléatoire en sondage, beaucoup plus nettement que ne le fait une approche modèle. Ce point de vue est de plus indispensable dès qu'on ne peut se contenter d'assimiler à un tirage simple sans remise les données dont on dispose.

En sondage on doit souvent estimer le total d'une variable Y sur une population finie ou une fonction de tels totaux, par exemple un rapport de deux totaux. Si l'on ne dispose d'aucune autre information sur la population finie, on ne peut que faire du sondage aléatoire simple. Heureusement, on dispose souvent d'informations auxiliaires sur la population, variables quantitatives ou qualitative connues pour tous les individus.

Ces informations auxiliaires peuvent améliorer la précision du sondage de deux manières :

- (1) en permettant de choisir des probabilités de tirage inégales entre les individus pour avoir des estimateurs de variance plus petite,
- (2) en permettant de corriger les estimations des totaux inconnus d'après les erreurs commises sur l'estimation des totaux connus, une fois l'échantillon observé.

L'ouvrage d'Yves Tillé aborde, sans lourdeur excessive ni simplification abusive, l'ensemble de ces démarches, avec les tenants et les aboutissants des méthodes. Il y a quelques erreurs typographiques dans le texte, dues apparemment à des opérations de copier/coller mais il y a peu d'erreurs dans les formules. L'index pourrait être un peu plus détaillé, ainsi le tirage systématique classique ou les conditions de Sen-Yates-Grundy n'y figurent-ils pas. L'exposé se lit agréablement. Il est précis et concis. Les notations sont proches de celles de Särndal *et al.* (1992) qui tendent à devenir la référence.

Après l'introduction du chapitre 1, le chapitre 2 rappelle l'histoire des sondages depuis le 18^{ème} siècle avec notamment l'argument récurrent de «l'exactitude ... considérée comme un principe de base de la science statistique». Les travaux de Bowley (1869-1957), en présentant notamment une vérification empirique du théorème central limite appliqué à l'échantillonnage, ébranlent ce principe. On peut regretter ici que l'auteur ne discute pas le partage coût/erreur dans les alternatives recensement/sondage, discussion qui clarifierait la situation de la statistique d'enquête dans certains services publics français. Les deux approches, basées sur le modèle ou sur le plan de sondages, sont discutées section 2.6. L'historique s'achève sur la calibration (Deville et Särndal, 1992), dernière avancée de la prise en compte d'information auxiliaire. Le chapitre 3 présente les fondements de la théorie des sondages. Les notions de plan de sondage, de tirage avec ou sans remise sont clairement énoncées et illustrées. Un paragraphe sur l'exhaustivité permet de percevoir la complexité de la théorie des sondages. Notons que le reste de l'ouvrage, à l'exception du paragraphe 3.14 sur l'optimalité et l'admissibilité, présente sensiblement moins de difficultés d'ordre mathématique. Le reste du chapitre 3 est consacré aux notions de base : plans simples de taille fixe, π -estimateur, estimateur de Hajek, estimation sur un domaine. Le chapitre 4 traite des plans simples avec et sans remise, des plans avec conservation des unités distinctes, de l'estimation d'autres fonctions d'intérêt que le total : ratio, proportion. Il est rarement facile d'obtenir des échantillons par tirages à probabilités inégales. Cette question n'est pas esquivée par Yves Tillé. Le chapitre se termine par la description de cinq algorithmes de tirages pour les plans simples sans remise. Notons à ce sujet qu'une douzaine d'algorithmes

