Introduction à la théorie des sondages - Cours 4

Gaspar Massiot
gaspar.massiot@ined.fr

2024-2025

Sources d'erreur en sondage

- Erreur d'échantillonnage
  - Taille de l'échantillon
  - Plan d'échantillonnage
  - Estimateur choisi
  - Variabilité du paramètre
- Erreur de mesure
  - Enquêté.e
  - Questionnaire
  - Saisie
- Non-réponse
  - Totale : entrée vide
  - Partielle : seule une partie du questionnaire est remplie.

Sources d'erreur en sondage

- Erreur d'échantillonnage ✓
  - Taille de l'échantillon ✓Précision L
  - Plan d'échantillonnage ✓SAS, stratifié, systématique, équilibré, allocations proportionnelle et de Neyman
  - Estimateur choisi ✓Horvitz-Thompson, Hájek
  - Variabilité du paramètre ✓Stratification, complété aujourd'hui
- Erreur de mesure Sujet du jour
  - Enquêté.e
  - Questionnaire
  - Saisie des données
- Non-réponse Sujet du cours 5
  - Totale : entrée vide
  - Partielle : seule une partie du questionnaire est remplie.

Sommaire

Chapitre 1

Rédiger un questionnaire

Questionnaire

Une fois les objectifs identifiés, il faut réaliser un questionnaire :

- Qui permette de répondre au mieux à la problématique de l'enquête
- Tout en s'assurant de la coopération de l'enquêté.e
- Et de sa compréhension des questions et de leur but
- Qui permette de la comparabilité avec d'autres sources

$\Rightarrow$ Une étape cruciale mais difficile !

Partie 1

Les outils pour la création de questionnaire

Comment faire un questionnaire ?

Les questionnaires des enquêtes historiques étaient fait à la main, ou à la machine à écrire.
Word est ainsi l'outil principal utilisé pour faire un questionnaire papier, même si on peut améliorer la mise en page avec d'autres outils.

Comment faire un questionnaire ?

Il existe des outils spécifiques pour créer des questionnaires, notamment en ligne :

- **Google Forms**, mais attention aux données personnelles !
- **LimeSurvey**, une alternative plus respectueuse
- De nombreux autres logiciels existent (logiciels métiers, outils développés, etc.)

Partie 2

Les questions à se poser

Les questions à se poser

Les questions à se poser lors de la rédaction du questionnaire :

- Durée du questionnaire cohérente avec le mode de collecte ?
- Quel(s) format(s) pour mes questions ? (différent selon l'objectif)
- Comment rendre mes questions non-ambigües ?
- Dans quel ordre mes questions sont-elles présentées ?
- Est-ce que cet ordre peut avoir un impact sur la réponse ?
- Pour un questionnaire papier ou internet, quelle présentation visuelle ?

Durée du questionnaire

Face-à-face :

- à domicile, 60 minutes maximum,
- sur site, le questionnaire doit être plus court (10 à 15 minutes).

Téléphone :

- 30 minutes maximum pour les téléphones fixes
- 20 minutes maximum pour les téléphones mobile

Auto-administré :

- le questionnaire papier peut être assez long à remplir (possibilité de le faire en plusieurs fois)
- Internet : on préconise un maximum de 15 minutes

Formats de questions

- Question ouverte (préféré en qualitatif)
- Question fermée (préféré en quantitatif)
  - choix binaire (oui / non)
  - choix multiple (région de résidence)
  - échelle d'évaluation :
    - Pas du tout, plutôt pas, plutôt d'accord ou tout à fait d'accord ?
    - Note de satisfaction de 1 à 10
  - classement des modalités (3 raisons principales dans l'ordre)
- Question semi-ouverte : "quel est votre secteur d'activité ?" avec du recodage

Des soucis d'interprétation

Des soucis d'interprétation

Premier exemple

Comment faire pour savoir si l'enquêté.e est cadre de son entreprise ?

- Lui demander explicitement : "êtes-vous cadre ?"

Des soucis d'interprétation

Premier exemple

Comment faire pour savoir si l'enquêté.e est cadre de son entreprise ?

- Lui demander explicitement : "êtes-vous cadre ?"
- Poser plusieurs questions qui permettent de l'identifier :
  - Quels horaires ?
  - Quel salaire ?
  - Encadrez-vous une équipe ?

Des soucis d'interprétation

Premier exemple

Comment faire pour savoir si l'enquêté.e est cadre de son entreprise ?

- Lui demander explicitement : "êtes-vous cadre ?"
- Poser plusieurs questions qui permettent de l'identifier :
  - Quels horaires ?
  - Quel salaire ?
  - Encadrez-vous une équipe ?
- Poser des questions sur des éléments spécifiques : "votre contrat mentionne t-il explicitement que vous êtes cadre ?"

Des soucis d'interprétation

Deuxième exemple

Imaginez la question "Combien de fenêtres y a t-il chez vous ?".
Quels problèmes peut-il y avoir ?

Des soucis d'interprétation

Deuxième exemple

Imaginez la question "Combien de fenêtres y a t-il chez vous ?".
Quels problèmes peut-il y avoir ?

- Je possède plusieurs logements.

Des soucis d'interprétation

Deuxième exemple

Imaginez la question "Combien de fenêtres y a t-il chez vous ?".
Quels problèmes peut-il y avoir ?

- Je possède plusieurs logements.
- Les porte-fenêtres / hublots / fenêtres de toit comptent ?

