Enquêtes et sondages
2003-2004
UV CNAM 18323 - STATISTIQUE B8

Intervenants : G.Saporta (CNAM), O.Marchese (IPSOS), S.Rousseau (INSEE)

Plan :

10 octobre : Introduction GS+OM
17 octobre : sondage aléatoire simple GS
24 octobre: sources d'erreur et biais OM
31 octobre: sondages à probabilités inégales GS
7 novembre : algorithmes de tirage GS
10 novembre: stratification GS
21 novembre: sondages à deux degrés et grappes GS
28 novembre : données manquantes et fusions de fichiers GS
5 décembre: Redressement (quotient, régression post-strates) GS
12 décembre: Effets et pratique des redressements OM

9 janvier: la méthode des quotas OM
16 janvier: panels GS
23 janvier: panels OM
30 janvier: méthodes d'enquête OM
6 février : méthodes d'enquête OM

Références

http://cedric.cnam.fr/~saporta/
http://www.agro-montpellier.fr/cnam-lr/statnet/
http://www.ipsos.fr/
http://www.cbs.nl/isi/iass/

P.ARDILLY	Les techniques de sondage (éditions Technip, 1994)
A.M. DUSSAIX, J.M. GROSBRAS	Exercices de sondages (Economica, 1992)
A.M. DUSSAIX, J.M. GROSBRAS	Les sondages (Que sais-je? N°701, 1996)
Y.TILLE	Théorie des sondages (Dunod, 2001)

ENQUETES et SONDAGES
UV 18323 - STATISTIQUE B8

2003-2004

Gilbert SAPORTA
Chaire de Statistique Appliquée
Conservatoire National des Arts et Mtiers
292 rue Saint Martin
75141 Paris cedex 03

saporta@cnam.fr
http://cedric.cnam.fr/~saporta

INTRODUCTION

- Aperçu du secteur
  - statistique publique

    CNIS

    INSEE – 7 000 employés
  - 400 Instituts privés

    (10 000 employés, dont 4 000 permanents)

    CA 2001: 1.275 milliards € (+6.25%)

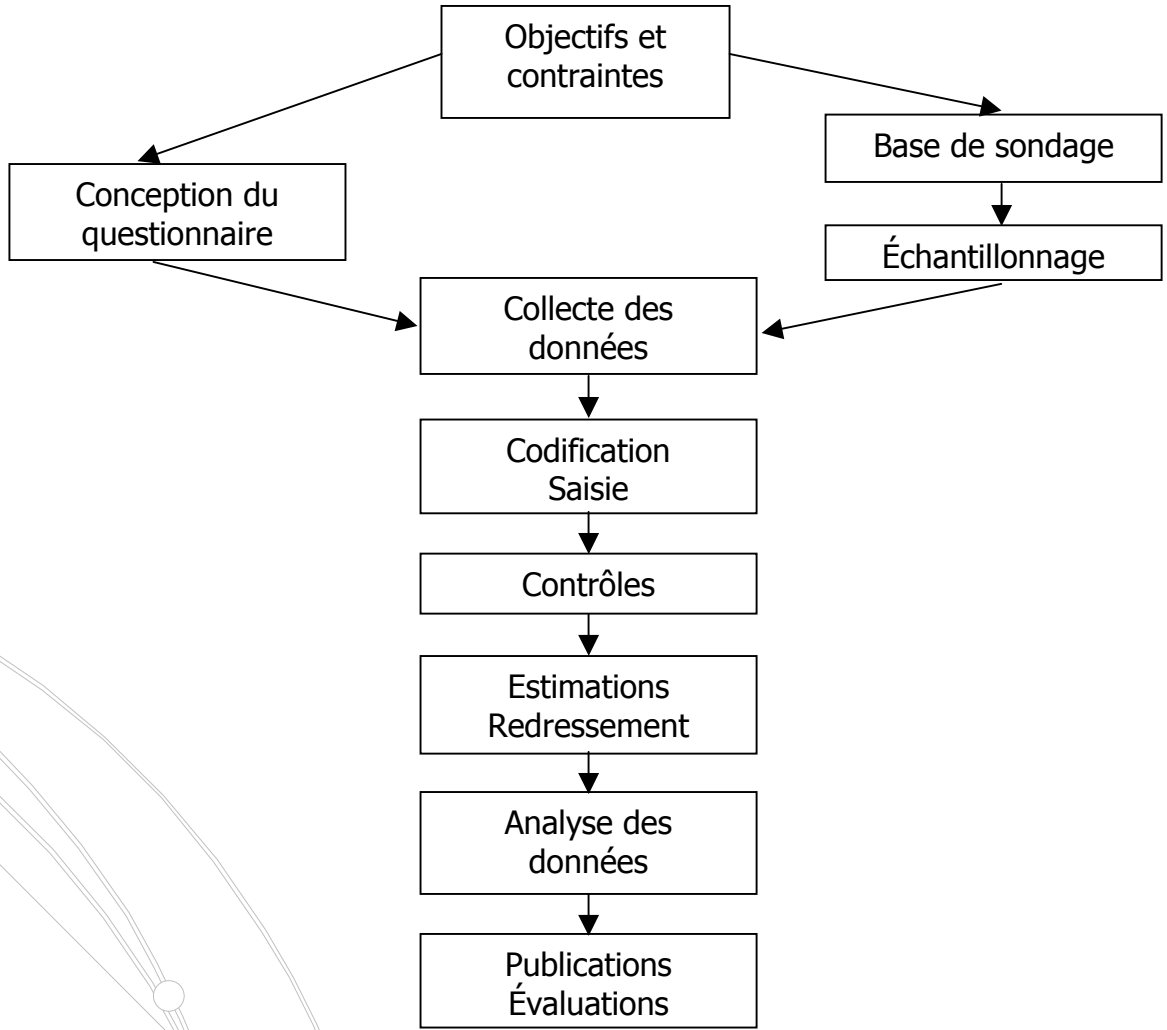
INTRODUCTION

Progression du CA des membres de Syntec Marketing et Opinion

INTRODUCTION

- Histoire récente
  - **1895** – Kiaer, dénombrements représentatifs
  - **1925** – Jensen
  - **1934** – Neyman, Sondages à 2 degrés
  - **1952** – Horvitz et Thompson, Sondages à probabilités inégales
  - **1936** – Election de Roosevelt
  - **1938** – Fondation de l'IFOP
  - **1965** – Ballottage De Gaulle

INTRODUCTION



LES TECHNIQUES DE SONDAGE

- Méthodes aléatoires:

  Plans de sondage

  - **Simples:** - à probabilités égales
    - à probabilités inégales
  - **Complexes:** - stratifié
    - en grappe
    - plusieurs degrés

LES TECHNIQUES DE SONDAGE

- Méthodes par choix raisonné ou judicieuse:
  - Quotas;
  - Itinéraires;
  - Unités – types;
  - Volontariat;
  - Échantillonnage sur place;
  - Sondage « à chaud ».

LES TECHNIQUES DE SONDAGE

- Problèmes essentiels:
  - Sélection de l'échantillon;
  - Agrégation des réponses
    - ✓ estimateur;
    - ✓ précision;

SONDAGE ALEATOIRE SIMPLE

- Notations:
  - Population ou base de sondage: **N**
  - Identifiant: **i**
  - Variable d'intérêt: **Y** ($Y_1, Y_2 \ldots\ldots Y_N$)

$$\overline{Y} = \frac{1}{N}\sum_{i=1}^{N} Y_i; \qquad T = \sum_{i=1}^{N} Y_i$$

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(Y_i - \overline{Y})^2; \qquad S^2 = \frac{1}{N-1}\sum_{i=1}^{N}(Y_i - \overline{Y})^2 = \frac{N}{N-1}\sigma^2$$

SONDAGE ALÉATOIRE SIMPLE

- Définition: tirage équiprobable sans remise de n unités;
- C_N^n échantillons possibles;
- π_i probabilité d'inclusion (plan de taille fixe): $\sum_{i=1}^{N} \pi_i = n$
- Équiprobabilité: $\rightarrow \pi_i = \frac{n}{N}$
- Remarque: $\pi_i = \sum_{s\ (i \in s)} p(s)$
- Taux de sondage: $\frac{n}{N} = f$

SONDAGE ALÉATOIRE SIMPLE

- Estimation du total et de la moyenne:

  $\bar{y}$ - estimateur de $\bar{Y}$

  N $\bar{y}$ - estimateur de T

  $E(\bar{y}) = \bar{Y}$; $E(N\bar{y}) = T$

- Démonstration avec les variables de Cornfield

$$\delta_i = \begin{cases} 1 \ si \ i \in s \\ 0 \ si \ i \notin s \end{cases}$$

$$E(\delta_i) = \pi_i$$

$$V(\delta_i) = \pi_i(1-\pi_i) \quad \text{cov}(\delta_i;\delta_j) = \pi_{ij} - \pi_i\pi_j$$

$$\frac{N}{n}\sum_{i\in s} y_i = \hat{T} = \sum_{i\in s}\frac{y_i}{\pi_i} = \sum_{i=1}^{N}\frac{Y_i}{\pi_i}\delta_i$$

$$E(\hat{T}) = \sum_{i=1}^{N}\frac{Y_i}{\pi_i}E(\delta_i) = \sum_{i=1}^{N} Y_i = T$$

y_i=variable aléatoire;

Y_i= variable non aléatoire

SONDAGE ALÉATOIRE SIMPLE

- **Variances:**

$$V(\bar{y}) = (1 - f)\frac{S^2}{n}$$

$$V(\hat{T}) = N^2(1 - f)\frac{S^2}{n}$$

Estimation de S^2:

$$s^2 = \frac{1}{n-1}\sum_{i \in s}(y_i - \bar{y})^2$$

$$E(s^2) = S^2$$

$$\Rightarrow \begin{cases} \widehat{V(\bar{y})} = (1-f)\dfrac{s^2}{n} \\ \widehat{V(\hat{T})} = N^2(1-f)\dfrac{s^2}{n} \end{cases}$$

SONDAGE ALÉATOIRE SIMPLE

- Intervalles de confiance estimés:

$$\bar{y}-2s\sqrt{\frac{1-f}{n}}<\bar{Y}<\bar{y}+2s\sqrt{\frac{1-f}{n}}$$

- Cas d'un pourcentage:

$$Y_i = \begin{cases} 1 \\ 0 \end{cases} \qquad \bar{Y}=P$$

$$V(p)=(1-f)\frac{P(1-P)}{n}\frac{N}{N-1}$$

$$\hat{V}(p)=(1-f)\frac{p(1-p)}{n-1}\simeq\frac{p(1-p)}{n}$$

Sources d’erreur et biais

Sources d'erreur et biais

- Utilisations des données d'enquête :
  - « Describers » & « Modelers »
- Sources d'erreur
- « Nonsampling errors »
  - Populations d'intérêt
  - Défaut ou excès de couverture
  - Non-réponse
  - Erreur de mesure
- Sources d'erreur et phases d'enquête
- L' «art» du sondeur

Utilisations des données d'enquête : « Describers » & « Modelers »

- Différents langages, différentes préoccupations

- Accent sur l'estimation des caractéristiques d'une population

  vs

- Accent sur la validation d'hypothèses théoriques

- Accent sur l'estimation de moyennes et proportions

  vs

- Accent sur l'exploration de structures de covariance

- Forte attention aux erreurs de non-observation (défauts de couverture, non-réponse)

  vs

- Forte attention aux erreurs d'observation (questionnaire)

Sources d'erreur {1/3}

- Erreur d'échantillonnage
  - Hétérogénéité des mesures parmi les individus de la population
- Défaut ou excès de couverture
  - Probabilité de sélection nulle ou non connue pour les individus de la population
- Non-réponse
  - Défaut de collecte de toute ou partie de l'information pour certains individus de l'échantillon
- Erreur de mesure
  - Influence de l'enquêteur sur les réponses des personnes interrogées
  - Incapacité (ou manque de volonté) des personnes interrogées à répondre aux questions : mémoire, impréparation, facteurs psychologiques, …
  - Défauts de l'instrument de mesure (questionnaire ou autre)
  - Effets du mode de recueil (face à face, téléphone, auto-administré papier ou Internet)

Sources d'erreur {2/3}

- Ces erreurs peuvent être liées les unes aux autres
  - Eg : Faire du « forcing » pour réduire la non-réponse peut amener à amplifier les erreurs de mesure
- En général, les efforts de modélisation et de mesure sont portés sur l'erreur d'échantillonnage et la non-réponse
- Souvent on ne sait que très peu – et parfois rien du tout - sur les erreurs d'observation et les défauts de couverture
- Or, cela peut s'avérer létal, car ces erreurs - qui ont essentiellement la nature de biais – ne diminuent pas lorsque la taille d'échantillon augmente

Sources d'erreur {3/3}

Moralité

- Les efforts visant à affiner une méthode de tirage ou l'expression d'un estimateur pour obtenir un gain de précision peuvent s'avérer bien illusoires si, par ailleurs, les erreurs d'observation, les défauts de couverture ou la non-réponse sont importants

- Dans une telle situation, une taille d'échantillon très importante ne sera pas non plus de nature à éviter la déroute
  - Lors de la Présidentielle américaine de 1936, le « vote de paille » organisé par le *Literary Digest* - portant sur près de deux millions de lecteurs - donnait une confortable avance à Alfred Landon (54%) … alors que Franklin Roosevelt allait recueillir 61% des suffrages !

« Nonsampling errors » : Populations d'intérêt

- Population objet de l'inférence (population of inference)
  - Ensemble des unités à étudier
- Population cible du sondage (target population)
  - Ensemble des unités étudiées
- Base de sondage (frame population)
  - Liste des unités utilisée pour la sélection de l'échantillon: l'« univers » auquel font référence la plupart des livres de statistique
- Population enquêtable (survey population)
  - Liste des unités accessibles, physiquement et mentalement prêtes à répondre, souhaitant répondre aux questions
  - Il s'agit bien évidemment d'une abstraction, puisque elle ne peut être observée indépendamment des opérations d'échantillonnage elles-mêmes
- Non-réponse
  - divergences entre « frame » et « survey population »
- Erreurs de couverture
  - divergences entre « frame » et « target population»

« Nonsampling errors » : Défaut ou excès de couverture {1/2}

- Ambiguïté du repérage des unités de la population
  - Une base de sondage se doit pour le moins d'être une liste d'identifiants de bonne qualité
- Manque d'exhaustivité
  - Chaque unité faisant partie du champ de l'enquête doit être présente dans la liste des identifiants
- Doubles comptes
  - Aucune unité doit être présente plusieurs fois dans la base (surtout si le nombre de fois n'est pas connu)
- Absence d'informations auxiliaires
  - Leur disponibilité peut être mise à profit pour améliorer soit la méthode de tirage, soit l'estimateur, soit les deux
- Vieillissement de la base elle-même
- Absence ou inaccessibilité de la base de sondage
  - (situation finalement pas si rare!)

