

ÉVALUATION DE L'HOMOGENÉITÉ DE TENDANCES PROVINCIALES AVEC DONNÉES D'ENQUÊTES COMPLEXES

Martin Pantel, Lenka Mach et Michelle Rotermann¹

RÉSUMÉ

Lors d'études analytiques, on s'intéresse souvent à la comparaison d'un paramètre pour plusieurs sous-ensembles de la population. Cet article utilise, à titre d'exemple, la comparaison entre les proportions de personnes de 20 ans et plus étant physiquement actives dans chacune des provinces canadiennes. L'intérêt porte particulièrement sur les tendances que suivent ces proportions au cours d'une année civile – certaines difficultés sont apportées par le chevauchement partiel des échantillons aux différents temps de l'année, par le grand nombre de comparaisons à être considérées, et par le plan d'échantillonnage complexe. À l'aide de logiciels courants, nous illustrons l'exécution d'une telle évaluation en utilisant les données de l'Enquête sur la santé dans les collectivités canadiennes.

1. INTRODUCTION

1.1 Enquêtes complexes

Les enquêtes menées auprès de grandes populations apportent souvent des défis analytiques intéressants. Les contraintes de temps et d'argent mènent normalement à un plan d'échantillonnage complexe, c'est-à-dire un plan où on impose une stratification et/ou une mise en grappes, et où les unités sont sélectionnées avec des probabilités inégales. Les techniques qui permettent de tenir compte de la probabilité inégale de sélection sont assez bien connues; on ajuste le poids de chaque répondant afin de refléter le nombre d'unités qu'il représente. La stratification et la mise en grappes ont un impact sur le calcul de la variance des estimations, et par conséquent sur les tests analytiques. L'impact de la non-réponse et de l'imputation doit également être considéré. Les techniques qui tiennent compte de ces difficultés sont parfois moins bien connues, et moins bien implantées dans les logiciels populaires.

1.2 Description du problème

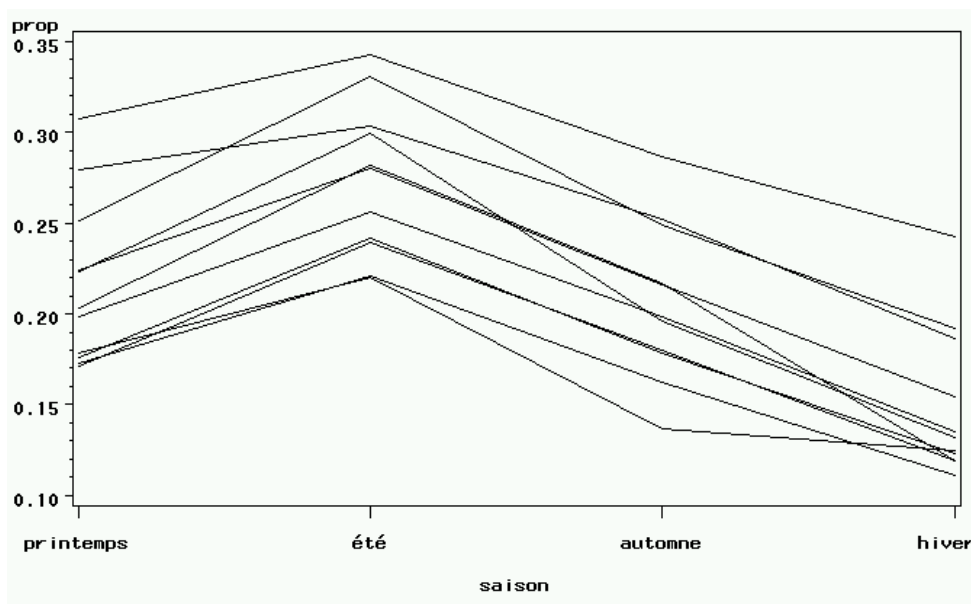
L'Enquête sur la santé dans les collectivités canadiennes (ESCC) est une enquête transversale menée par Statistique Canada. Pour chaque cycle de l'enquête, on vise l'obtention d'un échantillon d'environ 130,000 individus – le projet qui fait l'objet de cette étude est basé sur deux cycles de l'ESCC, référant aux années 2001 et 2003. Les strates de l'ESCC sont définies selon les 133 régions sanitaires au Canada, mais les unités primaires d'échantillonnage sont sélectionnées selon une de trois stratégies. Pour la majorité des strates, la base aréolaire de l'Enquête sur la population active (également menée par Statistique Canada) est utilisée. Il s'agit d'un plan d'échantillonnage stratifié à plusieurs degrés, où l'individu constitue l'unité finale d'échantillonnage. Dans les autres régions sanitaires, la stratégie pour sélectionner les logements est basée sur le numéro de téléphone; on utilise soit une liste de numéros, soit une méthode de composition aléatoire.