Des soucis d'interprétation

Deuxième exemple

Imaginez la question "Combien de fenêtres y a t-il chez vous ?".
Quels problèmes peut-il y avoir ?

- Je possède plusieurs logements.
- Les porte-fenêtres / hublots / fenêtres de toit comptent ?
- Je n'ose pas dire que mon logement est petit / aveugle

Des soucis d'interprétation

Deuxième exemple

Imaginez la question "Combien de fenêtres y a t-il chez vous ?".
Quels problèmes peut-il y avoir ?

- Je possède plusieurs logements.
- Les porte-fenêtres / hublots / fenêtres de toit comptent ?
- Je n'ose pas dire que mon logement est petit / aveugle
- Je ne sais pas, mon logement est trop grand

Influence de l'ordre

L'ordre dans lequel sont posées les questions peut jouer. Un exemple classique :

1. Pensez-vous que l'URSS doive envoyer des journalistes aux USA pour informer leurs citoyens ?
2. Pensez-vous que les USA doivent envoyer des journalistes en URSS pour informer leurs citoyens ?

vs

1. Pensez-vous que les USA doivent envoyer des journalistes en URSS pour informer leurs citoyens ?
2. Pensez-vous que l'URSS doive envoyer des journalistes aux USA pour informer leurs citoyens ?

Influence de l'ordre

Cela arrive dans de nombreuses enquêtes :

1. Souhaitez-vous retrouver un emploi ?
2. Avez-vous fait des démarches pour trouver un emploi dans les 15 derniers jours ?

vs

1. Avez-vous fait des démarches pour trouver un emploi dans les 15 derniers jours ?
2. Souhaitez-vous retrouver un emploi ?

Des soucis de présentation

Zone de réponse

a. Quand avez-vous commencé vos études à l'université de Washington ?

b. En quel mois et quelle année avez-vous commencé vos études à l'université de Washington ?

Mois Année

c. En quel mois et quelle année avez-vous commencé vos études à l'université de Washington ?
Merci de répondre en utilisant deux chiffres pour le mois, et quatre chiffres pour l'année

Mois Année

d. En quel mois et quelle année avez-vous commencé vos études à l'université de Washington ?
Merci de répondre en utilisant deux chiffres pour le mois, et quatre chiffres pour l'année

MM AAAA

Des soucis de présentation

Question TIC

→ 1. Au cours des trois derniers mois, en dehors de chez vous, vous avez utilisé internet...

Plusieurs réponses possibles

- ☐ sur votre lieu de travail.
- ☐ sur votre lieu d'études
- ☐ chez des membres de votre famille, des amis, voisins
- ☐ dans un autre lieu (cybercafé, bibliothèque, hôtel, aéroport, etc.)

Des soucis de présentation

Question TIC

→ 1. Au cours des trois derniers mois, en dehors de chez vous, vous avez utilisé internet...

Plusieurs réponses possibles

- ☐ sur votre lieu de travail.
- ☐ sur votre lieu d'études
- ☐ chez des membres de votre famille, des amis, voisins
- ☐ dans un autre lieu (cybercafé, bibliothèque, hôtel, aéroport, etc.)

Comment séparer la non-réponse et le fait de ne pas utiliser Internet dehors ?

Partie 3

Des pistes de solutions

Modules classiques

Il existe des jeux de questions que l'on pose classiquement :

- l'INSEE utilise le Tronc Commun des Ménages (TCM) pour lister les occupants du ménage
- un module Profession et Catégorie Socio-Professionnelles (PCS) permet de coder la profession suivant une nomenclature dédiée

Modules classiques

Il existe des jeux de questions que l'on pose classiquement :

- en santé, il existe des modules sur la santé mentale ou physique, et une question sur la dépendance GALI (Global Activity Limitation Indicator) : "Êtes-vous limité(e), depuis au moins six mois, à cause d'un problème de santé, dans les activités que les gens font habituellement ?"
  1. Oui, fortement limité(e)
  2. Oui, limité(e), mais pas fortement
  3. Non, pas limité(e) du tout

Effectuer un ou des tests de questionnaire

Toujours tester le questionnaire pour :

- Affiner la rédaction des items
- Améliorer la fluidité du questionnaire
- Estimer le temps nécessaire pour y répondre
- Vérifier les filtres (questions imbriquées)

Effectuer un ou des tests de questionnaire

Toujours tester le questionnaire pour :

- Affiner la rédaction des items
- Améliorer la fluidité du questionnaire
- Estimer le temps nécessaire pour y répondre
- Vérifier les filtres (questions imbriquées)

Risques liés à un questionnaire non testé :

- Incompréhensions
- Lassitude
- Abandons
- Réponses incomplètes

Lignes directrices

Travaux de Rebecca Morrison (Census Bureau, USA).
Lignes directrices concernant :

- la formulation des questions ;
- la clarté de la mise en forme ;
- la parcimonie de l'information ;
- la clarté et l'accessibilité des instructions ;
- la clarté du cheminement.

Lignes directrices

Enoncé

1. Formuler les demandes sous forme de questions ou d'énoncés impératifs, et non sous forme de fragments de phrase ou de mots clés.
2. Décomposer les questions complexes en une série de questions simples.

Lignes directrices

Espaces de réponse et options

3. Utiliser des espaces blancs sur fond de couleur pour mettre en relief les espaces de réponse.
4. Utiliser des espaces de réponse semblables pour les demandes d'un même type d'information.
5. Indiquer clairement l'unité de mesure.
6. Décider s'il faut ou non fournir aux répondants les données qu'ils ont déclarées antérieurement après avoir pesé les avantages et les risques éventuels en matière de qualité des données et les risques éventuels de divulgation.
7. Fournir des cases à cocher « Inscrire "X" si aucun(e) » s'il faut faire la distinction entre la non-réponse partielle et les valeurs nulles déclarées.