« Nonsampling errors » : Défaut ou excès de couverture {2/2}

- L'erreur de couverture est une fonction
  - de la proportion de population non couverte par la base de sondage
  - de la différence dans la valeur de la variable d'intérêt entre « frame » et « target population»

- $Y_c = Y + (N_{nc} / N) * (Y_C - Y_{nc})$

  où Y représente la valeur auprès des N unités de la target population

  Y_C représente la valeur auprès des N_C unités couvertes par la « frame population »

  Y_{nc} représente la valeur auprès des N_{nc} unités non couvertes par la « frame population »

- L'erreur de couverture
  - est liée à la variable d'intérêt
  - n'est pas une propriété de l'échantillon

« Nonsampling errors » : Non-réponse {1/3}

- Comme pour le défaut de couverture dû au manque d'exhaustivité de la base de sondage, la non-réponse
  - nous met dans l'impossibilité d'observer la valeur de la variable d'intérêt
  - engendre un biais non mesurable, puisque l'on ne sait pas si les unités observées sont comparables aux unités non observées

- A différence du défaut de couverture, la non réponse
  - est d'ampleur mesurable, à partir de l'échantillon tiré (taux de non-réponse calculable)
  - peut être complète ou partielle (l'individu sélectionné répond à certaines questions et pas à d'autres)

- En diminuant la taille de l'échantillon, la non-réponse occasionne une perte de précision (quelles que soient les hypothèses formulées sur le profile des non-répondants)

« Nonsampling errors » : Non-réponse {2/3}

- Le taux de non-réponse est souvent interprété comme LA mesure de qualité de l'estimation de la variable d'intérêt
  - or, il ne s'agit que d'une composante de l'erreur et ne peut pas en donner seul la mesure
- L'erreur dû à la non-réponse est une fonction
  - du taux de non-réponse
  - de la différence dans la valeur de la variable d'intérêt entre répondants et non-répondants
- $y_r = y_n + (nr / n) * (y_r - y_{nr})$
- L'erreur de couverture
  - est liée à la variable d'intérêt
  - n'est pas une propriété de l'échantillon

« Nonsampling errors » : Non-réponse {3/3}

- Une expression plus complète de la variable d'intérêt estimée devrait être

$$y_r = y_n + (nc / n) * (y_r - y_{nc}) + (ni / n) * (y_r - y_{ni}) + (rf / n) * (y_r - y_{rf})$$

où y_{nc} représente la valeur auprès des nc unités non contacté

y_{ni} représente la valeur auprès des ni unités incapables de fournir une réponse

y_{rf} représente la valeur auprès des rf unités refusant l'interview

avec nc + ni + rf = nr

« Nonsampling errors » : Erreur de mesure

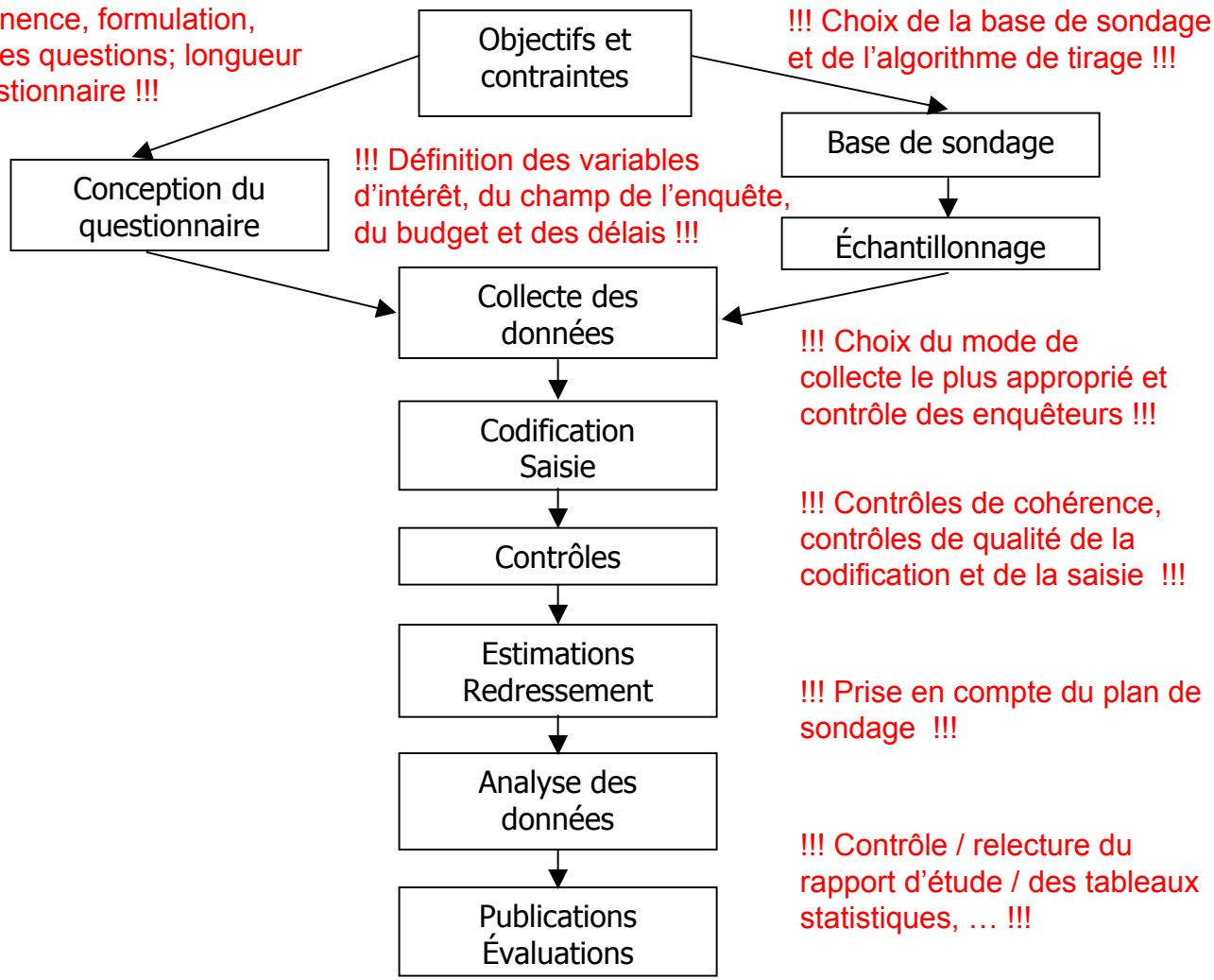
- Il y a erreur de mesure lorsque la valeur de la variable d'intérêt collectée pour un individu est différente de la vraie valeur attachée à ce même individu. Quelques cas (liste non ordonnée et non exhaustive !) :

- Questions faisant appel à la mémoire des personnes interrogées
- Questions portants sur des sujets sensibles (revenus, comportements sexuels, consommation de drogues, …)
- Mécanismes psychologiques liés à l'interaction enquêteur/enquêté
- Interprétation des réponses de la part de l'enquêteur
- « Suggestions » de l'enquêteur à l'enquêté
- Mauvaise compréhension de la question (surtout en cas de traduction des questions depuis une langue étrangère)
- Formulation de la question, effets d'ordre, …
- Fatigue due à la durée d'interviews
- Autres effets enquêteur : le sexe, l'âge de l'enquêteur, sa façon de se présenter … ne sont pas sans conséquences sur la qualité des réponses obtenues

Sources d'erreur et phases d' enquête

- Couverture
- Non-réponse
- Échantillonnage
- Erreurs de mesure

- Saisie
- Codification
- Traitement
- Présentation



L' «art» du sondeur

- La théorie statistique nous aide à mesurer et à réduire l'erreur d'échantillonnage
- L'«art» du sondeur, praticien d'enquête, consiste à juger de l'importance du non mesurable
- La pratique de cet « art » requière la compréhension
  - des causes qui sont à l'origine des erreurs
  - de leur importance relative
  - des effets générés
  - des coûts relatifs aux efforts de réduction des erreurs
- Juger de l'importance du non mesurable est un « art » qui ne doit pas se transformer en alibi pour arrêter tout effort de modélisation et mesure de l'erreur

- Lecture minimale
  - Ardilly, P. (1994), Les techniques de sondage, Editions Technip, Paris
    - Chapitre I. Aspects universels, principes de base
- Pour aller plus loin
  - Groves, R.M. (1989), Survey errors and survey costs, Wiley, New York
    - Chapitres I,III,IV,VII (si vous n'avez pas la force de le lire en entier)

SONDAGE A PROBABILITÉS INÉGALES

- Les plans simples équiprobables ne sont utilisés qu'en l'absence de toute autre information
- Tirage à probabilités inégales: une manière d'utiliser de l'information auxiliaire
- Infinité de plans à probabilités inégales et sans remise

SONDAGE A PROBABILITÉS INÉGALES

- Estimateur de Horvitz-Thompson ou des valeurs dilatées pour un total:

$$\hat{T} = \sum_{i \in s} a_i y_i = \sum_{i=1}^{N} a_i Y_i \delta_i$$

$$E(\hat{T}) = \sum_{i=1}^{N} a_i Y_i E(\delta_i) = \sum_{i=1}^{N} a_i \pi_i Y_i$$

Pour que $\hat{T}$ soit sans biais: $E(\hat{T}) = \sum_{i=1}^{N} Y_i$

$$a_i \pi_i = 1$$

SONDAGE A PROBABILITÉS INÉGALES

Théorème:

$\hat{T} = \sum_{i \in s} \frac{y_i}{\pi_i}$ est le seul estimateur linéaire sans biais

de T

Pour une moyenne $\overline{Y}$ $\quad \hat{\overline{Y}} = \frac{1}{N} \sum_{i \in s} \frac{y_i}{\pi_i}$

SONDAGE A PROBABILITÉS INÉGALES

Exemple (Ardilly) : nombre d'habitants Y inconnu, nombre de logements X connu.
Estimation du nombre moyen d'habitants par tirage à probabilités proportionnelles au nombre de logements

Communes	Nombre de logements = X	Nombre d'habitants = Y	Probabilité d'inclusion
(1) Antibes	48 812	70 688	0,99
(2) Cagnes	23 227	41 303	0,47
(3) St Laurent du Var	12 383	24 475	0,25
(4) Vence	9 341	15 364	0,19
(5) Villefranche/Mer	4 915	8 123	0,10
Moyenne	19 736	31 991	–

SONDAGE A PROBABILITÉS INÉGALES

Echantillons de deux communes:

Échantillon s	$\hat{\bar{Y}}(s)$	$\bar{y}(s)$ (SAS)
1,2	31 856	55 996
1,3	33 860	47 582
1,4	30 453	43 026
1,5	30 526	39 406
2,3	37 156	32 889
2,4	33 748	28 334
2,5	33 822	24 713
3,4	35 753	19 920
3,5	35 826	16 299
4,5	32 419	11 744
Espérance	31 991	31 991

SONDAGE A PROBABILITÉS INÉGALES

- Si N est inconnu:

$$N = \sum_{i=1}^{N} 1$$

- L'estimateur de N est donc:

$$\hat{N} = \sum_{i \in s} \frac{1}{\pi_i}$$

- D'où:

$$E\left(\sum_{i \in s} \frac{1}{\pi_i}\right) = N$$

SONDAGE A PROBABILITÉS INÉGALES

- Estimateur de Hajek:

$$\hat{\bar{Y}} = \left(\sum_{i \in s} \frac{1}{\pi_i} \right)^{-1} \sum_{i \in s} \frac{y_i}{\pi_i}$$

- Poids aléatoires de somme 1.
- Estimateur légèrement biaisé

SONDAGE A PROBABILITÉS INÉGALES

- Un cas gênant:

$$Y_i = C$$

$$\hat{\bar{y}} = \frac{1}{N}\sum_{i\in s}\frac{Y_i}{\pi_i} = \frac{C}{N}\sum_{i\in s}\frac{1}{\pi_i}$$

Comme $\sum_{i\in s}\frac{1}{\pi_i} \neq N$ alors $\hat{\bar{y}} \neq C$

- Mais: $E(\hat{\bar{y}}) = C$

SONDAGE A PROBABILITÉS INÉGALES

- Variance:

$$V(\hat{T}) = \sum_{i=1}^{N} \frac{Y_i^2}{\pi_i}(1-\pi_i) + \sum_{i \neq j}^{N} \sum \frac{Y_i}{\pi_i} \frac{Y_j}{\pi_j}(\pi_{ij} - \pi_i \pi_j)$$

si n fixe formule de Yates-Grundy :

$$V(\hat{T}) = \frac{1}{2} \sum_{i \neq j}^{N} \sum \left(\frac{Y_i}{\pi_i} - \frac{Y_j}{\pi_j} \right)^2 (\pi_i \pi_j - \pi_{ij})$$

SONDAGE A PROBABILITÉS INÉGALES

- Estimation de la variance (par Horvitz-Thomson):

Première formule:

$$\widehat{V(\hat{T})} = \sum_{i \in s} y_i^2 \frac{1-\pi_i}{\pi_i^2} + \sum_{i \neq j \in s} \sum y_i y_j \frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j \pi_{ij}}$$

peut être < 0

Deuxième formule:

$$\widehat{V(\hat{T})} = \frac{1}{2} \sum_{i,j \in s} \sum \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j} \right)^2 \frac{\pi_i \pi_j - \pi_{ij}}{\pi_{ij}}$$

SONDAGE A PROBABILITÉS INÉGALES

- La formule de Yates Grundy montre que l'on a intérêt à tirer proportionnellement aux valeurs d'une variable auxiliaire X corrélée (positivement!) à Y.
- Intéressant en cas d'effet taille (CA, nb d'employés, bénéfice…)

SONDAGE A PROBABILITÉS INÉGALES

- Calcul des probabilités d'inclusion
  - $$\pi_i = \frac{n x_i}{\sum_{i=1}^{N} x_i}$$
  - Exemple: tirage de 3 individus parmi 6 proportionnellement à

