Par Adam Baïz, économiste et enseignant (PhD Mines ParisTech)

En interrogeant 5000 personnes chaque semaine, l'enquête Datacovid permet d'ausculter les comportements et les perceptions des Français face à la crise actuelle. Mais 5000 personnes interrogées pour une population de 67 millions, n'est-ce pas aller un peu vite en besogne ?

En statistiques, il y a **deux façons d'apprécier l'état d'une population**. Soit on interroge un à un tous ses membres, à la façon d'un recensement, mais c'est très coûteux et parfois impossible à faire. Soit on interroge seulement un échantillon, en veillant bien sûr à ce qu'il soit représentatif. Pour viser cette représentativité, deux principales techniques existent.

La première consiste à tirer à l'aveugle un certain nombre d'individus dans la population générale : ce sondage dit *aléatoire* permet de reconstituer *grosso modo* toutes les caractéristiques visibles comme invisibles de la population. Ainsi si vous piochez au hasard 50 personnes parmi les 500 lycéens d'un établissement, il est fort à parier que cet échantillon aura *grosso modo* les mêmes caractéristiques visibles (comme la proportion de filles) et invisibles (comme la motivation à étudier) que la population totale des 500 lycéens. Et si l'on dit *grosso modo*, c'est parce qu'on pourrait accidentellement avoir quelques *epsilon* de différences entre les caractéristiques de l'échantillon et celles de la population : heureusement, les statistiques vous garantissent que ces *epsilon* se réduiront en augmentant la taille de votre échantillon.

La seconde technique consiste à réaliser ce que l'on appelle un sondage par strates, c'est-à-dire que l'on segmente la population en des catégories (ou *strates*) jugées homogènes, et on pioche les individus dans chaque catégorie, en tenant compte de leurs poids respectifs. Par exemple, si l'on sait que 60% des lycéens sont des filles, et qu'on estime que la distinction « fille/garçon » est pertinente pour analyser les choix post-bac, on pourrait tirer (aléatoirement) 30 filles dans le groupe des filles d'une part, et 20 garçons dans le groupe des garçons d'autre part. Ainsi la proportion de filles dans l'échantillon est-elle respectée de façon certaine. Cette méthode est privilégiée par les instituts de sondage, comme Ipsos pour Datacovid, puisqu'elle permet d'avoir des estimations robustes avec moins de personnes interrogées. Il reste que les catégories peuvent constituer un biais en soi : en effet, les choix post-bac pourraient être très hétérogènes entre filles (et entre garçons) et relever plutôt d'autres critères.

Mais passons, et admettons que nous avons notre échantillon représentatif. Que faire à présent ? En théorie, il suffit de faire des statistiques sur l'échantillon, et d'*inférer* qu'elles sont les mêmes sur la population générale, à une marge d'erreur près (cf. les

fameux intervalles de confiance). En pratique, un nouveau problème surgit : la non-réponse. Et plus exactement : la non-réponse totale, lorsqu'un individu échappe finalement à l'enquête (à cause d'un refus d'entrer dans le questionnaire, d'une absence lors du passage de l'enquêteur, etc.), et la non-réponse partielle, lorsque l'individu répond mais seulement à une partie des questions. **Dans un cas comme dans l'autre, il faut procéder à ce que l'on appelle un redressement.** C'est essentiel car autrement pourrait se produire un biais de sélection : par exemple, si on occultait les non-répondants à l'enquête sur les choix post-bac, on pourrait s'apercevoir que les répondants, en étant les élèves les plus motivés (à répondre à des sondages comme à travailler en général), biaisent les résultats.

D'une part, pour redresser la non-réponse partielle, on procède généralement à une imputation : on cherche des liens entre les caractéristiques des individus et les réponses obtenues, on teste leur robustesse (test du Chi-Deux, V de Cramer, etc.), et on impute alors aux réponses manquantes la modalité la plus répandue ou la plus vraisemblable. Par exemple, une lycéenne dont les parents sont cadres ne répond pas à la question « Travaillez-vous chaque soir chez vous ? », mais l'on observe que 80% des lycéennes répondant à cette question et dont les parents sont cadres répondent « Oui », alors on pourra raisonnablement lui imputer la réponse « Oui », ou un *mix* « 80% Oui et 20% Non ». **De même, pour redresser la non-réponse totale, la démarche consiste généralement à considérer des individus similaires** (au regard de certaines caractéristiques) et à prêter aux individus non-répondants les mêmes réponses. C'est comme si, finalement, on comptait plusieurs fois la réponse de certains individus répondants, pour compenser la présence d'individus non-répondants. **Il s'agit de la méthode de la repondération**. A noter que la modification des poids peut parfois altérer la structure de l'échantillon (en amenant par exemple à considérer trop de réponses « filles de cadres ») : le calage sur marges consiste, pour finir, à ré-estimer ces poids de façon à retrouver les bons effectifs totaux (par exemple le nombre total de filles), ou à retrouver certains totaux connus sur l'ensemble de la population (par exemple la moyenne du lycée au bac). Nous y revenons un peu encore dans la section ci-dessous.