L'ESCC recueille de l'information portant sur l'état actuel de santé, les comportements associés à la santé et l'utilisation de services de santé, ainsi qu'un profil sociodémographique. Une variable binaire a été dérivée pour identifier les répondants dont le montant d'activité physique est au-delà d'un seuil prédéterminé. On peut également identifier la province de résidence du répondant, ainsi que le temps de l'année où l'entrevue a été réalisée. On suppose que le niveau d'activité physique de la population diffère selon le temps de l'année, puisque plus de gens sont actifs pendant l'été que pendant l'hiver. On accepte également que la proportion de personnes actives diffère selon la province; certaines provinces

¹ Martin Pantel (Division des méthodes d'enquêtes sociales), Lenka Mach (Division des méthodes d'enquêtes sociales), Michelle Rotermann (Division de la statistique de la santé), Statistique Canada, Tunney's Pasture, Ottawa, Canada, K1A 0T6

canadiennes peuvent être plus propices que d'autres à l'activité physique, surtout étant donné la saison. La question est alors posée : la tendance que suit la proportion de personnes actives au cours des quatre saisons de l'année est-elle la même pour chacune des régions canadiennes? (On entend ici par « régions » les 10 provinces et l'ensemble des territoires). Les détails concernant la définition de la variable binaire et son utilisation pour calculer les proportions sont disponibles dans Tremblay et Rotermann (2005). La figure 1 montre les tendances que suivent les proportions d'une saison à l'autre, et ce pour chacune des 11 régions.

Figure 1 : Proportion de personnes actives dans chaque région selon la saison



L'échantillon pour l'ESCC est choisi pour une année entière, ensuite divisé de façon aléatoire en douze sous-groupes qui sont contactés au cours de l'année. Les questions portant sur l'activité physique réfèrent aux trois mois qui précèdent l'entrevue. Par exemple, un répondant contacté au mois de juin doit tenir compte des activités faites aux mois de mars, avril, mai et juin. Dans ce projet, une saison est définie comme un regroupement de trois mois; le printemps compte les mois de mars, avril et mai, l'été compte les mois de juin, juillet et août, etc. Ainsi, il existe une dépendance dans les données, puisqu'une personne active peut avoir été identifiée comme telle pour plus d'une saison. La méthode d'auto-amorçage, également connue comme la méthode « bootstrap », a été utilisée pour évaluer la variance des estimations. Un avantage de cette méthode, en plus de tenir compte de la stratification et de la mise en grappes, est qu'elle tient compte de cette dépendance.

Dans cet article, on présente trois différentes méthodes permettant l'évaluation de l'homogénéité des tendances que suivent les proportions provinciales d'une saison à l'autre. D'abord, à la section 2.1, une évaluation utilisant la statistique de Wald est utilisée – elle permet de faire une évaluation globale de l'homogénéité. À la section 2.2, la technique de Bonferroni permet d'évaluer de façon individuelle la différence entre chaque province et le reste du pays. Enfin dans la section 2.3, une évaluation basée sur un modèle de régression est explorée, qui permet d'effectuer en même temps une évaluation globale et une évaluation des différences individuelles.

