Théorie des sondages

Semestre 6

Table des matières

Première partie

1

Une fois l'échantillon sélectioné on dispose de l'information suivante :
$Y_{i1}, Y_{i2}, \dots Y_{in}$. Cherchons toujours θ, il convient de combiner ces petites n valeurs recuillies sur l'échantillon pour obtenir une expression dont la valeur numérique est proche de celle de θ. La formule qui aggrège les petis n valeurs s'appelle l'estimateur de θ, on le note : $\hat{\theta}$ il s'agit d'une fonction calculée a partir des données de l'échantillon, et la procédure qui permet de sauter le pas, c'est à dire de passer des odnnées recueillies sur les échantillons à la vraie valeur

inconnue dans la population s'appelle l'inférence statistique (coeur même de la théorie des sondages).

2 Les mesures des erreurs d'échantillonage

Une fois que la fonction qui agrège $\hat{\theta} = g(Y_{i1}, Y_{i2}, \dots Y_{in})$ a été choisie il est nésessaire d'evaluer sa pertinence en d'autre termes il convient de répondre à la question suivante : Sommes nous proche de la valeur θ lorsque l'on calcule g a partir de l'échantillon selectionné ?
En fait il n'est pas possible de répondre précisément à cette question pour la simple et bonne raison que l'on ne connait pas θ, mais la statistique permet d'apporter des éléments de réponse en exploitant l'aspect probabiliste des choses. Un des points fondamentaux qu'il faut a présent bien comprendre, réside dans la nature de l'aléa introduit dans notre problème, l'aléa se situe exclusivement au niveau des identifiants des individus de l'échantillon.

Ce qui est aléatoire sont les $i1, i2, \dots in$ et non les $Y_{i1}, Y_{i2}, \dots Y_{in}$, ainsi si l'on réalise une enquete sur les revenus, le sondeur considère dans son enquete que chaque individu dans la population est capable de fournir son revenu, en revanche l'échantillon s d'individus interrogés sur leur revenu aura une compisition aléatoire. Notre estimateur g (ou $\hat{\theta}$) est donc aléatoire fonction de l'échantillon s. La réponse à notre question initiale fait appel à trois notions : le biais, la variance, et l'erreur quadratique moyenne.

2.1 Le biais

Le biais permet de détecter la présence éventuelle d'erreur systématique, il correspond donc à la différence entre l'espérance mathématique de l'estimateur et le paramètre lui même.

$$\text{Biais}(\hat{\theta}) = \text{E}(\hat{\theta}) - \theta$$

Le biais constitue une premiere mesure d'erreur d'échantillonage.

2.2 La variance

On présent bien au travers de l'exemple précédent que la notion de moyenne ne suffit pas à mesurer la qualité d'un échantillonage et qu'il faut une autre grandeur d'avantage liée à la dispersion des valeurs de $\hat{\theta}$. On calcule la variance des estimateurs $\hat{\theta}_s$ lorsque l'aléa est l'échantillon, s.

$$\begin{aligned}
\text{V}(\hat{\theta}) &= \text{E}\left[\left(\hat{\theta} - \text{E}(\hat{\theta})\right)^2\right] \\
&= \sum_{i=1}^{s} P(i)\left[\left(\hat{\theta} - \text{E}(\hat{\theta})\right)^2\right]
\end{aligned}$$

La présence du terme quadratique fait toute la différence avec le biais puisque les écarts positifs et négatifs ne se compensent plus. En terme de sondage $\sigma_{\hat{\theta}}$ et $\text{V}(\hat{\theta})$ mesurent la précision et constituent après le biais une seconde mesure de l'erreur d'échantillonage, plus ils sont grands, moins le plan de sondage est bon.

Si $\sigma_{\hat{\theta}}$ et $\mathrm{V}(\hat{\theta})$ sont grands il faut agir sur l'expression de $\hat{\theta}$ soit sur les probabilités de tirage ($P(s)$). Le meilleur estimateur $\hat{\theta}$ est celui qui a la plus petite variance compte tenu du budget dont on dispose.
Pour terminer, il faut insister une fois encore sur le fait que ni l'espérance, ni la variance ne peuvent renseigner sur l'écart exact entre la valeur de $\hat{\theta}$ obtenue et la vrai valeur de θ inconnue.

