PANEL SOCIO-ECONOMIQUE

"Liewen zu Lëtzeburg"

Document PSELL n° 65 octobre 1994

La Variance des Estimateurs d'un Panel Ménage

La Méthode des Groupes Aléatoires appliquée au Panel Luxembourgeois

Marlis Riebschläger

CEPS/Instead
WALFERDANGE
Grand-Duché de Luxembourg

1994

Document produit par le :

CEPS/Instead

CENTRE D'ETUDES DE POPULATIONS, DE PAUVRETE
ET DE POLITIQUES SOCIO-ECONOMIQUES

B.P. 65, WALFERDANGE
tét. (352) 33 32 33 - 1 Fax. 33 27 05

Président: Gaston Schaber

ISBN 2-87987-029-1

Sommaire

1 Introduction

Le calcul de la variance d'une estimation quelconque est nécessaire pour différents raisons: La connaissance de la variance est nécessaire pour examiner, par exemple, si la différence qu'on a obtenue dans un échantillon est assez grande pour en conclure qu'il y a une différence dans la population à partir de laquelle l'échantillon a été tiré. Pour pouvoir faire des tests ou calculer des intervalles de confiance, il faut donc connaître la variance de l'estimateur examiné. Une autre application de la variance est le jugement sur l'efficacité des estimateurs. S'il y a deux estimateurs sans biais pour le même paramètre, par exemple, on préfèrera celui qui est le plus efficace, c'est-à-dire celui qui a la variance la plus petite.

2 Composition de la variance dans un panel

Une variable quelconque Y caractérisant une population de taille N peut être décrite de la manière suivante:

$$Y = \sum_{i=1}^{N} Y_i. \tag{1}$$

Pour un échantillon aléatoire de taille n, l'estimation peut s'écrire:

$$\hat{Y} = \sum_{i=1}^{n} Y_i w_i, \text{ où } w_i \text{ est le poid de l'individu c.-à-d. du ménage } i^1. \tag{2}$$

Pour calculer la variance de l'estimation dans un échantillon d'un panel ménage, il faut prendre en considération les dépendances des réponses. Dans le cas des analyses longitudinales, il est évident que l'indépendance des réponses n'est pas assurée puisqu'il s'agit des mêmes individus ou ménages.

Mais il en va de même pour les analyses en coupe dès lors qu'on examine les réponses des individus appartenant à un même ménage; dans un tel cas de figure, la suspicion du caractère "dépendant" des réponses sera encore renforcée pour toutes les questions subjectives.

Dans le cas des analyses longitudinales, on peut calculer des changements au cours du temps, par exemple, à l'aide des différences entre observations, si c'est raisonnable (ce qui n'est pas toujours le cas)[2]. Mais en adoptant cette solution, on n'inclut que les individus ou ménages qui ont participé et répondu aux périodes examinées; dans ce cas, on perd évidemment quelques observations (les individus / ménages qui n'ont pas répondu pour toutes les années analysées).

Dans le cas des analyses en coupe, on peut essayer d'estimer les corrélations des réponses. Pour faire cela, il est utile de considérer le processus de la sélection des ménages comme aléatoire, et de regarder les caractéristiques des ménages et des individus comme des valeurs fixes. De cette façon, on pourrait évaluer les estimations des variances et des corrélations sur base des probabilités de sélection et de l'attrition dans les vagues suivantes. Mais le calcul analytique ou même approximative de la probabilité d'appartenir à l'échantillon devient de plus en plus difficile en raison du processus complexe du developement des relations dans un panel.

[1] Si w_i est l'inverse de la probabilité de sélection de l'individu c.-à-d. du ménage (i $w_i = \frac{1}{\pi_i}$), l'estimateur est sans biais.

[2] voir l'exemple d'application ci-dessous

Il y a cependant une méthode alternative et simple pour calculer la variance d'une estimation dans le cadre d'une analyse en coupe ou longitudinale. Cette méthode – la méthode des groupes aléatoires – a été proposée par Wolter (1985) et appliquée par Rendtel (1991) aux données du panel ménage allemand (SOEP).

