ECHANTILLONNAGE ET ESTIMATION POUR LES ENQUETES CONTINUES : QUE MESURE-T-ON ?

J.-C. DEVILLE

ENSAI/CREST - Laboratoire de Statistique d'Enquête

Avertissement : ce texte a été écrit en 1996, à quelques détails près. On projetait alors deux opérations majeures avec collecte continue : le recensement (à rénover) et l'enquête Emploi (à refondre).

J'avais à l'époque évoqué de nombreuses fois les difficultés conceptuelles de ce type de données, indiquant qu'on ne devait pas se leurrer sur la nature des choses qu'on mesure.

Près de 5 ans après, ce texte me paraît toujours très actuel et il est toujours et vraisemblablement définitivement inachevé .

L'idée d'une collecte « continue » des données semble être une tendance « moderne » en matière d'enquête. Deux projets importants pour le début du siècle prochain devraient faire appel à cette technique : l'enquête Emploi refondue et le recensement permanent de la population. Les réflexions qui suivent s'appliquent essentiellement à la première. Cependant, bien des transpositions sont applicables à la seconde, même si des décisions récentes ont fortement remis en question l'idée d'une collecte continue, qui en était pourtant à la base.

Il est d'autre part illusoire de chercher ici une solution nette et définitive aux problèmes de définition et de choix d'estimateurs. On procédera uniquement de façon simplifiée et peut-être abusive. L'ambition est de donner des pistes d'études pour un véritable travail méthodologique sérieux. Il n'existe pas de catalogue de recettes applicable à des enquêtes continues où puiser une solution toute faite ; il y a un problème qu'il s'agit dans un premier temps de poser correctement. Même s'il est alors à moitié résolu (comme on dit), encore faut-il chercher sa solution. Celle-ci trouvée sur un schéma simplifié (excluant la non-réponse, par exemple), une mise au point opérationnelle prendra encore du temps, car il est d'expérience commune que les « détails » de mise au point engendrent souvent des délais insoupçonnés dans une « étude préalable ».

Cette « étude » est donc très simplifiée. En particulier, on ne cherchera pas à y intégrer des schémas d'échantillonnage ou d'estimation tenant compte du

renouvellement de la population. L'exposé de Caron-Ravalet dans cette même session fait assez complètement le tour de ce qui se pratique en cette matière. En général, pour obtenir des résultats analytiquement compréhensibles, on supposera que les sondages sont des sondages aléatoires simples. On fera aussi un usage assez libre de l'indépendance entre échantillons, ce qui se justifie si les taux de sondage sont assez faibles. Pour l'enquête Emploi « classique », ceci peut se justifier si on s'intéresse à une strate et qu'on admette qu'on y a réalisé un sondage aléatoire simple d'aires, qu'on ne se préoccupe pas des questions de compensation de non réponse ni d'introduction d'information auxiliaire dans les estimateurs. Bref, on laisse tomber franchement les choses les plus délicates ! Le plan de l'exposé voudrait être le suivant.

On examinera d'abord ce qu'apporte le concept d'observation continue et comment on dégage un formalisme sous-jacent de cette pratique. On verra ensuite les problèmes d'estimations liés à un recueil continu - biais, variance, estimation de variance etc... Il faudrait regarder ensuite les problèmes de recouvrement d'échantillon en rappelant quelques bons vieux principes : estimation de variation ou de niveau ; estimations composites. On se posera enfin le problème de l'articulation entre estimations trimestrielles, mensuelles et annuelles avec nuances selon que l'on veut s'intéresser à des niveaux ou à des variations.

(En fait, seul le début de ce programme est présent dans ce texte ; l'exposé de Caron-Ravalet dans la présente session des JMS couvre une bonne partie de la suite).