Lignes directrices

Désordre visuel

8. Utiliser les diverses polices de caractère uniformément et à une seule fin dans un questionnaire.
9. Regrouper les éléments de données et leurs espaces de réponse/options de réponse.
10. Évaluer la nécessité de tout graphique, image ou diagramme, afin de confirmer leur utilité pour les répondants.

Lignes directrices

Cheminement clair

11. Formater l'instrument uniformément, en tirant parti des habitudes de lecture familières.
12. Indiquer clairement le début de chaque question et de chaque partie.
13. Regrouper les éléments de données similaires.
14. Utiliser des espaces blancs pour séparer les questions et faciliter le cheminement dans les questionnaires.
15. Aligner les questions et les espaces de réponse/options de réponse.
16. Utiliser de bonnes caractéristiques visuelles pour mettre l'accent sur les instructions « Passez à ».
17. Informer les répondants du cheminement lorsqu'une question continue sur une autre page.

Lignes directrices

Clarté des instructions

18. Intégrer des instructions propres aux questions dans le questionnaire là où elles sont nécessaires. Éviter de placer les instructions sur une feuille, brochure ou page Web distincte.
19. Envisager de reformuler les instructions importantes sous forme de questions.
20. Dans la mesure du possible, utiliser une date réelle plutôt qu'un délai vague pour faire référence aux dates d'échéance.

Lignes directrices

Utilisation de matrices

21. Limiter l'utilisation de matrices. Considérer le niveau de familiarité possible des répondants avec les tableaux et les matrices avant de décider ou non de les utiliser.
22. Si une matrice est nécessaire, aider les répondants à traiter l'information en réduisant le nombre d'éléments de données recueillis et en établissant un cheminement clair.

Un exemple de question matrice

Source : creerunquestionnaire.fr

Lignes directrices

Travaux de Rebecca Morrison (Census Bureau, USA).
Lignes directrices concernant :

- la formulation des questions (questions simples ou à l'impératif) ;
- la clarté de la mise en forme (uniformité, contraste) ;
- la parcimonie de l'information (graphiques, matrices) ;
- la clarté et l'accessibilité des instructions (un seul document, unité de mesure, date) ;
- la clarté du cheminement (aligner, regrouper, changement de page).

Chapitre 2

La collecte d'une enquête

La collecte

- Le **mode de collecte** employé a non seulement un impact sur la conception du questionnaire mais peut aussi impacter la qualité des données par son effet sur l'enquêté.e et l'enquêteurice par exemple.
- La collecte est aussi le moment où l'on peut contrôler les erreurs de mesure dues à la saisie des données et aux enquêté.es/enquêteurices.
  - En mettant en place des processus de contrôle pendant le **suivi de la collecte** par exemple.
  - Tests automatiques sur la qualité des réponses,
  - Sélection aléatoire d'enquêteurices à inspecter, etc.
- La qualité des données peut aussi être étudiée au moment du **bilan de la collecte**.

Partie 1

Modes de collecte

Mode de collecte

Plusieurs modes de collecte :

- CAPI : enquêteur en face à face
- CATI : enquêteur par téléphone
- CAWI : collecte par Internet
- Papier : questionnaire auto-administré

Ces modes peuvent être proposés concurentiellement ou séquentiellement : on parle alors de "multimode", pour modes multiples.

Mode de collecte

Il existe d'autres modes de collecte :

- Collecte passive : capteurs de poussière, de pollution, GPS...
- Collecte spécifique : échantillons sanguins, examens médicaux...

Ils sont souvent conçus pour les besoins d'une enquête spécifique.

Multimode

Lorsque plusieurs modes sont proposés :

- Effet de mode :
  - On ne répond pas pareil au téléphone et sur papier
  - Biais de désirabilité
  - Exemple classique : consommation de drogues des adolescents
- Effet de sélection :
  - Ce ne sont pas les mêmes personnes qui répondent sur Internet que sur papier
  - Personnes plus jeunes et plus diplômées, ayant accès à Internet.

Multimode à l'INSEE

Dans les enquêtes de l'INSEE :

- CAPI : la plupart des enquêtes
- CATI : intermédiaires de l'Emploi, TIC
- CAWI : TIC, Recensement
- Papier : Recensement

Volonté de développer le recours au multimode dans les enquêtes de la statistique publique.

Partie 2

Suivi de collecte

Suivi de collecte

Principe :

- Au cours de la collecte, on sait qui a répondu ou pas
- Les profils peuvent être équilibrés ou non
- On peut concentrer les relances, par exemple téléphoniques
- On peut même débloquer des échantillons dits de réserve

Dispositif mis en oeuvre par les prestataires de collecte, et certains instituts comme Statistiques Canada.

Suivis spécifiques

Pour certains cas, des suivis spécifiques sont organisés :

- Enquêtes qui suivent des populations spécifiques (ENRJ [1] sur les ressources des jeunes adultes de 18 à 24 ans).
- Enquêtes qui ont des taux de collecte peu satisfaisants : par exemple, BDF [2] en Guyane (mars-avril 2017).

Dans ce cas, les équipes statistiques

- étudient la structure des répondants par rapport à l'échantillon ;
- s'assurent que les objectifs pour chaque sous-population seront atteints.