$x_1=300$ $x_2=90$ $x_3=70$ $x_4=50$ $x_5=20$ $x_6=20$

SONDAGE A PROBABILITÉS INÉGALES

- Unités sélectionnées d'office et unités tirées au hasard.
- Infinité de plans de sondage pour des π_i fixés.
- D'après Tillé une bonne procédure de tirage doit vérifier 4 critères:
  1. Exactitude
  2. Taille fixe
  3. Généralité
  4. Sans remise

SONDAGE A PROBABILITÉS INÉGALES

- Contraintes sur les π_{ij}
  - Strictement positives (sinon estimation de variance délicate)
  - Indépendantes de l'ordre du fichier
  - $\pi_{ij} < \pi_i \pi_j$
  - Variance inférieure à celle du plan avec remise
- Facilité de mise en œuvre
  - Algorithme rapide
  - Séquentiel

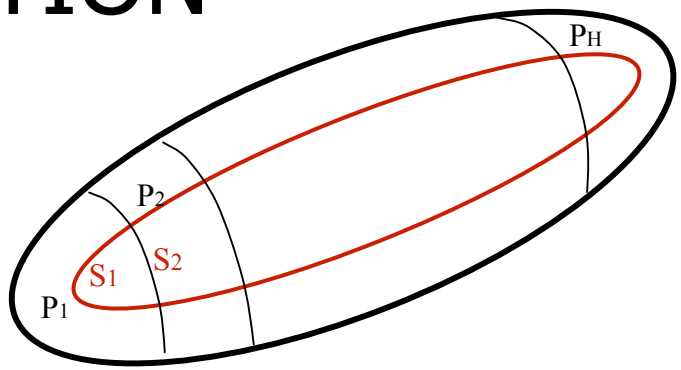
SONDAGE A PROBABILITÉS INÉGALES

- Plus de 50 méthodes! Aucune ne satisfait tous les critères.
- Quelques techniques simples:
  - Tirage avec remise et conservation des unités distinctes mais taille non fixe
  - Rejet de l'échantillon si il y a des doublons mais proba d'inclusion non proportionnelles aux x_i
  - Tirage successif sans remise:
    - On recalcule les probas d'inclusion après tirage de chaque individu. Si j est tiré: $$\pi_i^{'} = \frac{\pi_i}{1-\pi_j}$$
    - Ne respecte pas les probas d'inclusion d'ordre 1

SONDAGE A PROBABILITÉS INÉGALES

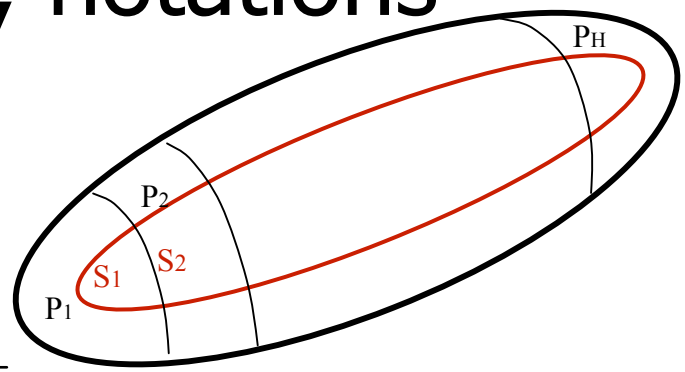
- Sondage systématique à probabilités inégales
- Simplicité
- Inconvénients:
  - certaines probabilités d'inclusion d'ordre 2 peuvent être nulles
  - Dépend de l'ordre du fichier
  - Tri aléatoire avant tirage?

STRATIFICATION



- Utilisation d'une information auxiliaire qualitative
- Toujours efficace

STRATIFICATION, notations



- **Strates:**

$N_1, N_2 \ldots\ldots N_h \ldots\ldots N_H$ $\quad N = \sum N_h$

$\bar{Y}_1, \bar{Y}_2 \ldots\ldots \bar{Y}_h \ldots\ldots \bar{Y}_H$ $\quad \bar{Y} = \sum \frac{N_h}{N} \bar{Y}_h$

$\sigma_1^2, \sigma_2^2 \ldots\ldots \sigma_h^2 \ldots\ldots \sigma_H^2$ $\quad \sigma^2 = \sum \frac{N_h}{N} \sigma_h^2 + \sum \frac{N_h}{N} \left(\bar{Y}_h - \bar{Y} \right)^2$

- **Échantillon:**

$n_1, n_2, \ldots\ldots n_h, \ldots\ldots n_H$ $\quad n = \sum n_h$

$\bar{y}_1, \bar{y}_2, \ldots\ldots \bar{y}_h, \ldots\ldots \bar{y}_H$ $\quad \bar{y} = \sum \frac{n_h}{n} \bar{y}_h$

$\hat{\sigma}_1^2, \hat{\sigma}_2^2, \ldots\ldots \hat{\sigma}_h^2 \ldots\ldots \hat{\sigma}_H^2$

STRATIFICATION

$$\sigma^2 = \sum \frac{N_h}{N} \sigma_h^2 + \sum \frac{N_h}{N} \left(\bar{Y}_h - \bar{Y} \right)^2 = \sigma^2_W + \sigma^2_B$$

- Variance totale=

  moyenne des variances (*variance intra*)

  +variance des moyennes (*variance inter*)

STRATIFICATION

- Estimateur sans biais de $\bar{Y}$ (Horvitz Thomson)

$$\hat{\bar{Y}}_{str} = \sum \frac{N_h}{N} \bar{y}_h$$

- Variance:

$$V(\hat{\bar{Y}}_{str}) = \sum \left(\frac{N_h}{N} \right)^2 V(\bar{y}_h) = \sum_{h=1}^{H} \left(\frac{N_h}{N} \right)^2 \frac{\sigma_h^2}{n_h} \frac{N_h - n_h}{N_h - 1}$$

$$= \frac{1}{N^2} \sum_{h=1}^{H} N_h (N_h - n_h) \frac{S_h^2}{n_h}$$

STRATIFICATION, répartition proportionnelle

- Échantillon dit « représentatif »:

$$\frac{n_h}{n} = \frac{N_h}{N} \Rightarrow \tau_h = \frac{n_h}{N_H} = \frac{n}{N} = \tau$$

- Taux de sondage constant dans chaque strate

$$\hat{\bar{Y}}_{str} = \sum_{h=1}^{H} \frac{N_h}{N} \bar{y}_h = \sum_{h=1}^{H} \frac{n_h}{n} \bar{y}_h = \bar{y} = \hat{\bar{Y}}_{prop}$$

STRATIFICATION, répartition proportionnelle

- variance :

$$V(\hat{\bar{Y}}_{prop}) = \frac{1}{N^2}\sum_{h=1}^{H} N_h(N_h - n_h)\frac{S_h^{\,2}}{n_h} = \frac{1}{N^2}\sum_{h=1}^{H}\frac{N_h - n_h}{n_h} N_h S_h^{\,2}$$

$$= \frac{1}{N^2}\sum_{h=1}^{H}\left(\frac{N_h}{n_h} - 1\right)N_h S_h^{\,2} = \frac{1}{N^2}\sum_{h=1}^{H}\left(\frac{N}{n} - 1\right)N_h S_h^{\,2} = \frac{N-n}{nN}\sum_{h=1}^{H}\frac{N_h}{N} S_h^{\,2}$$

- Si τ est faible:

$$V(\hat{\bar{Y}}_{prop}) = \frac{N-n}{nN}\sum_{h=1}^{H}\frac{N_h}{N} S_h^{\,2} \simeq \frac{N-n}{nN}\sum_{h=1}^{H}\frac{N_h}{N}\sigma_h^{\,2} = \frac{N-n}{N}\frac{\sigma^2_{\,w}}{n}$$

STRATIFICATION, répartition proportionnelle

- Variance de l'estimateur du SAS sans remise:

$$V(\bar{y}) = (1 - \frac{n}{N})\frac{S^2}{n} = \frac{N-n}{N}S^2 \simeq \frac{N-n}{N}\sigma^2$$

- Avec les mêmes probabilités d'inclusion d'ordre 1, l'échantillon stratifié représentatif est plus efficace qu'un échantillon simple de même taille dès que les $\bar{Y}_h$ sont différents.

STRATIFICATION optimale

- Répartition optimale:

$$V(\widehat{\overline{Y}_{str}}) = \frac{1}{N^2}\sum \frac{N_h\left(N_h - n_h\right)}{n_h} S_h^{\,2}$$

avec $S_h^{\,2} = \frac{N_h}{N_h - 1}\sigma_h^{\,2}$

c_h – coût unitaire d'une observation

$$\begin{cases} \min \sum \frac{N_h\left(N_h - n_h\right)}{n_h} S_h^{\,2} \\ \sum n_h c_h = c_0 \end{cases}$$

$$\sum \frac{N_h^{\,2}}{n_h} S_h^{\,2} - \underbrace{\sum N_h S_h^{\,2}}_{fixe}$$

STRATIFICATION optimale

- Solution:

$$\frac{N_h^2 S_h^2}{n_h^2}$$ proportionnel à c_h

$$\frac{n_h}{N_h} \propto \frac{S_h}{\sqrt{c_h}}$$

Si c_h constant:

$$n_h = n \frac{N_h S_h}{\sum N_h S_h}$$ - Répartition de Neyman

STRATIFICATION

- Exemple n° 1: présondage de 155 unités

Strates	**1**	**2**	**3**	**4**	
N_h	3750	3272	1387	2475	10 884
n_h	50	45	30	30	155
$\overline{y}_h$	12.6	14.5	18.6	13.8	
σ_h^2	2.8	2.9	4.8	3.2	

STRATIFICATION

- Exemple n° 1:

$$\widehat{\overline{Y}} = \sum \left(\frac{N_h}{N} \right)^2 \overline{y}_h = \frac{3750 \text{x} 12.6 + ... + 2475 \text{x} 13.8}{10884} = 14.21$$

$$V(\widehat{\overline{Y}}) \simeq \sum \left(\frac{N_h}{N} \right)^2 \frac{\widehat{\sigma}_h^{\,2}}{n_h} = 0.02059 = (0.14)^2$$

Intervalle de confiance à 95% pour $\overline{Y}$:

$14.21 \pm 2 \text{x} 0.14$ soit: $\left[13.93 < \overline{Y} < 14.49 \right]$

Pour T: 154662 ± 3047

STRATIFICATION

- Exemple n° 1:

$$\sigma^2 = \sum \frac{N_h}{N} \sigma^2{}_h + \sum \frac{N_h}{N} \left(Y_h - Y \right)^2$$

On estime: σ^2_h par $\frac{n_h}{n_{h-1}} \hat{\sigma}^2{}_h$

$\overline{Y}_h$ par $\overline{y}_h$

$\overline{Y}$ par

$$\hat{\sigma}^2 = 6.06 = \left(2.46 \right)^2$$

STRATIFICATION

- Suite: Répartition de Neyman pour n=1000:

$N_1S_1 = 6275$ $n_1 = 1000 \times 6275/19\,312 = 325$

$N_2S_2 = 5572$ $n_2 = 288$

$N_3S_3 = 3038$ $n_3 = 157$

$N_4S_4 = 4427$ $n_4 = 229$

$19\,312$

Variance: $\frac{1}{N^2}\sum\frac{N_h\left(N_h - n_h\right)}{n_h}S_h^2 = 0.0029 = \left(0.0542\right)^2$

$\overline{Y}$ connu à $\pm$ 2 x 0.0542 soit $\pm$ 0.108

T connu à $\pm$ 1179

STRATIFICATION

- Échantillon simple à 1000:

$$\frac{\sigma^2}{n} \times \frac{N-n}{N-1} = 0.0055 = (0.0742)^2$$

$\overline{Y}$ connu à $\pm$ 0.15; T connu à $\pm$ 1615

- Échantillon stratifié représentatif:

$n_1 = 345$

$n_2 = 301$

$n_3 = 127$

$n_4 = 227$

STRATIFICATION

- Comment stratifier?
  - Remarque préalable: dans un sondage à probabilité inégale π_i proportionnel à Y_i annule la variance.
  - Nombre de strates: le maximum mais…
  - Répartition dans les strates:
    - Si S_h inconnu : répartition proportionnelle
    - si S_h connu: Neyman
    - sinon, hypothèse fréquente $\frac{S_h}{\overline{Y}_h} = c$ d'où n_h proportionnel à la somme de la variable étudiée ou d'une variable corrélée.
    - Exemple: échantillon d'entreprises proportionnel au CA ou à l'effectif de la strate.

STRATIFICATION

- Variable de stratification: en théorie Y; sinon, variable bien corrélée avec Y.

- Limites de strates optimales:

  méthode de Dalenius et Hodges. Regrouper des classes selon le cumul de la racine des effectifs

STRATIFICATION

- Estimation d'une proportion p
- Même démarche: une proportion est une moyenne particulière

$$\hat{p}_{str} = \sum_{h=1}^{H} \frac{N_h}{N} f_h$$

$$V(\hat{p}_{str}) = \sum_{h=1}^{H} \left(\frac{N_h}{N} \right)^2 \frac{p_h(1-p_h)}{n_h} \frac{N_h - n_h}{N_h - 1}$$

$$\hat{V}(\hat{p}_{str}) \simeq \sum_{h=1}^{H} \left(\frac{N_h}{N} \right)^2 \frac{f_h(1-f_h)}{n_h} \left(1 - \frac{n_h}{N_h} \right)$$

SONDAGE A DEUX DEGRÉS

- Définition: tirage de m unités primaires puis de n_i unités secondaires

- Avantages:
  - frais de déplacement réduit
  - absence de liste autorisée
- Mais: précision moindre: effet de grappe.

