STA108 – Introduction

Cours n°1 du 28/09/2020

Sylvie Rousseau

le cnam

Organisation pratique

Planning : http://emploidutemps.cnam.fr/

- Code : CESTA108HT11
- Période : du 21/09/2020 au 22/01/2021
- Cours et ED / TP : les lundis et vendredis, de 18h30 à 20h30, en salle 21.1.12
- A consulter régulièrement pour les précisions de salle et les changements

Site du cours : http://maths.cnam.fr/spip.php?article54

- Programme prévisionnel
- Supports de cours, ED, TP, projet
- Bibliographie

Evaluation

- Un examen
- Un projet

Plan prévisionnel du cours

Cours	Date		Sujet	ED	Date		Sujet
Cours 1	L	21/09/20	Introduction				
Cours 2	L	28/09/20	Sondage aléatoire simple	ED 1	V	02/10/20	Sondage aléatoire simple
				ED 2	L	05/10/20	Sondage aléatoire simple
Cours 3	V	09/10/20	Sondages à probabilités inégales	ED 3	L	12/10/20	Plans à probabilités inégales
Cours 4	V	16/10/20	Algorithmes de tirage	ED 4	L	19/10/20	Plans à probabilités inégales
Cours 5	V	23/10/20	Plans stratifiés	ED 5	L	26/10/20	Plans stratifiés
Cours 6	V	30/10/20	Questionnaires, enquêteurs et enquêtés	ED 6	L	02/11/20	Plans stratifiés
Cours 7	V	06/11/20	Plans par grappes et plans à deux degrés	ED 7	L	09/11/20	TP- Algorithmes de tirage
Cours 8	V	13/11/20	Redressement (quotient, régression, post-strates)	ED 8	L	16/11/20	Plans par grappes
Cours 9	V	20/11/20	Modes de recueil (avec et sans enquêteur) + point projet	ED 9	L	23/11/20	Plans à plusieurs degrés
Cours 10	V	27/11/20	Cadrage d'une enquête : méthode des quotas , théoriques et redressement	ED 10	L	30/11/20	Plans à plusieurs degrés
Cours 11	V	04/12/20	Données manquantes et non-réponse	ED 11	L	07/12/20	Redressement
Cours 12	V	11/12/20	Les panels + Point projet	ED 12	L	14/12/20	Redressement
				ED 13	V	18/12/20	TP - Redressement
Cours 13	V	08/01/21	Enrichissement des données d'enquête - Process de réalisation d'une étude	ED 14	L	11/01/21	TP – Correction de la non-réponse
Cours 14	V	15/01/21	Paradonnées, outliers, qualité d'une enquête + Point projet	ED 15	L	18/01/21	Compléments et révisions
Cours 15	V	22/01/21	Recensement de la population				

Introduction aux sondages

1)Introduction

2)Représentativité

3)Erreur totale et erreur d'échantillonnage

4)Notion d'information auxiliaire

5)Sondages aléatoires et sondages empiriques

6)Formalisme : notations et concepts

Les sondages

Abondance de sondages dans notre quotidien

- En particulier de sondages d'opinions
- Au centre de polémiques
  - http://www.apmep.asso.fr/IMG/pdf/bull-474-_fine_piedenoir.pdf
  - http://www.senat.fr/notice-rapport/2010/r10-054-notice.html

Mais une discipline beaucoup plus vaste, méconnue, assez récente aussi

Le secteur

Deux acteurs principaux :

- La statistique publique : Insee et services statistiques ministériels

- Les instituts privés : IPSOS, TNS, IFOP, BVA, CSA, Médiamétrie, ...

La statistique publique

- Près de 7 000 personnes dont 5 300 à l'Insee
- Exemples d'opérations :
  - Enquêtes annuelles de recensement, recensement agricole
  - Enquêtes emploi, logement, budget, patrimoine, compétences des élèves, victimation, insertion professionnelle, ...
  - Estimation des volumes de certaines productions, chiffres d'affaires, effectifs employés
  - Enquêtes sur le comportement d'innovation, les consommations d'énergie, ...
  - Calculs d'indices de prix à la consommation, à la production, ...
- Une organisation ternaire, garante de plusieurs principes forts comme l'indépendance, l'impartialité, le secret, la qualité, la pertinence :
  - rôles du CNIS et de l'ASP

Les instituts de sondage

Élections américaines de 1936 : Franklin D.Roosevelt (démocrate) vs Alf Landon (républicain)

