ANNALES D'ÉCONOMIE ET DE STATISTIQUE. – N° 5 — 1987

Sur la durée d'observation dans les enquêtes à carnets de compte

Jean-Claude DEVILLE *

RÉSUMÉ. – Les enquêtes de consommation comportent souvent un relevé des dépenses des ménages au cours d'une période déterminée. A l'aide d'un modèle sur les dates et le montant des achats, on trouve des conditions sur la répartition des observations dans le temps, pour estimer sans biais la dépense totale annuelle ainsi qu'une formule exacte de la variance de cette estimation. On en déduit des règles permettant d'améliorer la précision des enquêtes en particulier dans le cas d'achats périodiques.

On the Observation Periods in Surveys of Household Spending Records

ABSTRACT. – Surveys of consumption often rest on household records of expenditures over a given period. With the aid of a model relating to dates and amounts of purchases, one can drive conditions for the repartition in time of the observations thats yield unbiased estimates of annual spending. An exact formula for the variance of the estimator follows. It is possible also to derive some rules that permit improving the accuracy of the surveys, especially those concerning periodic purchases.

* J. C. DEVILLE : INSEE, Département Population Ménage, 18, boulevard A. Pinard, 75675 Paris Cedex 14.

1 Introduction

Le recueil de données individuelles relatives à la consommation des ménages nécessite une méthodologie très élaborée. Dans les enquêtes sur la consommation alimentaire ou sur les budgets des familles, telles que l'INSEE les réalise, la technique utilisée est la suivante : un échantillon aléatoire de ménages reçoit un « carnet de comptes » destiné à enregistrer tous les achats effectués pendant une période déterminée. Celle-ci est fixée par le plan de sondage et varie d'un ménage à l'autre. Sa durée, toutefois, est généralement la même pour tout le monde, fixée par le caractère plus ou moins pénible de l'enquête, la lassitude qu'elle peut provoquer et donc, in fine, par la fiabilité des résultats qu'on peut en escompter. Une certaine latitude subsiste cependant. Ainsi, dans la préparation de l'enquête sur les budgets de 1985, il fallait étudier les avantages éventuels d'une durée d'observation portée à 14 jours au lieu de 10 pour l'enquête antérieure. Le but était d'estimer une dépense totale annuelle par groupe de produits ou par catégorie de population.

La méthodologie des enquêtes à carnets de compte est analysée par GLAUDE [1982], GLAUDE et MOUTARDIER [1983]. Pour l'enquête sur les budgets de familles, une documentation complète est donnée dans INSEE [1983 et 1984]. Dans cet article, on présente un modèle théorique bâti autour de quelques hypothèses plausibles relatives aux périodicités annuelles ou hebdomataires des achats et sur la liaison entre leur fréquence et leurs montants. Le comportement des ménages est décrit de façon très grossière pour être « manipulable » analytiquement. L'objectif est uniquement de chiffrer un ordre de grandeur des effets d'une décision à prendre, et non pas une estimation statistique au sens habituel qui ne présenterait guère d'intérêt.

Les outils probabilistes sont élémentaires et peuvent être trouvés dans DEVILLE [1984], SNYDER [1976] ou, de façon très compacte, dans KOROLIUK [1983]. L'objectif de l'article est d'indiquer un facteur d'efficacité dans les sondages temporels et l'ambition de modéliser le comportement des ménages est très limitée.

On trouvera des développements plus fins basés sur une théorie microéconomique dans quelques rares articles. Dans HAUSER et WISNIEWSKI [1982] on utilise la théorie des processus de Markov pour un modèle destiné à être utile dans les études de marché. Dans SINGER [1982] ou HECKMANN et SINGER [1984], on incorpore, dans une démarche analogue, quelques éléments de la théorie moderne des processus ponctuels (AALEN [1978], GILL [1980] pour un exposé très mathématisé ou l'excellente synthèse de ANDERSEN et BORGAN [1985]). Sur un problème analogue (la traite des vaches) on pourra signaler l'approche de BARTLETT [1986]. Bien que le recours aux processus ponctuels ou aux mesures poissonniennes puisse augmenter la portée du modèle ici présenté, nous préférons nous limiter à une approche élémentaire.

