Estimation anticipée de variance pour des enquêtes sur échantillon-maître

Chantal BRUTEL
Insee

Cet article décrit une méthode permettant d'estimer la précision d'une enquête dès l'instant où l'échantillon est tiré.On peut alors se servir de la procédure informatique de tirage comme d'un outil interactif de mise au point et parvenir ainsi à une meilleure précision de l'échantillonnage.

Lors de chaque Recensement de Population, il est procédé au tirage d 'un Echantillon Maître , c'est à dire une réserve de logements, tirée dans le fichier du Recensement et sur laquelle on dispose de variables pour chaque logement.Il est obtenu de la manière suivante :

Après avoir stratifié le fichier du Recensement en circonscriptions administratives (Unités Urbaines, cantons ruraux .....), on procède à un premier degré de sondage en tirant un échantillon de telles zones: les Unités Primaires. On procède alors au tirage des logements selon un plan de sondage à un ou plusieurs degrés selon les strates.

L'Echantillon Maître est la base dans laquelle sont tirés tous les échantillons d'enquêtes auprès des ménages sur la période intercensitaire(exception faite de l'enquête Emploi).

Un échantillon d'enquête contient donc les mêmes variables que l'Echantillon Maître et le Recensement.On peut alors comparer des totaux extrapolés sur un échantillon à la valeur exacte, et cela, si on le désire au niveau de chaque strate.

Après avoir présenté les estimateurs que l'on peut construire simplement lorsque l ' on dispose de totaux de variables simples sur l'exploitation exhaustive du Recensement, on exposera quelques idées permettant de résoudre des problèmes plus complexes.Ceux ci se présentent lorsque l'on s'intéresse à des variables recodifiées ou simulées ou à des statistiques plus élaborées que de simples totaux de variables.

On s'attardera sur l'une de ces solutions à cause de son intérêt particulier, tant au niveau théorique que pratique.

Enfin on présentera les tests effectués et les résultats auxquels on est arrivés.

I] PREMIERES ESTIMATIONS DE VARIANCE ,LEURS LIMITES.

1) Variance de totaux de variables simples.

- Avant même de chercher à construire des estimateurs de variance , il peut être intéressant et naturel d' avoir une première idée de la qualité de notre échantillon. Ceci est simplement réalisable en comparant le total extrapolé $\hat{X}$ de variables présentes dans la base de l' échantillon, avec le total issu de l'exploitation exhaustive du Recensement: X.On vérifie ainsi si l'échantillon est "bien ou mal tombé".

A partir de là,il est possible de construire un estimateur sans biais de la variance du total extrapolé d'une variable par :

$$(X - \hat{X})^2$$

Cet estimateur n'est bien sûr pas des meilleurs car calculé à un niveau trop global, mais il ouvre des perspectives.

En effet,suivant toujours cette idée de total extrapolé et si de plus on tient compte du fait que l'échantillon est stratifié ,il est possible d'affiner cette première estimation de variance.

Si pour une strate h,on dispose du total de notre variable d'intérêt issu de l'exploitation exhaustive du Recensement: Xh et du total extrapolé de l'échantillon $\hat{X}h$,on peut construire comme ci dessus, un estimateur de la variance du total de X dans la strate h: $(Xh - \hat{X}h)^2$. En sommant sur les strates, on obtient un estimateur de la variance du total de la variable dans notre échantillon:

$$VAR\text{-}Il (\hat{X}) = \sum_h (Xh - \hat{X}h)^2$$ où h est l'indice de strate

2) Variance de fonctions de totaux.

Ce type d'estimateurs permet de résoudre le cas de statistiques un peu plus complexes: celles qui sont des fonctions de totaux déja tabulés auxquelles on est amené à porter un intérêt particulier.Les exemples abondent: proportion des plus de 65 ans dans la population totale, taux de chômage (ratio chomeurs/actifs) etc.Le point est de diposer sur l'exploitation exhaustive du Recensement du total de ces grandeurs au niveau de chaque strate.

Par linéarisation de la fonction nous allons voir comment il est possible de construire une estimation de variance acceptable tant au niveau théorique que pratique.