1. Constantes de temps et mesures continues.

Dans de multiples opérations statistiques, en particulier les recensements et les enquêtes sur l'emploi, la philosophie classique consiste à considérer qu'on mesure l'état de la population à une date précise fixée. L'optique est celle de la photographie instantanée (si je puis risquer cette « quasi-champignaquerie »). Le cas extrême est celui du recensement à la Turque, qui met la population dans une position très inconfortable ; en effet, on y décrète le couvre-feu, pour permettre aux militaires, seuls autorisés à circuler ce jour-là, de recenser la population. Et sinon, personne ne bouge !

En pratique, même en Turquie, le dispositif de collecte tolère cependant un certain flou sur ce point, et le recueil des données s'étale sur une période plus longue. La raison de ce laxisme est évidemment la commodité d'organisation de la collecte, le fait de recourir à un plus petit nombre d'enquêteurs mieux formés et, de façon générale, la diminution des coûts.

Qu'on puisse le tolérer repose sur deux idées. La première est que ce qu'on veut mesurer ne varie pas ou ne varie que peu au cours de la période de recueil. Cette faible variation s'articule elle-même sur deux points :

- On considère que pour la très grosse majorité des cas (plus de 99,9% disons) les questions posées pour construire les variables d'intérêt ont un sens stable pour la période de référence ou en son voisinage (où habitez-vous, par exemple).

- Surtout, on considère comme négligeable (disons 0.1%) la proportion des unités susceptibles d'être en train de « bouger » au moment de la période de référence de collecte.

La seconde est, qu'éventuellement, on peut faire appel aux souvenirs des personnes enquêtées pour reconstituer leur situation à la date de référence.

L'optique (toujours elle !) est d'augmenter le temps de pause pour recueillir plus de lumière en estimant que le sujet ne bougera pas. On sait bien que ce qui est possible pour le portrait (ne bougeons plus, c'est le recensement Turc) ne l'est guère pour la photo sportive. Une enquête d'opinion « à chaud », une enquête « sortie des urnes » doivent obligatoirement être réalisées dans un délai très court (un à deux jours au plus). Une enquête sur le logement peut être étalée sur plusieurs mois sans qu'on y voie de véritable inconvénient. Un peu comme en physique, on peut parler, dans les enquêtes sociales, de « constantes de temps » qui mesurent la durée pendant laquelle un phénomène est considéré comme établi et stable, au niveau individuel (la chose en question à un sens) comme au niveau collectif (la chose en question ne varie pas). Pour les enquêtes de population ou sur l'emploi, il semble que la pratique nous dise que la « constante » de temps est de l'ordre de la semaine ou de la quinzaine environ. L'expérience de l'enquête « Transitions » [*DTS*] est à cet égard assez instructive : on y observait des changements de situations assez rapides, focalisés sur le début du mois et accessoirement sur le milieu du mois, indiquant qu'une période de référence ou de collecte supérieure à deux semaines pouvait induire des biais de mesure gênants.

On admettra donc, que pour une opération statistique donnée, il existe une « constante de temps » à l'intérieur de laquelle la situation est

- bien définie
- peu variable
- collectable sans biais de mémoire notable.

Pour les enquêtes Emploi, on peut admettre qu'elle est de l'ordre de la semaine ou de la quinzaine. Pour le Recensement de la population, elle doit pouvoir s'établir à un ou deux mois. Pour certaines variables des enquêtes « Budget de Famille », ou « Budget-Temps », elle ne dépasse pas la journée (d'où les techniques de collecte

très particulières de ces enquêtes). Pour une enquête démographique du genre de l'enquête sur les familles, elle peut atteindre un an.

On dira qu'il y a une *observation continue si la période de collecte de l'information est supérieure à cette constante de temps et que l'information collectée se rapporte à la situation au moment de l'enquête* (ou à une situation relative à une date de référence moins éloignée de la date de collecte que la constante de temps).