2 Un modèle relatif aux achats

Les achats successifs sont modélisés par un processus de Poisson non homogène dont l'intensité temporelle est $kp(t)\,dt$. Le nombre d'achat $N(T)$ survenant au cours d'une période T suit donc une loi de Poisson de paramètre $k\int_T p(t)\,dt = kp(T)$. Les variables aléatoires entières $N(T_1)$ et $N(T_2)$ relatives à deux périodes disjointes sont indépendantes. La fonction p du temps est supposée périodique, de période égale à 1 an (qui est l'unité de temps). Elle vérifie en outre $\int_A p(t)\,dt = 1$ ou A désigne un intervalle d'un an. Quant à k, c'est un paramètre aléatoire au niveau de la population qui fixe l'intensité du processus des achats pour chacun des individus qui la compose. A k fixé (et donc pour chaque individu de la population), le nombre annuel d'achat suit une loi de Poisson d'espérance égale à k (et de variance valant également k).

Toujours conditionnellement à k, le montant d'un achat réalisé à l'instant t est une variable aléatoire indépendante du processus générateur des instants d'achats et, naturellement, de la période aléatoire d'observation T. Les montants des achats successifs $M(t_i)$, pour $i=1$ à $N(T)$, sont considérés comme échantillonnés dans un processus du second ordre $M(t)$ de moyenne $\bar{M}(t)$ indépendant du processus $N(t)$ conditionnellement à k.

On notera $\bar{M}_k(t)$ et $V_k(t)$ l'espérance et la variance de $M(t)$ conditionnellement à k, et on admettra que ces fonctions ont une périodicité annuelle. On supposera enfin que les corrélations entre $M(t)$ et $M(s)$ ne dépendent pas de k.

On désire estimer une dépense totale annuelle. Intuitivement, mais cela sera justifié, la dépense moyenne par unité de temps sera, à k fixé, de l'ordre de $\bar{M}_k(t)\,kp(t)$. On posera

$$D(t) = k\,M(t),$$
$$\bar{D}_k(t) = k\,\bar{M}_k(t) = E(D(t)\,|\,k),$$

et

$$\bar{D}(t) = ED(t)$$

A k fixé, la dépense totale annuelle sera :

$$\bar{D}_k = \int_A \bar{D}_k(t)\,p(t)\,dt \tag{1}$$

La quantité à estimer devrait donc être :

$$\bar{D} = E\,\bar{D}_k = E\int_A k\,M(t)\,p(t)\,dt.$$

Pour estimer la dépense annuelle, on procède à des observations supposées indépendantes auprès d'individus de la population. Pour chacun d'eux on observe le processus d'achats et le montant de ceux-ci sur une période T de longueur fixée $|T|$. La loi de T (disons de la date de début d'observation) est définie par le statisticien et fait partie intégrante du plan de sondage de l'enquête. Celui-ci se modèlise donc de la façon suivante : échantillonnage dans la population (assimilé à un sondage aléatoire simple avec remise) et affectation à chaque unité échantillonnée d'une période d'observation considérée comme indépendante de l'échantillonnage.

Les observations seront donc générées par le mécanisme suivant :

* un échantillon d'individus est obtenu par n réalisations indépendantes d'un aléa dans un espace probabilisé. On peut noter les individus $\omega_i (i = 1 \text{ à } n)$.

* Pour chaque i une variable aléatoire inobservable $k(\omega_i)$ paramétrise un processus de Poisson d'intensité $k(\omega_i) p(t)\, dt$ où p est périodique de période 1.

* Notons S l'ensemble des « sauts » du processus de Poisson relatif à ω_i. Pour chaque t de S, $M(t)$ est une variable aléatoire de $\mathbb{R}^+$ dont la loi ne dépend que de $k(\omega_i)$. Les $M(t)$ successifs ont des caractéristiques au second ordre définies par :

$$V_k(t) = \mathrm{Var}(M_t | k) = CV^2 \, E(M_t | k)^2 = CV^2 \, \bar{M}_k^2(t) \tag{2}$$

(hypothèse simplificatrice de coefficient de variation constant)

$$\begin{aligned} \mathrm{Cov}(M_t, M_s | k) &= C_k(t, s) \\ &= (V_k(t) V_k(s))^{1/2} r(t, s) \\ &= CV^2 \bar{M}_k(t) \bar{M}_k(s) r(t, s) \end{aligned} \tag{3}$$

* Indépendamment de k, de S et des $M(t)$, une variable $T(\omega_i)$ détermine un intervalle aléatoire de longueur fixe $|T|$. On peut voir cette variable comme paramétrée par [0, 1[.

