Analyse fondée sur le plan

On emploie une approche fondée sur le plan lorsque l'objectif de l'analyse est d'estimer des statistiques descriptives pour une population finie *U*. Une population finie consiste en un ensemble de valeurs fixe où la seule source de variation aléatoire est le plan d'échantillonnage utilisé pour sélectionner un échantillon dans cette population finie. Les paramètres de population finie décrivent cette population dont nous tirons notre échantillon. Ces quantités représentent toujours la population à un moment donné (au temps d'échantillonnage).

Par exemple, supposons qu'un chercheur souhaite faire une inférence concernant la population des personnes âgées de 25 à 34 ans vivant dans les dix provinces canadiennes pendant la période de référence de l'enquête et qu'il souhaite estimer la proportion de ces individus qui avaient un emploi pendant l'année de référence. Dans ce cas, la population cible est finie et se compose de *N* individus, la variable d'intérêt est binaire (y_i=1 si l'individu *i* était employé pendant l'année de référence, y_i=0 dans le cas contraire) et le paramètre d'intérêt est fixe mais de quantité inconnue : $\theta_N = \sum_{i=1}^{N} y_i \Big/ N$. L'estimateur habituel de θ_N est donné par $\hat{\theta}_N = \sum_{i=1}^{n} w_i y_i \Big/ \sum_{i=1}^{n} w_i$, où *n* est le nombre d'individus dans l'échantillon et w_i représente le poids d'enquête de l'individu *i*. Ce poids indique combien d'individus de la population sont représentés par l'individu *i* dans l'échantillon. Lorsque $w_i = 1/\pi_i$, où π_i est la probabilité de sélectionner l'individu *i* dans l'échantillon, $\hat{\theta}_N$ est appelé estimateur de Horvitz-Thompson. Typiquement, on obtient le poids w_i en appliquant plusieurs ajustements au poids du plan de base. Ces ajustements tiennent compte de la non-réponse par certains individus échantillonnés et assurent le calage avec les chiffres de population connus.

La variance de la distribution d'échantillonnage de $\hat{\theta}_N$ est $V_p\left(\hat{\theta}_N\right) = E_p\left(\left[\hat{\theta}_N - E_p\left(\hat{\theta}_N\right)\right]^2 \right)$ où l'indice inférieur *p* indique que la prédiction est donnée pour le plan d'échantillonnage *p*, pour tous les échantillons possibles qui peuvent être sélectionnés en vertu du plan donné *p*. Les plans d'échantillonnage complexes impliquent une stratification, une répartition en grappes et des étapes de sélection multiples stages qui rendent l'estimation de $V_p\left(\hat{\theta}_N\right)$ complexe. De plus, la variance est aussi affectée par les ajustements pour non-réponse et le calage. Plusieurs méthodes d'estimation de la variance sont décrites au Chapitre 9 de Lohr (1999).

Pour la plupart des méthodes d'estimation de la variance, il existe dans la littérature des résultats théoriques qui indiquent les conditions auxquelles $\left(\hat{\theta}_N - \theta_N\right) \Big/ \sqrt{\hat{V}_p\left(\hat{\theta}_N\right)}$ suit asymptotiquement une distribution standard normale. Pour plus de détails et des références voyez, par exemple, la Section 9.5 de Lohr (1999).

Analyse fondée sur le modèle

Les analystes qui adoptent une approche fondée sur le modèle souhaitent faire des inférences sur des populations plus générales que la population finie fixe dont l'échantillon a été sélectionné. Leur but est de découvrir une vérité universelle, représentée par des paramètres de modèle. En théorie de l'échantillonnage, nous appelons ce modèle la *superpopulation* qui a généré la population finie que nous étudions.

Un modèle de superpopulation décrit la relation entre les variables au moyen d'un modèle, ξ. Ici, le concepteur du modèle ne s'intéresse pas à la population finie *U* à un moment donné mais plutôt aux paramètres du modèle de superpopulation. Par exemple, considérons un modèle de régression linéaire $y_i = \boldsymbol{x}_i\boldsymbol{\beta} + \varepsilon_i, i = 1,\ldots,N, \varepsilon_i,\ldots,\varepsilon_N \; i.i.d.N\left(0,\sigma^2\right)$, où le modèle décrit une relation entre la variable dépendante, le logarithme du revenu et des caractéristiques comme le niveau d'éducation, l'âge, les caractéristiques de l'emploi, l'expérience, le sexe, etc. Il faut distinguer entre le vecteur des coefficients de régression de la population finie, $\boldsymbol{\beta}_N$, défini comme une fonction des quantités de la population finie, et le paramètre de modèle $\boldsymbol{\beta}$ (Pfeffermann, 1993).