sont décrits et commentés dans ce livre, permettant d'obtenir des échantillons suivant le plan voulu. Le chapitre 5 étudie l'incorporation d'une information auxiliaire au niveau du plan de sondage. L'exposé traite d'abord de tirages avec remise, puis du plan de Poisson, de taille aléatoire, et du plan de taille fixe à entropie maximale, extension du plan de Poisson. Le sondage systématique à probabilités inégales est introduit ensuite, ainsi que le problème du calcul de la variance dans ce plan ; le sondage systématique classique est traité au chapitre 9 comme exemple de sondage en grappe. L'échantillonnage par scission, développé récemment par Deville et Tillé (1998), est exposé au chapitre 6. Son principe est d'écrire le vecteur des probabilités d'inclusion comme une combinaison convexe de vecteurs, puis tirer un de ces vecteurs et sonder suivant le vecteur obtenu. De nombreux exemples ainsi que la réécriture de méthodes classiques comme des méthodes de scission permettent de comprendre ce mode de description et de réalisation de tirages aléatoires. Le problème, souvent très difficile, de l'estimation de la variance est envisagé ensuite ; plusieurs approximations bien fondées sont fournies. Le chapitre 7 aborde la stratification. L'exposé est clair et concis. Notons le paragraphe sur la stratification mobile, méthode qui permet de tirer des unités à probabilités égales et pouvant donner un estimateur de petite variance. Jusqu'à ce point de l'exposé, l'information auxiliaire a consisté en une seule variable. Une façon d'exploiter la connaissance de plusieurs variables auxiliaires est de construire un plan équilibré, c'est-à-dire un plan qui estime sans erreur les totaux des variables auxiliaires ou qui minimise une fonction de coût des estimations de ces différents totaux. C'est le sujet du chapitre 8. L'auteur présente notamment la méthode du cube, issue de ses récents travaux avec Deville (2000). Elle est largement illustrée sur plusieurs plans précédemment examinés. Différentes approximations de l'estimation de la variance sont proposées et commentées. Le chapitre 9 est consacré aux plans par grappes, à plusieurs degrés et à deux phases. Le traitement est classique mais des estimateurs récents de la variance sont proposés, comparés et discutés. Si une information auxiliaire n'a pu être prise en compte dans la construction du plan de sondage, elle peut encore être utilisée pour corriger les poids des unités échantillonnées. C'est le sujet des chapitres 10 et 11. Au chapitre 10, le cas de variables auxiliaires qualitatives est d'abord envisagé à travers la post-stratification et son extension, l'estimation par calage sur marges. Quand l'information auxiliaire est quantitative et univariée, on a classiquement recours à l'estimation par ratio (ou par le quotient) et à l'estimation par différence. Ces méthodes sont examinées. Enfin, l'estimateur par régression est traité et les trois méthodes : par différence, par ratio et par régression, sont comparées en deux pages très claires. Le chapitre 11 poursuit et généralise l'étude au cas où l'on dispose de plusieurs variables auxiliaires. L'estimateur par régression généralisé est défini et ses liens avec le calage sur marge ou l'estimation par le quotient sont explicités. Ensuite on passe à l'estimation par calage, définie par Deville et Särndal (1992). Ici encore l'exposé est complet, sans longueurs. Le dernier paragraphe du chapitre comporte une réflexion théorique sur l'estimateur par la régression optimale et l'estimation conditionnelle. Observons que l'estimateur par régression suppose une dépendance linéaire de la variable d'étude par rapport aux variables

auxiliaires. Dans des travaux récents et qui semblent prometteurs, Breidt et Opsomer (2000) s'affranchissent de la linéarité globale de la dépendance de la variable d'intérêt sur les variables auxiliaires à l'aide de régressions linéaires locales. L'estimateur par calage ou par régression généralisée ne conduit pas à une estimation simple de la variance. La variance peut souvent s'estimer après linéarisation, cette technique fait l'objet du chapitre 12. Celui-ci commence par une présentation des ordres en probabilité. On aborde ensuite les techniques basées sur un développement de Taylor. L'exposé est concis et élégant. Il donne les techniques efficaces et les applique à de nombreux cas. Un paragraphe est consacré aux fonctions estimantes et aux fonctions d'intérêt implicites. Le chapitre se conclut par l'estimation de la variance d'un estimateur calé. Observons que la technique du bootstrap pour l'estimation de la variance n'est pas considérée dans ce livre. Le problème de la non-réponse fait l'objet du chapitre 13. On y trouve notamment des exposés concis mais utilisables de son traitement par repondération et par imputation.

La plupart des chapitres sont suivis d'exercices. Beaucoup sont corrigés en fin d'ouvrage. La bibliographie (229 références) comporte principalement des références historiques et théoriques. Si l'on cherche des exemples d'applications des techniques de sondage, on pourra voir la bibliographie de Lohr (1999), ouvrage récent essentiellement intuitif et peu mathématisé mais riche de problèmes concrets. Toutes les démonstrations ne sont pas données dans l'ouvrage d'Yves Tillé, mais les références des démonstrations manquantes figurent dans la bibliographie. L'exposé est largement autonome; il fait appel à une culture statistique générale plus qu'à des connaissances préalables précises.

