Philippe DUTARTE
Agrégé de mathématiques

Peut-on croire un sondage ?

Après le « traumatisme » du premier tour de l'élection présidentielle de 2002, sont parus le même jour dans le journal « le Monde » deux articles au ton assez différent. On peut penser qu'une formation citoyenne aux sondages passe par la compréhension des arguments exposés dans ces deux articles[1].

L'analyse de ces articles et des exercices auxquels ils peuvent donner lieu, se fera lors de la troisième séance du stage le 27/04/07. Afin de gagner en efficacité, nous vous demandons d'ici là de bien vouloir :

- Lire les **deux articles** de presse.
- Répondre « à chaud » au petit **questionnaire** qui suit.
- Lire le **lexique** et le compléter le cas échéant.
- Étudier les **quatre exercices** proposés et pour chacun, juger de leur intérêt, du niveau auquel ils sont abordables, proposer d'autres rédactions (en particulier au niveau du collège).

Deux articles parus dans « *Le Monde* » le même jour, après le premier tour de 2002

Le Monde

Article 1 : « Sondages et regrets »

Roland Cayrol (politologue)

Article publié dans *Le Monde* daté du 26 avril 2002

Le folklore électoral suppose désormais que l'on s'en prenne aux sondeurs. Personne, en politique, n'a désormais perdu; seuls les sondages ont failli. Plus grave : voici que des électeurs, qui se sont abstenus le 21 Avril ou ont dispersé leurs voix, regrettent tout haut : *« si j'avais su, ah si les sondages m'avaient dit... ! »*.

Une profession entière (que l'auteur de ces lignes n'engage pas) a eu beau s'échiner, sur toutes les antennes et dans toutes les colonnes, à prévenir : les sondages sont des photos, ils n'ont pas de caractère prédictif, ils comportent une marge d'erreur, de plus en plus de Français se décident au dernier moment (17% ce dimanche même, d'après leurs déclarations), l'hésitation, l'indécision, le « zapping » sont plus forts que jamais dans notre histoire électorale, il y aura forcément des surprises le jour du vote ; tout cela est balayé, l'élection faite : les sondeurs se sont (encore) trompés, les sondeurs nous ont menti.

[1] Voir **Jeanne Fine** – « ***Les sondages : délaissés par les statisticiens et malmenés par les politologues*** » – Communication au colloque « Statistique et sondages », Toulouse 21 mars 2007 : http://jeannefine.free.fr/Sondages-Toulouse2007/ .

Ne tombons pas, à l'inverse, dans l'arrogance ou l'autosatisfaction : le sondage est une technique de mesure approximative, empirique, qui connaît des revers relatifs, et qui doit sans cesse être remise sur le métier. Les méthodes d'échantillonnage, d'interrogation, de redressement des données, doivent en permanence être revisitées, améliorées, mises en phase avec les évolutions de l'opinion.
Mais enfin, une profession a-t-elle à rougir de son travail dans cette campagne électorale ?
Reprenons les deux sondages publiés dans les dix derniers jours (en citant ici ceux de CSA, pour ne pas faire parler indûment des confrères), et rappelons le résultat final du vote.

	Sondage 10/11 avril	**Sondage 17/18 avril**	**Résultat 21 avril**	**Ecart final**
	%	**%**	**%**	
Abstention, Blancs et nuls	35	31	30,4	0,6
- Chirac	21	19,5	19,7	0,2
- Le Pen	12	14	16,9	2,9
- Jospin	19	18	16,1	1,9
- Bayrou	5,5	6	6,9	0,9
- Laguiller	8	7	5,8	1,2
- Chevènement	7	6,5	5,3	1,2
- Mamère	6,5	5	5,2	0,2
- Besancenot	1,5	3	4,3	1,3
- Saint-Josse	3,5	4	4,2	0,2
- Madelin	3,5	3,5	3,9	0,4
- Hue	5,5	5	3,4	1,6
- Mégret	3	2,5	2,4	0,1
- Taubira	1	2,5	2,3	0,2
- Lepage	1,5	1,5	1,9	0,4
- Boutin	1	1,5	1,2	0,3
- Gluckstein	0,5	0,5	0,5	0

On le voit à la lecture de notre tableau : les écarts entre sondage et vote, faibles dans l'ensemble (l'écart moyen par candidat est de 0,8 point), sont parfaitement explicables par la marge d'erreur de la technique, sans compter même l'inévitable différence entre les jeudi et vendredi (date de la dernière enquête) et le dimanche. Les candidats les plus éloignés des mesures sont Le Pen (moins de trois points, et il reste trois jours de campagne) et Jospin (moins de deux), tous les autres s'étagent entre 0 et 1,5 point.

Mieux : l'évolution entre le sondage effectué dix jours avant le vote et celui réalisé trois jours avant montre les évolutions en cours, et qui vont se poursuivre : Jospin, Laguiller, Hue, Chevènement, Mamère ou l'abstention sont à la baisse, Le Pen, Bayrou, Taubira, Besancenot, Saint-Josse sont en hausse.