Là encore, l'enquête « Transitions » a montré les limites d'une observation rétrospective : entre l'« Enquête Trimestrielle Emploi » (ETE) de septembre et la situation à la date de l'ETE reconstituée par « Transitions », on notait un taux de divergence de l'ordre de 4,8 %, soit presque aussi important que la proportion de personnes ayant une transition (6,3%).

Ces réflexions nous mènent à un formalisme nécessaire pour préciser ce que l'on va chercher à mesurer (variable d'intérêt) et, bien, sûr, sur la façon (estimateur) de le mesurer. On va même aller jusqu'à formaliser la question en termes de population variable, car ça ne coûte pas plus d'un centime en plus.

2. Variables d'intérêt : où les choses ne sont plus ce qu'on croyait.

On notera U_t la population au temps t et $U = \bigcup_{t \in T} U_t$ la réunion de tous les individus susceptibles d'être observés (champ de l'opération) sur T ensemble des temps possibles. A tout individu k de U est associée une fonction $y(t)$ à valeurs dans R^q décrivant l'ensemble des variables collectables utiles à l'étude. Elle est définie pour les valeurs de t ; une de ses coordonnées, $e(t)$, « existence à la date t », vaut 1 si $k \in U_t$ et 0 sinon, 1 si k est « vivant » en t et 0 sinon. Si $e_k(t)=0$ alors $y_k(t) = 0$. Toutes les statistiques d'intérêt vont s'écrire en fonction de totaux formés sur les variables $y_k(t)$. Par exemple, la taille de la population à l'époque t vaut

$$N_t = \sum_{U_t} 1 = \sum_{U} e_k(t).$$

Les principales autres statistiques d'intérêt peuvent être caractérisées comme suit :

a) Les statistiques datées : N_t, $Y(t) = \sum_U y_k(t)$, $\overline{Y}(t) = Y(t) / N_t$ etc.

C'est l'optique « classique » de la statistique décrite plus haut, objectif des recensements « classiques » ainsi que des enquêtes Emploi annuelles « classiques ».

b) Les cumuls temporels. Soit T une partie de I (généralement un intervalle). On peut ressentir le besoin d'évaluer une quantité de la forme :

$$Y_T = \int_T Y(t)dt = \sum_{k \in U} \int y_k(t)dt = \sum_{k \in U} y_{k,T}$$

Remarque : Notez l'intérêt de la convention introduite plus haut : les sommes portent toujours sur la même population.

Ce type de statistique peut surprendre un peu les habitués de photos de la population « shootés » aux recensements et enquêtes Emploi « classiques ». Néanmoins, il fonctionne déjà, à la satisfaction générale, dans de nombreux domaines. Dans le cadre des enquêtes « classiques », tout ce qui fait appel à un carnet de recueil (budget, consommation alimentaire, temps, transport) a pour but de chiffrer un cumul. Au niveau individuel, le carnet n'est, au fond, que destiné à retracer un cumul sur la période d'observation. Au niveau global, compte tenu de divers facteurs dont les variations périodiques saisonnières, mensuelles, hebdomadaires, on est généralement intéressé par le cumul des consommations sur une année. Comme on le sait, on essaie de parvenir à ce but par un étalement des observations par vague sur l'ensemble de l'année. Pour une approche du problème, voir *[JCD]*.

c) Des choses plus compliquées liées aux études longitudinales.

Le concept fonctionne aussi dans les études démographiques et plus généralement dans les statistiques de type épidémiologique ou longitudinal. Le développement des modèles de durée leur donne une importance accrue. A titre d'exemple, on peut étudier la « mortalité » de la façon suivante. Pour une unité k, on définit la variable $d(t)$ en la fixant à 0 pour $t = t_0 = \inf T$ et en augmentant $d(t)$ de 1 chaque fois que $e(t)$ passe de 1 à 0 ; $d(t)$ est le nombre d'entrée en « décès » de k entre t_o et t.

