Méthodes statistiques pour les SCIENCES SOCIALES

Flora Chanvril-Ligneel
Viviane Le Hay

ellipses

Introduction

Les données

Ce chapitre introductif précise quelles données utiliser et, sans viser l'exhaustivité, fournit quelques pistes de recherche pour qui souhaite fonder son étude sur des données existantes. Dans la mesure où l'ouvrage s'appuie en grande partie (mais pas seulement cependant) sur des exemples scientifiques issus d'enquêtes par sondage (c'est-à-dire d'une interrogation d'un échantillon d'individus par le moyen d'un questionnaire en vue de saisir leurs comportements, pratiques, attitudes, etc.), nous avons souhaité en outre, dans cette introduction, préciser succinctement quelques spécificités liées à ce mode de recueil de l'information sociale et les précautions indispensables à prendre lors de son emploi. Il présente en particulier trois notions fondamentales liées à l'échantillonnage : la marge d'erreur, les principaux biais de la mesure par ce moyen et les méthodes de redressement[1].

I. Quel type de données ?

I.1. Définitions : recensement et sondage, données individuelles et agrégées

Pour répondre à une question sociale, on peut penser de prime abord que l'idéal est de disposer de données exhaustives, c'est-à-dire d'une information portant sur l'ensemble de la population composant son objet d'étude ou sa population d'intérêt (les lycéens français, les électeurs marseillais, etc.). C'est le cas lorsque l'on procède par ce que l'on appelle un *recensement*. Si le plus connu d'entre eux en France est le *Recensement général de la population*, toute enquête

---

1 Nous n'évoquerons pas ici la question de la protection des données personnelles et de l'anonymisation, très importante dans le cadre de l'utilisation de données. Nous renvoyons néanmoins le lecteur au site Internet de la Commission Nationale de l'Informatique et des Libertés (CNIL) <http://www.cnil.fr/> (consulté le 5 juin 2014), ainsi qu'au chapitre écrit par Roxane Silberman : « Chapitre 12 : La protection des données individuelles en France et la recherche en sciences sociales », Alain Chenu, Laurent Lesnard (dir.), *La France dans les comparaisons internationales : guide d'accès aux grandes enquêtes statistiques en sciences sociales*, Paris, Presses de Sciences Po, 2011, p. 183-204.
De la même manière, ce manuel de méthodologie n'est pas un ouvrage de sociologie de la mesure de l'opinion et du débat qui entoure cette question. Des éléments de réflexion à cet égard sont disponibles dans les articles suivants : Loïc Blondiaux, « Ce que les sondages font à l'opinion publique », *Politix*, vol. 10, n°37, 1997, p. 117-136 ; Pierre Bourdieu, « L'opinion publique n'existe pas », *Questions de sociologie*, Paris, Les Éditions de Minuit, 1980, p. 222-235 ; Claude Dargent, *Sociologie des opinions*, Paris, Armand Colin, 2011, 240 p. ; Gérard Grunberg, Nonna Mayer, Paul Sniderman (dir.), *La démocratie à l'épreuve : une nouvelle approche de l'opinion des Français*, Paris, Presses de Sciences po, 2002, 349 p. ; Nonna Mayer, *Sociologie des comportements politiques*, Paris, Armand Colin, 2010, 288 p. ; Paul M. Sniderman, « Les nouvelles perspectives de la recherche sur l'opinion publique », *Politix*, n°41, 1998, p. 123-175.

par questionnaire qui vise à interroger l'entièreté d'une population d'intérêt se nomme en réalité un recensement[1].

À moins que l'objet d'enquête porte sur une population de petite taille et facile d'accès, les coûts de collecte associés à ce mode de recueil sont cependant considérables, en termes financiers et humains, mais aussi en temps. Dans la plupart des cas, on ne dispose pas de l'ensemble de ces moyens : on préfère alors procéder par sondage, ce qui signifie que l'on s'appuiera sur une extraction, un échantillon de la population-mère. Les données ainsi obtenues sont donc partielles et ne représentent qu'une partie de la population d'intérêt. On parle dans ce cas de *données d'échantillon* et la qualité de l'échantillonnage mis en œuvre (c'est-à-dire des techniques déployées pour procéder à l'extraction[2]) est alors essentielle pour obtenir des données fiables et analysables ensuite du point de vue scientifique. Il s'agit en effet cette fois d'obtenir un miroir en modèle réduit de la population étudiée de telle sorte qu'il lui soit le plus fidèle, et donc le plus représentatif possible. Il existe bien entendu une palette d'outils et de précautions à respecter pour atteindre cet objectif indépassable.

Par ailleurs, les données peuvent être classées en deux grands types :

- les données individuelles (micro) : elles portent sur les unités statistiques de base et sont issues la plupart du temps d'enquêtes par sondage. On pense en tout premier lieu aux personnes physiques interrogées par questionnaire ;
- les données agrégées (macro) : elles sont issues de l'agrégation de données individuelles et proviennent généralement d'organismes institutionnels producteurs de données (voir ci-dessous). Il peut s'agir de quartiers, d'établissements scolaires, de divisions administratives – communes, départements, régions –, etc.

S'il est aisé de transformer des données individuelles en données agrégées, l'inverse n'est pas vrai. Prenons un exemple fictif : un fichier constitué d'individus d'une commune renseignant leur statut professionnel et leur quartier d'habitation. Le tableau correspondant serait le suivant (tableau 1).

1 Ainsi, le recensement de la population, assuré par l'Institut national de la statistique et des études économiques (INSEE) en France, se fonde sur deux questionnaires, l'un portant sur le logement (feuille de logement), l'autre sur les individus qui le composent (bulletin individuel). Il consiste non seulement à compter les personnes qui vivent sur le territoire, mais également à connaître leurs principales caractéristiques sociodémographiques (âge, sexe, composition des familles, niveau de formation, etc.) et conditions de vie (conditions de logement en particulier, ainsi que conditions d'emploi, etc.). Notons néanmoins que les techniques de recueil ont été modifiées à partir de 2004 (annualisation).

2 Se reporter à la partie II de ce chapitre pour une présentation générale des principales méthodes d'échantillonnage existantes.

Tableau 1 : Exemple de données individuelles, statut professionnel en trois catégories et quartier d'habitation (extraction)

Individus	Statut professionnel	Quartier d'habitation
1	Chômeur	A
2	Actif non chômeur	B
3	Inactif	B
4	Actif non chômeur	B
5	Actif non chômeur	B
6	Actif non chômeur	C
7	Actif non chômeur	A
8	Chômeur	C
9	Inactif	C
10	Inactif	C
11	Inactif	A
12	Actif non chômeur	A
13	Inactif	A
14	Actif non chômeur	A
15	Chômeur	B
16	Chômeur	B
17	Inactif	C
18	Chômeur	A
19	Actif non chômeur	A
20	Actif non chômeur	C
21	Actif non chômeur	A
22	Chômeur	B
23	Chômeur	A
24	Chômeur	C
25	Actif non chômeur	C
26	Inactif	A
27	Actif non chômeur	B
28	Actif non chômeur	C
29	Inactif	B
30	Actif non chômeur	A
etc.	etc.	etc.

Sources : données fictives

Il est possible d'agréger facilement la variable statut professionnel par quartier d'habitation afin d'obtenir les taux de chômage dans chaque quartier (tableau 2).

Tableau 2 : Exemple de données agrégées, taux de chômage selon le quartier d'habitation, %

Quartier	Taux de chômage
A	14
B	11
C	8

Sources : données fictives

En revanche, si l'on disposait comme unique source d'informations du tableau 2, il serait impossible de retrouver les détails individuels (ici le statut professionnel de chaque habitant).

I.2. Où trouver des données ?

