C'EST BON A SAVOIR

LA PRECISION DES ENQUETES REVISITEE

Anne-Marie DUSSAIX
ESSEC, B.P. 105, F-95021 CERGY-PONTOISE Cedex
☎ (33) 01 34 43 30 74
Fax (33) 01 34 43 30 01
e-mail : dussaix@edu.essec.fr

Dans les enquêtes par sondage, la précision des résultats s'exprime en particulier en fonction de la taille d'échantillon globale (cf. par exemple Ardilly, 1994 ; Dussaix et Grosbras, 1994). Cependant, lorsqu'il s'agit d'estimer la moyenne ou le total d'une variable quantitative Y, on sait bien que la précision des résultats est également liée au nombre d'individus dans l'échantillon pour lesquels la variable Y étudiée prend des valeurs différentes de zéro : par exemple, l'estimation des quantités achetées d'un produit donné à partir d'un échantillon de n individus sera très imprécise si peu d'individus dans l'échantillon déclarent avoir acheté le produit.

La note qui suit a pour objectif de mettre en évidence cette relation, qui n'est généralement pas explicitée. Elle met aussi en évidence les relations qui existent entre les moyennes, variances, coefficients de variation et précisions calculées sur l'échantillon total et sur le sous-échantillon des individus pour lequel les valeurs de Y sont différentes de zéro. Elle a donc un intérêt pour les utilisateurs de données d'enquêtes et également un intérêt pédagogique dans la mesure où elle développe certains résultats de base.

On va se limiter, dans ce qui suit, au cas où la variable Y étudiée est une variable quantitative, positive ou nulle.

Remarque :

Ceci constitue un cas particulier de l'estimation sur des cibles (ou domaines d'étude). Le sous-ensemble, constitué par les individus de la population pour lesquels la variable étudiée prend des valeurs strictement positives, est donc noté d (pour domaine).

On peut citer de nombreux exemples de tels domaines dans les enquêtes :

- ménages ayant acheté tel produit ou telle marque dans les panels de consommateurs ;
- pharmacies détenant une spécialité pharmaceutique donnée dans les panels de pharmacies ;
- auditeurs d'une station de radio donnée dans une enquête sur l'audience de la radio, etc...

Ces domaines particuliers sont l'objet de deux types d'analyse :

- estimation de quantité moyenne achetée, de vente moyenne par pharmacie détenant la spécialité, de durée moyenne d'écoute par auditeur ...

- étude des caractéristiques des acheteurs, des auditeurs, ...

On fera l'hypothèse simplificatrice et classique selon laquelle l'échantillon est aléatoire simple.

On notera :

- n la taille de l'échantillon total ;
- n_d le nombre d'individus de l'échantillon appartenant au domaine étudié i.e. ; tels que $y_i > 0$ (on les appellera dans la suite individus acheteurs) ;
- $\bar{y}$ la moyenne de la variable étudiée dans l'échantillon (s^2 la variance) ;
- $\bar{y}_d$ la moyenne de la variable étudiée sur les individus de l'échantillon appartenant au domaine (s_d^2 la variance) ;
- $p = n_d/n$ la proportion d'individus dans l'échantillon appartenant au domaine.

On obtient aisément les relations suivantes :

- entre les <u>moyennes</u> $\bar{y}$ et $\bar{y}_d$

(1) $\quad \bar{y} = p\bar{y}_d$

- entre les <u>variances</u>[1] s^2 et s_d^2

$$s^2 = \frac{1}{n-1}\left[(n_d - 1)s_d^2 + p(n - n_d)\bar{y}_d^2\right]$$

soit

(2) $$s^2 \cong ps_d^2 + p(1-p)\bar{y}_d^2$$

- entre les <u>coefficients de variation</u> $CV = s/\bar{y}$ et $CV_d = s_d / \bar{y}_d$

(3) $$CV^2 \cong \frac{1}{p} CV_d^2 + \frac{1-p}{p}$$

Le tableau suivant donne la valeur du coefficient de variation CV sur l'échantillon total en fonction du coefficient de variation CV_d pour les individus acheteurs et du taux p d'individus acheteurs (pour $CV_d = 0,5$; 1 ; 2 et p = 0,01 ; 0,05 ; 0,10 ; 0,20 ; 0,50 ; 0,80).