- Landon donné gagnant par 2 millions d'abonnés au Literary Digest
- Pronostic favorable à Roosevelt selon un échantillon de 50 000 personnes interrogées par Gallup

⇨Naissance des instituts de sondage

  - Généralisation aux études de marché

- *Près de 400 instituts d'étude de marché et d'opinion en France*
- *Environ 12 000 personnes, hors enquêteurs*

Un peu d'histoire

- 1895 – Kiaer : débat sur les dénombrements représentatifs
- 1925 – Jensen : une reconnaissance des sondages
- 1934 – Neyman : la théorie
- *1936 – Election de Roosevelt*
- *1938 – Fondation de l'IFOP par J.Stoetzel*
- 1952 – Horvitz et Thompson : sondages à probabilités inégales
- *1965 – Ballottage De Gaulle annoncé par l'IFOP*

Remarque : en démocratie, les sondages politiques sont validés par les résultats des élections. Et ça ne marche pas toujours … (ex : prédiction à tort de la défaite d'Harry Truman en 1948)

Recensement et sondage

Recensement : Observation exhaustive de tous les éléments d'une population

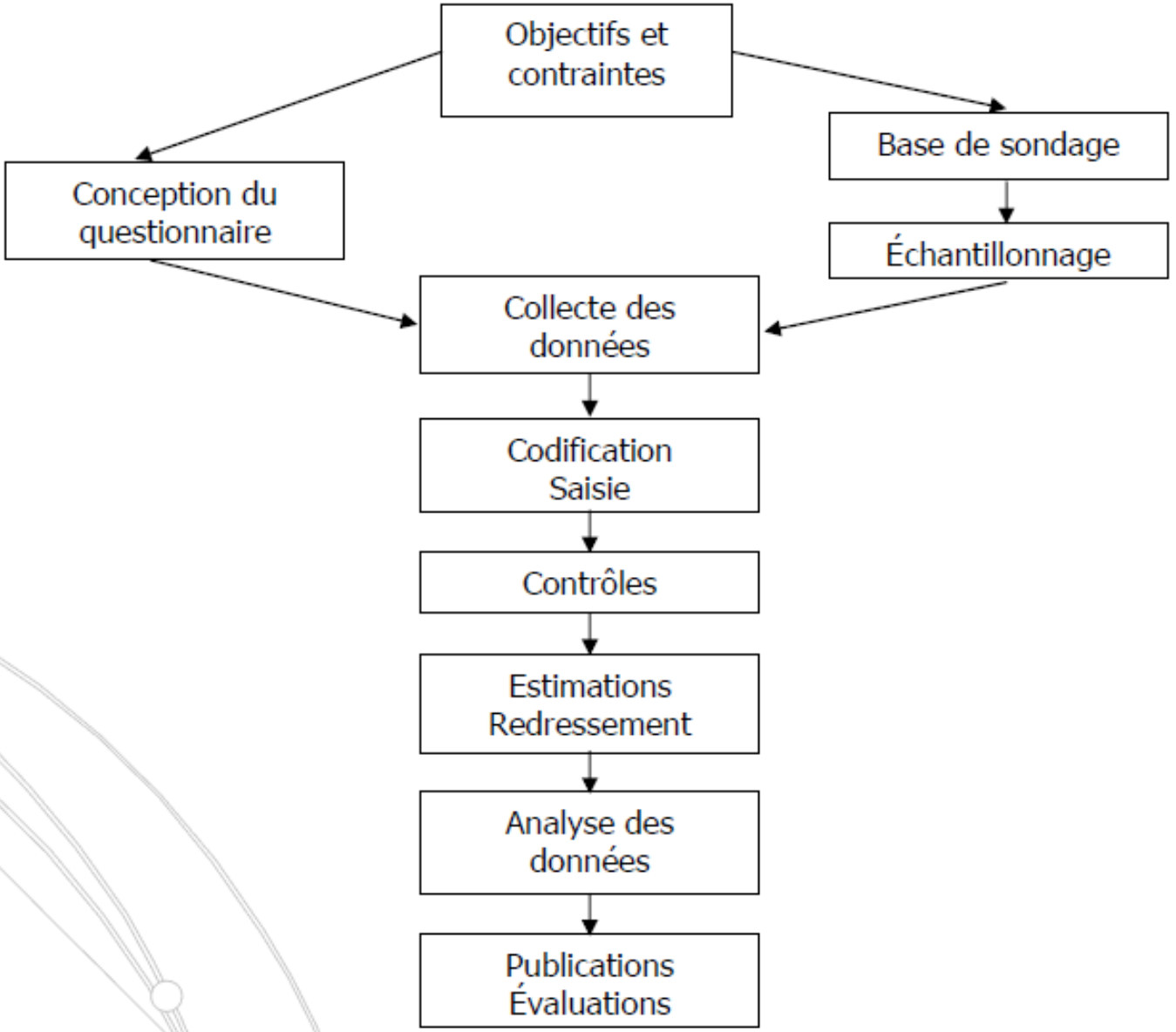
Sondage (Y. Tillé, 2001) : Méthode qui consiste à *prélever une partie (un échantillon) d'un ensemble* et à l'analyser afin d'extrapoler les résultats de l'échantillon à un ensemble de référence