Il existe de nombreuses sources, en particulier en ligne, donnant accès à des données sociales portant sur des objets d'étude infinis[1]. En voici certaines d'entre elles, couramment utilisées par les auteurs du fait de leurs spécialités thématiques. Il va sans dire qu'au vu de l'ampleur des données accessibles, le catalogue pourrait être très largement étoffé. En outre, le mouvement des *open data* met à disposition de plus en plus de données et rendra certainement assez vite caduque ce premier inventaire. Celui-ci reste néanmoins susceptible de guider en première instance le lecteur peu familier des méandres de la statistique sociale disponible en ligne.

Où trouver des données agrégées ?[2]

- pour les données concernant principalement **la France**, plusieurs organismes généralistes utiles existent, en particulier l'Institut national de la statistique et des études économiques (INSEE[3]), l'Institut national d'études démographiques (INED[4]), le Portail de la statistique publique[5], la Plateforme ouverte des données publiques françaises (DATAGOUV[6]) ou encore le Centre de recherche pour l'étude et l'observation des conditions de vie (CREDOC[7]);
- pour les **données européennes ou internationales**, les équivalents étrangers de l'INSEE sont indiqués sur son site Internet[8], et d'autres organismes sont également importants à connaître, comme l'Organisation de coopération et de développement économiques (OCDE[9]), EUROSTAT[10], la division statistique de l'Organisation des Nations Unies (ONU[11]) ou encore la Data world bank[12] ;
- pour les **données cartographiques**, de multiples sources, généralistes ou spécifiques, existent également. Citons par exemple l'atelier de cartographie de Sciences Po[13], ainsi que deux outils de cartographie électorale, CARTELEC[14] et VIZLAB[15].

1 Les listes qui suivent n'ont pas la prétention d'être exhaustives. Elles permettent néanmoins d'avoir une vision globale des principales sources de données existantes. Notons par ailleurs que les droits d'accès sont plus ou moins stricts selon les sources citées.

2 Nous indiquons ici les liens vers la page d'accueil de chacun des organismes ou banques de données. Pour les premiers, un onglet « bases de données » ou « statistiques » est généralement disponible dès l'ouverture de cette page d'accueil.

3 <http://www.insee.fr/fr/>, consulté le 19 mai 2014.

4 <http://www.ined.fr/>, consulté le 19 mai 2014.

5 <http://www.statistique-publique.fr/>, consulté le 19 mai 2014.

6 <http://www.data.gouv.fr/>, consulté le 19 mai 2014.

7 <http://www.credoc.fr/>, consulté le 19 mai 2014.

8 <http://www.Insee.fr/fr/Insee-statistique-publique/default.asp?page=sites-statistiques/instituts-nationaux-statistiques.htm>, consulté le 19 mai 2014.

9 <http://www.oecd.org/fr/>, consulté le 19 mai 2014.

10 <http://epp.eurostat.ec.europa.eu/portal/page/portal/eurostat/home/>, consulté le 19 mai 2014.

11 <http://www.un.org/fr/databases/index.shtml>, consulté le 19 mai 2014.

12 <http://data.worldbank.org/>, consulté le 19 mai 2014.

13 <http://cartographie.sciences-po.fr/>, consulté le 19 mai 2014.

14 <http://www.cartelec.net/>, consulté le 19 mai 2014.

15 <http://cdsp.sciences-po.fr/vizlab/?locale=fr>, consulté le 19 mai 2014.

Où trouver des données individuelles ?

- **en France**, le réseau Quételet[1] et le Centre de données socio-politiques de Sciences Po (CDSP[2]) archivent et diffusent les données d'enquête par questionnaire en sciences humaines et sociales ;
- le consortium interuniversitaire pour la recherche en science politique et en sociologie (ICPSR[3]) ainsi que l'institut de Leibniz pour les sciences sociales (GESIS[4]) archivent et diffusent les données **à l'international**. Chaque grande enquête comparative possède également son propre site Internet[5], en particulier les enquêtes *Eurobaromètres* (EB[6]), celles de l'*International social survey programme* (ISSP[7]), les *European value studies* (EVS[8]) et *World value studies* (WVS[9]), les enquêtes *European social survey* (ESS[10]), ainsi que les autres baromètres régionaux (Latino barometer[11], African barometer[12], Asian barometer[13], Arab barometer[14]). Enfin le Pew research center[15] donne également accès à des données utiles.

Voici enfin les principaux instituts de sondage en France : BVA[16], CSA[17], GFK[18], IFOP[19], IPSOS[20], LH2[21], OPINION WAY[22], TNS-SOFRES[23].

1 <http://www.reseau-quetelet.cnrs.fr/>, consulté le 19 mai 2014. Il est intéressant de noter que le réseau Quételet dispose également d'un outil spécifique, la Base De Questions (BDQ), permettant de rechercher une question spécifique d'un des questionnaires archivés à l'adresse suivante <http://bdq.reseau-quetelet.cnrs.fr/fr/Accueil>, consulté le 19 mai 2014.

2 <http://cdsp.sciences-po.fr/>, consulté le 19 mai 2014.

3 <http://www.icpsr.umich.edu>, consulté le 19 mai 2014.

4 <http://zacat.gesis.org>, consulté le 12 juin 2014.

5 Pour une présentation détaillée des trois principales enquêtes européennes (outre l'Enquête sociale européenne) : Pierre Bréchon, « Les Grandes enquêtes internationales (Eurobaromètres, Valeurs, ISSP) : apports et limites », *Année sociologique*, vol. 52, n°1, 2002, p. 105-130.

6 <http://ec.europa.eu/public_opinion/index_en.htm>, consulté le 19 mai 2014

7 <http://www.issp.org/>, consulté le 19 mai 2014.

8 <http://www.europeanvaluesstudy.eu/>, consulté le 19 mai 2014.

9 <http://www.worldvaluessurvey.org/>, consulté le 19 mai 2014.

10 <http://www.europeansocialsurvey.org/>, consulté le 19 mai 2014.

11 <http://www.latinobarometro.org>, consulté le 19 mai 2014.

12 <http://www.afrobarometer.org/>, consulté le 19 mai 2014.

13 <http://www.asianbarometer.org/>, consulté le 19 mai 2014.

14 <http://www.arabbarometer.org/>, consulté le 19 mai 2014.

15 <http://www.pewresearch.org/>, consulté le 19 mai 2014.

16 <http://www.bva.fr>, consulté le 19 mai 2014.

17 <http://www.csa.eu/>, consulté le 19 mai 2014.

18 <http://www.gfk.com/fr>, consulté le 19 mai 2014.

19 <http://www.ifop.com/>, consulté le 19 mai 2014.

20 <http://www.ipsos.fr/>, consulté le 19 mai 2014.

21 <http://www.lh2.fr/>, consulté le 19 mai 2014.

22 <http://www.opinion-way.com/>, consulté le 19 mai 2014.

23 <http://www.tns-sofres.com/>, consulté le 19 mai 2014.

II. Les données issues d'enquêtes par sondage

L'élaboration d'une enquête par questionnaire (d'un sondage) nécessite la plupart du temps de procéder à un échantillonnage de la population d'intérêt pour n'en interroger qu'un sous-ensemble que l'on souhaite le plus fidèle possible (il n'est généralement pas envisageable d'interroger la totalité des sujets qui composent son objet d'étude). Cela consiste à en extraire une partie appelée échantillon. Celui-ci est généralement de petite taille comparée à celle de la population-mère. Pour pouvoir conclure du point de vue de la population d'intérêt à partir de cet échantillon, il importe que ce dernier soit constitué avec le plus grand soin : de la qualité de cette phase technique dépend en effet la fiabilité de l'échantillon et donc les résultats du sondage. On parlera ici de la *représentativité* de l'échantillon, qui conditionne en grande partie les possibilités d'analyse. Différentes méthodes d'échantillonnage, aux caractéristiques techniques spécifiques et adaptées à des contextes d'utilisation particuliers, peuvent être employées. Nous en esquissons un rapide portrait ici, afin que le lecteur puisse saisir la complexité de cette étape, qui requiert une grande maîtrise et un long travail de réflexion avant toute mise en œuvre concrète.