1. https://drees.solidarites-sante.gouv.fr/sources-outils-et-enquetes/1-enquete-nationale-sur-les-ressources-des-jeunes
2. Budget de famille https://www.insee.fr/fr/metadonnees/source/serie/s1194

Corrections

En cas de soucis :

- Prolongation de la collecte ;
- Déblocage d'un échantillon complémentaire ou de réserve ;
- Priorisation de certains profils.

Remarque : en amont, on peut essayer d'organiser la collecte par vague pour faciliter le suivi.

Partie 3

Bilan de la collecte

Bilan de la collecte

On a trois catégories d'individus ou de logements :

- **Répondants** : ces individus sont considérés comme répondants, i.e les informations qu'ils ont fourni pendant la collecte seront mobilisées pour l'exploitation.
- **Non répondants** : on aurait souhaité que ces individus nous donnent des informations car ils nous intéressent dans le cadre de l'enquête, mais ce n'est pas le cas, pour différentes raisons.
- **Hors champ** : les réponses données ou non par ces individus ne nous intéressent pas, car ils sont hors champ de l'enquête. En revanche, estimer leur nombre peut être intéressant, surtout si le champ est complexe.

Résultats

Quelques exemples de résultats :

- Enquête réussie et validée. Considérée comme répondante.
- Enquête réussie partiellement : seules certaines questions ont été répondues.
- Logement vacant
- Impossible à joindre
- Déjà enquêté par le prestataire

Grande question : s'agit-il de non-réponse (l'individu n'a pas voulu répondre) ou de hors-champ (l'individu n'est pas concerné par l'enquête) ?

Chapitre 3

Compléments sur l'échantillonnage

Partie 1

Tirage systématique et stratification implicite

Compléments sur l'échantillonnage

Algorithme de tirage systématique

On cherche à effectuer un tirage de taille fixe n dans une population N. Chaque unité k de $\mathcal{U}$ dispose d'une probabilité d'inclusion simple π_k.

L'ordre des unités dans la base de sondage est fixé : on définit le cumul des probabilités d'inclusion $a_k = \sum_{k'=1}^{k} \pi_{k'}$.

L'algorithme de tirage systématique est alors le suivant :

1. On tire un réel η dans une loi uniforme sur $[0\,;1]$.
2. On sélectionne toutes les unités k vérifiant :

$$a_{k-1} \leq \eta + j - 1 < a_k$$

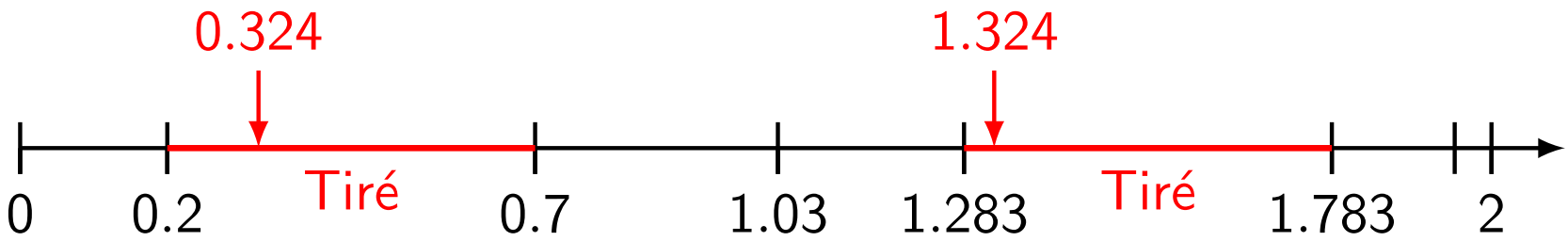
où j parcourt $1, \ldots, n$.

Compléments sur l'échantillonnage

Exemple de tirage systématique

$N = 7 \quad n = 2 \quad \sum_{k=1}^{7} \pi_k = 2 \quad \eta = 0,324$

k	1	2	3	4	5	6	7
π_k	0,2	0,5	0,33	0,25	0,5	0,166	0,05
a_k	0,2	0,7	1,03	1,283	1,783	1,950	2,00



L'échantillon tiré est $s = \{2, 5\}$.

Compléments sur l'échantillonnage

Propriétés du tirage systématique

1. Le sondage est à taille fixe et respecte les π_k.
2. C'est un algorithme efficace : un seul parcours de la base de sondage est nécessaire.
3. Selon l'ordre du fichier, des probabilités d'inclusion doubles π_{kl} peuvent être nulles : les estimateurs de variance de l'estimateur de Horvitz-Thompson sont alors biaisés.

Compléments sur l'échantillonnage

Stratification implicite

Quand la base de sondages est triée selon une ou plusieurs variables, mettre en œuvre un algorithme de tirage systématique sur l'ensemble de la base induit une **stratification implicite**.

En termes de précision, on obtient en effet un plan de sondage approximativement équivalent à un **sondage stratifié** :

1. dans les strates composées par les variables de tri ;
2. avec un SAS au sein de chaque strate ;
3. et une allocation proportionnelle.