La méthode peut s'appliquer aux événements renouvelables comme le chômage, la natalité etc...). Le taux de mortalité (pendant T) sera : $\sum_U (d_k(t_1) - d_k(t_0)) / \sum_U \int_T e_k(t)dt$,

c'est-à-dire le nombre d'événements divisé par le temps total d'exposition au risque.

C'est grâce à des techniques de ce genre qu'est exploitée, par exemple, la partie « fécondité » de l'enquête Famille. La méthode peut se raffiner si l'on veut estimer de façon non paramétrique la fonction de risque du phénomène. On a alors à former des objets qui sont de la nature d'intégrale produit *(voir par exemple [ABGK])*.

Dans toute cette description, insistons, il n'est question que de la population, et à aucun moment de technique d'échantillonnage ou de recueil de données destinées à répondre au problème de leur estimation.

d) Des variations

- Brutes :

$$Y(t_2) - Y(t_1) = \sum_{U_{t_2}} y_k(t_2) - \sum_{U_{t_1}} y_k(t_1) = \sum_{U} [y_k(t_2) - y_k(t_1)]$$

A noter que, dans la ligne du *b*, si on invente une dérivée $\delta y_h(t)$ (à l'aide en particulier de masses de Dirac pour tenir compte des discontinuités), cette variation est du style :

$$\sum_{U} \int_{t_1}^{t_2} \delta y_k(t)\, dt = \sum_{U} \Delta_T y_k$$

- « A champ constant » :

$$\Delta_T y_k \;=\; \Delta_T = \sum_{U_{t_2} \cap U_{t_1}} \sum_{U} e_k(t_1) e_k(t_0) (\Delta_T y_k)$$

ou, alternativement :

$$\Delta_T y = \sum_{\cap U_t} \Delta_T y_k = \sum_{U} \left(\prod_{t \in T} e_k(t)\right) \Delta_T(y_k)$$

On peut aussi fabriquer des variations de statistiques plus complexes, ratios, fractiles d'une distribution, indice de Gini etc...

On peut aussi, pour des variables de la nature de flux, où l'optique naturelle est de construire des statistiques du type *b* - cumul temporel -, avoir besoin de chiffrer des évolutions de type $Y_T - Y_S$.

On conçoit bien que certaines des statistiques envisagées jusqu'ici sont, par nature, plus adaptées à un mode de recueil temporel périodique alors que certaines autres demanderont plutôt une collecte continue. La suite de cette note exploratoire va chercher à cerner ce que l'on peut faire quand on choisit une collecte continue.

3. *Collecte continue : statistiques relatives à une période élémentaire.*

Par définition, une collecte continue peut très difficilement avoir pour but une estimation directe des valeurs Y_t pour tous les *t*, ni même pour un ensemble fini de *t* séparés par la « constante de temps » du phénomène. Pour l'enquête Emploi, cela signifierait qu'on veut évaluer des Y_t pour des *t* séparés d'une semaine ou d'une quinzaine. Cela voudrait dire, en particulier, qu'on dispose d'un échantillonnage suffisant par sa taille (représentativité en un certain sens !) et par sa structure (représentativité en un autre sens !) pour estimer Y_t de façon fiable.

Une période élémentaire d'observation sera un intervalle de temps au cours duquel le recueil des données permettra des extrapolations fiables. Pour être un peu concret, dans le cas de l'enquête Emploi, la période élémentaire d'observation devrait être le trimestre, pour le recensement (continu), l'année.

Tentons une définition formelle de la période élémentaire d'observation. On admettra, d'abord, que les seules quantités observables sont des $y_k(t_k)$, t_k date (éventuellement de référence !) de l'observation de *y* chez Monsieur *k*. Le cas de données rétrospectives ou de carnet de compte peut éventuellement s'adapter assez facilement à ce formalisme, mais ceci est peu important pour la suite et nous laisserons de côté ce problème.

