Article

Calcul d'un indice des loyers par post-stratification

RITSCHARD, Gilbert, DOZIO, Alessandro

Abstract

Cet article discute d'un certain nombre de difficultés théoriques et pratiques qui se présentent lorsqu'on calcule un indice des prix des loyers sur la base d'un échantillon post-stratifié.

Calcul d'un indice des loyers par post-stratification

GILBERT RITSCHARD et ALESSANDRO DOZIO*

1 INTRODUCTION

Dans l'activité pratique de la statistique appliquée on est souvent confronté à des contraintes qui rendent difficile l'utilisation de formulations théoriques optimales. Dans le cas d'une enquête visant à estimer l'évolution des prix de biens ou de services répartis en catégories, dans notre exemple il s'agira du prix des loyers, il peut s'avérer impossible de pré-stratifier l'échantillon. Tel est le cas lorsqu'on ne dispose pas d'un registre des logements qui permette de sélectionner des strates représentatives pour minimiser la taille de l'échantillon. Ou encore, des contraintes de nature institutionnelle peuvent interdire l'emploi de méthodes qui ne garantiraient pas la protection et la sécurité des données relevées conformément à la législation relative à la statistique fédérale. Il faut dès lors procéder à un tirage aléatoire et déterminer une stratification a posteriori de l'échantillon.

Cet article discute d'un certain nombre de difficultés théoriques et pratiques qui se présentent lorsqu'on calcule un indice des prix des loyers sur la base d'un échantillon post-stratifié.

En amont des problèmes liés à l'échantillonnage se pose évidemment le problème du choix d'un indice pertinent. S'agissant de loyers, deux approches sont a priori possibles. La première consiste à calculer le rapport entre le loyer moyen de la période courante et celui de la période de référence, ou de manière équivalente le rapport des dépenses allouées au loyer par l'ensemble de la population considérée. La seconde consiste à calculer la moyenne d'indices simples.

La première façon de procéder n'a de sens que si les logements considérés sont homogènes, par exemple du point de vue de la taille, du degré de vétusté ou encore de la localisation géographique. Elle est ainsi légitime pour évaluer l'indice des loyers d'une catégorie spécifique de logements. En tant qu'indice synthétique englobant aussi bien les variations de loyers des studios que ceux des grands appartements de luxe, le rapport de loyers moyens présente, en plus du problème conceptuel d'interprétation du «loyer moyen», l'inconvénient d'accorder une importance relative trop grande aux variations des loyers élevés.

* Adresser toute correspondance à GILBERT RITSCHARD, Département d'Econométrie, 102, bd Carl Vogt, CH-1211 Genève 4, e-mail: ritschar@uni2a.unige.ch.

Le second principe, qui permet une pondération indépendante du niveau du loyer, suppose par contre que l'on dispose de l'indice simple pour chaque logement. Ceci pose évidemment des problèmes eu égard au renouvellement de la population de référence. La méthode n'est en particulier pas applicable telle quelle si l'on envisage de se fonder sur des échantillons obtenus par rotation d'effectifs.

Pratiquement, la solution retenue consiste alors en un indice hybride, obtenu en considérant pour chaque catégorie un indice élémentaire sous forme de rapport des loyers moyens de la catégorie, et en prenant comme indice synthétique global la moyenne pondérée de ces indices élémentaires. Il nécessite naturellement une stratification de l'ensemble des logements. Un tel indice est généralement composé par l'intégration de plusieurs indices simples qui se réfèrent à des catégories spécifiques de logements comme, par exemple, le nombre de pièces, la date de construction ou de rénovation de l'immeuble ou l'unité géographique de localisation des logements.

L'article est consacré à l'estimation de cet indice de prix des loyers à partir d'un échantillon post-stratifié. Il s'agit essentiellement d'évaluer le biais et la dispersion de la version empirique de l'indice afin de pouvoir mieux juger de la significativité statistique des indices publiés. Il n'est pas inutile de rappeler ici que le niveau d'un indice des loyers est utilisé fréquemment dans la vie pratique, par exemple lors de séances de conciliation suite à des oppositions aux augmentations du loyer. Il s'agit donc d'une mesure qui peut avoir des conséquences réelles sur le budget d'un ménage et il importe qu'on en précise au mieux la significativité.

Dans la section 2 on établit les propriétés statistiques de l'indice empirique calculé sur la base d'un échantillon post-stratifié. On considérera donc le cas où la taille des sous-échantillons, et par conséquent des pondérations, sont aléatoires. Les résultats asymptotiques fournis permettront notamment d'établir des intervalles de confiance pour l'indice. La section 2.1 introduit les notations utilisées et donne quelques résultats relatifs à la distribution des tailles aléatoires des sous-échantillons définis par une stratification a posteriori. La section 2.2 étudie les propriétés statistiques de la moyenne d'un sous-échantillon d'une stratification a posteriori. La section 2.3 examine le rapport de deux moyennes. Finalement, la section 2.4 donne les propriétés de l'indice synthétique défini comme moyenne pondérée des rapports de moyennes.

Dans un second temps, nous présentons dans la section 3 des considérations d'ordre plus pragmatique en relation avec l'application des formules présentées en 2. Dans la pratique, il est important de pouvoir calculer facilement l'indice et son intervalle de confiance pour limiter les sources d'erreur dans le traitement des données et pour automatiser autant que possible les étapes du calcul. En 3.1 nous appliquons les formules établies à la section 2 aux données de la récente enquête semestrielle lausannoise sur les loyers. Des taux de croissance plus élevés sont simulés en 3.2. L'estimation de la covariance entre périodes induite par les observations appariées des échantillons constitue la partie la plus fastidieuse du traitement. Ainsi, au point 3.3, nous évaluons les conséquences de la non prise en compte de cette covariance dans les estimations du biais et de la variance de l'indice.

2 PROPRIETES STATISTIQUES DE L'INDICE

On désigne par m la taille de la population et, pour chaque strate h de la population, $h=1,2,...,c$, on note m_h sa taille, μ_h sa moyenne, σ_h^2 sa variance, et $p_h = m_h / m$ la probabilité que la $i$ème observation X_i appartienne à cette $h$ème strate. Par ailleurs, on note $\sigma_h^{*2} = (m_h / (m_h - 1))\, \sigma_h^2$ la variance modifiée qui permet d'alléger la présentation de nombreux résultats de la théorie des sondages (cf. notamment COCHRAN, 1977). Nous considérons essentiellement des tirages sans remises et, en désignant par n la taille de l'échantillon tiré, notons $f^{pc} = (m - n) / (m - 1)$ et $f^{pc*} = (m - n) / m$ les facteurs de correction utilisés respectivement avec σ^2 *et* σ^{*2}. Les indices que l'on se propose d'estimer s'écrivent formellement:

$$\text{Indice pour une strate } h\text{: } r_{h(t/0)} = \frac{\mu_{ht}}{\mu_{h0}} \tag{1}$$

$$\text{Indice synthétique global: } I_{t/0} = \sum_{h=1}^{c} p_h \frac{\mu_{ht}}{\mu_{h0}} \tag{2}$$

Des estimateurs naturels de ces indices sont donnés par leurs versions empiriques obtenues en remplaçant les moyennes μ_{ht} et μ_{h0} par les moyennes d'échantillon et les poids p_h par les proportions au sein de l'échantillon. Afin d'établir les propriétés statistiques de ces estimateurs, nous discutons successivement ci-après de la taille aléatoire des sous-échantillons résultant d'une stratification a posteriori, de la moyenne d'un échantillon de taille aléatoire, du rapport de deux moyennes et enfin de l'estimateur de l'indice synthétique global. Les résultats sont résumés sous la forme d'un lemme et de trois théorèmes dont les démonstrations sont données en annexe.

