Rien ne va plus en statistique... Tout va bien en statistique...

Jean-Marie Parnaudeau

En quelques mois, dans différents pays, les résultats de référendums ou d'élections ont surpris : le Brexit en Grande-Bretagne, l'élection de Donald Trump et la primaire de la droite et du centre en France. Surprise parce que le résultat final n'était pas en adéquation avec les « chiffres » des différents sondages d'intention de votes.

Faut-il jeter les sondages ? Les sondages sont-ils truqués ? Les sondés répondent-ils n'importe quoi ? Les votants votent-ils pour faire mentir les sondages ? Examinons cette histoire de plus près.

En simplifiant beaucoup, mais les lecteurs de Corol'aire connaissent bien ce qui suit, la théorie des sondages est enseignée « à la Bernoulli ». On dispose d'une urne contenant des boules toutes identiques à l'exception de la couleur, supposons des blanches et des noires. La répartition des couleurs est inconnue et on suppose qu'il y a beaucoup de boules (sinon on vide la boite et on compte !). On souhaite avoir une estimation de la proportion de boules blanches (ou de noires peu importe) dans l'urne. On prélève (EAS[1]) un certain nombre de boules et on calcule la proportion de boules blanches dans cet échantillon. Cette proportion est une estimation ponctuelle de la proportion de boules blanches dans l'urne. On peut aussi donner un intervalle de confiance (voir, par exemple, les programmes de terminale S ou ES) concernant cette proportion, la demie amplitude de cet intervalle est souvent nommée, par les médias, marge d'erreur ; terme malheureux, car il laisse croire à des erreurs de méthodes ou de calculs. C'est, en résumant, ce qui est enseigné au lycée. Terminons en disant que les boules sont honnêtes et que, si la composition de l'urne est inconnue, elle est supposée fixe.

Pour appliquer cette méthode mathématique, tous les sondages devraient être effectués suivant ce protocole, mais on comprend bien qu'en pratique, ce serait techniquement compliqué, voire infaisable. Comment procède-t-on dans la réalité ?

En France, les sondages (électoraux au moins) sont fondés, non pas sur une méthode aléatoire comme celle décrite ci-dessus, mais sur la méthode des quotas. Cette méthode fait partie des méthodes de sondage dite par choix raisonné. L'idée est de considérer que si un échantillon est un modèle réduit, construit à partir de certains critères, d'une population, alors il est représentatif de cette population. Le terme statistique exact est variables de contrôle et non critères ; ces variables de contrôle doivent vérifier au moins deux conditions, être fortement corrélées avec la (ou les) variable(s) étudiée(s) et avoir une distribution statistique connue. On commence donc par construire une partition de la population. Le plus souvent, pour le cas qui nous intéresse, les critères retenus sont la proportion hommes/femmes, l'âge, la répartition géographique (grande villes, petites villes...) et les classes sociales (PCS). À partir des études de l'INSEE, on connait la proportion de la population pour chacune des parties ainsi construites. Tout échantillon sera construit en respectant cette répartition. Par exemple, si 4% de la population appartient à une partie, alors 4% de l'échantillon sera constitué d'individus de cette partie. Pour cette méthode, l'aléatoire n'intervient que dans le choix d'individus dans chaque partie et la détermination des parties se base sur « ce que l'on sait déjà par ailleurs », d'où le nom de méthode par choix raisonné. Il n'y a pas de théorie mathématique, mais pour « les milieux autorisés », ça fonctionne bien[2].

Notons que, la méthode des quotas n'est pas celle utilisée par certains médias (télévision

1. EAS signifie échantillon aléatoire simple, c'est à dire que l'on prélève les boules, une par une, au hasard avec remise après chaque prélèvement, on peut ainsi appliquer la loi binomiale en toute rigueur.
2. Sur la définition légale des sondages et du vocabulaire associé, on pourra consulter http://www.vie-publique.fr/actualite/faq-citoyens/sondages-opinion/ . On pourra remarquer que, sur ce site officiel, la méthode des quotas comporte les mots "fiable" et "précis", alors que la méthode aléatoire "consiste en une sélection au hasard d'un nombre élevé de personnes appartenant à la population de référence".

d'information en continue ou site internet). Il s'agit alors d'échantillon dit spontané ; en simplifiant : « je vois, si je suis pour ou contre, je clique » ou bien « j'entends, je téléphone ou je SMS ».