Soit $S = F(X1,...,Xi,...)$ une statistique qui est une fonction suffisamment régulière des totaux de Q variables. Pour chacune d'elles on dispose des données du Recensement, et en particulier des totaux Xih dans chaque strate. L'échantillon fournit des estimateurs $\hat{X}ih$ de ces totaux ainsi qu'un estimateur $\hat{X}i = \sum_h \hat{X}ih$ de X.

L'estimateur habituel de S est:

$$\hat{S} = F (\hat{X}1,...,\hat{X}i,...)$$

Par linéarisation de F on obtient que :

$$S - \hat{S} \simeq \sum_{i=1}^{Q} Ai (Xi - \hat{X}i)$$

où Ai est la dérivée partielle de F par rapport à Xi prise au point (X1,...,Xi,...) ou éventuellement, au point ($\hat{X}1,...,\hat{X}i,...$).

On voit qu'alors :

$$VAR\text{-}I2 (\hat{S}) \simeq VAR (\sum_{i=1}^{Q} Ai \hat{X}i)$$

Il suffit alors de calculer, au niveau de chaque strate, le total $Yh = \sum Ai\ Xih$ de la variable synthétique obtenue par combinaison linéaire des Xi ainsi que son estimateur sur l'échantillon $\sum_i Ai\ \hat{X}ih = \hat{Y}h$.

La variance de $\hat{S}$ s'estime en additionnant les $(Yh - \hat{Y}h)^2$.

(On pourra se reporter à l'article d'ECONOMIE ET STATISTIQUES de Novembre 1986 : La précision des enquêtes emploi, J.C DEVILLE et N ROTH).

3) <u>Lorsque l'on ne dispose plus des totaux par strate du Recensement</u>.

Si l' on désire connaître la précision d'un échantillon d'enquête pour des variables ou statistiques sur lesquelles on ne dipose pas de totaux par strate sur l'exploitation exhaustive du Recensement , il faut trouver d'autres méthodes d'estimation que celles précédemment présentées. De telles variables peuvent être des variables recodifiées ou des variables simulées aléatoirement. Il est possible d'envisager par exemple :

- nombre de ménages comprenant au moins un enfant entre 7 et 13 ans.
- achats en vêtements,simulés à partir d'un modèle faisant intervenir la taille du ménage, la catégorie socio - professionnelle,le lieu de résidence .....

La partie suivante expose les méthodes envisageables pour traiter de tels cas.

II] LES DOUBLES ECHANTILLONS.

En retournant à l'exhaustif du Recensement et en tabulant par strates nos variables ou statistiques d'intérêt, (quelques 23 millions de logements) on pourrait appliquer les estimations de variance précédentes.La lourdeur évidente de l'opération nous oblige à envisager d'autres solutions plus élégantes et moins coûteuses.

La première idée qui vient à l'esprit s'apparente à la méthode des demi-échantillons.Il s'agit de disposer de deux échantillons indépendants tirés dans l'échantillon maître.

On aurait ainsi pour chacun d' entre eux : $\hat{Y}h1$ et $\hat{Y}h2$ deux estimateurs de Yh , ce qui fait que:

$1/2 \left(\hat{Y}h1 - \hat{Y}h2 \right)^2$ estime sans biais la variance de $\hat{Y}h$

En sommant par strate , on obtiendrait l'estimateur suivant:

$$VAR\text{-}II \left(\hat{Y} \right) = 1/2 \sum_h \left(\hat{Y}h1 - \hat{Y}h2 \right)^2$$

Mais tout ceci reste utopique dans la mesure où pour les strates dans lesquelles on tire les logements à plusieurs degrés, on se trouve dans des Unités Primaires qui sont fixes.Ceci induit donc des corrélations entre les deux estimateurs $\hat{Y}h1$ et $\hat{Y}h2$.

On n'estimerait alors que la partie de la variance conditionnelle à la partie fixe de l 'Echantillon Maître.L'autre partie de la variance ne serait pas simple à estimer.

Il faudrait en effet tirer lors de chaque enquête un double échantillon maître dans les strates où on fait un tirage à plusieurs degrés.Pratiquement cette solution n'est guère envisageable étant donnée la lourdeur du tirage d'un Echantillon Maître.