2.1 Tailles aléatoires lors de stratification a posteriori

Pour procéder à une stratification a posteriori on considère à chaque tirage i une variable X_i, par exemple le loyer payé par la $i$ème personne interrogée, et une variable de contrôle K_i, telle que, par exemple, le nombre de pièces du logement. On dispose donc d'un échantillon (X_1, K_1), (X_2, K_2), ..., (X_n, K_n). La stratification a posteriori consiste alors à partitionner l'ensemble des X_i, $i = 1,2,...,n$, en c classes selon les modalités que peuvent prendre les K_i. Formellement, cette partition, notée $\{\mathbf{E}_1, \mathbf{E}_2, ..., \mathbf{E}_c\}$, est définie comme suit:

$$\mathbf{E}_1 = \{ X_i \mid K_i \in \mathbf{K}_1 \}$$
$$\mathbf{E}_2 = \{ X_i \mid K_i \in \mathbf{K}_2 \}$$
$$\dots$$
$$\mathbf{E}_c = \{ X_i \mid K_i \in \mathbf{K}_c \}$$

où $\{\mathbf{K}_1, \mathbf{K}_2, \ldots, \mathbf{K}_c\}$ est une partition fixée a priori, donc non aléatoire, des modalités des K_i. Notons que la variable de contrôle K_i peut être la variable X_i elle-même, ou une autre variable. Elle ne doit pas nécessairement être métrique.

Les tailles n_h des sous-échantillons sont ici aléatoires, non indépendantes, vérifiant notamment la contrainte $\Sigma_h n_h = n$, où n est la taille fixée (non aléatoire) de l'échantillon. Individuellement, pour des tirages sans remises, chaque n_h suit une loi hypergéométrique de paramètres n et p_h. Le lemme suivant résume les propriétés statistiques des n_h.

Lemme 1: Soit un échantillon de taille n obtenu par tirages sans remises. Les tailles n_h, $h = 1, 2, \ldots, c$, des sous-échantillons résultant d'une stratification a posteriori vérifient les propriétés suivantes

$$\mathrm{E}(n_h) = np_h \tag{3}$$

$$\mathrm{Var}(n_h) = f^{pc} np_h (1 - p_h) \tag{4}$$

$$\mathrm{Cov}(n_h, n_s) = -f^{pc} np_h p_s, \quad h \neq s \tag{5}$$

$$\mathrm{E}\left(\frac{1}{n_h}\right) = \frac{1}{np_h} + f^{pc} \frac{(1 - p_h)}{n^2 p_h^2} \tag{6}$$

2.2 *Variance de la moyenne de sous-échantillons de tailles aléatoires*

On se propose à présent d'étudier la variance de la moyenne $\overline{X}_h$ des X_i d'un sous-échantillon $\mathbf{E}_h$ d'une stratification a posteriori.

Théorème 1: La variance $\mathrm{Var}\,(\overline{X}_h)$ *de la moyenne* $\overline{X}_h$ *des* X_i *d'un sous-échantillon obtenu par stratification a posteriori est*

$$\mathrm{Var}\,(\overline{X}_h) = \sigma_h^{*2}\left(\frac{1}{np_h} - \frac{1}{mp_h}\right) + \sigma_h^{*2} f^{pc}\left(\frac{1 - p_h}{n^2 p_h^2}\right) \tag{7}$$

Le premier terme de (7) correspond à la variance de $\overline{X}_h$ pour une taille certaine $n_h = np_h$. Le second terme représente ainsi l'accroissement de la variance due à l'incertitude quant à la taille aléatoire n_h. On notera que ce terme est en $1 / n^2$, et qu'il devient donc rapidement négligeable lorsque n croît. De même, on note qu'il est d'autant plus grand que p_h est petit. Il convient donc de lui prêter en particulier attention lorsque les strates comprennent une petite proportion de la population, soit notamment lorsqu'on a un grand nombre de strates.

Le tableau 1 illustre l'évolution de la variance de $\overline{X}_h$ et de ses composantes pour des échantillons tirés dans une population de taille $m = 100'000$.

Sur le plan pratique, il s'agira évidemment d'estimer $\mathrm{Var}(\overline{X}_h)$. On remplacera pour cela les paramètres inconnus par leurs estimations:

paramètre:	*estimation:*
p_h	$\hat{p}_h = n_h / n$
m_h	$\hat{m}_h = m\hat{p}_h$
σ_h^{*2}	$\hat{\sigma}_h^{*2} = \frac{\hat{m}_h}{\hat{m}_h - 1} \frac{1}{n_h - 1} \sum_{i=1}^{n_h} (x_{ih} - \overline{x}_h)^2$

Notons que l'estimation ne peut évidemment être calculée que si $n_h \geq 2$. Si l'on obtient un n_h nul ou égal à un, il conviendra soit de procéder à un regroupement de strates, soit d'exclure la strate correspondante de l'analyse.

Tableau 1. Evolution de *Var* $(\overline{X}_h)$ pour $\sigma^{*2} = 1$ et $m = 100'000$

n	p_h	np_h	$m_h = mp_h$	$\frac{1}{np_h} - \frac{1}{m_h}$	% (5 / 9)	$\frac{(m-n)(1-p_h)}{(m-1)\,n^2\,p_h^2}$	% (7 / 9)	$\mathrm{Var}(\overline{X}_h)$	% (6 + 8)
1	2	3	4	5	6	7	8	9	10
50	.05	2.5	5,000	.3998	72.5	.1519	27.5	.5521	100.0
	.10	5	10,000	.1999	84.8	.0360	15.2	.2361	100.0
	.20	10	20,000	.1000	92.6	.0080	7.4	.1080	100.0
	.50	25	50,000	.0400	98.0	.0008	2.0	.0408	100.0
	.80	40	80,000	.0250	99.5	.0001	0.5	.0251	100.0
100	.05	5	5,000	.1998	84.1	.0380	15.9	.2382	100.0
	.10	10	10,000	.0999	91.8	.0090	8.2	.1091	100.0
	.20	20	20,000	.0500	96.2	.0020	3.8	.0520	100.0
	.50	50	50,000	.0200	99.0	.0002	1.0	.0202	100.0
	.80	80	80,000	.0125	99.8	.0000	0.2	.0125	100.0
500	.05	25	5,000	.0398	96.4	.0015	3.6	.0417	100.0
	.10	50	10,000	.0199	98.2	.0004	1.8	.0205	100.0
	.20	100	20,000	.0100	99.2	.0001	0.8	.0101	100.0
	.50	250	50,000	.0040	99.8	.0000	0.2	.0040	100.0
	.80	400	80,000	.0025	100.0	.0000	0.0	.0025	100.0

2.3 Rapport de deux moyennes

Cette section présente les propriétés statistiques du rapport de deux moyennes d'échantillon

$$g(\bar{X}, \bar{Y}) = \frac{\bar{X}}{\bar{Y}}$$

Les résultats donnés par le théorème ci-dessous s'appliquent notamment à l'estimateur $\hat{r}_{h(t0)} = X_{ht}/X_{h0}$ de l'indice défini en (1) pour une strate.

Théorème 2: Soit $\bar{X}$ et $\bar{Y}$ les moyennes de deux échantillons et $r = \mu_x/\mu_y$ le rapport de leurs espérances mathématiques. Le rapport $\bar{X}/\bar{Y}$ possède, si on se limite aux termes d'ordre 1 / n, les propriétés suivantes

$$\mathrm{E}\left(\frac{\bar{X}}{\bar{Y}}\right) = r + r\left(\frac{\mathrm{Var}(\bar{Y})}{\mu_y^2} - \frac{\mathrm{Cov}(\bar{Y}, \bar{X})}{\mu_x \mu_y}\right) \tag{8}$$

$$\mathrm{Var}\left(\frac{\bar{X}}{\bar{Y}}\right) = \frac{1}{\mu_y^2}\left(\mathrm{Var}(\bar{X}) - 2r\mathrm{Cov}(\bar{X}, \bar{Y}) + r^2\,\mathrm{Var}(\bar{Y})\right) \tag{9}$$

$$= \frac{1}{\mu_y^2}\mathrm{E}\left((\bar{X} - \frac{\mu_x}{\mu_y}\bar{Y})^2\right) \tag{10}$$

Le second terme de (8) représente le biais du rapport $\bar{X}/\bar{Y}$ en tant qu'estimateur de $r = \mu_x/\mu_y$. On peut relever en particulier que, dans le cas où μ_x et μ_y sont tous deux positifs et que $\bar{X}$ et $\bar{Y}$ sont positivement corrélés, la dispersion du dénominateur introduit une sur-estimation du rapport des moyennes des populations, tandis que l'accroissement de la corrélation tend plutôt à une sous-estimation.