En termes statistiques, les sondages publiés rendent donc bien compte de la physionomie du scrutin. Au passage, les objections de la Commission des Sondages, sur la fiabilité relative des mesures concernant les "petits" candidats, sont balayées, comme les statisticiens pouvaient s'y attendre (la marge d'erreur est plus faible pour eux, en valeur absolue).

D'accord, objectera-t-on, les chiffres sont corrects, mais l'ordre d'arrivée entre le deuxième et le troisième n'est pas le bon ! Revenons-y : le dernier sondage permettait d'admettre comme possible (certes, pas forcément probable), que Le Pen (14%, et à la hausse) finît par dépasser Jospin (18%, et à la baisse). Tous les commentaires, toutes les interviews de sondeurs, ont d'ailleurs tourné autour de cette question, dans les derniers jours. A partir de là, on pouvait "pronostiquer" que Jospin resterait, de peu, qualifié pour le second tour, ou que ce serait Le Pen – mais il s'agissait dès lors de prédictions, exercice auquel les sondeurs ne sont certes pas plus habiles que n'importe qui !
Ajoutons ceci : même le jour du scrutin - toutes évolutions d'intentions de vote achevées -, même avec des échantillons plus importants, les sondages ne sont pas en mesure de dire qui, de deux candidats qui se tiennent à 0,8 point, est celui qui est en tête. La marge d'erreur de la méthode l'interdit évidemment.
Ce n'est pas à dire que les sondages soient peu utiles pour suivre l'évolution d'une campagne. Si l'on n'a pas été surpris par les scores atteints par, disons Chevènement, Mamère, Bayrou, Laguiller, Hue ou Madelin, alors que leurs partisans les imaginaient volontiers à des niveaux différents, c'est bien grâce aux sondages, qui ont rendu compte des hauts et des bas des uns et des autres…
On s'en prend aux sondages à cause de l'ordre d'arrivée Le Pen / Jospin, alors qu'ils ne pouvaient en dire plus. On leur en veut d'avoir "laissé croire" aux électeurs que le second tour opposerait "forcément" Chirac à Jospin. Mais qui l'a cru, pendant cinq ans : l'opinion publique, ou les sondages qui la mesuraient ? Qui l'a laissé penser jusqu'au bout : les sondeurs, ou les candidats en présence ?
Il ne faut pas hésiter à "tirer sur les sondeurs", quand ils commettent des erreurs de méthode, comme, sans doute, lors des dernières élections municipales. Il faut les rappeler à l'ordre d'un travail incessant pour améliorer encore leurs techniques, notamment leurs méthodes de redressement des données brutes. Mais il est absurde de vouloir leur faire "porter le chapeau" de comportements d'électeurs qui voudraient les voir coupables de … leurs propres choix, ou des errements de leur camp. Après cinq ans de cohabitation, après l'oubli par les partis de leurs engagements essentiels, la coupure de la gauche avec les catégories populaires, l'incapacité des leaders à organiser un premier tour qui fut de sélection et non de sanction, les trois quarts des électeurs ont préféré l'abstention, ou un vote contre Chirac et Jospin : doit-on vraiment en blâmer les sondeurs ?

Roland Cayrol est Directeur de recherche à la
Fondation Nationale des Sciences Politiques, Directeur de l'institut CSA.

Article 2 : « Faute de contrôles... »

Michel Lejeune (statisticien)

Article publié dans *Le Monde* daté du 26 avril 2002

L'ÉLECTION présidentielle 2002 vient d'écrire la page la plus sombre des sondages « à la française ». Pour les rares scientifiques qui savent comment sont produites les estimations, il était clair que l'écart des intentions de vote entre les candidats Le Pen et Jospin rendait tout à fait plausible le scénario qui s'est réalisé. En effet, certains des derniers sondages indiquaient 18 % pour Jospin et 14 % pour Le Pen. Si l'on se réfère à un sondage qui serait effectué dans des conditions idéales (tirage aléatoire absolu, taux de réponse 100 %, aucune fausse déclaration), on obtient sur de tels pourcentages une incertitude de plus ou moins 3 % étant donné la taille de l'échantillon qui n'est en fait que de 700 environ en raison des abstentions.

