Objectifs et notations

Myriam Maumy-Bertrand[1]

[1] IRMA, Université de Strasbourg
Strasbourg, France

Master 2ème Année 05-11-2010

Sommaire

Objectif des méthodes de sondage

Les méthodes de sondage ont pour objectif de

tirer, dans une population, des échantillons destinés à estimer avec la meilleure précision des paramètres d'intérêt.

Le **tirage équiprobable avec remise** qui conduit à des échantillons de v.a.i.i.d. est la base de la statistique que vous avez étudiée jusqu'alors et également le modèle de la statistique mathématique.

Malheureusement pour vous, ce tirage ne correspond absolument pas à la pratique et n'est au mieux qu'une approximation commode pour le praticien !

En réalité...

Les sondages réels portent sur des populations finies et sont effectués par **tirage sans remise**, pour ne pas interroger deux fois le même individu. Les échantillons ne sont donc pas constitués de v.a.i. et de plus, le **tirage ne se fait pas toujours avec les mêmes probabilités** ! ! !

Le but de ce cours

Ce cours a pour objectif de donner une initiation à la **théorie des sondages aléatoires** et ne prétend pas couvrir le sujet.

Les erreurs de l'échantillonnage

En particulier, il faut savoir que les erreurs dues à l'échantillonnage ne sont qu'une partie de l'erreur globale qui comprend les erreurs de mesure, de non-réponse, etc.

Il existe des méthodes non-aléatoires

Bien des sondages sont effectués avec des méthodes non-aléatoires comme la **méthode des quotas** qui ne sera pas abordée ici. Pour de plus amples renseignement sur ce sujet, nous renvoyons au livre de Pascal Ardilly, *Les techniques de sondage*, Editions technip, 2006.

Sommaire

Notations qui concernent la population

- Nous notons U la **population** (U comme Univers).
- Nous notons N la **taille de la population**. N est supposée connue, ce qui n'est pas toujours vrai...Nous verrons comment estimer N dans un TD !
  La population est aussi appelée **base de sondage**.
- Chaque **individu de la population** est noté i.

Variable d'intérêt

- Nous notons Y la variable d'intérêt dont les valeurs sont $Y_1, Y_2, \ldots, Y_N$.
- La variable Y n'est pas une v.a. !
- Nous supposons que Y_i est obtenue sans erreur si l'individu i est sélectionné.
- Nous ne traiterons pas dans le cadre de ce cours, que le cas où Y est une variable unidimensionnelle numérique, éventuellement binaire lorsqu'il s'agira d'estimer une proportion.

Moyenne et total

Nous nous intéresserons à l'estimation de quantités dépendant de Y comme la moyenne de Y de la population ou le total des valeurs T_Y noté T quand il n'y aura pas d'ambiguïté qui sont définis de la façon suivante par :

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

et

$$T = \sum_{i=1}^{N} Y_i = N\overline{Y}.$$

Variance et variance corrigée

Nous notons σ^2 la **variance de** Y et nous la définissons par la relation suivante

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}\left(Y_i - \overline{Y}\right)^2$$

et nous notons S^2 la **variance corrigée de** Y et nous la définissons par la relation suivante

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

Il peut paraître curieux d'utiliser la variance corrigée quand il ne s'agit pas d'un échantillon mais cela conduit à des formules plus simples.

Échantillon

Un **échantillon** est un sous-ensemble de n unités de la population. Il y a C_N^n échantillons distincts possibles notés $\mathcal{S}$.

Remarque

Il faut faire attention à l'utilisation du mot « échantillon ». Dans la théorie des sondages, il faut se demander si il s'agit d'un échantillon avec remise ou d'un échantillon sans remise. De plus, il faut aussi se demander si il s'agit d'un échantillon aléatoire (nous allons définir par la suite cette notion) ou non.

Taux de sondage

$f = \frac{n}{N}$ est appelé le **taux de sondage**.

Plan de sondage

Un **plan de sondage**, noté $p(\cdot)$, est une loi de probabilité sur l'ensemble de tous les échantillons possibles telle que

$$p(s) \geqslant 0, \text{pour tout } s \subset U \text{ et } \sum_{s \subset U} p(s) = 1.$$

Maintenant nous pouvons définir l'échantillon aléatoire.

Échantillon aléatoire

L'**échantillon aléatoire** $\mathcal{S}$ est un ensemble aléatoire d'étiquettes dont la distribution de probabilité est

$$\mathbb{P}(\mathcal{S} = s) = p(s), \text{pour tout } s \subset U.$$

Probabilité d'inclusion (d'ordre 1)

Dans un sondage aléatoire, chaque unité i de la population a une probabilité de tirage ou **probabilité d'inclusion** notée π_i bien définie qui ne doit pas être nulle sous peine de ne pouvoir faire des estimations sans biais.

Remarque

Notons que π_i est égale à la somme des probabilités des échantillons qui contiennent l'unité i

$$\pi_i = \sum_{\mathcal{S}(i \in \mathcal{S})} p(s)$$

Probabilités d'inclusion d'ordre 2

Une **probabilité d'inclusion d'ordre 2**, notée π_{ij} donne la probabilité que les unités i et j appartiennent à l'échantillon.

Propriétés

Si le plan est de taille fixe n, nous avons que

1. $$\sum_{i=1}^{N} \pi_i = n$$
2. $$\sum_{k=1}^{N} \pi_{k,l} = n\pi_l.$$

Variable de Cornfield

Une **variable de Cornfield** correspond à une indicatrice, notée δ_i, qui indique la sélection de l'unité *i* de la population dans l'échantillon.

Propriétés sur les variables de Cornfield

1. $$\mathbb{E}(\delta_i) = \pi_i$$
2. $$\mathbb{V}\text{ar}(\delta_i) = \pi_i(1 - \pi_i)$$
3. $$\mathbb{C}\text{ov}(\delta_k; \delta_l) = \pi_{kl} - \pi_k\pi_l \quad \text{si } k \neq l.$$
4. $$\sum_{k=1}^{N} \mathbb{C}\text{ov}(\delta_k; \delta_l) = 0.$$