On peut songer à corriger le biais en multipliant les estimations par le facteur $f^{biais} = (1 + \mathrm{Var}(\bar{Y})/\mu_y^2 - \mathrm{Cov}(\bar{X}, \bar{Y})/\mu_x\mu_y)^{-1}$. Pratiquement, comme μ_x, μ_y, $\mathrm{Var}(\bar{Y})$ et $\mathrm{Cov}(\bar{X}, \bar{Y})$ sont inconnus, on doit les remplacer par leurs estimations pour obtenir une estimation du biais. Dans le cas où le numérateur et le dénominateur sont des moyennes d'échantillons appariés de taille n, on utilise par exemple les estimations non biaisées classiques

$$\hat{\mathrm{Var}}(\bar{Y}) = f^{pc}\frac{1}{n(n-1)}\sum_{i=1}^{n}(Y_i - \bar{Y})^2 \tag{11}$$

$$\hat{\mathrm{Cov}}(\bar{X}, \bar{Y}) = f^{pc} \frac{1}{n\,(n-1)} \sum_{I=1}^{n} (X_i - \bar{X})\,(Y_i - \bar{Y}) \tag{12}$$

Pour des échantillons non appariés indépendants, la covariance est nulle. L'estimation de Var($\bar{Y}$) s'obtient en remplaçant dans l'expression ci-dessus n par n_y, la taille de l'échantillon sur les Y.

Dans le cas d'échantillons choisis par rotation d'un certain pourcentage $(1 - \lambda)$ des individus, seul un sous-ensemble des observations restent appariées. On estime dans ce cas la covariance Cov($\bar{X}$, $\bar{Y}$) à partir des $n_a = \lambda n$ observations appariées avec l'estimateur:

$$\hat{\mathrm{Cov}}(\bar{X}, \bar{Y}) = \lambda^2 f_a^{pc} \frac{1}{n_a\,(n_a - 1)} \sum_{i=1}^{n_a} (X_i - \bar{X}_a)\,(Y_i - \bar{Y}_a) \tag{13}$$

où les moyennes $\bar{X}_a$ et $\bar{Y}_a$, ainsi que le facteur de correction f_a^{pc} portent sur les données appariées.

Remarquons cependant que tant la variance de $\bar{Y}$ que la covariance entre $\bar{X}$ et $\bar{Y}$ sont en $1 / n$ et diminuent donc avec la taille de l'échantillon. Pour de grands échantillons le biais devient négligeable.

De même, pour évaluer la variance, on remplace les moyennes μ_x et μ_y par les moyennes observées d'échantillon $\bar{x}$ et $\bar{y}$, et, dans le cas de données appariées le terme, $\mathrm{E}((\bar{X} - \frac{\mu_x}{\mu_y} \bar{Y})^2)$ par son estimation

$$f^{pc} \frac{1}{n(n-1)} \sum_{i=1}^{n} \left(x_i - \frac{\bar{x}}{\bar{y}} y_i \right)^2 \tag{14}$$

Pour des échantillons indépendants, le terme précédent ne peut pas être calculé. On utilise alors l'expression (9), dans laquelle Cov($\bar{X}$, $\bar{Y}$) est nulle en vertu de l'indépendance, et l'on remplace les variances par leurs estimations.

A titre d'exemple, on a généré deux ensembles de $m = 100$ valeurs choisies aléatoirement selon une loi normale $N(2000, 150^2)$ pour le premier, et $N(1000, 100^2)$ pour le second. Les caractéristiques des deux populations ainsi constituées sont:

$$\mu_x = 2003.70 \quad \mu_y = 986.45$$
$$\sigma_x = 140.010 \quad \sigma_y = 92.959 \quad \sigma_{xy} = 768.28$$

$$r = \mu_x / \mu_y = 2.031$$
$$\sigma_x^2 - 2(\mu_x / \mu_y)\sigma_{xy} + (\mu_x / \mu_y)^2 \sigma_y^2 = 52'134$$
$$\sigma_x^2 + (\mu_x / \mu_y)^2 \sigma_y^2 = 55'813.22$$

Un échantillon de $n = 15$ valeurs a ensuite été tiré dans chacun de ces deux ensembles. Les valeurs d'échantillon sont

$$\begin{array}{rrrrrr} \bar{x} = 1971.44 & & \bar{y} = 947.28 & & & \\ \hat{\sigma}_x = 171.971 & & \hat{\sigma}_y = 76.530 & & \hat{\sigma}_{xy} = & 5392.51 \\ \mathrm{Var}(X) = 1122.04 & & \mathrm{Var}(Y) = 494.62 & \mathrm{Cov}(X,Y) = & & 43.98 \\ \hat{\mathrm{V}}\mathrm{ar}(X) = 1692.78 & & \hat{\mathrm{V}}\mathrm{ar}(Y) = 335.24 & \hat{\mathrm{C}}\mathrm{ov}(X,Y) = & & 308.66 \end{array}$$

et l'on trouve

$$\begin{aligned} \hat{r} = \bar{x}/\bar{y} &= 2.081 \\ \hat{\sigma}_x^2 - 2(\bar{x}/\bar{y})\hat{\sigma}_{xy} + (\bar{x}/\bar{y})^2 \hat{\sigma}_y^2 &= 32'495.5 \\ \hat{\sigma}_x^2 + (\bar{x}/\bar{y})^2 \hat{\sigma}_y^2 &= 54'940.8 \end{aligned}$$

En considérant les échantillons comme appariés, le facteur multiplicatif permettant de corriger le biais vaut

$$f^{biais} = \left(1 + \frac{494.62}{(986.45)^2} - \frac{43.98}{2003.7 \cdot 986.45}\right)^{-1} = 0.999514$$

et son estimation

$$\hat{f}^{biais} = \left(1 + \frac{335.24}{(986.45)^2} - \frac{308.66}{1971.44 \cdot 947.28}\right)^{-1} = 0.999792$$

Ces coefficients sont très proches de un. Le biais de $\hat{r} = \bar{X}/\bar{Y}$ en tant qu'estimateur de $r = \mu_x/\mu_y$ représente ainsi une quantité négligeable. Le calcul donne

$$\begin{aligned} \mathrm{Biais}(\hat{r}) &= 0.00099 \\ \hat{\mathrm{B}}\mathrm{iais}(\hat{r}) &= 0.00043 \end{aligned}$$

Si l'on admet l'indépendance des échantillons, c'est-à-dire si l'on ne tient pas compte de la covariance entre $\bar{X}$ et $\bar{Y}$, ces deux valeurs sont un peu plus grandes, mais restent très petites. On trouve respectivement $\mathrm{Biais}(\hat{r}) = 0.00104$ et $\hat{\mathrm{B}}\mathrm{iais}(\hat{r}) = 0.00077$. On peut relever que l'on obtient ces valeurs négligeables malgré la taille modeste ($n = 15$) des échantillons retenus.

La variance et l'écart-type de $\hat{r}$ sont

$$\begin{aligned} \mathrm{Var}(\hat{r}) &= \left(\frac{(100-15)}{99 \cdot 15} \frac{52'134}{(986.45)^2}\right) = 0.003070 \\ \sigma_{\hat{r}} &= 0.0554 \end{aligned}$$

et les estimations

$$\hat{\mathrm{V}}\mathrm{ar}(\hat{r}) = \left(\frac{(100-15)}{99 \cdot 15}\,\frac{32'495.5}{(947.28)^2}\right) = 0.002073$$
$$\hat{\sigma}_{\hat{r}} = 0.0455$$

L'erreur standard est donc ici, avec des échantillons de taille 15, de l'ordre de 0.05 / 2 = 2.5% de la valeur du ratio estimé.