2. ÉVALUATION DE L'HOMOGÉNÉITÉ

2.1 Évaluation globale selon le test de Wald

Les tests utilisant la statistique de Wald sont souvent utilisés pour évaluer les paramètres de régressions multiples, mais peuvent être utilisés de façon plus générale pour évaluer des hypothèses portant sur des vecteurs de paramètres. Les premières explications de la méthode sont dans Wald (1943), tandis que Korn et Graubard (1990) offrent une bonne discussion de la méthode dans le contexte d'une régression avec des données d'enquête complexe. En utilisant les notations habituelles en théorie

d'échantillonnage, la formule générale pour la statistique de Wald, pour évaluer si un vecteur μ est égal à 0, est la suivante :

$$W = \hat{\mu}' \hat{\Sigma}^{-1} \hat{\mu} \quad (1)$$

Sous certaines hypothèses, on peut montrer qu'avec un très grand nombre d'unités primaires d'échantillonnage, cette statistique suit une distribution khi carré avec d degrés de liberté (dénotée $\chi^2(d)$), où d réfère à la longueur du vecteur μ .

Considérons notre exemple : disons que la proportion de personnes actives pendant la saison j dans la région i est dénotée p_{ij} , et que la proportion de personnes actives dans le pays entier pendant la saison j est p_j . La différence entre la proportion pendant l'été et celle pendant le printemps est donc $p_{i2} - p_{i1}$ pour la région i , et $p_2 - p_1$ pour le pays. On veut savoir si ces deux différences sont égales pour les 11 régions, et ce pour les trois changements de saison (été – printemps, automne – été, hiver – automne). Cependant, il est seulement nécessaire de comparer 10 des provinces; les résultats des comparaisons pour la 11^{ième} province suivent directement des 10 premières. Notre hypothèse nulle H_0 est donc :

$$H_0 : \begin{bmatrix} (p_{i2} - p_{i1}) - (p_2 - p_1) \\ (p_{i3} - p_{i2}) - (p_3 - p_2) \\ (p_{i4} - p_{i3}) - (p_4 - p_3) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} \quad i = 1, \dots, 10 \quad (2)$$

Ainsi, notre hypothèse nulle propose que 30 différences sont toutes égales à 0. Le logiciel SUDAAN a été utilisé pour estimer les p_{ij} et les p_j . Ce logiciel permet de tenir compte de la pondération inégale et du plan complexe de sondage; cependant, les données de sortie ne sont pas dans le format voulu. Plutôt que d'avoir un vecteur de 30 estimations de différences, on obtient un vecteur de 60 proportions (les totaux marginaux sont également estimés), qu'on appelle \hat{m} . On obtient aussi la matrice de covariance de taille 60x60, qu'on appelle \hat{S} . Il a ensuite fallu construire une matrice de contrastes (appelons-la C) de taille 30x60, afin de définir correctement les 30 différences d'intérêt. En implantant ces éléments dans notre formule pour la statistique de Wald à l'équation (1), on obtient :

$$W = [C\hat{m}]' [C\hat{S}C']^{-1} [C\hat{m}] \quad (3)$$

Les éléments \hat{m} , \hat{S} et C obtenus, les manipulations permettant de calculer la statistique W ont été effectuées en SAS par l'entremise du langage matriciel interactif (PROC IML). Le résultat est de $W=94,50$, qu'on compare avec le percentile désiré d'une variable $\chi^2(30)$. Pour un niveau de confiance de 95%, on obtient $\chi^2_{0,95}(30) = 43,77$, donc on conclut que les tendances ne sont pas toutes homogènes.

2.2 Évaluations individuelles selon le test de Bonferroni

Puisqu'on a déterminé que les tendances ne sont pas toutes homogènes, il serait intéressant d'examiner où se trouvent ces différences. Le test de Bonferroni permet d'effectuer une série de comparaisons tout en fixant le niveau de confiance global à une valeur désirée. De plus, cette technique est facilement adaptable aux données provenant d'enquêtes à plan complexe. Afin d'effectuer ce test avec le logiciel SUDAAN, le trio de différences pour chaque région a été comparé avec le reste du Canada, tel que décrit ci-dessous (4), quoique avec le test de Wald on ait comparé la proportion dans chaque province à celle dans le pays entier, tel que décrit en (2). Lafortune et Roberts (2005) montrent que les deux comparaisons sont exactement équivalentes lorsqu'on compare des proportions. Malgré que la théorie ne s'applique pas directement aux différences de proportions, les vérifications empiriques utilisant les données de ce projet ont montré que les résultats étaient effectivement les mêmes dans les deux approches. Si l'hypothèse est définie comme dans (4), les évaluations peuvent toutes être faites dans une procédure DESCRIPT de SUDAAN, en utilisant des contrastes. Par contre, une évaluation de l'hypothèse (2) exigerait des manipulations à l'extérieur de SUDAAN, semblables à celles faites pour le test de Wald – sauf qu'il faudrait faire ces manipulations pour chacun des 33 tests! Ainsi, l'hypothèse évaluée est la suivante :