On envisage alors une troisième solution qui fait l'objet à elle seule de la partie suivante.C'est en effet la seule qui au stade actuel de la réflexion, s'avère appliquable, et qui de ce fait a été retenue pour être intégrée dans la chaîne d'échantillonnage du prochain Echantillon Maître.

III] LE DOUBLE ECHANTILLON DE REFERENCE.

1) Estimateur de variance à partir d'un échantillon de référence tiré dans le R.P.

Etant donné le caractère génant sur le plan pratique des solutions précédentes,on envisage de disposer d'un échantillon de référence,qui servirait de façon universelle pour tous les échantillons d'enquête.

Puisque dans cet échantillon on n'a pas l'intention de tirer des enquêtes on peut l' obtenir selon un plan de sondage qui disperse les observations au plan géographique et qui de ce fait assure une plus grande efficacité dans chaque strate.On envisage par exemple un sondage aléatoire simple dans chaque strate ou un sondage systématique sur le fichier du Recensement.Cet échantillon doit être de taille suffisamment grande par rapport à un échantillon d'enquête, mais pas trop pour pouvoir tabuler des variables complexes à faibles coûts.Une taille de l'ordre de 50 000 logements semble correcte.

Pour chaque strate, on dispose alors de l'estimateur du total de nôtre variable d'intérêt dans l'échantillon d'enquête: $\hat{Y}h$ et du total estimé de cette même variable dans l' échantillon de référence: $\hat{Y}h0$.

La quantité $I = \sum_h (\hat{Y}h - \hat{Y}h0)^2$ a une variance égale à :

$$V (I) = \sum_h Vh + Vh0 = VAR (\hat{Y}) + \sum_h Vh0$$

Si on dipose d'un estimateur $\hat{V}0$ de la variance dans l'échantillon de référence , $I - \hat{V}0$ estime sans biais la variance dans notre échantillon d' enquête pour la variable d'intérêt.

$$VAR\text{-}III1 (\hat{Y}) = \hat{V}0 - \sum_h (\hat{Y}h - \hat{Y}h0)^2$$

Cet estimateur bien que plaisant comporte un inconvénient dans la mesure où il dépend de l'estimation de la variance dans l' échantillon de référence, en laquelle on peut n'avoir qu'une confiance relative.

2) Estimateur basé sur deux échantillons de référence indépendants.

Il est possible d'améliorer considérablement l'estimation de variance présentée ci-dessus si, au lieu de tirer un échantillon on en tire deux.

Ces deux échantillons, tirés tous deux dans le fichier du Recensement ,doivent l'être de manière indépendante et selon le même plan de sondage stratifié.Compte tenu de la remarque précédemment faite il suffit de tirer ces deux échantillons selon un plan de sondage aléatoire simple dans chaque strate ou un sondage systématique avec le même nombre de logements dans chaque strate.(une taille de 25000 logements par échantillon semble convenable).

On dispose alors dans le premier échantillon de référence d'un estimateur du total de la variable d'intérêt $\hat{Y}o$, ainsi que d'un estimateur de ce même total dans le deuxième échantillon de référence $\hat{Y}o'$.

$\hat{Y}o - \hat{Y}o'$ estime 0 avec une variance égale à 2 Vo

(en effet, étant donnée la manière dont ces échantillons ont été tirés leurs variance sont égales).

et $\hat{Y} - 1/2 (\hat{Y}o + \hat{Y}o')$ estime 0 avec la variance V + Vo/2

il en résulte donc que :

$[\hat{Y} - 1/2 (\hat{Y}o + \hat{Y}o')]^2 - [1/2 (\hat{Y}o - \hat{Y}o')]^2$ estime V sans biais.

soit exprimé autrement:

$(\hat{Y} - \hat{Y}o) * (\hat{Y} - \hat{Y}o')$ estime sans biais la variance de $\hat{Y}$.