2.4 Indice avec pondérations a posteriori

On considère à nouveau la partition a posteriori des données échantillonnées, et l'on s'intéresse à l'estimateur $\hat{I}_{t/0}$ de l'indice synthétique (2). On retient comme estimateur la moyenne des indices des strates pondérés selon les tailles $n_h = n_{h0}$, où n_{h0} représente la taille (aléatoire) du $h$ème sous-groupe échantillonné en 0, c'est-à-dire à la période de référence. Formellement, on a donc

$$\hat{I}_{t/0} = \frac{1}{n}\sum_h n_h \frac{\bar{X}_{ht}}{\bar{X}_{h0}} \tag{15}$$

Théorème 3: L'indice $\hat{I}_{t/0} = \sum_h (n_h/n)\,(\bar{X}_{ht}/\bar{X}_{h0})$, avec pondérations définies selon la stratification a posteriori de la période de référence 0, possède les propriétés statistiques

$$\mathrm{E}(\hat{I}_{t/0}) = \sum_h p_h r_h + f^{pc*}\frac{1}{n}\sum_h r_h\, b_h^* \tag{16}$$

$$\mathrm{Var}(\hat{I}_{t/0}) = f^{pc}\frac{1}{n}\left(\sum_h r_h^{*2}\, p_h\,(1-p_h) - \sum_h\sum_{s\neq h} r_h^*\, r_s^*\, p_h\, p_s\right) + \frac{1}{n}\sum_{h=1}^{c} p_h\,\beta_h^*\left(1 - \frac{n + f^{pc}\,(1-p_h)/p_h}{m}\right) \tag{17}$$

où

r_h = μ_{ht}/μ_{h0}
b_h^* = $\sigma_{h0}^{*2}/\mu_{h0}^2 - \sigma_{h0t}^*/(\mu_{ht}/\mu_{h0})$
r_h^* = $r_h\,(1 - b_h^*/m_h)$
β_h^* = $(\sigma_{ht}^{*2} - 2r_h\sigma_{h0t}^* + r_h^2\,\sigma_{h0}^{*2})/\mu_{h0}^2$

Le premier terme de $E(\hat{I}_{t/0})$ n'est rien d'autre que l'indice $I_{t/0}$. Le second terme donne donc le biais de l'estimateur. Il correspond à $\sum_h E_{n_h} (n_h / n)\text{Biais}(\hat{r}_h \mid n_h)$, c'est-à-dire à l'espérance de la moyenne pondérée des biais des ratios empiriques $\hat{r}_h = \bar{X}_{ht} / \bar{X}_{h0}$.

En ce qui concerne la variance, le premier terme reflète la part de variance due à l'incertitude sur les poids n_h / n, et le second la part due à l'incertitude sur les indices des strates.

Tableau 2. Strates et taux de croissance théoriques

h	strate 1	strate 2	strate 3	strate 4	
$k = x_0$	$(-\infty, 800)$	[800, 1000)	[1000, 1200)	[1200, ∞)	total
m_h	13	35	31	21	100
α_h	1.1	1.6	1.2	2	

A titre d'exemple, nous avons généré un ensemble de 100 valeurs x_0 selon une loi $N(1000, 200^2)$. Celles-ci ont été stratifiées en prenant comme variable de contrôle la variable $k = x_0$. Les strates, et la répartition des 100 valeurs obtenues selon ces strates sont décrites au tableau 2. A partir de ces données, nous avons engendré une série de 100 valeurs x_t avec un modèle de la forme $x_{iht} = \alpha_h x_{ih0} + u_{ih}$, où les u_{ih} sont i.i.d. $N(0, 50^2)$, et les α_h sont des coefficients précisés dans le tableau 2 pour chaque strate h. On dispose ainsi de $m = 100$ couples (x_{ih0}, x_{iht}) dont le tableau 3 résume les caractéristiques.

Tableau 3. Caractéristiques de la population ($m = 100$)

	strate 1	strate 2	strate 3	strate 4
p_h	0.13	0.35	0.31	0.21
μ_{ht}	747.241	1'431.193	1'307.375	2'676.409
σ^*_{ht}	85.364	98.289	76.118	291.530
μ_{h0}	683.752	896.504	1'093.878	1'327.762
σ^*_{h0}	73.587	58.799	56.967	147.986
σ^*_{h0t}	4'979.36	5'124.76	3'099.07	42'508.14
r_h	1.0929	1.5694	1.1952	2.0157
r^*_h	1.0927	1.5694	1.1952	2.0157
b^*_h	0.0018	0.0003	0.0005	0.0005
β^*_h	0.0061	0.0026	0.0025	0.0015
$1 - (n + f^{pc}(1 - p_h) / p_h) / m$	0.6993	0.7359	0.7331	0.7215

Un échantillon de $n = 25$ données appariées a ensuite été tiré au hasard dans cette population. Le tableau 4 en résume les caractéristiques.

Tableau 4. Caractéristiques de l'échantillon ($n = 25$)

	strate 1	strate 2	strate 3	strate 4
n_h	3	5	10	7
$\hat{p}_h$	0.12	0.2	0.4	0.28
$\hat{\mu}_{ht}$	770.73	1'366.04	1'297.31	2'574.97
$\hat{\sigma}^*_{ht}$	119.202	115.198	78.884	162.432
$\hat{\mu}_{h0}$	713.19	877.80	1'068.26	1'274.40
$\hat{\sigma}^*_{h0}$	112.454	79.214	55.045	58.482
$\hat{\sigma}^*_{h0t}$	13'309.48	8'121.62	3'878.32	9'351.76
$\hat{r}_h$	1.0807	1.5562	1.2144	2.0205
$\hat{r}^*_h$	1.0806	1.5561	1.2144	2.0206
$\sigma_{\hat{r}_h}$	0.0397	0.0212	0.0131	0.0119
$\hat{\sigma}_{\hat{r}_h}$	0.0103	0.0266	0.0087	0.0122
$\hat{b}^*_h$	0.0006	0.0014	0.0001	0.0007
$\hat{\beta}^*_h$	0.0004	0.0041	0.0011	0.0016
$1-(n+f^{pc}(1-\hat{p}_h)/\hat{p}_h)/m$	0.6944	0.7197	0.7386	0.7305

Avec les indications données dans les tableaux 3 et 4 on détermine sans peine la valeur de $I_{t/0}$ et de son estimation

$$I_{t/0} = 1.4946$$
$$\hat{I}_{t/0} = 1.4924$$

ainsi que le biais et la variance de l'estimateur $\hat{I}_{t/0}$

$$\text{Biais } (\hat{I}_{t/0}) = 0.00012$$
$$\text{Var } (\hat{I}_{t/0}) = 0.003314 + 0.000081 = 0.003395$$
$$\sigma_{\hat{I}_{t/0}} = 0.0583$$

Le biais représente environ 0.2% de l'écart-type. On note par ailleurs que la valeur estimée diffère de moins du dixième d'un écart-type tant de l'espérance mathématique $E(\hat{I}_{t/0}) = 1.4947$, que de la vraie valeur de l'indice.

On obtient une estimation du biais et de l'écart-type de l'estimateur utilisé en remplaçant dans (16) et (17) les paramètres p_h, μ_{h0}, μ_{ht}, σ^*_{h0}, σ^*_{ht} et σ^*_{h0t} par leurs estimations. On trouve

$$\hat{\text{Biais}}\,(\hat{I}_{t/0}) = 0.0003$$
$$\hat{\text{Var}}\,(\hat{I}_{t/0}) = 0.003944 + 0.000051 = 0.003995$$
$$\hat{\sigma}_{\hat{I}_{t/0}} = 0.0632$$

Ces estimations restent proches des vraies valeurs calculées précédemment en exploitant les informations sur l'ensemble de la population.

3 APPLICATION AUX DONNEES LAUSANNOISES

Pour illustrer l'application des formules présentées dans la première partie nous avons utilisé les premiers résultats de l'enquête semestrielle lausannoise sur les loyers. L'enquête se base sur un échantillon rotatif aléatoire d'environ 4'500 logements relevé aux mois de mai et de novembre de 1993. Cette nouvelle enquête fait suite à la dernière révision de l'indice suisse des prix à la consommation (IPC). Auparavant, c'était l'Office fédéral de la statistique qui procédait à la constitution d'un très important échantillon, de l'ordre de quelques 120'000 appartements, dans le cadre du calcul de l'IPC. Depuis la dernière révision, base 100 = mai 1993, l'échantillon de la confédération a été considérablement réduit et ne permet plus de calculer un indice régionalisé des prix des loyers. Les collectivités publiques des cantons ou des grandes villes qui souhaitent continuer le calcul d'un tel indice doivent actuellement procéder à leur propre enquête, comme c'est le cas pour la ville de Lausanne.

