Découpage optimal d'une variable quantitative pour la stratification





Une expérience sur les données d'entreprises françaises - Arnaud Fizzala

Application aux Enquêtes Mensuelles de Branches - Laurent Costa

Insee
Mesurer pour comprendre

Une expérience sur les données d'entreprises françaises

Présentation de l'étude

Les échantillons des enquêtes auprès des entreprises menées par l'Insee sont le plus souvent tirés selon des plans de sondage aléatoires simples stratifiés par activité et tranche de taille.

On utilise généralement 7 tranches d'effectif dites « classiques » (10 à 19, 20 à 49, 50 à 249, 250 à 499, 500 à 999, 1 000 à 4 999, 5 000 et +).

L'étude consiste à trouver un meilleur découpage en tranches et à évaluer les gains de précision produits.

On se base sur une extraction Sirus des unités légales marchandes, exploitantes de 10 salariés ou plus.

Objectif et cadre de l'étude

L'objectif est de découper la variable effectif salarié en sept tranches de façon à minimiser la taille d'échantillon nécessaire pour atteindre une précision fixée a priori de l'estimateur de l'effectif salarié de l'ensemble des unités.

Nous nous plaçons dans un cadre simplifié.

- Nous ne distinguons pas les différents secteurs d'activité des entreprises;
- Nous ne tenons pas compte des phénomènes de non réponse pour nos calculs;
- Nous nous limitons au cas où la variable de stratification est confondue avec la variable d'intérêt.

Description des données

Tranche d'effectif	Nombre d'unités	Somme des effectifs	Dispersion des effectifs
10 à 19	109 232	1 455 314	2,8
20 à 49	64 673	1 980 756	8,4
50 à 99	18 704	1 267 209	14,0
100 à 249	10 468	1 599 923	41,0
250 à 499	3 316	1 138 863	68,5
500 à 999	1 483	1 013 769	135,5
1 000 à 4 999	1 084	2 069 807	902,9
5 000 et plus	129	1 864 320	20 433,1
Total	**209 089**	**12 389 961**	**641,5**

Forte asymétrie et dispersion croissante avec le nombre de salariés.

Les méthodes de découpage optimal de strates

3 méthodes implémentées dans le package R *stratification* à disposition sur internet :

2 méthodes « théoriques » (hypothèses : taux de sondage négligeables et distribution uniforme de la variable de stratification dans chaque strate)

- Dalenius (basée sur la distribution);
- Géométrique (basée sur les valeurs extrêmes);

1 méthode « empirique »

- LH : on teste un grand nombre de limites de strate et on conserve les meilleures.

Nous appliquons ces méthodes sur nos données.

Résultats – méthodes de découpage optimal

Cv	Sondage stratifié classique	Sondage stratifié méthode Dalenius	Sondage stratifié méthode géométrique	Sondage stratifié méthode LH
1%	666	615	611	602
2%	272	257	255	251
3%	195	188	187	184
4%	168	164	163	160
5%	156	152	151	150
6%	149	146	145	144
7%	144	144	143	140
8%	140	141	140	137
9%	140	137	138	137
10%	138	137	136	136

Optimisation du seuil d'exhaustivité avec la méthode LH

Pour établir les résultats précédents, nous avons imposé que les entreprises de 5 000 salariés ou plus soient tirées d'office dans l'échantillon.

Si l'on souhaite intégrer une strate exhaustive sans a priori sur le seuil d'exhaustivité, on peut, pour la méthode LH, imposer que la $L^{ème}$ strate soit tirée exhaustivement.

L'algorithme fonctionne comme précédemment, mais au moment du calcul de l'allocation, il impose $n_L=N_L$.

Optimisation du seuil d'exhaustivité avec la méthode LH

Cv	classique	méthode LH sans optimisation seuil exhaustivé	méthode LH avec optimisation seuil exhaustivé	Seuil d'exhaustivité optimal	N_h dans la strate exhaustive optimale
1%	666	602	601	4 760	142
2%	272	251	201	11 774	32
3%	195	184	105	16 055	23
4%	168	160	70	22 868	16
5%	156	150	51	32 712	13
6%	149	144	36	87 294	2
7%	144	140	28	87 294	2
8%	140	137	22	87 294	2
9%	140	137	18	87 294	2
10%	138	136	15	87 294	2

Le nombre de strates : une marge de gain importante

Dans ce qui précède, nous nous sommes limités à sept strates. Nous proposons d'améliorer nos résultats en optimisant le nombre de strates :

Nous calculons pour L (nombre de strates) variant de 2 à 50 la taille d'échantillon n_L permettant d'atteindre l'objectif de précision fixé a priori.