Ceci étant vrai pour chaque strate , on a donc un estimateur de la variance de la variable d'interêt dans l'échantillon d'enquête :

$$VAR\text{-}III2\ (\ \hat{Y}\ ) = \sum_{n} (\ \hat{Yh} - \hat{Yho}\ ) * (\ \hat{Yh} - \hat{Yho'}\ )$$

Cette formule est d'autant plus sympatique qu'elle est très simple de mise en oeuvre et peut s'appliquer à toutes sortes de variables et de statistiques, comme nous allons le voir dans la partie suivante.

IV] TESTS PRATIQUES ET RESULTATS.

Tous les tests pratiques ont été réalisés à partir de 16 échantillons de l'enquête "LOYERS et CHARGES".Le choix de cette enquête est lié au fait que d'une part, elle est réalisée trimestriellement et que d'autre part, à chaque trimestre, l'échantillon d'enquête est renouvelé par huitième.Pour cette enquête on dispose donc d'un nombre assez important de sous-échantillons, ce qui est pratique dans l'optique d'un test.En effet, on peut ainsi mesurer la dispersion vraie d'échantillonnage en calculant l'écart type sur 16 réalisations indépendantes du même échantillonnage.

Les développements qui précèdent étaient basés sur une stratification fixe, celle de l' Echantillon Maître.En fait tous les résultats restent valides pour des regroupements de strates.Or l'exploitation exhaustive du Recensement fournit des résultats à un degré de finesse moindre que le découpage en strates utilisé pour les enquêtes.On a donc construit une stratification en 85 "strates d'étude" qui vérifient les deux propriétés suivantes :

- chaque "strate d'étude" est un regroupement de strates de sondage.
- les résultats du Recensement sont disponibles au niveau de chaque "strate d'étude".

Ceci a permit d'utiliser la méthode présentée page 3, pour les variables qui s'y prètent, c'est à dire, l'estimateur basé sur les statistiques de la forme: $(\ Yh - \hat{Yh}\ )^2$.

La constitution des deux échantillons de référence pose des problèmes d'une autre nature.Il était illusoire dans le cadre d'un travail exploratoire d'utiliser l'ensemble de l'exhaustif pour en tirer des échantillons stratifiés.On est parti en fait de l'échantillon "officiel" au centième pour les construire.Or l'échantillon au 1/100 avait été constitué par tirage systématique , trié par Région et tranche d'Unités Urbaines, permettant de constituer une stratification selon ces variables.

On a pu construire une stratification en 166 strates compatible entre l' Echantillon Maître et l'échantillon au centième du Recensement.

Les 85 "strates d'étude" sont un regroupement de ces 166 strates.On a donc pu utiliser la méthode des deux échantillons de référence soit sur 166 strates soit sur 85 strates.

Le tableau ci-dessous récapitule les différentes méthodes testées.

spécificités / méthodes	nbre de strates	Informations disponibles et estimateurs
méthode 1	85	Totaux R.P par strates $\sum_{n} (Yh - \hat{Yh})$
méthode 2	85	Les deux échantillons de référence $(\hat{Yh} - \hat{Yoh})*(\hat{Yh} - \hat{Yoh})$
méthode 3	166	" "

1) <u>Estimateurs de variance pour des totaux tabulés par strates</u>.

La méthode 1 a été testée sur les totaux suivants :

NI : nombre d'individus.
NA : nombre d'actifs.
NE16 : nombre d'enfants de moins de 16 ans.
NE24 : nombre d'enfants de moins de 24 ans.
N65 : nombre de personnes de plus de 65 ans.

Les résultats obtenus sont présentés dans le tableau suivant ainsi que la moyenne, l'écart-type et le coefficient de variation de l'estimateur sur les 16 échantillons testés.Sont également présentés les écarts-type des estimations du total extrapolé sur les 16 échantillons (voir note de bas de page).