3.1 Calcul de l'indice et de ses caractéristiques

Une stratification a posteriori est définie en fonction du nombre de pièces (de 1 à 5 et plus), de la date de construction ou de rénovation de l'immeuble (plus de 20 ans d'âge; entre 11 et 20; et 10 ans ou moins) et de la catégorie de logement, soit à loyer libre ou subventionné. Notons que les logements à loyer subventionné ne représentent qu'environ le 13% du total de l'échantillon. Pour les logements subventionnés de 1 ou de 5 pièces le nombre d'observations n'est en particulier pas suffisant pour pratiquer la distinction selon la date de construction ou de rénovation. Nous avons pour ces deux cas regroupé les logements indépendamment de leur âge. On retient ainsi une stratification en 26 classes, soit 11 pour les loyers subventionnés et 15 pour les loyers libres. Le tableau 6 reproduit les principales caractéristiques de l'échantillon.

Tableau 5. Indice des loyers de novembre 1993

	$\hat{I}_{t/0}$	$\hat{\sigma}_{\hat{I}_{t/0}}$	$\hat{B}$iais	n
Loyers subventionnés	100.9293	.9662	.0678	592
Loyers libres	99.9254	.3836	.0122	3'939
Ensemble	100.0566	.3566	.0195	4'531

Le tableau 5 donne, pour les loyers subventionnés, les loyers libres et l'ensemble des logements, les estimations de l'indice synthétique $I_{t/0}$ du mois de novembre 1993, calculées selon la formule (15) avec mai 1993 comme base 100. Le tableau donne également les estimations du biais et de l'écart-type de l'estimateur calculées à partir des expressions (16) et (17). Les biais et écarts-types rapportés sont exprimés dans les mêmes unités que l'indice, soit multipliés par 100. Notons que l'indice global, calculé sur la base des 26 strates, s'exprime comme une moyenne des indices des loyers subventionnés et des loyers libres pondérés selon leurs effectifs respectifs. En calculant l'indice global sans distinguer entre loyers subventionnés et libres, c'est-à-dire en ne considérant que 15 strates, cette relation n'est plus respectée. Pour les mêmes données, on obtient par exemple une valeur de l'indice inférieure à celle de l'indice des loyers libres, et donc non cohérente avec la désagrégation entre loyers libres et subventionnés.

On peut relever que le biais (estimé) est très faible. Il représente dans tous les cas moins du dixième de l'écart-type. L'intervalle de confiance à 95% pour l'indice synthétique (tableau 7) indique clairement que les variations de loyers observées ne sont pas statistiquement significatives. Les loyers sont restés stables, tant pour les loyers subventionnés que pour les loyers libres, et donc aussi pour l'ensemble.

Bien que l'augmentation des loyers subventionnés ne soit pas statistiquement significative, elle semble correspondre aux répercussions des fortes hausses du début des années '90 qui ont été, pour cette catégorie de logement, décalées dans le temps.

Notons que l'écart-type et le biais sont évidemment plus élevés pour la catégorie des loyers subventionnés en raison de l'effectif plus réduit de l'échantillon. C'est l'effet du facteur $1 / n$ dans les expressions (16) et (17). A titre indicatif, on peut mentionner encore que la part de variance des indices simples due à l'incertitude sur les poids des strates est respectivement de 0.37% pour les loyers subventionnés, de 0.25% pour les loyers libres, et de 0.29% pour l'indice global.

Tableau 6. Echantillon poststratifié de logements lausannois, (base 0 = mai 1993, t = novembre 1993)

h	strate pièces	âge	m_h	n_h	$\hat{p}_h$	$\hat{\mu}_{h0}$	$\hat{\mu}_{ht}$	$\hat{r}_h$
loyers subventionnés			6'438	592	.131			
1	1	tous	874	60	.013	314	316	1.008
2	2	plus de 20	1'135	121	.027	346	355	1.025
3		11-20	427	8	.002	612	618	1.009
4		10 et moins	286	30	.007	426	436	1.023
5	3	plus de 20	1'258	169	.037	536	539	1.005
6		11-20	458	9	.002	1'001	982	0.981
7		10 et moins	812	70	.015	674	675	1.002
8	4	plus de 20	236	57	.013	797	781	0.980
9		11-20	294	8	.002	1'074	1'047	0.974
10		10 et moins	563	42	.009	1'028	1'054	1.025
11	5	tous	95	18	.004	1'179	1'225	1.039
loyers libres			50'962	3'939	.869			
12	1	plus de 20	4'523	377	.083	554	550	0.992
13		11-20	1'685	30	.007	609	624	1.025
14		10 et moins	2'444	182	.040	627	619	0.987
15	2	plus de 20	7'832	687	.152	699	693	0.992
16		11-20	2'403	148	.033	732	736	1.005
17		10 et moins	5'437	431	.095	853	841	0.986
18	3	plus de 20	8'151	727	.160	871	885	1.016
19		11-20	3'071	177	.039	958	964	1.006
20		10 et moins	5'864	426	.094	1'047	1'043	0.997
21	4	plus de 20	2'834	253	.056	1'036	1'050	1.014
22		11-20	1'408	81	.018	1'258	1'296	1.030
23		10 et moins	2'701	208	.046	1'434	1'405	0.980
24	5	plus de 20	1'066	99	.022	1'386	1'395	1.006
25		11-20	539	40	.009	1'737	1'755	1.011
26		10 et moins	1'004	73	.016	1'802	1'756	0.975
ensemble			57'400	4'531	1			

Tableau 7. Intervalles de confiance à 95%

	borne inférieure	centre	borne supérieure
Loyers subventionnés	98.97	100.86	102.76
Loyers libres	99.16	99.91	100.67
Ensemble	99.34	100.04	100.74

3.2 Simulation de taux de croissance plus élevés

Les loyers n'ayant pas évolué significativement durant cette période de stabilité des taux hypothécaires, nous avons procédé à une simulation pour illustrer le comportement de l'indice en présence de taux de croissance plus élevés. La simulation se limite à la catégorie des loyers libres.

Deux échantillons avec des taux de croissance α de 70% et de 240% ont été générés selon la forme $x_{iht} = \alpha x_{ih0} + u_i$. Les taux choisis n'ont évidemment qu'un but illustratif et ne prétendent pas illustrer une quelconque réalité historique. Pour respecter la rotation de l'échantillon, on a retenu pour x_{ih0} les loyers de mai 93 pour les données appariées, et les loyers de novembre 93 pour les données renouvelées. Le taux de croissance est identique pour chaque strate mais les u_i sont assignés différemment selon le niveau du loyer de la période de base. La simulation avec $\alpha = 70\%$ se base sur les hypothèses suivantes. Pour les loyers de moins de fr. 500.–, les u_i sont indépendamment et identiquement distribués selon une loi normale de moyenne 0 et d'écart-type fr. 50.–. Pour les loyers entre fr. 500.– et fr. 1'000.–, les u_i suivent une loi $N(0,100^2)$; entre fr. 1'000.– et 1'500.–, une loi $N(0,150^2)$; et pour les loyers de plus de fr. 1'500.–, une loi $N(0,200^2)$. La seconde simulation avec $\alpha = 240\%$ utilise les mêmes distributions mais avec des écarts-types doublés. Les estimations de l'indice, ainsi que celles de l'écart-type et du biais, sont rapportées dans le tableau 8.

Tableau 8. Indice des loyers pour des taux de croissance simulés

Loyers libres	$\hat{I}_{t/0}$	$\hat{\sigma}_{\hat{I}_{t/0}}$	Biais	n
$\alpha = 70\%$	170.3032	.7173	.0223	3'939
$\alpha = 240\%$	340.6064	1.4346	.0446	3'939

On remarque que tant la dispersion que le biais augmentent. Un examen des formules (16) et (17) montre que l'accroissement du biais, qui est indépendant de la dispersion des strates à l'époque t, résulte de l'accroissement de la valeur des indices par strate r_h Quant à la variance, qui est due essentiellement à la dispersion des r_h, elle croît avec les r_h, mais aussi avec la dispersion au sein des strates. Dans notre exemple, elle croît dans les mêmes proportions que le biais car les r_h, les écarts-types σ_{ht} et les covariances σ_{h0t} sont tous approximativement multipliés par le même facteur α. Si l'on compare les résultats de la simulation avec les valeurs données au tableau 5 pour les loyers libres, on note que le biais et l'écart-type sont multipliés respectivement par environ 1.85 et 3.7, soit un peu plus que par la valeur de α. La différence résulte des aléas supplémentaires u_i qui ont notamment pour effet de limiter l'augmentation des covariances induite par le taux de croissance.