- M unités primaire de taille N_i

$$N = \sum_{i=1}^{M} N_i \qquad T_i = \sum_{j=1}^{N\,i} Y_{ij}$$ - total de l'UP n°i

SONDAGE A DEUX DEGRÉS

Tirage aléatoire simple à chaque degré.

$$\hat{T} = \frac{M}{m} \sum_{i \in s} \left(\frac{N_i}{n_i} \sum_{j \in S_i} y_{ij} \right)$$

Remarque: inutile de connaître N pour estimer T.

$$V\left(\hat{T}\right) = \underbrace{M^2 \left(1 - \frac{m}{M}\right) \frac{S_1^2}{m}}_{\substack{Variance \\ inter \\ UP}} + \underbrace{\frac{M}{m} \sum N_i^2 \left(1 - \frac{n_i}{N_i}\right) \frac{S_{2,i}^2}{n_i}}_{\substack{Variance \\ intra \\ UP}}$$

$$S_1^2 = \frac{1}{M-1} \sum_{i=1}^{M} \left(T_i - \bar{T}\right)^2 \qquad S_{2,i}^2 = \frac{1}{N_i - 1} \sum_{i=1}^{N_i} \left(Y_{ij} - \bar{Y}_i\right)^2$$

SONDAGE A DEUX DEGRÉS

- S_1^2 estimé par $s_1^2 = \frac{1}{m-1}\sum_{i=1}^{m}\left(\hat{T}_i - \frac{\hat{T}}{M}\right)^2$
- idem pour $S^2_{2,i}$
- Remarque:
  - Si n_i proportionnel à N_i: taille d'échantillon aléatoire

$$n_i = n_0 \frac{N_i}{N} \qquad E(n_s) = E\left(\sum_{i \in S_i} n_0 \frac{N_i}{N}\right) = \sum_{k \in U_i} n_0 \frac{N_i}{N} \frac{m}{M} = \frac{n_0 m}{M}$$

SONDAGE A DEUX DEGRÉS

- Sondage autopondéré:

  - m unités primaires tirées à probabilités proportionnelles à leur taille
  - tirage d'échantillons de taille fixe n_0
  - probabilités d'inclusion constantes

$$\pi_i = \frac{N_j}{N} m \frac{n_0}{N_j} = \frac{m n_0}{N}$$

- Estimateur de la moyenne: N peut être inconnu

$$\hat{\bar{Y}} = \bar{y}$$

CAS PARTICULIER: SONDAGE EN GRAPPES

- Définition: toutes les US sont observées dans les UP tirées.

Nécessité de grappes: hétérogènes
de faible taille
nombreuses
de tailles voisines

Le tirage systématique est un tirage d'une grappe.

SONDAGE EN GRAPPES

- Cas général :tirage de grappes à probabilités inégales

  - Estimation du total: $$\hat{T} = \sum_{i=1}^{m} \frac{T_i}{\pi_i}$$

  - Estimation d'une moyenne $$\hat{\bar{y}} = \frac{1}{N} \sum_{i=1}^{m} \frac{N_i \bar{Y}_i}{\pi_i}$$

pb si N inconnu: utiliser l'estimateur de Hajek

SONDAGE EN GRAPPES

- Tirage de grappes à probabilités égales

$$\pi_i = \frac{m}{M}$$

  - taille d'échantillon aléatoire

$$E(n_s) = E(\sum_{i \in S_i} N_i) = \sum_{k \in U_i} N_i \frac{m}{M} = \frac{Nm}{M}$$

$$\hat{T} = \frac{M}{m} \sum_{i \in s} T_i \qquad V(\hat{T}) = M^2 \left(1 - \frac{m}{M}\right) \frac{S_1^2}{m} \qquad \hat{\bar{y}} = \frac{M}{m} \frac{1}{N} \sum_{i=1}^{m} N_i \bar{Y}_i$$

SONDAGE EN GRAPPES

- Tirage de grappes à probabilités proportionnelles à la taille

$$\pi_i = m\frac{N_i}{N} \qquad \hat{\bar{y}} = \frac{1}{m}\sum_{i=1}^{m}\bar{y}_i$$

$$\hat{V}(\hat{\bar{y}}) = \frac{1}{m(m-1)}\sum_{i=1}^{m}(1-m\frac{N_i}{N})(\bar{y}_i - \hat{\bar{y}})^2$$

$$E(n_s) = E(\sum_{i\in S_i} N_i) = \sum_{i\in U_i} N_i \frac{N_i m}{M} = \frac{m}{M}\sum_{i=1}^{M} N_i^2$$

MÉTHODES DE REDRESSEMENT OU DE PONDÉRATION

- Principe:

utiliser a posteriori une information supplémentaire corrélée avec la variable à étudier.

- Information:

variables de contrôle dont on connaît soit des caractéristiques globales, des caractéristiques par classes, pour chaque individu.

MÉTHODES DE REDRESSEMENT OU DE PONDÉRATION

- Estimation par le quotient ou redressement par variable quantitative

Exemple:

- Échantillon de 80 hypermarchés - On veut estimer le CA moyen $\overline{Y}$
- On a $\overline{y} = 110,2 MF$
- On sait que le nombre moyen $\overline{X}$ de caisses dans la population des hypermarchés est 28.
- Dans l'échantillon $\overline{x} = 28.8$

$$\hat{\overline{Y}} = 110.2 \times \frac{28}{28.8} = 107.1$$

Estimation par le quotient

Formule générale: $$\bar{y}_q = \bar{y}\frac{\bar{X}}{\bar{x}}$$

Remarque: en général estimation biaisée, mais biais négligeable si n>1000.

Hypothèse de proportionnalité (règle de 3)

Estimation par le quotient

- Calcul du biais:

$$\overline{y_q} = \overline{X}\frac{\overline{y}}{\overline{x}} = \overline{X}\frac{\overline{y}-\overline{Y}+\overline{Y}}{\overline{x}-\overline{X}+\overline{X}} = \overline{Y}\frac{1+\dfrac{\overline{y}-\overline{Y}}{\overline{Y}}}{\underbrace{1+\dfrac{\overline{x}-\overline{X}}{\overline{X}}}_{\varepsilon}}$$

Développement limité:

$$\overline{y_q} \simeq \overline{Y}\left(1+\frac{\overline{y}-\overline{Y}}{\overline{Y}}\right)\left[1-\frac{\overline{x}-\overline{X}}{\overline{X}}+\left(\frac{\overline{x}-\overline{X}}{\overline{X}}\right)^2\right] \simeq \overline{Y}\left[1+\frac{\overline{y}-\overline{Y}}{\overline{Y}}-\frac{\overline{x}-\overline{X}}{\overline{X}}\times\frac{\overline{y}-\overline{Y}}{\overline{Y}}-\frac{\overline{x}-\overline{X}}{\overline{X}}+\left(\frac{\overline{x}-\overline{X}}{\overline{X}}\right)^2\right]$$

$$E\left(\bar{y}_q\right) \simeq \bar{Y}\left[1 - \frac{\mathrm{cov}\left(\bar{x};\bar{y}\right)}{\bar{X}\bar{Y}} + \frac{V\left(\bar{x}\right)}{\bar{X}^2}\right]$$

Si probabilité égale et sans remise:

$$E\left(\bar{y}_q\right) = \bar{Y} + \frac{N-n}{Nn}\bar{Y}\left[\frac{s_x^2}{\bar{X}^2} - \frac{\mathrm{cov}\left(x;y\right)}{\bar{X}\bar{Y}}\right]$$

Biais en 1/n.

Biais nul si la droite de régression passe par 0.

- Erreur quadratique moyenne

$$E(\bar{y}_q - \bar{Y})^2 = \frac{N-n}{Nn}\left(s_y^2 - 2\frac{\bar{Y}}{\bar{X}} s_{xy} + (\frac{\bar{Y}}{\bar{X}})^2 s_x^2 \right) \quad \text{estimé par} \quad \frac{N-n}{Nn}\frac{1}{n-1}\sum_{i=1}^{n} z_i^2$$

Avec $z_i = y_i - rx_i \quad \text{où} \quad r = \frac{\bar{y}}{\bar{x}}$

Complément: estimation d'un ratio

- Exemple: Tirage de n exploitations agricoles (élevage): X_i nombre de vaches, Y_i production
- Rendement par vache: $R = \frac{\bar{Y}}{\bar{X}}$ estimé par $r = \frac{\bar{y}}{\bar{x}}$
- Rapport de deux variables aléatoires
- Développement limité

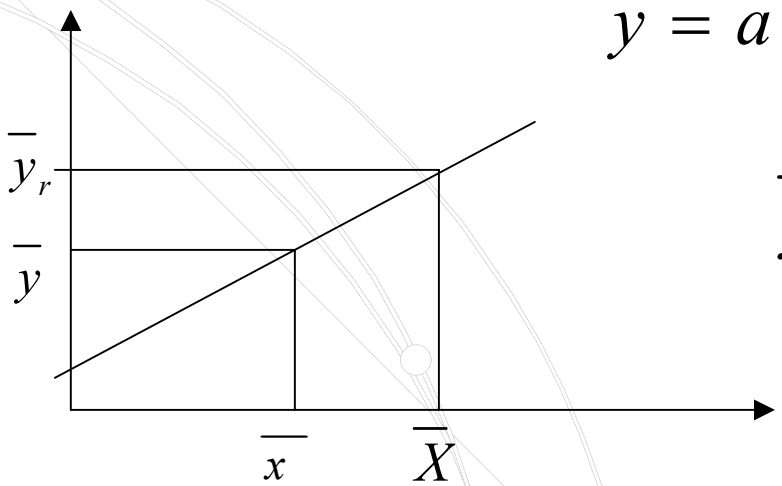
$$E(r) \simeq R + \frac{N-n}{Nn} R \left(\frac{s_x^2}{\bar{X}^2} - \frac{s_{xy}}{\bar{X}\bar{Y}} \right)$$

Estimation par la régression

On connaît pour chaque individu de l'échantillon une variable de contrôle x_i et aussi la valeur moyenne $\overline{X}$ sur la population .

Hypothèse:

$$y = a + bx$$



$$\overline{y}_r = \overline{y} + b\left(\overline{X} - \overline{x}\right)$$

Post-stratification; redressement sur critère qualitatif

Exemple:

n=1000; on veut estimer le pourcentage de fréquentation du cinéma.

On s'aperçoit que la fréquentation du cinéma est liée à la posssession de TV.

On sait que $\tau_{télé} = 80\%$.

Post-stratification; redressement sur critère qualitatif

Cinéma / Tele	Oui	Non	Total
Oui	20	680	700
Non	80	220	300
Total	100	900	

(800) X 8/7

(200) x 2/3

Après redressement:

Cinéma / Tele	Oui	Non	Total
Oui	23	777	800
Non	53	147	200
Total	76	924	

Généralisation: calage sur marges

- Redressement sur plusieurs critères
  - Méthode itérative de Deming et Stephan (RAS)

    On ajuste alternativement sur chaque marge (succession de règles de 3)
  - Macro CALMAR de l'INSEE

Post-stratification pour une variable numérique

$$\hat{T}_{post} = \sum N_h \bar{y}_h \qquad \hat{\bar{y}}_{post} = \frac{1}{N} \sum N_h \bar{y}_h$$

Exemple: enquête concernant les revenus

X=classe d'âge; Y=revenu

<20	21-35	36-50	>50
15%	30%	30%	25%
6000	9000	15.000	12.000

$\bar{y} = 11.100$

On sait que les proportions sont:

20 35 30 15

$\hat{\bar{y}}_{post} = 10650$

Post-stratification pour une variable numérique

$$V\left(\hat{\bar{y}}_{post}\right) = \underbrace{V\left[E\left(Y/n_h\right)\right]}_{0} + E\left[V\left(Y/n_h\right)\right]$$

Conditionnellement aux n_h:

$$\sum\left(\frac{N_h}{N}\right)^2 V\left(\bar{y}_h\right) = \sum\left(\frac{N_h}{N}\right)^2 \frac{N_h - n_h}{N_h n_h} S_h^2 = \sum\left(\frac{N_h}{N}\right)^2 S_h^2 \frac{1}{n_h} - \frac{1}{N}\sum\left(\frac{N_h}{N}\right) S_h^2$$

En prenant l'espérance:

$$\sum\left(\frac{N_h}{N}\right)^2 S_h^2 E\left(\frac{1}{n_h}\right) - \frac{1}{N}\sum\left(\frac{N_h}{N}\right) S_h^2$$

Calcul de $E\left(\frac{1}{n_h}\right)$

$$P_h = \frac{N_h}{N} \qquad p_h = \frac{n_h}{n}$$

$$n_h = n\frac{n_h}{n} = np_h = n\left(p_h - P_h + P_h\right) = nP_h\left(1 + \frac{p_h - P_h}{P_h}\right)$$

Développement limité

$$\frac{1}{n_h} = \frac{1}{n\,P_h} \times \frac{1}{1 + \underbrace{\frac{p_h - P_h}{P_h}}_{\varepsilon}}$$

$$\frac{1}{n_h} \simeq \frac{1}{nP_h}\left[1 - \varepsilon + \varepsilon^2\right] = \frac{1}{nP_h}\left[1 - \frac{p_h - P_h}{P_h} + \left(\frac{p_h - P_h}{P_h}\right)^2\right]$$

En prenant l'espérance :

$$E(p_h) = P_h \qquad V(p_h) = \frac{N-n}{Nn} P_h (1-P_h)$$

$$E\left(\frac{1}{n_h}\right) = \frac{1}{nP_h}\left[1 + \frac{N-n}{Nn^2} \times \frac{Q_h}{P_h}\right]$$

$$V(\bar{y}_{post}) = \sum P_h^2 S_h^2 \left[\frac{1}{nP_h} + \frac{N-n}{Nn^2} \times \frac{Q_h}{P_h}\right] - \frac{1}{N}\sum P_h S_h^2$$

$$= \frac{N-n}{Nn}\sum P_h S_h^2 + \frac{1}{n}\frac{N-n}{Nn}\sum Q_h S_h^2$$

$$V(\bar{y}_{post}) = \left(\frac{1-f}{n}\right)\sum \frac{N_h}{N} S_h^2 + \frac{1-f}{n^2}\sum\left(1 - \frac{N_h}{N}\right) S_h^2$$

- Pour avoir une bonne post-stratification
  - Variable de redressement bien corrélée
  - N grand
  - $(N-N_h)/N$ petit donc grandes strates
  - Effectifs N_h connus

MÉTHODES DE REDRESSEMENT OU DE PONDÉRATION

Remarque:

- ne pas utiliser que des variables socio-décisionnelles;
- dangers de redressement sur critères multiples.

Propriétés:

- l'estimateur est sans biais, mais il faut connaître avec certitude les poids des strates.
- sa variance est plus petite si le critère de post-stratification est très lié à la variable d'intérêt, si n est grand et s'il n'y a pas trop de strates.

Questions sensibles ou indiscrètes: la méthode des questions aléatoires

Première technique:

On tire ou sort dans une urne avec θ boules blanches et 1- θ boules noires la question

Si blanc: question A: « Avez-vous fraudé le fisc? »

Si noire: question $\overline{A}$: « Je n'ai pas fraudé »

On veut estimer P_A.