Ecart type de totaux tabulés par strate

unités: Milliers de personnes

Variables / Echantillons	NI	NA	NE16	NE24	N65
LC01	986	545	535	628	423
LC02	987	506	443	597	366
LC03	1172	690	592	785	396
LC04	1011	709	530	658	330
LC05	1418	845	613	851	366
LC06	976	643	484	683	385
LC07	1220	541	528	760	275
LC08	1064	560	581	664	363
LC09	986	543	574	652	370
LC10	1272	636	597	776	351
LC11	1083	751	761	734	367
LC12	943	475	524	628	375
LC13	1000	653	538	618	381
LC14	1082	538	625	706	341
LC15	1238	675	769	839	384
LC16	1014	619	540	727	436
Moyenne	1090	621	577	707	367
Ecart-type	136	99	87	79	38
Coefficient de variation	12	16	15	11	10
Ecart-type des estimateurs du total	1024	614	471	436	385

Il semble que la moyenne des estimateurs sur les 16 échantillons surestime l'écart-type des estimations du total pour les variables NI, NA, NE16 et NE24.Par contre pour la variable N65, il est surprenant de constater que la tendance est inversée.

2) Tests sur des fonctions régulières de totaux.

Le test du calcul de variance de telles statistiques s'avère intéressant puisque les trois méthodes sont appliquables.

Note: Les écarts-type des estimations du total extrapolé sur les 16 échantillons ont été calculés par la formule suivante:

$$\sqrt{1/16 \sum_{i=1}^{i=16} (\hat{X}i - X)^2}$$, X:total issu du Recensement

Les fonctions ainsi testées sont les suivantes :

- S1 : nombre d'enfants en âge scolaire: différence entre le nombre d'enfants de moins de 16 ans et le nombre d'enfants de moins de 6 ans.

- S2 : Proportion des plus de soixante cinq ans dans la population totale.

Ecart-type de fonctions régulières de totaux

Unités:Milliers de personnes (pour S1)

Fonctions / Echantillons	nbre d'enfants en âge scolaire				% des plus de 65ans			
	S1	méth1	méth2	méth3	S2	méth1	méth2	méth3
LC01	7760	440	432	476	14.47	0.81	0.78	0.79
LC02	7744	366	347	443	12.62	0.62	0.64	0.72
LC03	8224	462	468	528	13.70	0.76	0.78	0.79
LC04	8112	403	398	461	13.73	0.73	0.78	0.77
LC05	8208	571	573	572	13.75	0.87	0.86	0.79
LC06	7520	396	385	379	13.58	0.79	0.77	0.80
LC07	8576	485	491	504	14.14	0.54	0.56	0.75
LC08	8416	396	392	460	13.15	0.71	0.68	0.72
LC09	8304	443	429	495	13.26	0.71	0.71	0.75
LC10	8544	599	617	514	13.85	0.74	0.81	0.73
LC11	8048	434	422	488	13.89	0.70	0.71	0.73
LC12	8416	421	412	440	13.89	0.67	0.66	0.67
LC13	8224	407	404	482	14.14	0.73	0.78	0.76
LC14	8704	426	432	475	13.28	0.70	0.71	0.77
LC15	9472	679	661	522	12.94	0.84	0.90	0.82
LC16	7744	546	553	478	14.26	0.87	0.82	0.74
Moyenne	8251	467	464	482	13.66	0.73	0.74	0.75
Ecart-type	468	87	91	44	0.50	0.087	0.086	0.037
Coefficient de variation	0.056	0.18	0.19	0.09	0.04	0.119	0.116	0.049

La comparaison des méthodes 1 et 2 montre que les estimations de variance obtenues sont similaires.

La comparaison des méthodes 2 et 3 (qui ne diffèrent que par le nombre de strates utilisées) montre que la méthode 3 est beaucoup plus stable.Par contre, on constate une légère surestimation des estimations de variance mais qui ne semble pas significative.

Pour la statistique S1, il est assez satisfaisant de remarquer que les moyennes des écart-types obtenus par les méthodes 1 à 3 sont proches de l'écart-type des valeurs des statistiques sur les 16 échantillons.Par contre ,pour la statistique S2,l'écart-type de "0.50" semble être sous-estimé.

3) Tests sur des variables recodifiées.

Pour ce type de variables, seule la méthode 3 a été testée.

Les variables complexes testées sont :

V1: nombre d'individus habitant dans un immeuble achevé avant 1948 ,c'est à dire une variable du type:X * indicatrice.

V2: nombre d'individus par piéce pour les ménages dont la personne de référence est ouvrière, soit un ratio réduit à un domaine particulier.