3.3 Non prise en compte de la covariance

Dans la pratique de la production d'un indice, il est important que l'application de la méthode théorique puisse facilement être mise en oeuvre car il est souvent essentiel de pouvoir automatiser le plus possible les procédures de calcul. Dans le cas d'un échantillon rotatif, qui remplace à chaque enquête une part donnée des observations, l'obstacle le plus grand à la réalisation de routines de calcul est posé par l'estimation de la covariance. Il est en effet nécessaire de sélectionner les cas appariés à chaque nouveau tirage et cette contrainte exige un traitement informatique spécifique. Pour déterminer dans quelle mesure l'exclusion du terme de la covariance affecte les estimations de l'indice et de ses caractéristiques, nous avons effectué les calculs sans tenir compte de la covariance. Compte tenu de la formulation des expressions (8) et (9), où le terme de la covariance apparaît avec un signe négatif, on doit s'attendre à une sur-estimation à la fois du biais et de la variance de l'indice. Les valeurs calculées (tableau 9) confirment cette attente.

Tableau 9. Sans tenir compte des covariances

Loyers libres	$\hat{I}_{t/0}$	$\hat{\sigma}_{I_{t/0}}$	$\hat{B}$iais
novembre1993	99.9254	.6560	.0327
$\alpha = 70$ %	170.3032	1.1519	.0558
$\alpha = 240$ %	340.6064	2.3037	.1116

Comme on pouvait s'y attendre, on sous-estime la précision des résultats en excluant les covariances induites par les données appariées. En effet, les valeurs du biais données dans le tableau 9 sont environ deux fois et demi celles calculées précédemment, tandis que les écarts-types sont approximativement multipliés par un facteur 1.7. Ces derniers représentent ici environ 0.66% de la valeur de l'indice contre 0.4% en tenant compte de la correction pour les covariances. On peut relever que l'évaluation de la précision relative reste indépendante du niveau de l'indice. Par ailleurs, la comparaison des simulations aux résultats pour novembre 1993 indique que l'écart-type et le biais sont ici proportionnels au taux de croissance, ce qui confirme que les différences observées précédemment résultent des covariances.

D'un point de vue théorique, les écarts entre les approximations obtenues en ne tenant pas compte des covariances et les valeurs réelles des estimations sont importantes. Les approximations vont cependant dans le sens des estimations plutôt conservatrices auxquelles on a souvent recours dans la pratique. A titre d'exemple, L'Office fédéral de la statistique (1993, p. 63) mentionne un écart-type de 0.75 pour l'indice des loyers qu'il calcule dans le cadre de l'IPC et ceci par rapport à un échantillon de taille comparable à celui que nous avons utilisé (de l'ordre donc de quelques 5'000 logements). Les autres

offices statistiques (Zurich, Bâle, Genève) qui établissent un indice des prix des loyers ne publient pas d'intervalles de confiance pour leurs mesures.

Il convient de souligner que pour des échantillons rotatifs tels que ceux considérés, l'effet réducteur des covariances s'amenuise avec les périodes. En effet, le pourcentage λ d'observations appariées dont dépend la covariance (13) diminue à chaque nouvelle enquête. Ainsi, pour un taux de rotation ρ, on peut avoir après t périodes $\lambda = 1 - t\rho$ et donc un échantillon totalement renouvelé après $1 / \rho$ périodes. Pour l'enquête lausannoise, par exemple, il ne devrait plus y avoir de données appariées après huit semestres et, par conséquent, plus de covariances à prendre en compte. Par contre, dans le cas d'échantillons pleinement appariés, le gain de précision mesuré par les termes incluant les covariances sera plus important encore que celui illustré ci-dessus.

En résumé, lorsqu'on dispose de données appariées, l'exclusion de la covariance conduit à une sous-estimation de la précision statistique de l'indice calculé qui est d'autant plus importante que la proportion de données appariées est forte. Le traitement informatique s'en trouve par contre considérablement allégé. Cette option, qui a finalement des implications limitées pour des échantillons rotatifs, facilite donc la tâche du praticien. Celui-ci ne peut, malheureusement, souvent pas consacrer beaucoup de temps aux développements théoriques les plus élaborés, mais doit néanmoins pouvoir garantir la fiabilité des statistiques qu'il produit. En dernière instance, pour un degré de significativité statistique souhaité, le choix d'appliquer complètement ou partiellement une méthode dépend de contraintes matérielles liées à l'environnement informatique, au nombre de collaborateurs et aux limites temporelles imposées à la production d'une statistique choisie.

4 CONCLUSION

Nous avons établi que l'estimation de l'indice des loyers est biaisé, mais que ce biais est négligeable comme l'illustre notamment l'application aux données de l'enquête lausannoise. Pour ce qui est de l'effet de la post-stratification, il concerne essentiellement la dispersion de l'estimateur. Là aussi, il apparaît que l'incertitude sur les tailles des strates n'a qu'un impact très modéré sur la variance de $\hat{I}_{t/0}$. Les implications de la covariance induite par les données appariées sont de ce point de vue plus conséquentes, mais disparaissent au cours du temps dans le cas d'échantillons rotatifs.

ANNEXE: DEMONSTRATIONS MATHEMATIQUES

A.1 Démonstration du lemme 1

Les deux premiers résultats sont classiques et ne sont pas démontrés ici. Pour (5), on fait dans un premier temps le calcul pour le cas de tirages indépendants. La covariance pour les tirages sans remises s'en déduit en appliquant le facteur de correction pour population finie.

Associons à chaque tirage i, $i = 1,2..,n$, les variables binaires

$$Y_{ih} = \begin{cases} 1 & \text{si } K_i \in \mathbf{K}_h \\ 0 & \text{sinon} \end{cases} \quad Y_{is} = \begin{cases} 1 & \text{si } K_i \in \mathbf{K}_s \\ 0 & \text{sinon} \end{cases} \tag{18}$$

Les variables Y_{ih} et Y_{is} ne sont pas indépendantes. On a en particulier $P(Y_{ih} = 1 \text{ et } Y_{is} = 1) = 0$, et donc

$$\mathrm{E}(Y_{ih}\, Y_{is}) = 0 \Leftrightarrow \mathrm{Cov}(Y_{ih}, Y_{is}) = -p_h\, p_s \tag{19}$$

Par contre, les tirages étant supposés indépendants, Y_{ih} et Y_{js} sont indépendants pour tout $i \neq j$. On a donc $P(Y_{ih} = 1 \text{ et } Y_{js} = 1) = p_h\, p_s$, et par conséquent (20)

$$\mathrm{E}(Y_{ih}\, Y_{is}) = p_h\, p_s \Leftrightarrow \mathrm{Cov}(Y_{ih}, Y_{is}) = 0$$

Comme

$$n_h = \sum_{i=1}^{n} Y_{ih} \quad \text{et} \quad n_s = \sum_{i=1}^{n} Y_{is} \tag{21}$$

on a

$$\begin{aligned} \mathrm{E}(n_h\, n_s) &= \sum_{i=1}^{n} \mathrm{E}(Y_{ih}\, Y_{is}) + \sum_{i=1}^{n} \sum_{j \neq i} \mathrm{E}(Y_{ih}\, Y_{is}) \\ &= 0 + n(n-1)\, p_h\, p_s \end{aligned}$$

En retranchant à ce moment croisé le produit des espérances $\mathrm{E}(n_h)\, \mathrm{E}(n_s) = n^2 p_h\, p_s$, on obtient la covariance entre n_h et n_s, soit

$$\mathrm{Cov}(n_h, n_s) = -np_h p_s \tag{22}$$

En multipliant ce dernier résultat par f^{pc} on établit (5).

Pour démontrer (6), on considère un développement limité de la fonction $f(n_h) = 1 / n_h$ autour de $\mathrm{E}(n_h) = np_h$, soit

$$f(n_h) = \frac{1}{np_h} - \frac{1}{n^2 p_h^2}(n_h - np_h) + \frac{1}{n^3 p_h^3}(n_h - np_h)^2 + R_3 \tag{23}$$

Comme $\mathrm{E}(n_h - np_h) = 0$ et $\mathrm{E}(n_h - np_h)^2 = \mathrm{Var}(n_h)$, l'espérance de l'expression précédente donne, si l'on néglige l'espérance du reste R_3, l'approximation suivante de l'espérance de $1 / n_h$

$$\mathrm{E}\left(\frac{1}{n_h}\right) = \frac{1}{np_h} + \frac{1}{n^3 p_h^3}\mathrm{Var}(n_h)$$

Dans le cas de tirages sans remises, on obtient le résultat (6) en remplaçant Var (n_h) par (4).