Aux États-Unis et en Grande-Bretagne, les méthodes utilisées sont différentes. Il s'agit de méthodes « aléatoires », le plus souvent par contact téléphonique ou par internet. Le mot aléatoire est entre guillemets parce que la base de données des instituts de sondages (liste de numéros de téléphone et/ou adresse courriel) ne correspond pas exactement à la base de sondage (population américaine en âge de voter) : individus non répertoriés ou individus comptés plusieurs fois. L'aléatoire ne porte que sur le numéro de téléphone ou l'adresse électronique et sur la réponse ou non de la personne contactée. Les échantillons obtenus sont la plupart du temps très différents d'un modèle réduit de la population, par exemple, la proportion femmes/hommes peut atteindre 65/35 très loin des 51/49. Les instituts procèdent alors à des redressements d'échantillons.

L'utilisation de techniques de redressement des sondages est une procédure utilisée par presque tous les instituts de sondages ; elle consiste à modifier les résultats bruts en utilisant des connaissances extérieures au sondage proprement dit. Ce qui signifie que les résultats rendus publics ne sont pas les résultats bruts. Pour certains, il s'agit de « magouille » ou de recettes secrètes mais, en France en tout cas, les méthodes de redressement sont publiques[3]. À titre d'illustration, les instituts de sondages savent que certains individus ne disent pas, pour différentes raisons, la vérité sur leur intention de vote et en ce sens ne sont pas honnêtes comme cela était le cas des boules ; pour être le plus proche possible de la « réalité », il faut donc en tenir compte et modifier[4] les résultats bruts.

Les sondages électoraux menés aux États-Unis donnaient Hillary Clinton gagnante, elle a obtenu la majorité des voix et pourtant elle a perdu. Pour mémoire, Hillary Clinton a obtenu 47,8% des voix et Donald Trump 47,3% et pourtant il a eu une majorité de grands électeurs. Examinons cette question de plus près.

Aux États-Unis, le président n'est pas élu au suffrage universel, comme en France, mais par les grands électeurs. Les grands électeurs sont élus le jour de l'« election day[5] ». Chaque État élit deux sénateurs (le sénat est donc composé de 100 sénateurs puisqu'il y a 50 états) et un ou plusieurs représentants à la chambre des représentants. Le nombre des représentants est fixé à 435, le nombre de représentants d'un état est, en gros, proportionnel au nombre d'habitants de l'État ; le nombre de grands électeurs varie de 3, pour les états les moins peuplés, à 55 pour la Californie. Il y a aussi 3 grands électeurs pour le district de Columbia, ils participent à l'élection du président mais ne siègent nulle part. Il y a donc en tout 538 grands électeurs. Pour être élu, il faut avoir au moins 270 grands électeurs. MAIS, les élections se font État par État suivant le principe du « the winner takes all », le gagnant rafle tout (sauf deux États où la répartition se fait à la proportionnelle). Dans chaque État on comptabilise les bulletins et celui ou celle qui a le plus de voix obtient tous les grands électeurs de cet État. Il suffit d'avoir la majorité plus une voix pour obtenir tous les grands électeurs d'un État, quelle que soit la taille de cet état.

Ce système de désignation explique que l'on peut être majoritaire en voix et minoritaire en nombre de grands électeurs et donc perdre l'élection présidentielle et qu'un parti majoritaire en voix peut être minoritaire dans une assemblée voire dans les deux. Cette méthode d'élection explique aussi que les grands « meetings » des candidats se soient déroulés dans les « swing states », les états pour lesquels le résultat était indécis et qui pouvaient faire « basculer » le résultat final, par exemple la Floride.

Dans une optique à la Bernoulli, on a une urne, dans laquelle se trouve 50 urnes de tailles et de compositions différentes, et, par exemple, si dans une urne on prélève 4 blanches et 3

3. Par exemple, on pourra consulter les notices, dans lesquelles la (ou les) méthode(s) de redressement sont indiquée(s) sur http://www.commission-des-sondages.fr/ ; l'étude de certaines de ces notices me semble d'une part un bon exercice de mathématiques, car certaines affirmations sont fausses mathématiquement, mais socialement acceptées (par exemple l'interprétation des intervalles de confiance) et d'autre part, un excellent exercice d'initiation au décryptage de l'information liée aux sondages et par suite à la citoyenneté.

