Nature et déterminants de l'erreur
d'échantillonnage dans les enquêtes par sondage

Pascal Ardilly
Insee, direction de la méthodologie

Lorsqu'on cherche à estimer une grandeur relative à une population à partir de données collectées seulement sur un échantillon, on s'expose à une erreur d'échantillonnage. Cette erreur a deux composantes : le biais, qui existe dès lors que la moyenne des estimations issues de tous les échantillons possibles diffère de la grandeur visée, et la variance, qui traduit une forme d'instabilité des estimations. Les sondages probabilistes permettent d'estimer l'erreur d'échantillonnage. Les sondages empiriques, par construction, ne le permettent pas. Il est entré dans l'usage de publier pour les sondages empiriques des estimations de l'erreur d'échantillonnage obtenues en « faisant comme si » il s'agissait de sondages probabilistes. C'est une pratique défendable dans certaines circonstances, mais un professionnalisme élevé est nécessaire pour la justifier.

Les statisticiens d'enquête ont pour mission de mettre en œuvre des méthodes permettant de connaître « au mieux » des grandeurs définies sur une population. Ces grandeurs sont le plus souvent des moyennes, des totaux ou des proportions. Par exemple, juste avant une élection, on souhaite connaître la proportion d'électeurs qui vont voter pour un candidat donné. On peut aussi s'intéresser chaque mois au nombre de personnes en recherche d'emploi, ou au chiffre d'affaire annuel moyen des entreprises innovantes. En situation idéale, les valeurs exactes de ces grandeurs – que l'on appelle des paramètres – peuvent être obtenues par recensement, c'est-à-dire par une collecte exhaustive des informations individuelles formant le paramètre. Mais le plus souvent, cette option est trop onéreuse et irréalisable en pratique: on doit alors se tourner vers la solution alternative, beaucoup moins lourde, qu'est l'enquête par sondage, c'est-à-dire une collecte de données auprès d'une partie seulement de la population (l'échantillon).

Toute enquête par sondage se construit en distinguant deux étapes essentielles : la sélection de l'échantillon d'une part et le choix de la méthode d'estimation d'autre part, c'est-à-dire le processus de traitement des données collectées qui va fournir une valeur supposée proche du paramètre inconnu – cette valeur est appelée « estimation du paramètre ». L'économie de moyens qu'offre le sondage a néanmoins une contrepartie qui est l'erreur d'échantillonnage, autrement dit un inévitable écart entre le paramètre et son estimation, laquelle va d'ailleurs dépendre de l'échantillon tiré. En fin d'opération, il est souhaitable d'apprécier la qualité de l'estimation produite en proposant une estimation de l'erreur d'échantillonnage.

La théorie des sondages offre aujourd'hui les moyens de quantifier l'erreur d'échantillonnage dans toutes les circonstances créées par la combinaison d'une méthode d'échantillonnage et d'une méthode d'estimation. Curieusement, alors que les outils probabilistes ont quelques siècles d'existence, l'approche spécifique aux sondages – une forme d'extrapolation pourtant bien naturelle en population finie – a créé une discipline nouvelle qui a connu un développement

flamboyant après la dernière guerre, sous l'impulsion de pères fondateurs parmi lesquels on compte Jerzy Neyman, Morris Hansen ou William Cochran.

Les deux composantes de l'erreur d'échantillonnage

L'erreur d'échantillonnage s'apprécie au travers de deux composantes: le biais et la variance de l'estimation (en toute rigueur, le terme « estimateur » serait plus correct, mais nous adopterons ici la terminologie la plus commune). Pour illustrer ces concepts, partons du problème posé par l'estimation de la proportion de votants pour un candidat donné à des élections présidentielles. Avant le scrutin, c'est évidemment un véritable enjeu que de prédire le nom du candidat qui l'emportera ! Il faut comprendre que la proportion exacte que l'on cherche à approcher au mieux et que l'on notera désormais P_{vrai}, est une proportion certes inconnue mais que l'on pourrait obtenir de façon exacte si on disposait de moyens suffisants pour questionner l'ensemble des électeurs …et sous condition évidemment que ceux-ci ne changent pas d'avis entre la date du sondage et celle de l'élection proprement dite. Au demeurant, le risque causé par un éventuel changement d'opinion n'est pas à mettre au compte de l'erreur d'échantillonnage, c'est là un tout autre phénomène qui tient au fait qu'un paramètre – comme une population d'ailleurs – évolue avec le temps.