Nous retenons le L_{opt} qui permet d'atteindre l'objectif de précision avec la plus petite taille d'échantillon n_{Lopt}.

Le nombre de strates : une marge de gain importante

Cv	Classique	LH avec 7 strates	LH avec L_{opt} strates	L_{opt}
1%	666	601	60	39
2%	272	201	37	29
3%	195	105	26	18
4%	168	70	21	18
5%	156	51	18	16
6%	149	36	16	12
7%	144	28	14	11
8%	140	22	13	11
9%	140	18	12	10
10%	138	15	11	9

Discussion – La variable de stratification n'est que corrélée à la variable d'intérêt

Nous avons testé les plans de sondage précédents sur des variables corrélées à l'effectif salarié.

Stratification	Taille d'échantillon	Effectif total	Effectif N-1 total	Chiffre d'affaires total
Classique (L=7)	666	1,0%	3,1%	16,1%
LH (L=7)	602	1,0%	2,8%	16,0%
LH (L=39)	60	1,0%	8,9%	61,8%

Discussion – autres points

On souhaite souvent garantir que plus d'une unité soit tirée dans chaque strate.

les objectifs de précision sont souvent multiples

- plusieurs variables sont relevées lors de l'enquête
- pour une même variable on produit plusieurs estimations en fonction de nos domaines d'intérêt

La présence de non-réponse rendra moins précises les estimations

Conclusion

Les trois méthodes conduisent à des résultats assez proches tant que le seuil d'exhaustivité est fixé à 5 000 salariés.

La méthode LH peut donner de meilleurs résultats car elle permet d'optimiser le seuil d'exhaustivité.

En jouant sur le nombre de strates, on peut encore améliorer les résultats, mais le plan de sondage devient alors très (trop?) lié aux données.

Il sera donc important de contrôler l'optimisation des limites de strates par des calculs de précision sur d'autres variables que la variable de stratification.

Application aux Enquêtes Mensuelles de Branches

Les Enquêtes Mensuelles de Branches

Mesurer l'évolution d'un mois sur l'autre de la production industrielle française par produit.

Permettent de constituer les séries témoin de l'Indice de la Production Industrielle (IPI).

La base de sondage des EMB contient pour chaque produit la liste des entreprises qui le fabriquent (strates exhaustives de l'EAP)

L'objectif de l'étude était double :

- optimiser, sur les 13 produits fabriqués par plus de 200 entreprises le plan de sondage actuel (cut off)
- étudier l'effet d'une extension du champ sur les résultats des enquêtes de branche sur 4 produits intermédiaires

Critères d'optimisation

Nous avons cherché à obtenir le meilleur coefficient de variation possible dans l'estimation de la facturation totale de chaque produit, en travaillant à taille d'échantillon égale à l'an dernier.

Comparaison sur 100 simulations la performance de différents plans de sondage.

$$CV_y = \frac{\frac{1}{100} \times \sum_{i=1}^{100} \sigma_i}{\frac{1}{100} \times \sum_{i=1}^{100} t_y^i} \qquad V(t_y^i) = \sigma_i^2 = \sum_h N_h^2 (1 - \frac{n_h}{N_h}) \frac{S_h^2}{n_h}$$

La base de sondage est d'abord divisée

- en une strate exhaustive, qui aurait été obtenu avec un *cut-off* à x %, avec x variable ;
- le reste de l'échantillon est sélectionné suivant différentes procédures (sans stratification, stratification par quartiles et stratification optimisée)

On dira qu'une approche est plus précise qu'une autre lorsque le coefficient de variation de celle-ci est plus faible