3 La méthode des groupes aléatoires

L'idée de la méthode des groupes aléatoires est la suivante:
Si l'on a plusieurs (R) échantillons indépendants, tirés de la même population, on peut en calculer une estimation $\hat{Y}_r, r = 1, \ldots, R$ pour chaque échantillon. A base de ces R observations indépendantes et distribuées identiquement, on peut estimer la dispersion des données.

La situation du panel ménage allemand était différente. L'échantillon était le résultat d'une seule expérimentation, ce qui fut aussi le cas au Luxembourg. Mais si l'on crée une partition aléatoire de l'échantillon initial, on peut considérer les sous-échantillons comme des répliques quasi-indépendantes de l'expérimentation de tirage.[3]

Pour la création des sous-échantillons il est important de garder les caractéristiques d'un panel. Comme l'échantillon initial du panel luxembourgeois est un échantillon aléatoire simple des ménages, il suffit de diviser cet échantillon initial en R sous-échantillons. De la sorte les personnes qui quittent un ménage afin de fonder leur propre ménage, restent dans le même groupe. Du fait des évolutions différentes, il n'est pas garanti qu'on garde le même nombre de ménages dans les sous-échantillons, mais on peut postuler que les nombres ne sont pas très différents.

Enfin il faut décider combien de groupes on veut créer. De l'un coté, il vaut mieux créer beaucoup de groupes afin d'estimer la variance au base d'un grand nombre d'observations quasi-indépendantes; de l'autre coté, les groupes ne doivent pas être trop petits. Le panel allemand a été divisé en 8 groupes. Cette décision a été prise avant tout à cause de la méthode de tirage de l'échantillon, qui comprend 2 stages. Après avoir tiré les unités premières sur base d'une partition régionale, on a tiré 8 ménages en moyenne dans chaque unité première. Il a été démontré que le chiffre '8' était utilisable pour estimer la variance.

Au Luxembourg la méthode de tirage était plus simple, de façon que le nombre n'est pas impliqué par la méthode de tirage. Mais bien que le nombre de ménages soit plus faible dans le panel luxembourgeois qu'en Allemagne, le nombre de groupes ne doit pas être inférieur à 8, afin d'avoir assez d'observations pour estimer la variance.

Sur base des R sous-échantillons G_r, on obtient R estimations $\widehat{Y}_r, r = 1, \ldots, R$ à partit desquelles on peut calculer un intervalle de confiance pour le paramètre Y par la statistique de l'ordre de la manière suivante:
$\widehat{Y_{(1)}}$ soit l'estimation la plus petite et $\widehat{Y_{(R)}}$ soit l'estimation la plus grande. Si l'on postule[4] que la médiane de la distribution des $\widehat{Y}_r$ est égale à Y, les intervalles de confiance se présentent de la manière suivante (Büning,Trenkler, 1978) :

$$P(\widehat{Y_{(1)}} \leq Y \leq \widehat{Y_{(8)}}) = 1 - 0.008$$

[3]Elles ne sont pas vraiment indépendantes, parce qu'on ne peut affecter les individus qu'à $R-1$ groupes indépendemment, en cas de R groupes.

[4]cela veut dire qu'on peut estimer Y par la médiane des observations.

$$P(\widehat{Y_{(2)}} \leq Y \leq \widehat{Y_{(7)}}) = 1 - 0.07 \tag{3}$$

$$P(\widehat{Y_{(3)}} \leq Y \leq \widehat{Y_{(6)}}) = 1 - 0.29.$$

La deuxième équation rend l'intervalle le plus proche du niveau de 95 %. Alors on peut accepter comme règle, que la valeur immédiatement inférieure à la valeur maximum et la valeur immédiatement supérieure à la valeur minimum rendent un intervalle qui couvre la caractéristique Y avec une probabilité de 93 %

$$CI_{93\%} = CI_R := [\widehat{Y}_{(2)}, \widehat{Y}_{(7)}].$$

Comme les groupes sont par définition approximativement de la même taille, Y peut être estimé par

$$\bar{\hat{Y}} = \frac{1}{R} \sum_{r=1}^{R} \hat{Y}_r.$$

Par conséquent, l'écart-type peut être estimé par

$$\widehat{\sigma_R} = \sqrt{\frac{1}{R(R-1)} \sum_{r=1}^{R} (\widehat{Y}_r - \bar{\hat{Y}})^2} = \sqrt{\frac{1}{8 \cdot 7} \sum_{r=1}^{8} (\widehat{Y}_r - \bar{\hat{Y}})^2}.$$