- *Sondage probabiliste : l'échantillon est sélectionné de manière aléatoire*
- *Sondage empirique : méthode des quotas, volontariat, etc.*

Particularités des sondages

- Deux enjeux :
  - Sélection de l'échantillon dans une population finie
  - Agrégation des réponses
    - Estimation
    - Précision
- D'autres questions :
  - Formulation du questionnaire
  - Choix du mode de collecte
  - ...

Dispositif d'une enquête



Introduction aux sondages

1)Introduction

2)Représentativité

3)Erreur totale et erreur d'échantillonnage

4)Notion d'information auxiliaire

5)Sondages aléatoires et sondages empiriques

6)Formalisme : notations et concepts

Représentativité ?

- Notion peu scientifique
- Souvent confondue avec l'équiprobabilité ou le respect de certaines proportions
  - Idée de l'échantillon comme modèle réduit de la population
- Un sondage à probabilités inégales peut être représentatif en un autre sens

Représentativité ?

Yves Tillé, Théorie des sondages (Dunod, 2001) :

«L'objectif d'un sondage est de fournir un certain nombre d'informations sur une population en n'examinant qu'une partie de celle-ci, appelée échantillon. On dit souvent qu'un échantillon est représentatif d'une population s'il en constitue le modèle réduit. La représentativité est ainsi évoquée en tant qu'argument de validité : un bon échantillon devrait ressembler autant que possible à la population à étudier de sorte que certaines catégories apparaissent en mêmes proportions dans l'échantillon et la population. Pourtant cette théorie, couramment véhiculée par les médias et même par certains ouvrages de méthodologie est erronée. Il est en effet souvent souhaitable de d'effectuer des tirages à probabilités inégales ou de sur représenter certaines fractions de la population. Pour estimer avec précision un paramètre, il faut aller chercher l'information de manière judicieuse plutôt que d'accorder la même importance à chaque unité»

Introduction aux sondages

1)Introduction

2)Représentativité

3)Erreur d'échantillonnage et erreur totale

4)Notion d'information auxiliaire

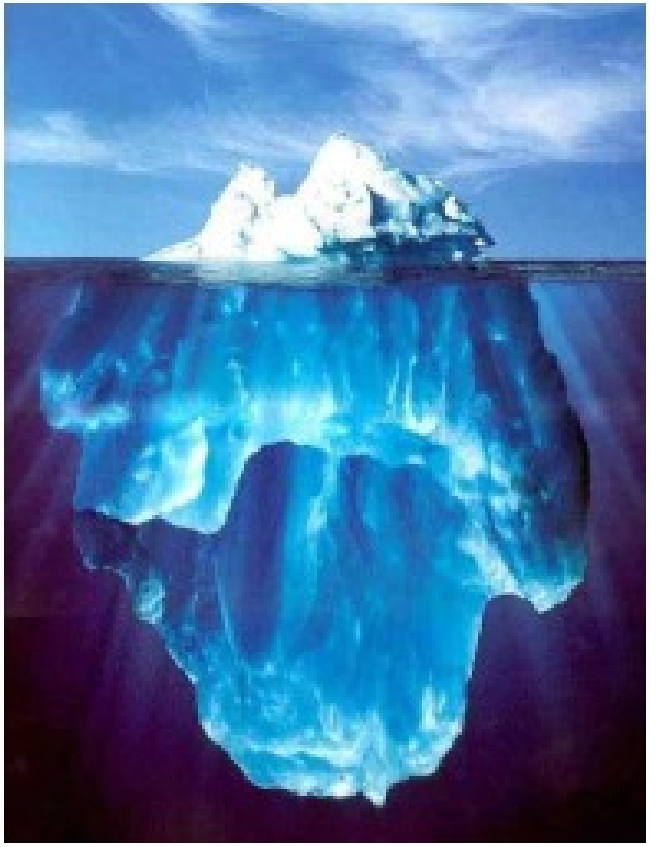
5)Sondages aléatoires et sondages empiriques