En pratique, on est loin de se trouver dans ces conditions : on ne peut pas réaliser un véritable tirage aléatoire parmi les électeurs, le taux de réponse est de 10 % ou 20 % en pratique dans le cas du téléphone, enfin, les fausses déclarations d'intentions ne sont pas négligeables, en particulier sur les intentions de vote pour l'extrême droite. Les sondeurs croient, ou feignent de croire, que grâce à l'utilisation de quotas et aux redressements des échantillons, leur précision pourrait être meilleure que, par exemple, ces 3 %.
Pour ce qui est des quotas, qui consistent à sélectionner un échantillon qui ait les mêmes structures en âge, sexe et profession que la population des électeurs, les calculs montrent que le gain, en conditions idéales toujours, est insignifiant.
En ce qui concerne les redressements, notamment sur des votes antérieurs, ils laissent espérer au mieux un gain de 10 %. Mais les fausses déclarations (on n'avouait guère plus son vote Le Pen lors d'élections antérieures), les problèmes de mémoire et autres imperfections réduisent à néant cet espoir théorique. Les sondeurs prêtent des vertus démesurées aux redressements, ils ne connaissent pas leurs limites. Pour comble, on a même pu lire qu'ils corrigeaient les fausses déclarations !
Deux jours avant le premier tour, j'étais consterné d'entendre un sondeur déclarer que le vote Le Pen constaté dans les échantillons devait être multiplié (c'est-à-dire redressé) par 2. Selon des règles mathématiques rigoureuses, ajoutait-il. En réalité, les traitements réalisés sont techniquement très simples et leur compréhension est à la portée de n'importe quel bachelier. Il faut savoir qu'il n'y a aucun, je dis bien aucun, statisticien spécialiste de la théorie des sondages dans les cinq instituts concernés. De fait, il n'y a aucune règle objectivable de façon scientifique dans le traitement des résultats bruts - ne serait-ce qu'en raison du choix subjectif des critères de redressement - mais plutôt des corrections qui relèvent plus de l'appréciation empirique des politologues que de la théorie statistique.
J'en veux pour preuve la surprenante proximité des résultats d'un institut à l'autre, ainsi que leurs faibles variations semaine après semaine. Les sept derniers sondages publiés donnaient tous Jospin à 18 % : pour tout observateur avisé, cette constance est statistiquement invraisemblable avec des échantillons de taille 1 000.
Lorsqu'on est amené à « redresser » le vote Le Pen par un coefficient 2, dans des conditions au demeurant imparfaites, il est évident qu'on obtient un résultat redressé fragile, entaché d'une forte incertitude. Je pose aux sondeurs responsables la question suivante : quelle aurait été l'issue du premier tour si l'on avait utilisé un coefficient 2,5 ou, pire, si l'on avait publié, comme le font les Anglo-Saxons, une marge d'erreur de 3 %, aussi optimiste soit-elle ?
Aujourd'hui la communauté scientifique doit se sentir en partie responsable de ce qui vient de se produire. Par dédain, elle n'a jamais voulu s'intéresser à la pratique des sondages ni vraiment, d'ailleurs, à la théorie. Le nombre de thèses dans le domaine, en France, se compte sur les doigts d'une seule main, car un doctorant en théorie des sondages perd pratiquement tout espoir de décrocher un poste universitaire.
Quant aux médias, ils sont les dupes, peut-être parfois les complices, des discours lénifiants des instituts. Je leur suggère, pour l'heure, de réclamer la mise sur pied d'une commission indépendante de statisticiens qui vienne décortiquer, a posteriori, la façon dont les chiffres publiés ont été produits. Au-delà, je plaiderais pour l'instauration de contrôles approfondis des sondages d'opinion, comme cela se fait pour les études d'audience des médias, alors qu'actuellement la commission des sondages opère de façon très limitée. Tout le monde y gagnera pour l'avenir, y compris les sondeurs.

Questionnaire (c'est pour un sondage...)

Après la lecture des deux articles précédents, pouvez-vous répondre « à chaud » au petit questionnaire suivant (nous en débattrons lors de la séance 3 du stage) ?

- Quel est, de ces deux articles, celui qui vous a paru le plus clair ?

☐ Article 1 (Cayrol) ; ☐ Article 2 (Lejeune).

- Avec lequel des deux articles vous sentez-vous le plus en accord ?

☐ Article 1 (Cayrol) ; ☐ Article 2 (Lejeune).

- Quelle est l'idée forte que vous avez retenu de la première lecture de ces deux articles ?

.....
.....
.....
.....

- Selon vous, quelle thèse principale défend chacun des deux articles ?

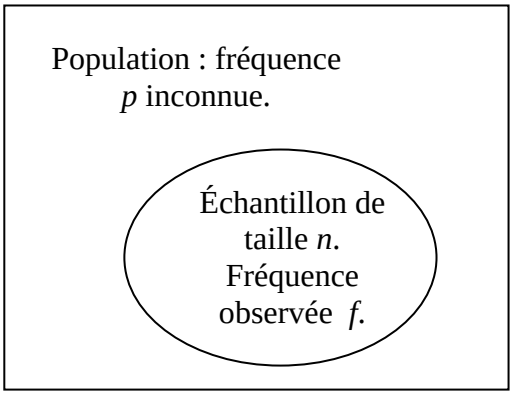
Article 1 (Cayrol) : .....
.....
.....
.....
Article 2 (Lejeune) : .....
.....
.....
.....

Lexique... et méthodologie

Sondage aléatoire simple

C'est le sondage enseigné au lycée, dans le cadre d'un « thème d'étude » en classe de seconde générale, ou dans le chapitre « intervalles de confiance » de certaines sections de BTS.
Il s'agit de tirer au hasard n éléments dans une population où la fréquence p d'un caractère est inconnue (par exemple le pourcentage p d'électeurs en faveur d'un candidat). L'expression « au hasard » signifie que chaque échantillon de taille n a la même probabilité d'être tiré.
On peut facilement appliquer la théorie des probabilités à ce type de sondage.

