SUR LE BESOIN DE PARAMÈTRES ANALYTIQUES POUR LES ÉCHANTILLONS COMPLEXES[1]

Leslie Kish[2]

Mon propos ici est de souligner l'importance et l'urgence de découvrir des expressions algébriques utiles pour le calcul de paramètres analytiques adaptés aux plans de sondage complexes. Je voudrais m'adresser aux statisticiens mathématiques car ce problème devrait les intéresser puisqu'il en présente toutes les caractéristiques. c'est-à-dire qu'il est important. non encore résolu et soluble.

Il n'existe pas à l'heure actuelle d'analyse mathématique adéquate des problèmes les plus importants et les plus difficiles de la pratique de l'échantillonnage Les manuels portent presque exclusivement sur le calcul de bonnes estimations d'agrégats, de moyennes de variables et de moyennes de quotients. Il est parfois question de la différence entre deux valeurs d'un de ces paramètres, mais ce problème n'est abordé que de façon passagère et sporadique. C'est à cela que se limitent nos outils statistiques pour l'étude d'échantillons complexes.

Depuis la naissance de la théorie de l'échantillonnage. les méthodes probabilistes ont acquis progressivement un rôle privilégié dans le cours des travaux d'échantillonnage, et ces méthodes sont appliquées à des plans de sondage qui sont à la fois économiques et complexes. Le volume de données de haute qualité recueillies au moyen d'enquêtes n'a cessé d'augmenter. grâce à ces techniques. et les spécialistes veulent soumettre leurs données empiriques à des analyses de plus en plus compliquées. Toutefois on ne dispose pas de la théorie statistique mathématique requise pour faire des jugements valables. Pour utiliser les paramètres analytiques habituels. on doit supposer que les éléments choisis sont indépendants· mais cette indépendance est absente dans les plans de sondage complexes. Ainsi. un chercheur peut se

---

[1] Exposé présenté au symposium.

[2] Leslie Kish. Institute for Social Research. The University of Michigan

voir contraint à renoncer à une analyse qu'il considère indispensable. Par contre, s'il est trop impatient ou mal renseigné pour faire preuve d'une telle prudence, il utilisera peut-être les formules conçues pour l'échantillonnage aléatoire simple qu'il peut trouver dans les manuels de statistique, ce qui conduit souvent à des erreurs très graves.

J'espère que les statisticiens mathématiques se rendront compte de l'importance des problèmes non encore résolus dans la recherche de paramètres analytiques pour les données recueillies à partir de plans de sondage complexes. Cette lacune cause plus d'erreurs fondamentales que tout autre écart des hypothèses habituelles

Ces problèmes sont importants, ils n'ont pas encore été résolus et ils sont intéressants. Sommes-nous actuellement en mesure de proposer des solutions? Je crois que oui pour trois raisons. Premièrement, la théorie statistique a beaucoup progressé au cours des dernières années. Deuxièmement, l'augmentation rapide de la quantité d'ordinateurs électroniques et de leur qualité fait que le temps est propice pour la résolution de quelques-uns de ces problèmes. Troisièmement, on peut constater un nouvel intérêt à l'égard d'une méthode générale qui fait espérer des progrès rapides aboutissant à des approximations utiles. Au Survey Research Center, nous sommes en train d'élaborer cette méthode de calcul de la variance des coefficients de régression et d'autres paramètres pour lesquels des formules n'ont pas encore été mises au point.

Cette méthode me fait penser à la manière dont Alexandre a "résolu" le problème du noeud gordien. D'un point de vue théorique, je ne sais pas si cette méthode résout le problème ou le contourne. Mais, comme elle promet de bonnes approximations de variances qu'il est très important de mesurer, les praticiens devraient l'accueillir avec enthousiasme et intérêt. Cette méthode permet de calculer des estimations d'intervalles de confiance de certains paramètres analytiques pour lesquels il n'existe pas de formule actuellement.

Tout ce qui précède est tiré intégralement d'un exposé que j'ai présenté au cours d'une séance conjointe de l'American Statistical Association et l'Institute of Mathematical Statistics en 1957. La situation n'a guère évolué depuis. La technique sur laquelle nos espoirs de trancher le noeud gordien étaient fondés en 1957 est souvent utilisée aujourd'hui· c'est la méthode des duplications successives par blocs (balanced repeated replication) (Kish et Frankel 1970, 1974). Toutefois, mon discours passionné sur les lacunes de la