A l'aide de cet estimateur, on peut calculer un intervalle de confiance habituel :

$$CI_{\widehat{\sigma_R}} := [\, \hat{Y} \pm \hat{\sigma_R} \cdot t_{R-1,1-\frac{\alpha}{2}} \,]$$

$$(\alpha = 0.05, R = 8 \Longrightarrow t_{R-1,1-\frac{\alpha}{2}} = 2.365)$$

Les groupes aléatoires peuvent de même servir à estimer des paramètres selon la méthode **jackknife**. Afin de calculer une estimation selon cette méthode la procédure suivante est appliquée :

- Pour l'estimation $\widehat{Y_r}$ on utilise le complément du groupe numéro r, c.-à-d. toutes les valeurs qui n'appartiennent pas au r-ième groupe[5] $\hat{Y}_{\bar{r}} := \sum_{i \in \bar{G}_r} Y_i w_i$.
- On calcule les pseudo-valeurs[6] $\widehat{Y_{\bar{r}}^{(+)}} := R\hat{Y} - (R-1)\hat{Y}_{\bar{r}}$.
- Enfin l'écart-type est calculé de la manière décrite ci-dessus,

$$\widehat{\sigma_J} = \sqrt{\frac{1}{R(R-1)} \sum_{r=1}^{R} (\widehat{Y_{\bar{r}}^{(+)}} - \bar{\hat{Y}})^2}.$$

[5]Ces valeurs sont très proches de la moyenne générale.

[6]celles-ci sont très proches des valeurs $\widehat{Y_r}$.

4 Un exemple d'application

Dans l'étude luxembourgeoise *Liewen zu Letzebuerg* on a posé des questions qui concernent l'évaluation personnelle de la situation monétaire. On pose la question générale "Dans votre ménage éprouvez-vous des difficultés à joindre les deux bouts?" et des questions plus détaillées; il s'agit des difficultés pour payer l'eau, le gaz, l'électricité, le chauffage, l'alimentation, le medecin et les vêtements.

Comme les réponses à ces questions-ci sont très proches, les résultats de la question suivante sont présentés: *Au cours des douze derniers mois, vous est-il arrivé d'avoir des difficultés pour payer les vêtements ?*

On compare d'abord le taux des ménages (c.-à.d. $Y = p$) qui ont des difficultés dans le groupe des ménages sans enfant avec le taux correspondant dans le groupe des ménages avec des enfants. Ensuite, on examine si le taux des ménages avec des difficultés demeure stable entre 1987 et 1990.

Dans le premier cas, il s'agit d'une comparaison de deux groupes indépendants[7], donc on peut faire un test conventionel. La méthode des groupes aléatoires est appliquée en plus afin de pouvoir comparer les deux.

Dans le deuxième cas, il s'agit de deux groupes dépendants . Il est évident que les ménages avec des enfants en 1987 ne sont pas précisément les mêmes en 1990, mais l'intersection est très grande. Mais grâce à la méthode des groupes aléatoires, on peut estimer la variance de la différence des taux des difficultés, en calculant ces différences dans les groupes aléatoires[8]. Comme les réponses dans les catégories originales de la question générale sont très faibles, celles-ci ont été regroupées de la manière suivante:

énormément de diff.
beaucoup de diff.
assez de diff.
pas trop de diff. } avec des diff. plus graves (1)

peu de diff.

aucune diff. } avec des diff. (2)

Comme la catégorie *peu de difficultés* peu être considerée comme *presque pas de difficultés* d'un côté et comme *avec des difficultés* de l'autre côté, elle a été regroupée de deux manières. En effet, les résultats sont différents. Les tableaux 1 et 2 et les représentations graphiques 1.1 et 1.2 les présentent.

Les catégories de réponses à la question qui concerne le paiement des vêtements ont été regroupées d'une manière analogue. Les tableaux 3 et 4 et les représentations graphiques 2.1 et 2.2 présentent ces résultats-là [9].

