Pour une plus grande transparence
sur la méthodologie des sondages électoraux

Jeanne Fine
Statisticienne, Toulouse

Si les sondages électoraux pratiqués par les Instituts de sondage sont scientifiquement fondés, il est légitime de demander une plus grande transparence sur la méthodologie adoptée. Les deux points de méthodologie sur lesquels nous souhaitons insister ici sont la « marge d'erreur » et le « redressement d'échantillon ». Pour mieux comprendre les débats les concernant, nous les insérons dans une présentation des écarts entre la théorie des sondages aléatoires et la pratique des sondages électoraux par quotas.

Une proposition de loi pour une plus grande transparence

A l'initiative d'Hugues Portelli (sénateur UMP) et de Jean-Pierre Sueur (sénateur PS), une Proposition de loi sur les sondages visant à mieux garantir la sincérité du débat politique et électoral, a été adoptée le 14 février 2011 à l'unanimité des sénateurs, ce qui est suffisamment rare pour être souligné (Portelli et Sueur, 2011). Un résumé des dispositions relatives à la transparence est donné en annexe. Il s'agit des derniers paragraphes de la présentation des quinze recommandations du rapport d'information au Sénat présenté en octobre 2010 (Portelli et Sueur, 2010). La Commission des lois de l'Assemblée Nationale a fortement amendé cette proposition de loi, en particulier pour la restreindre aux sondages électoraux. Le rapport de la discussion à l'Assemblée Nationale d'Étienne Blanc (enregistré le 1er juin 2011) est donné en référence (Blanc, 2011). Deux ans plus tard, la proposition de loi ainsi amendée n'a toujours pas été présentée à l'Assemblée !

La législation actuelle sur la publication et diffusion des sondages électoraux date de 1977 et une « commission des sondages », autorité administrative indépendante, a été créée pour contrôler le respect de la loi. On pourra consulter cette loi, modifiée en 2002, sur le site de la commission des sondages donné en référence. Pour davantage d'informations sur le droit des sondages, en France et à l'étranger, on renvoie à l'article de Romain Rambaud dans ce numéro et à son site ; voir en particulier le texte de son intervention à l'IFOP le 2 juillet 2013 (Rambaud, 2013).

Pour résumer les positions, la demande d'une plus grande transparence de la proposition de loi est soutenue par les statisticiens mais combattue par la plupart des Instituts de sondage qui préfèrent une autorégulation à une nouvelle loi ! Pour les statisticiens, si la méthode utilisée est scientifique, la méthodologie doit être clairement publiée et les résultats doivent pouvoir être vérifiés par la communauté scientifique. Si ce n'est pas le cas, il faut tout aussi clairement annoncer qu'il s'agit d'avis de politologues fondés sur des sources diverses et personnelles et que les résultats n'ont pas de fondement statistique.

La proposition de loi votée au Sénat intervenait dans le contexte d'un rapprochement entre

statisticiens et sondeurs. On pourra consulter, sur le site de la Société Française de Statistique (SFdS), la page du groupe « Enquêtes et modèles » concernant les débats méthodologiques des sondages électoraux ; cf. en particulier le compte-rendu du « séminaire sur le bon usage des sondages d'intentions de vote » organisé avec le CEVIPOF en mars 2012 (SFdS, Groupe Enquêtes et Modèles, 2012). Nous n'en sommes plus aux débats houleux qui se sont tenus autour d'une table ronde sur les sondages électoraux lors du « Colloque francophone sur les sondages » en novembre 2007 à Marseille !

La réglementation des sondages électoraux concerne les sondages ayant un rapport direct ou indirect avec une élection réglementée par le code électoral ainsi qu'avec l'élection des représentants au Parlement européen. Notons que ces sondages d'intention de vote, lorsqu'ils sont réalisés quelques jours avant une élection et bien qu'ils se présentent comme un « état de l'opinion », donnent une « prévision » qui pourra être comparée au résultat officiel. Ce n'est pas le cas des sondages de notoriété politique ou des sondages d'opinion dont il n'est jamais possible d'évaluer la qualité !

