Introduction aux méthodes de sondage

Myriam Maumy-Bertrand[1]

[1] IRMA, Université de Strasbourg
Strasbourg, France

Master 2ème Année 05-10-2011

Ce chapitre s'appuie essentiellement sur deux livres :

- « Éléments de statistiques »,
  de Jean-Jacques Droesbeke,
  Université de Bruxelles, 2001.
- « Les techniques de sondage »
  de Pascal Ardilly,
  éditions Technip, 2006.

Sommaire

Définition

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

Définition

Sondage : *Méthode qui permet de construire un échantillon et qui prélève qu'une partie des informations existantes.*

Le **recensement** semble a priori plus naturel. Ne sommes-nous pas sûrs d'avoir plus de précision et un résultat de meilleure qualité en relevant toutes les informations possibles ?

Historique

- **17ème siècle :** Changement d'idée ! L'école d' « arithmétique politique » anglaise et l'apparition des premiers résultats de la théorie des probabilités suscitent de l'intérêt pour les relevés partiels.
- **Fin du 19ème siècle :** Le concept de **représentativité** d'un échantillon est discuté.
- **Après 1925 :** La question qui se pose naturellement est : « Comment choisir un échantillon ? »

Une polémique commence.

Définition

L'échantillonnage probabiliste : *Attribuer à chaque individu les mêmes « chances » d'être choisi.*

Opposition entre :

- les tenants de l'**échantillonnage probabiliste** ;
- et ceux dont l'objectif est plus pragmatique : Comment obtenir, en un temps minimal et au moindre coût, un échantillon dont les qualités soient suffisantes pour « représenter » une population ?

- Comme dans toutes les réalisations humaines, ces deux solutions présentent des qualités mais aussi des défauts.
- Une démarche scientifique, soucieuse de passer correctement du partiel à l'exhaustif, plaide généralement pour l'**échantillonnage probabiliste**.
- Nous essayerons d'en donner les raisons dans ce cours.
- Si le **recensement** se révèle théoriquement meilleur qu'un **sondage** parcequ'il est exhaustif, il n'en est pas moins vrai que le **sondage** se voit souvent préféré.

Pourquoi le **sondage** se voit souvent préféré à un recensement ?

Parce qu'il est :

- plus facile à obtenir,
- d'un coût moins élevé.

Autres avantages

Le **sondage** est souvent confié à une équipe réduite et spécialisée. Ainsi nous évitons des erreurs dues :

- au caractère inexpérimenté des personnes chargées de l'enquête
- aux difficultés d'être exhaustif.

Si un échantillon probabiliste est synonyme d'une rigueur scientifique plus grande, les **sondages** dits **judicieux** comme

- les sondages par choix raisonné,
- les sondages selon la **méthode des quotas**, qui sera présentée dans ce cours

sont parfois applicables là où les premiers ont échoués.

- Ne condamnons pas systématiquement les enquêtes effectuées par les organismes privés qui recourent souvent à la **méthode des quotas**.
- Comme nous venons de le dire, la **méthode des quotas** constitue parfois la seule façon d'agir pratiquement. Elle peut être même justifiée sous des hypothèses de travail bien spécifiques.
- Le **caractère empirique des méthodes** implique que ces sondages doivent être organisés par des personnes compétentes et intègres. Leur interprétation doit se faire avec circonspection.

Exemple

Les « sondages à chaud » effectués au cours des émissions télévisées pour mesurer le degré de persuasion de l'orateur politique invité.

Exemple

Les « sondages » obtenus par certains journaux qui sollicitent des réponses de leurs lecteurs.

Nous ne pouvons en aucun cas imaginer que ces relevés partiels et partiaux traduisent l'opinion de toute une population !

Dans ce cours, nous trouverons :

une présentation des méthodes qui proposent, d'un point de vue statistique, des propriétés qui permettent d'atteindre un objectif précis :
Comment, à partir d'une information partielle, pouvons-nous obtenir certaines conclusions au niveau de la population toute entière ?

- Si nous nous intéressons à certains paramètres de la population, il est **IMPOSSIBLE** d'en connaître la valeur exacte sur la base de l'information fournie par l'échantillon.

- Néanmoins, l'échantillon peut donner une **estimation** de ces paramètres, comme nous allons le voir dès le chapitre deux.

Sommaire

Situation générale :

Population U composée de N individus ou éléments appelés **unités statistiques**. N est la **taille** de la population U, supposée finie.

Exemples de population :

L'ensemble des touristes d'un pays, l'ensemble des ménages d'un pays, la production de pièces mécaniques d'une usine...

Nous pouvons dresser une liste exhaustive des éléments de la population U, appelée **base de sondage** où chaque élément est représenté soit par son nom soit par un numéro compris entre 1 et N.