2.3 L'erreur quadratique moyennei (EQM) (MSE)

On peut également construire un indicateur de précision qui englobe les notions de biais et de variance pour se faire, il suffit de calculer l'erreur quadratique moyenne tel que :

$$\mathrm{EQM} = \mathrm{E}\left[\hat{\theta} - \theta\right]^2$$

Cette erreur quadratique moyenne est un indicateur synthétique de qualité permettant de répondre à la question complexe suivante : est-il préférable d'avoir un biais fort et une variance faible ou un biais faible et une variance forte ?
Il n'y a evidement pas de bonne réponse à cette question, mais la pratique montre que l'on cherche en général à éviter en priorité les biais forts.

Pour conclure, l'erreur d'échantillonage est donc le fait que les résultats numériques publiés à la suite d'un sondage dépendent des individus qui composent l'échantillon. Elle est présente, et même une caractéristique des enquetes par sondage. Cette erreur d'échantillonage se mesure par le biais, la variance ou l'erreur quadratique moyenne, elle ne peut pas être évitée.

Exemple 1. Considérons une population composée de 4 entreprises $N = 4$ et l'on s'interesse au chiffre d'affaire mensuel moyen de cette population $\overline{\text{CA}}$ mensuel. Le CA mensuel des entreprises est le suivant :

$$\text{CA}_1 = 6000€ \quad \text{CA}_2 = 12000€ \quad \text{CA}_3 = 8000€ \quad \text{CA}_4 = 6000€$$

Supposons que pour raison de contrainte de budget, je ne peux interroger que 2 entreprises parmis les 4, $n = 2$.

$$\theta = \frac{CA_1 + CA_2 + CA_3 + CA_4}{N} = 8000$$

Nombre d'échantillons possibles :

$$C_4^2 = \frac{4!}{2!(4-2)!} = 6$$

Ces 6 échantillons sont les suivants :

$$\begin{array}{ll} s_1 = (1,2) & P(s_1) = 0.25 \\ s_2 = (1,3) & P(s_2) = 0.25 \\ s_3 = (1,4) & P(s_3) = 0.2 \\ s_4 = (2,3) & P(s_4) = 0.1 \\ s_5 = (2,4) & P(s_5) = 0.1 \\ s_6 = (3,4) & P(s_6) = 0.1 \end{array}$$

Parce que l'on juge que l'entreprise 1 est particulièrement coopérative sur ce sujet, on veut lui donner une probabilité de tirage supérieure aux trois autres, si bien que les trois échantillons s_1, s_2, s_3 sont un peu plus probables que les autres.
Et par soucis de simplicité on choisi comme estimateur la moyenne simple dans l'échantillon. Autrement dit, on suppose que : $\hat{\overline{\text{CA}}} = \overline{y}$ La moyenne simple calculée pour les entreprises des échantillons.

$$\begin{array}{ll} s_1 & \overline{\text{CA}}(s_1) = \frac{6000 + 12000}{2} = 9000 \\ s_2 & \overline{\text{CA}}(s_2) = \frac{6000 + 8000}{2} = 7000 \\ s_3 & \overline{\text{CA}}(s_3) = \frac{6000 + 6000}{2} = 6000 \\ s_4 & \overline{\text{CA}}(s_4) = \frac{12000 + 8000}{2} = 10000 \\ s_5 & \overline{\text{CA}}(s_5) = \frac{12000 + 6000}{2} = 9000 \\ s_6 & \overline{\text{CA}}(s_6) = \frac{8000 + 6000}{2} = 7000 \end{array}$$

Calculons le biais : $\theta = \overline{\text{CA}} = 8000$

$$\begin{aligned} \text{E}(\theta) &= \sum_{s=1}^{6} P(s) \cdot \hat{\theta}(s) \\ &= 0.25 * 9000 + 0.25 * 7000 + 0.2 * 6000 + 0.1 * 10000 + 0.1 * 9000 + 0.1 * 7000 = 7800 \end{aligned}$$

En moyenne l'estimateur est de 7800, on peut donc calculer le biais :

$$\text{Biais} = 7800 - 8000 = -200$$

Variance :

$$0.25(9000 - 7800)^2 = 360000$$

3 Les bases de sondage

3.1 Les propriétés des bases de sondages

Pour pouvoir réaliser un tirage probabilisite, c'est à dire un tirage pour lequel par définition chaque individu de la population a une probabilité connue et fixée à l'avance de faire partie de l'échantillon à enqueter, il est absolument nécessaire de disposer d'une liste de toutes les unités d'échantillonage faisant partie du champs de l'enquete, cette liste est appellée base de sondage, doit avoir trois qualités essentielles.

