h e g

Haute école de gestion de Genève
Geneva School of Business Administration

Échantillonnage

Dr Christophe Hebeisen
christophe.hebeisen@a3.epfl.ch

HEG - Économie d'Entreprise
cours 4 et 5

11 et 18 mars 2009

Objectifs

- Connaître la différence entre sondage probabiliste et sondage empirique
- Estimation d'une quantité d'intérêt :
  - Connaissance du biais
  - Connaissance de la variance d'échantillonnage
  - Convergence et erreur quadratique moyenne

Introduction

Échantillonnage et taux de sondage

L'échantillonnage consiste à déterminer comment sélectionner les personnes qui seront incluses dans l'enquête.

De façon générale, on considère une population de taille N finie dans laquelle on veut tirer un échantillon de taille n finie. Le rapport

$$f = \frac{n}{N}$$

est appelé taux de sondage.

L'échantillonnage (le tirage) peut être effectué avec ou sans remise.

Classes de sondage

Deux classes de sondage

1. Sondages probabilistes (ou « aléatoires ») :
   Chaque individu de la population a une probabilité connue *à priori* d'appartenir à l'échantillon. Cette probabilité est appelée probabilité d'inclusion.

2. Sondages empiriques (ou « non aléatoires ») :
   La probabilité d'inclusion des individus dans l'échantillon est inconnue.

Sondages empiriques

Malgré leurs défauts, les méthodes non aléatoires sont très utilisées dans les sondages d'opinion et les études de marché, essentiellement pour deux raisons :

1. La rareté ou la non-disponibilité des bases de sondage
2. Le coût et les délais de réalisation

Principales méthodes empiriques

Les principales méthodes empiriques sont :

- les quotas
- les itinéraires
- l'emplacement
- le volontariat
- la méthode boule de neige

La méthode des quotas

Définition

« Les différents caractères que l'on peut observer dans une population n'étant pas indépendants entre eux, un échantillon identique à la population dans laquelle il est prélevé en ce qui concerne la distribution statistique de certains caractères importants sera également peu différent de la population en ce qui concerne la distribution statistique des caractères qui ne sont pas contrôlés. »

J. Desabie (1966)

Méthode des quotas = construire un échantillon qui soit un modèle réduit de la population étudiée.

La méthode des quotas

Le principe de la méthode est le suivant :

- choisir quelques caractéristiques (= quotas) dont on connaît la distribution dans la population étudiée
- donner aux enquêteurs un plan de travail (souvent accompagné de consignes de recherche) qui lui impose le respect de certaines proportions au sein de ses interviews

L'échantillon obtenu est représentatif de la population *par rapport* aux variables choisies (il respecte les proportions constatées dans la population).

Difficulté : réaliser ses fins de quotas pour éviter de se retrouver dans une situation délicate pour ses derniers interviewés.

Important : contrôler *a posteriori* le travail des enquêteurs par contre-enquête.

La méthode des quotas

La feuille de quotas aura p.ex. l'allure suivante :

Région	Lausanne
Habitat	plus de 130'000 habitants
100 interviews à réaliser	
Sexe de l'interviewé	
homme	40 ** ** ** **
femme	60 ** ** ** ** ** **
Âge de l'interviewé	
$<$ 18 ans	20 ** **
18 à 60 ans	60 ** ** ** ** ** **
$>$ 60 ans	20 ** **
Secteur d'activité	
banque	50 ** ** ** ** **
autre	50 ** ** ** ** **

La méthode des quotas

Les inconvénients de la méthode sont ceux des méthodes non aléatoires en général :

- existence de biais (même avec les consignes) ; la qualité des enquêtes repose sur celle du travail de l'enquêteur
- impossibilité de calculer des marges d'erreur

Les avantages sont essentiellement :

- des coûts et des délais de réalisations plus faibles que ceux d'une enquête aléatoire
- des résultats que l'on peut qualifier de fidèles

Sondages probabilistes

Chaque membre de la population possède une probabilité connue *a priori* d'être inclus dans l'échantillon (= probabilité d'inclusion).

Contrairement à la méthode des quotas (mimer la population étudiée), les méthodes probabilistes permettent de sur- ou sous-échantillonner certaines catégories de la population.

Par exemple, des personnes possédant une **caractéristique rare** importante pour l'étude devront être sur-représentées dans l'échantillon. Les estimations seront ensuite corrigées par pondération.

But : s'assurer que toutes les tendances ou caractéristiques importantes de la population soient prises en compte dans l'échantillon, afin que ce dernier soit représentatif.