Objectif :

Relever auprès de chaque individu de la population U, à l'aide d'un questionnaire ou d'un autre moyen de collecte (que connaissez-vous comme autre moyen ?), la valeur d'une ou de plusieurs variables d'intérêt.

Soit une variable Y, appelée **variable d'intérêt**, dont les valeurs associées à chaque unité de sondage sont notées $y_1, \ldots, y_N$.

Retour aux trois exemples

- $U =$ touristes $\Rightarrow$ $Y =$budget dépensé
- $U =$ ménages $\Rightarrow$ $Y =$ revenu du ménage
- $U =$ pièces produites $\Rightarrow$ $Y =$ caractère défectueux ou non de la pièce.

Remarque

Les deux premières variables sont **quantitatives**, la dernière ne l'est pas ! Quelle est donc sa nature ?

Remarque

La dernière variable est **dichotomique**.

Dans le cas de notre dernier exemple, pour représenter cette variable, nous la quantifions au moyen d'une **fonction indicatrice** selon laquelle la valeur Y prise par la pièce est égale à 1 si la pièce est défectueuse, 0 sinon.

Remarque

Ce processus se généralise dès que nous avons à faire à une **variable dichotomique** !

Notations

Afin de ne pas alourdir la terminologie : deux simplifications.

1. Représenter chaque individu par son indice k et non par u_k.
   Notation : $k \in U$.
2. Assimiler les individus interrogés par le questionnaire et leurs caractéristiques.

Définition

*Nous utilisons le terme « **population** » pour désigner*

- *aussi bien l'ensemble des individus susceptibles d'être soumis à l'étude*
- *que la distribution des valeurs d'une variable associée à ces individus.*

Remarque

Pour simplifier la présentation de ce cours, la variable étudiée sera **quantitative**.

Dans ce cas, une population est résumée **partiellement** par des paramètres comme :

- la moyenne $\mu = \frac{1}{N} \sum_{k \in U} y_k$,
- le total $T = \sum_{k \in U} y_k = N\mu$,
- la variance $\sigma^2 = \frac{1}{N} \sum_{k \in U} (y_k - \mu)^2$.

- Par exemple, l'objectif d'une étude peut consister à s'interroger sur leur valeur.
- Leur valeur exacte ne peut être connue que si nous disposons de toutes les valeurs de la population.
- Dans le cas contraire, nous procédons à une estimation présentée au moyen d'un exemple.

Nous nous intéressons à une **population de tailles** (en centimètres) dont la **moyenne** μ est **inconnue**.

Si nous prélevons un échantillon, il semble opportun de calculer sa **moyenne arithmétique** avec l'espoir que cette dernière nous permette d'estimer la valeur de μ.

Remarque

En général, la **moyenne arithmétique** de l'échantillon sera distincte de la moyenne μ de la population.

Définition

Nous disons que **la première est une estimation de la seconde**, *notée* $\widehat{\mu}$.

Définition

L'erreur que nous commettons en remplaçant μ *par* $\widehat{\mu}$ *résulte du fait qu'une partie de la population a été omise, appelée* **erreur d'échantillonnage.**

Deux questions

Deux questions se posent naturellement :

1. Comment choisir un **estimateur d'un paramètre** à partir des valeurs fournies par l'échantillon ?
   Nous venons d'évoquer l'estimateur pour la moyenne, mais d'autres cas sont envisageables comme un estimateur pour le total, pour la variance, pour la médiane...
2. Est-il possible de quantifier l'**erreur d'échantillonnage** afin de se rendre compte de son importance en fonction de la connaissance que nous avons sur la population et l'échantillon ?

- Il est évident que les réponses à ces questions dépendent de la manière de prélever l'information.
- Et c'est ce sujet que nous allons aborder dans le paragraphe suivant !

Sommaire

Définition

échantillon : *Il est composé de n individus, où n varie de* 1 *à N. L'échantillon est dit alors de* **taille** *n. Le prélèvement des individus est, en général, de telle sorte que les individus soient distincts des uns des autres.*

Remarque

Dans ce cas, l'échantillon $\mathcal{S}$ est un sous-ensemble de la population U (U comme univers). Nous noterons : $\mathcal{S} \subset U$.

Remarque

$\mathcal{S}$ est utilisé pour rappeler que nous parlons d'un échantillon car en langue anglaise, un échantillon se dit « sample ».

- **échantillon :** des variables Y_i pour lesquelles $i \in \mathcal{S}$.
- **Cas étudié ici :** une seule variable d'intérêt notée Y.
- **échantillon** $\{Y_i, i \in \mathcal{S}\}$ **:** une série statistique dont nous pouvons déterminer des paramètres comme la moyenne $\widehat{\mu}$ et la variance $\widehat{\sigma^2}$:

$$\widehat{\mu} = \frac{1}{n}\sum_{i\in\mathcal{S}} Y_i, \quad \widehat{\sigma^2} = \frac{1}{n}\sum_{i\in\mathcal{S}}(Y_i - \widehat{\mu})^2.$$

Définition

Taux de sondage :

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

avec n : taille de l'échantillon et N : taille de la population.