Le biais …

Considérons maintenant un axe sur lequel on place la valeur de P_{vrai}, et imaginons tous les échantillons possibles que l'on peut construire à partir de la population complète (le nombre de ces échantillons est gigantesque dès que la population dépasse quelques dizaines d'individus) : chaque échantillon donne lieu à une estimation de proportion qui lui est propre et qui se positionne sur l'axe au moyen, mettons, d'une petite croix. Il est clair que toutes ces petites croix se répartissent entre les valeurs zéro et un : en effet, certains échantillons sont composés uniquement de partisans du candidat-cible (estimation égale à 100%, soit 1) et certains échantillons, tout à fait à l'opposé, ne rassembleront que des opposants au candidat – l'estimation de la proportion étant alors égale à 0%. La théorie dit que la probabilité de tomber « par hasard » sur un de ces échantillons extrêmes est très faible …mais elle n'est pas nulle ! Il est par ailleurs bien clair que plus l'échantillon est de grande taille, moins on a de chance de rencontrer une situation aussi atypique. Entre ces valeurs extrêmes, compte tenu des très grandes tailles de population auxquelles on a à faire, on trouve un quasi-continuum de situations : on se convainc facilement que, quelle que soit une estimation donnée a priori et comprise entre 0 et 100 %, il existe toujours (au moins) un échantillon qui conduit à une estimation qui en est extrêmement proche.

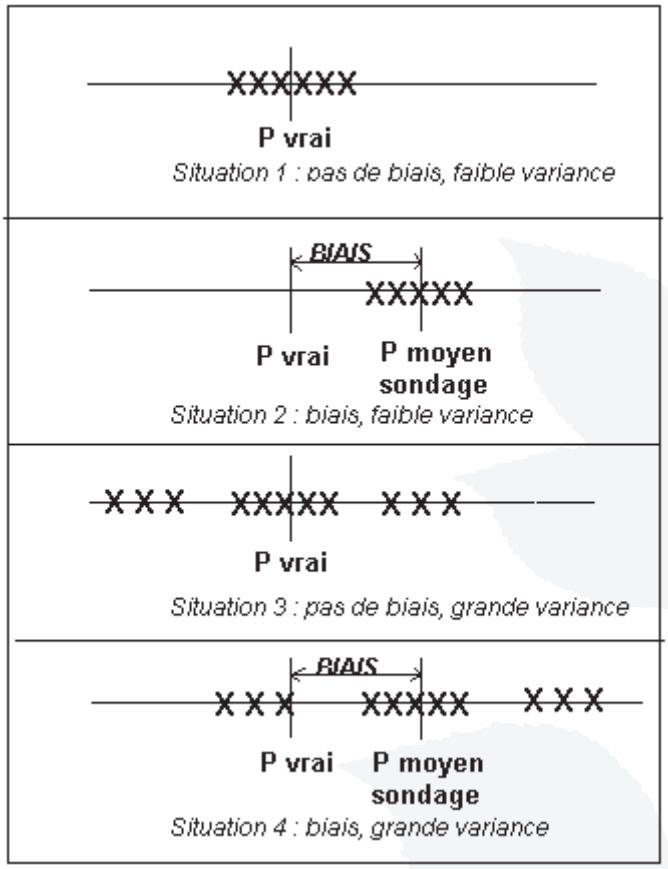
Par définition, un biais existe dès lors que la moyenne des estimations, formée à partir de l'ensemble de tous les échantillons que l'on peut constituer dans la population, diffère du paramètre: graphiquement, c'est le cas lorsque la moyenne de toutes les petites croix – on va l'appeler $P_{sondage}^{moyen}$ – n'est pas égale à P_{vrai}. Dans le cas contraire, on a un plan de sondage sans biais. Lorsqu'on parle de moyenne des estimations, il faut comprendre que la moyenne en question est pondérée par les probabilités de sélectionner les échantillons considérés (techniquement, on parle « d'espérance mathématique »). Le schéma des petites croix est certes pratique pour comprendre la notion de biais mais cette approche purement graphique se justifie plus particulièrement lorsque tous les échantillons que l'on peut construire ont la même probabilité d'être sélectionnés. Très souvent, ce n'est pas le cas car on privilégie certains échantillons par rapport à d'autres. C'est alors comme si notre axe constituait le fléau d'une balance articulé au niveau de la vraie proportion P_{vrai} et que chaque petite croix positionnée sur ce fléau avait un poids égal à la probabilité de tirer l'échantillon auquel elle est associée. Si le plan de sondage est sans biais, la balance est équilibrée; dans le cas contraire le fléau va pencher, à droite s'il y a

surestimation, à gauche s'il y a sous-estimation. Les sondeurs ont généralement de l'aversion à mettre en place des plans de sondage biaisés – ce qui est néanmoins inévitable lorsqu'il y a de la non-réponse parce que les mécanismes probabilistes de la non-réponse ne sont ni contrôlés ni connus par le sondeur.

