Cerema

Direction territoriale Centre-Est

Eléments pour évaluer la précision obtenue dans l'estimation des indicateurs produits à partir de l'EMC2 de la Grande Région Grenobloise[1]

[1] Ce document s'inspire largement du document « Eléments pour évaluer la précision obtenue dans l'estimation des indicateurs donnés par les enquêtes ménages déplacements », Certu, déc. 2006

1 Introduction

1.1 A l'approche de la vérité : le principe de l'inférence statistique et ses outils

L'EMC2 de la Grande Région Grenobloise 2021 a été réalisée pour mesurer la mobilité quotidienne des habitants de ce territoire. Cet échantillon aléatoire de ménages a été interrogé pour estimer des indicateurs de mobilité ; les valeurs produites par cet unique échantillon ne sont donc que des estimations de « vraies valeurs » qu'on aurait trouvées si on avait interrogé tous les ménages de la zone d'enquête : un autre échantillon de même taille, et sélectionné aléatoirement de la même manière, pourrait produire pour un même indicateur une autre valeur puisqu'il a de fortes chances d'être composé de ménages et d'individus différents.

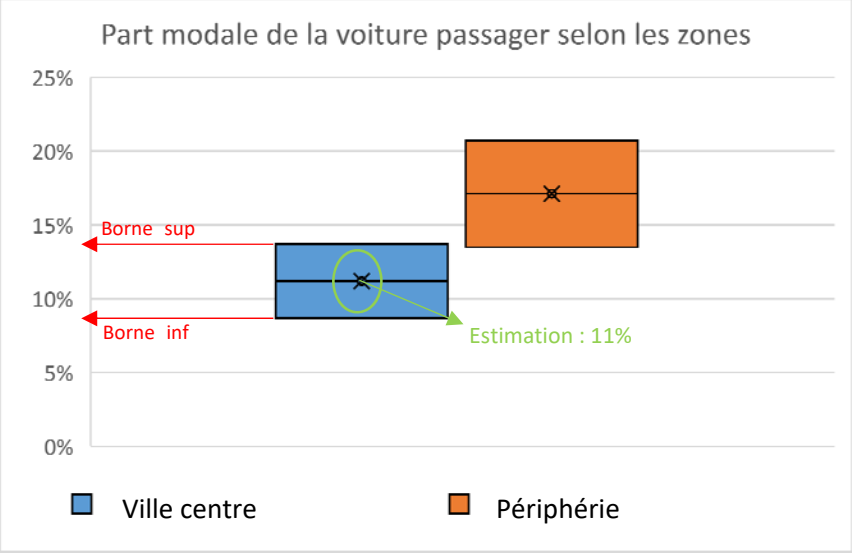
C'est pourquoi la théorie des sondages propose de calculer, en plus d'une estimation ponctuelle produite par un échantillon donné, un « intervalle de confiance » pour cette estimation. Cet intervalle prend en compte l'imprécision due à cette variabilité de composition de l'échantillon (dite « variance d'échantillonnage »), et uniquement celle-là. La théorie avance, sous certaines hypothèses, que si on tirait l'ensemble des échantillons possibles (de même taille), la « vraie valeur » aurait x % de chance d'être contenue dans cet intervalle. En pratique, on calcule souvent des intervalles de confiance à 95%. La théorie précise également que la « vraie valeur » a plus de chance se trouver aux alentours du milieu de l'intervalle, qu'à ses bornes : les valeurs extrêmes dues à une composition « particulière de l'échantillon » peuvent exister mais la probabilité qu'elles surviennent est très faible. On comprend bien que plus la taille de l'échantillon est grande, plus on a d'information et plus les résultats seront précis, c'est-à-dire que l'intervalle sera plus réduit autour de l'estimation. En pratique, lorsqu'on travaille sur des sous-populations, il devient très périlleux d'estimer des indicateurs avec des tailles d'échantillon inférieures à 30 car les hypothèses relatives à la loi des grands nombres ne s'appliquent plus correctement. D'autres éléments sont également à prendre en compte pour assurer la validité de la mesure, ils seront détaillés ensuite.