Sous une approche *fondée purement sur le modèle* les poids d'enquête ne sont pas utilisés dans l'analyse et les inférences sont faites par rapport à la variance sous le modèle. Si le modèle est bon pour toutes les unités de la population, alors les estimateurs fondés sur le modèle des paramètres du modèle sont optimaux. En revanche, si le modèle n'est pas bon, les estimateurs fondés sur le modèle auront probablement de très mauvaises propriétés. Il est donc recommandé d'employer une *approche commune fondée sur le plan et le modèle* qui utilise à la fois la randomisation sous le plan et sous le modèle. Avec cette approche, les poids d'enquête et la *variance totale*, sous le plan et sous le modèle, sont utilisés dans l'analyse.

Dans de nombreuses situations, la variance totale peut être approximée par la variance du plan $V_p\left(\hat{\theta}\right)$. Les approches adaptées à l'inférence concernant les paramètres de modèle dans le cas de données d'enquête complexes sont discutées par Binder et Rogers (2003).

Korn et Graubard (1999) et Heeringa et al. (2010) offrent des références très utiles pour les chercheurs qui utilisent des données d'enquête dans leurs analyses.

Plans informatifs et ignorables

…alors, puisqu'un échantillon aléatoire simple d'un échantillon aléatoire simple est lui-même un échantillon aléatoire simple, les problèmes d'inférence peuvent être traités de façon classique.

Barnard (1973)

Même si Barnard (1973) avait raison de noter qu'un échantillon aléatoire simple d'un échantillon aléatoire simple est aussi un échantillon aléatoire simple, la plupart des échantillons d'enquête ne sont pas des échantillons aléatoires simples! Si les probabilités de sélection dépendent de *y* par plus que les covariables, alors on risque des conclusions erronées si on ne tient pas compte du plan (Pffeferman 1993, 1996, 2003, Rao et Scott 1981).
Le choix de l'approche appropriée pour analyser des données d'enquête est lié aux concepts des plans d'échantillons *ignorables* et *informatifs*. Comme l'expliquent Binder et Roberts (2001), un plan est ignorable pour une analyse donnée si l'inférence basée sur toutes les informations connues, informations de plan incluses, est équivalente à l'inférence basée sur les mêmes informations à l'exclusion des résultats des variables aléatoires liées au plan. Un plan d'échantillon *informatif* produit des échantillons pour lesquels la distribution d'une variable d'intérêt est différente de sa distribution dans la population (Binder, Kovacevic, Roberts, 2005). Les plans non informatifs sont tous ignorables, mais l'inverse n'est pas vrai.

Effets de plan de sondage, intervalles de confiance et tests statistiques

Kish (1965) suggérait qu'une façon de résumer l'effet du plan d'échantillonnage sur la variance de l'estimation était d'examiner le ratio suivant, qu'il avait nommé *effet de plan de sondage* (En anglais, Design Effect, *deff*) :

$deff\left(\hat{\theta}_N\right)=\frac{V\left(\hat{\theta}_N \text{ sous le plan } p \text{ avec } n \text{ observations}\right)}{V\left(\hat{\theta}_N \text{ sous le EAS avec } n \text{ observations}\right)}=\frac{V_p\left(\hat{\theta}_N\right)}{V_{SRS}\left(\hat{\theta}_N\right)}$, où $\hat{\theta}_N$ est un estimateur de θ_N convergent par rapport au plan et fondé sur les poids d'enquête des plans de sondage respectifs.

Par conséquent, le *deff* est la mesure de la précision gagnée ou perdue en raison de l'utilisation d'un plan d'échantillonnage complexe au lieu d'un échantillonnage aléatoire simple (EAS). Habituellement, la stratification augmente la précision, tandis que la répartition en grappes et la pondération diminuent la précision. Généralement, pour l'échantillonnage

stratifié en grappes à plusieurs degrés, les valeurs du *deff* sont plus élevées que 1, ce qui signifie une perte de précision en comparaison avec l'EAS. Les valeurs du *deff* tendent à être plus grandes pour les estimations des chiffres de population, les moyennes et les proportions, et peuvent être sensiblement plus faibles pour les estimations du coefficient de régression (Kish, 1995; Park et Lee, 2004; Heeringa et coll., 2010).

