Proceedings Chapter

Comportement d'indices obtenus par post-stratification

RITSCHARD, Gilbert

Comportement d'indices obtenus par post-stratification

Gilbert Ritschard

1. Introduction

L'étude présentée ici est née du souci de l'Office d'études socio-économiques et statistiques de la Ville de Lausanne de pouvoir donner des intervalles de confiance pour les indices de loyers qu'il publie depuis 1993, soit depuis que l'Office fédéral de la statistique (1993) a renoncé à publier des indices de loyers locaux. Les indices calculés par la Ville de Lausanne sont obtenus, comme la plupart des indices de loyers, par post-stratification. On observe les loyers payés pour un échantillon d'appartements que l'on stratifie a posteriori selon, par exemple, le nombre de pièces, la date de construction ou de rénovation de l'immeuble ou l'unité géographique de localisation des logements. Pour chaque strate, on calcule le rapport des loyers moyens payés à deux dates distinctes, et l'indice synthétique est le résultat d'une moyenne pondérées de ces rapports. En fait, il s'agit d'une estimation, et le problème examiné ici est celui de son comportement statistique compte tenu de la procédure d'échantillonnage retenu. Il s'inscrit donc dans la lignée des travaux que Robert Fortet a mené dans le cadre de la théorie des sondages.

Nous rappelons dans un premier temps les résultats asymptotiques établis dans Ritschard & Dozio (1995). Nous présentons ensuite les enseignements de simulations qui ont été menées pour en vérifier empiriquement la validité.

Avant d'aborder les problèmes liés à l'échantillonnage, il nous parait important de dire quelques mots sur la pertinence de la forme d'indice considérée. S'agissant de loyers, on peut songer en effet a priori à deux approches. 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 des indices simples de chaque logement.

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 petits 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.

Le second principe, qui permet une pondération des unités observées indépendante du niveau du loyer, suppose par contre que l'on dispose de l'indice simple pour chaque logement. Ceci pose des problèmes eu égard notamment au renouvellement de la population de référence. L'indice ne peut pas, par exemple, être calculé de cette manière lorsqu'on procède, comme c'est l'usage, à des rotations d'effectifs.

L'indice retenu est une forme intermédiaire obtenue 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 permet ainsi d'éviter la distorsion introduite par la première approche et les limites d'application de la seconde. Un tel indice nécessite naturellement une stratification de l'ensemble des logements en groupes homogènes.

La suite de l'article est consacrée aux propriétés statistiques de l'estimateur de cet indice lorsque la stratification est réalisée a posteriori, c'est-à-dire dans le cas où l'échantillon est choisi indépendamment des critères de stratification, ceux-ci ne s'appliquant qu'après coup sur l'échantillon. Les aspects examinés sont essentiellement le biais et la dispersion de l'estimateur de l'indice.

La section 2 présente les propriétés statistiques théoriques de l'indice empirique calculé sur la base d'un échantillon post-stratifié. La particularité du cas considéré tient à la taille aléatoire des sous-échantillons, et par conséquent des pondérations. 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 le biais et la variance asymptotique 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.

Le comportement de l'indice est ensuite étudié empiriquement à la section 3 par le biais de simulations. Celles-ci montrent en particulier que la formule de variance asymptotique est fiable, même pour des échantillons de petite taille ($n = 20$). Elles permettent également de confirmer la normalité de la distribution de l'estimateur.

2. Propriétés 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 corrections utilisés respectivement avec σ^2 et σ^{*2}.

Les indices que l'on se propose d'estimer s'écrivent formellement :

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

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

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 dis-

pose 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 $\{\mathbf{E}_1, \mathbf{E}_2, \ldots, \mathbf{E}_c\}$, est définie comme suit:

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

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 $\sum_h n_h = n$, 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 paramètre *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, $h = 1,2,...,c$, *des sous-échantillons résultant d'une stratification a posteriori vérifient les propriétés suivantes*

$$\mathrm{E}(n_h) = np_h \tag{3}$$
$$\mathrm{Var}(n_h) = f^{pc} np_h(1-p_h) \tag{4}$$
$$\mathrm{Cov}(n_h, n_s) = -f^{pc} np_h p_s \tag{5}$$
$$\mathrm{E}\left(\frac{1}{n_h}\right) = \frac{1}{np_h} + f^{pc}\frac{(1-p_h)}{n^2 p_h^2} \tag{6}$$

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 une attention particulière lorsque certaines strates comprennent une petite proportion de la population, soit notamment lorsqu'on a un grand nombre de strates.