On recueille Π = Proba de Oui = $\theta P_A + (1-\theta)(1-P_A)$

$\widehat{\Pi}$ % de « Oui »

$$\widehat{P}_A = \frac{\widehat{\Pi} - (1-\theta)}{2\theta - 1}$$

$$V\left(\widehat{P}_A\right) = \frac{1}{(2\theta-1)^2} V\left(\widehat{\Pi}\right) \simeq \frac{P_A(1-P_A)}{n} + \frac{1}{n} \frac{\theta(1-\theta)}{(2\theta-1)^2}$$

Inconvénient: $\overline{A}$ aussi indiscrète que A!

Deuxième technique:

Si blanche, question A sensible

Si noire, question B banale

$$\Pi = P_A\theta + P_B(1-\theta) \qquad \widehat{P}_A = \frac{\widehat{\Pi} - (1-\theta)P_B}{\theta}$$

$$V(\widehat{P}_A) \simeq \frac{\Pi(1-\Pi)}{n\theta^2} + \frac{P_B(1-P_B)(1-\theta)^2}{n\theta^2}$$

P_B peut être connu à l'avance ou estimé par une autre enquête.

Exemple:

A: combien de fois avez-vous avorté?

B: nombre idéal d'enfants?

Effets et pratique des redressements

Effets et pratique des redressements

- Précisions de langage
- Redresser pour quoi faire ?
- Une pratique qu'il ne faut pas banaliser
- Redresser sur quoi et comment
  - Le choix des critères
  - Les contrôles à opérer
- La pratique des redressements
  - dans les études marketing
  - dans les études politiques
- Peut-on se fier aux redressements ?
- Bibliographie

Précisions de langage

- « Extrapolation »
  - Le poids comme coefficient d'extrapolation : passage des « effectifs échantillon » aux « effectifs population »
- « Pondération »
  - Redistribution de poids à effectif échantillon constant, visant à corriger une sur/sous -pondération de strates *décidée lors de l'établissement du plan de sondage*
- « Redressement »
  - Redistribution de poids à effectif échantillon constant - généralement fondée sur des critères multiples -, visant à corriger une sur/sous -représentation de catégories de la population *constatée a posteriori*

Redresser pour quoi faire ? {1/2}

- Prise en compte du plan de sondage
  - Pondération de strates d'échantillon
  - Pondération selon la taille des unités primaires (eg ménages/individus)
- Prise en compte d'informations sur la population (post-stratification)
  - Correction de distorsions dues à des erreurs de non-observation (erreurs de couverture et/ou de non-réponse)
  - Ce type de correction est plus courant pour les échantillons non-probabilistes (eg quota), ou dans les échantillons probabilistes entachés d'importants erreurs de non-observation

Redresser pour quoi faire ? {2/2}

- Ne pas redresser revient à attribuer aux non-répondants le comportement moyen de l'ensemble des répondants, ce qui constitue souvent une grossière erreur
- Il est bien connu que les non-répondants se trouvent plus particulièrement dans des catégories sociales spécifiques (personnes âgées, femmes, personnes à faible niveau d'instruction, …)
- D'habitude il est préférable attribuer aux non-répondants le comportement moyen des répondants appartenant aux mêmes catégories sociales

Une pratique qu'il ne faut pas banaliser

- Le redressement est trop souvent considéré comme une simple étape « informatique », permettant de caler mécaniquement la structure de l'échantillon sur celle de la population étudiée
- Cela fini par devenir une pratique de « maquillage d'échantillon », ayant pour but de corriger les écarts entre quotas demandés et quotas réalisés
- Comme toute autre phase de l'enquête, le redressement doit être préparé en amont : il faut penser à poser les bonnes questions, codées de façon homogène aux données de référence les plus récentes, en prenant garde aux unités statistiques (ménages vs individus, entreprises vs établissements, …)

Le choix des critères

- Les variables de redressement doivent être le plus corrélées possible aux thématiques de l'étude (afin de réduire la variance des estimateurs) : des méthodes de segmentation (eg CHAID) sont parfois utilisées dans leur sélection, mais la plupart des fois quelques bons tris croisés suffisent

- Les variables de redressement doivent être peu nombreuses, et doivent être agrégées de façon pertinente (afin d'éviter des effets mal maîtrisés)

- Les non-répondants aux questions utilisées dans le redressement doivent être éliminés ou laissés à leur poids (éviter des hypothèses trop fortes à leur égard)

Les contrôles à opérer

- Il est important d'opérer une validation préalable de la structure brute d'échantillon, sur un ensemble de variables critiques, qu'elles aient fait l'objet de quotas ou qu'elles soient utilisées comme simples variables de contrôle
- Bien sûr les variables à utiliser dépendent du sujet de l'étude : nombre de personnes au ménage, présence d'enfants, type et équipement du logement, « restitution » du vote à une élection antérieure, …
- Après redressement, il faut vérifier la distribution des poids générés : min, max, quantiles et courbes de fréquence, indicateurs de forme du type

$$100 * (\sum \text{poids})^2 / n \sum \text{poids}^2$$

(*) Cela vaut 100 si tous les poids sont égaux, entre 50 et 70 s'il y a une forte dispersion; à moins de 50 le redressement est à revoir …

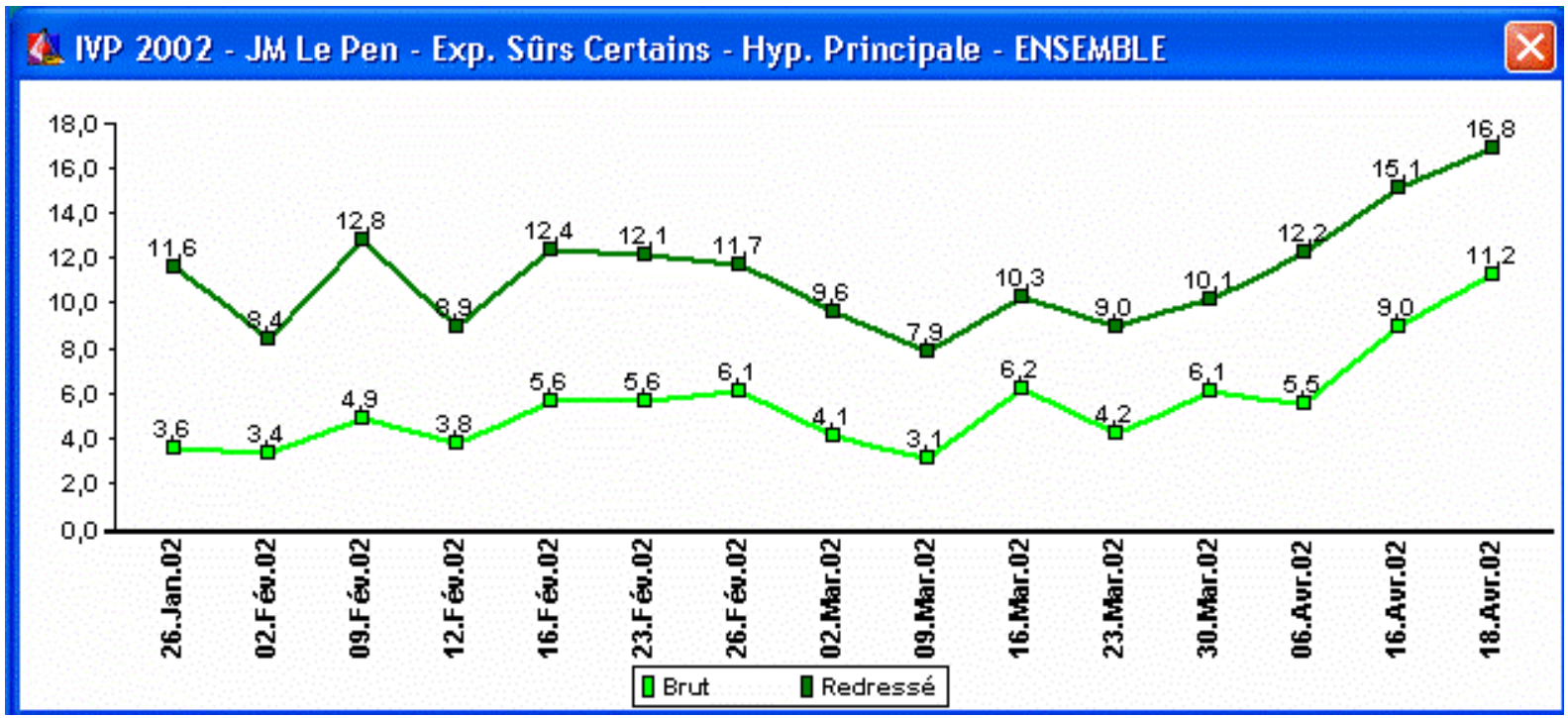
La pratique des redressements dans les études marketing

- Région croisée par tranche d'unité urbaine, sexe, âge, CSP de la « personne de référence » ou de l'individu, présence d'enfants, niveau d'études, pratiques médias … sont les variables le plus souvent utilisées dans les études marketing

- Le plus important c'est de :
  - ne pas jouer au « perroquet »
  - toujours utiliser des données de référence fiables et à jour
  - veiller à redresser en plusieurs étapes s'il le faut : d'abord une première pondération - eg ménage ou pays -, ensuite un calage sur marges portant sur les variables « individu »
  - rester aussi critiques que possible sur les éventuels erreurs de mesure commises

La pratique des redressements dans les études politiques

- Ce qui a été dit pour les études marketing reste bien sûr valable
- Pour le reste, un exemple vaut mieux que mille discours :



Peut-on se fier aux redressements ?

- Le redressement est indispensable
  - Correction des erreurs de non-observation
  - Standardisation des structures à des fins de comparaison
- Le redressement ne peut pas corriger les erreurs de mesure
  - Déclarations de revenus
  - Restitutions et intentions de vote
  - …
- Le redressement peut augmenter les biais
  - Poids délirants > limitation des poids (eg. de 0,25 à 4,00)
  - Disponibilité de données de référence fiables et récentes, codées de façon homogène

Bibliographie

- Ardilly, P. (1994), Les techniques de sondage, Editions Technip, Paris
  - Chapitre III. Amélioration des estimateurs (redressements)
- Lejeune, M., éd. (2001), Traitements des fichiers d'enquêtes. Redressements, injections de réponses, fusions, PUG, Grenoble
- Brossier, G., Dussaix,A.-M., éd. (1999), Enquêtes et sondages, Dunod, Paris
  - Chapitre 5. Méthodes de redressement et de calage

La méthode des quotas

La méthode des quotas

Principe de la méthode
Point de départ et but recherché

Le point de départ : toutes les méthodes d'échantillonnage aléatoire supposent l'existence d'une base de sondage à partir de laquelle on tire aléatoirement (mais avec probabilité connue) un échantillon sans biais dont la taille a été déterminée à la suite de considérations sur le niveau de précision souhaité

Or, pour la majorité des enquêtes d'opinion comme des études de marché on ne dispose pas de base de sondage

Le but recherché : il s'agit de se rapprocher le plus possible d'un tirage rigoureusement aléatoire

Principe de la méthode
Caractère « intuitif » de la méthode

- On suppose que si l'échantillon reproduit fidèlement certaines caractéristiques de la population étudiée (et peut donc être considéré, par abus de langage, « représentatif »), alors il sera également à même de reproduire d'autres caractéristiques non contrôlées et/ou contrôlables qui constituent l'objet même de l'enquête

- … si la population se compose de 50% d'hommes, on imposera à l'enquêteur chargé de réaliser 10 interviews un quota de 5 hommes pour 10 personnes enquêtées … si la même population comporte 10% d'agriculteurs, il devra y avoir une et une seule interview d'agriculteur …

Principe de la méthode
A quoi ressemble une feuille de quotas ?

10 interviews Auprès d'électeurs inscrits	Répartition des interviews	1 2 3 4 5 6 7 8 9 10
Sexe		
Homme	5	1 2 3 4 5
Femme	5	1 2 3 4 5
Age		
18 – 34 ans	3	1 2 3
35 – 49 ans	2	1 2
50 – 64 ans	2	1 2
65 ans et plus	3	1 2 3
CS de la personne interrogée		
Agriculteur	1	1
Artisan / Petit commerçant	1	1
Prof. Lib. / Cadre supérieur	1	1
Prof. Intermédiaire, employé, ouvrier	4	1 2 3 4
Retraité, étudiants, autre inactif	3	1 2 3

Principe de la méthode
La recherche des personnes à interroger

- La recherche peut être d'autant plus longue que l'on approche la fin de la feuille : la dernière personne est déterminée de manière unique par les modalités restantes

- Tout le métier de l'enquêteur consiste à ne pas se faire piéger et réaliser correctement ses « fin de quotas »

- Définir des quotas revient à définir une stratification multiple sur la population. La différence avec l'échantillon probabiliste est que au lieu de tirer les unités de sondage on laisse à l'enquêteur le soin de les trouver lui-même au hasard de ses pérégrinations (cela prend un sens un peu différent en face à face et au téléphone)

Le biais est créé par les conditions mêmes du travail de l'enquêteur

A différentes heures de la journée les différentes catégories de population présentent des probabilités différentes et inconnues d'être touchées par l'enquêteur

La probabilité d'être touché varie également avec l'accessibilité des personnes à interroger : digicodes à l'entrée des immeubles en face à face, numéros sur liste rouge au téléphone …

Critique de la méthode
Impossibilité de calculer l'erreur d'échantillonnage

La probabilité qu'a un individu de la population d'appartenir à l'échantillon est inconnue : il est alors impossible d'évaluer la variance d'échantillonnage et donc de mesurer la précision des estimations

Deux réactions possibles :

- De nombreux auteurs considèrent que cette méthode est inutilisable
- D'autres auteurs, faute de mieux, adoptent l'hypothèse d'un tirage à probabilités égales; hypothèse qui n'est, vraisemblablement, jamais vérifiée

Critique de la méthode
Si c'est aussi « compliqué » …

- … les estimateurs employés sont biaisés;
- … les calculs de taille d'échantillon que l'on réalise en employant les formules du sondage à probabilités égales ne peuvent être que des approximations plus ou moins grossières
- … le plan de sondage doit être accompagné d'une série de consignes données à l'enquêteur visant à la fois à :
  - réduire le biais d'observation;
  - se rapprocher le plus possible des conditions de tirage à probabilités égales

Critique de la méthode
… pourquoi continue-t-on ?