Lorsque l'information sur le plan de sondage n'est pas disponible pour estimer $V_p\left(\hat{\theta}_N\right)$, mais que l'on connaît l'effet de plan de sondage, alors les analystes peuvent utiliser le *deff* pour ajuster $\hat{V}_{SRS}\left(\hat{\theta}_N\right)$ produit par un logiciel standard et ainsi obtenir des intervalles de confiance ou des tests statistiques pour les proportions, les moyennes et les totaux de la population finie (voir la section 7.5 dans Lohr, 1999). Par exemple, un intervalle de confiance de 95 % pour une moyenne de la variable *y*, lorsque la taille de l'échantillon *n* est grande, peut être estimé comme $\hat{\bar{y}} \pm 1.96\sqrt{deff\left(\hat{\bar{y}}\right)}\sqrt{\hat{S}^2 \big/ n}$. L'estimateur ponctuel $\hat{\bar{y}} = \hat{\theta}_N = \sum_{i=1}^{n} w_i y_i \Big/ \sum_{i=1}^{n} w_i$ est un estimateur de $\bar{Y}_N = \theta_N = \sum_{i=1}^{N} y_i \Big/ N$ convergent par rapport au plan, et w_i est le poids d'enquête individuel *i,* et $\hat{S}^2 = \sum_{i=1}^{n} w_i \left(y_i - \hat{\bar{y}}_N\right)^2 \Big/ \left(\sum_{i=1}^{n} w_i - 1\right)$ est un estimateur de la variance de la population finie de la variable *y*, $S^2 = \sum_{i=1}^{N} \left(y_i - \bar{Y}_N\right)^2 \Big/ (N-1)$.

Lorsque l'on utilise des données d'enquêtes complexes pour ajuster un modèle de régression, les estimations de la variance des coefficients de régression individuels produits par un logiciel standard doivent être ajustées par les effets de plan de sondage individuels pour obtenir les intervalles de confiance $\hat{\beta}_k \pm t\sqrt{deff\left(\hat{\beta}_k\right)}\sqrt{\hat{V}_{SRS}\left(\hat{\beta}_k\right)},\ k = 1, \ldots, p.$ L'estimateur $\hat{\beta}_k$ est théoriquement l'estimation par la méthode des moindres carrés pondérés du coefficient de population finie $\beta_{N,k}$ et aussi du paramètre du modèle β_k. Les poids utilisés pour effectuer l'estimation par la méthode des moindres carrés pondérés sont les poids d'enquête et non les poids pour corriger l'hétérogénéité de la variance de l'erreur. Dans le cas d'un plan de sondage complexe, on peut habituellement effectuer une approximation du nombre des degrés de liberté pour la loi *t* de Student par $\kappa = \#grappes - \#strates$ (voir la section 3.5 dans Heeringa et coll., 2010).

Quant à eux, Korn et Graubard (1990) proposent d'utiliser la méthode de Bonferroni pour calculer l'inférence simultanée des paramètres de régression. En ce qui concerne le test global *F*, il faudrait une estimation

convergente par rapport au plan de la matrice de variance-covariance *p*x*p*, $\hat{\boldsymbol{V}}_p\left(\hat{\boldsymbol{\beta}}_k\right)$, ou une matrice des effets de plan de sondage *p*x*p* et $\hat{\boldsymbol{V}}_{SRS}\left(\hat{\boldsymbol{\beta}}_k\right)$.

Pour les tableaux de contingence, plusieurs méthodes ont été suggérées pour prendre en compte le plan de sondage pour le test de l'homogénéité des proportions de la population ou de l'indépendance des variables. La correction de premier ordre suggérée par Rao et Scott (1981; 1984) utilise les effets de plan de sondage pour estimer les proportions des cellules, rangées et colonnes. Pour obtenir plus de détails, consultez l'exemple 10.3.3 dans Lohr (1999) ou 6.4.4 dans Heeringa et coll. (2010).