1. Elle doit permetre de reperer l'unité sans aucune ambiguité, une bonne base de sondage est une bonne base d'identifiant.
2. Elle doit être exhaustive, cela signifie que chaque unité faisant partie du champ de l'enquete doit etre necessairement présente dans la liste des identifiants. Si ce n'est pas le cas, on parle de base de sondage incomplete ou de défaut de couverture.
3. Elle doit être sans double compte, aucun individu ne doit être présent deux fois dans la base de sondage, même sous deux idenfiants différents.

De facon générale il est extremement difficile en pratique de s'affranchir du manque d'exhaustivité (la plupart des bases de sondage est incomplete) et de la présence de double compte. L'important est de juger de leur impact et de ne conserver que les bases faiblement imparfaites. A ces trois propriéés on en rajoute souvent une quatrième qui sans être indispensable, peut s'averer très bénéfique, c'est la suivante : une information auxilière de bonne qualité (information supplémentaire dont on va pouvoir se servir pour améliorer soit la qualité des estimateurs ou la méthode de tirage).En fait on appelle information auxilière, toute var quantitatve ou qualitative autre que la variable d'interet Y et autre que les variables necessaires et suffisantes à l'identification des individus de la population.

3.2 Abscence de base de sondage

Lorsque l'on s'interesse a une population, il est tout a fait possible que on ne puisse pas trouver de base de ce sondage reproduisant cette population ou qu'une telle base existe mais soit jugée de mauvaise qualité, il peut également arriver que l'on renonce à l'utilisation de base e bonne qualité pour des raisons d'ordre pratique, parmis ces raisons :

- La base existe mais ne peut pas etre donnée
- La base est trop volumineuse et le matériel informatique dont on dispose ne permet pas son traitement, cet obstacle ne doit pas etre négligé, car on oublie trop souvent que si les échantillons sont de petite taille, les bases elles sont souvent très grande.

Si pour une raison quelquonque on ne dispose pas de base de sondage acceptable permetant de réaliser un échantillo,nage de bonne qualité, trois solutions sont envisageables :

1. Ne pas utiliser dutout de base de sondage, sortir du cadre probabiliste, réaliser un sondage empirique.
2. On peut rechercher des bases de sondage, non plus d'individus irectement succeptibles de fournir l'information, mais accepter de passer par

un niveau intermédiaire de groupe d'individus, par un tirage à plusieurs degrès, et en réalisant des recenssements intermédiares dans les groupes sélectionés, on peut échantilloner rigoureusement des unités de la population qui nous intéresse. Cette façon de faire conduit fréquement sur ce que l'on appelle les sondages de type aréolaire, pour lequels on échantillonne en premier lieu des aires géographiques.

3. Si le sujet auquel on s'interesse s'y prete, on peut reccueillir l'information en passant par une population intermédiaire, qui conduit aux unités d'observation, population intermédaire d'une autre nature que celle des individus de la population sur laquelle portera l'inférence.

3.3 Les différents types de bases

Celon letype d'enquetes que l'on est amené a réaliser, on peut opter pour une base de liste, c'est à dire une base constitué par une liste d'indentifiants d'indivius, ou pour une base aréolaire, constituée par des aires géographiques, bien délimités et bien identifiés, la base aréolaire permet d'effectuer des sondages en grappe, où une grappe est constituée par l'ensemble des individus de l'aire. On attribue trois avantages comparatifs principaux aux sondages aréolaires par rapport aux sondages réalisés à partir d'une base de liste.
Avantage des bases aréolaires

1. L'aire est une entité relativement stable qui permet de mieux prenre en compte l'évolution de la structure réelle de la population qu'elle contient.
2. Le repgroupement géographique est y par définition : maximal, ce qui permet de limiter les couts de déplacement
3. Les taux de réponse sont en général meilleur que dans les sondages par liste