En supposant que le tirage est effectué avec remise (on peut faire cette hypothèse si la taille n du sondage est faible devant celle de la population), on a la situation du schéma de Bernoulli. Si n est « assez grand », l'approximation de la loi binomiale par la loi normale conduit à un « intervalle de confiance » : si on observe la fréquence f sur l'échantillon de taille n, on démontre que la fréquence correspondante inconnue p dans la population est située dans l'intervalle $\left[f - 1{,}96\sqrt{\frac{p(1-p)}{n}},\ f + 1{,}96\sqrt{\frac{p(1-p)}{n}} \right]$ avec 95 % de « confiance ». Cette expression signifie que sur un grand nombre d'échantillons de taille n, dans environ 95 % des cas, p est effectivement dans l'intervalle ci-dessus.

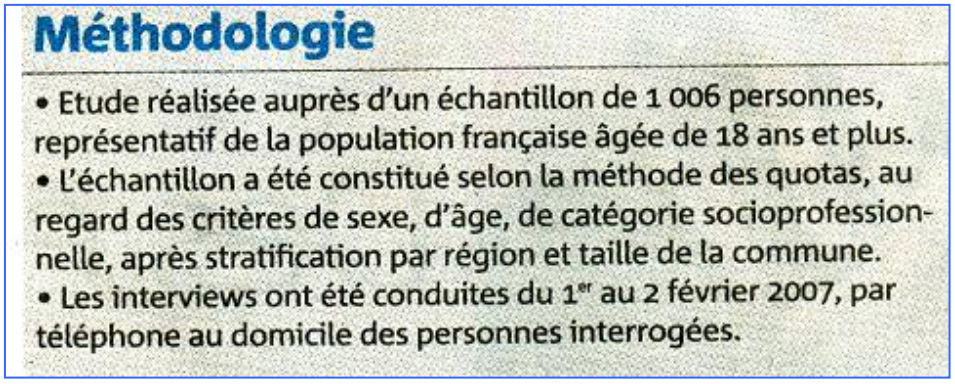


Comme p est inconnu, en remarquant que pour tout p de l'intervalle [0, 1], on a $1{,}96\sqrt{\frac{p(1-p)}{n}} \leq \frac{1{,}96}{2\sqrt{n}} \leq \frac{1}{\sqrt{n}}$, on peut majorer l'intervalle de confiance à 95 % en donnant une « fourchette » à plus de 95 % de confiance sous la forme $\left[f - \frac{1}{\sqrt{n}},\ f + \frac{1}{\sqrt{n}} \right]$.

C'est cette « fourchette » qui peut être expérimentée en seconde, par simulation.

Sondage par quotas

C'est la méthode pratiquée par les instituts de sondages français (voir les encarts méthodologiques publiés dans la presse, quand ils le sont). Cette méthode ne contient rien d'aléatoire (du moins maîtrisé) et par conséquent sa fiabilité ne peut être mathématiquement calculée, puisqu'on ne peut pas utiliser le calcul des probabilités. La fiabilité de la méthode des quotas n'est qu'empirique, fondée sur « l'expérience des sondages précédents ».



Méthodologie

- Etude réalisée auprès d'un échantillon de 1 006 personnes, représentatif de la population française âgée de 18 ans et plus.
- L'échantillon a été constitué selon la méthode des quotas, au regard des critères de sexe, d'âge, de catégorie socioprofessionnelle, après stratification par région et taille de la commune.
- Les interviews ont été conduites du 1er au 2 février 2007, par téléphone au domicile des personnes interrogées.

En étant assez optimiste, on peut considérer que la méthode des quotas conduit à une marge d'erreur, à 95 %, de l'ordre de celle d'un sondage aléatoire simple, c'est-à-dire environ $\frac{1}{\sqrt{n}}$, soit, pour n = 1000 personnes interrogées, $\frac{1}{\sqrt{1000}} \approx 3$ %. C'est la raison pour laquelle, même si elle n'est pas pratiquée par les instituts de sondages français, l'enseignement des mathématiques de la méthode aléatoire est instructif (et cette méthode est utilisée dans les contrôles de qualité).

☺ Voir l'**exercice 1**.

La méthode des quotas consiste à choisir un certain nombre de critères jugés importants pour le sujet du sondage : sexe, âge, catégorie socioprofessionnelle, région, taille de la commune..., puis à calculer le pourcentage de personnes appartenant à chaque catégorie selon les données du recensement.
Il s'agit alors d'obtenir autant de réponses que chaque quota ainsi calculé pour un échantillon de taille *n*. Pour atteindre les quotas de chaque catégorie, il ne s'agit pas de tirages au sort organisés de façon à maîtriser les probabilités (c'est beaucoup plus économique que d'interroger des personnes réellement au hasard). Évidemment des biais existent (que l'on peut chercher à corriger de façon plus ou moins empirique...), en particulier parce que répondent les personnes joignables qui veulent bien répondre. C'est un peu comme si un biologiste voulant tester un nouveau produit sur une souris le faisait sur la première souris qu'il peut attraper dans la cage. Il y a toutes les chances pour que cette souris soit la plus faible de toutes, la moins vive.