* On observe des instants t qui appartiennent à T et les montants $M(t)$ correspondants

3 Quantité à estimer et estimateur

Le but est d'estimer la dépense moyenne dans la population c'est-à-dire l'espérance de la variable aléatoire :

$$DT = \sum_{t \in S} M(t). \tag{4}$$

En conditionnant par le nombre N(A) d'achat dans l'année, les t de S suivent indépendamment la même loi de densité $p(t)\, dt$ (Deville [1984]).

On a donc :

$$E(DT \mid k, N(A)) = N(A) \int \bar{M}_k(t) p(t)\, dt \tag{5}$$

d'où

$$E(DT \mid k) = k \int \bar{M}_k(t) p(t)\, dt = \int \bar{D}_k(t) p(t)\, dt \tag{6}$$

soit enfin :

$$E(DT) = \bar{D}$$

Comme on pouvait s'y attendre la quantité à estimer est donc bien $\bar{D}$. Si la dépense annuelle était connue pour chaque individu de l'échantillon, $\bar{D}$ serait estimé par la moyenne empirique des dépenses annuelles DT. En fait, chaque dépense au niveau individuel est-elle même l'objet d'une estimation à cause de l'aspect temporel du plan de sondage. On retiendra pour chaque individu de l'échantillon l'estimateur suivant de sa dépense annuelle :

$$\hat{D} = \frac{1}{|T|} \sum_{t \in S \cap T} M(t). \tag{7}$$

$\hat{D}$ est la dépense moyenne par unité de temps au cours de la période d'observation. L'estimateur final de $\hat{D}$ sera la moyenne sur l'échantillon de $\hat{D}$. L'écart quadratique moyen de cet estimateur vaudra 1/2 ou $n^{-1}E(\hat{D} - \bar{D})^2$. On cherchera donc la loi de T permettant un échantillonnage sans biais, puis à évaluer la variance de l'estimation sous ces conditions. Tout revient donc à étudier les propriétés statistiques de la vatiable aléatoire $\hat{D}$ qui dépend, rappelons-le, des aléas k, T, S et $M(t)(t \in S)$.

4 Condition pour que l'estimateur soit sans biais

On a immédiatement :

$$E(\hat{D} \mid T, S, k) = \frac{1}{|T|} \sum_{t \in S \cap T} E(M(t) \mid T, S, k).$$

Conditionnellement à N(T) les t de $S \cap T$ suivent indépendamment sur T une loi de densité $p(t)\, dt / p(T)$. Il vient donc :

$$E(\hat{D} \mid T, k, N(T)) = \frac{N(T)}{|T| p(T)} \int_T \bar{M}_k(t) p(t)\, dt = \frac{N(T)}{k p(T)} \bar{D}_k(T)$$

en notant :

$$\bar{D}_k(T) = \frac{1}{|T|}\int_T \bar{D}_k(t)p(t)\,dt$$

Comme $EN(T) = kp(T)$, il vient :

(9) $$E(\hat{D}\,|\,T, k) = \bar{D}_k(T)$$

et

(10) $$E(\hat{D}\,|\,T) = \bar{D}(T) = \frac{1}{|T|}\int_T \bar{D}(t)p(t)\,dt$$

$$E(\hat{D}) = E\,\bar{D}(T)$$

Si en particulier $\bar{D}(t)$ est constant on a :

$$E(\hat{D}) = \bar{D}E\bar{p}(T) \qquad \text{avec} \quad \bar{p}(T) = \frac{1}{|T|}p(T)$$

Pour que l'estimateur soit sans biais il faut et il suffit donc que $E\bar{p}(T) = 1$ dans ce cas particulier et que $E\bar{D}(T) = \bar{D}$ dans le cas général. Ces espérances ne dépendent plus que de la loi de densité $l(t)\,dt$ choisie pour l'échantillonnage temporel.

PROPOSITION : L'estimateur est sans biais dans les cas suivants :

i. quelle que soit la loi de T si $\bar{D}(t)p(t)$ est constante;

ii. quelle que soit la loi de T si $\bar{D}(t)p(t)$ est périodique avec une période dont $|T|$ est un multiple entier;

iii. avec une loi uniforme uniquement si on désire un résultat valable pour toute fonction M ou tout toute fonction *p*.