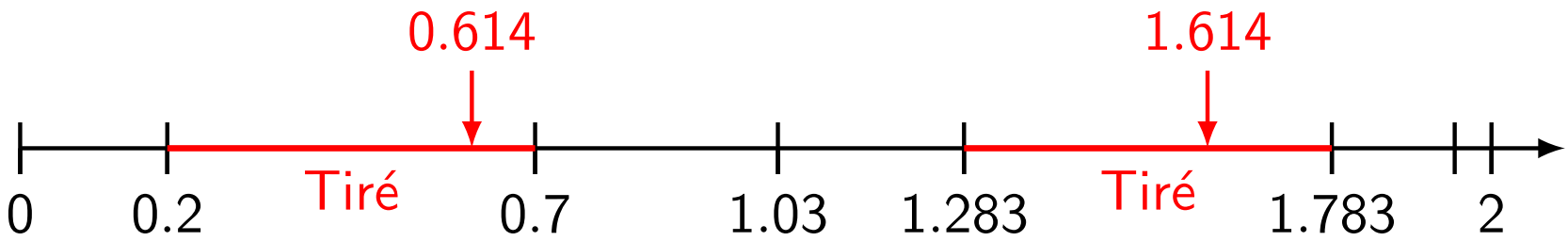
Un tirage systématique sur fichier trié ne peut donc qu'améliorer la précision de tous les estimateurs de l'enquête.

Compléments sur l'échantillonnage

Retour sur l'exemple de tirage systématique

$N_H = 3 \quad N_F = 4 \quad n = 2 \quad \sum_{k=1}^{7} \pi_k = 2 \quad \eta = 0,614$

k	1	2	3	4	5	6	7
Sexe	H	H	H	F	F	F	F
π_k	0,2	0,5	0,33	0,25	0,5	0,166	0,05
a_k	0,2	0,7	1,03	1,283	1,783	1,950	2,00



L'échantillon tiré est $s = \{2, 5\}$, stratifié entre hommes et femmes.

Compléments sur l'échantillonnage

Arbitrage entre précision des estimateurs et estimation sans biais de la précision

Intérêt : Quand la stratification devient trop fine, les estimateurs deviennent instables. On peut alors recourir à une stratification implicite par tirage systématique.

Arbitrage : Certaines probabilités d'inclusion double devenant nulle, les estimateurs de variance sont biaisés. On gagne certes en variance, mais on ne peut plus l'estimer sans biais.

En pratique, on préfère souvent une variance plus faible, même si cela signifie ne plus pouvoir l'estimer sans biais.

Partie 2

Tirage équilibré

Retour sur l'échantillon "représentatif"

Lorsque l'on réalise un sondage aléatoire simple, on ne connaît pas la structure de l'échantillon obtenu : ratio homme/femme, etc.

Pour pallier ce problème, on peut stratifier ou faire un tirage systématique.

Mais comment faire si l'on souhaite une structure précise pour :

- Sexe ;
- Âge ;
- Région...

Tirage équilibré

Cela demanderait trop de strates : à chaque fois qu'on rajoute un critère, il faut le croiser avec tous les autres, ce qui augmente très rapidement le nombre de strates.

Une autre méthode est possible : l'échantillonnage équilibré

- On choisit des variables X pour la structure : qualitatives ou quantitatives
- On sélectionne un échantillon s qui est correct sur ces variables X, c'est à dire que :

$$\hat{X}_{HT} = T(X)$$

- Si ce n'est pas possible, on cherche à être le plus proche possible.

En pratique

Comment utiliser le tirage équilibré en pratique ?

- Méthode réjective : tirer des échantillons jusqu'à obtenir un échantillon qui convienne. Problème : quelles sont les vraies probabilités de sélection ?
- Méthode du Cube, qui respecte les π_i. Méthode assez complexe, qui est implémentée :
  - en SAS, via la macro Cube : `https://www.insee.fr/fr/information/2021904`
  - en R, par exemple dans les packages *sampling* et *BalancedSampling*.

Partie 3

Sondage à deux degrés

Introduction

On parle de sondage à deux degrés lorsque :

- On réalise un premier degré de tirage (par exemple : on sélectionne des logements) en suivant un plan de sondage spécifique.
- L'échantillon obtenu permet de constituer la base de sondage du second degré (ici, l'ensemble des occupants des logements)
- On définit alors un plan de sondage pour faire un tirage dans cette base (on sélectionne des individus)
- Le dernier plan de sondage peut dépendre de l'étape de tirage du premier degré.

Echantillon-maître

Cas classique :

- Le premier degré est constitué de communes ou de zones géographiques.
- L'échantillon de premier degré correspond à un territoire sur lequel l'enquête va se faire.
- On échantillonne ensuite des logements au sein de ce territoire.
- Remarque : on parle d'échantillon-maître lorsque ce tirage est fait pour plusieurs enquêtes (Octopusse 2009-2019).

Avantages et inconvenients

Avantages :

- Gestion : on sait où vont être les enquêtes.
- Coût : on n'a pas besoin de couvrir tout le territoire / tous les logements.

Inconvénients :

- Complexité : plus difficile à mettre en oeuvre qu'un sondage dispersé.
- Réinterrogation : plus de risques de ré-enquêter la même personne.
- Imprécision : on diminue la précision des estimations (voir après).

Précision ?

Imaginons une île où il y aurait deux communes très éloignées :

- A : ville de 1 000 habitants où tout le monde est très riche.
- B : ville de 1 000 habitants où tout le monde est très pauvre.

On veut estimer le revenu moyen sur l'île en interrogeant 50 personnes.

Cas 1 : dispersé

On pourrait réaliser un sondage stratifié selon la commune :

- A : 25 personnes interrogées : $Y_1 \approx \ldots \approx Y_{25} \approx 100$
- **et** B : 25 personnes interrogées : $Y_{26} \approx \ldots \approx Y_{50} \approx 0$

Chaque échantillon donne $\bar{Y} \approx 50$. On est très précis. Cependant, il faut un enquêteur dans chaque ville.