[7] en négligeant les relations potentielles entre les ménages dans un groupe, ce qui semble être raisonnable dans ce cas

[8] ou selon la méthode jackknife; comme les résultats qu'on obtient par les deux méthodes sont très proches, seulements ceux des groupes aléatoires sont présentés

[9] Comme le calcul des résultats a été réalisé de la même manière que dans les tableaux 1 et 2, les tableaux 3 et 4 sont présentés de façon moins détaillée.

Notation:

n : taille du groupe
$\hat{p}$: taux de ménages

$$\widehat{\sigma_{class}} = \widehat{\sigma_{\text{Diff}}} = \sqrt{\text{Var}(\hat{p}_2 - \hat{p}_1)}$$

$$= \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

$$\widehat{\sigma_R} = \sqrt{\frac{1}{8 \cdot 7} \sum_{r=1}^{8} (\widehat{p_r} - \widehat{p})^2}$$

$$\widehat{CI_{class}} = (\hat{p} \pm \widehat{\sigma_{\text{Diff}}} \cdot t_{n_1+n_2-2,1-\frac{\alpha}{2}})$$

$$CI_R = [\hat{p}_{(2)}, \hat{p}_{(7)}]$$

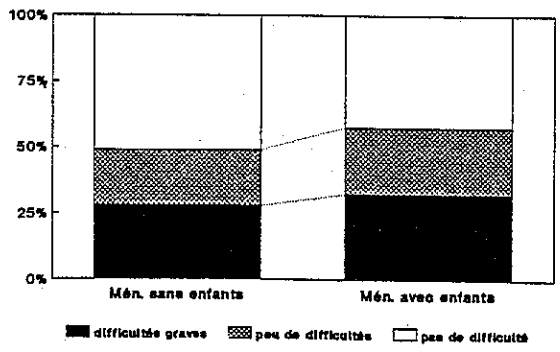
$$CI_{\widehat{\sigma_R}} = [\, \hat{p} \pm \hat{\sigma_R} \cdot t_{R-1,1-\frac{\alpha}{2}} \,]$$

signification : $\begin{cases} + & \text{si l'intervalle couvre zéro} \\ - & \text{sinon} \end{cases}$

Groupe aléatoire	1987 sans enfants	1987 avec enfants	1987 Différ.	1990 sans enfants	1990 avec enfants	1990 Différ.	Différence 1987-1990 sans enfants	Différence 1987-1990 avec enfants
1	23	44	21	19	43	24	4	1
2	26	33	7	15	25	10	11	8
3	32	38	6	18	32	14	14	6
4	27	34	7	20	17	-3	7	17
5	35	36	1	26	26	0	9	10
6	29	20	-9	28	29	1	1	-9
7	31	22	-9	22	25	3	9	-3
8	24	31	7	16	29	13	8	2
n	980	656		1041	632			
$\hat{p}$	28	32	3.9	20	28	7.8	7.9	4
$\widehat{\sigma_{class}}$			2.32			2.17		
$\widehat{\sigma_R}$			3.23			3.01	1.33	2.7
CI_{class}			$[-0.6, 8.4]$			$[3.5, 12.1]$		
CI_R			$[-9.0, 7.0]$			$[0.0, 14.0]$	$[4.0, 17.0]$	$[-3.0, 10.0]$
$CI_{\widehat{\sigma_R}}$			$[-3.8, 11.5]$			$[0.6, 14.9]$	$[4.8, 11.1]$	$[-2.4, 10.4]$
signifiance			$-$			$+$	$+$	$-$

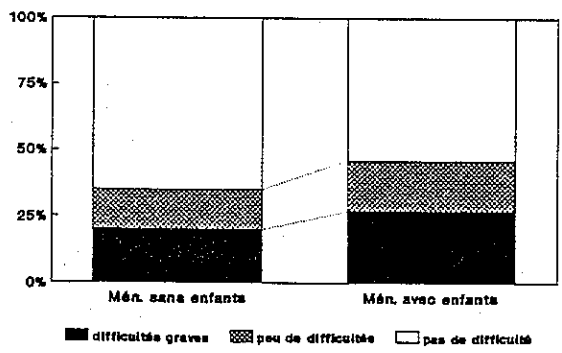
Tableau 1: Taux de ménages *avec des difficultés plus graves* Regroupement(1))