Si le redressement de la non-réponse partielle et celui de la non-réponse totale peuvent se faire indifféremment dans cet ordre ou l'ordre inverse, il est important d'avoir à l'esprit que l'association d'un individu non-répondant à un individu répondant est loin d'être anodine. Il se pourrait en effet que certains individus, dits non-substituables, ne ressemblent à aucun autre : en cas de non-réponse, le plus prudent restera alors de faire des relances…

Maintenant que nous avons inféré des statistiques sur la population globale, nous arrivons à la dernière étape de ce papier de blog : l'estimation sur petits domaines. De quoi s'agit-il ? C'est simple, vous réalisez une enquête sur la santé des Français, à partir d'un échantillon

tiré ici et là sur le territoire national, vous obtenez vos statistiques nationales, et là, vous êtes pris d'un doute : que se passe-t-il dans chacune des régions de France ? et si les réalités locales étaient différentes de la moyenne nationale ? Si vous avez stratifié votre échantillon (cf. ci-dessus) selon les régions, et que vous avez interrogé suffisamment d'individus dans chaque région, tout va bien. Autrement, rien ne vous garantit que les personnes tirées dans chaque région soient assez nombreuses ou suffisamment représentatives de leurs régions respectives. En d'autres termes, ce n'est pas parce que l'échantillon est représentatif à l'échelle national qu'il est représentatif, *par morceau*, à l'échelle de chaque région. Si vous passez outre cet avertissement, et que vous utilisez des estimateurs classiques (du type Horvitz-Thompson), vous risquez de tomber sur des statistiques régionales avec des variances très élevées (et donc très peu fiables). Ceci étant dit, on peut quand même se débrouiller, et recourir aux méthodes dites d'*estimation sur petits domaines*.

A nouveau, deux principales techniques statistiques peuvent être considérées : le calage sur marges régionales d'une part, et une modélisation explicite selon la méthode de Fay-Herriot d'autre part. Ces deux techniques exigent d'identifier au préalable des variables dont les « vraies » valeurs (ce que l'on appelle les variables auxiliaires qui permettent de caractériser des *marges*) sont connues au niveau de chaque *domaine* (la région par exemple) et qui sont liées à la variable à estimer. Par exemple, si la variable d'intérêt est la santé, ces variables *auxiliaires* pourraient être diverses variables sociodémographiques (âge, sexe, catégorie socio-professionnelle, diplôme, type de famille, niveau de vie, etc.) ou des variables plus spécifiques, relatives à certains aspects de la vie (travail, santé, lieu de vie, lien social, vie civique, etc.). L'important est de connaître la véritable distribution de chaque variable auxiliaire au niveau géographique considéré, et de vérifier que les variables auxiliaires sont suffisamment corrélées à la variable d'intérêt, indépendamment du domaine (voir l'annexe 1 pour les sources de données possibles).

Pour le dire simplement, le calage sur marges régionales vise à pondérer dans chaque région les individus interrogés, de façon à ce que l'agrégation de leurs caractéristiques retombe sur les agrégats régionaux connus (par exemple la part de cadres dans chaque région). La deuxième méthode, dite de *Fay-Herriot*, revient à faire une estimation économétrique hybride, qui tienne en particulier compte de la taille de chaque région, et des spécificités régionales. Ces deux méthodes sont relativement robustes, et cohérentes entre elles : pour en savoir plus sur leur mise en œuvre technique, et les façons d'en tester la robustesse, nous vous invitons à vous reporter aux annexes et aux suggestions de lecture.

En résumé, pour produire des statistiques fiables, et en dérouler des analyses robustes, vous avez besoin d'un échantillon représentatif et redressé. Et pour produire des

indicateurs locaux, vous pouvez aussi vous appuyer sur des bases de données auxiliaires et certaines techniques statistiques. Pour les données DataCovid, la première étape (et la plus coûteuse !) est déjà franchie : à vous de jouer pour le reste !

Suggestions de lecture (pour la technique et la pratique) :

- « Comment redresser une enquête thématique ?« , Béatrice Neiter et Benoît Buisson, Document de travail, Insee, 2010.
- « Panorama des principales méthodes d'estimation sur les petits domaines », Pascal Ardilly, Document de travail, Insee, 2006.
- « Méthode d'estimation sur petits domaines : l'exemple de la régionalisation d'indicateurs de bien-être subjectif », Adam Baïz et Pierre Villedieu, Document de travail n° 46 CGDD/SDES, 2020.