On peut remercier Yves Tillé de nous avoir fourni un livre non seulement rigoureux mais aussi riche de remarques pratiques déduites de la solide théorie qui y est exposée. Ce livre intéressera le mathématicien qui y verra comment la théorie des sondages utilise l'outil mathématique. Il intéressera aussi le méthodologiste qui ne peut se contenter d'appliquer toujours les mêmes techniques alors que les problèmes qui lui sont soumis changent.

Références

- BREIDT F. J. et OPSOMER J. D. (2000), Local polynomial regression estimators in survey sampling. *The Annals of Statistics*, 28, 4, 1026-1053.
- DEVILLE J.-C. et SARNDAL C.-E. (1992), Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87, 376-382.
- DEVILLE J.-C. et TILLÉ Y. (1998), Unequal probability sampling without replacement through a splitting method, *Biometrika*, 85, 89-101.
- DEVILLE J.-C. et TILLÉ Y. (2000), Balanced sampling by means of the cube method, Document de travail, Rennes, CREST-ENSAI.
- LOHR S. L. (1999), *Sampling : Design and Analysis*. Pacific Grove, Duxbury Press.
- SARNDAL C.-E., SWENSON B. et WRETMAN J. H. (1992), *Model Assisted Survey Sampling*., Springer Verlag, New York.

Yves Aragon

Traitements des fichiers d'enquêtes redressements, injections de réponses, fusions

Michel LEJEUNE (éditeur scientifique)

Presses Universitaires de Grenoble, 2001.

ISBN : 2-7061-1029-5

Ce petit livre (125 pages) rassemble les communications d'un colloque qui s'est tenu à Grenoble en avril 2000. Il ne présente pas le défaut classique de ce type d'ouvrage, à savoir une grande dispersion entre des communications très diverses artificiellement rassemblées. Ici le thème est très resserré autour d'un sujet, la manipulation des fichiers d'enquêtes. Ce thème est décliné en deux parties : le redressement d'une part et la fusion de fichiers d'autre part. Le tout est précédé d'un article de synthèse clair et bien documenté de L. Lebart.

La partie redressement (4 communications : A.M Dussaix, G. Roy et A. Vanheuverzwyn, P. Guilbert, J.M. Grosbras et C. Lagarenne) est la mieux connue des deux techniques. Parce qu'elle est plus ancienne, elle a fait l'objet de nombreux travaux théoriques et pratiques. Elle est abordée ici surtout au travers d'une question délicate : que faire quand les coefficients du redressement deviennent exagérément grands ? Les limiter arbitrairement au risque de perdre en « représentativité », les limiter en reportant sur d'autres variables au risque de concentrer tous les poids vers des valeurs limites, les laisser tels quels au risque de générer des valeurs aberrantes, jeter l'enquête, ... etc. Le lecteur trouvera ici des exemples d'enquêtes et des réflexions utiles à l'éclairer sur sa propre pratique.

La partie fusion de fichiers (5 communications : M. Lejeune, G. Santini, T. Aluja-Banet et S. Thio, H. Chavenon et H. Haering, J. Sousselier) est la plus nouvelle. En ce domaine, la réflexion méthodologique est clairement en retard sur la pratique. En effet, avec la multiplication des fichiers d'enquêtes, la fusion de fichiers est utilisée de plus en plus souvent, en particulier dans les enquêtes issues du marketing. Les communications présentées ici sont, pour l'essentiel, centrées autour d'une question : comment évaluer la pertinence statistique de ces méthodes ? Autrement dit que peut-on attendre de données ainsi reconstituées artificiellement ? Les conclusions de ces études sont fort stimulantes et ne peuvent que nous inviter à intensifier la réflexion sur cette pratique.

Comme le lecteur l'aura deviné, ce « petit » livre mérite d'être largement connu et reconnu. En premier lieu par les praticiens parce que sa problématique est ancrée dans le quotidien de la pratique statistique et que sa lecture en est très

COMPTES RENDUS DE LECTURE

aisée. Mais il devrait aussi être lu par des statisticiens plus théoriciens qui y trouveront des questions d'actualité sur lesquelles il y a matière à réflexion. C'est par ce type d'approche, je pense au colloque comme au livre, que la statistique montre qu'elle est une science vivante, moderne et passionnante.