Sondage aléatoire stratifié (proportionnel / optimal)

Ces méthodes permettent d'améliorer la fiabilité du sondage aléatoire simple. Elles demeurent cependant des méthodes aléatoires, avec toutes les exigences d'un tirage « au hasard » et ne sont donc pas comparables à la méthode des quotas, malgré certaines confusions parfois entretenues (en particulier par les termes de « stratification » ou de « représentatif »). On suppose que pour toutes les personnes de la population, on peut avoir accès aux informations correspondant aux critères sélectionnés (sexe, âge, catégorie socioprofessionnelle, région, taille de la commune...). On effectue alors une partition de la population selon les critères retenus.
Un échantillon aléatoire stratifié proportionnel sera obtenu par tirage au sort dans chaque sous-ensemble (« strate ») de la population, en quantités proportionnelles aux effectifs de chaque sous-ensemble. On montre, qu'à taille d'échantillon égale, la précision peut-être considérablement améliorée par rapport à un sondage aléatoire simple, en particulier si les proportions, pour le caractère étudié, sont très différentes selon les strates.
Un échantillon aléatoire stratifié optimal tient compte de la dispersion selon les strates de la variable faisant l'objet de l'enquête. On gagne en précision en abandonnant la simple proportionnalité et en interrogeant davantage dans les strates à forte dispersion que dans celles très homogènes.

Marge d'erreur

Pour un sondage portant sur 1000 personnes, on parle parfois de « marge d'erreur de plus ou moins 3 % ». L'expression « marge d'erreur » pourrait laisser croire qu'au delà de cette marge, on a la certitude de ne pas trouver le pourcentage réel que l'on cherche à estimer. Ce qui est faux. Il vaudrait mieux parler de marge « d'incertitude à 95 % de confiance ».
Rappelons que pour la méthode des quotas, il est impossible d'évaluer sérieusement, c'est-à-dire mathématiquement, la marge d'incertitude. Pour un sondage aléatoire simple de 1000 personnes, la marge d'incertitude à 95 % de confiance, à partir d'une fréquence *f*

calculée sur le sondage, est de plus ou moins $1{,}96\sqrt{\frac{p(1-p)}{1000}}$, que l'on peut approcher par $1{,}96\sqrt{\frac{f(1-f)}{1000}}$. Le tableau suivant donne quelques calculs :

Fréquence f calculée sur un sondage	10% ou 90%	20% ou 80%	30% ou 70%	40% ou 60%	50%
Marge d'incertitude à 95% de confiance pour un sondage aléatoire de taille 1000.	1,86%	2,48%	2,84%	3,04%	3,10%

On constate qu'évaluer la marge d'incertitude à 3% (à 95% de confiance) est bien adapté pour des fréquences observées entre 30% et 70%. Pour de petites ou de fortes fréquences, cette marge de 3% est une majoration parfois importante de l'incertitude.
A noter que si f est trop petite ou trop grande, la formule précédente cesse d'être valable, il faut avoir $n \times f$ et $n \times (1-f)$ au moins supérieurs à 5 pour appliquer raisonnablement l'approximation d'une loi binomiale par une loi normale.

Échantillonnage

L'échantillonnage est la manière dont est constitué l'échantillon. Il faut distinguer l'échantillonnage aléatoire de celui qui ne l'est pas (comme dans la méthode des quotas). Avec un échantillonnage aléatoire, on peut utiliser le calcul des probabilités et donc estimer l'incertitude (encore faut-il ne pas avoir de non réponses, de fausses déclarations, d'enquêteurs peu sérieux...). Avec un échantillonnage non aléatoire, on ne peut pas estimer le risque d'erreur.
Il ne faut pas confondre « aléatoire » et « aveugle ». Quand on interroge les gens au téléphone par la méthode des quotas, on procède de façon « aveugle », donc avec une certaine dose d'aléatoire mais sans savoir laquelle ni dans quel sens. Quand on parle de méthode « aléatoire », il s'agit d'un « hasard complet », maîtrisé de sorte à s'assurer que chaque individu de la population a les mêmes chances d'être interrogé. Il faut pour cela une procédure très contrôlée : par exemple numéroter tous les individus et tirer les numéros au sort selon un procédé dont on sait qu'il respecte l'équiprobabilité, par exemple un générateur de nombres aléatoires.

Échantillon représentatif

Voilà une expression qui, si elle n'est pas précisée, peut signifier à peu près n'importe quoi.
Un échantillon constitué selon la méthode des quotas est évidemment « représentatif » des critères correspondants aux quotas (sexe, âge, catégorie socioprofessionnelle, région, taille de la commune...) selon lesquels il a été fabriqué. Mais on n'a aucun moyen de savoir jusqu'à quel point il est « représentatif » de ce pour quoi il a été prélevé, c'est-à-dire le sujet du sondage, l'opinion, le pourcentage que l'on cherche à évaluer. L'expression « représentatif de la population française », que l'on lit souvent dans la presse, prête évidemment à confusion. On a l'impression que l'échantillon est « représentatif » de tout ce que l'on veut.
En statistique, on désigne plutôt par « échantillon représentatif », un échantillon où le hasard permet d'éviter les biais inconnus et d'appliquer le calcul des probabilités. La méthode optimale pour obtenir un échantillon « représentatif » est celle du sondage aléatoire stratifié optimal.