Première approche

Sélection d'un échantillon de taille similaire à 2014. Par cutoff jusqu'à un certain seuil puis tirage aléatoire simple sans stratification du produit.

serie	25%	50%
1624ZXR0	0,0307	0,0213
1812Z1199010	0,0404	0,0268
1812Z2125000	0,0172	0,0113
2229A2R1	0,0196	0,0140
2511Z1235040	0,0171	0,0119
2511Z2R010	0,0345	0,0232
2511Z3R011	0,0411	0,0268
2512Z1R0	0,0300	0,0194
2550B1R0	0,0261	0,0175
2561Z1R0	0,0261	0,0173
3312Z5R1	0,0555	0,0357
3320A111003A	0,0501	0,0323
3320C1R0	0,0740	0,0469

serie	25%	50%	65%
2223Z2145000	0,05132	0,02225	0,01614
2399Z1131000	0,02583	0,02008	
2573A1R0	0,02774	0,01950	0,01499
3250A1R2	0,04684	0,02574	

Approche avec stratification simple

Sélection d'un échantillon de taille similaire à 2014. Par cutoff jusqu'à un certain seuil puis tirage aléatoire simple en stratifiant la partie non exhaustive de chaque produit en fonction des quartiles de la distribution de ses facturations .

Serie	25%	50%
1624ZXR0	0,0246	0,0204
1812Z1199010	0,0293	0,0245
1812Z2125000	0,0127	0,0102
2229A2R1	0,0161	0,0133
2511Z1235040	0,0134	0,0113
2511Z2R010	0,0267	0,0218
2511Z3R011	0,0300	0,0244
2512Z1R0	0,0208	0,0175
2550B1R0	0,0200	0,0164
2561Z1R0	0,0208	0,0166
3312Z5R1	0,0363	0,0310
3320A111003A	0,0371	0,0299
3320C1R0	0,0542	0,0438

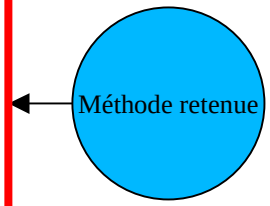
serie	25%	50%	65%
2223Z2145000	0,02187	0,01887	0,01583
2399Z1131000	0,02268	0,01954	
2573A1R0	0,01872	0,01643	0,01439
3250A1R2	0,03322	0,02302	

Approche avec stratification optimisée

Sélection d'un échantillon de taille similaire à 2014. Par cutoff jusqu'à un certain seuil puis tirage aléatoire simple en stratifiant la partie non exhaustive de chaque produit selon la méthode de Dalenius .

serie	NAT	25%	50%
1624ZXR0	0,024364	0,024431	0,020234
1812Z1199010	0,028073	0,028033	0,024320
1812Z2125000	0,012334	0,012314	0,010234
2229A2R1	0,016451	0,015983	0,013313
2511Z1235040	0,013240	0,013163	0,011258
2511Z2R010	0,024769	0,024691	0,021699
2511Z3R011	0,027428	0,027471	0,023645
2512Z1R0	0,019114	0,019173	0,017189
2550B1R0	0,018002	0,018204	0,016229
2561Z1R0	0,019520	0,019602	0,016511
3312Z5R1	0,031929	0,032311	0,030134
3320A111003A	0,032252	0,031108	0,029457
3320C1R0	0,047693	0,047133	0,043289

serie	NAT	25%	50%
2223Z2145000	0,0208239	0,0205830	0,0185661
2399Z1131000	0,0232168	0,0229327	0,0194501
2573A1R0	0,0180270	0,0181649	0,0164007
3250A1R2	0,0263711	0,0263276	0,0222579



Comparaison des précisions des méthodes

Nous cherchons ainsi à estimer la variation du chiffre d'affaires dans le produit déclaré par l'ensemble des entreprises de l'EAP entre 2011 et 2012.