A.2 Démonstration du théorème 1

En utilisant les propriétés des espérances conditionnelles (voir par exemple MOOD et al., 1974, 158-159), on peut écrire:

$$\mathrm{E}(X) = \mathrm{E}_y(\mathrm{E}(X \mid Y)) \tag{24}$$

$$\mathrm{Var}(X) = \mathrm{Var}_y\Big(\mathrm{E}(X \mid Y)\Big) + \mathrm{E}_y\Big(\mathrm{Var}(X \mid Y)\Big) \tag{25}$$

En supposant des tirages sans remises (et à probabilités égales), on a

$$\mathrm{E}(\bar{X}_h \mid n_h) = \mu_h \tag{26}$$

$$\begin{aligned} \mathrm{Var}(\bar{X}_h \mid n_h) &= \left(1 - \frac{n_h}{m_h}\right)\frac{\sigma_h^{*2}}{n_h} \\ &= \frac{\sigma_h^{*2}}{n_h} - \frac{\sigma_h^{*2}}{m_h} \end{aligned} \tag{27}$$

Comme μ_h ne dépend pas de n_h, on a alors $\mathrm{Var}(\overline{X}_h) = \mathrm{E}_{n_h}(\mathrm{Var}(\overline{X}_h \mid n_h))$. Le dernier terme de (27) est non aléatoire. Dans le premier, seul le dénominateur n_h est aléatoire. Ainsi, en utilisant le résultat (6), l'espérance de $\mathrm{Var}(\overline{X}_h / n_h)$ vaut

$$\mathrm{E}_{n_h}\left((\mathrm{Var}(\overline{X}_h \mid n_h)\right) = \sigma_h^{*2}\left(\frac{1}{np_h} - \frac{1}{m_h}\right) + \sigma_h^{*2} f^{pc} \frac{(1-p_h)}{n^2 p_h^2}$$

On obtient (7) en notant que $m_h = mp_h$. Le théorème est ainsi démontré.

A.3 Démonstration du théorème 2

Il s'agit d'étudier la distribution d'un rapport. Considérons le développement limité au second ordre de la fonction g autour de $(\mathrm{E}(\overline{X}), \mathrm{E}(\overline{Y})) = (\mu_x, \mu_y)$, soit, en notant grad g le vecteur colonne des dérivées premières de la fonction g, et $\partial^2 g / \partial \overline{X} \partial \overline{Y}$ la matrice des dérivées secondes,

$$\begin{aligned} g(\overline{X}, \overline{Y}) &= g(\mu_x, \mu_y) + {}^t\mathrm{grad}\, g(\mu_x, \mu_y)\begin{pmatrix}\overline{X}-\mu_x \\ \overline{Y}-\mu_y\end{pmatrix} \\ &\quad + \frac{1}{2}(\overline{X}-\mu_x, \overline{Y}-\mu_y)\frac{\partial^2 g(\mu_x, \mu_y)}{\partial \overline{X}\partial \overline{Y}}\begin{pmatrix}\overline{X}-\mu_x \\ \overline{Y}-\mu_y\end{pmatrix} + R_3 \qquad (28) \\ &= \frac{\mu_x}{\mu_y} + \frac{1}{\mu_y}(\overline{X}-\mu_x) - \frac{\mu_x}{\mu_y^2}(\overline{Y}-\mu_y) \\ &\quad + \frac{\mu_x}{\mu_y^3}(\overline{Y}-\mu_y)^2 - \frac{1}{\mu_y^2}(\overline{X}-\mu_x)(\overline{Y}-\mu_y) + R_3 \qquad (29) \end{aligned}$$

En prenant l'espérance de (29) et en négligeant le reste R_3, on établit l'approximation (8) d'ordre $1/n$ de l'espérance du rapport $\overline{X}/\overline{Y}$.

Pour la variance, on note tout d'abord qu'elle diffère de $\mathrm{E}((\overline{X}/\overline{Y})^2)$ par le carré du biais de $\overline{X}/\overline{Y}$ en tant qu'estimateur de r. Ce biais, le second terme de (8), est en $1/n$. Son carré, qui est donc en $1/n^2$ peut être négligé, et l'on a alors, à l'ordre $1/n$, $\mathrm{Var}(\overline{X}/\overline{Y}) = \mathrm{E}((\overline{X}/\overline{Y} - \mu_x/\mu_y)^2)$.

En négligeant R_3 et le terme du second ordre dans le développement (29), l'écart $g(\overline{X}/\overline{Y}) - g(\mu_x/\mu_y)$ vaut

$$\frac{\overline{X}}{\overline{Y}} - \frac{\mu_x}{\mu_y} = \frac{1}{\mu_y^2}\left(\mu_y(\overline{X}-\mu_x) - \mu_x(\overline{Y}-\mu_y)\right) \qquad (30)$$

En mettant au carré et en prenant l'espérance, on obtient l'approximation suivante d'ordre 1 / n de la variance

$$\mathrm{Var}\left(\frac{\bar{X}}{\bar{Y}}\right) = \frac{1}{\mu_y^4}\left(\mu_y^2 \mathrm{Var}(\bar{X}) - 2\mu_x\mu_y \mathrm{Cov}(\bar{X}, \bar{Y}) + \mu_x^2 \mathrm{Var}(\bar{Y})\right) \tag{31}$$

La forme (9) s'en déduit en simplifiant par μ_y^2. Cette expression peut également s'écrire sous la forme (10). En effet

$$\mathrm{E}\left((\bar{X} - \frac{\mu_x}{\mu_y}\bar{Y})^2\right) = \mathrm{E}\left(\left((\bar{X} - \mu_x) - \frac{\mu_x}{\mu_y}(\bar{Y} - \mu_y)\right)^2\right)$$

Le théorème est ainsi démontré.

A.4 Démonstration du théorème 3

Le développement de $\mathrm{E}(\hat{I}_{t/0})$ et $\mathrm{Var}(\hat{I}_{t/0})$ en termes d'espérances et de variances conditionnelles donne

$$\mathrm{E}(\hat{I}_{t/0}) = \mathrm{E}_{n_1, \ldots, n_c}\left(\mathrm{E}(\hat{I}_{t/0} \mid n_1, \ldots, n_c)\right) \tag{32}$$

$$\mathrm{Var}(\hat{I}_{t/0}) = \mathrm{Var}_{n_1, \ldots, n_c}\left(\mathrm{E}(\hat{I}_{t/0} \mid n_1, \ldots, n_c)\right) + \mathrm{E}_{n_1, \ldots, n_c}\left(\mathrm{Var}(\hat{I}_{t/0} \mid n_1, \ldots, n_c)\right) \tag{33}$$

avec

$$\mathrm{E}(\hat{I}_{t/0} \mid n_1, \ldots, n_c) = \frac{1}{n}\sum_h n_h \mathrm{E}\left(\frac{\bar{X}_{ht}}{X_{h0}} \mid n_h\right) \tag{34}$$

$$\mathrm{Var}(\hat{I}_{t/0} \mid n_1, \ldots, n_c) = \frac{1}{n^2}\sum_h n_h^2 \mathrm{Var}\left(\frac{\bar{X}_{ht}}{X_{h0}} \mid n_h\right) \tag{35}$$

Il s'agit donc de déterminer l'espérance et la variance de (34) et l'espérance de (35). Commençons par le calcul de $\mathrm{E}(\sum_h (n_h/n) E(\bar{X}_{ht}/X_{h0}))$. Selon le Théorème 2

$$\mathrm{E}\left(\frac{\bar{X}_{ht}}{\bar{X}_{h0}} \mid n_h\right) = r_h\left(1 + \frac{\mathrm{Var}(\bar{X}_{h0} \mid n_h)}{\mu_{h0}^2} - \frac{\mathrm{Cov}(\bar{X}_{h0}, \bar{X}_{ht} \mid n_h)}{\mu_{ht}\,\mu_{h0}}\right). \tag{36}$$