Fig. 1.1 : Difficultés générales 1987



Source : PSELL

Fig. 1.2 : Difficultés générales 1990

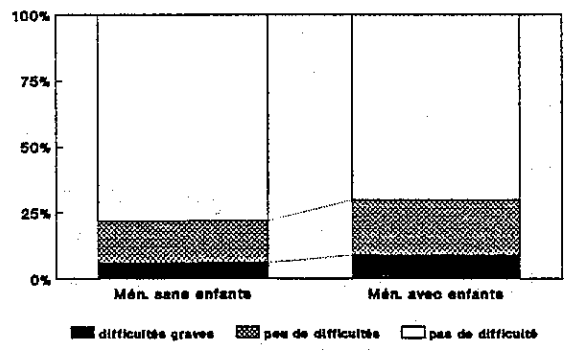


Source : PSELL

Groupe aléatoire	1987 sans enfants	1987 avec enfants	1987 Différ.	1990 sans enfants	1990 avec enfants	1990 Différ.	Différence 1987-1990 sans enfants	Différence 1987-1990 avec enfants
1	38	64	26	31	63	32	7	1
2	51	63	12	30	49	19	21	14
3	51	62	11	33	52	19	18	10
4	44	56	12	33	36	3	11	20
5	55	63	8	50	44	-6	5	19
6	54	52	-2	43	50	7	11	2
7	57	43	-6	34	40	6	23	3
8	44	63	19	27	46	19	17	17
n	980	656		1041	632			
$\hat{p}$	49	58	9	35	47	12.4	14.1	10.8
$\widehat{\sigma_{class}}$			2.50			2.48		
$\widehat{\sigma_R}$			3.42			3.98	2.20	2.80
CI_{class}			$[4.1, 13.9]$			$[7.6, 17.3]$		
CI_R			$[-2.0, 19.0]$			$[3.0, 19.0]$	$[7.0, 21.0]$	$[2.0, 19.0]$
$CI_{\widehat{\sigma_R}}$			$[1.9, 18.1]$			$[3.0, 21.8]$	$[8.9, 19.3]$	$[4.2, 17.4]$
signifiance			$+-+$			$+$	$+$	$+$

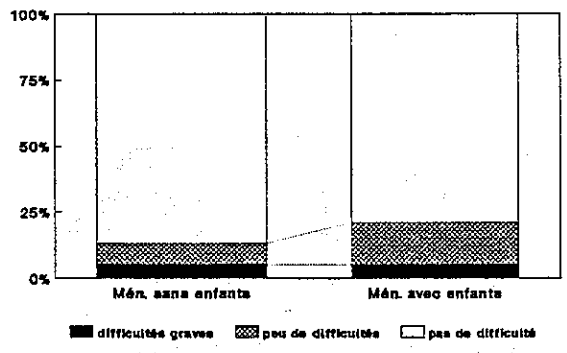
Tableau 2: Taux de ménages *avec des difficultés* (Regroupement(2))

Fig. 2.1 : Diff. p. payer des vêtements 1987



Source : PSELL

Fig. 2.2 : Diff. p. payer des vêtements 1990



Source : PSELL

	1987	1990	Différ. 1987-1990	
	Différence sans — avec enfants	Différence sans — avec enfants	sans enfants	avec enfants
$\hat{p}$	3	0	1	4
$\widehat{\sigma_{class}}$	1.35	1.10		
$\widehat{\sigma_R}$	1.03	1.36	0.96	1.16
CI_{class}	$[-0.4, 5.7]$	$[-2.2, 2.2]$		
CI_R	$[1.0, 6.0]$	$[-3.0, 6.0]$	$[0.0, 3.0]$	$[1.0, 7.0]$
$CI_{\widehat{\sigma_R}}$	$[0.6, 5.4]$	$[-3.2, 3.2]$	$[-0.4, 4.2]$	$[1.3, 6.7]$
signifiance	$-++$	$-$	$-$	$+$

