Analyse statistique - Séance 5 : Enquêtes par questionnaire et représentativité statistique

Aubin Poissonnier-Beraud

Villes et environnements urbains - Université Lumière Lyon 2

2025

Population et échantillon

Thomé Cécile, « “On fait juste attention.” La mesure du retrait comme méthode contraceptive dans les enquêtes en France depuis les années 1970 », Population, 2023, vol. 78, n1, p. 29-50.

- Pourquoi les premières enquêtes des années 1960 n'étaient pas de « véritables enquêtes » ?
- Quel principal obstacle ont rencontré les enquêtes des années 1970 pour la mesure du retrait ?
- Quelles stratégies d'enquête ont permis ou permettraient d'améliorer la mesure de cette méthode contraceptive ?

Population et mesure du retrait

Tableau A.1. Mesure du retrait dans neuf enquêtes françaises entre 1970 et 2016

Année	Nom de l'enquête	Population interrogée	Mesure du retrait (en gras, la période concernée)	Désignation de la méthode (le cas échéant)	Méthode considérée comme principale dans le traitement des données	Taux de recours au retrait comme méthode principale (chez les personnes concernées)	
						Hommes	Femmes
1970	Enquête Simon	1 375 femmes et 1 250 hommes de 20 ans et plus	Il ne s'agit pas de la méthode utilisée actuellement, mais de **méthodes ayant déjà été utilisées au cours de la vie**. Un tableau des méthodes était proposé.	« Interruption du rapport avant l'éjaculation »	-	50 % (au cours de la vie)	46 % (au cours de la vie)
1971	Enquête sur la régulation des naissances (Ined / Insee)	2 890 femmes non célibataires ayant moins de 47 ans	Une première question ouverte est posée concernant « ce qu'on peut faire pour éviter une naissance », puis des questions sur les informations médicales reçues sur le sujet. Puis : « **Depuis que vous êtes mariée**, avez-vous fait quelque chose, vous-même ou votre mari, pour éviter une naissance ? (Si oui) Qu'avez-vous fait ? (donnez-moi le numéro de la liste) ». Le retrait est situé en 3e position.	« Retrait, l'homme se retire à temps »	Déclaratif	-	30 % (depuis le mariage)
1978	Enquête mondiale de fécondité (Ined/Insee)	3 011 femmes de 20 à 44 ans, mariées ou non	Un paragraphe introductif visant à normaliser l'utilisation d'une contraception est lu aux enquêtées. Les méthodes sont citées et explicitées (le retrait cité en 2e position) pour en tester la connaissance, puis on demande à l'enquêtée : « Est-ce que votre mari (partenaire) et vous-même employez **actuellement** une méthode pour éviter d'avoir un enfant ? », et si oui laquelle ou lesquelles. En cas de réponse négative, une question de rattrapage est prévue : « Ni vous, ni votre mari, ne prenez aucune précaution ? »	« L'homme peut aussi pratiquer le retrait, c'est-à-dire se retirer à temps (avant l'éjaculation) »	Celle utilisée en milieu de cycle. Si plusieurs méthodes sont déclarées, hiérarchisation (retrait 6e sur 8).	-	18 % (actuellement)
1988	Enquête fécondité (Ined/Inserm)	3 188 femmes de 18 à 49 ans	Les méthodes sont citées et explicitées (le retrait est cité en 2e position) pour en tester la connaissance et l'utilisation au cours de la vie, puis on demande à l'enquêtée : « Employez-vous **actuellement** une méthode, vous ou votre conjoint, pour éviter d'avoir un enfant ? », et si oui laquelle. En cas de réponse négative une question de rattrapage est prévue : « Ni vous, ni votre conjoint ne prenez aucune précaution ? »	« L'homme peut aussi pratiquer le retrait, c'est-à-dire se retirer à temps (avant l'éjaculation) »	Celle utilisée en milieu de cycle. Si plusieurs méthodes déclarées, hiérarchisation (retrait 6e sur 10).	-	9 % (actuellement)