Annexe 1 : exemples de variables auxiliaires envisageables pour exploiter les données DataCovid au niveau local

Variable (nb. de modalités)	Champs	Source / Base de donnée
Sexe (2)	Tous les individus	Recensement (2010-2013)
Âge (6)	Tous les individus	Recensement (2010-2013)
CSP (8)	15 ans et plus	Recensement (2010-2013)
Type de famille (5)	Tous les ménages	Recensement (2010-2013)
Diplôme (4)	Non scolarisés de 15 ans et plus	Recensement (2010-2013)
Situation vis-à-vis du travail (8)	15 ans et plus	Recensement (2010-2013)
Niveau de vie (10)	Tous les individus	Filosofi (2010-2013)
Part des 75 ans et plus seuls (3)	Tous les individus	Recensement (2011)
Indice de mortalité comparatif (4)	Tous les individus	État-civil – Recensement (2012)

Participation présidentielles 2012 (4)	Tous les individus	Ministère de l'intérieur (2012)
Accès aux équipements (3)	Tous les individus	Base permanente des équipements-Recensement de la population (2013)

Les annexes suivantes sont tirés de ce papier (Baïz et Villedieu, 2020).

Annexe 2 : le calage sur marges régionales

La première méthode correspond à un calage sur des marges régionales. Soit X1, X2, …, Xk des variables auxiliaires dont les valeurs xi,k sont connues pour tous les individus i de l'échantillon {1, …, n} de l'enquête et qui sont liées à la variable d'intérêt Y. Pour chacune de ces variables, est également connu le total X^{tot} au niveau du domaine *a*. Pour chaque région, le calage utilise cette information pour re-pondérer les observations de l'échantillon total national de taille n de façon à retrouver les marges régionales associées aux différentes variables auxiliaires. Formellement, cela signifie qu'il est question de minimiser la distance entre les poids initiaux de l'enquête (car ils possèdent *a priori* de bonnes propriétés statistiques) et les nouveaux poids de façon à ce que ces derniers respectent les équations de calage:

$\forall k \in \{1,\dots,K\}, \sum_{i=1}^{n} x_{i,k} * w_{i,a} = X_{k,a}^{tot}$

Cette procédure doit être réalisée autant de fois qu'il y a de domaines (= régions). A partir du jeu de poids ainsi obtenu, fonction de la région, l'estimation du total de la variable d'intérêt *Y* dans la région *a* est obtenue par la somme des *y* de l'échantillon national, pondérés par leurs poids associés à cette égion :

$Y_a^{est} = \sum_{i=1}^{n} y_{i,a} * w_{i,a}$

Les macros CALMAR (CALage sur MARges) de l'Insee, en accès libre, vous permettront d'aller directement à la mise en œuvre pratique.

Annexe 3 : la modélisation explicite au niveau individuel (méthode Fay-Herriot)

Cette méthode, d'abord introduite par Battese, Harter et Fuller, repose, elle, sur une modélisation explicite, ici au niveau individuel. En plus des aléas classiques d'un modèle économétrique (terme $e_{a,i}$), le modèle fait ici intervenir un effet aléatoire ou « effet domaine

» qui permet de capter de potentielles spécificités (terme va) des individus de chaque domaine (ici les régions, dénotées par a) :

Ya,i=Xa,iTβ+va+ea,i

A partir de ce modèle appartenant à la classe des modèles linéaires mixtes, l'estimateur final, c'est-à-dire la valeur calculée à partir des données et qui permet de se rapprocher de la « vraie » valeur de la variable d'intérêt, peut s'écrire sous la forme suivante :

Y¯^aFH=γay¯a+(X¯a–xa)Tβ~+(1–γa)X¯aTβ~

Cette forme fait apparaître la nature hybride de l'estimateur puisqu'il s'agit d'une moyenne pondérée entre un estimateur synthétique (équivalent à la méthode) et d'un estimateur plus « direct » privilégiant les observations du domaine considéré. Le poids (noté gamma « γ ») accordé à cette partie directe sera d'autant plus grand que la taille du domaine est grande dans l'échantillon, et que cette région se distingue effectivement des autres. Enfin, cette méthode présente l'intérêt de permettre le calcul de l'erreur quadratique moyenne (EQM) en prenant en compte l'aléa du modèle (mais pas l'aléa de sondage).

Les macros STAT CANADA de *Statistics Canada*, disponibles seulement sous convention, vous permettront d'aller directement à la mise en œuvre pratique.