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

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

Le tableau 1 illustre l'évolution de la variance de $\overline{X}_h$ et de ces 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 $n_h \geq 2$. Lorsque n_h est nul ou égal à un, il conviendra soit de procéder à un regroupement de strates, soit d'exclure la strate correspondante de l'analyse.

2.3. Rapport de deux moyennes

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

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

Les résultats donnés par le théorème ci-dessous s'appliquent notamment à l'estimateur $\hat{r}_{h(t/0)} = \overline{X}_{ht} / \overline{X}_{h0}$ de l'indice défini en (1) pour une strate.

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

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

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

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

Le second terme de (8) représente le biais du rapport $\overline{X} / \overline{Y}$ 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 $\overline{X}$ et $\overline{Y}$ 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 $f^{biais} = (1 + \mathrm{Var}(\overline{Y}) / \mu_y^2 - \mathrm{Cov}(\overline{X},\overline{Y}) / \mu_x\mu_y)^{-1}$. Pratiquement, comme μ_x, μ_y, $\mathrm{Var}(\overline{Y})$ et $\mathrm{Cov}(\overline{X},\overline{Y})$ 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

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

$$\hat{\mathrm{C}}\mathrm{ov}(\overline{X},\overline{Y}) = f^{pc}\frac{1}{n(n-1)}\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y}) \tag{12}$$

Pour des échantillons non appariés indépendants, la covariance est nulle. L'estimation de $\mathrm{Var}(\overline{Y})$ 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, seuls un sous-ensemble des observations restent appariées. On estime dans ce cas la covariance $\mathrm{Cov}(\overline{X},\overline{Y})$ à partir des $n_a = \lambda n$ observations appariées avec l'estimateur:

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

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

Remarquons cependant que tant la variance de $\overline{Y}$ que la covariance entre $\overline{X}$ et $\overline{Y}$ 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, $\mathrm{E}\left((\overline{X} - (\mu_x / \mu_y)\overline{Y})^2\right)$ 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 $\mathrm{Cov}(\overline{X},\overline{Y})$ est nulle en vertu de l'indépendance, et l'on remplace les variances par leurs estimations.

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

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

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

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

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

où

$$\begin{aligned} r_h &= \mu_{ht} / \mu_{h0} \\ b_h^* &= \sigma_{h0}^{*2} / \mu_{h0}^2 - \sigma_{h0t}^* / (\mu_t \mu_0) \\ r_h^* &= r_h (1 - b_h^* / m_h) \\ \beta_h^* &= (\sigma_{ht}^{*2} - 2 r_h \sigma_{h0t}^* + r_h^2 \sigma_{h0}^{*2}) / \mu_{h0}^2 \end{aligned}$$

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

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.

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 est donnée 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 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	

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.

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 $\hat{I}_{t/0}$

$$\text{Biais}(\hat{I}_{t/0}) = 0.00012$$

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

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 $\mathrm{E}(\hat{I}_{t/0}) = 1.4947$, que de la vraie valeur de l'indice.

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

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

$$\hat{\mathrm{B}}\mathrm{iais}(\hat{I}_{t/0}) = 0.00003$$
$$\hat{\mathrm{V}}\mathrm{ar}(\hat{I}_{t/0}) = 0.003944 + 0.000051 = 0.003995$$
$$\hat{\sigma}_{\hat{I}_{t/0}} = 0.0632$$

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

3. Simulations

Nous présentons ici quelques simulations que nous avons menées afin d'analyser empiriquement la distribution de l'estimateur $\hat{I}_{t/0} = \sum_h (n_h / n)(\overline{X}_{ht} / \overline{X}_{h0})$ de $I_{t/0}$. Dans la perspective de pouvoir construire des intervalles de confiances à partir de l'expression asymptotique donnée pour la variance au théorème 3, il s'agit d'une part de vérifier la pertinence de cette formule pour des échantillons de tailles finies et, d'autre part, de s'assurer que la distribution de $\hat{I}_{t/0}$ est approximativement normale.