6)Formalisme : notations et concepts

Il n'y a pas que l'erreur d'échantillonnage

Bien souvent, on se focalise sur un type d'erreur, ***l'erreur d'échantillonnage*** (due au fait que l'on travaille sur *un échantillon*) d'autant qu'elle peut, dans certains cas, être formalisée et correctement évaluée

Mais quand bien ***même on interrogerait toute la population, il resterait quand même des sources d'erreurs*** : valeurs manquantes, compréhension des questions, interactions enquêteur/enquêté, etc.



L’erreur totale dans les enquêtes par sondage

L’erreur totale est la somme de deux types d’erreurs :

- Erreur d’échantillonnage : due à l'aléa de tirage
- Autres sources d’erreurs : non directement liées à l’échantillonnage

Très peu d’ouvrages de méthodologie intègrent la notion d’erreur totale

- *Robert M. Groves « Survey Errorsand Survey Costs », Wiley Interscience*

Erreur d'échantillonnage
« Sampling error »

- Se produit lorsqu'on estime une caractéristique de la population à partir d'un échantillon
  - Différence entre l'estimation calculée à partir de l'échantillon et la « vraie » valeur qui aurait été obtenue avec un recensement auprès de la population entière
  - Dans un recensement, il n'y a pas d'erreur d'échantillonnage
- Caractéristiques
  - L'erreur d'échantillonnage diminue quand la taille de l'échantillon augmente
  - Elle dépend de la taille de la population étudiée
  - Elle dépend de la variabilité de la caractéristique de la population qu'on étudie
  - On peut en réduire la portée grâce à un plan de sondage approprié
  - Elle peut être mesurée si l'échantillon est choisi selon un plan de sondage probabiliste

Erreurs non liées à l'échantillonnage
« Non-sampling errors »

Ce type d'erreur s'observe aussi dans un recensement

- **Erreur d'observation** : Se rapporte aux différentes phases de la collecte
  - *Erreur de sur couverture* : on sélectionne des individus qui n'ont rien à faire dans l'échantillon
  - *Erreur de mesure* : différences entre ce qui est mesuré et les valeurs réelles. Par exemple : problème de compréhension de la question, malaise à répondre sur des sujets sensibles, influence de l'enquêteur, etc.
  - *Erreur de production* : saisie, codage, traitement, transcription des résultats
- **Erreur de non observation** : on n'observe pas les valeurs pour certains individus
  - *Sous couverture* : l'échantillon omet certains éléments de la population cible. Ces individus, appartenant à un ou plusieurs groupes de la population, ont une probabilité nulle d'être sélectionnés et ne peuvent « parler » pour leur groupe d'appartenance
  - *Erreur de non réponse* : on distingue la non réponse totale (les individus n'ont pas voulu participer à l'enquête) et la non réponse partielle (seules certaines questions sont restées sans réponses). La non-réponse est souvent un gros problème car les caractéristiques des non-répondants diffèrent généralement de celles des répondants.

Erreurs non liées à l'échantillonnage
« Non-sampling errors »

- Elles peuvent survenir à toutes les étapes du processus d'enquête, hors échantillonnage
- Elles peuvent être liées les unes aux autres
  - Faire du «forcing» pour réduire la non-réponse peut amener à amplifier les erreurs de mesure
- Elles s'observent aussi dans un recensement
  - Et peuvent y être plus importantes que dans une enquête par sondage (avec des procédures de contrôle ou de formation pouvant être plus approfondies, en lien notamment avec la taille réduite du nombre d'unités interrogées)
- Elles sont difficiles à mesurer
  - En général, les efforts 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 très dommageable, car ces erreurs -qui ont essentiellement la nature de biais –ne diminuent pas lorsque la taille d'échantillon augmente