Sondages probabilistes

Les méthodes d'échantillonnage probabilistes peuvent être à un ou plusieurs niveaux (ou degrés).

- Un niveau : l'échantillonnage se fait par rapport à l'ensemble de la population considérée
  - sondage aléatoire simple (SAS)
- Deux niveaux ou plus : la population est tout d'abord découpée en plusieurs groupes mutuellement exclusifs, puis l'échantillonnage s'effectue indépendamment au sein de ces groupes.
  - sondage stratifié
  - sondage par grappes

Estimation d'une quantité d'intérêt

Objectif

But : estimation d'un paramètre θ (inconnu) au sein d'une population de taille N (finie et connue). Cet estimateur est noté $\hat{\theta}$.

Rappel (statistiques III) : un estimateur $\hat{\theta}$ est considéré comme une variable aléatoire dont

- le comportement en moyenne est l'espérance, $E(\hat{\theta})$
- la variance, $\mathrm{Var}(\hat{\theta})$, est une mesure de dispersion des estimations

Rappels sur l'espérance et la variance

Si X est une variable aléatoire discrète pouvant prendre les valeurs x_i, avec des probabilités respectives $p(x_i)$, alors pour toute fonction g on aura

$$E\big(g(X)\big) = \sum_i g(x_i)p(x_i)$$

Il en découle les formules suivantes pour l'espérance :

- $E(X) = \sum_i x_i p(x_i)$
- $E(X^n) = \sum_i x_i^n p(x_i)$
- $E(aX + b) = aE(X) + b$, où a et b sont des constantes
- $E(X + Y) = E(X) + E(Y)$

Rappels sur l'espérance et la variance

La variance de X est la quantité

$$\mathrm{Var}(X) = E\big((X - E(X))^2\big) = \ldots = E(X^2) - \big(E(X)\big)^2$$

Propriété de la variance :

$$\mathrm{Var}(aX + b) = a^2\,\mathrm{Var}(X)$$

La racine carrée de la variance de X est appelée l'écart-type de X, qui se note σ :

$$\sigma = \sqrt{\mathrm{Var}(X)}$$

Exercice : calculer l'espérance et la variance de la variable X, résultat d'un lancer de dé non truqué

Solution

Solution : X peut prendre les valeurs 1, 2, 3, 4, 5 et 6 avec probabilité respective $p = \frac{1}{6}$. On aura donc

$$\begin{aligned} E(X) &= 1 \cdot \tfrac{1}{6} + 2 \cdot \tfrac{1}{6} + 3 \cdot \tfrac{1}{6} + 4 \cdot \tfrac{1}{6} + 5 \cdot \tfrac{1}{6} + 6 \cdot \tfrac{1}{6} = \tfrac{7}{2} \\ E(X^2) &= 1^2 \cdot \tfrac{1}{6} + 2^2 \cdot \tfrac{1}{6} + 3^2 \cdot \tfrac{1}{6} + \ldots + 6^2 \cdot \tfrac{1}{6} = \tfrac{91}{6} \\ \mathrm{Var}(X) &= \tfrac{91}{6} - \left(\tfrac{7}{2}\right)^2 = \tfrac{35}{12} \end{aligned}$$

Paramètre d'intérêt

Exemples de paramètres d'intérêt θ d'une population de taille N finie.

- une moyenne

$$\theta = \frac{\sum_{i=1}^{N} Y_i}{N} = \bar{Y}$$

Exemple : $Y_i =$ note d'un test, taille d'un individu, etc.

- une proportion

$$\theta = \frac{\sum_{i=1}^{N} Y_i}{N} = \bar{Y} \text{ où } Y_i = 0 \text{ ou } 1$$

Exemple : $Y_i = 0$ (homme), $Y_i = 1$ (femme)

- un total

$$\theta = \sum_{i=1}^{N} Y_i = T$$

Exemple : $Y_i =$ prix d'un article dans un magasin

Principe de base

L'estimateur est construit sur le même modèle que le paramètre à estimer. Par exemple, si s est l'échantillon de taille n, alors

$$\theta = \sum_{i=1}^{N} \alpha_i Y_i \qquad \Longrightarrow \qquad \hat{\theta} = \sum_{i \in s} w_i(s) Y_i$$

$w_i(s)$ = poids attaché à l'individu i de l'échantillon s (= nombre d'individus de la population qu'il représente)

De la même manière, si le paramètre d'intérêt est une variance, alors l'estimateur est une expression quadratique.

Définition
Un plan de sondage est constitué d'une méthode d'échantillonnage et de l'expression d'un estimateur.