...et la variance

La variance est définie comme la moyenne des carrés des écarts entre les estimations possibles et leur espérance mathématique. Techniquement, à l'image exacte de ce qui se fait pour définir l'espérance, la moyenne des carrés est pondérée par les probabilités de tirer les différents échantillons que l'on pourrait sélectionner. La variance traduit une forme d'instabilité des estimations (ce que le biais ne permet pas du tout d'apprécier): si l'estimation est numériquement très sensible à l'échantillon tiré, il y aura une grande variance. Reprenant notre approche graphique, on est confronté à une grande variance lorsque les croix sont très étalées le long de l'axe, et à une petite variance si elles sont concentrées à un endroit quelconque de l'axe (qui n'est pas nécessairement égal à P_{vrai}, ni même proche de ce paramètre – cet aspect renvoie à la notion de biais). Mais là encore, la réalité va un peu au-delà de ce que traduit le graphique : si certaines petites croix sont très éloignées de la vraie valeur P_{vrai} mais que les échantillons qui les créent ont extrêmement peu de chance d'être sélectionnés, alors la variance va rester faible. C'est là que l'on perçoit la faiblesse de ces approches « en moyenne » : un « mauvais » hasard, certes peu probable mais pas impossible à rencontrer, peut conduire à une estimation extrême sans que la variance n'en soit vraiment affectée. Si l'estimation est grossièrement excessive, contraire au sens commun, on adoptera évidemment des méthodes de correction, éventuellement on refera l'enquête mais un échantillon simplement « un peu trop » déséquilibré ne sera pas forcément détecté par un indicateur comme la variance. Des techniques spécifiques dites de redressement permettent de réduire sensiblement ce risque, mais ceci est une autre histoire...

Le biais et la variance, considérés conjointement, constituent une information très importante qui permet d'apprécier l'erreur d'échantillonnage, constituant elle-même une composante de la qualité d'une enquête par sondage. L'ampleur du biais ne préjuge pas de celle de la variance, si bien qu'on peut trouver en pratique des plans de sondage de toutes sortes, que l'on classerait schématiquement en quatre catégories (voir graphique 1).



Graphique 1 : Les quatre situations dans lesquelles un plan de sondage peut se trouver

La situation 1 est la meilleure et la situation 4 est la pire. Les expressions mathématiques du biais et de la variance dépendent à la fois de la méthode d'échantillonnage et de la procédure d'estimation. S'ajoute l'impact – toujours « négatif » – du phénomène de non-réponse, qui génère systématiquement une dégradation à la fois du biais et de la variance (la non-réponse diminue la taille de l'échantillon, donc augmente la variance). Dans tous les cas, dans l'approche de la théorie classique des sondages, il est nécessaire de connaître le contexte probabiliste qui préside au tirage des individus de l'échantillon. Autrement dit, il faut être capable de dire avec quelle probabilité un individu a été échantillonné (on parle de « probabilité de sélection »). Dans le cas contraire, on ne peut pas estimer l'erreur d'échantillonnage de manière rigoureuse. Il faut également faire des hypothèses de comportement pour traiter la non-réponse et comme il s'agit justement d'un processus que l'on ne contrôle pas du point de vue probabiliste, le biais est inévitable et le calcul de variance est nécessairement entaché d'erreur.

La problématique des intervalles de confiance

La variance n'est pas le concept le plus abouti en matière d'erreur d'échantillonnage : en effet, le plus souvent, on diffuse des intervalles de confiance pour juger de la qualité d'une estimation. Dans l'approche classique, l'intervalle de confiance à x% (classiquement 95%) est constitué par une limite inférieure et une limite supérieure, calculées de telle sorte qu'il y ait x chances sur 100 (classiquement 95 chances sur 100) pour que le paramètre recherché (P_{vrai} dans notre exemple) soit compris entre ces limites. Les bornes de cet intervalle sont aléatoires car elles dépendent de l'échantillon tiré. La théorie de l'intervalle de confiance fait appel à quelques hypothèses techniques dont voici les deux principales : premièrement, la taille d'échantillon doit être « assez grande » (ce qui est généralement satisfait) et deuxièmement, l'estimation doit être sans biais – ce qui permet de centrer l'intervalle sur l'estimation du paramètre auquel