Population et mesure du retrait

Tableau A.1 (suite). Mesure du retrait dans neuf enquêtes françaises entre 1970 et 2016

Année	Nom de l'enquête	Population interrogée	Mesure du retrait (en gras, la période concernée)	Désignation de la méthode (le cas échéant)	Méthode considérée comme principale dans le traitement des données	Taux de recours au retrait comme méthode principale (chez les personnes concernées)	
						Hommes	Femmes
2000	Cocon 2000 « Cohorte sur la contraception » Inserm/Ined	2863 femmes ayant entre 18 et 44 ans	Toutes les méthodes sont citées (le retrait est cité en 2e position), question sur la « **méthode actuelle** ». Une question de rattrapage était prévue si aucune méthode n'était déclarée : « Pouvez-vous me dire si ces phrases vous concernent en ce moment » : « 7. Mon partenaire se retire avant la fin du rapport ».	« Retrait du partenaire avant l'éjaculation »	Hiérarchisation	-	2,1 % (actuellement)
2006	CSF « Contexte de la sexualité en France » Ined/Inserm	6824 femmes et 5540 hommes de 18 à 69 ans	Question concernant le **dernier rapport sexuel**. Les méthodes ne sont pas citées, mais il y a une question de rattrapage ouverte : « Pourquoi n'utilisiez-vous pas de méthodes pour éviter une grossesse ? » dont l'une des modalités est « N'utilise pas de méthode, mais fait quand même attention »	-	-	3,3 % (dernier rapport sexuel)	2,9 % (dernier rapport sexuel)
2010	Fecond 2010 « Fécondité - contraception - dysfonctions sexuelles » Inserm	5275 femmes et 3373 hommes de 15 à 49 ans	Toutes les méthodes sont citées (le retrait est cité en 5e position) pour déterminer l'utilisation au cours de la vie et l'utilisation « **actuellement** ».	« Retrait du partenaire avant l'éjaculation »	Déclaratif	2,3 % (7,7 % en tout) (actuellement)	3,6 % (actuellement)
2013	Fecond 2013 « Fécondité - contraception - dysfonctions sexuelles » Inserm	4453 femmes et 1587 hommes de 15 à 49 ans	Les méthodes ne sont pas citées, mais les « méthodes naturelles » sont mentionnées dans la question : « **Actuellement**, est-ce que vous ou votre partenaire utilisez un moyen pour éviter une grossesse y compris une méthode naturelle, et si oui le ou lesquels ? ». Une question de rattrapage mentionne le retrait : « Parmi les phrases suivantes, lesquelles vous concernent actuellement ? […] 5. Vous/votre partenaire se retire avant l'éjaculation […] »	-	Déclaratif	4,5 % (actuellement)	5,1 % (actuellement)
2016	Baromètre santé, volet « Contraception »	4315 femmes âgées de 15 à 49 ans	Les méthodes ne sont pas citées, mais les « méthodes naturelles » sont mentionnées dans la question : « **Actuellement**, est-ce que vous ou votre partenaire utilisez une méthode pour éviter une grossesse, y compris les méthodes naturelles, et si oui laquelle ? ». Il n'y a pas de question de rattrapage.	-	Hiérarchisation (retrait 12e sur 17)	-	2,7 %[a] (actuellement)

(a) Merci à Mireille Le Guen, Nathalie Lydié et Delphine Rahib d'avoir fourni ce chiffre.

Sources

Les premières slides de ce powerpoint ont été récupérées sur le site de Martin Chevalier, administrateur à l'INSEE. Tous ses supports de cours sont disponibles en accès libres sur le site : https://teaching.slmc.fr/.