Une période élémentaire d'observation $T = [t_0, t_1]$ sera caractérisée par :

- un échantillon s_T tiré de la population $U_T = \underset{t \in T}{U} U_t$ selon un plan de sondage P_T parfaitement défini et contrôlé (disons un échantillon extrapolable d'aires = réunions de sous-échantillons dans la terminologie habituelle de l'enquête Emploi).

En pratique, U_t devrait peu varier et un échantillon tiré dans U_{t_0} devrait faire l'affaire. En réalité (noter la différence avec « en pratique » !), on aura à utiliser une base de sondage datée, antérieure à t_0, et il y a quelques autres menus problèmes.

C'est la raison pour laquelle, dans cette première approximation, nous allons faire comme si U_t était constante, au moins pour *t* dans *T*, et comme on le verra plus loin, pour des périodes adjacentes. Sinon, ça devient tout de suite assez compliqué et on risque de se perdre (le lecteur avisé aura remarqué, qu'au fond, on ne fait qu'exploiter la différence de constante de temps entre les variables d'intérêt et les variables d'appartenance à la population).

- une répartition temporelle des observations des unités k de s_T . A chacune des unités k est affectée une date d'observation t_k aléatoire mais contrôlée par le statisticien. Formellement donc, on travaillera avec la loi des $(t_k, k \in s_T)$, conditionnellement à s_T.

Une condition naturelle qu'on puisse requérir sur les t_k est qu'elles suivent la même loi (l'étiquette k ne doit pas être informative). En revanche, la loi jointe des t_k ne doit surtout pas être simpliste : le tirage i.i.d. des t_k serait évidemment une catastrophe. Deux objectifs apparaissent naturellement :

1) que les observations soient équilibrées dans le temps : si T est décomposé en sous-périodes (semaines), les t_k doivent être contrôlées de façon à respecter une certaine répartition entre sous-périodes. Les sous-périodes constituent une sorte de critère de stratification supplémentaire.

2) que certaines observations (en pratique, géographiquement proches) soient effectuées au cours de la même sous-période.

Dans une pratique (un peu idéalisée) de l'enquête Emploi, ceci conduit, par exemple, à une affectation aléatoire équilibrée d'aires à des semaines de collecte, ou, dans un schéma nécessairement plus complexe, à des quinzaines ou des périodes plus longues. Dans un tel schéma, chaque unité k se trouve dotée d'une probabilité d'inclusion π_k et d'une date d'observation t_k de loi connue, disons uniforme sur T pour fixer les idées (attention : nous laissons au lecteur le soin d'imaginer ce qui se passe si l'on met des probabilités non uniformes sur T avec une densité $a(t)$). Cette procédure serait parfaitement justifiée s'il se passe beaucoup de choses à certaines sous-périodes connues du calendrier, mois de septembre par exemple).

Prenons le problème à l'envers et voyons ce qu'on estime si on forme la statistique la plus naturelle qui soit :

$$\hat{Y} = \sum_{s_T} \frac{y_k(t_k)}{\pi_k}$$

L'espérance de $\hat{Y}$ pour commencer.

On a :

$$E(\hat{Y} / s_T) = \sum_{s_T} \frac{1}{\pi_k} \underset{(t_k)}{E}\, y_k(t_k)$$

Or : $Ey_k(t_k) = \int_T y_k(t)dt = y_{k,T} = T\bar{y}_{k,T}$ (Définition).

D'où :

$$E(\hat{Y}) = \sum_U y_{k,T} = \int_T (\sum_U y_k(t))dt = \int_T Y_t dt = T\overline{Y}_T$$

L'espérance de $\hat{Y}$ est donc la valeur moyenne, au cours de la période, du total de Y. Fondamentalement, l'observation en continu ne permet pas d'estimer d'autres statistiques que des cumuls sur des périodes élémentaires (et naturellement des fonctions plus ou moins complexes de ces cumuls).