4 Les différents types d'erreurs recontrées dans les enquetes par sonage

4.1 L'erreur d'échantillonage

cf sec 1

4.2 Les erreurs d'observation

Dans la pratique, il existe une seconde famille d'erreurs, que l'on appelle les erreurs d'observation ou les erreurs de mesures, qui tient au fait que la valeur que l'on reccueil lors de l'enquete peut etre une valeur Y_i^* différente de la vrai valeur Y_i. Ce type d'erreur survient dans les questions sensibles. Sur de tels sujets, l'erreur d'observation est volontairement introduite par l'enqueté, mais il existe parallèlement à cela une longue liste de source d'erreurs d'observations qui n'ont pas d'origine volontaire, parmis lesquelles on peut citer les erreurs de bonne foi de l'enqueté. Ensuite il y a les erreurs introduites par l'enqueteur qui interpete les questions et souffle potentiellement les réponses, il s'agit là d'un défaut de formation de l'enqueteur qui n'a pas a influencer la réponse de l'enqueté

4.3 Le défaut de couverture et les non réponses

On peut distinguer une troisième famille d'erreurs liée à l'existence d'une base de sondage incomplète et a la non réponse de certains individus aux questions posées. En effet, une base de sondage incomplète est une situation qui donne lieu à un défaut de couverture de la population, dans ce cas il y à dès l'origine un biais de l'estimateur que l'on ne peut pas mesurer. Il convient néanmoins de distinguer le défaut d'une base comprenant des doubles compte du défaut de couverture, dans le premier cas (a double compte), toute l'information est disponible dans le fichier, mais elle est mal utilisée alors que dans le défaut de couverture, l'information n'existe pas dutout. Concernant la non exhaustivité d'une base est un voisin de la non réponse dans la mesure où ces deux sources d'erreur sont dues a l'existence d'individus pour lequel on ne peut pas tirer d'information sur la valeur de Y. Néanmoins la non réponse se différencie sur deux points par rapport aux erreurs liées au défaut de couverture. L'empleur du phénomène de nom réponse est mesurable à partir de l'échantillon sélectioné alors que l'étendue d'un défaut de couverture ne l'est pas nécessairement. La taille de l'échantillon qui importe dans la mesure de l'échantillon avec une base incomplète et la taille totale de l'échantillon (taille qui peut être fixée à l'avance.) Alors que la taille de l'échantillon qui importe pour le calcul de la précision lorsqu'il y a non réponse est l'effectif des répondant, qui est aléatoire. Pour ce qui concerrne la non réponse on peut avoir a faire en une réponse partielle ou une non réponse complète. Finalement l'erreur totale que l'on connait

$$\begin{aligned}\text{Erreur totale} = &\boxed{\text{Erreur d'échantillonage}}\\ &+ \text{Erreur de mesure/observation}\\ &+ \text{Erreurs dues au défaut de couverture}\\ &+ \boxed{\text{Erreurs dues à la non réponse}}\end{aligned}$$

En général on a tendance naturellement a faire porter l'effort sur l'erreur d'échantillonage et sur l'erreur de non réponse .En général les défauts de modélisation et de mesure portentn sur l'erreur d'échantillonage et sur la non réponse, malheursement on ne sait que très peu de choses sur les erreurs d'observation et les défauts de couverture si ce n'est qu'ils sont de la nature de biais. Ce constat n'est donc pas très réjouissant car dans la mesure au contrairement aux errerus d'échantillonage, ces erreurs ne diminuent pas avec la taille de l'échantillon

Deuxième partie

Le sondage aléatoire simple

Il s'agit d'une méthode de tirage qui consiste à tirer dans la population de taille N un échantillon de taille fixée n sans remise à partir des seuls identifiants des individus de manière à ce que, chaque individu de la population ait la même probabilité d'inclusion et cela sans manipulation au préalable de la population. C'est exactement ce que l'on réalise lorsqu'on sélectione dans une urne des boules sans remise. Le sondage aléatoire simple attribue a chaque échantillon s qui peut etre formé, la même probabilité de sortie $p(s)$ qui est égale à l'inverse du

nombre d'échantillons distincts que l'on peut constituer dans la population, cette propriété remarquable le caractérise il ne nécéssiste pas non plus d'informations auxilières lors de sa mise en oeuvre.