Quelques éléments de définition

Une enquête statistique est un dispositif d'observation qui :

- porte sur un échantillon d'unités (ménages, entreprises, etc.) considérées comme représentatives d'une population,
- repose sur un questionnaire et un ensemble de codifications (modalités de réponse pré-codées, nomenclatures),
- permet un traitement quantitatif des données recueillies

Exemple : l'enquête Emploi en continu

L'enquête Emploi en continu (EEC) est un des dispositifs les plus importants de la statistique publique :

- échantillon rotatif, environ 100 000 personnes de 15 ans ou plus interrogées chaque trimestre (en France métropolitaine),
- questionnaire de 50 pages, mesure du chômage selon la définition du Bureau international du travail (BIT),
- publication trimestriel du taux de chômage France métropolitaine, un des principaux indicateurs économiques nationaux et internationaux.

Enquête statistique et base de données

Toute enquête statistique peut être abordée comme une base de données mais toute base de données ne présente pas les caractéristiques d'une enquête statistique.

Exemples

- Données administratives : déclarations annuelles de données sociales (DADS), fichiers de demandeurs d'emploi, fichier des infractions police gendarmerie, etc.
- « Mégadonnées » (big data) collectées automatiquement : données de téléphone portable, données de caisse des supermarchés, etc

Ces bases de données n'ont pas été pensées à l'origine pour produire de l'information statistique : leur exploitation est complexe et nécessite des précautions

Construire un questionnaire

L'élaboration du questionnaire d'une enquête statistique constitue un arbitrage entre deux contraintes contradictoires :

- l'exhaustivité des thèmes abordés, précision de l'information collectée, comparabilité avec d'autres dispositifs existants,
- la durée de passation, caractère compréhensible des questions et absence d'« imposition de problématique ».

Comprendre les contraintes de l'élaboration d'un questionnaire permet de comprendre les choix qui ont été faits et d'être à même de « donner du sens aux données ».

Les principales contraintes dans la rédaction d'un questionnaire

Longueur du questionnaire : fortement dépendant du mode de collecte (cf. infra).
Choix du vocabulaire et formulation des questions :

- être compris par le public visé (exemple : enquête de l'UNICEF sur les enfants de 6 à 18 ans),
- ne pas influencer le répondant (biais de désirabilité, de cohérence, etc.).

Choix des questions et de leur forme :

- recueillir l'information la plus plus précise prossible,
- être économe en temps et en énergie pour le ou la répondant.e,
- l'important de l'ordre des questions

Les types de questions

Les questions fermées sous forme de réponse unique, réponses multiples ou de classement des préférences :

- Facilité de codification et de traitement, rapidité pour l'enquêté
- Information restrictive et sans nuance, risque de réponse « au hasard » ou de tentative de deviner la « bonne » réponse.

Les questions ouvertes :

- Grande liberté pour l'enquêté, possibilité d'obtenir des réponses non-prévues à la conception de questionnaire, exploitations originales (statistique textuelle).
- Difficulté de traitement, relativement coûteux pour l'enquêté (risque de non-réponse partielle).

Une phase cruciale de l'enquête

La collecte va permettre de juger du succès d'une enquête. Son déroulement détermine le niveau de (non-)réponse à l'enquête, ainsi que la qualité des informations recueillies. Le choix du mode de collecte, effectué très en amont, est déterminant.

- en face-à-face (assistée par ordinateur)
- par téléphone
- par courrier et par dépôt-retrait
- par internet

Le rôle des enquêteurices

L'enquêteur a un rôle important à jouer pour assurer le succès d'une enquête, en particulier quand il ou elle s'écarte des consignes de collecte :

- création d'une relation de confiance avec l'enquêté : ne pas montrer la lettre officielle mais parler de la télévision ou du sport, etc.
- reformuler ou passer des questions « stupides » : questions non-comprises, questions absurdes.

Statistiques descriptives et inférentielles

La méthode d'enquête par questionnaire s'inscrit dans un type de raisonnement statistique particulier, la représentativité, qui se fixe au début du XXe siècle.