- … ce n'est pas parce que l'on ne connaît pas la précision d'une estimation que cette estimation est mauvaise

- … de façon empirique nous avons d'innombrables exemples de résultats issus d'échantillons par quotas fort comparables à ceux fournis par des échantillons aléatoires

Enquête par quotas et sondage aléatoire
Une étude empirique (1/2)

- En 1953, à l'initiative de la London School of Economics, méthode aléatoire et méthode par quotas ont été comparées. L'échantillon aléatoire était tiré à partir des listes électorales, l'autre échantillon devait respecter trois quotas : le sexe, l'âge regroupé en quatre classes, la classe sociale en trois postes.

- Le questionnaire était le même dans les deux enquêtes et portait sur des variables socio-démographiques, les niveaux de revenu et d'instruction, les loisir.

Enquête par quotas et sondage aléatoire
Une étude empirique (2/2)

La comparaison des résultats des deux enquêtes a montré que dans ce cas :

- L'échantillon sur quotas donnait des estimations plus biaisées sur les variables socio-démographique que sur les variables purement sociologiques (loisirs, consommation)
- Pour ces variables sociologiques, si le biais était très faible, en revanche il est presque toujours dans le sens de la surestimation
- En l'absence de quotas sur le secteur économique, il y a sous-estimation des travailleurs de l'industrie

Enquête par quotas et sondage aléatoire
Quelques enseignements

- Il est important de contrôler les variables socio-démographiques susceptibles d'être corrélées aux variables d'intérêts

- Le secteur d'activité économique doit également être contrôlé, en imposant des quotas à priori ou par post-stratification (redressement)

- Il faut toujours se méfier du syndrome du perroquet, lorsqu'il s'agit de définir les variables que l'on souhaite contrôler par des quotas

Réalisation du plan de sondage
La nécessité de consignes précises

- Afin de canaliser les agissements de l'enquêteur la feuille de quotas doit être accompagnée par des consignes précises, visant à se rapprocher le plus possible des conditions d'un tirage à probabilités égales

- Il s'agit de rapprocher autant que possible les différentes probabilités que les individus ont d'être interrogés : par exemple, pour une enquête comportant des interviews d'actifs, il est important de travailler en semaine après 18h00, les samedis et dimanches …

Ce qui est demandé à l'enquêteur professionnel

- Savoir éviter les refus
- Être disponible pour un travail sans horaires fixes
- Savoir éviter d'interroger , dans une zone donnée, des personnes se ressemblant trop ou vivant dans les mêmes conditions
- Ne pas hésiter à renoncer à une interview si la personne contactée ne correspond pas aux quotas
- Respecter les consignes de dispersion géographique des interviews
- Brasser large à l'intérieur des cellules de quota : si un quota rassemble ouvriers et employés, ne pas se contenter d'interroger que des ouvriers ...

Réalisation du plan de sondage
La nécessité d'enquêteurs professionnels 2/2

- Ce que l'on demande à l'enquêteur travaillant par quotas c'est en quelque sorte de se transformer en un instrument de tirage quasi aléatoire qui, par ses cheminements au hasard de la zone qu'il exploite, réussi à constituer une sélection d'interviews proche de l'équiprobabilité

- Cette « fiction » rejoint plus ou moins la réalité du terrain, selon le niveau de formation des enquêteurs et la qualité du travail de préparation effectué : si les quotas que l'on impose à l'enquêteur reflètent correctement la structure de la zone qu'il a à exploiter, le bon enquêteur réalise rapidement la série d'interviews qui lui sont confiées

La plupart des enquêtes réalisées adoptent des quota marginaux

Lorsqu'il dispose de quotas marginaux l'enquêteur travaille beaucoup plus rapidement, même si le risque de se faire piéger par des fins de quotas irréalisables le guette

Ce risque est souvent moins fort lorsque l'on doit réaliser un seul quota croisé

La plupart des enquêtes par quotas se font en deux degrés, le premier degré correspondant à un tirage de zone géographique. Si pour ces unités primaires en général on dispose des données statistiques marginales, les distributions croisées sont, elles, rarement disponibles

Les quotas doivent être :

- Pertinents et liés aux variables d'intérêt, notamment dans les enquêtes ad hoc
- Connus au niveau géographique le plus fin possible (sources statistiques disponibles)
- Aisément identifiables en termes de recherche sur le terrain et de validation en début d'interview
- Indépendants entre eux : si deux quotas sont indépendants, l'éventuelle déformation de l'un n'implique pas la déformation de l'autre
- Aussi peu nombreux que possible : le contrôle d'autres variables liées au sujet des l'enquête peut toujours être opéré par post-stratification (redressement)

Peut-on se fier aux échantillons par quotas ?

Quoique empirique, la méthode des quotas peut donner des résultats très satisfaisants

Elle présente l'avantage d'être plus rapide et moins coûteuse que l'enquête aléatoire

En raison des risques de biais dont elle est affectée, elle doit faire l'objet d'une préparation minutieuse

- → sur le plan statistique : sources utilisées, définition des critères de recherche, définition des critères de redressement
- → Au niveau du terrain : sélection et formation adéquates des enquêteurs, clarté des documents, précision des consignes de travail

Bibliographie

- Ardilly, P. (1994), Les techniques de sondage, Editions Technip, Paris
  - Chapitre II.6. Sondages empiriques
- Deroo, M., Dussaix,A.-M. (1980), Pratique et analyse des enquêtes par sondage, PUF, Paris
  - Chapitre 7. Une méthode empirique : la méthode des quota
- Dussaix,A.-M., Grosbras, J.-M., (1993), Les sondages : principes et méthodes, PUF, Paris (Que sais-je ? n°701)
  - Chapitre 5. La méthode des quotas
- Jacquart, H. (1988), Qui ? Quoi ? Comment ? ou la pratique des sondages, Eyrolles, Paris
  - Chapitre 6. L'échantillon par quotas ou échantillon proportionnel

LES PANELS

- Panel= échantillon permanent d'individus interrogés régulièrement sur leurs comportements ou leurs opinions
- Quelques exemples:
  - Panels de consommateurs
  - Panels de distributeurs
  - Panels de téléspectateurs
  - Echantillon démographique de l'INSEE (700 000)
  - Enquêtes emploi, loyers et charges (INSEE)
  - Panels de professionnels: médecins, pharmaciens, dentistes, agriculteurs.

PANELS :
CONSOMMATEURS ou DISTRIBUTEURS ?

PANEL DE CONSOMMATEURS

Permet de connaître ce qui acheté :

- quantités, prix
- promotion (?)
- acheteurs : profils

PANEL DE DISTRIBUTEURS

Permet de connaître ce qui est vendu :

- quantité, prix
- promotion
- circuits, enseignes

Un Rapide Historique des Panels

- **1929 :** le premier panel détaillant aux USA crée par Arthur Charles Nielsen
- **1959 :** le premier panel détaillant en France créé par Nielsen
- **1954 :** le premier panel de consommateurs en France crée par Stafco
- **1969 :** création de Sécodip = fusion de Stafco et Cécodis
- **1994/95 : la révolution du Scanning**

Les Sociétés de Panels Consommateurs

- Sécodip

  le panel Consoscan scannérisé de 8000 foyers
  qui a remplacé depuis le 01/95 deux panels traditionnels

  un panel de 1000 foyers avec bébés de 0 à 36 mois

  un panel de 3300 automobilistes

- Nielsen

  le panel Homescan scannérisé

Metascope SOFRES

- **Metascope**
  Le Métascope est un Access Panel **constitué d'un échantillon de 30 000 foyers**, soit 80 000 individus, représentatifs de la population des ménages français en termes de :
  région, habitat, profession,
  catégorie sociale du chef de ménage,
  âge du chef de ménage,
  nombre de personnes au foyer.
- La base de sondage est consultée mensuellement par voie postale à l'aide d'un questionnaire auto-administré. **Elle est renouvelée à hauteur de 6 000 foyers par an, par douzième mensuel.** Les panélistes sont recrutés en face-à-face à domicile, par téléphone, ou par voie postale à partir de fichiers spécifiques pour toucher des cibles larges ou très fines.

- **Automobile / Transports**
  Descriptif et suivi du Parc Automobile
  Le financement des automobiles
  Suivi des achats de pneus
  Description et suivi du Parc des deux-roues à moteur

  Banques / Assurances
  Baromètre des contrats d'assurance détenus par les foyers : Assurance Fidélité Transfert
  Suivi de l'impact des actions publicitaires des banques et des compagnies d'assurances
  Suivi du marché des ouvertures de comptes
  Suivi des comportements, besoins et attentes des PME-PMI à l'égard de la banque

Equipement de la maison

Suivi des achats de revêtements de sols, d'arts de la table et ustensiles de cuisson

Grande Consommation : alimentaire / entretien / hygiène-beauté

Etudes d'image et attitudes
Test de produits ou de concepts
Carnets de comportement

Pharmacie / Santé

Suivi des achats de lunettes correctrices, solaires et lentilles de contact
Profil, descriptif du profil, des comportements d'achats et des habitudes des utilisatrices de soin du corps

Tourisme / Loisirs

Descriptif des jardins et suivi des achats
Suivi de la demande touristique des Français
Suivi des achats de photos d'identité, de livres, de cassettes vidéo
Suivi des achats de cartes routières, de guides touristiques, atlas et plans de ville

Audience télé (Médiamètrie)

MEDIAMAT

Médiamat est l'outil de référence de la mesure d'audience de la télévision en France au niveau national. Il permet de mesurer avec précision les comportements du public en général et des principales catégories qui le composent, tous les jours, pour chaque programme diffusé par les chaînes nationales.

Répondre aux problématiques du marché

Médiamat permet de :

- connaître quotidiennement les scores de chaque programme pour différentes catégories de publics ; ces données sont livrées avec un historique de plus de dix ans ;
- analyser avec une grande précision le comportement d'individus face à la télévision : fidélité à une émission, écoute de plusieurs émissions, nombre d'expositions à une campagne publicitaire ;
- définir le public d'une émission, par exemple, selon les principaux critères socio-démographiques ;
- savoir, par exemple, si la cible touchée est large ou sélective, si l'auditoire a manifesté un intérêt fort ou faible, voire différé via l'enregistrement sur magnétoscope, s'il a regardé une émission entièrement ou partiellement.

Méthodologie

Médiamat est constitué d'un panel de 3 100 foyers, soit 8 000 individus de 4 ans et plus représentatifs des ménages résidant en France et possédant un téléviseur dans leur résidence principale.

Comment Médiamétrie mesure l'audience de la télévision ?

Médiamétrie installe dans chaque foyer faisant partie du panel Médiamat, un ou plusieurs[1] audimètres munis de télécommande à touches individuelles qui enregistrent en permanence et à la seconde près :

Toutes les utilisations du ou des téléviseur(s) du ménage :

- la marche et l'arrêt du téléviseur ;
- l'écoute des différentes chaînes ;
- l'utilisation du magnétoscope ;
- l'utilisation du téléviseur pour des jeux vidéo ou comme moniteur.

Toutes les audiences de chacun des membres du foyer et de leurs invités :

- chaque membre du foyer dispose de sa propre touche individuelle qu'il enclenche pour signaler sa présence dans la pièce où le téléviseur est allumé ;
- les invités du foyer déclarent également leur présence.

Les audimètres de chaque foyer sont reliés au réseau téléphonique ; le centre informatique de Médiamétrie les appelle toutes les nuits, recueille les données stockées et procède aux calculs des indicateurs d'audience quotidiens pour les mettre à la disposition des souscripteurs dès 9 heures du matin.

mediametrie

Objectifs

- Fournir des estimations des parametres de la population à différentes périodes
- Fournir des estimations sur une période de temps
- Mesurer des évolutions
- Mesurer des composantes d'évolution au niveau individuel
- Agréger des données au niveau individuel sur une période donnée
- Mesurer des fréquences, des durées pendant une période donnée
- Cumuler des échantillons

Panels ou échantillons indépendants?

- Un panel
  - Limite les erreurs d'observation dues aux défaillances de la mémoire
  - Donne une meilleure précision pour mesurer des évolutions

Cas de deux enquêtes successives, avec mêmes unités

- Différence de moyennes

  Estimation de m_2-m_1 (mêmes variances, grands échantillons, taux de sondage faible)

$$V(\bar{y}_2 - \bar{y}_1) = V(\bar{y}_2) + V(\bar{y}_1) - 2\,\mathrm{cov}(\bar{y}_1, \bar{y}_2)$$

$$= \frac{\sigma^2}{n} + \frac{\sigma^2}{n} - 2\rho\frac{\sigma^2}{n}$$

$$= \frac{2\sigma^2}{n}(1 - 2\rho)$$

Si échantillons indépendants :

$$V(\bar{y}_2 - \bar{y}_1) = 2\frac{\sigma^2}{n}$$

Panel efficace si $\rho > 0$

Cas de deux enquêtes successives, avec mêmes unités (2)

- Différence de moyennes

Test d'évolution

H_0 $m_2=m_1$

H_1 $m_2\neq m_1$

$$\bar{y}_2 - \bar{y}_1 \sim N\left(0 ; \sigma\sqrt{\frac{2}{n}(1-2\rho)}\right)$$

Mais ρ et σ inconnus.

Test de Student pour données appariées

$d_i = y_{i2} - y_{i1}$ différences individuelles

$\bar{d} = \bar{y}_2 - \bar{y}_1$ $\quad s_d = \sqrt{\frac{1}{n-1}\sum (d_i - \bar{d})^2}$

$$\frac{\sqrt{n}\,\bar{d}}{s_d} \sim T_{n-1} \text{ si } H_0 \text{ vraie}$$

Cas de deux enquêtes successives, avec mêmes unités (3)

- Différence de pourcentages

Test de Mc Nemar

exemple: on interroge à deux reprises, après une action, 600 clients d'une société pour connaître leur taux de satisfaction.

La proportion de satisfaits est passée de 41.7% à 46.7% **.** S'il s'agissait de deux échantillons indépendants de 600 individus, cette différence ne serait pas jugée significative.

On commettrait une grave erreur en appliquant les formules des échantillons indépendants : il faut ici connaître pour chaque individu son état aux deux enquêtes, que l'on peut résumer par le tableau de contingence 2x2 croisant les effectifs des deux variables.

Cas de deux enquêtes successives, avec mêmes unités (4)

- Différence de pourcentages

Test de Mc Nemar

T1 \ T2	Satisfaits	Non satisfaits
satisfaits	200	50
Non satisfaits	80	270

Mais pour tester la significativité de cette différence, il faut en réalité comparer les effectifs des individus ayant changé d'avis.