Biais et Variance

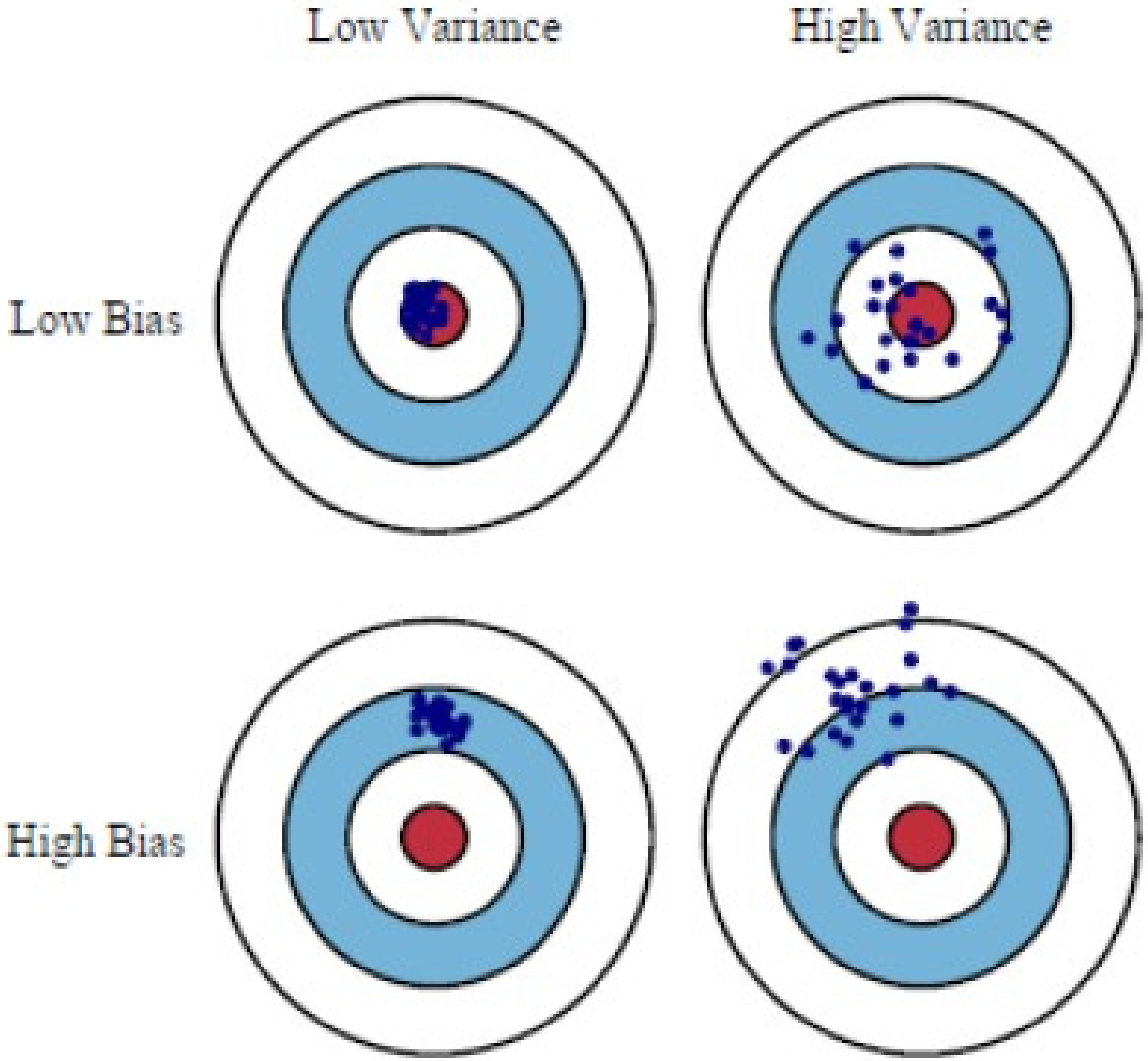
Le biais est la différence entre la valeur réelle du paramètre dans l'ensemble de la population et la valeur estimée à partir de tous les échantillons possibles selon un plan de sondage donné.

- le biais est souvent peu sensible à la taille de l'échantillon

La variance mesure la variabilité des estimations d'un échantillon à l'autre (on observe une valeur différente si l'on répète plusieurs fois la même enquête)