Démonstration et commentaire : le point (*i*) est évident et correspond, pratiquement, à un flux de dépenses complètement homogène dans le temps sans caractère périodique ni saisonnier. Pour démontrer (*ii*) il suffit de remarquer que $\bar{D}(T)$ est la moyenne de $\bar{D}(t)p(t)$ sur plusieurs périodes et vaut par suite la moyenne sur toute l'année (à condition que la période soit un sous-multiple exact de l'année, ou approximativement, que l'année contienne un grand nombre de périodes, ce qui correspond au cas pratique des 52 semaines de l'année). Cette condition montre que si la période de la fonction $\bar{D}(t)p(t)$ est d'une semaine, le jour de début d'observation n'a pas d'importance sur le biais (ni sur la variance) à condition de fixer la durée d'observation à un nombre entier de semaines.

Passons au point (*iii*). Il correspond au cas ou on ignore tout de *p* ou de M, ou bien, plus concrétement, au cas ou une même enquête doit fournir des données sur les produits assez différents. Si $l(t) = 1$ pour tout *t*, on a :

$$E\bar{D}(T) = \frac{1}{|T|}\int_0^1 du \int_u^{u+|T|} \bar{D}(t)p(t)\,dt$$

$$= \frac{1}{|T|} \int_0^1 du \int_0^{|T|} p(u+t)\,\bar{D}(u+t)\,dt$$

$$= \frac{1}{|T|} \int_0^{|T|} dt \int_0^1 p(u+t)\,\bar{D}(u+t)\,du$$

Comme p et $\bar{D}(t)$ ont par hypothèse la période 1, la dernière intégrale vaut $\bar{D}$ d'où le résultat. Inversement, si $p(t)\,\bar{D}(t)$ n'est pas constante il suffit de prendre $l(t) > 1$ pour les t ou cette fonction est plus grande que $\bar{D}$ et nulle ailleurs pour qu'il y ait un biais. □

Pour beaucoup de postes de consommation il est naturel de supposer p périodique hebdomadaire et de supposer $\bar{M}$ constant (ou peu variable). Ceci dit on supposera dans toute la suite que $\hat{D}$ est sans biais et même que $l(t) = 1$ ce qui correspond à une répartition uniforme des enquêtes dans l'année.

5 La variance de $\hat{D}$

On utilise encore une technique de déconditionnements successifs. Pour k, T et S fixés on a :

$$(11)\quad \mathrm{Var}(\hat{D} \mid k, S, T) = \frac{1}{|T|^2} \mathrm{Var}\Big(\sum_{t \in S \cap T} E(M_t \mid k, S, T) \Big)$$

$$= \frac{1}{|T|^2} \Big(\sum_t V_k(t) + \sum_{\substack{t, s \\ t \neq s}} C_k(t, s) \Big).$$

En conditionnant par N(T), et en observant que la loi conditionnelle des t et s est celle de variables indépendantes dans la loi de densité $p(t)\,dt/p(T)$:

$$(12)\quad \mathrm{Var}(\hat{D} \mid k, S, N(T)) = \frac{1}{|T|^2} \left(\frac{N(T)}{p(T)} \int_T V_k(t)\,dt + \frac{N(T)(N(T)-1)}{p(T)^2} \iint_{T \times T} C_k(t, s)\,p(t)\,p(s)\,dt\,ds \right)$$

Le terme $\mathrm{Var}\,E(\hat{D} \mid k, T, N(T))$ est nul d'après (8).

Le déconditionnement par rapport à N(T) fait apparaître deux termes (variance de l'espérance et espérance de variance). Le premier vaut, d'après (8) encore,

$$(13)\qquad \mathrm{Var}(E(\hat{D} \mid T, k, N(T)) \mid T, k) = \frac{\bar{D}_k^2(T)}{k\,p(T)}$$

Le second, vu les moments d'ordre 1 et 2 de la loi de Poisson vaut :

$$(14)\quad \mathrm{E}(\mathrm{Var}(\hat{\mathrm{D}}\,|\,k,\,\mathrm{T},\,\mathrm{N}(\mathrm{T}))\,|\,k,\,\mathrm{T}) = \frac{1}{|\mathrm{T}|^2}\left(k\int_{\mathrm{T}} \mathrm{V}_k(t)\,dt + k^2 \iint_{\mathrm{T}\times\mathrm{T}} \mathrm{C}_k(t,\,s)\,p(t)\,p(s)\,dt\,ds \right)$$

Dans le cas particulier où $\bar{\mathrm{M}}_k(t)$ est un bruit blanc, on obtient, d'après (13) et (14)

$$(15)\quad \mathrm{Var}(\hat{\mathrm{D}}\,|\,k,\,\mathrm{T}) = \frac{\bar{p}(\mathrm{T})}{|\mathrm{T}|}\left(k\,\mathrm{V}_k + \frac{\mathrm{D}_k^2}{k} \right)$$