Suite au vote de la proposition de loi au Sénat, le journal Le Monde publie le 9 mars 2011 un article de Roland Cayrol intitulé *Il est inutile de légiférer sur la question des sondages, la suspicion contre les professionnels est injuste* dans lequel le sondeur appelle à une « réflexion adulte sur la publication des sondages ». Dans son édition du 17 mars 2011, *Le Monde* publie un article d'Hugues Portelli et Jean-Pierre Sueur répondant point par point aux arguments de Roland Cayrol et, en réponse à sa proposition d'autorégulation, *Le Monde* propose de publier un sondage d'intentions de vote et d'en faire un décryptage aussi complet que possible, article intitulé *Les secrets de fabrique d'un sondage politique*, réalisé par Ipsos et commenté par Brice Teinturier.

Pour mieux comprendre notre demande d'une plus grande transparence sur la méthodologie des sondages électoraux, en particulier sur la « marge d'erreur » et le « redressement d'échantillon », nous insérons les éléments des débats les concernant dans une présentation des écarts entre la théorie des sondages aléatoires et la pratique des sondages par quotas. Nous nous référerons au sondage du *Monde* du 17 mars 2011 pour illustrer notre propos. Pour ceux que les développements plus techniques ne rebutent pas, nous renvoyons à Fine (2007) dans lequel nous décryptons deux articles publiés dans *Le Monde* suite au premier tour des présidentielles de 2002 et, directement en lien avec le sujet qui nous intéresse ici, à Riandey (2012) et à Ardilly (2010). Cette dernière référence est une annexe au rapport d'information sur les sondages au Sénat intitulée *des éléments sur la précision des sondages* et rédigée par un statisticien de l'INSEE spécialiste de la théorie et de la pratique des enquêtes par sondage.

Méthode aléatoire et méthode des quotas

L'association Pénombre propose réflexions et échanges sur l'usage du nombre dans le débat public, par des articles accessibles sur leur site et par des rencontres que les adhérents organisent. Dans le document préparatoire à la 12ème rencontre « Nocturne », du 5 juin 2009, sur le thème : *Les sondages d'opinion, outils de mesure à utiliser avec précaution ?* (Pénombre, 2009), Alain Tripier, praticien des sondages, écrit (p.14) :

> La théorie des probabilités nous enseigne que pour connaître les événements qui peuvent survenir dans une population donnée, il est possible de n'interviewer qu'une petite partie de celle-ci, à condition de respecter des règles de sélection rigoureuses garantes de sa représentativité.
>
> Un premier choix cornélien se pose alors au praticien entre :
> • la méthode aléatoire (ou probabiliste) préférée par les anglo-saxons ;
> • la méthode dite des « quotas » quasi hégémonique en France.

Nous sommes en désaccord avec cette présentation car la théorie des probabilités ne traite que de la méthode aléatoire. La méthode des quotas est une méthode empirique qui n'a pas de fondement théorique. Si elle est utilisée en France, alors que la méthode aléatoire est préférée par les anglo-saxons, c'est peut-être parce que la formation en probabilités et statistique a été trop longtemps négligée en France au contraire des pays anglo-saxons !

Précisons auparavant le vocabulaire. Nous n'avons en français qu'un seul mot « sondages » pour désigner les « sondages d'opinion » (*poll* en anglais, *sondeo* en espagnol) et la « théorie des sondages » (*sampling* en anglais, *muestreo* en espagnol). Les sondages d'opinion sont généralement réalisés par la « *méthode des quotas* » qui consiste à construire un « échantillon représentatif » de la population selon quelques variables de quotas : sexe, tranches d'âge…, c'est-à-dire, un échantillon respectant les « proportions » des différentes catégories des variables de quotas de la population. La « théorie des sondages » repose sur *l'échantillonnage aléatoire* et l'estimation en population finie ; elle a des fondements probabilistes. Le recueil de données peut être réalisé avec ou sans questionnaire, auprès d'une population humaine, d'une population de ménages, d'entreprises, d'objets… Pour le statisticien, tout échantillon aléatoire dûment contrôlé est « représentatif » de la population dans la mesure où il permet d'estimer les paramètres inconnus de la population et d'en estimer la précision. Le *sondage aléatoire simple à probabilités égales* est le plus simple et sert de référence pour mesurer les améliorations apportées par des échantillons aléatoires plus complexes.