Ce qu'il doit absolument retenir, c'est qu'un échantillon, par construction et quels que soient les efforts déployés pour l'élaborer avec la plus grande rigueur, n'en demeure pas moins un échantillon. Puisque la population d'intérêt dans son ensemble n'aura pas été touchée, il s'accompagne de fait d'une incertitude dans sa construction (un échantillon n'est qu'une approximation) et d'une erreur de mesure contre laquelle on ne peut rien, si ce n'est ne pas l'ignorer.

En outre, il ne s'agit pas de l'unique erreur de mesure de ce type d'outil de collecte de l'information. Nous présenterons également deux biais supplémentaires importants de l'enquête par questionnaire, qu'il convient d'avoir en tête avant toute analyse de ce type de matériau : une telle connaissance permet en effet d'éviter les surinterprétations et les mésusages malheureux.

Précisons néanmoins d'ores et déjà que parler d'incertitude (et de biais) n'invalide en aucun cas la démarche, à partir du moment où l'on s'efforce d'une part de la minimiser, d'autre part d'en tenir compte dans son approche et de faire avec.

II.1. Les principales méthodes d'échantillonnage[1] et leur biais intrinsèque : nous avons affaire à des échantillons !

Il existe deux grandes familles d'échantillonnage :

- l'échantillonnage *aléatoire*, aussi appelé *probabiliste* : ici, nous sommes dans l'univers du hasard. Il consiste à tirer au sort, au sein de la population

[1] Le lecteur souhaitant en apprendre davantage sur les méthodes d'échantillonnage et leurs implications mathématiques peut se référer à Pascal Ardilly, *Les techniques de sondage*, Paris, Technip, 2006, 675 p. Une présentation plus accessible au lecteur débutant est consultable dans : Jean-Paul Bozonnet, Pierre Bréchon, « Établir un échantillon représentatif », Pierre Bréchon (dir.), *Enquêtes qualitatives, enquêtes quantitatives*, Grenoble, Presses universitaires de Grenoble, 2011, p. 123-143. Enfin, nous invitons le lecteur à consulter les sites Internet des grandes enquêtes internationales et les sites de diffusion d'enquête (précédemment indiqués) qui fournissent un vivier considérable d'exemples d'échantillonnage (chaque enquête présente en détail sa méthode de collecte).

d'intérêt et selon des techniques plus ou moins originales, novatrices et complexes, les personnes que l'on va effectivement contacter pour être interrogées. La probabilité pour un individu de figurer dans l'échantillon est appelée *probabilité d'inclusion* et découle de la loi des grands nombres. Celle-ci dépend du type d'échantillonnage aléatoire utilisé mais présente l'intérêt d'être connue. C'est ainsi que l'on peut estimer l'incertitude, c'est-à-dire l'imprécision qui accompagne le sous-ensemble ainsi constitué. Dans ses formules les plus simples qui consistent à s'appuyer sur des listes d'individus au sein desquelles on tire au sort (chaque personne figurant dans la liste ayant la même chance ou probabilité d'être choisie), cette technique nécessite une connaissance fine de la population-mère et implique la possibilité d'accéder à des listes nominatives, ce qui n'est pas toujours possible, en France en particulier. C'est alors qu'il faut user d'inventivité dans la méthode de prélèvement aléatoire (mais c'est généralement très coûteux financièrement) ou procéder de façon non probabiliste ;

- l'échantillonnage *non aléatoire*, aussi appelé *empirique* : les méthodes aléatoires présentant un coût de mise en œuvre élevé, la méthode empirique des quotas est souvent utilisée en France. Lorsque l'on procède de la sorte, il s'agit de reproduire dans l'échantillon les proportions observées, dans la population d'intérêt, de certaines caractéristiques jugées déterminantes. Par exemple, la population française comporte environ 52% de femmes. Un échantillon de 1000 personnes (souhaitant reproduire la population française) utilisant un quota de sexe devra comporter 520 femmes et 480 hommes. De la sorte, ces catégories de population ne seront ni sur-, ni sous-représentées dans l'échantillon. Plus l'on mobilisera de quotas différents, plus il s'avèrera délicat de tirer un échantillon correspondant exactement aux distributions de chacun d'entre eux. L'objectif sera donc de s'en rapprocher au plus près pour aboutir à un échantillon le moins déformé possible. Il conviendra également de ne pas multiplier à l'excès les critères retenus. Classiquement et parce qu'ils sont connus dans la population générale grâce au recensement de la population effectué par l'Insee, ce sont les quotas d'âge, de sexe et de profession, assortis d'un contrôle par la région et la catégorie d'agglomération d'appartenance, auxquels les sondages d'opinion ont recours. L'univers académique a coutume, quand cela est possible, d'y ajouter un critère de niveau de diplôme (la question du coût entre en ligne de compte pour tout rajout de quota). Cependant, par cette méthode d'échantillonnage, la probabilité d'inclusion est inconnue et ne peut donc être, tout au moins en théorie, calculée. La conséquence la plus forte pour ce qui nous concerne est qu'il serait alors impossible d'estimer l'incertitude (l'erreur) qui accompagne notre échantillon. Ceci explique le rejet et la méfiance généralement suscités par cette méthode aux fondements non scientifiques dans l'univers des statisticiens.

Revenons un instant sur la question de la mesure de l'incertitude en fonction de ces deux grandes familles d'échantillonnage. Comme nous l'avons souligné, les fondements mathématiques et probabilistes sur lesquels s'appuient les démonstrations liées à l'inférence statistique (c'est-à-dire l'extrapolation des tendances observées dans son échantillon à l'ensemble de la population mère), à la marge d'erreur et au calcul d'intervalles de confiance (*cf.* chapitre 4) impliquent un échantillonnage aléatoire et s'appliquent donc, en théorie, uniquement dans ce cadre. En d'autres termes, un échantillon par quota ne pourrait faire l'objet de traitements statistiques ayant recours à ces méthodes et aucune mesure de l'incertitude ne serait permise.

Nous souhaitons néanmoins nuancer cette première conclusion et défendons au contraire une thèse inverse. Une longue tradition de sondage par quota nous indique même la pertinence de ce type de pratiques. En cela, nous rejoignons Jean-Paul Bozonnet et Pierre Bréchon : « Malgré l'absence de fondement scientifique, cette technique "marche" aussi bien que la méthode aléatoire. Sa fiabilité est démontrée par la constance des résultats obtenus lors de sondages successifs sur des questions qui ne sont pas soumises aux aléas de l'opinion. (...) Il est vrai qu'en toute rigueur la méthode aléatoire est la seule à autoriser le calcul d'une marge d'erreur. Mais, si on considère que la procédure des quotas est en fait une reconstitution de l'aléa à moindre coût, on peut en déduire que les marges d'erreur valables pour la méthodologie aléatoire le sont aussi pour les quotas »[1].