- la variance diminue quand la taille de l'échantillon augmente

Biais et Variance



le cnam

Introduction aux sondages

1)Introduction

2)Représentativité

3)Erreur totale et erreur d'échantillonnage

4)Notion d'information auxiliaire

5)Sondages aléatoires et sondages empiriques

6)Formalisme : notations et concepts

Utilisez au mieux les sources d'information auxiliaire

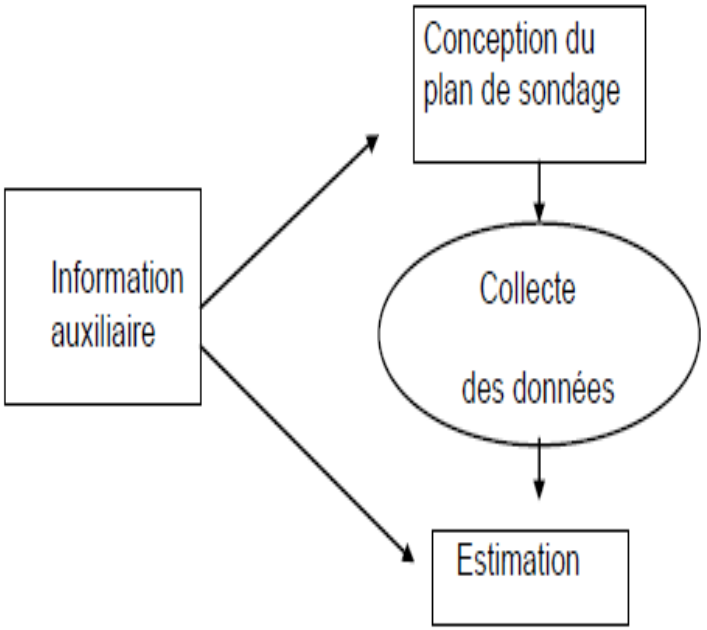
Un principe fondamental : *lorsqu'on dispose d'information auxiliaire, il faut chercher à l'utiliser pour obtenir des estimations plus précises*

Cette information peut être utilisée à deux moments :

1. ***Au moment du tirage***, on utilise des techniques de stratification, de tirage proportionnel à un critère de taille, ou de tirage équilibré

2. ***Au moment de l'estimation***, on utilise des techniques dites de redressement, de calage sur marges notamment

FIG. 1 - *Les deux étapes de l'utilisation de l'information auxiliaire*



Le sondage aléatoire simple

- Un usage par défaut, en l'absence de toute information auxiliaire
- Modèle de référence, le plus simple à formaliser :
  - tous les individus ont la même probabilité d'appartenir à l'échantillon
- Référence : il sert d'étalon pour les autres plans de sondage
- Brique de base pour des plans de sondage usuels
  - par exemple, les sondages stratifiés et les sondages à deux degrés sont des assemblages de sondages simples.

Introduction aux sondages

1)Introduction

2)Représentativité

3)Erreur totale et erreur d'échantillonnage

4)Notion d'information auxiliaire

5)Sondages aléatoires et sondages empiriques

6)Formalisme : notations et concepts

Méthode probabiliste (aléatoire) et méthode empirique

- **Un sondage est probabiliste ou aléatoire** si chaque individu de la population a une probabilité donnée, connue d'avance et non nulle, d'appartenir à l'échantillon. Cette probabilité est appelée probabilité d'inclusion ou probabilité de sélection

  La validation des méthodes aléatoires est basée sur le calcul des probabilités, qui permettent de construire des intervalles de confiance

- **Les sondages empiriques** sont ceux qui ne permettent pas de calculer la probabilité d'inclusion des individus. Il s'agit principalement des méthodes de quotas ou encore de la méthode d'unités-types, le volontariat, etc.

  La validation des méthodes empiriques s'obtient par expérimentation en comparant les résultats avec des recensements ou des résultats sur la population

Méthode probabiliste (aléatoire) et méthode empirique

- Une démarche scientifique plaide généralement pour l'échantillonnage probabiliste
- Les sondages empiriques sont principalement utilisés par les instituts privés pour plusieurs raisons
  - Réglementaires :
    - ils ne disposent pas de bases de sondage
    - ils n'ont pas de moyens coercitifs
  - Économiques
    - les enquêtes aléatoires sont beaucoup plus rapides et moins chères

⇨ Un équilibre à trouver entre la rapidité et le maintien de la qualité de collecte : pour effectuer un bon sondage avec quotas, il faut se rapprocher le plus possible d'un tirage aléatoire des individus

Plans de sondage probabilistes

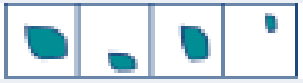
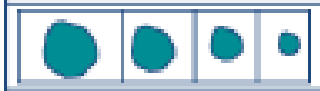
- Sondage aléatoire simple sans remise
- Plans à probabilités inégales
- Plans stratifiés
- Plans par grappes
- Plans à plusieurs degrés
- ...

Méthodes à choix raisonné

- Méthode des quotas
- Méthode des unités-types (méthode de Politz)
- Méthode des itinéraires
- Volontariat
- Échantillonnage sur place
- Échantillon « boule de neige »
- ...