Cas de deux enquêtes successives, avec mêmes unités (5)

- **Test de Mc Nemar (suite)**

T1 \ T2	Satisfaits	Non satisfaits	
satisfaits	p_{11}	p_{12}	$p_{1.}$
Non satisfaits	p_{21}	P_{22}	$P_{2.}$
	$p_{.1}$	$p_{.2}$	

$H_0 : p_{1.} = p_{.1}$

Comme $p_{1.} = p_{11} + p_{12}$ et $p_{.1} = p_{11} + p_{21}$ H_0 revient à tester $p_{12} = p_{21}$

$\Rightarrow$ **test du khi-deux** : sous l'hypothèse nulle $p_{12} = p_{21}$ est estimé par $(n_{12}+n_{21})/2$

Cas de deux enquêtes successives, avec mêmes unités (6)

- **Test de Mc Nemar (fin)**

  La statistique de test est donc :

$$\frac{\left(n_{12}-\frac{n_{12}+n_{21}}{2}\right)^2+\left(n_{21}-\frac{n_{12}+n_{21}}{2}\right)^2}{\frac{n_{12}+n_{21}}{2}}$$

  Qui se simplifie en :

$$\frac{(n_{12}-n_{21})^2}{n_{12}+n_{21}}$$

  On compare cette quantité à un χ^2_1 , ou sa racine carrée à une variable normale centrée réduite.

  Ici on trouve $\frac{(n_{12}-n_{21})}{\sqrt{n_{12}+n_{21}}}=\frac{80-50}{\sqrt{80+50}}=2.63$

  $\Rightarrow$ augmentation significative de la satisfaction.

Biais et problèmes liés aux panels

- Biais de sélection
  - Recrutement
  - Non-réponses (lassitude)
- Biais de conditionnement
  - Effet de panel (apprentissage)
- Naissance d'individus (défaut de couverture)
- Solution: renouvellement du panel

Renouvellement partiel d'un panel

- Pour estimer m_2-m_1 :
  - En théorie
    - si $\rho > 0$: pas de renouvellement
    - si $\rho < 0$: renouvellement total
- Pour estimer m_2:
  - Utiliser l'information de la vague 1
  - k taux de renouvellement
  - Estimateur combinaison linéaire de l'estimation à la vague 2 de la partie nouvelle, et d'un estimateur par régression sur la partie reconduite
  - k optimal > 50%

Les panels

Les panels

- Des panels pour quoi faire ?
- Les utilisations les plus appropriées
- Questions pour un panel
- Panels et échantillons *ad hoc*
- Recrutement des panels
- Gestion des panels
- Contrôles qualité
- Vrais et faux panels
- Peut-on se fier aux panels ?
- Bibliographie

Des panels pour quoi faire ?

- L'élément distinctif des études utilisant des panels c'est de collecter les *mêmes données* de façon répétée auprès d'un *même échantillon* représentatif de la population de référence
- Le plus souvent ces données sont de nature comportementale plus qu'attitudinale
- La fréquence de la collecte ainsi que les outils employés dépendent :
  - du sujet de l'enquête
  - de l'utilisation des données
  - du niveau de précision recherché
  - du budget disponible
  - du rythme de reporting demandé

Les utilisations les plus appropriées

- Par rapport à des échantillons indépendants, des mesures répétées sur les mêmes répondants produisent des résultats plus précis pour ce qui est des *évolutions* étudiées
- Les panels rendent également possibles des analyses de *parcours individuels* (évolutions dans le temps de comportements micro)
- Ils constituent un instrument privilégié pour limiter les *erreurs d'observation* sur les questions faisant appel à la mémoire des individus
  - Par leur caractère périodique, les panels permettent de relever l'information juste au moment opportun

Questions pour un panel

- Qui, Quoi, Combien, A quel prix, Où, Quand, Quoi d'autre ? Où d'autre ?
  - Quantifier
  - sur de larges échantillons
  - au travers de mesures répétées dans le temps
  - avec une fréquence raisonnablement élevée
  - pour des utilisateurs multiples

Panels et échantillons *ad hoc*

- D'une façon générale ce qui bon pour tout échantillon est bon pour un panel
  - Définition précise de la population de référence
  - Base de sondage adéquate
  - Plan de sondage efficace
  - Mode de collecte approprié
  - Choix de redressement pertinent
- Il y a cependant des règles spécifiques à respecter
  - Une fois les éléments constitutifs établis, il vaut mieux ne plus y toucher
  - Si des biais viennent à être connus, souvent il est préférable de les garder inchangés plutôt que chercher à les corriger
  - Les règles de maintenance du panel doivent faire l'objet d'une étude rigoureuse dès le départ
- D'abord il s'agit de minimiser le biais, ensuite de le maintenir constant

Recrutement des panels (1/4)

- Après avoir décidé du plan d'échantillonnage
- sélectionné l'échantillon
- contacté les individus sélectionnés et avoir décrit les tâches à accomplir
- faut-il encore recevoir l'accord des panélistes et s'assurer de leur collaboration

- Les trois premiers points ne diffèrent guère entre panels et échantillons *ad hoc*
  - les taux de réponse sont comparables à ceux obtenus pour ces derniers
- Le dernier est spécifique au recrutement des panels
  - même s'il varient en fonction de la complexité des tâches et du temps demandé au panélistes, les taux de recrutement des panels sont souvent bien inférieurs aux taux de réponse des enquêtes *ad hoc*

Recrutement des panels (2/4)

- Lors du recrutement des panélistes, il n'est jamais souhaitable de minimiser les tâches à accomplir
  - Cela ne peut que se traduire par un taux d'abandon plus fort lors des premières expériences du panéliste (comme cela arrive aux enquêteurs qui, en phase de contact, « trichent » sur la durée du questionnaire)
  - Cela fini par coûter cher, créer des problèmes de gestion du panel et accentuer les difficultés rencontrées pour le maintien de la « représentativité » du panel
- Le dimensionnement du dispositif de recrutement doit prendre en compte les sous-populations les plus difficiles à recruter (par leur « rareté » ou par leur faible propension à participer à ce type d'étude)
  - Afin d'éviter trop de contacts inutiles dans les « cibles » les plus faciles, les phases de qualification (« screening ») et de recrutement sont souvent séparées

Recrutement des panels (3/4)

- Compte tenu des faibles taux de recrutement et de leur forte variabilité selon les catégories de population, le mode d'échantillonnage retenu le plus souvent est celui par quotas
- C'est une pratique courante que d'avoir recours à un « establishment survey », parfois réalisé lors d'études omnibus. Cela fourni un échantillon de contacts parfaitement qualifiés pour le recrutement proprement dit, qui a lieu dans un deuxième temps
- La conformité des pratiques de recrutement avec le code ESOMAR comme avec les règlementations nationales (Informatique et Liberté, …) est un point à ne pas négliger : droits d'accès, sécurité des données, périmètre d'utilisation des données des panélistes, …

Recrutement des panels (4/4)

- Selon les ressources disponibles et le type de recrutements à réaliser, le mode de contact peut être le courrier, le téléphone, le face à face, le on-line ou un mix de plusieurs modes

- Le choix du mode de contact dépend également du type de formation spécifique requise pour le panéliste

Gestion des panels (1/4)

- Le recrutement d'un panel est une affaire très coûteuse. S'assurer du niveau de collaboration le plus élevé possible de la part des panélistes constitue l'enjeu majeur de la gestion de panel
- La continuité et la cohérence de la collecte dépendent de cette relation de collaboration qu'il convient de lier avec le panéliste
- La relation avec les panélistes est fondée sur un contrat qui doit être respecté par les deux parties
  - Il faut s'abstenir de demander aux panélistes des tâches qui n'ont pas été définies lors du recrutement
  - Il faut également s'abstenir d'augmenter le temps convenu

Gestion des panels (2/4)

- Les « incentives » ne doivent pas être perçus comme la rémunération d'un travail (principe du volontariat); ils ne doivent pas non plus être perçus comme dérisoires
- Ils doivent primer la *qualité* et la *régularité* de la collaboration et doivent inciter le panéliste à prolonger la *durée* de la relation
- Les « incentives » ne doivent pas être en relation avec le thème de l'étude, ni être de nature à modifier le comportement des panélistes au cours du temps

Gestion des panels (3/4)

- La communication est également un élément important de la relation avec les panélistes
  - L'utilisation de newsletters ou de sites internet dédiés est devenue monnaie courante dans l'animation des panels
  - D'autres formes de communication, tels que des serveurs vocaux interactifs ou des contacts avec des animateurs (téléphoniques ou en face à face) sont également utilisés
- Elle doit valoriser l'intérêt des résultats obtenus grâce au panel
- Elle peut accomplir des fonctions utilitaires telles que rappeler des dates ou des moments importants de la vie du panel et/ou servir à la formation continue des panélistes, en illustrant les « meilleures pratiques »

Gestion des panels (4/4)

- Définition du « contrat »
- attribution des « incentives »
- animation
- fréquence de sollicitation
- modalités des recrutements complémentaires
- « purge » des non-répondants

constituent les moments forts de la gestion des panels

Contrôles qualité

- La qualité d'un panel est toujours jugée sur la cohérence des évolutions mesurées
- La conformité des pratiques des panélistes aux consignes données est un élément primordial de la qualité du recueil
- Des procédures de précaution et de contrôle très strictes doivent être mises en œuvre durant toute la durée de vie d'un panel
  - Souvent les données collectées auprès d'un panéliste qui vient d'être recruté ne sont pas exploités
  - Les pratiques atypiques (par rapport à la moyenne de l'échantillon ou à l'historique de l'individu) font l'objet de validation auprès des déclarants
  - La consommation de produits « de base » est également vérifiée et constitue une source de contrôle indirect des données collectées

Vrais et faux panels

- « Access panels », « mégabases »

- Des panels comme répertoires d'adresses qualifiées

- à la dérive des répertoires d'adresses qualifiées présentés comme « panels »

Peut-on se fier aux panels ? (1/2)

Les erreurs de couverture touchent les panels ni plus ni moins que les échantillons *ad hoc*

Les problèmes liés à la non-réponse – complète ou partielle – se posent souvent de façon plus aigue dans les panels

Mais c'est surtout sur le terrain des erreurs de mesure que les panels connaissent les plus grandes difficultés

Peut-on se fier aux panels ? (2/2)

Le mot « panel » peut recouvrir des réalités très variées

Depuis la notion d'échantillon permanent permettant de mieux estimer des évolutions

à celle d'échantillon prêt à l'emploi donnant accès à des sous-populations rares pour la réalisation d'enquêtes *ad hoc*

le chemin est long.

Mais la frontière avec les « mégabases » devrait rester infranchissable, un peu comme celle séparant l'échantillon par quota de l'échantillon de volontaires.

Bibliographie

- Ardilly, P. (1994), Les techniques de sondage, Editions Technip, Paris
  - Chapitre IV.3.1. Les panels
- Deroo, M., Dussaix,A.-M. (1980), Pratique et analyse des enquêtes par sondage, PUF, Paris
  - Chapitre 8. Les panels
- Pinet,B. (1980), Méthodes et pratique des panels, Technique & Vulgarisation, Paris
- Blanchard,D., Lesceux, D., (1995), Les panels. De la guerre des panels à la révolution du scanning, Dunod, Paris

Méthodes d'enquête
Les études on-line

Les études on-line

- Pourquoi maintenant ?
- Un développement majeur
- Les « fondamentaux » restent les mêmes
- Un auto-administré d'un genre nouveau
- Les défauts de couverture
- La participation et ses écueils spécifiques
- L'échantillonnage : comment fait-on ?
- Les systèmes CAWI
  - Principales caractéristiques
  - Limites actuelles
- Autour des systèmes CAWI
  - Les interactions avec les sites
  - Le suivi on-line des études
- Des structures de coût inédites
- Bibliographie

Pourquoi maintenant ?

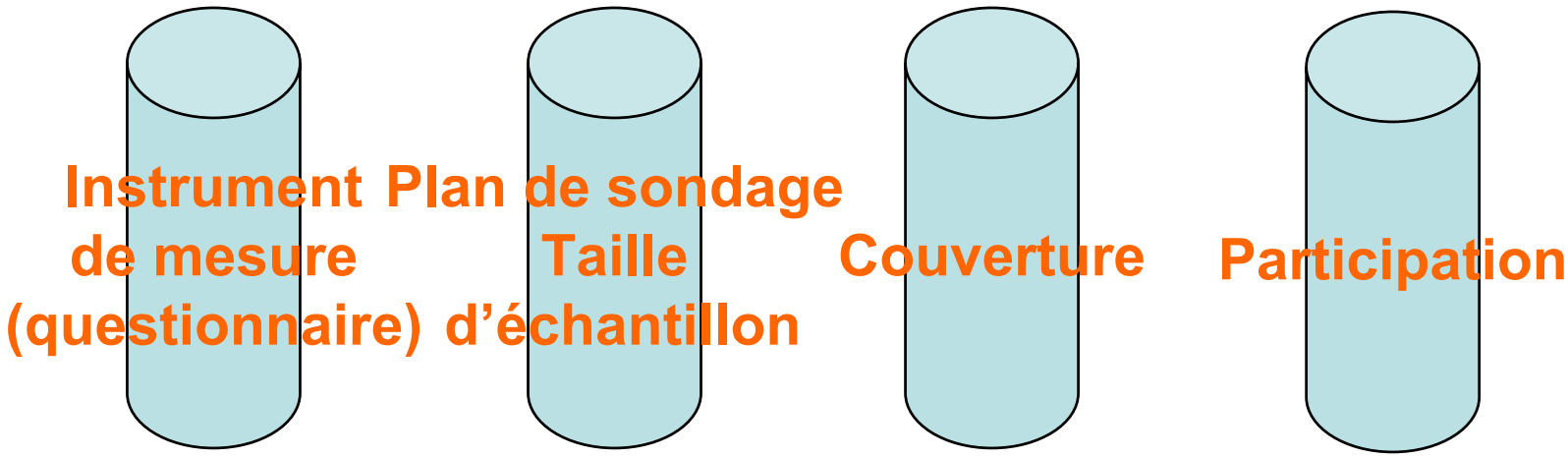
- Il y a cinq ans
  - tout le monde en parlait …
  - rares étaient ceux qui en faisaient

- Aujourd'hui
  - Les technologies commencent à arriver à maturité
  - Les taux d'équipement commencent à être convenables
    - En entreprise
    - Dans les ménages
  - Les clients des instituts en demandent