Présentation de la méthode aléatoire

Sondage aléatoire simple

Dans le cadre d'un sondage d'intention de votes, un sondage aléatoire à probabilités égales de taille 1000 consisterait à établir la *base de sondage*, c'est-à-dire, la liste de la population adulte française inscrite sur les listes électorales, et à utiliser un générateur de nombres aléatoires pour sélectionner 1000 personnes de cette base de sondage. La sélection doit être réalisée *selon une procédure aléatoire dûment contrôlée* qui garantit que tous les échantillons de taille 1000 ont la même probabilité d'être tirés. Il faut arriver à contacter ces 1000 personnes sélectionnées, il faut qu'elles acceptent de répondre au questionnaire, qu'elles comprennent les questions posées, qu'elles aient une opinion sur ces questions, qu'elles répondent de façon sincère.

Marge d'incertitude

Lorsque toutes ces conditions sont réunies, au niveau de confiance de 95%, c'est-à-dire, avec une probabilité de se tromper de 5%, on peut considérer que la proportion qui nous intéresse sur la population (proportion d'intention de vote pour tel candidat par exemple) peut être estimée à partir de la proportion p observée sur l'échantillon à 3% près (c'est-à-dire, p plus ou moins 3%).
C'est à tort que cette marge d'incertitude de 3% est appelée « marge d'erreur ». Il ne s'agit pas d'une erreur de calcul mais de tenir compte de la fluctuation naturelle d'une proportion calculée sur un échantillon de taille 1000 : elle fluctue d'un échantillon de taille 1000 à un autre échantillon de taille 1000. Pour un échantillon de taille n, la marge d'incertitude (au niveau de confiance de 95%) est égale à l'inverse de la racine carrée de n (10% pour un échantillon de taille 100, 5% pour un échantillon de taille 400, 3% pour un échantillon de taille 1000 et 1% pour un échantillon de taille 10 000). Cette règle simple, enseignée aujourd'hui en classe de seconde, est utilisée pour des proportions comprises entre 20% et 80%. Il y a une diminution de l'incertitude lorsque les proportions sont plus proches de 0 ou de 1 mais cette marge de 3% est un bon repère pour les résultats du sondage se rapportant à *l'ensemble de la population*.

Lorsque les résultats concernent *une sous-population*, il faut consulter la taille de l'échantillon correspondant afin de réévaluer la marge d'incertitude (5% pour un échantillon de taille 400, 10% pour un échantillon de taille 100).

Il est important de noter que, lorsque le taux de sondage (égal au rapport de la taille de l'échantillon sur la taille de la population) est faible, ce qui est le cas ici, la marge d'incertitude ne dépend que de la taille *n* de l'échantillon et non du taux de sondage.

Avec un tel échantillon aléatoire à probabilités égales de taille 1000, on obtient également les proportions d'hommes, de femmes, des différentes tranches d'âge… à 3% près. Il s'agit donc d'un échantillon à peu près « proportionnel » à la population selon toutes les catégories, connues ou non, c'est-à-dire d'un échantillon « représentatif » de la population selon toutes les variables, connues ou non. C'est l'immense efficacité du sondage aléatoire : **le hasard assure une représentativité auquel aucun expert ne peut prétendre en s'affranchissant d'une sélection aléatoire.**

Redressement

Si l'on connaît par ailleurs avec certitude la répartition de la population selon les tranches d'âge et que ces proportions ne sont pas exactement respectées dans l'échantillon, on peut « redresser l'échantillon » en majorant le poids des individus des tranches sous-représentées au détriment de celles surreprésentées. Ce *redressement* améliore les estimations des proportions d'intention de vote sans pour autant changer fondamentalement la marge d'incertitude. Notons que c'est l'échantillon qui est redressé avec un système de pondérations, ce qui influe évidemment sur les estimations. Mais le redressement n'est justifié théoriquement que si les pondérations s'appuient sur une connaissance « précise » des informations utilisées concernant la population.

Sondage aléatoire stratifié

Si l'information concernant ces variables (sexe, âge…) est indiquée dans la base de sondage *pour chaque individu de la population*, il est possible de réaliser un *sondage aléatoire stratifié proportionnel* selon ces variables. Cela consiste à tirer un échantillon aléatoire simple dans chaque catégorie avec des tailles d'échantillons proportionnelles aux tailles des catégories dans la population. Là encore, on pourra noter une amélioration des estimations des proportions d'intention de vote sans que les marges d'incertitude soient vraiment réduites.