- Ainsi, compte tenu du paragraphe précédent, les valeurs de $\widehat{\mu}$ et de $\widehat{\sigma^2}$ sont des **estimateurs (et non des estimations) des paramètres** correspondants de μ et de σ^2 (définis dans le slide 21) de la population U.

- Maintenant, notre attention va porter sur leurs **qualités** et la possibilité de mesurer l'**erreur d'échantillonnage** (définie dans le slide 24). Dans ce but, nous allons préciser dans quelles circonstances cet objectif est atteint.

Sommaire

Définition

Méthode de sondage ou de tirage : *Définir la façon dont nous devons prélever des individus (des valeurs) dans une population afin de constituer un échantillon.*

Remarque

Nous distinguons deux grandes catégories de sondages :

- les sondages probabilistes ;
- les sondages empiriques.

Définition

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* **probabilités d'inclusion** *ou* **probabilité de sélection**. *Nous parlons alors d'***échantillon aléatoire** *ou d'***échantillon probabiliste**.

Définition

Les **sondages empiriques** *sont ceux qui ne permettent pas de calculer la probabilité d'inclusion des individus. Comme nous l'avons déjà dit il s'agit principalement des méthodes de quotas ou encore de la méthode d'unités-types.*

Définition

Un **plan de sondage** *est l'association des deux éléments suivants :*

- *la* **méthode de tirage**
- *et l'***expression de l'estimateur***.*

Remarque

Pour une méthode de tirage donnée, il existe de nombreux estimateurs concurrents.
Réciproquement, un estimateur donné peut être appliqué à des échantillons sélectionnés selon des méthodes de tirage différentes.

- Sélectionner un échantillon avec une méthode de tirage aléatoire équivaut à doter la population d'une **distribution de probabilités** qui définit un certain nombre de propriétés des valeurs retenues dans l'échantillon ou des paramètres construits à partir de ces derniers.

- Ainsi, si nous observons une valeur dans la population, la donnée observée y est considérée comme la valeur d'une variable aléatoire Y admettant pour loi la distribution de probabilités de la population.

Individus sélectionnés avec la même probabilité
$\Rightarrow$ **Sondage Aléatoire à Probabilités Égales** (SAPE).

Individus sélectionnés avec des probabilités différentes les unes des autres
$\Rightarrow$ **Sondage Aléatoire à Probabilités Inégales** (SAPI).

Pour chaque méthode de tirage, deux manières de réaliser un tel sondage selon que nous effectuons un prélèvement **avec** ou **sans** remise, ce qui fait déjà **quatre méthodes de tirage aléatoire**.

Sommaire

Dans ce cours, nous nous intéresserons qu'à l'estimation d'un paramètre initial θ qui a la forme d'une moyenne, d'une proportion ou d'un total.
Il est à noter qu'une proportion est une moyenne particulière. μ et T sont tous deux des sommes pondérées des valeurs Y_i de tous les individus de la population. Nous disons qu'il s'agit de paramètres **fonction linéaire** des Y_i.
Par conséquent les estimateurs utilisés seront construits sur le même modèle que le paramètre qu'ils cherchent à estimer.
Nous chercherons donc des estimateurs linéaires, qui font intervenir non seulement les Y_i des individus i de l'échantillons mais aussi un **système de poids** par individu i, notés $W_i(S)$, où $\mathcal{S}$ désigne l'échantillon de taille n.

Ainsi, l'estimateur $\widehat{\theta}$ est de la forme :

$$\widehat{\theta} = \sum_{i \in \mathcal{S}} W_i(\mathcal{S}) Y_i.$$

$W_i(\mathcal{S})$ est le **poids de sondage** (W comme weight en anglais). Toutes nos préoccupations tourneront autour du choix de ces poids.
Le problème du sondage est un **problème de pondération** : chaque individu de l'échantillon représente un certain nombre d'individus de la population et c'est pour cette raison qu'il doit **être dilaté** de $W_i(\mathcal{S})$.

Dans la suite de ce cours, sauf mention contraire, nous supposerons que la **taille de la population, notée N, est connue**.
Si ce n'est pas le cas, nous serions amenés à produire un **estimateur sans biais de N** pour pouvoir estimer des moyennes dans le cas de plans à probabilités d'inclusion quelconque.
Dans le cas des plans à probabilités d'inclusion égales et de taille fixe, il est nécessaire de connaître N pour estimer sans biais un total.