- Les *statistiques descriptives* correspondent (dans un sens restreint) aux calculs réalisés sur des populations exhaustives.
- On appelle *population de référence* l'agrégat que l'on souhaite étudier : les étudiant·es, les entreprises du CAC40, les député·es de l'Assemblée Nationale etc.
- Les *statistiques inférentielles* correspondent aux opérations réalisées sur une partie de la population de référence, appelée l'*échantillon*, dans le but de tirer des conclusions **fiables** sur celle-ci. La fiabilité de nos affirmations dépend de la *représentativité* de notre échantillon – ce qui implique de mettre en place des procédures d'*échantillonnage* et de *pondération* adéquates – et de ses *effectifs*.
- Attention, un échantillon représentatif **n'est pas une miniature de la population de référence**. La représentativité dépend des méthodes d'échantillonnage et de redressement; ses définitions sont multiples.

La pondération, soit le fait d'associer des poids (coefficients) aux individus statistiques, renvoie à cet objectif de représentativité.

Les méthodes d'échantillonnage

Les méthodes d'échantillonnage permettent de construire des échantillons à partir d'une populaiton.

- Le *plan de sondage* décrit la façon dont les individus ont été sélectionnés : aléatoirement, par quotas, par effets boule de neige etc.
- Le *poids* d'un individu correspond au nombre d'individus que l'individu de l'échantillon représente dans la population. Si l'on interroge 1 individu sur 100, le poids est alors de 100.
- Le plan de sondage peut adopter des stratégies pour maximiser l'information sans augmenter la taille de l'échantillon. Le principe de *la stratification* consiste par exemple à intentionnellement sur-sélectionner les individus rares. Cette sur-sélection risque d'introduire un biais qui sera compensé par des poids inverses à la *probabilité d'inclusion* des individus.

Les méthodes de redressement

Par ailleurs, tout un ensemble de méthodes de redressement post-collecte permettent de corriger certains biais que l'on connaît. La correction de la *non-réponse* ou le redressement par *quotas* en sont des cas classiques. Elles se traduisent aussi par une modificatin des poids.

Paramètre et estimation

- L'essentiel de l'analyse quantitative en science sociale repose sur l'*estimation* de *paramètres* à l'aide d'*estimateurs* ainsi que sur l'élaboration de *modèles* confirmés ou infirmés par des *tests d'hypothèse*.
- On peut vouloir décrire la structure d'âge d'une population en calculant la moyenne de l'âge (le paramètre M). Comme on dispose rarement d'information sur l'ensemble de la population, on réalise une estimation de la valeur réelle du paramètre grâce à un outil statistique, la moyenne observée (m). La théorie statistique permet de construire des *intervalles de confiance* renseignant sur la qualité de cette estimation. Elle peut être très précise ou au contraire imprécise.
- On peut vouloir savoir si le niveau de qualification professionnelle est dépendant du genre, c'est-à-dire si le niveau de qualification diffère en fonction du genre des individus. On doit ici aussi souvent partir d'un échantillon observé et mettre en place des thèses d'hypothèse pour conclure, selon un certain *niveau de risque* ou *seuil de confiance*, s'il existe ou non un lien entre les variables étudiées.

La typologie des familles

Typologie des familles par classe d'âge en 2003			
	Type de famille (en %)		
Catégorie d'âge	pas de frères et soeurs	1 à 3 frères et soeurs	plus de 3 frères et soeurs
18 à 30 ans	9%	70%	20%
31 à 60 ans	7%	55%	38%
61 ans et plus	14%	57%	29%
Ensemble	9%	59%	32%

Lecture : en 2003, 7% des personnes âgées de 31 à 60 ans n'avaient pas de frères et soeurs. Dans l'ensemble, 9% de la population française toutes catégories d'âge confondues n'avaient pas de frères et soeurs
Champ : personnes majeures habitant en France métropolitaine
Source : Histoire de Vie 2003 (N = 2000) | A. POISSONNIER | 2023

- Les personnes âgées de 18 à 30 ans sont sur-représentées parmi les familles d'1 à 3 frères et soeurs (70% contre 59% dans l'ensemble) et sous-représentées parmi les familles de plus de 3 frères et soeurs (20% contre 32% dans l'ensemble).
- Les personnes âgées de 31 à 60 ans sont sous-représentées parmi les familles d'1 à 3 frères et soeurs (55% contre 59% dans l'ensemble) et sur-représentées parmi les familles de plus de 3 frères et soeurs (38% contre 32% dans l'ensemble).