Pour ces simulations, nous avons considéré une population fictive de 1'000 logements répartis en 6 catégories. Les loyers x_{h0i} (en francs suisses) de la période 0 ont été générés aléatoirement selon une loi normale $N(\mu_h, \sigma_h^2)$, dont la moyenne μ_h et l'écart-type σ_h ont été fixés arbitrairement pour chaque strate h. Les loyers x_{hti} de la période t ont été calculés à partir des valeurs générées pour la période 0 en postulant un taux de croissance g_h différent pour chaque strate et un bruit d'écart-type proportionnel au niveau du loyer initial. Formellement, les x_{hti} ont été générés comme suit : $x_{hti} = (1 + g_h)x_{h0i} + u_{hi}$, où u_{hi} est $N\left(0, (0.05x_{h0i})^2\right)$.

Tableau 5. Valeurs numériques des paramètres utilisés

h	n_h	μ_h	σ_h	g_h
1	45	800	150	5%
2	101	1'000	200	4%
3	185	1'300	300	8%
4	352	1'700	325	10%
5	236	2'300	375	8%
6	81	3'000	400	7%

Les valeurs numériques postulées des paramètres sont données au tableau 5, le tableau 6 donne les caractéristiques des loyers générés et la figure 1 montre leur distribution.

Tableau 6. Caractéristiques de la population générée

h	μ_{h0}	σ_{h0}	μ_{ht}	σ_{ht}	$r_{t/0}$
1	763.74	170.99	807.17	184.04	1.057
2	1’009.66	204.57	1’043.78	212.67	1.034
3	1’295.29	285.84	1’399.63	316.66	1.081
4	1’691.20	324.73	1’860.07	371.47	1.100
5	2’307.21	358.93	2’490.07	398.45	1.079
6	3’062.94	414.63	3’284.12	451.56	1.072

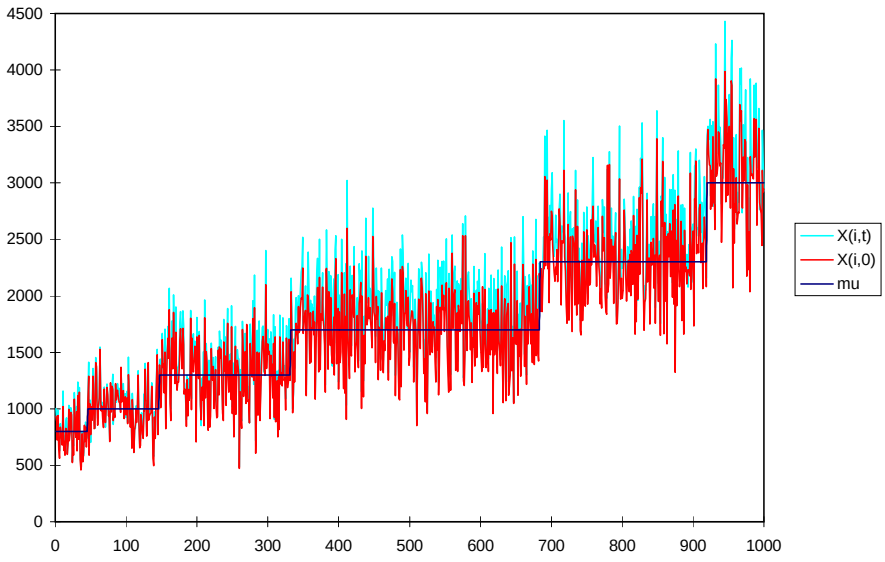

Figure 1. Données: 1’000 loyers, 6 catégories de logements, 2 périodes

Sur la base de ces données, nous avons procédé à deux séries de simulations. Dans la première, nous avons tirés aléatoirement 200 échantillons de $n = 100$ logements, et dans la seconde, 200 échantillons de $n = 20$ logements. Nous avons considéré le cas de données appariées en retenant pour chaque logement échantillonné son loyer à la période 0 et celui en t. Les estimations $\hat{I}_{t/0}$ de l’indice et $\hat{\sigma}_{\hat{I},\infty}$ de son

écart-type asymptotique ont été calculées pour chaque échantillon en tenant compte de cet appariement. On dispose ainsi de deux séries de 200 estimations de l'indice et de son écart-type.