Gildas Brossier

Accelerated Life Models. Modeling and Statistical Analysis

V. BAGDONAVICIUS et M. NIKULIN

Chapman & Hall/CRC2, 334 pages

ISBN : 1-58488-186-0

Beaucoup d'ouvrages ont été écrits sur le sujet de l'analyse des durées, mais l'originalité de celui-ci réside dans deux traits particuliers : d'une part, il présente de manière beaucoup plus exhaustive qu'à l'habitude l'ensemble de tous les modèles existants, paramétriques et semi-paramétriques, et propose des généralisations nouvelles de ces modèles ; d'autre part, il met bien en évidence les relations qu'il peut y avoir entre ces diverses familles de modèles : les modèles de Cox, à hasards proportionnels, ceux de la vie accélérée et les modèles de fragilité. Il montre que certaines classes considérées comme distinctes ont en réalité des intersections non vides. Il met aussi en évidence les spécificités de chacun des modèles, qui font qu'on a intérêt, selon les cas concrets considérés, à employer certaines familles plutôt que d'autres et propose des tests d'adéquation de ces divers modèles. Il faut noter que les durées multidimensionnelles ne sont pas considérées dans cet ouvrage. Une troisième particularité intéressante de l'ouvrage tient à ce que le point de vue adopté n'est pas purement industriel, ni purement biomédical, mais allie les deux en faisant profiter chacun des deux contextes des développements accomplis dans l'autre. Par exemple, comme cela est indiqué dans l'introduction, les variables explicatives susceptibles d'avoir un effet ralentissant ou, au contraire, accélérant la durée jusqu'à l'événement attendu, et qui interviennent dans les modèles de régression, sont appelées des « stress » dans le contexte industriel et des « covariables » dans le domaine biomédical, et cette différence de dénomination correspond à une différence conceptuelle. Bien entendu, quoique l'accent soit mis dans cet ouvrage sur les domaines industriel et biomédical, l'application de ces modèles concerne aussi la sociologie, l'économie et de manière générale toutes les disciplines dans lesquelles on analyse des durées. On peut regretter cependant que des exemples numériques ne viennent pas illustrer l'emploi de ces modèles ; mais de nombreuses références sont données (plus de deux cent trente) dans lesquelles on peut en trouver en attendant un deuxième volume qui en proposera. La structure de l'ouvrage est la suivante.

Tout d'abord, un premier chapitre fait le tour des modèles paramétriques en les classant selon que le taux d'incidence, c'est-à-dire le taux de mort ou de panne, au sens large, le taux d'incidence pouvant être par exemple un taux de « guérison », est constant, monotone croissant, monotone décroissant en baignoire (en \cup), ou en pont (en \cap). Outre les modèles couramment utilisés

comme, exponentiel, Weibull, gamma, log-normal, log-logistique, gaussien inverse, Gompertz-Makeham, les auteurs proposent des modèles moins connus comme le Weibull généralisé, le Weibull exponentié, et le modèle de Birnbaum et Saunders.

Au deuxième chapitre, sont proposés des modèles de régression de deux grandes familles : les *modèles accélérés* (AFT) et les *modèles à hasards proportionnels* (PH), le terme de « hasard » qui intervient partout en survie provient d'un anglicisme très répandu, le « hasard » ou « taux de hasard » signifiant en réalité le taux d'incidence, c'est-à-dire la probabilité instantanée que l'événement d'intérêt (mort, guérison ou panne, etc..., selon le cas) survienne à l'instant t , sachant qu'il n'était pas encore survenu jusque là. Le principe de Sedyakin, introduit à ce stade, veut que le taux d'incidence sous un certain stress à moment donné ne dépende pas seulement de ce stress mais aussi de l'état d'usure du sujet à cet instant résumé par sa probabilité de survie à cet instant. Il y a deux différences essentielles entre les deux familles de modèles, accélérés, plutôt utilisés dans l'industrie et à hasards proportionnels, plutôt utilisés dans le domaine biomédical : dans les modèles à hasard proportionnels, le taux d'incidence à l'instant t ne dépend que de la valeur des covariables à l'instant t , alors que dans les modèles accélérés obéissant au principe de Sedyakin, ce taux d'incidence dépend non seulement de la valeur des covariables (appelées stress) à cet instant mais aussi de l'usure du sujet. L'autre différence est que les stress sont complètement exogènes dans l'industrie, et maîtrisés par l'expérimentateur, qui peut, de ce fait, proposer un plan d'expérience adéquat pour en mesurer l'impact sur la durée de vie de la machine considérée. D'ailleurs des plans d'expérience sont proposés dans l'ouvrage (chapitre 9, p. 194). En revanche, dans le domaine biomédical, les covariables sont parfois exogènes, comme par exemple les traitements à comparer lorsqu'on poursuit un essai clinique, mais elles ne le sont pas toutes. En effet, des mesures de facteurs biologiques, comme la pression sanguine ou certains taux biologiques sont aussi des covariables mesurées chez les patients. Or ces covariables reflètent en quelque sorte un certain état d'usure du sujet, et dans ce sens jouent le rôle dévolu dans le cas des modèles accélérés à la dépendance du taux d'incidence par rapport à la survie restante du sujet.