En particulier, estimer $Y(t_1)$, même compte tenu de la constante de temps assimilant t_1 à une sous-période est hors de portée des techniques basées sur l'extrapolation liées au plan de sondage. Techniquement, l'estimation de $Y(t_1)$ est une estimation de type « petit domaine ». On renforce un estimateur en important des données grâce à un modèle.

4. *Estimation relative à une date fixée vue comme un petit domaine.*

Donnons un exemple simple (mais parfaitement réaliste). Postulons, au cours de la période T, une évolution linéaire de Y(t), donc de la forme :

$$Y(t) = Y_0 + B(t - t^*) + \varepsilon(t)$$

$t^* = \frac{t_0 + t_1}{2}$, milieu de la période, $\varepsilon(t)$ résidus et B minimisant $\int_T \varepsilon(t)^2 dt$. Pour simplifier les notations, prenons l'origine du temps en t^*. Pour simplifier (encore !) le problème, supposons qu'on a des probabilités d'inclusion π_k égales et que la stratification des t_k nous assure l'équilibrage des deux premiers moments :

$$\sum_{s_T} t_k = 0$$

$$\sum_{s_T} t_k^2 = n\int_T t^2 dt = n\frac{T^2}{12}$$

On peut estimer B à partir des données. Dans le cas particulier que nous examinons, cela donnera :

$$\hat{B} = \frac{\sum_{s_T} (y_k(t_k) - \bar{y})\, t_k}{\sum_{s_T} t_k^2} = \frac{12}{T^2} \frac{1}{n} \sum_{s_T} t_k y_k(t_k).$$

Il est facile de voir que $\hat{B}$ est un estimateur sans biais de B, moyenne des quantités observées sur l'échantillon $z_k = t_k y_k(t_k)$.

De ce fait, on estimera, sous ce qui devient alors un modèle, $Y(t)$ par :

$$\hat{Y}(t) = \frac{1}{T} \hat{Y}_T + \hat{B}\, t = \hat{\bar{Y}} + \hat{B}\, t \quad (ici\ \hat{\bar{Y}}_T = \bar{y})$$

En particulier, une estimation en fin de période devient possible (et même au-delà si l'on veut !). Cela semble miraculeux mais il y a un prix !

D'abord, la fiabilité du modèle linéaire même sur une période aussi courte que 13 semaines. On peut envisager certaines améliorations de ce côté :

- en changeant un peu la forme fonctionnelle du régresseur : l'utilisation d'une fonction symétrique *S(t)* « sigmoïdale » permettrait de prendre moins de risques « au bord » de l'intervalle.
- on peut aussi perdre un ou deux degrés de liberté pour affiner l'ajustement.
- on pourrait, enfin, chercher à utiliser des modèles de type série chronologique, en introduisant soit une variable décalée, soit une autocorrélation des résidus $\varepsilon(t)$, ce qui revient un peu au même.

Ensuite, il va falloir parler de variance et d'estimation de variance.

Variance : Regardons d'abord ce qui concerne $\hat{Y}$. On a :

$$Var(\hat{Y}) = Var(E\,\hat{Y} \,|s_T) + E\, Var(\hat{Y} / s_T)$$

$$= \mathrm{Var} \sum_{s_T} \frac{1}{\pi_k} y_{k,T} + E_P[\, Var_{(t_k)} (\sum y_k(t_k) / \pi_k)^2]$$

Le premier terme est la variance, pour le plan, de la moyenne individuelle des $y_k(t)$. Cette quantité est inobservée. Néanmoins, elle est de l'ordre de la variance d'un total à date fixée, t^* par exemple.