théorie des distributions de statistiques analytiques doublement complexes n'a pas mobilisé les enthousiasmes des statisticiens mathématiques. Je comprends maintenant pourquoi il en est ainsi. la sagesse étant le fruit de l'expérience. Premièrement, les problèmes que les statisticiens, comme les autres scientifiques. tentent de résoudre sont choisis non en fonction d'un critère de nécessité, mais en fonction d'un critère de faisabilité à un moment défini. (La mise au point de bombes nucléaires est un bon exemple.) Deuxièmement, les problèmes de la théorie des distributions de paramètres d'échantillons complexes semblent trop difficiles à surmonter. Troisièmement. les solutions comporteraient elles-mêmes trop de paramètres pour être utiles. Le point de vue que j'ai exprimé en 1978 (Kish 1978) et qui est encore le mien aujourd'hui est donc plus modéré· "Les nouvelles méthodes de calcul permettent d'obtenir des approximations des variances qui semblent satisfaisantes pour des besoins pratiques. Toutefois il serait préférable d'avoir une théorie mathématique des distributions de paramètres analytiques (tels que les coefficients de réqression) qui est fondée non sur des hypothèses d'indépendance. mais sur des corrélations complexes entre les observations recueillies pour un échantillon. On peut oser espérer qu'un jour il y aura de nouvelles découvertes. mais les résultats ne seront pas universellement utilisables. à cause des complexités mathématiques. et surtout à cause du fait que le nombre de paramètres nécessaire sera trop qrand pour les applications pratiques.

Je veux maintenant décrire sans ambaqes sept points importants au sujet des échantillons complexes. Ces points ne sont pas tous qénéralement connus ou acceptés. mais je vous demande d'y croire, de les mettre en pratique et de les enseiqner. comme je le fais moi-même.

1. Les effets des plans de sondaqe complexes doivent être analysés séparément pour les estimations ponctuelles et les juqements de probabilité (comme. par exemple. les intervalles de confiance ou les tests d'hypothèses). Dans le cas des estimations ponctuelles, nous avons. pour tous les plans de sondaqe. une méthode cohérente pour évaluer les paramètres à l'aide des estimateurs semblables pondérés selon la probabilité (estimateurs H-T). En revanche, les juqements de probabilité tels que les intervalles de confiance sont très sensibles aux effets du plan de sondaqe, surtout dans l'échantillonnaqe en qrappes.

a. "Les paramètres d'un échantillon (moyennes, coefficients de régression etc.) approchent la valeur de la population mère à mesure que la taille de l'échantillon augmente.

b. La convergence est généralement ralentie par les effets du plan.

c. Les effets du plan varient pour différents paramètres, pour différentes variables et pour différents plans de sondage" (Kish et Frankel 1974).

L'article cité plus haut présente également les arguments les plus convaincants à l'appui de ces propositions: la preuve en est d'ailleurs abondante (voir par exemple, Verma et coll. 1980). Néanmoins, deux statisticiens célèbres ont trahi complètement notre pensée dans des descriptions de notre article: "Les auteurs arrivent à la conclusion importante que les estimations d'intervalles de confiance d'un paramètre inconnu ne sont pas sensibles numériquement à la non-indépendance des observations introduite par les techniques d'enquête complexes telles que l'échantillonnage stratifié en grappes". Hélas, cette erreur est citée par d'autres théoriciens qui n'ont pas lu notre réponse à l'intention des praticiens des méthodes d'enquête: "Ces auteurs n'ont pas du tout compris le message principal que nous ne cessons de répéter: les estimations d'intervalles de confiance sont très sensibles numériquement au manque d'indépendance entre les observations dans les techniques d'enquête complexes telles que l'échantillonnage stratifié en grappes" (Kish et Frankel, 1974).

Le fait que cette erreur soit partagée tant par des non-statisticiens que par des théoriciens de l'échantillonnage pose des problèmes à nous, praticiens des techniques d'enquête. Pour cette raison, nous travaillons à l'heure actuelle à fournir une explication plus claire de notre point de vue.

2. Faut-il tenir compte des erreurs d'échantillonnage dans le calcul de paramètres analytiques à partir des données d'enquêtes complexes? Ou est-ce que quelques-uns d'entre nous se sont consacrés à la solution d'un problème négligeable, voire insignifiant? Je comprends mieux saint Sébastien quand les flèches d'une foule de païens enragés lui pleuvaient sur le corps (même un acuponcteur se plaindrait). Au premier rang, il y a les spécialistes des études de marché et des maisons de sondage qui ne nous écoutent pas, bien que certains aient appris à mettre un $\sqrt{\ }$ entre 2 et (pq/n). Au deuxième rang certains démographes écrivent qu'à cause de leurs grands échantillons et de leurs grandes erreurs de mesure, ils n'ont pas le temps de s'occuper des