Le chapitre trois est consacré à un thème assez nouveau en survie et qui concerne les modèles de dégradation. Il s'agit en fait d'un modèle multi-états où n'est plus considérée seulement la survenue d'un événement terminal, mais le passage par des niveaux progressifs de dégradation ou de vieillissement. Ce chapitre assez court donne une idée de la difficulté du problème qui est l'un des thèmes majeurs sur lesquels travaillent actuellement les auteurs, qui donnent au dernier chapitre du livre (chapitre treize) beaucoup de résultats sur l'inférence statistique dans ces modèles.

Les chapitres suivants, de quatre à dix, traitent de l'estimation dans les modèles présentés. Au chapitre quatre on trouve l'estimation par maximum de vraisemblance pour des modèles paramétriques en présence de censure droite. Seul le dernier paragraphe traite le cas de l'estimation dans un cadre semi-paramétrique. Au chapitre cinq, les auteurs traitent de l'estimation

paramétrique, toujours avec des censures droites, de modèles accélérés paramétriques, avec ou sans régression, alors qu'au chapitre six, ils traitent le cas semi-paramétrique de modèles accélérés. Dans les chapitres 7 et 8, les auteurs donnent les estimateurs et leurs propriétés pour le modèle à hasards proportionnels de Cox (PH) au chapitre 7, et pour sa généralisation (GPH) au chapitre 8. Le modèle de Cox généralisé (GPH) consiste à introduire une dépendance multiplicative du taux de hasard par rapport au taux d'usure de la manière suivante : au lieu d'être simplement le produit de deux fonctions, un taux de hasard de base et une fonction des covariables, il est multiplié par une fonction du taux d'incidence cumulé. Au chapitre neuf, l'estimation est faite pour les modèles accélérés avec changement d'échelle et de forme (CHSS), et au chapitre 10 pour les modèles à hasards additifs généralisés (GAH), multiplicatifs généralisés (GAMH), additifs et partiellement paramétriques (PPAR) et le modèle à risques additifs d'Aalen (AAR).

Le chapitre 11 est plus typiquement lié à la fiabilité industrielle en examinant des processus de production instables et en leur appliquant les plans d'expérience proposés au chapitre 9 pour des modèles accélérés ainsi que des modèles à hasards proportionnels généralisés.

Le chapitre 12 est intéressant car il propose des tests d'adéquation pour les grandes familles de modèles envisagées. Pour le modèle de Sedyakin généralisé (GS), un test est proposé contre le modèle à hasards proportionnels fondé sur un plan d'expérience de stress en escaliers capable de mettre en évidence la dépendance du taux d'incidence par rapport au degré d'usure de l'élément. Les tests d'adéquation pour le modèle à hasard proportionnel sont fondés sur l'absence de mémoire (AM), sous-entendu du taux d'incidence, testé contre le modèle (GS), qui, lui, en a une. Le chapitre 13 concerne l'inférence statistique pour les modèles de dégradation présentés au chapitre 3.

Enfin, un appendice très utile figure à la fin du volume rappelant la plupart des notions mathématiques utiles pour la compréhension des passages les plus techniques de l'ouvrage. Le lecteur pourra ainsi s'y référer sans avoir à consulter les livres spécialisés sur ces sujets qui sont cités en référence.

En conclusion, je pense que ce livre actualise les recherches dans le domaine très porteur des études en survie. Il est très riche car il ne se focalise pas sur un type particulier de modèle, mais au contraire fait le lien entre les grandes familles de modèles épars, en donne des généralisations et propose pour eux des tests d'adéquation. C'est un livre destiné à devenir un ouvrage de référence dans le domaine de l'analyse des durées.

Catherine Huber