En suivant la ligne de ces auteurs, nous recommandons donc au lecteur :

- 1. d'admettre que quelle que soit la façon dont il a constitué son échantillon il ne s'agit que d'un échantillon et que par conséquent, il ne peut pas être une copie à l'exact identique de sa population-mère. Même s'il s'en rapproche, il existe par conséquent une incertitude, même très faible, appelée marge d'erreur, que l'on peut plus ou moins bien reconstituer et qui varie en fonction de trois grands critères (l'encart 1 développe la notion de marge d'erreur) ;
- 2. à défaut de pouvoir les calculer rigoureusement quel que soit son mode d'échantillonnage et plutôt que de ne rien faire, mieux vaut appliquer les marges d'erreur que l'on connaît dans le cas de la procédure aléatoire à d'autres types d'échantillons[2]. Ces derniers devront toutefois être élaborés en fonction de critères de rigueur déjà éprouvés. Un échantillon élaboré « au petit bonheur la chance », c'est-à-dire sans réflexion approfondie préalable sur les procédures à mettre en œuvre et sans travail de documentation fouillée sur les méthodes existantes, n'entre pas dans cette catégorie. Nous excluons également ici tout échantillon au sein duquel la personne interrogée n'est pas préalablement et rigoureusement sélectionnée par l'analyste. Nous pouvons citer ici l'exemple des expériences de type *vote de paille* : poser une question à plusieurs millions d'individus sans prendre en compte le moindre critère statistique dans leur

1 Jean-Paul Bozonnet et Pierre Bréchon, *op. cit.*, p. 135 et 140.

2 Des études ont été effectuées pour montrer qu'il est possible d'approximer la marge d'erreur dans le cas d'échantillons construits de manière non aléatoire, notamment Jean-Claude Deville, « Une théorie des enquêtes par quota », *Techniques d'enquête*, vol. 17, n°2, 1991, p. 163-181.

sélection fournit un résultat à coup sûr erroné, alors qu'un sondage fondé sur un échantillon de taille modeste mais élaboré de façon plus fiable donne un résultat plus juste[1]. Ainsi les questions posées par les médias à leurs lecteurs/téléspectateurs (de type « la question du jour »), n'ont pour le coup aucune espèce de validité scientifique ou de représentativité au-delà des personnes qui ont bien voulu répondre ;

- 3. contrairement à d'autres méthodes raisonnées davantage sujettes à caution, la méthode des quotas, à condition qu'elle soit menée en toute rigueur, peut être appréhendée comme une forme de reproduction de l'aléatoire. À ce titre, nous considérons que mieux vaut tenir compte des marges d'erreur que l'on connaît dans le cadre de l'échantillon probabiliste plutôt que de faire comme si l'incertitude n'existait pas !

Encart 1 : ATTENTION PRUDENCE !
La marge d'erreur liée à l'échantillonnage

Si échantillonner permet de réduire fortement les coûts d'enquête par rapport à un recensement (imaginons un sondage portant sur 2000 personnes représentatives de la population française versus un recensement exhaustif de cette même population française !), le simple fait d'effectuer ensuite des calculs et des analyses à partir d'un échantillon et d'en tirer des conclusions sur la population d'intérêt introduit inexorablement une erreur de mesure. Mais dans le cas d'un échantillonnage aléatoire et selon nous d'une sélection par quota, cette dernière est quantifiable et donc contrôlable. C'est ce que l'on appelle la marge d'erreur, dont la construction sera définie mathématiquement et plus en détail dans le chapitre 4.

Que donne cette marge d'erreur dans le cas de la mesure d'une proportion (par exemple, le pourcentage d'intentions de vote pour le candidat A à une élection donnée, le pourcentage de personnes ayant regardé la TV la veille de l'interview, etc.) ? Les tableaux 3a et 3b détaillent le niveau de celle-ci.

Partons d'un exemple classique d'analyse électorale afin de mieux comprendre à quoi correspond cette marge d'erreur concrètement. Supposons que nous avons effectué un sondage auprès de 1000 individus français, en âge de voter et inscrits sur les listes électorales et que nous obtenons 52% d'intentions de vote pour le candidat A en vue du second tour d'une élection présidentielle. Laquelle de ces trois propositions est vraie ?[2]

1 En particulier celle de 1936 opposant le vote de paille du *Literary Digest* (établi auprès de 2,4 millions de personnes et mesurant à tort une défaite de Roosevelt) et le sondage mis en place par l'institut Gallup (établi quant à lui auprès de 5000 personnes choisies au hasard sur le territoire américain et mesurant à raison une victoire de Roosevelt).

2 Il s'agit à ce stade d'un cas d'école, qui ne tient pas compte d'une éventuelle dissimulation (réponse volontairement amplifiée, minimisée, voire erronée) de la personne interrogée par exemple.

- 1. 52% des Français en âge de voter et inscrits sur les listes électorales ont l'intention de voter pour le candidat A ;
- 2. 52% des personnes interrogées (françaises, en âge de voter et inscrites sur les listes électorales) ont l'intention de voter pour le candidat A ;
- 3. au seuil de confiance de 95%, entre 49% et 55% des Français en âge de voter et inscrits sur les listes électorales ont l'intention de voter pour le candidat A.

La première interprétation est doublement fausse : elle fait de « l'inférence sauvage » en extrapolant directement le résultat mesuré dans l'échantillon à la population d'intérêt, sans aucune prise en compte de l'incertitude. La deuxième est en partie juste car elle prend en compte le fait que le résultat est mesuré parmi les 1000 personnes interrogées mais elle oublie néanmoins de préciser l'incertitude. La troisième et dernière est donc la seule manière rigoureuse de présenter un résultat issu d'une enquête par sondage : la marge d'erreur, décrivant l'incertitude liée à l'échantillonnage effectué, est indiquée clairement et l'inférence peut ainsi être effectuée (passage des « personnes interrogées » aux « Français »).

Mais comment obtient-on cette marge d'erreur entre 49% et 55% ?

Ces valeurs proviennent de la lecture du tableau 3a ci-dessous.

Lecture : au seuil de confiance de 95% (tableau 3a), pour une proportion estimée d'environ 50% (ici 52% : dernière ligne du tableau) et un échantillon d'environ 1000 personnes (3e colonne du tableau), on peut lire que l'on obtient une **marge d'erreur de 3,1 points de pourcentage** au-dessus et en-dessous (**arrondie ici à 3 points**).

C'est ainsi que l'on aboutit, partant d'une proportion mesurée dans l'échantillon de 52%, à l'intervalle borné entre **49%** (52% - 3 points) et **55%** (52% + 3 points).

Le fait notable ici est bien de relever que si le résultat de ce sondage peut sembler, **en apparence**, favorable au candidat A, ce n'est en réalité par le cas si l'on tient compte de la marge d'erreur. Ainsi, ce candidat peut tout à fait obtenir *seulement* 49% des suffrages, sans qu'il soit possible de remettre en cause la qualité du sondage réalisé.