$$H_0 = \begin{bmatrix} (p_{i,2} - p_{i,1}) - (p_{CAN-i,2} - p_{CAN-i,1}) \\ (p_{i,3} - p_{i,2}) - (p_{CAN-i,3} - p_{CAN-i,2}) \\ (p_{i,4} - p_{i,3}) - (p_{CAN-i,4} - p_{CAN-i,3}) \end{bmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} \quad i = 1, \dots, 11 \quad (4)$$

Pour chacune des onze régions, trois tests t ont été effectués, une pour chaque différence. En tout, on a 33 comparaisons donc 33 tests individuels. Si on effectuait une seule comparaison à un niveau de confiance de 95%, on comparerait la valeur- p obtenue du test à la valeur critique de 0.05. Puisqu'on a plutôt 33 comparaisons simultanées, on doit ajuster la valeur critique : afin d'avoir un niveau de confiance de 95% dans l'ensemble de nos conclusions, on doit comparer chacune des valeurs- p à la valeur critique de $0.05/33=0.0015$. Selon cette évaluation, trois différences se sont démarquées. Le tableau 1 identifie ces différences et donne leur valeur- p . La dernière colonne du tableau indique si la différence régionale est plus (+) grande ou moins (-) grande que la différence pour le reste du pays.

Tableau 1 : Différences significatives selon le test de Bonferroni

Test	Valeur- p	Direction
$p_{CB,2} - p_{CB,1} = p_{CAN-CB,2} - p_{CAN-CB,1}$	0.0011	-
$p_{SK,3} - p_{SK,2} = p_{CAN-SK,3} - p_{CAN-SK,2}$	0.0014	+
$p_{TN,4} - p_{TN,3} = p_{CAN-TN,4} - p_{CAN-TN,3}$	0.0004	-

2.3 Évaluation selon un modèle de régression

Les évaluations globale et individuelles des différences entre les proportions selon les tests de Wald et Bonferroni exigent un effort considérable de programmation. Il serait utile de pouvoir obtenir les mêmes résultats à l'aide d'une méthode plus directe. On peut le faire en ajustant des modèles de régression – encore une fois, le logiciel SUDAAN peut être utilisé afin de pouvoir tenir compte de la complexité du plan. D'abord, on ajuste un modèle linéaire sans interactions, c'est-à-dire avec seulement les effets principaux de la province et de la saison. On peut l'exprimer selon la formule suivante :

$$p_{ij} = \beta_0 + \sum_{i=1}^{10} \beta_{1i} D_{prov=i} + \sum_{j=1}^3 \beta_{2j} D_{saison=j} \quad (5)$$

où p_{ij} est la proportion de personnes actives dans la province i pendant la saison j , et $D_{prov=i}$ est une variable indicatrice qui prend la valeur 1 pour la province i , et la valeur 0 autrement. (Une définition semblable s'applique à la variable $D_{saison=j}$). Il y a donc 14 paramètres β , soit l'ordonnée à l'origine, et un pour chaque variable indicatrice identifiant chaque région et chaque saison (sauf les niveaux de référence, qui sont l'Ontario et l'hiver dans cet exemple).

Le deuxième modèle est le modèle saturé – on inclut toutes les interactions possibles entre les provinces et les saisons. Il comprend donc 44 paramètres : l'ordonnée à l'origine, les 10 régions (avec l'Ontario comme référence), les 3 saisons (avec l'hiver comme référence), et les 30 interactions entre les régions et saisons. Le logiciel SUDAAN permet de calculer des statistiques qui décrivent la justesse d'un modèle – en comparant les statistiques obtenues pour le modèle (5) à celles obtenues pour le modèle saturé, on pourrait déterminer si l'ajout des interactions améliore la qualité du modèle. Ensuite, les différences individuelles (telles que décrites pour le test de Bonferroni dans la section 2.2) pourraient être évaluées avec ce même modèle, en définissant des contrastes parmi ses coefficients.