La forme générale de la variance conditionnelle à k et T de $\hat{\mathrm{D}}$ résulte de l'addition de (13) et (14) mais prend une forme un peu plus sympathique si on utilise les hypothèses (2) et (3) :

$$(16)\quad \mathrm{Var}(\hat{\mathrm{D}}\,|\,k,\,\mathrm{T}) = \frac{\mathrm{CV}^2}{k\,|\mathrm{T}|^2}\left(\int_{\mathrm{T}} \bar{\mathrm{D}}_k^2(t)\,p(t)\,dt + k\iint_{\mathrm{T}\times\mathrm{T}} \bar{\mathrm{D}}_k(t)\,\bar{\mathrm{D}}_k(s)\,r(t,\,s)\,p(t)\,p(s)\,dt\,ds \right) + \frac{1}{kp(\mathrm{T})}\bar{\mathrm{D}}_k^2(\mathrm{T})$$

Sous cette forme il est difficile d'aller plus loin. On pourra admettre que $\bar{\mathrm{D}}_k(t)$ varie lentement par rapport aux oscillations de p sur une période d'observation de longueur T. Techniquement, cela permet de « sortir » les termes en D des intégrales et d'aboutir à la formule approximative suivante :

$$(17)\quad \mathrm{Var}(\hat{\mathrm{D}}\,|\,k,\,\mathrm{T}) \simeq \frac{\bar{\mathrm{D}}_k^2(t)}{k}\cdot\frac{\bar{p}(\mathrm{T})}{|\mathrm{T}|}(1+\mathrm{CV}^2(1+k\bar{r}\,|\,\mathrm{T}\,|\,\bar{p}(\mathrm{T}))$$

où $\bar{r}$ désigne la valeur moyenne supposée indépendante de T de $r(t,\,s)$ sur $\mathrm{T}\times\mathrm{T}$. Dans cette formule $k\bar{p}(\mathrm{T})\,|\,\mathrm{T}|$ n'est autre que le nombre moyen d'achat $\bar{\mathrm{N}}(\mathrm{T})$ au cours de la période, de sorte qu'on peut écrire :

$$(18)\quad \mathrm{Var}(\hat{\mathrm{D}}\,|\,k,\,\mathrm{T}) = \frac{\bar{\mathrm{D}}_k^2(t)}{k}\,\frac{\bar{p}(\mathrm{T})}{|\mathrm{T}|}(1+\mathrm{CV}^2(1+\bar{r}\,\bar{\mathrm{N}}(\mathrm{T})).$$

Dans l'hypothèse où les montants des achats sont indépendants on a $\bar{r}=0$ et on trouve une formule analogue à (15) utilisant les hypothèses (2) et (3) à ceci près que les dépenses D_k dépendent du temps. L'hypothèse plausible sur $\bar{r}$ est une valeur faiblement négative : après un gros achat on aura peu de chance d'observer un achat de même nature d'un montant élevé. La

variance d'estimation de la dépense totale sera donc plus faible que dans la cas où on postule l'indépendance des montants des achats successifs. Dans la suite nous admettrons qu'on peut négliger ce terme.

Partant de (18) et en utilisant (10) avec la même approximation que dans (18), on obtient :

$$\text{(19)} \qquad \operatorname{Var}(\hat{\mathrm{D}} \mid \mathrm{T}) = \frac{\bar{p}(\mathrm{T})}{|\mathrm{T}|}(1+\mathrm{CV}^2)\,\mathrm{E}\frac{\bar{\mathrm{D}}_k^2(t)}{k} + \bar{p}(\mathrm{T})^2 \operatorname{Var}\bar{\mathrm{D}}_k(t).$$

La formule exacte lorsque $\bar{\mathrm{M}}_k(t)$ est un bruit blanc reste la même en gommant les « t » dans (19).