La moyenne et l'écart-type des 200 valeurs $\hat{I}_{t/0}$ obtenues pour des échantillons de taille $n = 100$ sont donnés au tableau 7. En comparant la moyenne des indices $\hat{I}_{t/0}$ à la vraie valeur $I_{t/0}$ calculée pour l'ensemble des 1'000 logements constituant la population, on constate, pour un indice exprimé ici en base 100, une très légère sur-estimation moyenne de 0.052. Ce faible biais est clairement non significatif en comparaison de l'écart-type des estimations. Celui-ci est en effet de l'ordre d'un demi point (0.492) et est proche de l'écart-type asymptotique (0.517) calculé avec la formule du théorème 3. Dans la pratique, on doit se contenter de l'estimation de l'écart-type asymptotique. Il est, de ce point de vue, utile de constater que la moyenne (0.524) des estimations de l'écart-type asymptotique est elle aussi très proche de l'écart-type observé. La différence de 0.032 reste notamment inférieure à l'écart-type (0.038) des estimations de ces écarts asymptotiques.

La simulation réalisée avec des échantillons de taille 20 montre (tableau 8) que la formule de variance asymptotique reste fiable pour évaluer l'écart-type des indices estimés avec de petits échantillons. Les estimations de l'indice sont évidemment moins fiables, leur écart-type dépassant ici un point en raison de la taille réduite des échantillons. On note cependant que, comme précédemment, l'écart-type des 200 indices calculés est proche de l'écart-type asymptotique, est que l'estimateur de ces écarts-types semble non biaisé.

Si l'on veut construire des intervalle de confiance, il s'agit encore de connaître la distribution de l'estimateur $\hat{I}_{t/0}$. Les figures 2 et 3 donnent, pour chacune des deux simulations, la répartition des quantiles empiriques en regard des quantiles de la loi normale $N(I_{t/0}, \sigma^2_{\hat{I},\infty})$, où $I_{t/0}$ est la valeur de l'indice pour l'ensemble de la population, et $\sigma^2_{\hat{I},\infty}$ la variance asymptotique de $\hat{I}_{t/0}$. On observe que les distributions empiriques sont proches de la loi normale. Il semble donc très raisonnable d'exploiter la loi normale pour la construction d'intervalles.

A titre d'exemple, pour le dernier des 200 échantillons de taille 100, l'indice calculé $\hat{I}_{t/0}$ vaut 108.526, et l'estimation de l'écart-type 0.552. L'intervalle de confiance à 95 % est alors dans ce cas $[107.444 ; 109.608]$. De même, pour le dernier échantillon de taille 20, $\hat{I}_{t/0}$ vaut 105.565 et $\hat{\sigma}_{\hat{I}}$ 1.266, ce qui donne, pour un degré de confiance de 95 % l'intervalle $[103.084 ; 108.046]$.

Tableau 7. Résultats pour 200 échantillons de taille 100

Indice		Ecart-type	
Réel			
Indice I	108.058	Ecart-type asymptotique de $\hat{I}$	0.517
Biais asymptotique	0.0018	(part de variance due aux n_h)	(12.9 %)
		Pour échantillons indépendants	2.773
		(part de variance due aux n_h)	(0.5 %)
Simulations			
Moyenne $\bar{I}$ des estimations $\hat{I}$	108.120	Ecart-type des 200 estimations $\hat{I}$	0.492
Biais moyen observé $\bar{I} - I$	0.062	Moyenne des écarts-types estimés	0.524
Minimum	106.751	(Ecart-type des écarts-types estimés)	(0.033)
Maximum	109.287		