Notons qu'un modèle de régression logistique est normalement utilisé lorsque la variable dépendante est binaire – avec le modèle (5), il serait théoriquement possible que la proportion modélisée par un ensemble de variables indépendantes ne serait pas dans l'intervalle $[0,1]$. Cependant, l'évaluation de la différence entre deux valeurs modélisées de p_{ij} sur une échelle logistique ne représente pas la même comparaison que celle donnée par les tests de Wald et Bonferroni. De plus, les proportions estimées tombent toutes entre 0.10 et 0.35; avec le grand nombre de répondants considérés, on s'attend que les

proportions modélisées seraient semblables à celles estimées. D'autres considérations de régression doivent être examinées de plus près; une évaluation plus détaillée sera présentée au colloque.

3. DISCUSSION ET CONCLUSION

Deux thèmes d'analyse populaires ont été touchés dans ce texte : la comparaison d'une caractéristique entre une sous-population et la population entière, et l'évaluation de l'homogénéité des tendances que suivent ces caractéristiques dans le temps. Pour tenir compte du plan complexe de l'enquête, les poids bootstrap ont été utilisés pour les estimations de variance. Souvent, ces poids constituent la seule source d'information sur l'échantillonnage qui soit à la disposition des analystes; les variables identifiant les strates et les unités primaires d'échantillonnage sont souvent omises des fichiers d'analyse pour des raisons de confidentialité, et les autres ajustements (par ex. pour la non-réponse et l'imputation) sont souvent ignorés, sauf dans les poids bootstrap.

Le test de Wald nous a révélé que les tendances que suivent les proportions de personnes actives au courant de l'année ne sont pas homogènes parmi les 11 régions, sans toutefois identifier où se trouvent cette ou ces différences. Le test de Bonferroni a un aspect un peu arbitraire, au sens où le nombre de comparaisons à effectuer détermine la valeur critique à laquelle on compare chacune des valeurs- p . Cependant, lorsqu'on sait qu'il existe une différence dans l'ensemble des comparaisons, il se révèle très utile pour l'identifier. En poursuivant avec ce test, on a pu déterminer les combinaisons de provinces et de saisons où les tendances que suivent les proportions de personnes actives diffèrent des tendances générales au Canada. Cette combinaison de tests a exigé un effort considérable de programmation – il serait utile de pouvoir exécuter le tout en une seule étape. L'utilisation d'un modèle de régression linéaire, qui tient compte des particularités apportées par un plan complexe de sondage, est prometteuse pour atteindre cet objectif.

REMERCIEMENTS

Les auteurs remercient sincèrement les deux examinateurs de leurs remarques et suggestions constructives. Ils remercient également Georgia Roberts et Yves Lafortune de leurs conseils et commentaires instructifs.

RÉFÉRENCES

- Agresti, A. (1996), *An Introduction to Categorical Data Analysis*, New York : Wiley.
- Korn, E.L., and Graubard, B.I. (1990), « Simultaneous Testing of Regression Coefficients with Complex Survey Data : Use of Bonferroni t Statistics », *The American Statistician*, Vol. 44, No. 4, p. 270-276.
- Lafortune, Y., et Roberts, G. (2005), « Comparing a rate in a sub-population to the rate in the full population : How it may be done when using survey data, and available software tools », rapport non publié, Ottawa, Canada : Statistique Canada.
- Lohr, S.L. (1999), *Sampling : Design and Analysis*, Pacific Grove, CA : Brooks/Cole.
- Research Triangle Institute (2004), *SUDAAN Language Manual, Release 9.0*, Research Triangle Park, NC : Research Triangle Institute.
- Tremblay, M., et Rotermann, M. (2005), « Seasonal Variation in Physical Activity by Province in Canada », article présenté à la conférence du American College of Sports Medicine, Nashville, USA.
- Wald, A. (1943), « Tests of statistical hypotheses concerning several parameters when the number of observations is large », *Transactions of the American Mathematical Society* Vol. 54, p. 426-482.