Le deuxième terme, $E_p V_2$, est une variance due au caractère aléatoire de la répartition des t_k. Tout se passe comme si l'on voulait mesurer $y_k(t^*)$ - date fixe - et qu'une imprécision de pointage temporel - t_k au lieu de t^* - provoque une erreur de mesure. On peut écrire :

$$V_2 = \sum_{s_T} \frac{1}{\pi_k^2} \int (y_k(t) - y_{k,T})^2 \, dt + 2 \sum_{k<l} \sum_{\in s_T} \frac{1}{\pi_k \pi_l} \iint (y_k(t_k) - y_{k,T})(y_l(t_l) - y_{l,T}) d(t_k, t_l)$$

où l'on note $d(t_k, t_l)$ la loi jointe de t_k et t_l.

S'il y a indépendance entre les t_k (ce qu'il faut éviter absolument !), le deuxième terme de V_2 est nul. Le premier a pour espérance :

$$\sum_U \frac{1}{\pi_k} v_k \qquad \text{avec} \qquad v_k = \int (y_k(t) - y_{k,T})^2 \, dt.$$

Il faut beaucoup se méfier de ce terme à cause du « gros » facteur $\frac{1}{\pi_k}$. Sous un modèle simple sur les $y_k(t)$, on peut évaluer les ordres de grandeur. Voir plus loin.

Le deuxième terme doit être négatif. Pour cela, il faut « programmer » une répartition des t_k qui assure une corrélation négative la plus grande possible entre tous les couples $(y_k(t_k), y_l(t_l))$. Comment faire ? Ce n'est pas évident a priori, d'autant plus que le fait de traiter la répartition temporelle en grappes n'arrange rien.

En effet, si, par exemple, on répartit indépendamment (ce qu'il ne faut surtout pas faire !) les «grappes » (= aires par exemple) dans le temps, on obtient le fait que t_k et t_l sont indépendants si k et l sont dans des grappes distinctes, $t_k = t_l$ s'ils sont dans la même. Les covariances sont nulles dans le premier cas, dans le second elles valent

$$\int (y_k(t) - y_{kT})(y_l(t) - y_{l,T}) \, dt = c_{kl}$$

Dire que $c_{kl} > 0$, c'est dire que Messieurs k et l, de la même aire, subissent la même évolution (transition vers le chômage) au cours de la période ; ça paraît plus probable, en matière d'emploi, que le contraire (Monsieur *k* est licencié, remplacé par Monsieur *l* son voisin !). Voir encore le modèle annoncé plus haut pour plus loin.

Fermons provisoirement le débat là-dessus. Juste un petit coup d'oeil sur l'Estimation de variance. Les $y_k(t)$ apparaissent comme des mesures avec erreurs des $y_{k,T}$, et celles-ci sont assez bien prises en compte dans l'estimation de variance.

Vérifions ça une fois de plus. Supposons qu'on sache estimer la variance d'un total pour l'estimateur de Horvitz-Thompson (POULPE) :

$$\hat{V}(\hat{Y}) = \sum_s \sum \Delta_{kl} y_k y_l$$

Appliquons cela aux données $y_k(t_k)$:

$$\hat{V}(\hat{Y}) = \sum_s \Delta_{kk} y_k(t_k)^2 + 2 \sum_{kl} \sum \Delta_{kl} y_k(t_k) y_l(t_l)$$

Il reste les aléas temporels, mais, si l'on prend l'espérance, il vient :

$$\underset{(t_k)}{E} \hat{V}(\hat{Y}) = \sum_s \sum \Delta_{kk} y_{kT} y_{lT} + \sum_s \Delta_{kk} v_k + 2 \sum_{kl} \sum c_{kl}$$

Or : $\Delta_{kk} = \frac{1}{\pi_k^2}(1 - \pi_k)$, ce qui fait que le second terme capture à peu près toute l'erreur de mesure s'il y a indépendance des t_k. Restent les covariances c_{kl} qu'on espère rendre négatives.