Tableau 3: Taux de ménages *avec des difficultés graves pour payer les vêtements* (Regroupement(1))

4.1 Les résultats théoriques

En examinant le premier cas (comparaison de deux groupes indépendants), on peut conclure que les intervalles de confiance qu'on calcule à l'aide des groupes aléatoires sont moins sensibles[10] que les intervalles classiques, c.-à-d. ils sont plus grands. Il en est de même pour les écarts-type. On peut aussi dire que les intervalles $CI_R = [p_{\hat{(2)}}, p_{\hat{(7)}}]$ sont les plus grands dans presque tous les cas. A l'exception d'un cas (tab. 4), les intervalles CI_{class} et CI_{σ_R} conduisent à des résultats identiques[11]. Mais dans la plupart des cas, les trois intervalles incluent ou excluent tous la valeur zéro.

On peut en conclure que les intervalles $\mathbf{CI_{\sigma_R}} = [\hat{\mathbf{p}} \pm \hat{\sigma_R} \cdot \mathbf{t_{R-1,1-\frac{\alpha}{2}}}]$ sont à préférer parce qu'on peut les calculer dans tous les cas, mais ils sont plus sensibles que les intervalles $\mathbf{CI_R} = [\mathbf{p}_{\hat{(2)}}, \mathbf{p}_{\hat{(7)}}]$.

4.2 Les résultats

En ce qui concerne les différences générales on peut dire que les ménages sans enfants indiquent moins de difficultés que les ménages avec des enfants. En 1990 la différence est significative selon chacune des trois méthodes et selon les deux versions de regroupement. En 1987 la différence entre les taux des ménages avec *des difficultés plus graves* (version 2) n'est pas assez grande pour être significante selon aucune méthode, mais l'inclusion de la catégorie *peu de difficultés* (version 1) produit néamoins une différence significative.

[10] à l'exception de la première comparaison, tab. 3)

[11] in- ou exclusion du zéro

	1987	1990	Différ. 1987-1990	
	Différence sans — avec enfants	Différence sans — avec enfants	sans enfants	avec enfants
$\hat{p}$	8	7	9	10
$\widehat{\sigma_{class}}$	2.23	1.90		
$\widehat{\sigma_R}$	2.64	2.05	1.43	2.34
CI_{class}	[3.6, 12.4]	[3.3, 10.7]		
CI_R	[−2.0, 16.0]	[4.0, 12.0]	[6.0, 16.0]	[4.0, 15.0]
$CI_{\widehat{\sigma_R}}$	[1.8, 14.2]	[2.2, 11.9]	[6.1, 12.9]	[3.8, 14.8]
signifiance	+ − +	+	+	+

Tableau 4: Taux de ménages *avec des difficultés pour payer les vêtements* (Regroupement(2))

Les chiffres concernant la question du paiement des vêtements sont un peu différents de ceux relatifs aux difficultés générales. Il faut remarquer d'abord que les taux des difficutés sont plus bas que ceux des difficultés générales.

Les indications des difficultés plus graves diffèrent très peu en 1987 et pas du tout en 1990. Mais les taux des ménages qui ont indiqué avoir *peu de difficultés* sont différents. Il en resulte des différences significatives selon le regroupement (2).

Les comparaisons des taux entre 1987 et 1990 montrent qu'en général les difficultés ont diminué [12]. Ces résultats coïncident avec ceux d'une analyse d'évolution des revenus des ménages entre 1985 et 1990 [Hausman], qui constate une augmentation générale des revenus au cours de cette période.

Bibliographie

- Büning, H., G.Trenkler, 1978: Nichtparametrische statistische Methoden, Berlin.
- Hausman, Pierre, 1994: Evolution des revenus des ménages 1985 - 1990, Document PSELL 56
- Rendtel, Ulrich, 1991: Die Schätzung von Populationswerten in Panelerhebungen. Allgemeines Statistisches Archiv, 75, S 280 - 299
- Wolter, Kirk, 1985: Introduction to Variance Estimation. Springer-Verlag. New York.

[12] à l'exception du cas : ménage sans enfant, difficultés graves pour payer les vêtements; ici le taux en 1987 était déjà très faible (6 %)