V3: différence entre le nombre d'individus par piéce entre les ménages dont la personne de référence est ouvrière et ceux dont la personne de référence est d'une autre C.S.P , soit une différence de ratios, chacun d'eux étant réduit à un domaine particulier.

Ecart-type de variables complexes

Unités:Millions de personnes(pour V1)

Variables / Echantillons	V1		V2		V3	
	valeur	écart type	valeur	écart type	valeur	écart type
LC01	20800	857	0.991	0.024	0.432	0.0027
LC02	21264	955	0.937	0.023	0.386	0.0026
LC03	22656	926	0.947	0.024	0.377	0.0024
LC04	18944	951	0.941	0.021	0.399	0.0025
LC05	20080	849	0.972	0.026	0.432	0.0028
LC06	21008	857	0.932	0.025	0.387	0.0028
LC07	21504	1023	0.961	0.020	0.400	0.0026
LC08	19552	903	0.949	0.024	0.395	0.0027
LC09	21360	1104	0.942	0.025	0.396	0.0028
LC10	21424	954	0.943	0.024	0.369	0.0026
LC11	20704	768	0.962	0.020	0.430	0.0025
LC12	20496	921	0.958	0.023	0.387	0.0026
LC13	20944	827	0.977	0.023	0.419	0.0027
LC14	20864	747	0.980	0.020	0.414	0.0022
LC15	21456	989	0.912	0.025	0.371	0.0029
LC16	20640	772	0.924	0.024	0.373	0.0026
Moyenne	20856	900	0.951	0.024	0.398	$26*10^{-4}$
Ecart-type	855	98	0.021	0.002	$22*10^{-4}$	$2*10^{-3}$
Coefficient de variation	0.04	0.10	0.022	0.08	0.006	0.06

Cette fois encore les résultats obtenus semblent satisfaisants au niveau de la stabilité de la méthode, si l'on s'en tient à la valeur du coefficient de variation.

Toutefois on voit apparaître une certaine surestimation de la variance: 900 au lieu de 855, $26*10^{-4}$ au lieu de $22*10^{-4}$ qui confirme l'observation faite au sujet de la méthode 3.

Cette remarque rejoint celle qui avait déja été émise aux paragraphes 1 et 2.Il semble que toutes les méthodes d'estimation de variance surestiment la dispersion obtenue sur 16 échantillons indépendants.

On peut toutefois se poser la question dans l'autre sens:peut-être la dispersion des 16 sous-échantillons est-elle artificiellement réduite pour diverses raisons peu claires.L'une d'entre-elles pourrait être le rejet de certains sous-échantillons qui "tombent" particulièrement mal.

V] CONCLUSION.

Tous les tests effectués (voir note de bas de page) ont permis de se rendre compte que l'estimation anticipée de variance est possible et ceci à des coûts relativement faibles.

Il serait cependant souhaitable d'élargir quelque peu ces tests afin d'affiner cette première impression.

Notamment, il semble que des tests sur un nombre plus élévé d'échantillons permettraient une analyse plus approfondie.Il serait également souhaitable de réaliser des tests sur des variables simulées par modèle linéaire ou non pour voir si la méthode des deux échantillons de référence donne des résultats tout aussi satisfaisants.

Enfin ,il faudrait intégrer ces procédures de calcul de variance à celle du tirage des échantillons d'enquête afin de parvenir à nôtre but, qui rappelons le, est une meilleure efficacité de l'échantillonnage.

Les tests présentés ont été effectués au cours d'un stage au sein de la Division "Méthodes Statistiques et Sondages" que dirige Mr J.C DEVILLE.

BIBLIOGRAPHIE

- L' ECHANTILLON MAITRE FAIT PEAU NEUVE. Courrier des statistiques N°29.Janvier 1984 . J.C DEVILLE et G.ROY.

- LA PRECISION DES ENQUETES EMPLOI .Economie et statis--tiques Novembre 1986 . J.C DEVILLE et N.ROTH.

- AN INTRODUCTION TO VARIANCE ESTIMATION. K.WOLTER. (Springer 1985).

- RAPPORT DE STAGE - E.N.S.A.E. (MARS 1990). C.BRUTEL.