Pour obtenir la variance de $\hat{\mathrm{D}}$, on aurait besoin de faire une hypothèse sur la corrélation entre $\bar{\mathrm{D}}_k(t)$ (ou $\bar{\mathrm{D}}(t)$) et $p(t)$ vues comme des variables aléatoires dépendant de l'aléa t. Pour éviter des formules inextricables et arbitraires on se limitera au cas où $\bar{\mathrm{D}}_k(t) = \mathrm{D}_k$ est constante sur toute l'année. On a alors au lieu de (9) : $\mathrm{E}(\hat{\mathrm{D}} \mid \mathrm{T}, k) = \mathrm{D}_k \bar{p}(\mathrm{T})$, et

$$\operatorname{Var}(\hat{\mathrm{D}}) = \mathrm{E}\operatorname{Var}(\hat{\mathrm{D}} \mid \mathrm{T}) + \operatorname{Var}\mathrm{E}(\hat{\mathrm{D}} \mid \mathrm{T}).$$

Le caractère sans biais de l'estimation équivaut à $\mathrm{E}\bar{p}(\mathrm{T}) = 1$, de sorte qu'on obtient :

$$\text{(20)} \qquad \operatorname{Var}(\hat{\mathrm{D}}) = \operatorname{Var}(\mathrm{D}_k) + \frac{1+\mathrm{CV}^2}{|\mathrm{T}|}\mathrm{E}\left(\frac{\mathrm{D}_k^2}{k}\right) + \operatorname{Var}\bar{p}(\mathrm{T}).\,\mathrm{E}(\mathrm{D}_k^2).$$

Pour la suite il est utile d'avoir une expression de la variance de $\hat{\mathrm{D}}$ conditionnelle à k. Celle-ci s'obtient à partir de (18), (9) et de l'hypothèse de bruit blanc sur le processus $\bar{\mathrm{D}}_k(t)$. On obtient :

$$\text{(21)} \qquad \operatorname{Var}(\hat{\mathrm{D}} \mid k) = \mathrm{D}_k^2\left(\frac{1+\mathrm{CV}^2}{k|\mathrm{T}|} + \operatorname{Var}\bar{p}(\mathrm{T})\right)$$

Le déconditionnement de (21) redonne, bien entendu (20).

6 Commentaires et discussion

La formule (20) montre que la variance de $\hat{\mathrm{D}}$ est liée à la variabilité des dépenses elles-mêmes que ce soit directement (termes en Var D_k) ou à comportement d'achat fixé (terme en CV^2). Même un échantillonnage parfait assurant $\operatorname{Var}\bar{p}(\mathrm{T}) = 0$ et T grand pour annuler le second terme, nous laisserait l'égalité « naturelle » $\operatorname{Var}\hat{\mathrm{D}} = \operatorname{Var}\mathrm{D}_k$.

Supposons maintenant qu'on s'intéresse à une catégorie dont la dépense totale est très peu variable. Ceci revient à poser $\mathrm{D}_k \simeq \bar{\mathrm{D}}$ et $\operatorname{Var}\mathrm{D}_k \simeq 0$. On obtient la formule simplifiée :

$$\operatorname{Var}\hat{\mathrm{D}} \simeq \bar{\mathrm{D}}^2\left(\frac{1+\mathrm{CV}^2}{|\mathrm{T}|}\mathrm{E}\left(\frac{1}{k}\right) + \operatorname{Var}p(\mathrm{T})\right)$$

qui est analogue à (21) si on suppose de plus que k varie peu dans la population à laquelle on s'intéresse. Sous cette forme on constate que la fréquence des achats (k grand) diminue la variance d'estimation de $\bar{D}$, de même, bien entendu, qu'un allongement de la période d'observation (terme en $1/|T|$).

Ces remarques sont assez triviales. Elles montrent tout au plus que le modèle est qualitativement correctement formulé. On pourrait aller même jusqu'à une tentative de quantification de ces effets à partir de données d'enquêtes antérieures ou d'opinions *a priori*.

Notre objectif, cependant, est avant tout de guider le choix de la durée d'observation $|T|$. Les dates d'enquêtes sont supposées réparties uniformément sur l'année ce qui est nécessaire pour une enquête sur les budgets ou toutes les consommations sont relevées. Il n'en serait peut-être pas de même pour une enquête sur l'habillement, les cadeaux, ou la santé qui induisent des consommations beaucoup plus saisonnières.

Dans la formule (20) la loi de T n'a aucune incidence sur le premier terme, qui correspond à la variance limite obtenue par une observation sur une période de durée infinie. On étudiera les deux termes suivant conditionnellement à k, c'est-à-dire à partir de la formule (21).

On y trouve un premier terme facile à interpréter. La précision augmente avec le nombre moyen d'achats $k|T|$ par période d'observation.

Le dernier terme mérite un examen plus approfondi. Il est nul si p est constante tout au long de l'année, mais il l'est également si p est périodique (par exemple hebdomadaire) et que T est un multiple entier de la période. Dans ce cas, le gain de précision dû au passage d'une observation sur 10 jours à une observation sur 14 jours peut-être appréciable. On peut le constater sur l'exemple qui suit.