1.2 Mettre du flou…pour s'en débarrasser ensuite : l'utilité de l'intervalle de confiance

L'enquête produit une grande diversité d'indicateurs sur des sous populations variées (âge, territoire, type d'usager etc.) que l'on cherche souvent à comparer entre elles pour mieux comprendre les comportements de mobilité. Cependant, compte tenu des principes de l'inférence énoncés ci-dessus, on ne peut ignorer que deux estimations ponctuelles différentes, ne mettent pas forcément en évidence une différence « significative » de comportement.

Par exemple, considérons la part modale de la voiture-passager dans une ville centre qui est de 11%, contre 17% dans sa périphérie ; représentons les intervalles de confiance concernés, ces proportions sont-elles « significativement différentes » ?

Il est facile de comprendre que si les intervalles se recouvrent, cela signifie que certains échantillons (que nous n'avons pas sélectionnés) pourraient donner des estimations semblables, surtout si les intervalles se recouvrent totalement. Par-contre, si seules des extrémités se recouvrent en partie, ou a fortiori s'il n'existe aucun recouvrement, alors les deux parts modales seront considérées comme différentes.



Au vu des résultats, on peut donc en conclure que les déplacements s'effectuent plus souvent en voiture en tant que passager dans la périphérie, que dans la ville centre. De la même manière, on peut aussi estimer les volumes de déplacements concernés (et non plus les parts) et leur intervalle de confiance.

La forme de l'indicateur influe sur la manière de calculer l'intervalle de confiance : la méthode diffère selon qu'on estime une part (ou proportion), une moyenne ou encore un volume. Elle peut se compliquer énormément lorsque l'indicateur est plus complexe.

2 Application de la théorie des sondages pour calculer les intervalles de confiance des indicateurs produits par l'enquête

Notre méthode consiste à calculer « l'erreur relative » à x % et d'en déduire ensuite la taille du demi-intervalle de confiance entourant la valeur estimée. La démarche pas-à-pas est explicitée pour chaque grande famille d'indicateurs dans le document Excel joint à cette note.

2.1 Les proportions

Une proportion est une part d'une sous population dans un total par exemple la part des déplacements effectués en voiture dans l'ensemble des déplacements, la part des personnes ne s'étant pas déplacées un jour de semaine, ou encore la part des ménages détenant une seule voiture.

Dans le cas d'un échantillon aléatoire, l'erreur relative à x % d'une proportion p se calcule ainsi :

$P_{\text{err alea}}(p) = t \times \frac{\sqrt{p(1-p)}}{\sqrt{n-1}} / p$, (ex : **p** la part des déplacements effectués en voiture)

n est le nombre total d'unités à partir duquel est construit la part (ex : nombre de déplacements)
t dépend du seuil de confiance choisi pour l'intervalle de confiance : on le fixe ici à 95% donc t = 1,96.

Si notre échantillon de ménages est bien aléatoire, nos échantillons d'individus ou de déplacements le sont moins : les individus, tout comme les déplacements, ne sont pas choisis complètement aléatoirement dans la population puisque ce sont toutes les personnes âgés de + 5 ans du ménage sélectionné[2] qui sont invitées à répondre, et les déplacements découlent directement de ces individus. Pour cette raison, les estimations

[2] Au téléphone, seulement une ou deux personnes du ménage sélectionné

relatives aux personnes ou aux déplacements sont moins précises. On parle alors d'« effet-grappe » (**coeffG**). Cet effet augmenter l'imprécision de l'estimation, donc la taille de l'intervalle de confiance :

$P_{err\ grappe}(p) = coeffG \times P_{err\ alea} = coeffG \times 1{,}96 \times \frac{\sqrt{p(1-p)}}{\sqrt{n-1}} / p$

Le coefficient de grappe est différent selon qu'on s'intéresse aux personnes ou aux déplacements, il a été estimé à partir de plusieurs enquêtes ménages :

- au voisinage de 2 pour les déplacements
- au voisinage de 1,2 pour les personnes[3]
- il est de 1 pour les ménages, c'est-à-dire qu'il n'y a pas d'effet-grappe

Dans notre exemple, si on calcule l'erreur relative de la part modale voiture (p), on en déduit