erreurs d'échantillonnage. Au troisième rang, on retrouve les spécialistes en psychologie mathématique. en économétrie et en biométrie, qui, malheureusement, tirent leurs modèles linéaires directement de la statistique mathématique. Au quatrième rang, ce qui est encore plus malheureux, il y a les statisticiens mathématiques, qui ont tendance à oublier que "la fin ne justifie pas leurs moyennes," à invoquer l'hypothèse selon laquelle les observations sont indépendantes et distribuées identiquement ou à utiliser des formules d'exorcisme bayésiennes pour chasser les mauvais esprits hors du plan de sondage. Au cinquième rang, ce qui est le plus décourageant, on retrouve les théoriciens de l'échantillonnage qui construisent des théorèmes prouvant que. dans les modèles complètement spécifiés de superpopulations arbitraires, il n'est pas nécessaire de se soucier de la provenance et de la méthode de sélection des unités, ni de pondérer les observations pour tenir compte des inégalités des probabilités de sélection. Ces théoriciens ont même réussi à convaincre certains praticiens qu'ils peuvent rester sur l'Olympe avec leurs modèles sans jamais descendre sur terre, là où se trouve la population.

Ces remarques forcément brèves vous permettent de constater que je suis extrémiste pour plusieurs raisons: a) les effets du plan mesurés pour les paramètres analytiques révèlent généralement que les modèles des échantillons les mieux stratifiés sont mal définis: b) on observe souvent que les poids de sélection ont des effets sur les échantillons et c) il existe des relations entre les variables explicatives et les variables expliquées dans chaque individu, mais chaque individu fait partie de populations réelles et ces dernières ont des effets sur les plans de sondage (j'examine ces points en détail dans un ouvrage intitulé Statistical Design for Social Research, sur lequel je travaille actuellement et qui sera publié par Wiley en 1985.)

Ma philosophie est catégorique, mais, en pratique, je suis moins dogmatique. Je reconnais qu'en pratique: a) il n'est jamais possible de dénombrer complètement une population cible et, par conséquent, on doit toujours construire des modèles pour faire des inférences: b) l'échantillonnage probabiliste est trop coûteux et impraticable pour la plupart des expériences et c) malgré le manque de randomisation dans la sélection d'unités ou les analyses, on parvient souvent à des résultats fiables grâce à la prudence, à la répétition de l'échantillonnage, à la qualité du plan de sondage, à la propriété d'additivité et, aussi, un peu au jeu du hasard.

3. Les paramètres analytiques reposent sur la notion de sous-classes et de leur comparaison. Au cours des trois dernières décennies, beaucoup d'ouvrages et d'articles utiles ont été publiés à propos des variances et des effets du plan de sondage au niveau des sous-classes. On peut y trouver de grandes quantités de résultats empiriques, plusieurs observations pratiques basées sur ces résultats (Kish 1980, Kish et coll. 1976, Verma et coll. 1980) ainsi que des théories récentes (Rust 1984, chapitre 6).

a) On doit distinguer les domaines proprement dits des classes recoupées, qui sont plus fréquentes et constituent notre centre d'intérêt ici.

b) La probabilité de sélection demeure constante pour les classes recoupées, mais la taille des échantillons est très variable.

c) Les estimations de totaux et de moyennes calculées à partir d'échantillons complexes doivent être conservées sous la forme de quotients et d'expressions conditionnelles.

d) Les effets du plan au niveau des classes recoupées s'approchent d'une valeur presque égale à 1 proportionnellement à mesure que la taille des sous-classes par grappe primaire s'approche de 1. Ce modèle approximatif doit être utilisé avec prudence et sous certaines réserves, mais il est préférable à toutes les échappatoires respectables qu'on trouve habituellement au sujet des effets du plan et selon lesquelles ces effets sont simplement soit égaux à 1 ou à une autre constante, soit identiques pour l'ensemble de l'échantillon. Un modèle agrégé peut souvent s'avérer meilleur qu'une série de calculs séparés et très variables.

4. Dans les comparaisons de moyennes appariées, les effets du plan sont souvent supérieurs à 1, mais beaucoup moins élevés que la somme de la variance de ces deux moyennes. Ce phénomène est attribuable à des covariances positives (et donc à une sorte d'additivité) et il a été observé régulièrement et souvent dans des comparaisons de classes recoupées et des enquêtes périodiques (Kish 1965, 14.1: voir aussi les publications susmentionnées).

5. En ce qui concerne la production de paramètres analytiques, plusieurs méthodes exploitent les possibilités des ordinateurs électroniques. Les méthodes linéarisées tayloriennes (méthodes delta), qui comprennent des programmes de différentiation automatisée, la méthode des duplications successives par blocs (balanced repeated replication) et la technique du jackknife se sont toutes révélées des outils importants pour l'estimation de la variance et des

effets du plan pour les échantillons complexes (Kish et Frankel 1970 et 1974; Woodruff et Causey 1978). La méthode bootstrap pourrait également se répandre à l'avenir (Rao 1984).