Biais

Nous n'avons explicitement parlé jusqu'ici que de l'incertitude due à l'échantillonnage. Plus sérieux est le problème d'estimations « biaisées » dues, par exemple, à une base de sondage inadéquate ou à un grand nombre de « non-réponses » : aussi grande que soit la taille de l'échantillon, les estimations se concentreront autour de valeurs éloignées des proportions de la population de référence que l'on cherche à estimer. Par exemple, un échantillon de personnes répondant à une enquête sur Internet n'est représentatif que de la population d'usagers d'Internet prenant le temps de répondre à ce type d'enquête ! Le problème de non-réponses fait l'objet de nombreuses recherches en théorie statistique des sondages aléatoires afin d'en atténuer les conséquences négatives.

Ce problème de non-réponses ne dépend pas de l'échantillonnage, il concerne aussi les enquêtes exhaustives. Prenons un exemple fictif : dans un lycée, 40% des élèves de terminales ont déjà consommé de la marijuana mais, lors d'une enquête à ce sujet, les non-consommateurs répondent bien ne pas en avoir consommé mais seulement 10% des consommateurs répondent

en avoir consommé, les autres refusant de répondre. La proportion de consommateurs issue de l'enquête est alors de 6.25% alors que la proportion réelle est de 40%. Le problème de réponses « non sincères » ne dépend pas non plus de l'échantillonnage. Si, dans notre exemple, les 90% de consommateurs qui n'ont pas répondu, répondent en fait qu'ils n'ont pas consommé de marijuana, la proportion de consommateurs issue de l'enquête sera de 4%.

Présentation de la méthode des quotas

La méthode des quotas est une méthode empirique qui n'a pas de fondement théorique. La méthode des quotas est un ersatz de *sondage aléatoire stratifié proportionnel*. La seule contrainte de la méthode des quotas est de construire un échantillon "proportionnel" pour quelques variables de quotas ; pour le sondage du Monde du 17 mars : sexe, tranches d'âge, catégories socioprofessionnelles, tailles d'agglomération de résidence et régions. L'échantillon n'est donc « représentatif » de la population que pour ces quelques variables. Il est facile de trouver un échantillon de taille 1000 représentatif de la population selon ces variables de quotas dans, par exemple, la sous-population des adhérents de n'importe quel parti politique.

Il est clair qu'une contrainte supplémentaire, peu discutée par les sondeurs, est de *se rapprocher le plus possible d'un sondage aléatoire*. Il faut donc trouver ce qui va jouer le rôle de base de sondage et faire des tirages aléatoires dans cette base. La base de sondage des électeurs a été remplacée par l'annuaire téléphonique dans le sondage du *Monde* du 17 mars et des numéros de téléphone ont été appelés « au hasard ». Des questions filtres sont alors posées afin de n'interroger que des personnes inscrites sur les listes électorales (voire même seulement des personnes inscrites et ayant l'intention de voter).

Dans le meilleur des cas, c'est-à-dire, sans compter les biais éventuels, la « marge d'incertitude » (à 95% de confiance) d'un sondage par quotas de taille 1000 est de l'ordre de 3%, marge d'incertitude d'un échantillon aléatoire simple (ou d'un sondage aléatoire stratifié proportionnel) de même taille. Certains sondeurs refusent que soit mentionnée la marge d'incertitude d'un échantillon aléatoire qu'ils cherchent à imiter alors que leur « marge d'erreur » est supérieure : « Évoquer de telles marges d'erreur conduit », selon Brice Teinturier, à « donner l'illusion de scientificité » et à « se rassurer à bon compte ». Pourtant, lors des présidentielles de 2002, les électeurs auraient été mieux informés si les sondeurs avaient annoncé les intentions de vote à 3% près ! Il aurait été plus facile de faire entendre des électeurs que l'ordre des trois premiers candidats n'était pas garanti ! De plus, si les résultats des sondages par quotas étaient accompagnés de marges d'incertitude (de sondages aléatoires de même taille) nous aurions beaucoup moins de publications erronées annonçant que tel candidat est passé devant tel autre alors que la différence entre les proportions d'intentions de vote pour les deux candidats est trop faible pour conclure. Il serait souhaitable que les Instituts de sondage fournissent une présentation des résultats de leurs sondages afin d'en limiter un traitement médiatique abusif ; certains le font déjà sur leur site.