Taux de réponse

La plupart des sondages sont effectués par téléphone. Dans ce cadre, Michel Lejeune évoque dans son article un taux de réponse de l'ordre de 10% à 20%. Avec un tel taux de non-réponses, le biais est sans doute non négligeable. Qui répond ? Qui refuse de répondre ? Le taux de non réponse n'est sans doute pas le même dans les différentes catégories d'opinion.

☺ Voir l'**exercice 2**.

Défaut de couverture

C'est un autre biais important. La population sondée est-elle la population visée ? Si le sondage est effectué par Internet, s'il l'est par téléphone pendant les heures de travail, ... ce n'est certainement pas le cas. De toutes manières, des pans entiers de la populations sont hors d'atteinte.

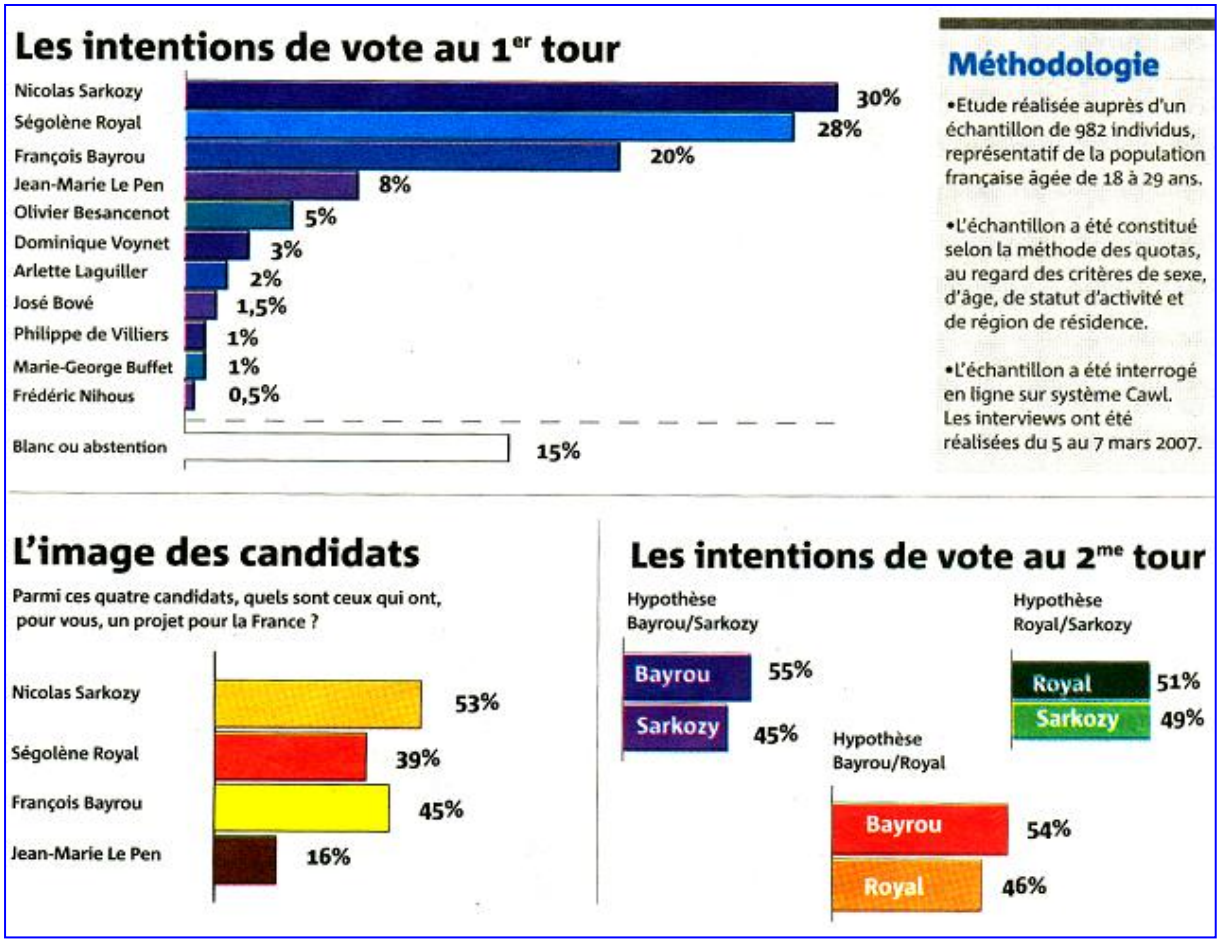
Fausses déclarations

C'est une source importante de biais pour des questions sensibles et souvent difficile à évaluer. Il existe des méthodes d'interrogation aléatoire, ou de recoupement avec d'autres questions.

☺ Voir l'**exercice 3**.

Redressement de l'échantillon

Il s'agit souvent de méthodes empiriques consistant à « corriger » certains biais constatés lors des études analogues précédentes. Par exemple, en 2002, certains sondeurs ont expliqué « redresser » les intentions de votes pour J.-M. Le Pen en les multipliant par 2.



Et encore ?

Il vous manque des mots ? Vous en avez ajouté d'autres ? C'est confus ?

.....
.....
.....
.....
.....
.....
.....

Exercice 1 : fourchettes.