Un encadrement de l'estimation

Sous certaines conditions de méthode d'échantillonnage, il est possible de fournir un encadrement de notre estimation par le biais d'un *intervalle de confiance*.

La valeur des bornes est donnée par la formule suivante :

$$Bornes = estimation \pm marge\ d'erreur \tag{1}$$

Soit, pour un intervalle dont le *niveau de confiance* est fixé à 95% :

$$Bornes = \bar{m} \pm 1.96 \frac{s}{\sqrt{n}} \tag{2}$$

Où $\bar{m}$ est la moyenne estimée dans notre échantillon, 1.96 un facteur lié à notre niveau de confiance, s l'écart-type estimé et n la taille de l'échantillon.

Ce qui donne un intervalle de confiance de la forme :

$$IC = [\bar{m} - 1.96 \frac{s}{\sqrt{n}}; \bar{m} + 1.96 \frac{s}{\sqrt{n}}] \tag{3}$$

Ce qui fait varier l'amplitude de l'intervalle

L'amplitude de l'intervalle (la distance entre les deux bornes) varie donc avec :

- Le niveau de confiance choisi : 50%, 90%, 95% ou encore 99%. Plus je veux être confiant dans ce que j'affirme, moins je peux me permettre d'être précis, donc mon IC s'agrandit.
  Par exemple, un IC à 90% aura des bornes de la forme :
  $Bornes = \bar{m} \pm \mathbf{1.645} \frac{s}{\sqrt{n}}$ plutôt que $Bornes = \bar{m} \pm \mathbf{1.96} \frac{s}{\sqrt{n}}$.
- La niveau de variabilité (écart-type et variance) de la variable.
- La taille de mon échantillon.

Interpréter un intervalle

L'intervalle de confiance à 95% de l'âge moyen peut s'interpréter comme tel :

> *J'ai une probabilité de 95% d'avoir raison (ou 5% de me tromper) en affirmant que mon intervalle de confiance construit autour de mon estimation de l'âge moyen (plutôt qu'une autre estimation tirée d'un autre échantillon) contient la vraie valeur de l'âge moyen dans la population.*

On évitera en revanche de dire :

> *Il y a une probabilité de 95% que la vraie valeur de l'âge moyen dans la population soit dans mon intervalle de confiance.*

Car cette probabilité est égale à 1 (si elle est dedans) ou 0 (si elle ne l'est pas), mais jamais autre chose.

L'incertitude des estimations

Typologie des familles par classe d'âge en 2003			
	Type de famille (en %)		
Catégorie d'âge	pas de frères et soeurs	1 à 3 frères et soeurs	plus de 3 frères et soeurs
18 à 30 ans	9% ± 2%	70% ± 0%	20% ± 1%
31 à 60 ans	7% ± 1%	55% ± 0%	38% ± 0%
61 ans et plus	14% ± 1%	57% ± 0%	29% ± 1%
Ensemble	9%	59%	32%

Lecture : en 2003, 7% des personnes âgées de 31 à 60 ans n'avaient pas de frères et soeurs. Dans l'ensemble, 9% de la population française toutes catégories d'âge confondues n'avaient pas de frères et soeurs
Champ : Individus de 18 ans et plus habitant en France métropolitaine
Source : Histoire de Vie 2003 (N = 2000) | A. POISSONNIER | 2023