Revenons maintenant à notre propos qui est le prix à payer pour estimer $y(t_1)$ à l'aide du modèle linéaire $Y(t) = Y(t^*) + B(t - t^*)$. A partir de ce qui est écrit plus haut, on peut écrire :

$$\hat{V}(\hat{Y}_{t_1}) = \hat{Y}_T + t\hat{B} \implies Var(\hat{Y}_{t_1}) = \frac{1}{T^2} Var(\hat{Y}_T) + t^2 Var(\hat{B}) + 2t Cov(\hat{Y}_T, \hat{B}).$$

Avec le contrôle supposé de la répartition des t_k (et les poids de sondage uniformes), la variance de $\hat{B}$ est celle du total des $z_k = t_k y_k(t_k)$, chose que l'on sait (plus ou moins) étudier et estimer. Il en va de même de $Cov(\hat{Y}_T, \hat{B})$, mais on peut aller plus loin sur ce terme : en cas de sondage aléatoire simple (ou de sondage aléatoire simple de grappes affectées uniformément au même t_k), cette quantité est évidemment nulle, car si $e_k = y_k(t_k) - A - Bt_k$ est le résidu, alors les variances-covariances de $\hat{A} = \hat{Y}$ et $\hat{B}$ sont proportionnelles à $\begin{pmatrix} \sum e_k^2 & \sum t_k e_k \\ & \sum t_k^2 \end{pmatrix}$, qui est une

matrice diagonale. On a donc, rigoureusement dans le cas qui vient d'être dit et sans doute approximativement dans le cas général :

$$Var(\hat{Y}(t_1)) = \frac{1}{T^2} Var(\hat{Y}) + t^2 Var(\hat{B})$$

et le second terme est le fameux prix à payer. Ceci dit, l'estimateur de $Y(t_1)$ prend une allure aussi sympathique que paradoxale d'estimateur pondéré. En effet, si l'on regarde l'estimateur de la moyenne, on aura (dans le cas du sondage aléatoire simple) :

$$\hat{\bar{Y}}(\frac{T}{2}) = \bar{y} + \hat{B}\frac{T}{2}$$

$$= \bar{y} + \frac{12}{T^2}\frac{1}{n}(\sum t_k y_k).\frac{T}{2} = \frac{1}{n}\sum\left(1 + \frac{6t_k}{T}\right)y_k$$

Comme t_k varie uniformément de $-\frac{T}{2}$ à $+\frac{T}{2}$, les poids seront corrigés par un facteur variant de - 2 à + 4. Ceci semble indiquer une variance assez importante (voir encore l'exercice plus loin).

Conclusion : Il reste du travail pour soigner la répartition des enquêtes à l'intérieur du trimestre (échantillonnage équilibré), à examiner la valeur pratique des estimateurs en bord de période, à examiner si le $\bar{Y}(t_*)$ (milieu de période) peut être amélioré avec un modèle linéaire un peu plus rusé par exemple :

$y_t = A + B^- t$ si $t < t_*$

$= A + B^* t$ si $t > t_*$.

Ceci n'engage en rien la question de fabrication d'estimateurs utilisant la rotation d'échantillons, point qui est étudié dans un autre article.

Il n'y a pas de suite, et ça m'étonnerait qu'il y en ait une un jour.

Références :

DEVILLE J.C, (1987), Sur la durée d'observation dans les enquêtes à carnets de compte, Annales d'économie et de statistique, n°5.

ANDERSEN, BORGAN, GILL, KEIDING (1995), Statistical Methods based on counting Processes, Springer-Verlag .

DETOUR C., THIESSET C., SCHUHL P. (1997), Résultats de l'enquête TRANSITIONS sur le marché du travail, Actes des journées de Méthodologie Statistiques du 18/19 octobre 1995 (INSEE Méthodes 59-61).

CARON, N. et RAVALET, Ph. (2000), Estimation dans les enquêtes répétées (VII èmes Journées de Méthodologie Statistique).

BOSREDON J. et FEVRIER Ph. (2000), Estimations dans l'enquête Emploi en continu (VII èmes Journées de Méthodologie Statistique).