4. L'auteur ne prend aucunement position sur le sujet, mais ne fait que décrire une pratique professionnelle.

5. Qui se déroule le premier mardi qui suit le premier lundi de novembre (ce n'est pas forcément le premier mardi de novembre…)

noires, alors on décide que toutes les boules sont blanches. On est loin de la loi binomiale… Ainsi, aux États-Unis, il devient aberrant, pour une élection présidentielle, de se fonder sur des sondages nationaux.

Enfin en France, les primaires de la droite et du centre. Cette élection, pour désigner le candidat de la droite et du centre à la prochaine élection présidentielle, s'est terminée par la victoire de François Fillon. Longtemps « très loin » dans les sondages, sa victoire a été une surprise médiatique.

Examinons cette question de plus près.

Les « études » estimaient le nombre de participants entre 2 et 5 millions, cela signifie que l'urne était de taille certes inconnue, mais variable entre les sondages et la primaire. Les programmes des candidats étant proches, le choix des participants pouvaient être modifié à tout instant, au gré des déclarations, même marginales, des uns et des autres ; ce qui signifie que les boules pouvaient changer de couleur au fil du temps et donc que la composition de l'urne était variable au cours du temps. Phénomène que les instituts de sondages qualifient de fluidité de l'électorat. Les participants, à cette élection, au dire encore des sondages, n'étaient pas tous des sympathisants de la droite et du centre, ce qui signifie qu'on rajoutait (ou on enlevait), à l'insu de tous, des boules de l'urne. Enfin, pour cette élection, aucun redressement n'était possible, puisque l'on ne disposait d'aucun recul historique. On est loin, même très loin, du modèle d'urne de Bernoulli…

Quelques points positifs tout de même, entre les deux tours de cette élection, la taille des échantillons a fortement augmenté, on a vu apparaître des sondages comportant jusqu'à 9000 personnes, ce qui revient à diviser la « marge d'erreur » par 3 par rapport à un échantillon usuel de 1000 personnes ; les commentateurs sont devenus plus réservés sur leurs affirmations suite à une publication et enfin si un sondage est une enquête statistique à un moment donné, la notion de tendance ou d'évolution des intentions de vote est davantage prise en compte. Remarquons aussi que depuis déjà quelques mois, certains médias, télévision et presse écrite, ne commandent plus de sondage.

Que conclure de tout cela ?

En statistique, comme partout, il y a des principes. Citons en deux. La stratégie du réverbère et le principe GIGO (Garbage In, Garbage Out).

La stratégie du réverbère, issue de l'histoire suivante : la nuit, un passant voit une personne regardant le sol avec attention. Intrigué, il s'approche et lui demande ce qu'il fait. « Je cherche mes clefs » lui répond-il. Le passant demande alors « Êtes-vous sûr de les avoir perdues ici ? » ; « Non, lui répond la personne, mais il n'y a qu'ici qu'il y a de la lumière ». Les élection américaines me semblent une bonne illustration de cette stratégie : faute de faire des sondages État par État puisque les élections se font État par État, il a été fait des sondages sur l'ensemble du pays (ou du moins les médias n'ont relayé que des sondages nationaux). Les élections américaines de 2000 n'avaient pas servi de leçon.

Le principe GIGO se traduit par le fait que si on veut estimer la composition d'une urne par prélèvement d'un échantillon, on s'assure au moins que cette urne est stable, on sait ce qu'il y a dedans et le contenu est inchangé au regard du protocole statistique. Tous les statisticiens sont d'accord, avant de faire une étude statistique, il faut que la population soit parfaitement identifiée : « Tous les échantillonnages visent à découvrir des renseignements au sujet d'une population particulière. Nous devons savoir clairement à quelle population nous nous intéressons[6]. »

C'était loin d'être le cas pour les élections de la droite et du centre.

Terminons sur deux conseils de lecture « L'enseignement des sondages à l'usage du plus grand nombre : quelques réflexions tirées de l'expérience » de Benoît RIANDEY et Isabelle WIDMER paru dans la revue Statistique et Enseignement 7 et l'article de Jeanne Fine et Jean-Louis Piednoir « Les sondages délaissés par les statisticiens et malmenés par les politologues » paru en 2008 dans le bulletin Vert N°474.

6. M.J Moroney Comprendre la statistique Ed Marabout 1970 page 119.

7. http://publications-sfds.math.cnrs.fr/ojs/index.php/StatEns/issue/view/1