Nous comparons deux estimateurs de cette différence :

- le premier est obtenu si les productions du produit en 2011 et 2012 sont issues à chaque fois d'un échantillon sélectionné par *cut-off* ;

- le second est obtenu si les facturations du produit chacune des deux années sont estimées par un échantillon sélectionné par sondage aléatoire simple

L'EQM de la méthode du *cut-off* sur les estimateurs des années d'enquête 2013 et 2014 prend en compte seulement le biais car la variance est nulle

$$\text{biais} = \left(\left(t^{2012}_{y,exhaustive} - t^{2011}_{y,exhaustive} \right) - \left(t^{2012}_{y,cutoff} - t^{2011}_{y,cutoff} \right)\right)^2$$

Dans la méthode du sondage stratifié par quartiles, l'EQM ne présente pas de biais mais une variance de la différence des estimateurs 2012 et 2011

En supposant indépendance des tirages et des variances environ égales :

$$EQM = V(y_{2012} - y_{2011}) = V(y_{2012}) + V(y_{2011}) \approx 2\, V(y_{2012}) = 2\, (\hat{y}_{2012}\, CV)^2$$

Comparaison des précisions des méthodes

On peut alors en dégager un intervalle de confiance à 95 % près de la précision de l'estimateur avec la nouvelle méthode qui nous permet de dire si le biais engendré par la méthode du cutoff est tolérable en comparaison aux résultats de la méthode par tirage stratifié non optimisé

série	Taux de croissance	Biais du taux de croissance	Intervalle de confiance du taux de croissance
2223Z2145000	-0,10564	-0,04920	[-0,134483 ; -0,076797]
2399Z1131000	0,20077	0,08639	[0,163283 ; 0,238257]
2573A1R0	-0,15638	-0,05492	[-0,180773 ; -0,131987]
3250A1R2	0,01240	0,00445	[-0,023556 ; 0,048356]

Conclusion

Les choix de la stratification et du seuil jouent beaucoup sur la précision obtenue.

Meilleure approche sur la précision : stratification optimale avec un seuil de couverture à 50%

Ce gain de précision par cette nouvelle méthode est traduit par les calculs des intervalles de confiance sur la variation des totaux du chiffre d'affaires entre les années 2011 et 2012.

Merci de votre attention !

Insee
18 bd Adolphe-Pinard
75675 Paris Cedex 14
www.insee.fr

Informations statistiques :
www.insee.fr / Contacter l'Insee
09 72 72 4000
(coût d'un appel local)
du lundi au vendredi de 9h00 à 17h00

Annexes

La méthode Dalenius – Principe

Dalenius montre que des bornes b_h optimales vérifient :

$$\frac{(b_h - \mu_{yh})^2 + S_{yh}^2}{S_{yh}} = \frac{(b_h - \mu_{yh+1})^2 + S_{yh+1}^2}{S_{yh+1}}$$

Comme S_{yh} et u_{yh} dépendent de la borne b_h, cette équation ne peut pas se résoudre facilement.

Dalenius propose de l'approcher par un algorithme basé sur le cumul des racines carrées des fréquences en introduisant l'hypothèse supplémentaire que les y_k sont repartis selon une loi uniforme sur la strate h.

La méthode Dalenius – Algorithme

On connait la distribution des y_k selon un découpage fin en J classes de mêmes longueurs.

1ère strate = J1 premières classes : $\sum_{j=1}^{J_1} \sqrt{N_j} \approx \frac{\sum_{j=1}^{J} \sqrt{N_j}}{L}$

2ème strate = J_2 classes suivantes : $\sum_{j=J_1+1}^{J_1+J_2} \sqrt{N_j} \approx \frac{\sum_{j=J_1+1}^{J_1+J_2} \sqrt{N_j}}{L}$

Etc...

La méthode géométrique – Principe

D'après l'article de Gunning et Horgan, la méthode est fondée sur une observation de Cochran, selon laquelle, dans le cas de bornes quasi-optimales, les coefficients de variation empiriques sont souvent approximativement les mêmes pour toutes les strates.

La méthode cherche donc à trouver les bornes $b_1,\ldots,b_{L-1}$ telles que :

$$\frac{S_{yh}}{\mu_{yh}} = \frac{S_{yh+1}}{\mu_{yh+1}}$$

La méthode géométrique – Algorithme

Sous l'hypothèse supplémentaire que les y_k sont repartis selon une loi uniforme dans chaque strate h, on a :

$$b_h = b_0 \left(\frac{b_L}{b_0}\right)^{\frac{h}{L}}$$

Pour mettre en œuvre l'algorithme, on trouve la valeur maximale b_L et la valeur minimale b_0 des y_k et on construit nos bornes selon la formule ci-dessus.