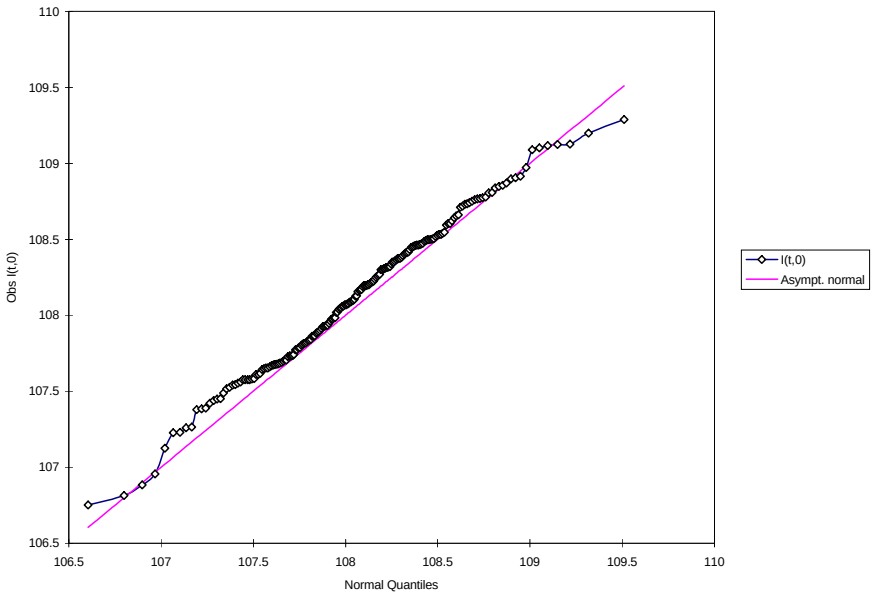

Figure 2. Normal QQ-plot pour les 200 échantillons de taille 100

Tableau 8. Résultats pour 200 échantillons de taille 20

Indice		Ecart-type	
Réel			
Indice I	108.058	Ecart-type asymptotique de $\hat{I}$	1.207
Biais asymptotique	0.0018	(part de variance due aux n_h)	(12.9 %)
		Pour échantillons indépendants	6.470
		(part de variance due aux n_h)	(0.5 %)
Simulations			
Moyenne $\bar{I}$ des estimations $\hat{I}$	108.039	Ecart-type des 200 estimations $\hat{I}$	1.266
Biais moyen observé $\bar{I} - I$	-0.019	Moyenne des écarts-types estimés	1.263
Minimum	104.339	(Ecart-type des écarts-types estimés)	(0.211)
Maximum	111.150		

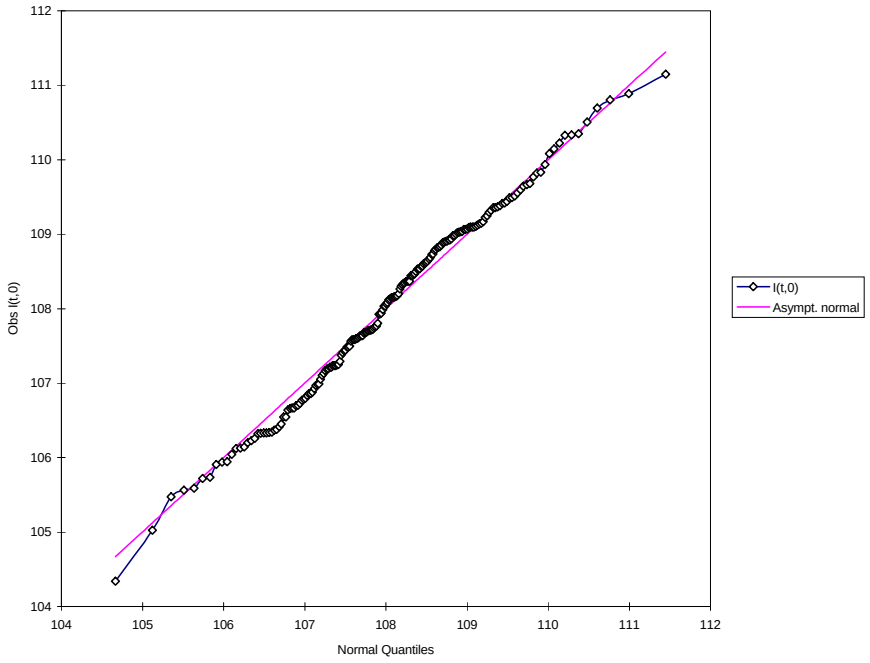

Figure 3. Normal QQ-plot pour les 200 échantillons de taille 20

4. Conclusion

La post-stratification se traduit par des pondérations aléatoires dans l'indice $\hat{I}_{t/0}$ considéré. Les résultats analytiques établis dans Ritschard & Dozio (1995) et qui sont repris dans la première partie de cet article, permettent d'isoler sous forme additive la variance de $\hat{I}_{t/0}$ due à l'incertitude quant à ces poids. La formule de variance proposée comprend par ailleurs des termes incluant les covariances entre observations d'une même strate qui reflètent l'effet dû à l'appariement de données. Pour des covariances positives, comme on les observe en général en cas d'appariement, cet effet est réducteur, c'est-à-dire qu'il tend à réduire la variance.