Comme

$$\mathrm{Var}(\bar{X}_{h0} \mid n_h) = \frac{\sigma^{*2}_{h0}}{n_h} - \frac{\sigma^{*2}_{h0}}{m_h}. \tag{37}$$

$$\mathrm{Cov}(\bar{X}_{h0}, \bar{X}_{ht} \mid n_h) = \frac{\sigma^{*}_{h0t}}{n_h} - \frac{\sigma^{*}_{h0t}}{m_h} \tag{38}$$

on a, en posant $b^*_h = \sigma^{*2}_{h0} / \mu^2_{h0} - \sigma^*_{h0t} / (\mu_{h0}\, \mu_{ht})$

$$n_h \,\mathrm{E}\left(\frac{\bar{X}_{ht}}{\bar{X}_{h0}} \mid n_h\right) = n_h r_h + r_h b^*_h \left(1 - \frac{n_h}{m_h}\right). \tag{39}$$

Comme $\mathrm{E}(n_h) = np_h$ et $m_h = mp_h$, $\mathrm{E}(\hat{I}_{t0})$, qui est l'espérance de (34), vaut

$$\mathrm{E}(\hat{I}_{t0}) = \sum_h p_h r_h + \left(1 - \frac{n}{m}\right) \frac{1}{n} \sum_h r_h b^*_h$$

ce qui établit (16).

Pour le calcul de la variance, considérons en premier lieu la variance de (34). En posant $r^*_h = r_h\,(1 - b^*_h / m_h)$, on peut réécrire (39) sous la forme

$$n_h \,\mathrm{E}\left(\frac{\bar{X}_{ht}}{\bar{X}_{h0}} \mid n_h\right) = r_h b^*_h + n_h r^*_h. \tag{40}$$

Seul le dernier terme est en n_h et est donc aléatoire. La variance de (34) est alors

$$\mathrm{Var}\left(\frac{1}{n} \sum_h n_h \,\mathrm{E}\left(\frac{\bar{X}_{ht}}{\bar{X}_{h0}} \mid n_h\right)\right) = \mathrm{Var}\left(\frac{1}{n} \sum_h n_h r^*_h\right). \tag{41}$$

soit, en tenant compte du Lemme 1,

$$\begin{aligned} \mathrm{Var}\left(\frac{1}{n} \sum_h n_h\, r^*_h\right) &= \frac{1}{n^2} \sum_h r^{*2}_h \,\mathrm{Var}(n_h) + \frac{1}{n^2} \sum_h \sum_{s \neq h} r^*_h r^*_s \,\mathrm{Cov}(n_h, n_s) \\ &= f^{pc} \frac{1}{n} \left(\sum_h r^{*2}_h p_h\,(1 - p_h) - \sum_h \sum_{s \neq h} r^*_h r^*_s p_h\, p_s\right). \end{aligned} \tag{42}$$

Considérons à présent l'espérance de $\sum_h (n_h/n)^2 \operatorname{Var}(\bar{X}_{ht}/\bar{X}_{h0} \mid n_h)$. Selon le Théorème 2, on a

$$\operatorname{Var}\left(\frac{\bar{X}_{ht}}{\bar{X}_{h0}} \mid n_h\right) = \frac{1}{\mu_{h0}^2}\left(\operatorname{Var}(\bar{X}_{ht} \mid n_h) - 2r_h \operatorname{Cov}(\bar{X}_{h0}, \bar{X}_{ht} \mid n_h) + r_h^2 \operatorname{Var}(\bar{X}_{h0} \mid n_h)\right)$$

En tenant compte de (37) et (38), et en posant $\beta_h^* = \left(\sigma_{ht}^{*2} - 2r_h\sigma_{h0t}^* + r_h^2\sigma_{h0}^{*2}\right)/\mu_{h0}^2$, on établit

$$n_h^2 \operatorname{Var}\left(\frac{\bar{X}_{ht}}{\bar{X}_{h0}} \mid n_h\right) = n_h\beta_h^* - n_h^2\frac{\beta_h^*}{m_h}. \tag{43}$$

Comme $\mathrm{E}(n_h) = np_h$ et, $\mathrm{E}(n_h^2) = f^{pc}np_h(1-p_h) + n^2p_h^2$, l'espérance mathématique de l'expression précédente vaut

$$\mathrm{E}\left(n_h^2 \operatorname{Var}\left(\frac{\bar{X}_{ht}}{\bar{X}_{h0}} \mid n_h\right)\right) = np_h\beta_h^* - \left(f^{pc}np_h(1-p_h) + n^2p_h^2\right)\frac{\beta_h^*}{m_h} \tag{44}$$

$$= np_h\beta_h^*\left(1 - \frac{(f^{pc}(1-p_h)/p_h + n)}{m}\right). \tag{45}$$

d'où

$$\mathrm{E}\left(\frac{1}{n^2}\sum_h n_h^2 \operatorname{Var}\frac{\bar{X}_{ht}}{\bar{X}_{h0}}\right) = \frac{1}{n}\sum_{h=1}^{c} p_h\beta_h^*\left(1 - \frac{(f^{pc}(1-p_h)/p_h + n)}{m}\right). \tag{46}$$

Finalement, la variance de $\hat{I}_{t/0}$ s'obtient en sommant (46) et (42). Le Théorème est ainsi démontré.

SUMMARY

The paper establishes the theoretical properties of a synthesized index number computed from an a posteriori stratified sample. The index number considered can be expressed as a weighted mean of the mean ratios computed for the subsamples. In the first part, we show that the index has a biais in $1/n$ and that its variance can be expressed as the sum of two terms reflecting the uncertainty on the size of the subsamples and the uncertainty on the ratio in each stratum. In the second part, we apply the theoretical results to the data from the recent survey of housing rent prices in Lausanne.

RESUME

Cet article établit les propriétés théoriques d'un indice synthétique calculé avec un échantillon stratifié a posteriori. L'indice considéré s'exprime comme une moyenne pondérée des rapports de moyennes calculées pour chaque sous-échantillon. Dans la première partie, nous montrons que cet indice comporte un biais qui dépend de $1 / n$ et que sa variance peut être exprimée par la somme de deux termes qui reflètent l'incertitude liée à la taille des sous-échantillons et l'incertitude des rapports de moyennes de chaque strate. Dans la deuxième partie nous appliquons les résultats théoriques aux données de la récente enquête sur les loyers lausannois.

ZUSAMMENFASSUNG

Ziel dieses Artikels ist es, die theoretischen Eigenschaften eines synthetischen Indexes, der aus einer a posteriori geschichteten Stichprobe berechnet wurde, herauszuarbeiten. Wir betrachten den Fall eines synthetischen Indexes, der als gewichteter Durchschnitt des für jede Unterstichprobe berechneten Durchschnittskoeffizienten ausgedrückt werden kann. Im ersten Teil zeigen wir, dass dieser Index verzerrt ist und dass die Verzerrung von $1 / n$ abhängig ist. Die Varianz des Indexes lässt sich als Summe von zwei Elementen schreiben, die die mit der Grösse der Stichprobe verbundene Ungewissheit sowie die Ungewissheit über den Durchschnittskoeffizienten jeder Schicht reflektieren. Im zweiten Teil wenden wir die theoretischen Resultate auf die Daten der kürzlich erfolgten Umfrage über Lausanner Mieten an.

REFERENCES

BARNETT, VIC (1991), Sample Survey, Principles and Methods, Edward Arnold, London.

CHANCELLERIE FEDERALE (1993), Législation relative à la statistique fédérale, Chancellerie fédérale, Berne.

COCHRAN, WILLIAM G. (1977), Sampling Techniques, Wiley, New York.

DEPARTEMENT FEDERAL DE L'ECONOMIE PUBLIQUE (1977), L'indice suisse des prix à la consommation: les nouvelles bases et méthodes de son calcul dès 1977, La vie économique, 89e numéro spécial.

GROSBRAS JEAN-MARIE (1987), Méthodes statistiques des sondages, Economica, Paris.

MOOD, A.M., F.A. GRAYBILL and D.C. BOES (1974), Introduction to the Theory of Statistics, McGraw-Hill, New York.

OFFICE FEDERAL DE LA STATISTIQUE (1993), Révision de l'indice suisse des prix à la consommation, Statistique de la Suisse, domaine 5 prix, Berne.