Un développement majeur

- Le on-line constitue un développement majeur dans la méthodologie d'enquête, d'une importance comparable à l'application de la théorie de l'échantillonnage dans les années 1940 ou au développement des études par téléphone dans les années 1970
  - Déconnexion des coûts du recueil de la taille d'échantillon
  - Réduction des délais du recueil
  - Élimination des frontières dans le recueil des données

Les « fondamentaux » restent les mêmes



Un auto-administré d'un genre nouveau

- Des possibilités nouvelles
  - Interaction dynamique avec la personne interrogée
  - Filtres, aiguillages, tirages aléatoires, …
  - Aide et instructions en ligne
  - Longues listes d'items, pour codification immédiate
  - Images et sons
- … et des risques nouveaux
  - Dépendance du butineur, de la définition de l'écran, du débit de la liaison à l'Internet, …
  - Dépendance de la dextérité informatique de la personne interrogée

Les défauts de couverture (1/3)

Pour les études en population générale la couverture reste largement insuffisante

- Des biais persistants
  - ++ jeunes
  - ++ hommes
  - ++ instruits
  - ++ professions supérieures
- Absence de bases de sondage
  - (listes exhaustives des populations de référence)

Les défauts de couverture (2/3)

Pour certaines populations spécifiques la couverture peut être tout à fait correcte

- Visiteurs de sites Web
- Professeurs universitaires
- Décideurs informatiques
- Salariés d'entreprises connectées à l'Internet
- …

Les défauts de couverture (3/3)

Des problèmes difficiles restent à résoudre

- Adresses multiples
  - Pas de correspondance « un à un » ménage <> adresse
- Manque de standardisation des adresses
  - Impossibilité de générer des adresses au hasard

La participation et ses écueils spécifiques

- « Ras-le-bol » des pop-up
- Junk mail (spam)
- Tarification de la liaison à l'Internet
  - Dans de nombreux cas, la personne interrogée paie pour répondre (sic!)
- Débit de la liaison à l'Internet
  - Parfois c'est vraiment trop long, et l'on peut avoir envie de laisser tomber (ou de ne pas y aller)

L'échantillonnage : comment fait-on ? (1/2)

Plusieurs méthodologies disponibles :

- E-mailing (personnalisé ou pas)
- Pop-up
  - On Entry
  - On Exit
  - On Entry / Exit
  - Avec e-mail automatique
- Bandeau …
  - …avec récupération d'informations du site

L'échantillonnage : comment fait-on ? (2/2)

Cookies et codes PIN

- Cookie sur pop-up
  - Pas de re-présentation après acceptation / refus
- Cookie sur questionnaire
  - Reprise du questionnaire en cours
  - Pas de double remplissage (sur un même poste)

Attention ! Parfois les navigateurs sont configurés pour refuser les cookies

- Code PIN
  - Identification de panélistes
  - Remplissage du questionnaire en plusieurs fois

Les systèmes CAWI
Principales caractéristiques (1/2)

- Tout type de question
  - Simple
  - Multiple
  - Quantité
  - Ouverte
- Déroulements complexes
  - Présentation d'items en aléatoire
    - « normal »
    - « prioritaire »
    - « pondéré »
  - Présentation de questions en aléatoire
  - Présentation de blocs de questions en aléatoire
  - Déroulement «horizontal» de questions

Les systèmes CAWI
Principales caractéristiques (2/2)

- Affichage de tableaux question/sous-questions sur un même écran)
- Affichage de plusieurs questions par écran
- Support questionnaires multi-langues
- Ouverture à plusieurs technologies Internet
  - Côté serveur
    - ASP
    - PHP
  - Côté client
    - Applet Java
    - Java script (ne nécessite pas de machine Java)
    - HTML
    - WAP
    - Flash

Les systèmes CAWI
Limites actuelles

- Limitation du nombre de contacts simultanés
  - par le nombre de licences (connexions autorisées) côté serveur
  - par la puissance des serveurs
- Possibilités de mise en page encore limitées
  - Tout ou presque est possible, mais par programmation

Autour des systèmes CAWI
Les interactions avec les sites

- Off-line
  - Alignement sur la charte graphique (logos, fonds, polices)
- On-line
  - Récupération d'informations depuis le site visité par l'internaute
  - Visite d'un site Web en cours de questionnaire
  - Visite d'un site Wap en cours de questionnaire (Par émulateur)
  - Listes additionnelles (Communes, profession…)
  - Intégration de logiciels de trade-off (ACA, …)

Autour des systèmes CAWI
Le suivi on-line des études

- Suivi de quotas
- Tris à plat
- Tris croisés
- Relecture (et codification) des questions ouvertes

Des structures de coût inédites

- Déconnexion (relative) coûts / nombre d'interviews
- Faible corrélation coûts / durée d'hébergement du questionnaire sur le serveur de production
- Peu d'achats extérieurs (pas d'enquêteurs), c'est surtout le temps passé par les programmeurs qui compte
  - Les éléments clefs :
    - La charte graphique
    - Le nombre de langues
    - La complexité du questionnaire

Bien sûr, cela n'est vrai qu'en dehors de l'éventuel achat d'adresses; et là beaucoup dépend de l'origine des adresses (panels, méga-bases, fichiers spécifiques, ...)

Bibliographie

- Grossnickle, J., Raskin, O. (2001). The Handbook of OnLine Marketing Research, New York, McGraw-Hill
- Dillman, D.A., (2000). Mail and Internet Surveys. The Tailored Design Method, New York, Wiley
- Des données de cadrage sur les nouvelles technologies figurent dans le « Tableau de bord de l'innovation" (décembre 2003) édité par le SESSI
  - http://www.industrie.gouv.fr/sessi/
- Le Collège Internet du **CESP** a publié, en 1997, une terminologie de la mesure d'audience d'Internet. La dernière version a été finalisée courant mai 2002
  - http://www.cesp.fr

Méthodes d'enquête
Enquêtes en face à face, par téléphone, par voie postale

Enquêtes en face à face, par téléphone, par voie postale

La relation enquêteur / enquêté

Quelle que soit l'approche épistémologique adoptée,

« objectiviste » - renvoyant à un chargé d'études neutre et détaché -

ou « constructiviste » - soulignant l'interaction inévitable du chargé d'études avec l'objet de son intérêt -,

il faut bien admettre que l'interrogation par questionnaire standardisé se situe dans le sillage de la première de ces deux approches.

Dans ce contexte, l'influence exercée par l'enquêteur sur l'enquêté doit être minimisée.

Les enquêtes en face à face
Avantages

- Possibilité de présentation d'éléments visuels
  - Échantillons de produits, maquettes de concepts
  - Descriptions illustrées de situations
  - Annonces publicitaires (« visuels », films)
- Utilisation de systèmes CAPI (Computer Assisted Personal Interviewing)
  - Filtres et aiguillages
  - Temps d'administration, global et par section
  - Dates et heures d'interview
  - Contrôle du « carnet de route »
- Possibilité d'obtenir des réponses autres que le choix entres différents items
  - Classement de cartes contenant des descriptions de produits, selon un niveau d'agrément …
- Possibilité pour l'enquêteur d'observer directement l'enquêté dans son environnement
  - Éléments relatifs au logement, au niveau de vie …

Les enquêtes en face à face
Avantages

- Taux de participation à l'enquête généralement élevés
  - Dépend du nombre de visites
  - de la durée du questionnaire
  - de l'éventuelle récompense (« incentive ») pour l'enquêté
  - et bien sûr du « métier » de l'enquêteur
- Non réponses partielles contenues
  - Possibilité de relance ou de clarification par l'enquêteur
- Faibles taux d'abandon en cours d'interviews
  - « S'il vous plaît, il ne reste qu'une minute … »

Les enquêtes en face à face
Inconvénients

- La façon de se présenter de l'enquêteur peut engendrer des refus de participer à l'enquête
  - Qualité du contact
  - Facteurs vestimentaires
  - Hostilité envers un groupe social, racisme, …
- La présence de l'enquêteur peut influencer les réponses données par l'enquêté
  - Recherche d'approbation
  - Évitement de réponses embarrassantes

Les enquêtes en face à face
Inconvénients

- Moindre dispersion de l'échantillon
  - La nécessité de limiter les déplacements des enquêteurs impose un nombre minimum de questionnaires à réaliser sur chaque point d'enquête, ce qui génère un effet de grappe
- Durées de terrain généralement assez longues
  - Cela se compte en jours ou en semaines
- L'enquêteur peut interpréter les réponses de l'enquêté
  - Perception sélective en fonction de ses propres opinions
  - Attente de réponses « probables » ou « logiques »
- Possibilité de tricherie de la part de l'enquêteur
  - Le mode de paiement de l'enquêteur, généralement au questionnaire complété, peut « pousser au crime »
  - Le contrôle de 10 à 20 % des interviews réalisées (« back-checks ») limite ces problèmes, sans pouvoir les éliminer

Les enquêtes par téléphone
Avantages

- Utilisation de systèmes CATI (Computer Assisted Telephone Interviewing)
  - Filtres et aiguillages
  - Temps d'administration, global et par section
  - Dates et heures d'interview
  - Contrôle du « carnet de route »
  - Gestion des adresses gérée par ordinateur
  - Fonctionnalités de « preview » et « predictive-dialing »
- Encadrement rapproché
  - Facilité de briefing centralisé
  - Présence de chefs d'équipe dans les salles
- Écoutes à distance
- Rapidité d'exécution, surtout pour les enquêtes par quotas
  - Cela se compte en jours, parfois même en heures

Les enquêtes par téléphone
Avantages

- Taux de participation à l'enquête généralement élevés
  - Dépend du nombre d'appels
  - de la durée du questionnaire
  - de l'éventuelle récompense (« incentive ») pour l'enquêté
  - et bien sûr du « métier » de l'enquêteur

- Certaines personnes répondent au téléphone plus qu'elles ne laissent rentrer des inconnus chez elles
  - Dans des zones où les problèmes d'insécurité sont le plus ressentis
  - Le soir, au moment où l'on cherche à interroger les actifs

- Non réponses partielles contenues
  - Possibilité de relance ou de clarification par l'enquêteur

Les enquêtes par téléphone
Inconvénients

- Pas de possibilité de supports visuels
  - Il n'y a que du son …
- Moindre attention des enquêtés (la télé allumée, les enfants qui pleurent, …)
  - Nécessité de réduire la durée des questionnaires
- L'influence de l'enquêteur est réduite par rapport au face à face, mais toujours présente
  - Recherche d'approbation
  - Évitement de réponses embarrassantes
- Comme en face à face, l'enquêteur peut interpréter les réponses de l'enquêté
  - Perception sélective en fonction de ses propres opinions
  - Attente de réponses « probables » ou « logiques »

Les enquêtes par téléphone
Inconvénients

- Qualité décroissante des bases de sondage
  - Listes « rouges »
  - Convergence « fixe » / « mobile »

  Le recours au « Random Digit Dialing » constitue une assez bonne réponse à ces problèmes

- « Ras-le-bol » des appels non sollicités
  - Il est relativement facile de filtrer les appels
  - ou de raccrocher à l'enquêteur

- Taux d'abandon en cours d'interviews plus élevé
  - Certes l'enquêteur peut limiter les abandons « S'il vous plaît, il ne reste qu'une minute … », mais au téléphone cela lui est plus difficile qu'en face à face

Les enquêtes par voie postale
Avantages

- Questionnaire auto-administré
  - La personne interrogée peut répondre à son rythme
  - en choisissant le moment qui lui convient le mieux
- Bon contrôle de l'échantillon (au niveau ménage)
- Possibilité d'administrer des questionnaires comportant de nombreuses questions
- Pas de présence d'enquêteur, donc aucune influence de celui-ci sur l'enquêté
- Possibilité de présenter des éléments visuels (dessins, photos, VHS/DVD …)
- Moins de tricheries possibles

Les enquêtes par voie postale
Inconvénients

- Aucune aide personnalisée au remplissage du questionnaire ni d'encouragement à en arriver au bout
  - Pour ce qui est de l'assistance, la mise en place d'un numéro vert est fortement recommandée
  - Pour l'encouragement, l'usage d'« incentives » est généralement conseillé
- Limitation dans la complexité du questionnaire, notamment en ce qui concerne les filtres et les aiguillages
- Non-réponse partielle plus fréquente, due à l'absence de relance de la part de l'enquêteur

Les enquêtes par voie postale
Inconvénients

- Obsolescence des listes d'adresses
- Risque de non ouverture du courrier ou de confusion avec du courrier publicitaire (souvent jeté d'emblée)
- Mauvais contrôle de l'échantillon (au niveau individu)
- Lenteur des retours
- Structure des répondants souvent trop « haut de gamme »
  - Plus instruits
  - Disposant de revenus plus élevés …

Comparaison des trois méthodes

	Face à face	Téléphone	Voie postale
Taux de participation	♥ ♥ ♥	♥ ♥	♥
Nombre de questions	♥ ♥ ♥	♥	♥ ♥
Complexité des questions	♥ ♥ ♥	♥ ♥ ♥	♥
Interaction enquêteur / enquêté	♥	♥ ♥	♥ ♥ ♥
Coût de l'interview	♥	♥ ♥	♥ ♥ ♥
Rapidité	♥ ♥	♥ ♥ ♥	♥

Autres modes et modes combinés

Autres modes

- Enquêtes en salle
  - en rue ou en sortie | entrée de magasin, bureau de vote, …
  - par fax, mail, internet, …

et modes combinés

- Dépôt / Rappel (« Drop-off » / « callback »)
- Phone / Mail / Phone
- Face à face, puis internet
- Téléphone, puis internet
- Voie postale, puis internet …

La liste est longue et amenée à changer tous les jours.

Question d'arbitrages

Le choix d'une méthode d'enquête n'est pas toujours une évidence.

Dans l'absolu, cela n'a pas de sens d'affirmer la supériorité d'une méthode sur une autre

Le sujet de l'étude, le budget, les délais, constituent un système de contraintes parfois difficile à appréhender.

Il est toujours question d'arbitrages.

« L'art du sondeur » consiste à trouver à chaque fois la meilleure solution à adopter ;

ou parfois simplement la moins mauvaise.

Bibliographie

- Lebart, L., éd. (1992), La qualité de l'information dans les enquêtes, Paris, Dunod
- Corbetta, P. (2003). Social Research, Theory, Methods and Technics, London, Sage
- Birn, R., éd. (2000), The International Handbook of Market Research Techniques, Second Edition, London, Kogan Page