[1] Ce résultat est obtenu en appliquant l'équation d'analyse de variance dans l'échantillon aux deux groupes constitués par les "acheteurs" et par les "non-acheteurs".

Coefficient de variation CV sur l'échantillon total

p	CV_d 0,5	1	2
0,01	11,14	14,11	22,34
0,05	4,90	6,25	9,95
0,10	3,39	4,36	7,00
0,20	2,29	3,00	4,90
0,50	1,22	1,73	3,00
0,80	0,75	1,22	2,29

Pour des valeurs faibles de p

$$CV^2 \cong \frac{1}{p}\left[CV_d^2 + 1\right]$$

Si le nombre d'individus n_d est suffisant (on admettra la condition $n_d > 30$), on peut utiliser l'approximation normale pour la construction des intervalles de confiance pour m et pour m_d[2]. Au degré de confiance 1-α :

$$m \in \left[\bar{y} - t_{(1-\alpha/2)}\frac{s}{\sqrt{n}}, \bar{y} + t_{(1-\alpha/2)}\frac{s}{\sqrt{n}}\right]$$

$$m_d \in \left[\bar{y}_d - t_{(1-\alpha/2)}\frac{s_d}{\sqrt{n_d}}, \bar{y}_d + t_{(1-\alpha/2)}\frac{s_d}{\sqrt{n_d}}\right]$$

où $t_{1-\alpha/2}$ est le fractile de la loi normale centrée réduite correspondant à la probabilité 1-α/2.

On en déduit alors la relation :

- entre la précision relative δ au degré de confiance 1-α pour l'estimation des quantités achetées en moyenne sur l'échantillon total

$$\delta = t_{1-\alpha/2}\frac{CV}{\sqrt{n}}$$

- et la précision relative δ_d sur les seuls individus du domaine

$$\delta_d = t_{1-\alpha/2}\frac{CV_d}{\sqrt{n_d}}$$

[2] On peut trouver par exemple dans Ardilly, 1994, les propriétés statistiques de l'estimateur $\bar{y}_d$.

En utilisant les résultats précédents, on obtient :

$$\delta^2 = \delta_d^2 + \left(t_{1-\alpha/2}\right)^2 \frac{1}{n}\frac{1-p}{p}$$

ou

$$(4) \qquad \delta^2 = \left(t_{1-\alpha/2}\right)^2 \frac{1}{n_d}\left[CV_d^2 + 1 - p\right]$$

On a donc toujours :

$$\delta^2 \geq \delta_d^2$$

Le résultat (4) est assez surprenant puisqu'il lie la précision relative obtenue dans l'estimation de la moyenne de Y sur l'ensemble de la population au nombre n_d d'individus pour lesquels $y_i > 0$. Il l'est moins si l'on se rappelle que la précision relative est le demi-intervalle de confiance divisé par la moyenne dans l'échantillon, moyenne très proche de zéro si, pour un nombre d'individus $n - n_d$ élevé, $y_i = 0$.

Exemple des quantités achetées d'un produit donné ; supposons que n = 10 000, p = proportion d'individus acheteurs = 1 %, CV_d = coefficient de variation des quantités achetées sur les individus acheteurs = 1. Dans ce cas, au degré de confiance 0,95,

$$\delta^2 = (1{,}96)^2 \frac{1}{100}(1 + 0{,}99) = 0{,}076 \text{ soit } \delta = 27{,}6\%$$

(la précision relative dans l'estimation de la moyenne, pour les seuls individus acheteurs est $\delta_d = 19{,}6\ \%$).

Malgré la taille d'échantillon élevée, l'estimation du total des quantités achetées sur l'ensemble de la population est affectée de la même "imprécision" : au degré de confiance 0,95, le total des quantités achetées est estimé avec une précision relative de 27,6 %.

En conclusion, cette simple application de résultats statistiques élémentaires permet de mettre en évidence le lien entre la précision dans l'estimation des moyennes ou totaux, au niveau de la population tout entière, et la fraction des individus acheteurs, auditeurs, ... dans l'ensemble de l'échantillon.

Références :

- Ardilly P., Les techniques de sondage, Technip, 1994.
- Dussaix A.M. et Grosbras J.M., Les sondages : principes et méthodes, Que sais-je n° 701, Presses Universitaires de France, 2è édition, 1994.