La méthode LH (Lavallée et Hidiroglou)

La méthode LH correspond à une adaptation des premières méthodes itératives proposées par Lavallée et Hidiroglou en 1988.

Kozak a proposé cette méthode itérative dans l'article *Optimal stratification using random search method in agricultural surveys* paru en avril 2004 dans la revue *Statistics in transition*.

Voir aussi l'article de Kozak et Verma *Approche de la stratification par une méthode géométrique et par optimisation : Une comparaison de l'efficacité* paru en décembre 2006 dans la revue canadienne Techniques d'enquête.

La méthode LH (Lavallée et Hidiroglou) – Principe

Contrairement aux méthodes précédentes, qui cherchent à approcher l'équation établie par Dalenius, cette méthode est basée sur des itérations.

Elle compare un grand nombre de limites de strates et retient celles qui affichent la plus petite variance pour l'estimation du total de la variable de stratification.

Il n'y a pas d'autres hypothèses simplificatrices (fh=0, loi uniforme des y_k...).

La méthode LH (Lavallée et Hidiroglou) – Algorithme

L'algorithme démarre avec des limites de strates initiales (spécifiées par l'utilisateur ou calculées par le package).

A chaque étape, une limite de strate b_h est sélectionnée aléatoirement et modifiée aléatoirement.

Si le nouvel ensemble de limites de strates est meilleur que le précédent, il remplace ce dernier.

On recommence jusqu'à ne plus obtenir un certain nombre de fois de meilleures limites de strates.

Étude au niveau du seuil naturel

Le seuil naturel représente le taux de couverture des entreprises d'un produit K qui ont un poids de tirage supérieur à 1 : $n \frac{x_i}{\sum_{i \in K} x_i} > 1$

série	seuil naturel (%)	CV sans stratification	CV avec stratification	Taille strate exhaustive
gros produits				
1624ZXR0	16,2395	0,0343	0,0256	4
1812Z1199010	24,1260	0,0409	0,0292	6
1812Z2125000	18,3080	0,0186	0,0128	9
2229A2R1	15,1182	0,0221	0,0166	5
2511Z1235040	20,9834	0,0182	0,0140	10
2511Z2R010	24,6609	0,0347	0,0267	7
2511Z3R011	22,2377	0,0431	0,0299	5
2512Z1R0	25,8502	0,0289	0,0204	9
2550B1R0	28,4156	0,0244	0,0195	11
2561Z1R0	23,0589	0,0267	0,0209	8
3312Z5R1	33,6904	0,0465	0,0345	7
3320A111003A	31,7058	0,0436	0,0350	7
3320C1R0	26,5622	0,0687	0,0534	4
4 produits enquêtés				
2223Z2145000	30,1634	0,0304	0,0212	9
2399Z1131000	16,5333	0,0275	0,0233	4
2573A1R0	22,5012	0,0287	0,0188	7
3250A1R2	24,3507	0,0470	0,0331	7

Tableaux supplémentaires

Dans l'approche avec stratification optimisée, nous avons alors décidé d'aller plus loin pour voir si nous obtenions une amélioration plus significative en augmentant le nombre de strates à 6 pour les quatre produits enquêtés.

serie	NAT	25%	50%
2223Z2145000	0,020787	0,019900	0,018482
2399Z1131000	0,022969	0,022506	0,019206
2573A1R0	0,017771	0,017941	0,016409
3250A1R2	0,025903	0,025887	0,021987

Dans l'étude de la comparaison de la précision des méthodes, nous avons également représenté les différents estimateurs sans forme d'évolution

série	$t_{2012} - t_{2011}$	biais	intervalle de confiance
2223Z2145000	-220 272	-75 179	[-280 415,78 ; -160 128,22]
2399Z1131000	222 877	63 657	[181 261, 68 ; 264 492,32]
2573A1R0	-102 158	-26 262	[-118 093,08 ; -86 222,92]
3250A1R2	19 403	4 934	[-36 855,2 ; 75 661,2]