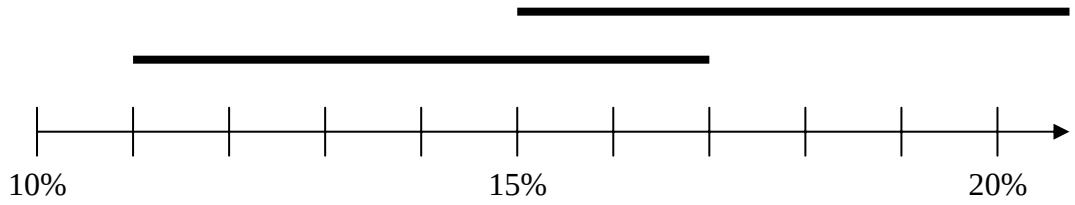
Énoncé

Voici un extrait d'article, publié dans le journal « Le Monde » par le statisticien Michel Lejeune, après le premier tour de l'élection présidentielle de 2002.
« Pour les rares scientifiques qui savent comment sont produites les estimations, il était clair que l'écart des intentions de vote entre les candidats Le Pen et Jospin rendait tout à fait plausible le scénario qui s'est réalisé. En effet, certains des derniers sondages indiquaient 18 % pour Jospin et 14 % pour Le Pen. Si l'on se réfère à un sondage qui serait effectué dans des conditions idéales [...], on obtient sur de tels pourcentages une incertitude de plus ou moins 3 % étant donné la taille de l'échantillon [...]. »

1. Si l'on tient compte de l'incertitude liée au sondage, entre quels pourcentages pourraient se situer réellement (à 95% de confiance) les deux candidats lorsque le sondage donne 18% pour l'un et 14% pour l'autre ?
2. Représenter sur un même graphique les deux « fourchettes » calculées à la question précédente. Peut-on prévoir l'ordre des candidats ?
3. Au premier tour de l'élection présidentielle de 2002, L. Jospin a obtenu 16,18% des voix et J.-M. Le Pen 16,86%.
Expliquer la phrase « l'écart des intentions de vote entre les candidats Le Pen et Jospin rendait tout à fait plausible le scénario qui s'est réalisé ».

Éléments de réponse

1. Pour L. Jospin, entre 15% et 21%. Pour J.-M. Le Pen, entre 11% et 17%.
2. Un dessin possible.



Si on utilise ces fourchettes, on ne peut pas prévoir l'ordre des candidats car elles ont une partie commune.
3. La phrase correspond au fait que les pourcentages obtenus à l'élection sont situés dans les fourchettes du sondage.

Quel est l'intérêt de l'exercice (du point de vue des maths et/ou de la citoyenneté) ?
.....
.....
.....
A quel niveau cet exercice est-il abordable, tel qu'il est rédigé ?
.....
.....
A quel niveau cet exercice est-il abordable, avec une autre rédaction (proposer) ?
.....
.....
.....
.....

Exercice 2 : non réponses

Énoncé

On a interrogé 1000 personnes auxquelles on demandait de répondre par « oui » ou par « non » à la question « à la prochaine élection, pensez-vous voter pour le candidat de l'extrême droite ? »
Les résultats sont les suivants :

Oui : **98** ; Non : **717** ; Non réponses : **185** ; Total : **1000**.

1. Quel est le pourcentage de « oui », par rapport aux personnes interrogées et par rapport aux répondants ?

2. On suppose que des études précédentes ont montré que parmi les électeurs d'extrême droite, **50%** refusent de répondre.
En admettant que ce pourcentage est ici respecté et qu'il n'y a aucune fausse déclaration, calculer le pourcentage d'électeurs d'extrême droite parmi les 1000 personnes interrogées.
3. Quel est, en utilisant le résultat de la question précédente, le pourcentage de non-réponses parmi les électeurs ne votant pas pour l'extrême droite ?

Éléments de réponse (c'est le cas de le dire...)

1. Il y a **9,8%** de « oui » par rapport aux 1000 personnes interrogées et $\frac{98}{98 + 717} \approx$ **12%** par rapport aux répondants.
2. Si l'on admet qu'il n'y a pas de fausses réponses, il y a 98 électeurs d'extrême droite qui répondent. En supposant que parmi les électeurs d'extrême droite 50% répondent et 50% ne répondent pas, il y a donc 98 + 98 = 196 électeurs d'extrême droite. Ce qui représente **19,6%** des 1000 personnes.
3. D'après la question précédente il y a 1000 – 196 = 804 électeurs ne votant pas pour l'extrême droite dont 717 ont répondus. Il y a donc 804 – 717 = 87 électeurs ne votant pas pour l'extrême droite qui n'ont pas répondu. (On peut faire aussi 185 – 98 = 87).
Le taux de non réponse chez ces électeurs est donc de $\frac{87}{804} \approx 10{,}8\%$.
Remarque : il peut être utile de visualiser la situation à l'aide d'un tableau à double entrée :

	électeurs d'extrême droite	non électeurs d'extrême droite	total
réponses	**98**	**717**	815
non réponses	98	87	**185**
total	196	804	**1000**

Quel est l'intérêt de l'exercice (du point de vue des maths et/ou de la citoyenneté) ?
.....
.....
.....
A quel niveau cet exercice est-il abordable, tel qu'il est rédigé ?
.....
.....
A quel niveau cet exercice est-il abordable, avec une autre rédaction (proposer) ?
.....
.....
.....
.....