Comparatif des plans de sondage classiques

- Pascal Ardilly, les techniques de sondage, Dunod, 2006

	Plan de sondage	Réalisation du tirage et estimation	précision	coût terrain
	Sondage aléatoire simple	=	=	=
	Sondage stratifié allocation quelconque	-	+	=
	Sondage stratifié alloc proportionnelle	-	+	=
	Sondage stratifié allocation optimale	--	+++	=
	Sondage 'quelconque' à plusieurs degrés	--	-	+
	Sondage en grappes	-	--	++
	Sondages à probabilité inégales	-	Si Yi proportionnel à Xi ++, sinon --	=
	Sondage équilibré	---	+++	=
	Sondage par quota	-	?	++

Introduction aux sondages

1)Introduction

2)Représentativité

3)Erreur totale et erreur d'échantillonnage

4)Notion d'information auxiliaire

5)Sondages aléatoires et sondages empiriques

6)Formalisme : notations et concepts

Population, base de sondage, variable d'intérêt

- **Population U** composée de N individus ou éléments appelés unités statistiques
  - N est la taille de la population U, supposée finie
  - Exemples de population : ensemble des touristes d'un pays, ensemble des ménages d'un pays, production de pièces mécaniques d'une usine…

- **Base de sondage** : liste exhaustive des éléments de la population U où chaque élément est représenté par son identifiant

  U = 1, …, k, …, N

- **Variable d'intérêt Y** dont les valeurs associées à chaque unité sont notées

  $y_1, y_2, \ldots y_N$

Paramètres d'intérêt

- But du sondage : estimer un total, une moyenne, une proportion = un paramètre d'intérêt θ sur la population entière :

$$\theta = f(y_k, k \in U)$$

- Cette fonction est appelée fonction d'intérêt. Elle est souvent linéaire, par exemple
  - le total : $t_y = \Sigma_{k \in U}\, y_k$
  - ou la moyenne : $\bar{Y} = \frac{1}{N}\sum_{k \in U} y_k$

    Remarque : estimer un total ou une moyenne n'est pas forcément équivalent car la taille de la population peut être méconnue

- D'autres fonctions plus complexes, par exemple :

  la variance : $\sigma_y^2 = \frac{1}{N}\sum_{k \in U}(y_k - \bar{Y})^2$ ou un ratio : $R = t_x / t_y$

Échantillon

- **Préférer les tirages sans remise**
- Un échantillon est un sous-ensemble non vide de U traditionnellement noté *s* (*« sample »)*
- Notation ensembliste : un échantillon non ordonné sans remise est représenté par un n-uplet listant les individus retenus d'après leur identifiant (combinaison de n unités de U prises sans répétition)
- S = Ensemble des échantillons possibles de U = ensemble des parties non vides de U
  - Par exemple, pour une population U={1,2,3} :
    - S = { {1}, {2}, {3}, {1,2}, {1,3}, {2,3}, {1,2,3} }
    - le nombre d'échantillons de taille quelconque, non ordonnés et sans remise est 2^N-1
- Si l'échantillon est de taille fixe, alors on notera n sa taille
- **Taux de sondage : f =n/N** avec
  - n : taille de l'échantillon
  - N : taille de la population

Plan de sondage

Un **plan de sondage** non ordonné sans remise p(s) est une loi de probabilités sur S :

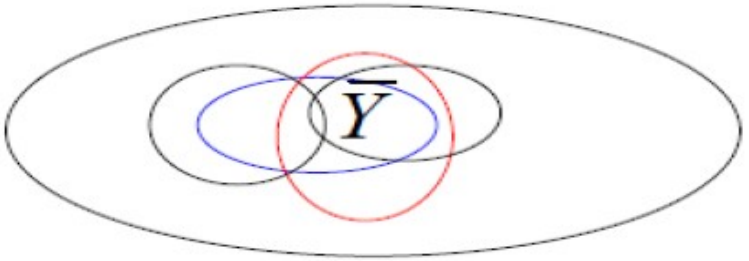
$$\Sigma_{s \in S}\, p(s) = 1 \text{ avec } p(s) \text{ compris entre } 0 \text{ et } 1 \text{ pour tout } s$$

Exemple : plan simple sans remise de n=2 unités dans une population U = {1,2,3,4} de taille N=4