conduit l'échantillon sélectionné. La seconde condition nous intéresse tout particulièrement car nous verrons que les estimations des sondages dits « empiriques » sont en toute rigueur biaisées. Sans le justifier davantage, signalons que l'intervalle de confiance dit « à 95% » se construit en général ainsi: la limite inférieure est égale à l'estimation issue de l'enquête moins deux fois la racine carrée de la variance, et la limite supérieure est égale à l'estimation issue de l'enquête plus deux fois la racine carrée de la variance.

La perception du sens de l'intervalle de confiance est souvent erronée: il ne faut surtout pas imaginer que le vrai paramètre P_{vrai} se trouve a priori « quelque part » de manière uniforme au sein de cet intervalle. Derrière cette notion, il y a des éléments probabilistes qui montrent que P_{vrai} se trouve plus probablement plus près du centre de l'intervalle que de ses extrémités: plus on se rapproche du centre de l'intervalle, plus on a de chances d'y trouver P_{vrai} (techniquement, l'estimation suit une loi probabiliste bien connue appelée loi de Gauss).

Par ailleurs, on peut discuter longtemps sur le choix de la probabilité de 95%, qui est extrêmement courant en pratique, on pourrait même dire systématique, mais néanmoins largement conventionnel : en effet, on pourrait tout aussi bien opter pour des intervalles de confiance associés à une autre probabilité. C'est évidemment un moyen d'influer – de manière plus ou moins insidieuse il est vrai – sur l'amplitude de l'intervalle de confiance. Ainsi, l'intervalle de confiance à 90% sera plus étroit que l'intervalle à 95%, ce qui donnera certes une apparence de meilleure précision mais par définition on augmente le risque d'une mauvaise prévision du résultat des élections. De ce point de vue, les sondages électoraux ont la spécificité d'être confrontés ultérieurement à la vraie valeur (au – sérieux – problème près des changements d'opinion…) et cet aspect est essentiel en terme de communication. A l'opposé, pour limiter les risques de mauvaise prévision, on peut construire un intervalle de confiance couvrant la vraie valeur avec plus de 95 chances sur 100 – mais l'intervalle va s'élargir: si on opte pour un intervalle de confiance à 99%, les limites de l'intervalle sont construites en calculant plus ou moins 2,6 fois la racine carrée de la variance (au lieu de 2): la largeur de l'intervalle augmente donc de 30% par un simple effet d'affichage !

Sondages probabilistes versus sondages empiriques et calculs d'erreur d'échantillonnage

Si on s'intéresse à la phase d'échantillonnage, on distingue traditionnellement deux classes de méthodes: d'une part les échantillonnages probabilistes, d'autre part les échantillonnages empiriques – la méthode empirique la plus connue étant la méthode des quotas.

Les sondages probabilistes ont pour caractéristique de permettre le calcul de la probabilité de sélection de chaque individu de la population couverte par l'enquête. L'échantillonnage probabiliste relève en effet de règles de sélection extrêmement précises dans une population au sein de laquelle chaque individu est clairement identifié. Une méthode de tirage totalement objective (un « algorithme » mathématique), sans aucune intervention humaine, permet d'associer à chaque individu de la population une probabilité connue d'être sélectionné. Pratiquement, c'est un programme informatique qui tire l'échantillon à partir d'un fichier informatique qui recense l'intégralité de la population (on parle de base de sondage).

A l'opposé, l'échantillonnage des sondages empiriques ne peut pas donner lieu à un calcul rigoureux des probabilités de sélection des individus enquêtés. L'échantillonnage empirique – disons de type quotas – relève en effet d'une sélection non totalement contrôlée, offrant une composante subjective parce que la sélection est concrètement réalisée par l'enquêteur en fonction des circonstances qu'il rencontre. Ce dernier dispose bien entendu de consignes pour limiter cette part d'appréciation subjective au moment de l'échantillonnage mais il conserve

– inévitablement – une liberté qui ne permet pas de maitriser les aspects probabilistes de la sélection. La qualité d'un échantillonnage empirique est donc plus difficile à apprécier avec les outils qu'offre la théorie classique des sondages: de fait, ils n'autorisent pas de mesure formelle de la qualité à partir des notions classiques de biais et de variance d'échantillonnage, ce qui constitue une critique récurrente des méthodes empiriques (qui ont par ailleurs des atouts d'une autre nature).