Biais

Le biais

Si nous associons une probabilité $p(s_i)$ à tout échantillon s_i de taille n fixée, nous avons vu que l'espérance de l'estimateur $\hat{\theta}$ du paramètre θ est la valeur

$$E(\hat{\theta}) = \sum_i p(s_i)\hat{\theta}(s_i)$$

Définition

Le biais de l'estimateur est la quantité

$$\text{Biais}(\hat{\theta}) = E(\hat{\theta} - \theta) = E(\hat{\theta}) - \theta$$

Un estimateur est dit sans biais (ou non biaisé) si le biais est nul, c'est-à-dire si $E(\hat{\theta}) = \theta$

But : diminuer le biais en ayant $E(\hat{\theta})$ le plus proche possible de θ

Exemple

Prenons une population de 4 individus $\{1, 2, 3, 4\}$. Le revenu mensuel de ces individus est donné par le tableau suivant :

$$\begin{aligned} R_1 &= 6'000 \\ R_2 &= 12'000 \\ R_3 &= 8'000 \\ R_4 &= 6'000 \end{aligned}$$

Le salaire moyen de cette population est donc

$$\theta = \bar{R} = \frac{R_1 + R_2 + R_3 + R_4}{4} = 8'000$$

Un sondage doit être conduit afin d'estimer le revenu moyen de cette population (le tableau ci-dessus n'est pas connu). Si un sondage ne peut se faire que sur 2 individus, pour des questions budgétaires, quels sont les échantillons possibles?

Exemple (suite)

$$s_1 = \{1, 2\} \qquad s_4 = \{2, 3\}$$
$$s_2 = \{1, 3\} \qquad s_5 = \{2, 4\}$$
$$s_3 = \{1, 4\} \qquad s_6 = \{3, 4\}$$

Estimons le revenu moyen par la moyenne simple :

$$\hat{\theta}(s_1) = \frac{R_1+R_2}{2} = 9'000 \qquad \hat{\theta}(s_4) = \frac{R_2+R_3}{2} = 10'000$$
$$\hat{\theta}(s_2) = \frac{R_1+R_3}{2} = 7'000 \qquad \hat{\theta}(s_5) = \frac{R_2+R_4}{2} = 9'000$$
$$\hat{\theta}(s_3) = \frac{R_1+R_4}{2} = 6'000 \qquad \hat{\theta}(s_6) = \frac{R_3+R_4}{2} = 7'000$$

Si tous les échantillons ont la même probabilité d'être pris, c'est-à-dire $p(s_i) = \frac{1}{6}$, l'espérance de la moyenne simple sur tous les échantillons possibles vaut

$$E(\hat{\theta}) = \sum_{i=1}^{6} p(s_i)\hat{\theta}(s_i) = 8'000 = \theta \text{ (biais nul)}$$

Exemple (suite)

Mais si l'on décide de favoriser l'individu 1 (il est p.ex. plus coopératif que les autres), on pourrait associer aux ensembles les probabilités suivantes :

$$p(s_1) = p(s_2) = 0.25,\ p(s_3) = 0.2,\ p(s_4) = p(s_5) = p(s_6) = 0.1$$

Rappelons que la somme des probabilités doit faire 1, ce qui est bien le cas ici :

$$2 \cdot 0.25 + 1 \cdot 0.2 + 3 \cdot 0.1 = 1$$

L'espérance de la moyenne simple vaut alors

$$E(\hat{\theta}) = \sum_{i=1}^{6} p(s_i)\hat{\theta}(s_i) = 7'800 \neq \theta$$

Exemple (suite)

Le biais vaut donc ici $7'800 - 8'000 = -200$, et l'erreur relative commise vaut

$$\frac{\text{biais}}{\text{paramètre}} = \frac{-200}{8'000} = -2,5\%$$

Remarquons encore qu'en ne tirant qu'un échantillon (cas le plus courant), la plus petite valeur possible du biais est $1'000$ (cas s_1, s_2, s_5 et s_6) ; l'erreur commise est alors de $12,5\%$.

Au pire, il est de $2'000$ (cas s_3 et s_4), ce qui arrive avec une probabilité de $\frac{1}{6} + \frac{1}{6} = \frac{1}{3}$ dans le cas équiprobable, et de $0.2 + 0.1 = 0.3$ dans le cas avec poids de préférence. L'erreur commise est de 25%.

Variance de l'échantillon

Variance de l'estimateur

On rappelle la variance de l'estimateur $\hat{\theta}$:

$$\mathrm{Var}(\hat{\theta}) = E\big(\hat{\theta} - E(\hat{\theta})\big)^2 = \sum_i p(s_i)\big(\hat{\theta}(s_i) - E(\hat{\theta})\big)^2$$

La variance et l'écart-type σ sont des mesures de précision de l'estimateur. Plus elles sont petites, plus l'estimation sera précise en moyenne.