$P_{err\ grappe}(p) = coeffG \times P_{err\ alea} = 2 \times 1{,}96 \times \frac{\sqrt{p(1-p)}}{\sqrt{n-1}} / p$

Enfin, on peut remarquer que la taille de l'intervalle de confiance dépend certes de la taille de l'échantillon comme dit précédemment, mais également du niveau de la proportion qu'on mesure : à taille d'échantillon donnée, plus on estime une sous-population rare (p petit), plus la taille de l'intervalle s'agrandit. Autrement dit, pour connaître une population rare avec une bonne précision, il est nécessaire d'augmenter la taille de l'échantillon, et donc de dépenser plus d'argent (« ce qui est rare, est cher »…).

2.2 Les moyennes

Ce type d'indicateurs recouvre par exemple la taille moyenne des ménages, le nombre moyen de véhicules par ménage ou taux de motorisation, ou encore le nombre moyen de déplacements par personne et par jour.

La précision relative d'une moyenne M (ex : nombre moyen de déplacements par personne) se calcule comme suit :

$P_{rel\ alea}(M) = t \times (\sigma/\sqrt{n}) / M$

t dépend du seuil de confiance choisi pour l'intervalle de confiance : on le fixe ici à 95 % donc t = 1,96.
σ est l'écart-type de la grandeur mesurée (ex : écart-type du nombre moyen de déplacements par personne)
n taille de l'échantillon concerné (ex : nombre de personnes).

Aucune formule mathématique ne permet de déduire précisément cette précision car il faut estimer la dispersion de la variable à étudier sur tous les échantillons possibles (σ), et donc tirer un grand nombre d'échantillons différents pour estimer cette variabilité. Comme dit précédemment, nous n'en sélectionnons qu'un seul.

[3] Dans le cas des enquêtes par téléphone, seuls 1 ou 2 individus par ménage sont interrogés donc l'effet grappe devrait être un peu moins fort (entre 1,05 et 1,1). Dans cette étude, tous les intervalles sont calculés avec les effets grappe du face-à-face, donc « au pire » pour les zones enquêtées par téléphone.

Le Cerema (ex-Certu à l'époque) a estimé par simulation pour différents paramètres la valeur de σ/M. Cette valeur est assez équivalente d'une enquête à l'autre, on s'appuie donc sur ces mesures empiriques pour estimer la précision relative des moyennes.

Indicateurs de type « moyenne »	**Population de référence (n)**	**σ/M**
Taille des ménage	ménages	0,60
Nombre d'actifs par ménage	ménages	0,85
Taux de motorisation	ménages	1,00
Taux d'occupation des VP - tous motifs	déplacements VP conducteurs	0,56
Mobilité – tous modes	personnes	0,80

Il faut également prendre en compte l'effet grappe (coeffG) pour les individus et les déplacements.

Soit pour notre exemple $P_{rel\ grappe}(M) = 1{,}2 \times (1{,}96 \times 0{,}8) \times (1 / \sqrt{n})$

Voici quelques ordres de grandeur :

		Nombre de ménages					
	Valeur indicateur	**15000**	**10000**	**5000**	**3000**	**2000**	**1000**
Taille moyenne des ménages	2,5	1,0%	1,2%	1,7%	2,1%	2,6%	3,7%
Nb de voitures par ménages	1	1,6%	2,0%	2,8%	3,6%	4,4%	6,2%

		Nombre de personnes					
	Valeur indicateur	**15000**	**10000**	**5000**	**3000**	**2000**	**1000**
Nb de dépalcements par personne	3,8	1,5%	1,9%	2,7%	3,4%	4,2%	6,0%

3 Aide à l'utilisation des calculs d'intervalle de confiance

Le fichier Excel joint propose un module de calcul simplifié pour chaque famille d'indicateurs. Ce module respecte les principes et valeurs exposés dans cette note.

Quel que soit le type d'indicateurs, nous avons calculé à chaque fois l'erreur relative, la valeur du « demi-intervalle » de confiance à ajouter/retrancher de la valeur estimée par l'échantillon :

IC = [estimation – demi intervalle ; estimation + demi intervalle]