Exercice 3 : quand le hasard évite les fausses déclarations

Énoncé

On souhaite interroger 1000 personnes sur leur intention de voter pour l'extrême droite.
Pour éviter les fausses déclarations, l'enquêteur préservera l'anonymat des réponses en proposant de réaliser en son absence la procédure suivante :

Lancer une pièce.

- *Si elle tombe sur pile, répondre par « oui » ou par « non » à la question :*

« avez-vous l'intention de voter pour l'extrême droite ? ».

- *Si elle tombe sur face, relancer la pièce, si elle tombe sur pile, répondre « oui », si elle tombe sur face, répondre « non ».*

De cette façon, l'enquêteur ne peut pas savoir à quoi correspond la réponse « oui » ou « non » qu'on lui transmet.
L'enquête donne 338 réponses « oui » et 662 réponses « non ».
1. On suppose qu'il y a autant de pile que de face aux premiers lancers, comme aux seconds lancers. Combien y en a-t-il dans chaque cas ?
2. On désigne par x la fréquence des électeurs d'extrême droite, montrer que x vérifie l'équation $500\,x + 250 = 338$.
3. Calculer la fréquence des électeurs d'extrême droite.

Éléments de réponse

1. Il y a 1000 premiers lancers avec 500 faces et 500 piles.
Il y a 500 seconds lancers avec 250 piles et 250 faces.
2. Les 338 réponses « oui » correspondent aux électeurs d'extrême droite parmi les 500 qui ont fait pile au premier lancer, soit $500 \times x$, auxquels s'ajoutent les 250 personnes qui ont fait pile au second lancer. D'où $500\,x + 250 = 338$.
3. On en déduit que $x = \dfrac{338 - 250}{500} = 17{,}6\%$.
Remarque 1 : bien sûr, il n'y aura pas exactement autant de pile que de face, mais l'exercice montre qu'on peut « estimer » la proportion x sans rien connaître des réponses individuelles.
Remarque 2 : en classe de première, on peut proposer cet exercice dans le cadre des probabilités, à résoudre par exemple à l'aide d'un arbre.

Quel est l'intérêt de l'exercice (du point de vue des maths et/ou de la citoyenneté) ?
.....
.....
.....
.....
A quel niveau cet exercice est-il abordable, tel qu'il est rédigé ?
.....
.....
A quel niveau cet exercice est-il abordable, avec une autre rédaction (proposer) ?
.....
.....
.....
.....
.....
.....

Exercice 4 : où l'on suspecte la méthode par quotas

Énoncé

A la suite du premier tour de l'élection présidentielle de 2002, le statisticien Michel Lejeune a mis en cause la méthode des quotas, utilisée en France pour effectuer les sondages, en affirmant :

« Les sept derniers sondages publiés donnaient tous Jospin à 18 % : pour tout observateur avisé, cette constance est statistiquement invraisemblable avec des échantillons de taille 1000. »

1. On suppose que le jour des sondages, il y avait effectivement 18% des électeurs en faveur de L. Jospin. On note X la variable aléatoire qui à tout tirage au hasard de 1000 électeurs associe le nombre de ceux en faveur de L. Jospin (la population étant très importante, on supposera que les tirages sont indépendants).
Quelle est la loi suivie par X ?

2. Les résultats des sondages sont donnés à 0,5% près. Pour quelles valeurs de X arrondira-t-on le résultat du sondage à 18% ?

3. A l'aide d'un tableur, on a calculé ci-dessous les probabilités $P(X \leq k)$ pour $k = 184$ et $k = 174$.
Quelle est la probabilité de l'événement : « le sondage affiche un résultat à 18% pour L. Jospin » ?

B1 = =LOI.BINOMIALE(184;1000;0,18;VRAI)

	A	B	C	D	E
1	P(X <= 184)	0,647			
2	P(X <= 174)	0,328			
3	Différence	0,319			
4					

4. En supposant que les 7 sondages ont été réalisés indépendamment et dans les mêmes conditions, quelle est la probabilité qu'ils donnent le même résultat de 18% ?

Éléments de réponse

1. La variable aléatoire X suit la loi binomiale de paramètres $n = 1000$ et $p = 0,18$.
2. Le résultat du sondage est arrondi à 18% pour X compris entre 175 et 184.
3. On a $P(175 \leq X \leq 184) = P(X \leq 184) - P(X \leq 174) \approx 0,319$.
4. La probabilité que les 7 sondages donnent le même résultat est $(0,319)^7 \approx 0,0003$.
On en déduit que les sondages n'étaient sans doute pas effectués de façon indépendante.

Quel est l'intérêt de l'exercice (du point de vue des maths et/ou de la citoyenneté) ?
.....
.....
.....
.....
A quel niveau cet exercice est-il abordable, tel qu'il est rédigé ?
En terminale S.
A quel niveau cet exercice est-il abordable, avec une autre rédaction (proposer) ?
(En seconde, par simulation...)
.....
.....
.....