Les plans de sondage probabilistes permettent des estimations en théorie sans biais mais en pratique et *in fine*, les estimations sont toujours biaisées parce qu'on est systématiquement confronté à un phénomène de non-réponse. Par ailleurs, s'il y a une seule règle statistique à énoncer ici, c'est celle qui traduit la diminution de la variance lorsque la taille de l'échantillon augmente : avec un tirage équiprobable d'individus (tirage « totalement au hasard ») et sans non-réponse, l'intervalle de confiance a une largeur qui varie (presque…) rigoureusement comme l'inverse de la racine carrée de la taille de l'échantillon: pour diviser par 2 (respectivement par 3) la largeur de l'intervalle, il faut donc un échantillon 4 fois plus gros (respectivement 9 fois plus gros). Si l'échantillonnage probabiliste est d'une autre nature (par exemple si on tire les individus avec des probabilités de sélection inégales) et si on prend en compte en sus la non-réponse, la règle doit être adaptée sur le plan mathématique mais en toute circonstance on retrouvera ce principe majeur et universel: grosso modo, quel que soit le plan de sondage, l'ordre de grandeur de la largeur de l'intervalle de confiance sera fonction de l'inverse de la racine carrée de la taille de l'échantillon répondant.

Quant aux plans de sondage empiriques visant une forme de « représentativité », c'est-à-dire en pratique correspondant aux enquêtes par quotas, le biais dépend très largement de la relation qui existe, au sein de chacune des sous-populations déterminées par les quotas, entre la variable qui permet de définir le paramètre (pour un sondage électoral, c'est une variable qui vaut 1 si l'individu déclare voter pour le candidat-cible, et 0 sinon) et la probabilité de sélection de l'individu par l'enquêteur.

L'existence de ce biais de sélection est assez intuitive et nous donnons ici un exemple un peu caricatural mais néanmoins éclairant : si on effectue une enquête sur l'emploi du temps par une méthode empirique, il y a un vrai risque de surestimation du temps moyen d'inactivité des individus. En effet, l'enquêteur qui n'a pas de consignes appropriées va plus facilement contacter des personnes qui sont plus fréquemment présentes à leur domicile, sauf à planifier de son propre chef un improbable passage sur le terrain à des horaires tardifs. L'échantillon sera probablement (plus ou moins) déséquilibré avec une sous-représentation des personnes qui travaillent beaucoup: il y a là, par nature, une corrélation positive génératrice de biais entre la probabilité de sélection et la variable « temps d'inactivité ». Ce type de biais a en particulier la très désagréable propriété de ne pas diminuer lorsque la taille de l'échantillon augmente. C'est pourquoi le sondage empirique, qui a par ailleurs des vertus indéniables dans certaines circonstances, est particulièrement sensible aux conditions de collecte et aux consignes données aux enquêteurs…c'est-à-dire au professionnalisme de la structure qui les gère !

Quant à la variance, la théorie classique des sondages ne permet pas de mener un calcul rigoureux en contexte empirique puisqu'on ne maitrise pas les aspects probabilistes de l'échantillonnage. Très succinctement, sur le plan qualitatif, on résumera le contexte en disant que les contraintes imposées sur la structure de l'échantillon par les quotas constituent un élément qui limite à l'évidence l'ampleur de la variance – mais d'une façon que l'on ne peut pas formaliser dans l'approche classique (il y a moyen d'y échapper en modélisant le comportement des individus, mais on entre sur un terrain d'une toute autre nature). Cette question doit d'ailleurs être rapprochée des possibilités offertes par les plans de sondage probabilistes utilisant des informations externes et cela nous renvoie également à d'autres développements.

En pratique, on trouve des intervalles de confiance publiés à l'occasion de sondages empiriques, électoraux ou portant sur d'autres thèmes. Il faut être clair sur le sens de ces calculs: d'une part ils négligent le biais, d'autre part ils résultent nécessairement d'une assimilation de l'échantillonnage pratiqué à un échantillonnage probabiliste. L'auteur de ces lignes peut accepter cette façon de procéder dans certaines circonstances, mais considère qu'elle reste largement conditionnée à la maitrise technique et opérationnelle du processus d'enquête, laquelle est par nature très dépendante du degré de professionnalisme de la structure qui en a la responsabilité.

Référence

Ardilly, P., Les techniques de sondage, Paris, Technip, 2006.