Tableau 3a : La marge d'erreur pour une proportion, pour un seuil de confiance de 95%, en fonction de la proportion et de la taille de l'échantillon

		Taille de l'échantillon				
		10	100	1000	10000	100000
Proportion mesurée	10% ou 90%	18,6	5,9	1,9	0,6	0,2
	20% ou 80%	24,8	7,8	2,5	0,8	0,2
	30% ou 70%	28,4	9,0	2,8	0,9	0,3
	40% ou 60%	30,4	9,6	3,0	1,0	0,3
	50%	31,0	9,8	3,1	1,0	0,3

Sources : calculs des auteurs

Tableau 3b : La marge d'erreur pour une proportion, pour un seuil de confiance de 99%, en fonction de la proportion et de la taille de l'échantillon

		Taille de l'échantillon				
		10	100	1000	10000	100000
Proportion mesurée	10% ou 90%	24,4	7,7	2,4	0,8	0,2
	20% ou 80%	32,6	10,3	3,3	1,0	0,3
	30% ou 70%	37,3	11,8	3,7	1,2	0,4
	40% ou 60%	39,9	12,6	4,0	1,3	0,4
	50%	40,7	12,9	4,1	1,3	0,4

Sources : calculs des auteurs

Il est également fondamental de noter que la marge d'erreur varie en fonction de trois critères (ces constats peuvent être vérifiés à la lecture des tableaux 3a et 3b) :

- la **taille de l'échantillon** : plus l'échantillon est grand, plus la marge d'erreur est faible. Mais l'évolution de celle-ci n'est pas linéaire : multiplier par 10 la taille de l'échantillon diminue par seulement 3,16 la marge d'erreur associée à la mesure. Cela justifie le fait que l'on ait bien souvent affaire à des sondages se fondant sur des échantillons d'environ 1000 personnes : cette taille d'échantillon constitue en quelque sorte le meilleur arbitrage entre le coût du sondage et la précision de la mesure obtenue ;
- le **niveau de la proportion mesurée** : plus celle-ci se rapproche de 50%, plus la marge d'erreur augmente. On le comprend intuitivement : plus les personnes interrogées apparaissent partagées sur une question, plus il est difficile de la chiffrer précisément sur la base d'un échantillon ;
- le **seuil de confiance** : plus il est proche de 100%, plus la marge d'erreur est importante. Il sera défini plus en détail au chapitre 4. Ce seuil de confiance est déterminé par l'analyste. Globalement, il permet de fixer la probabilité avec laquelle la proportion que l'on cherche à mesurer est effectivement égale à la proportion mesurée dans l'échantillon plus ou moins la marge d'erreur. En d'autres termes, dans notre exemple, le seuil de confiance de 95% consiste à dire qu'il y a 95 chances sur 100 que la proportion en faveur du candidat A se situe entre 49% et 55% : il reste encore 5 chances sur 100 que cela soit au-delà de ces limites, mais l'analyste accepte de prendre un tel risque. Mécaniquement et on peut le comprendre intuitivement, le fait de durcir ce seuil de confiance fait augmenter la marge d'erreur que l'on calcule (*cf.* tableau 3b). En sciences sociales, ce seuil à 95% est généralement celui que l'on pratique (étant entendu que dans d'autres usages, ce seuil peut être considéré comme insuffisant ou trop généreux – en aéronautique par exemple ! –).
- Il est enfin important de relever que la marge d'erreur ne dépend en aucune manière de la taille de la population d'intérêt. Cela signifie, de façon sans doute contre-intuitive, qu'un échantillon de 1000 personnes dûment constitué aboutira à un niveau de

précision équivalent quelle que soit la taille de la population mère : avec un tel échantillon, la marge d'erreur est rigoureusement identique que l'on travaille par exemple sur la population chinoise dans son ensemble ou bien sur la population d'une commune de 50 000 habitants.

La loi n°2002-214 du 19 février 2002 définit dans ses articles 2 et 3 le cadre de publication des sondages, et en particulier les précautions « d'interprétation [à prendre vis-à-vis] des résultats publiés », faisant directement référence aux marges d'erreur :

« Article 2

Modifié par Loi n°2002-214 du 19 février 2002 art. 1er (JORF 20 février 2002)

La publication et la diffusion de tout sondage tel que défini à l'article 1er doivent être accompagnées des indications suivantes, établies sous la responsabilité de l'organisme qui l'a réalisé :

Le nom de l'organisme ayant réalisé le sondage ;

Le nom et la qualité de l'acheteur du sondage ;

Le nombre des personnes interrogées ;

La ou les dates auxquelles il a été procédé aux interrogations ;

Une mention indiquant le droit de toute personne à consulter la notice prévue par l'article 3.

Article 3

Modifié par Loi n°2002-214 du 19 février 2002 art. 2 (JORF 20 février 2002)

Avant la publication ou la diffusion de tout sondage tel que défini à l'article 1er, l'organisme qui l'a réalisé doit procéder au dépôt auprès de la commission des sondages instituée en application de l'article 5 de la présente loi d'une notice précisant notamment :

L'objet du sondage ;

La méthode selon laquelle les personnes interrogées ont été choisies, le choix et la composition de l'échantillon ;

Les conditions dans lesquelles il a été procédé aux interrogations ;

Le texte intégral des questions posées ;

La proportion des personnes n'ayant pas répondu à chacune des questions ;

Les limites d'interprétation des résultats publiés ;

S'il y a lieu, la méthode utilisée pour en déduire les résultats de caractère indirect qui seraient publiés.

La commission des sondages peut ordonner la publication par ceux qui ont procédé à la publication ou à la diffusion d'un sondage tel que défini à l'article 1er des indications figurant dans la notice qui l'accompagne ou de certaines d'entre elles.

Toute personne a le droit de consulter auprès de la commission des sondages la notice prévue par le présent article. »[1]

De l'importance de ne pas être trop précis : adieu les pourcentages précisant les décimales !

Dans le cadre d'une enquête par sondage, il est courant d'utiliser la répartition des réponses aux questions en pourcentages, ventilés en fonction de sous-groupes ou non. Dans ce cas, **il n'est pas nécessaire, si ce n'est trompeur, de préciser les chiffres après la virgule**[2]. En effet, rappelons-nous qu'il existe une marge d'erreur ! Et ces valeurs décimales sont dans la grande majorité des cas (à moins d'un échantillon de très grande taille) compris dans cette marge d'erreur ! Moralité : rien ne sert de vouloir être trop précis !

Prenons un exemple (tableau 4).

Tableau 4 : Distribution statistique d'une variable exprimée en pourcentages et marge d'erreur associée

Souhaitez-vous dans l'avenir devenir propriétaire ?		Marge d'erreur à 95%
Non, pas du tout	32	± 2,8 points
Non, pas vraiment	11	± 1,9 points
Oui, plutôt	22	± 2,6 points
Oui, tout à fait	33	± 2,9 points
NSP	2	± 0,9 points

Sources : *Les Français et le logement*, TNS Sofres pour les *États généraux du logement*, avril-mai 2011

Dans cet exemple, les marges d'erreur sont comprises entre 1 et 3 points de pourcentage. Compte tenu de ces dernières et de l'incertitude qui accompagne les résultats, il s'avère donc inutile voire fallacieux d'indiquer les chiffres après la virgule pour les pourcentages associés à la distribution de cette variable : ainsi, savoir qu'il y a en réalité 32,3% ou 32,7% des personnes interrogées qui ne souhaitent pas du tout devenir propriétaire constitue une précision bien peu indispensable si on garde à l'esprit la marge d'erreur qui accompagne cette proportion mesurée...

1. <http://www.commission-des-sondages.fr/lois/lois.htm>, consulté le 14 mai 2014.
2. De plus, préciser les chiffres après la virgule ne facilite pas la lecture ! Nous conseillons vivement au lecteur le texte suivant : Olivier Temam, « Épisode n°1. Des chiffres : point trop n'en faut », *Courrier des statistiques*, Hors-série « Savoir compter, savoir conter », décembre 2009, p. 5-7.

II.2. Le « répondant fantôme » : un deuxième biais de mesure de l'enquête par questionnaire

Deux grands défis sont à relever lorsque l'on construit un échantillon. Le premier consiste à réussir à capter la totalité de la population d'intérêt et à ne pas laisser de côté une partie de celle-ci. S'il existe des questions de disponibilité et de plus ou moins grande facilité dans la prise de contact, la difficulté majeure réside dans les unités statistiques jamais saisies au cours de l'échantillonnage. C'est ce que l'on appelle le « répondant fantôme ».