Cas 2 : concentré

On pourrait réaliser un sondage à deux degrés :

- A : 50 personnes interrogées : $Y_1 \approx \ldots \approx Y_{50} \approx 100$
- **ou** B : 50 personnes interrogées : $Y_1 \approx \ldots \approx Y_{50} \approx 0$

Chaque échantillon donne $\bar{Y} \approx 100$ ou $\bar{Y} \approx 0$. La moyenne sur tous les échantillons est bien de 50. On est sans biais et très peu précis ; mais à moindre coût.

Premier degré

Dans la pratique, on ne veut pas être dans ce cas. Comment choisir les zones géographiques (premier degré) ?

- Il faut des zones suffisamment grandes pour pouvoir enquêter les logements nécessaires, mais peu étendues pour réduire le coût.
- Il faut des zones hétérogènes en intra pour assurer une diversité.
- Il faut des zones similaires entre elles pour limiter la perte de précision liée au fait de choisir A plutôt que B $\rightarrow$ il faudrait plutôt prendre C et D comme chacune la moitié de A et la moitié de B ; mais c'est très étendu !

Auto-pondéré

Quel plan de sondage utiliser ?

- Notre principal objectif : tous les logements doivent avoir le même poids.
- Objectif secondaire : chaque zone a la même charge de collecte.
- Solution : on tire à probabilités inégales les zones puis uniformément les logements.

Définitions

Le sondage uniforme stratifié consiste à tirer le même nombre m d'individus dans chaque strate.

Définitions

Le sondage à probabilités inégales est un sondage où les π_i sont différents entre eux.

Une méthode classique : le sondage à probabilités inégales selon une variable X disponible dans la base de sondage. On veut ici que π_k et X_k soient proportionnels : plus X est fort, plus il y a de chances qu'on sélectionne un individu. On a :

$$\pi_k = n\frac{X_k}{\sum_{k\in U} X_k}$$

Sondage auto-pondéré

1. On échantillonne des zones à probabilités inégales selon X, le nombre de logements ;
2. On sélectionne 20 logements dans la zone.

Pourquoi ils ont la même probabilité ?

- Une zone de 2 000 logements a une chance sur 20 d'être sélectionnée. Ensuite, chaque logement a une chance sur 100 : au total, une chance sur 2 000.
- Une zone de 20 000 logements a une chance sur 2 d'être sélectionnée. Ensuite, chaque logement a une chance sur 1 000 : au total, une chance sur 2 000.

Lien avec l'équilibrage

Lorsque l'on tire un échantillon-maître pour plusieurs enquêtes, voire plusieurs années, on favorise souvent un tirage équilibré au premier degré afin de :

- Garantir que tous les types des zones soient présentes dans l'échantillon, selon les variables utilisées pour l'équilibrage (richesse, structures de population, taux de chômage...)
- Limiter la perte en précision liée au premier degré.

L'échantillon-maître INSEE

À l'INSEE, les enquêtes sont réalisées dans l'échantillon maître (dit Nautile) :

- Environ 500 UP (Unités Primaires) qui sont des communes ou groupes de communes ;
- Zones tirées pour les enquêtes entre 2020 et 2029 ;
- Le tirage a été équilibré sur des totaux de population, des revenus, etc.

Concrètement, les enquêteurs sont fixes vont réaliser leurs enquêtes au même endroit pendant 10 ans.

Zones exhaustives

Une problématique régulière avec ce type d'échantillon concerne les très grandes communes :

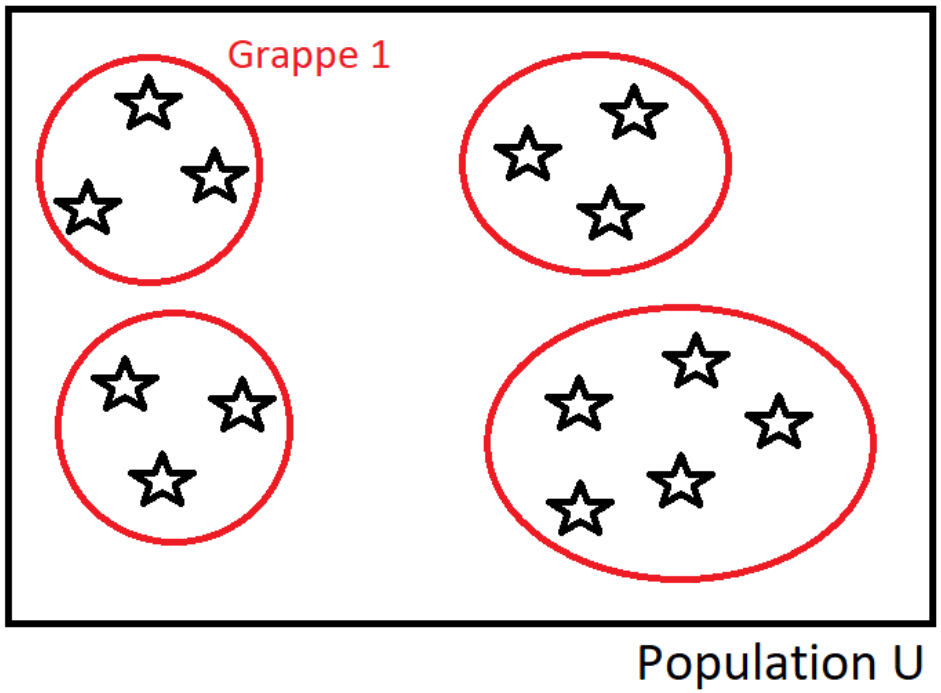
- Si on tire proportionnellement à leur taille, la probabilité dépasse 1 ;
- On les inclut donc automatiquement dans l'échantillon de zones ;
- Mais alors combien de logements tirer ? On travaille au prorata entre zones exhaustives et zones non exhaustives ;
- Exemples : Paris, Lyon, Marseille...