Supposons que p soit nulle du dimanche au jeudi et vaille une constante (3,5/365 exactement) le vendredi et le samedi. On peut modéliser ainsi de gros achats de fin de semaine (supermarché ou biens durables par exemple). Pour une durée d'observation de 10 jours, les dates de début d'observation suivant une loi uniforme, on voit facilement que dans 3 cas sur 7 (début d'observation le dimanche, le lundi ou le mardi) on n'a que 2 journées d'achat possibles, dans 2 cas sur 7 on en a 3 (périodes débutant le mercredi ou le samedi) et dans 2 cas sur 7 on en a 3 (périodes débutant le mercredi ou le samedi) et dans 2 cas sur 7 on en trouve 4. Avec $|T| = 10/365$ on évalue $\bar{p}(T)$:

$$\bar{p}(T) = \frac{365}{10} \times 2 \times \frac{3,5}{365} = 0,7 \quad \text{dans 3 cas sur 7}$$

$$= 1,05 \text{ dans 2 cas sur 7}$$

$$= 1,4 \quad \text{dans 2 cas sur 7}$$

La moyenne de $\bar{p}(T)$ vaut 1 par construction et comme prévu. On trouve que $\mathrm{Var}\,\bar{p}(T) \simeq 0,085$.

Si les achats sont concentrés sur une seule journée de la semaine le même calcul donne $\mathrm{Var}\,\bar{p}(T) = 0,099$. Pour évaluer l'importance relative de ce

terme, on peut essayer de chiffrer l'autre terme dans (21). Admettons une fréquence d'environ un achat par semaine et donc $k \simeq 50$ d'où $1/k \simeq 0{,}02$. Avec $|T| = 10/365$ et une valeur de CV raisonnable de l'ordre de 0,5, le premier terme vaut 0,913. Dans l'hypothèse d'un suivi d'observation sur 10 jours, la « composante périodique » $\mathrm{Var}\,\bar{p}(T)$ augmente la variance d'estimation d'environ 10%.

Si la précision de la collecte reste la même quand on porte la durée d'observation de 10 à 14 jours, le gain de précision sur l'estimation de la dépense totale (à k fixé) peut donc être de l'ordre de 35%, passant de 0,998 à 0,650 (à un facteur près). Sur ces 35 points, 25 peuvent être attribués à l'allongement de la période d'observation (et donc à un plus grand nombre d'achats) et 10 à la disparition de la composante périodique. Il s'agit malgré tout d'un cas assez extrême; l'utilisation de la formule complète (20) ou l'application en cas d'achats plus rares (k petit) minimiserait son importance. Concrètement, on obtiendra donc un gain de précision plus important pour des populations homogènes (à milieu social donné par exemple) et pour des achats à périodicité « très » hebdomadaire (la patisserie et le cinéma plus que le pain ou le tabac).

7 Un peu plus généralement

On peut aller un peu plus loin dans l'analyse du terme $\mathrm{Var}\,\bar{p}(T)$. Supposons qu'on ait une décomposition :

$$p(t) = p_0(t) + p_1(t)$$

où P_0 sera « lisse » et p_1 à périodicité courte (hebdomadaire pour fixer les idées. On admettra qu'une année compte exactement 52 semaines). On écrit les séries de Fourier de p_0 et p_1 (supposées rapidement convergentes) :

$$p_0(t) = 1 + \sum_{n=1}^{\infty} a_n \sin(2\pi nt + \alpha_n)$$

$$p_1(t) = \sum_{m=1}^{\infty} b_m \sin(2\pi \mathrm{N} mt + \beta_m)$$

avec $\mathrm{N} = 52$ pour une périodicité hebdomadaire. L'idée que p_0 est lisse conduit à l'hypothèse de coefficient de Fourier très petits à partir d'un certain rang. On approxime cette idée en posant $a_n = 0$ si $n \geqslant 50$. La série de Fourier de p est la somme des séries représentant p_0 et p_1. Si t est aléatoire et suit une loi uniforme sur (0, 1) il est immédiat (c'est en fait l'identité de Parseval) que :

$$\mathrm{Var}\,p(t) = \frac{1}{2}\left(\sum a_n^2 + \sum b_m^2\right)$$

La mise en moyenne mobile de p conduisant à $\bar{p}(T)$ se traduit par une série de Fourier modifiée (le paramètre t désigne ici le *milieu* de l'intervalle aléatoire T) :

$$\frac{1}{|T|}\int_{t-(|T|/2)}^{t+(|T|/2)} p(u)\,du = 1 + \frac{1}{|T|}\int_{t-(|T|/2)}^{t+(|T|/2)} \sum_{n=1}^{\infty} C_n \sin(\omega_n t + \gamma_n)$$

avec $\omega_n = 2\pi n$, $C_n = a_n$ si $n < 50$ et b_m si $n = Nm$ pour n multiple de $52 = N$, et $\gamma_n = \alpha_n$ ou β_m selon le cas.