Ce répondant fantôme correspond à deux situations distinctes et cumulatives :

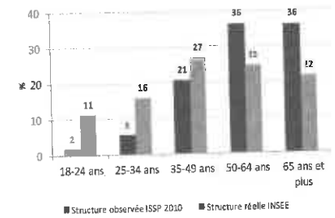
- l'erreur de couverture : elle correspond aux unités statistiques non couvertes (non prises en compte) lors de l'échantillonnage et qui ne peuvent donc pas être mesurées. La probabilité d'inclusion est alors nulle ou inconnue pour ces unités. Par exemple, si l'on procède à une enquête par téléphone, les personnes qui ne disposent pas d'un numéro (fixe ou portable) mais qui font pourtant partie de la population d'intérêt ne pourront jamais être jointes et constitueront l'erreur de couverture ;
- la non-réponse totale : elle est constituée par les refus de répondre à l'enquête, c'est-à-dire par les personnes contactées mais qui ne consentent pas à participer au dispositif. Ce biais se rajoute donc dans un deuxième temps à l'erreur de couverture. Ces refus répondent à des motivations variées : une méfiance globale envers les sondages, des personnes peu présentes à leur domicile dans le cas d'enquêtes en face-à-face et qu'on ne parvient pas à toucher, ou encore des individus qui déclarent ne pas avoir le temps, ne pas être intéressés par le sujet de l'enquête, etc. Quelle que soit la motivation de celui-ci, le refus de répondre n'est jamais socialement neutre et constitue de ce fait un biais d'échantillonnage !

Ces deux biais peuvent se résumer ainsi : si l'on somme le taux de réponse, le taux de refus et le taux de non couverture, on obtient un total de 100%. Le taux de couverture correspond donc à la somme du taux de réponse et du taux de refus.

Prenons un exemple comparant la distribution des classes d'âge dans l'enquête ISSP[1] de 2010 et celle dans le recensement de l'INSEE de la même époque (figure 1).

1 <http://issp-france.upmf-grenoble.fr/>, consulté le 12 juin 2014. Il s'agit d'une des grandes enquêtes sociologiques internationales de référence.

Figure 1 : Structure par âge dans l'enquête ISSP 2010 et structure réelle issue du recensement de l'INSEE



Sources : Enquête ISSP 2010, données non pondérées ; INSEE, estimations de population au 1er janvier 2010

On constate une sous-représentation des 18-24 ans (-9 points), des 25-34 ans (-11 points) et des 35-49 ans (-6 points) dans l'enquête contre une surreprésentation des 50-64 ans (+11 points) et des 65 ans et plus (+14 points). Pour déterminer la cause de ce déficit des classes d'âge les plus jeunes, il faudrait étudier les individus n'ayant pas répondu à l'enquête afin de déterminer si cette sous-représentation proviendrait plutôt d'une non-réponse totale ou d'une erreur de couverture.

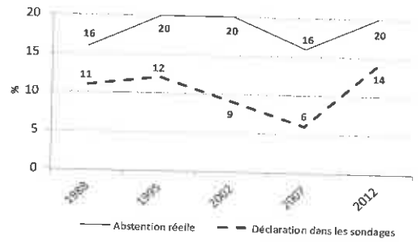
II.3. La désirabilité sociale : un troisième biais de l'enquête par questionnaire

Le second défi est lié aux biais de réponse à certaines questions. Ils sont de plusieurs sortes : la non-réponse partielle, le sentiment d'incompétence ou la désirabilité sociale. Le premier correspond aux individus qui ne répondent pas à toutes les questions du questionnaire, soit parce qu'ils ne savent pas quoi répondre, soit parce qu'ils refusent ponctuellement de répondre, pour des motivations variées (question jugée indiscrète par exemple). Le sentiment d'incompétence quant à lui peut intervenir sur des questions abordant des sujets complexes sur lesquels les personnes interrogées peuvent parfois s'estimer peu légitimes pour répondre. Le soin pris dans la formulation de ladite question sera central pour neutraliser cet écueil.

Concentrons-nous maintenant sur le biais de désirabilité sociale. Elle peut être définie de la manière suivante : un individu cherche généralement à présenter une bonne image de lui-même dans les réponses qu'il donne aux questions posées dans l'enquête par sondage, en particulier lorsque celles-ci sont considérées trop normatives. L'attention portée à la formulation de ces questions est aussi fondamentale pour réduire au maximum ce biais. Néanmoins, un sujet dit « sensible » (par exemple : une enquête sur la consommation de stupéfiants ou la mesure de l'abstention dans un sondage d'intention de vote) peut se révéler difficile à mesurer correctement dans une enquête par sondage.

Prenons l'exemple de l'abstention telle que mesurée dans les sondages post seconds tours des élections présidentielles françaises depuis 1988 et comparons-la à l'abstention réelle telle que fournie par le ministère de l'Intérieur pour chaque scrutin (figure 2).

Figure 2 : Abstention réelle et déclaration dans les sondages, seconds tours des scrutins présidentiels en France depuis 1988



Sources : Enquêtes *postélectorales* du CEVIPOF / ministère de l'Intérieur

L'abstention mesurée dans les sondages postélectoraux est systématiquement sous-estimée (entre -4 et -11 points) : cette tendance liée à la désirabilité sociale indique qu'il s'avère plus délicat (car moins valorisé socialement) de déclarer s'être abstenu qu'être allé voter à un scrutin électoral.

II.4. Comment rectifier l'échantillonnage *a posteriori* ?

Il est tout à fait possible de corriger *a posteriori* des biais intervenant dans la construction de l'échantillon. En effet, nous avons étudié précédemment différentes déformations possibles intervenant lors de l'échantillonnage et menant à des sur ou sous-représentations de certaines catégories de population. Ces déformations peuvent être atténuées : on parle alors de correction ou de redressement apporté à l'échantillon concerné. Des poids sont alors affectés aux unités statistiques constitutives de l'échantillon afin de réduire, le cas échéant, la sur ou sous-représentation initiale, en fonction de certains critères jugés importants. Les unités statistiques sont, avant redressement, toutes affectées d'un poids identique (coefficient=1) ; après redressement elles pèseront plus ou moins fortement dans les statistiques effectuées[1]. De la sorte, les réponses au questionnaire des individus insuffisamment représentés dans l'échantillon seront davantage pris en compte dans les pourcentages (coefficient>1) et inversement (coefficient<1). Attention toutefois, cette procédure n'est valable qu'à la condition d'aboutir à des poids de redressement proches de 1 (et donc déformant de façon marginale l'échantillon initial). Dans le cas contraire, cela signifie que l'échantillon obtenu n'a pas été élaboré dans des conditions de rigueur suffisantes.

Prenons l'exemple du tableau suivant (tableau 5) tiré de l'enquête *European Social Survey* portant sur les personnes résidant sur le territoire français et âgées de 15 ans et plus.

[1] Nous recommandons, dans un premier temps en tous les cas, d'appliquer ces procédures à des fins descriptives (tris à plat, tris croisés) et de les éviter en cas de modélisation (régressions, etc.).

Tableau 5 : Distribution conjointe de l'âge et du sexe dans une enquête par sondage, données brutes

	Hommes	Femmes	Marge observée
15-17 ans	27	21	48
18-24 ans	66	73	139
25-34 ans	91	141	232
35-49 ans	212	238	450
50-64 ans	229	247	476
65 ans et plus	177	206	383
Marge observée	802	926	1728

Sources : *European Social Survey*[1], Vague 5

L'enquête datant de 2010, on peut facilement trouver sur le site de l'INSEE les données du recensement correspondantes (tableau 6) qui fournissent les marges réelles[2] dans la population, c'est-à-dire la répartition réelle des hommes et des femmes par tranche d'âge.