Partie 4

Sondage par grappes

Grappes

Le sondage par grappes repose sur l'idée que la population U est répartie entre groupes, qu'on appelle "grappes". On réalise ensuite un échantillon de grappes, et on va enquêter l'intégralité des occupants de la strate.



Grappes

Quelques exemples :

- On peut échantillonner des bâtiments pour enquêter tous les occupants (Recensement de la Population) ;
- On peut enquêter tous les occupants d'un ménage tiré au hasard

Attention :

- Ne pas confondre avec le sondage stratifié : on tire certaines grappes et non dans chaque strate.
- Ressemble au sondage à deux degrés

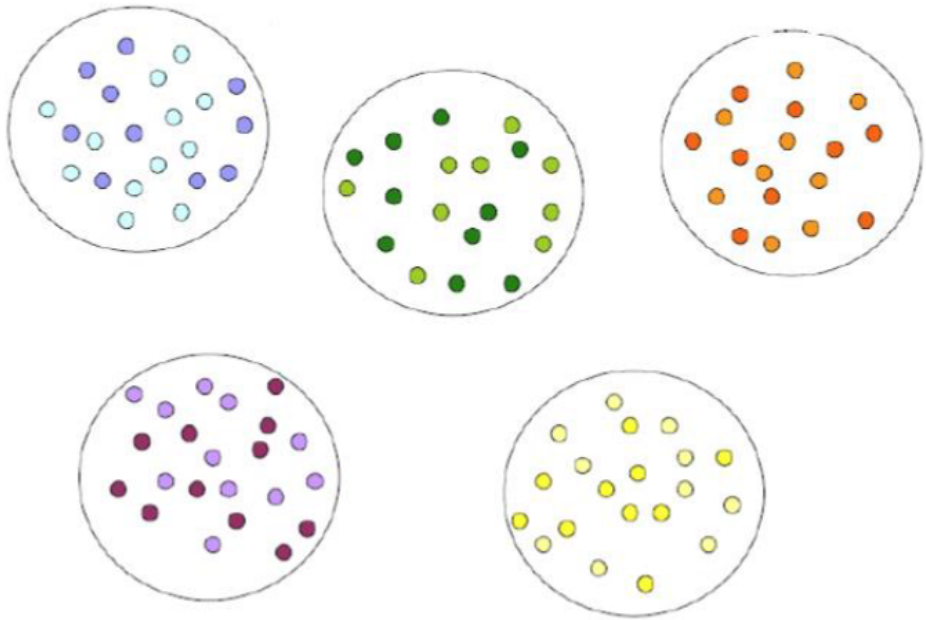
Effet de grappe

On appelle "effet de grappe" le phénomène de perte de précision due à la similarité des comportements entre individus d'une même grappe.
Lorsque toutes les grappes sont de même taille N_g, L'effet de grappe peut être mesuré par un coefficient appelé "coefficient de corrélation intra-grappe" noté ρ :

$$\rho = \frac{1}{N_g - 1}\left[N_g \frac{\sigma^2_{y,inter}}{\sigma^2_y} - 1\right]$$

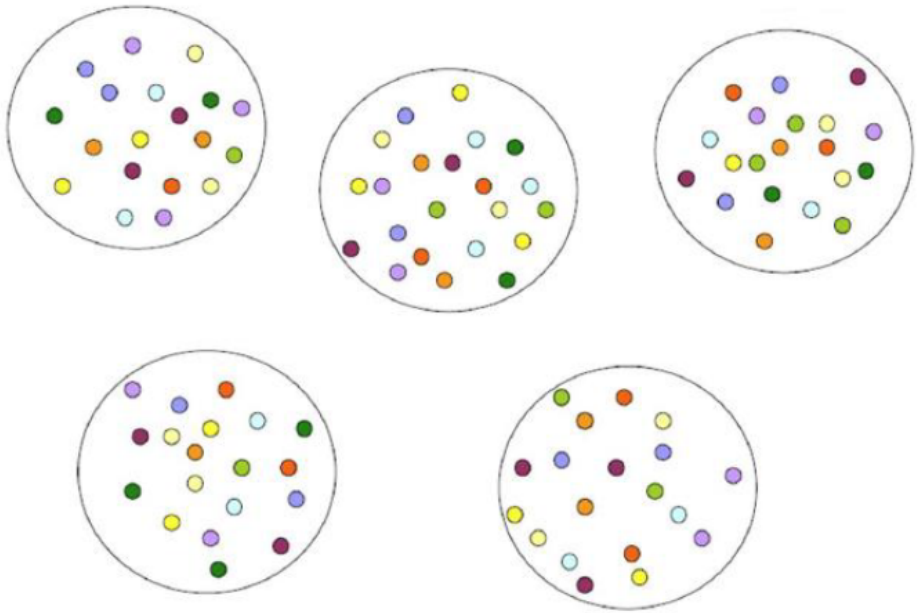
Effet de grappe

Quand ρ est grand, les grappes regroupent des individus qui se ressemblent :



Effet de grappe

Quand ρ est petit, à l'inverse, les grappes regroupent des individus différents entre eux :



Effet de grappe

Utiliser un sondage par grappes conduit à une précision dégradée ; on appelle *design effet* le ratio entre la précision obtenue et celle obtenue avec un SAS, et sa formule est :

$$Deff = \frac{Var_{Grappes}}{Var_{SAS}} = 1 + \rho(N_g - 1)$$

Ce ratio est d'autant plus grand que les grappes sont grandes (N_g) et homogènes (ρ). Malheureusement, souvent, les grappes sont des zones géographiques qui regroupent des individus qui se ressemblent (c'est le cas dans l'Enquête Emploi en Continu de l'Insee, par exemple).