Dans la pratique, il apparaît que l'incertitude sur les tailles des strates n'a qu'un impact modéré sur la variance de $\hat{I}_{t/0}$. Pour les cas simulés à la section 3, l'accroissement de la variance par rapport au cas avec pondérations fixes est par exemple de l'ordre de 15 % (12.9/(100 - 12.9)). Les implications de la covariance induite par les données appariées sont de ce point de vue plus conséquentes. On observe en particulier une réduction de l'ordre de 80 % dans les situations simulées. Notons cependant que si l'on procède à des renouvellements par rotation de l'échantillon, ce qui est en général le cas pour le calcul d'indices de loyers, ce gain de variance s'estompe au cours du temps.

Les simulations présentées à la section 3 ont permis d'illustrer la fiabilité de la formule établie pour la variance. Ils ont également permis de confirmer que le biais reste non significatif même pour de petits échantillons (20 observations réparties en 6 classes). L'examen de la distribution des indices calculés pour les 200 échantillons générés montre, par ailleurs, que l'on peut très raisonnablement exploiter la loi normale pour la construction d'intervalles de confiance.

Annexe : Démonstrations mathématiques

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, $i = 1,2,\ldots,n$, les variables binaires

$$Y_{ih} = \begin{cases} 1 & \text{si } K_i \in \mathbf{K}_\text{h} \\ 0 & \text{sinon} \end{cases} \qquad Y_{is} = \begin{cases} 1 & \text{si } K_i \in \mathbf{K}_\text{s} \\ 0 & \text{sinon} \end{cases} \tag{18}$$

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

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

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

$$\mathrm{E}(Y_{ih} Y_{is}) = p_h p_s \quad \Leftrightarrow \quad \mathrm{Cov}(Y_{ih}, Y_{is}) = 0 \tag{20}$$

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 $E(n_h)E(n_s) = n^2 p_h p_s$, on obtient la covariance entre n_h et n_s, soit

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

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 $\mathrm{Var}(n_h)$ 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 :

$$\mathrm{E}(X) = \mathrm{E}_Y(\mathrm{E}(X|Y) \tag{24}$$

$$\mathrm{Var}(X) = \mathrm{Var}_Y\left((\mathrm{E}(X|Y) + \mathrm{E}_Y(\mathrm{Var}(X|Y)\right) \tag{25}$$

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

$$\mathrm{E}(\overline{X}_h|n_h) = \mu_h \tag{26}$$

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

Comme μ_h ne dépend pas de n_h ,on a alors $\mathrm{Var}(\overline{X}_h) = \mathrm{E}_{n_h}\left(\mathrm{Var}(\overline{X}_h|n_h)\right)$. 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

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

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 $\left(\mathrm{E}(\overline{X}), \mathrm{E}(\overline{Y})\right) = (\mu_x, \mu_y)$, soit, en notant $\operatorname{grad} g$ 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,

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

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 $\mathrm{E}\left((\overline{X}/\overline{Y})^2\right)$ 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 $1/n$, $\mathrm{Var}(\overline{X}/\overline{Y}) = (\mathrm{E}\left((\overline{X}/\overline{Y} - \mu_x/\mu_y)^2\right)$.

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

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

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

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

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

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

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

avec

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

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

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

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

Comme

$$\mathrm{Var}(\overline{X}_{h0}|n_h) = \frac{\sigma^{*2}_{h0}}{n_h} - \frac{\sigma^{*2}_{h0}}{m_h} \quad (37)$$

$$\mathrm{Cov}(\overline{X}_{h0}, \overline{X}_{ht}|n_h) = \frac{\sigma^{*}_{h0t}}{n_h} - \frac{\sigma^{*}_{h0t}}{m_h} \quad (38)$$

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

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

Comme $\mathrm{E}(n_h) = np_h$ et $m_h = mp_h$, $\mathrm{E}(\hat{I}_{t/0})$, qui est l'espérance de (34), vaut

$$\mathrm{E}(\hat{I}_{t/0}) = \sum_h p_h r_h + \left(1 - \frac{n}{m}\right)\frac{1}{n}\sum_h r_h b^*_h$$

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

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

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

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

soit, en tenant compte du Lemme 1,

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

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

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

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

$$n_h^2\,\mathrm{Var}\left(\frac{\overline{X}_{ht}}{\overline{X}_{h0}} | n_h\right) = n_h\beta_h^* - n_h^2\frac{\beta_h^*}{m_h}. \qquad (43)$$

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

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

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

d'où

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

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

Références

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

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

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.

RITSCHARD, G. ET A. DOZIO (1995), Calcul d'un indice des loyers par post-stratification, *Revue suisse d'économie politique et de statistique*, 131(4), à paraître.