Tableau 6 : Distribution conjointe de l'âge et du sexe dans les données du recensement

	Hommes	Femmes	Marge réelle
15-17 ans	1161378	1107668	2269046
18-24 ans	2781778	2723917	5505695
25-34 ans	3826668	3899428	7726096
35-49 ans	6487588	6613155	13100743
50-64 ans	5875465	6204313	12079778
65 ans et plus	4357972	6181894	10539866
Marge réelle	24490849	26730375	51221224

Sources : INSEE, estimations de population au 1er janvier 2010

Si les caractéristiques des personnes interrogées au cours de la vague 5 de l'*European Social Survey* se distribuaient exactement comme celles du recensement de l'INSEE, on obtiendrait les marges suivantes (appelées marges théoriques ou marges selon l'INSEE dans notre exemple) pour les données ESS (tableau 7).

Pour la ligne « 15-17 ans », la marge selon l'INSEE est obtenue ainsi :

$$77 \approx 2269046 * \frac{1728}{51221224}.$$

De même, pour la colonne « Hommes », la marge selon l'INSEE est obtenue ainsi :

$$826 \approx 24490849 * \frac{1728}{51221224}.$$

De manière générale, la marge théorique est obtenue par produit en croix :

$$Marge\ théorique \approx Marge\ réelle * \frac{Total\ observé}{Total\ population}.$$

[1] <http://www.europeansocialsurvey.org/>, consulté le 12 juin 2014. Il s'agit d'une autre des grandes enquêtes sociologiques internationales dont l'échantillon est constitué selon une procédure aléatoire.

[2] Les marges d'un tableau (ou effectifs marginaux) correspondent aux totaux pour les différentes catégories (en ligne ou en colonne). À ne pas confondre avec la marge d'erreur !

Tableau 7 : Distribution conjointe de l'âge et du sexe dans une enquête par sondage, données brutes, et marges théoriques

	Hommes	Femmes	Marge observée	Marge théorique
15-17 ans	27	21	48	77
18-24 ans	66	73	139	186
25-34 ans	91	141	232	261
35-49 ans	212	238	450	442
50-64 ans	229	247	476	408
65 ans et plus	177	206	383	356
Marge observée	802	926	1728	
Marge théorique	826	902		

Sources : *European Social Survey*, Vague 5 ; INSEE, estimations de population au 1er janvier 2010

Comment prendre en compte la différence entre marge observée et marge théorique ? Des coefficients de pondération, appelés aussi coefficients de calage, peuvent être construits à partir des marges théoriques afin de corriger ce biais. Celles-ci nous informent sur la sous ou la surreprésentation de certaines catégories de population. Ainsi les 15-17 ans sont sous-représentés dans l'échantillon ESS tandis que les femmes sont surreprésentées. Les individus appartenant à une catégorie donnée, par exemple les femmes âgées de 15-17 ans, se verront attribués le même coefficient de pondération. Ici nous avons 12 catégories, qui aboutissent donc à la construction de 12 coefficients de pondération (tableau 8). Pour la catégorie des femmes âgées de 15 à 17 ans, le coefficient 1,55 est obtenu ainsi :

$$1{,}55 \approx \frac{77}{48} \times \frac{902}{926}.$$

De manière générale, le coefficient de pondération est obtenu par :

$$Coefficient\ de\ pondération \approx \frac{Marge\ théorique\ n°1}{Marge\ observée\ n°1} \times \frac{Marge\ théorique\ n°2}{Marge\ observée\ n°2}.$$

Cela signifie que les réponses aux questions de l'enquête des femmes âgées de 15 à 17 ans seront affectées d'un poids de 1,55 car elles sont en sous-effectif dans l'échantillon. De la sorte, leurs réponses compteront davantage. À l'inverse, dans la mesure où les femmes de 65 ans et plus sont trop nombreuses dans l'échantillon au vu de leur représentation dans la population mère, les réponses de ces dernières joueront moins (poids=0,90). Par cette procédure, il s'agit bien de rétablir la juste reproduction (en proportion) de sous-populations particulières dans un échantillon.

Ce raisonnement est facilement généralisable au cas où l'on souhaite redresser sur plus de deux variables.

Tableau 8 : Coefficients de pondération par catégorie

	Hommes	Femmes
15-17 ans	1,64	1,55
18-24 ans	1,38	1,30
25-34 ans	1,16	1,09
35-49 ans	1,01	0,96
50-64 ans	0,88	0,83
65 ans et plus	0,96	0,90

Sources : *European Social Survey*, Vague 5 ; INSEE, estimations de population au 1er janvier 2010

Les coefficients de pondération supérieurs à 1 signifient que les catégories correspondantes étaient sous-représentées dans l'échantillon ESS tandis que ceux inférieurs à 1 correspondent aux catégories qui étaient surreprésentées. Ils permettent ainsi de rétablir l'équilibre et de réduire le biais initial. Qu'en est-il du nouveau tableau issu de ce redressement (tableau 9) ?

Tableau 9 : Distribution conjointe de l'âge et du sexe dans une enquête par sondage, données pondérées par calage sur marges théoriques

	Hommes	Femmes	Marge calée
15-17 ans	44,36	32,61	76,97
18-24 ans	90,86	95,00	185,85
25-34 ans	105,32	154,27	259,59
35-49 ans	214,50	227,64	442,14
50-64 ans	201,98	205,94	407,91
65 ans et plus	169,29	186,25	355,53
Marge calée	826,31	901,69	1728,00

Sources : *European Social Survey*, Vague 5 ; INSEE, estimations de population au 1er janvier 2010

Les nouveaux effectifs croisés sont calculés en pondérant les effectifs bruts par les coefficients définis dans le tableau 8. Ainsi pour la catégorie des femmes âgées de 15 à 17 ans, on obtient :

$$32{,}61 \approx 21 \times \frac{77}{48} \times \frac{902}{926}.$$

De manière générale,

$$Effectif\ pondéré \approx Effectif\ brut \times \frac{Marge\ théorique\ n°1}{Marge\ observée\ n°1} \times \frac{Marge\ théorique\ n°2}{Marge\ observée\ n°2}.$$

Les nouvelles marges calées sont proches des marges théoriques mais pas forcément égales à celles-ci, du fait de la prise en compte de plusieurs variables pour le calcul des coefficients de pondération.

La méthode de redressement présentée dans l'exemple filé ci-dessus s'appelle la stratification *a posteriori*, à ne pas confondre avec la stratification *a priori*. Elle est adaptée à des variables de calage de type qualitatives (ici l'âge en classes et le sexe). D'autres méthodes existent, en particulier pour les variables quantitatives, comme le redressement par le quotient ou par la régression[1].

III. Les exercices d'application

III.1. Logement et travaux d'isolation

Dans une enquête réalisée par téléphone auprès d'un échantillon aléatoire de Franciliens représentatif des résidents d'Ile-de-France âgés de 18 ans et plus, un institut de sondage pose la question suivante en janvier 2008 et en janvier 2009 (données fictives) : « Personnellement, êtes-vous prêt à faire des travaux pour

[1] Pour en savoir plus : Pascal Ardilly, « Chapitre III. Amélioration des estimateurs (redressements, correction de non-réponse) », *Les techniques de sondage*, Paris, Technip, 2006, p. 273-472.

mieux isoler votre logement pour favoriser le développement durable en Ile-de-France ? »[1].

En janvier 2008, 51% des personnes interrogées répondent positivement ; en janvier 2009, elles sont 57%.

En utilisant le tableau de marges d'erreur ci-dessous (tableau 10), peut-on dire que l'évolution entre 2008 et 2009 est significative, ou qu'elle est due à la marge d'erreur :

- pour un échantillon de 400 personnes ?
- pour un échantillon de 800 personnes ?
- pour un échantillon de 1500 personnes ?