On obtient :

$$\bar{p}(T) = 1 + \frac{2}{|T|}\sum_{n=1}^{\infty} \frac{C_n}{\omega_n} \sin\left(\frac{|T|}{2}\omega_n\right) \sin(\omega_n t + \gamma_n)$$

d'où :

$$\mathrm{Var}\,\bar{p}(T) = \frac{2}{|T|}\sum_{n=1}^{\infty} \frac{C_n^2}{(2\pi n)^2} \sin^2(|T|\pi n)$$

Sous cette forme on voit que :

- si $|T| = \dfrac{q}{N}$ (q entier : l'observation porte sur un nombre entier de semaines) et que $n = Nm$, les termes correspondant à la variation « haute fréquence » de p disparaissent;
- si $|T|$ est petit et que les a_n décroissent rapidement on peut écrire :

$$\sin^2(|T|\pi n) \simeq |T|^2 \pi^2 n^2.$$

Les termes correspondants sont donc sensiblement égaux a : $a_n^2/2$ c'est-à-dire identiques à ceux qui interviennent dans $\mathrm{Var}\,p(t)$. Dans les expressions (20) ou (21), la variance d'estimation due au caractère non homogène du processus d'achats (modélisé par $p(t)$) se décompose donc en une partie « basse fréquence » irréductible liée à la saisonnalité et une partie « haute fréquence » qu'on peut annuler par le choix d'une durée d'observation qui soit un multiple entier de la période fondamentale de ces hautes fréquences.

8 Conclusion

Les variations annuelles de l'intensité des achats sont un facteur d'imprécision dans les enquêtes de consommation par carnets de comptes. Idéalement on le supprime par une enquête réalisée sur toute l'année. Pratiquement, la collecte ne pouvant guère excéder un petit nombre de semaines, on obtient une estimation sans biais de la consommation en répartissant les dates de

début d'enquête uniformément sur toute l'année. On obtient une réduction parfois sensible de la variance d'estimation en fixant une durée de collecte à un nombre entier de semaine dans les cas ou la périodicité des achats est hebdomadaire.

● Références bibliographiques

AALEN, O. (1978). — « Non Parametric Inference for a Family of Counting Processes », *Annals of Statistics*, vol. 6, n° 4.

ANDERSEN, P. K. et BORGAN, O. (1985). — « Counting Processes Models for Life History Data: a Review », *Scandinavian Journal of Statistics*, vol. 12.

BARTLETT, R. F. (1986). — « Sampling with Trend and Correlation », *Revue Canadienne de Statistique*, vol. 14, n° 3.

DEVILLE, J.-C. (1984). — *Processus aléatoire du second ordre, Première partie*, ENSAE, Paris.

GILL, R. D. (1980). — *Censoring and Stochastic Integrals*, Mathematical Centre Tracts Amsterdam.

GLAUDE, M. (1983). — « L'importance des erreurs d'observation dans les enquêtes françaises sur les Budgets de Familles », *Communication à l'Institut International de Statistique*, Madrid.

GLAUDE, M. et MOUTARDIER, M. (1982). — « Les Budgets des Ménages », *Économie et Statistique*, n° 140.

HAUSER, J. R. et WISNIEWSKI, K. J. (1982). — « Dynamic Analysis of Consumer Response to Marketing Strategies », *Management Science*, vol. 28, n° 5.

HECKMAN, J. J. et SINGER, B. (1984). — « Econometric Duration Analysis », *Journal of Econometrics*, 24.

I.N.S.E.E. Division « Condition de Vie des Ménages » (1983-1984). — *Enquête Budgets de Famille* 1978-1979, Dossier d'enquête, t. 1, 1983, t. 2, 1984.

KOROLIUK, V. (1983). — *Aide mémoire de théorie des probabilités et de statistique mathématique*, Mir, Moscou.

SINGER, B. (1982). — « Aspects of Non-Stationarity », *Journal of Econometrics*, vol. 18.

SNYDER, P. L. (1976). — *Random Point Processes*, Wiley.