Sans pouvoir les mesurer, de nombreuses autres sources d'erreurs s'ajoutent à celle due à l'échantillonnage : le développement des téléphones mobiles affaiblit considérablement la qualité de l'annuaire téléphonique comme base de sondage (d'où le recours à des appels de mobiles) ; il faut réaliser 7000 appels pour obtenir 847 personnes qui acceptent de répondre et qui ont exprimé une intention de vote (mais les non-réponses ne sont pas considérées comme un problème pour les sondeurs puisque les individus qui ne répondent pas peuvent être aisément remplacés) ; l'intention de vote n'est définitive que pour la moitié des répondants, ce qui affaiblit encore un peu plus la précision des résultats publiés.

Les sondeurs utilisent des redressements. Rien de choquant s'il s'agit de *redressement d'échantillon* selon des critères affichés *a priori* (permettant de retrouver la composition sociologique de

l'électorat ou même, bien que ce soit moins fiable, permettant de retrouver la composition électorale d'élections précédentes). Cf. dans ce numéro la contribution de Jean Chiche à ce sujet. En revanche, si la méthode de redressement est déterminée *a posteriori* et dépend des résultats eux-mêmes, on ne peut plus dire que l'enquête dispose d'un fondement scientifique.

Pour une plus grande transparence sur la méthodologie des sondages électoraux

Les Instituts de sondage doivent remettre une notice méthodologique détaillée à la Commission des sondages (cf. le document en annexe sur ce que devrait, selon le projet de loi, contenir cette notice). Il serait souhaitable que cette notice soit en libre accès sur Internet ; nombreux sont les citoyens capables de les lire. Comme nous l'avons proposé plus haut, il serait souhaitable de disposer, en plus de cette notice, d'une présentation succincte des résultats par les Instituts de sondage ; cela pourrait éviter une utilisation abusive des résultats par les médias.

Nous n'avons évoqué ici que les aspects statistiques des sondages préélectoraux en supposant que les questions sont rédigées sans ambiguïté. La construction des questionnaires d'un sondage est une opération difficile : la rédaction des questions influe fortement sur les réponses. Plus généralement, le fait de demander leur opinion sur un sujet, à des personnes qui n'ont pas d'opinion a priori sur le sujet, pose un problème de fond que les sociologues sont nombreux à souligner.

Depuis plusieurs années, les sondages politiques et électoraux, au lieu d'apporter une information utile aux citoyens, perturbent trop souvent le débat démocratique. Une nouvelle loi sur une plus grande transparence pourrait limiter l'inflation sondagière et améliorer la qualité de ce débat.

Références

Ardilly, P. (2010). Éléments sur la précision des sondages. Annexe 2 du Rapport d'information au Sénat sur les sondages n° 54 enregistré le 20 octobre 2010. http://www.senat.fr/rap/r10-054/r10-05421.html#toc216

Blanc, É. (2011). Rapport de la discussion à l'Assemblée Nationale de la proposition de loi sur les sondages, enregistré le 1er juin 2011. http://www.assemblee-nationale.fr/13/rapports/r3502.asp

Commission des sondages : http://www.commission-des-sondages.fr/

Fine, J. (2007). Les sondages : délaissés par les statisticiens et malmenés par les politologues http://jeannefine.free.fr/Sondages-Toulouse2007/documents/doc_JFine.pdf

Pénombre (2009). Les sondages d'opinion, outils de mesure à utiliser avec précaution ?, Document préparatoire à la 12ème Nocturne de Pénombre 5 juin 2009 http://www.penombre.org/IMG/File/document-preparatoire.pdf

Portelli, H. et Sueur, J.-P. (2010). Rapport d'information au Sénat n° 54 enregistré le 20 octobre 2010. Sondages et démocratie. Pour une législation plus respectueuse de la sincérité du débat politique http://www.senat.fr/rap/r10-054/r10-0540.html#toc0

Portelli, H. et Sueur, J.-P. (2011). Proposition de loi sur les sondages visant à mieux garantir la sincérité du débat politique et électoral, adoptée par le Sénat le 14 février 2011. http://www.senat.fr/leg/tas10-063.html

Rambaud, R. (2013). Le droit des sondages, conférence à l'IFOP http://droitdessondages.blog.lemonde.fr/2013/07/02/02072013-le-droit-des-sondages-fait-son-entree-a-lifop/

Riandey, B. (2012) Pédagogie statistique des sondages électoraux. http://www.sfds.asso.fr/ressource.php?fct=ddoc&i=1321

SFdS - Groupe Enquêtes et Modèles (2012). Sondages électoraux : débats méthodologiques http://www.sfds.asso.fr/286-Sondages_Electoraux#AncreSem