Tableau 10 : Marges d'erreur au seuil de confiance de 95%

Taille de l'échantillon	Proportions évaluées en pourcentage									
	5% ou 95%	10% ou 90%	15% ou 85%	20% ou 80%	25% ou 75%	30% ou 70%	35% ou 65%	40% ou 60%	45% ou 55%	50%
100	4,4	6,0	7,1	8,0	8,7	9,2	9,5	9,8	9,9	10,0
300	2,5	3,5	4,1	4,6	5,0	5,3	5,5	5,7	5,8	5,8
400	2,2	3,0	3,6	4,0	4,3	4,6	4,8	4,9	5,0	5,0
500	2,0	2,7	3,2	3,6	3,9	4,1	4,3	4,4	4,5	4,5
800	1,5	2,1	2,5	2,8	3,0	3,2	3,3	3,4	3,5	3,5
1000	1,4	1,9	2,3	2,5	2,8	2,9	3,1	3,1	3,2	3,2
1500	1,1	1,6	1,9	2,1	2,3	2,4	2,5	2,5	2,6	2,6
2000	1	1,3	1,6	1,8	1,9	2,0	2,1	2,2	2,2	2,2
5000	0,6	0,9	1,0	1,1	1,2	1,3	1,4	1,4	1,4	1,4

Sources : calculs des auteurs

III.2. Le vote Front national

Le tableau 11 présente des résultats issus d'une enquête par sondage.

Tableau 11 : Distribution conjointe du sexe et du vote pour Marine Le Pen au premier tour de l'élection présidentielle de 2012, aux inscrits, données brutes

	Hommes	Femmes	Marge observée
Marine Le Pen	154	136	290
Autres candidats	804	966	1770
Abstention, blancs, nuls	200	244	444
Marge observée	1158	1346	2504

Sources : Enquête *post-électorale* CEVIPOF, 2012

On sait par ailleurs que 53% des inscrits sont des femmes[2] et que les résultats réels du vote au premier tour de l'élection présidentielle de 2012 sont les suivants[3] : parmi les inscrits, 14% ont voté pour Marine Le Pen, 64% pour un autre candidat et 22% se sont abstenus ou ont voté blanc ou nul.

Pondérer le tableau 11 en utilisant la technique de post-stratification.

1 Question tirée de l'enquête « Les Franciliens et le développement durable : jusqu'où sont-ils prêts à aller ? » réalisée par l'institut TNS Sofres en novembre 2008.

2 *Sources* : INSEE, fichier électoral 2010 et estimation de population au 31 décembre 2010.

3 *Sources* : ministère de l'Intérieur, 2012.

III.3. Les corrigés

[III.1.] Le tableau 12 résume les marges d'erreur calculées pour les différentes tailles d'échantillon (400, 800 et 1500) et pour les deux dates (2008 et 2009), ainsi que les intervalles qui en résultent[1].

Tableau 12 : Marges d'erreur et intervalles (de confiance) au seuil de 95% pour les trois tailles d'échantillon proposées

Taille de l'échantillon	Proportion		Marge d'erreur		Intervalle (de confiance)	
	2008	2009	2008	2009	2008	2009
400	51%	57%	5,0	5,0	[46%;56%]	[52%;62%]
800			3,5	3,5	[47,5%;54,5%]	[53,5%;60,5%]
1500			2,6	2,6	[48,4%;53,6%]	[54,4%;59,6%]

Sources : simulation des auteurs

Les deux échantillons les plus petits (respectivement 400 et 800) aboutissent dans notre exemple à une zone de chevauchement (respectivement de 4 points avec [52% ; 56%] et de 1 point avec [53,5% ; 54,5%]) : l'évolution de +6 points mesurée entre janvier 2008 et janvier 2009 sur l'opinion sur les travaux d'isolation du logement pour favoriser le développement durable en Ile-de-France n'est donc pas significative au seuil de 95% pour de tels échantillons. En revanche, il n'y a pas de zone de chevauchement due aux marges d'erreur pour un échantillon de 1500 personnes. Pour cet échantillon l'évolution de 51% à 57% est significative au seuil de 95%.

Une même évolution peut donc amener à des conclusions différentes en termes de significativité en fonction de la taille de l'échantillon. Il est à noter que le niveau des pourcentages comparés (51% et 57% versus 11% et 17% par exemple) conduirait également à des conclusions différentes pour certaines tailles d'échantillon (ici 800 où l'évolution serait non significative dans un cas et significative dans l'autre pour un même écart).

Cet exercice permet d'insister encore une fois sur l'importance de préciser les marges d'erreur dans toute publication de données issues de sondage !

1 Se référer au chapitre 4 pour plus de détails sur ces intervalles appelés intervalles de confiance en référence au seuil de confiance considéré, ici 95%.

[III.2.] Les marges théoriques s'obtiennent directement à partir des proportions fournies par l'INSEE pour le sexe et par le ministère de l'Intérieur pour les résultats réels du vote. On obtient ainsi le tableau 13.

Tableau 13 : Distribution conjointe du sexe et du vote pour Marine Le Pen au premier tour de l'élection présidentielle de 2012, aux inscrits, données brutes et marges théoriques

	Hommes	Femmes	Marge observée	Marge théorique
Marine Le Pen	154	136	290	351
Autres candidats	804	966	1770	1603
Abstention, blancs, nuls	200	244	444	551
Marge observée	1158	1346	2504	
Marge théorique	1177	1327		

Sources : Enquête *post-électorale* CEVIPOF, 2012 ; INSEE, fichier électoral 2010 et estimation de population au 31 décembre 2010 ; ministère de l'Intérieur, 2012

Nous constatons une sous-déclaration dans l'enquête post-électorale du vote pour la candidate du Front national et de l'abstention, ainsi qu'une légère surreprésentation des femmes.

On en déduit facilement les coefficients de pondération à appliquer (tableau 14) et les marges calées par ces coefficients (tableau 15).

Tableau 14 : Coefficients de pondération par catégorie

	Hommes	Femmes
Marine Le Pen	1,23	1,19
Autres candidats	0,92	0,89
Abstention, blancs, nuls	1,26	1,22

Sources : Enquête *post-électorale* CEVIPOF, 2012 ; INSEE, fichier électoral 2010 et estimation de population au 31 décembre 2010 ; ministère de l'Intérieur, 2012

Tableau 15 : Distribution conjointe du sexe et du vote pour Marine Le Pen au premier tour de l'élection présidentielle de 2012, aux inscrits, données pondérées par calage sur marges théoriques

	Hommes	Femmes	Marge calée
Marine Le Pen	189,19	162,09	351,29
Autres candidats	739,81	862,35	1602,16
Abstention, blancs, nuls	252,19	298,49	550,68
Marge calée	1181,20	1322,93	2504,13

Sources : Enquête *post-électorale* CEVIPOF, 2012 ; INSEE, fichier électoral 2010 et estimation de population au 31 décembre 2010 ; ministère de l'Intérieur, 2012

IV. Les articles d'application conseillés

- Nicole Dubois, Élise Aubert, « Valeur sociale des personnes : deux informations valent-elles mieux qu'une ? », *Revue internationale de psychologie sociale*, vol. 23, n°1, 2010, p. 57-92.
- Viviane Le Hay, « Chapitre 11 / Le panel électoral français 2007. Enjeux de méthode », Bruno Cautrès, Anne Muxel (dir.), *Comment les électeurs font-ils leur choix ? Le panel électoral français 2007*, Paris, Presses de Sciences Po, 2009, p. 259-284.
- Lionel Prouteau, François-Charles Wolff, « Adhésions et dons aux associations : permanence et évolutions de 2002 à 2010 », *Économie et statistique*, n°459, 2013, p. 27-57.