- Il y a C_4^2 = 6 échantillons possibles, tous équiprobables
- S= { {1,2}, {1,3}, {1,4}, {2,3}, {2,4}, {3,4} }
- p(s) =1/6 pour tout s
- π_k = 1/2 = n/N

Fluctuation ou erreur d'échantillonnage

Fluctuations d'échantillonnage : avec les mêmes probabilités d'inclusion, répéter x fois un sondage donnera potentiellement x résultats différents



$\overline{y}_1, \overline{y}_2, \ldots, \overline{y}_q$

Attention : ***l'aléa se situe exclusivement au niveau du choix des individus dans l'échantillon***. Les valeurs de Y sont considérées connues

Approche différente de celle adoptée en économétrie par exemple où les valeurs de Y sont des variables aléatoires dont on observe une réalisation

En sondage, l'estimateur est aléatoire, non pas par la nature des variables mesurées, mais par la composition de l'échantillon retenu

Espérance et variance

Soit θ le paramètre d'intérêt qu'on estime à partir des valeurs prises sur l'échantillon :

$$\theta = f(y_k, k \in U) \qquad \hat{\theta}(s) = f(y_k, k \in s)$$

L'espérance de l'estimateur est définie par :

$$E(\hat{\theta}) = \sum_{s \in S} \hat{\theta}(s) \times p(S = s)$$

Le **biais** désigne la quantité : $B(\hat{\theta}) = E(\hat{\theta}) - \theta$

Un estimateur est dit sans biais si : $E(\hat{\theta}) = \theta$

Sa **variance** se calcule par :

$$Var(\hat{\theta}) = \sum_{s \in S} \left[\hat{\theta} - E(\hat{\theta})\right]^2 \times p(S = s) = E\left[\hat{\theta} - E(\hat{\theta})\right]^2 = E\left[\hat{\theta}\right]^2 - E^2(\hat{\theta})$$

L'erreur quadratique moyenne vaut : $EQM(\hat{\theta}) = E\left[\hat{\theta} - \theta\right]^2 = Var(\hat{\theta}) + B^2(\hat{\theta})$

Application à l'estimation de la moyenne

La moyenne de Y sur la population U :

$$\bar{y} = \frac{1}{N} \sum_{k \in U} y_k$$

est estimée sur un échantillon s par :

$$\hat{\bar{y}} = \frac{1}{n} \sum_{k \in S} y_k$$

Un exemple de sondage aléatoire simple

Principe

Sélectionner dans une population de taille N un échantillon de taille fixée n sans remise, tel que tous les individus aient la même probabilité d'être retenus

Exemple :

Soit un sondage aléatoire simple de 2 unités dans une population de 5 entreprises.

Le paramètre d'intérêt est la moyenne des montants des ventes.

On sait que : $y_1 = 5$, $y_2 = 8$, $y_3 = 10$, $y_4 = 12$ et $y_5 = 15$

La vraie moyenne vaut : (5 + 8 + 10 + 12 + 15) / 5 = 10

Sur chaque échantillon {k,l}, la moyenne estimée est calculée par : $(y_k + y_l) / 2$

Soit sur les 10 échantillons possibles :

y1	5	5	5	5	8	8	8	10	10	12
y2	8	10	12	15	10	12	15	12	15	15
Y	*6.5*	*7.5*	*8.5*	*10*	*9*	*10*	*11.5*	*11*	*12.5*	*13.5*

Un exemple de sondage aléatoire simple

Biais

On dit que l'estimateur de la moyenne est sans biais quand l'espérance de cet estimateur est égal à la vraie moyenne de la population

Sur l'exemple, on vérifie que

$$(6,5 + 7,5 + 8,5 + 10 + 9 + 10 + 11,5 + 11 + 12,5 + 13,5) / 10 = 10$$

Autrement dit, pour un sondage aléatoire simple, la moyenne empirique est un estimateur sans biais de la moyenne

Attention : « sans biais » signifie que les résultats sont bons « en moyenne » mais pas que le résultat obtenu à partir d'un échantillon le soit ...

Bibliographie

J.ANTOINE, Histoire des sondages (Odile Jacob, 2005)

P.ARDILLY, Les techniques de sondage, 2ème édition (Technip, 2006)

Y.TILLÉ, Théorie des sondages (Dunod, 2001)

P.ARDILLY, Y.TILLÉ, Exercices corrigés de méthodes de sondage (Ellipses, 2003)

A.M. DUSSAIX, J.M. GROSBRAS, Exercices de sondages (Economica, 1992)

SYNTEC, Etudes Marketing et Opinion (Dunod, 2007)