Partie 5

Méthodes empiriques

Méthodes des quotas

La méthode des quotas est une méthode de sondage non probabiliste souvent utilisée par les instituts de sondage :

- On fixe le quota que l'on souhaite atteindre pour chaque catégorie : femmes ; cadres ; habitants en Bretagne . . .

Méthodes des quotas

La méthode des quotas est une méthode de sondage non probabiliste souvent utilisée par les instituts de sondage :

- On fixe le quota que l'on souhaite atteindre pour chaque catégorie : femmes ; cadres ; habitants en Bretagne . . .
- On appelle de nombreuses personnes : une personne souhaitant participer ne peut le faire que si elle rentre dans le quota.

Méthodes des quotas

La méthode des quotas est une méthode de sondage non probabiliste souvent utilisée par les instituts de sondage :

- On fixe le quota que l'on souhaite atteindre pour chaque catégorie : femmes ; cadres ; habitants en Bretagne . . .
- On appelle de nombreuses personnes : une personne souhaitant participer ne peut le faire que si elle rentre dans le quota.
- On obtient à la fin un échantillon avec la structure souhaitée.

Méthodes des quotas

Cette méthode n'est pas un sondage "probabiliste" au sein du cours, mais est utilisée dans la plupart des enquêtes d'opinion et politiques. Elle a plusieurs avantages :

- Ne demande pas de disposer à l'avance d'une base de sondage ;
- Respecte une logique proche de celle d'un sondage stratifié à allocation proportionnelle (ce qui améliorerait la précision) ;
- Moins coûteux car permettant de solliciter de nombreux individus

Méthodes des quotas

Cela pose de nombreux problèmes :

- Les premiers à répondre ont plus de chance d'être dans les quotas ;
- Comment trouver les derniers répondants ? Il faut une femme de 15 à 30 ans à Paris qui soit agricultrice
- Quid de ceux qui n'ont jamais répondu ?
- Pire : l'utilisation des *access panel*, c'est à dire des panélistes auto-recrutés en ligne, perturbe encore plus ces étapes.

$\Rightarrow$ *La singulière fabrique des sondages d'opinion*, M. Lejeune.

Méthodes RDS

Les méthodes RDS (Respondant-Driven Sample), parfois appelés boule de neige, reposent sur la logique suivante :

- On tire ou sélectionne un premier échantillon de petite taille ;

Méthodes RDS

Les méthodes RDS (Respondant-Driven Sample), parfois appelés boule de neige, reposent sur la logique suivante :

- On tire ou sélectionne un premier échantillon de petite taille ;
- On enquête ces individus ; chacun d'entre eux fournit un ou plusieurs contacts appartenant à la population-cible ;

Méthodes RDS

Les méthodes RDS (Respondant-Driven Sample), parfois appelés boule de neige, reposent sur la logique suivante :

- On tire ou sélectionne un premier échantillon de petite taille ;
- On enquête ces individus ; chacun d'entre eux fournit un ou plusieurs contacts appartenant à la population-cible ;
- On va enquêter ces nouveaux individus ;

Méthodes RDS

Les méthodes RDS (Respondant-Driven Sample), parfois appelés boule de neige, reposent sur la logique suivante :

- On tire ou sélectionne un premier échantillon de petite taille ;
- On enquête ces individus ; chacun d'entre eux fournit un ou plusieurs contacts appartenant à la population-cible ;
- On va enquêter ces nouveaux individus ;
- On répète ces étapes jusqu'à avoir obtenu un échantillon suffisamment grand ;

Méthodes RDS

Les méthodes RDS (Respondant-Driven Sample), parfois appelés boule de neige, reposent sur la logique suivante :

- On tire ou sélectionne un premier échantillon de petite taille ;
- On enquête ces individus ; chacun d'entre eux fournit un ou plusieurs contacts appartenant à la population-cible ;
- On va enquêter ces nouveaux individus ;
- On répète ces étapes jusqu'à avoir obtenu un échantillon suffisamment grand ;

Ces méthodes sont utilisées à l'Ined pour cibler des populations rares (HSH, usagers de drogues, minorités, etc.).

La méthode des itinéraires

La méthode des itinéraires permet de tirer "aléatoirement" un logement à enquêter au sein d'une commune lorsque l'on ne dispose pas d'une liste des logements. Elle fonctionne de la façon suivante :

- On tire au hasard un point GPS ;
- L'enquêteur s'y rend : il suit ensuite une directive fixée à l'avance, par exemple tourner à droite à la deuxième rue, et enquêter un logement sur 15 ;

Cette méthode est utilisée pour des enquêtes sociales comme par exemple l'ESS (Enquête Sociale Européenne) dans certains pays.