Les paramètres analytiques comportent généralement des effets du plan supérieurs à 1 et ces résultats sont significatifs dans tous les sens de ce mot, mais ces effets du plan sont également moins élevés que ceux calculés pour les moyennes. Certaines régularités ont été mises en évidence dans les relations entre les effets du plan sur divers coefficients et les effets du plan sur les moyennes.

Pour nous aider à voir clair, il nous faut non seulement plus de travaux empiriques, mais aussi plus de résultats fondés sur la théorie de l'échantillonnage et l'analyse de modèles. Je dois avouer que je suis déçu que, depuis nos premiers travaux, nous n'avons pas vu paraître plus de publications sur la théorie et les modèles qui seraient directement applicables à l'inférence à partir de données réelles. Il est nécessaire d'étudier les fondements empiriques des effets du plan, mais pour satisfaire notre besoin intellectuel de comprendre, il nous faut aussi une théorie mieux étayée et de meilleurs modèles. Même nos besoins pratiques ne sont pas complètement satisfaits par les travaux empiriques sur les effets du plan parce que les résultats dépendent à la fois des variables, du type d'estimations calculées pour le plan de sondage utilisé et de la population mère dans laquelle les données ont été recueillies. Cette source de variation à quatre dimensions est trop complexe et ce qui nous manque, c'est une théorie qui permet d'élaborer des modèles qui simplifient la réalité.

6. L'analyse de données qualitatives est un domaine important qui s'étend rapidement et plusieurs travaux ont été faits sur l'application de ces méthodes aux données d'enquêtes complexes (Fay 1982; Landis et coll. 1982; Koch et coll. 1975). Ces travaux abordent également des questions d'analyse de la variance qui étaient le point de départ de quelques-uns des premiers modèles, mais n'ont pas été approfondies par la suite (Kempthorne et Wilk 1955; Tukey et Cornfield).

7. Quant à l'avenir, j'ai bon espoir que de nouveaux éléments de théorie donneront naissance à des applications pratiques, sauf dans deux domaines. Premièrement, la statistique mathématique n'a pas et ne pourra pas nous donner une théorie complète des distributions qui soit directement utilisable parce

qu'il y a trop de paramètres dans la double complexité des paramètres analytiques d'enquêtes complexes. Deuxièmement, les concepteurs de modèles sont incapables de faire disparaître cette complexité. Ils pourront toutefois nous aider à trouver des méthodes d'inférence plus exactes et compréhensives et améliorer l'utilisation et la présentation des paramètres analytiques d'enquêtes complexes.

BIBLIOGRAPHIE

[1] Fay, R. (1982). Contingency table analysis for complex sample designs: CPLX. Proceedings of the Section on Survey Research Methods, American Statistical Association, pp. 44-53.

[2] Kish, L. (1957). Some unsolved problems of complex samples. Communication présentée à une séance conjointe de l'Amercian Statistical Association etde l'Institute for Mathematical Statistics.

[3] Kish, L., et Frankel, M.R. (1970). Balanced repeated replications for standard errors. JASA, 65, pp. 1071-1094.

[4] Kish, L., et Frankel, M.R. (1974). Inference from complex samples. JRSS (B), 36, pp. 1-74.

[5] Kish, L. (1980). Design and estimation for domains. The Statistician (Londres) 29, pp. 209-222.

[6] Kish, L., Groves, R.M, et Krotki (1976). Sampling errors for fertility. Occasional paper 17, Londres: World Fertility Surveys, 61 pages.

[7] Koch, G., Freeman, D., et Freeman, J. (1975). Strategies in the multivariate analysis of data from complex surveys. International Statistical Review, 43, pp. 53-59.

[8] Landis, J.R., Lepkowski, J., Eklund, S., et Stehouwer, S. (1982). A statistical methodology for analyzing data from a complex sample survey. Vital and Health Statistics, Series 2 - no. 92. DHHS Publ. no. 82-1366. Public Health Service, Washington, U.S. Government Printing Office.

[9] Rao, J.N.K. (1984). Bootstrap inference with stratified samples. (non encore publié).

[10] Rust, K.F. (1984). Techniques for Estimating Variances for Sample Surveys, The University of Michigan, thèse de Ph.D.

[11] Verma, V., Scott, C., et O'Muircheartaigh, C. (1980). Sample designs and sampling errors for the World Fertility Survey. JRSS (a), 143, pp. 431-473.