Amélioration de la précision:

- chercher une meilleure formule pour $\hat{\theta}$
- modifier la méthode d'échantillonnage

But : chercher à réduire la variance (le plus souvent en premier).

Exemple

Exemple précédent, avec poids de préférence. Nous avions :

$$\hat{\theta}(s_1) = 9'000 \qquad \hat{\theta}(s_4) = 10'000$$
$$\hat{\theta}(s_2) = 7'000 \qquad \hat{\theta}(s_5) = 9'000$$
$$\hat{\theta}(s_3) = 6'000 \qquad \hat{\theta}(s_6) = 7'000$$

$$p(s_1) = p(s_2) = 0.25,\ p(s_3) = 0.2,\ p(s_4) = p(s_5) = p(s_6) = 0.1$$

et $E(\hat{\theta}) = 7'800$

Nous calculons

$$\hat{\theta}(s_1) - E(\hat{\theta}) = 1'200 \qquad \hat{\theta}(s_4) - E(\hat{\theta}) = 2'200$$
$$\hat{\theta}(s_2) - E(\hat{\theta}) = -800 \qquad \hat{\theta}(s_5) - E(\hat{\theta}) = 1'200$$
$$\hat{\theta}(s_3) - E(\hat{\theta}) = -1'800 \qquad \hat{\theta}(s_6) - E(\hat{\theta}) = -800$$

$$\begin{aligned} \mathrm{Var}(\hat{\theta}) &= 0.25 \cdot (1200)^2 + 0.25 \cdot (-800)^2 + \ldots + 0.1 \cdot (-800)^2 \\ &= 1'860'000 \end{aligned}$$

Erreur quadratique moyenne

Erreur quadratique moyenne

Définition

L'erreur quadratique moyenne (EQM ; angl. MSE : Mean Square Error) est la quantité $\mathrm{EQM}(\hat{\theta}) = E\big[(\hat{\theta} - \theta)^2\big]$.

On peut montrer que

$$\mathrm{EQM}(\hat{\theta}) = \mathrm{Var}(\hat{\theta}) + \big(\mathrm{Biais}(\hat{\theta})\big)^2$$

Parmi plusieurs estimateurs d'un même paramètre, certains peuvent être biaisés, d'autres non, certains ont même variance, d'autres non. L'EQM est un indicateur de la qualité d'un estimateur prenant en compte les deux notions.

But : choisir celui qui a l'EQM la plus petite.

Exemple 1

Reprenons l'exemple précédent.

$$\begin{aligned} \mathrm{EQM}(\hat{\theta}) &= \mathrm{Var}(\hat{\theta}) + \big(\mathrm{Biais}(\hat{\theta})\big)^2 \\ &= 1'860'000 + (-200)^2 = 1'900'000 \end{aligned}$$

Calcul avec $\mathrm{EQM}(\hat{\theta}) = E\big[(\hat{\theta} - \theta)^2\big]$

$$\begin{aligned} \mathrm{EQM}(\hat{\theta}) &= (9'000 - 8'000)^2 \cdot 0.25 + \\ &\quad (7'000 - 8'000)^2 \cdot 0.25 + \\ &\quad (6'000 - 8'000)^2 \cdot 0.2 + \\ &\quad (10'000 - 8'000)^2 \cdot 0.1 + \\ &\quad (9'000 - 8'000)^2 \cdot 0.1 + \\ &\quad (7'000 - 8'000)^2 \cdot 0.1 \\ &= 1'900'000 \end{aligned}$$

Exemple 2

Regardons l'exemple du support de cours : on a un paramètre θ et deux estimateurs $\hat{\theta}_1$ et $\hat{\theta}_2$ de ce paramètre. On connaît

$$\begin{aligned} \mathrm{Var}(\hat{\theta}_1) &= 100 \qquad & \mathrm{Biais}(\hat{\theta}_1) &= 0 \\ \mathrm{Var}(\hat{\theta}_2) &= 50 \qquad & \mathrm{Biais}(\hat{\theta}_2) &= 5 \end{aligned}$$

On aura donc

$$\begin{aligned} \mathrm{EQM}(\hat{\theta}_1) &= 100 + 0^2 &= 100 \\ \mathrm{EQM}(\hat{\theta}_2) &= 50 + 5^2 &= 75 \end{aligned}$$

Choix de l'estimateur dans ce cas : $\hat{\theta}_2$