# *Petit lexique des sondages politiques*

## Niveau

Document d'information et de réflexion pour le professeur (mais lisible par un élève de terminale).

## Situation étudiée

Peut-on croire un sondage politique ?

## Objectifs

### *Contenus mathématiques*

Échantillonnage aléatoire.
Intervalle de confiance.
Loi des grands nombres, loi binomiale, loi normale.

### *Enjeux citoyens*

Les sondages sont un élément omniprésent de notre vie politique.
Quelle est la place des mathématiques dans les sondages ? Comprendre les notions mathématiques intervenant.
En quoi notre enseignement peut-il permettre d'éduquer le citoyen dans ce domaine ?
Que se cache-t-il derrière les mots « quotas », « stratifié », « représentatif », « 95% de confiance » ... ?

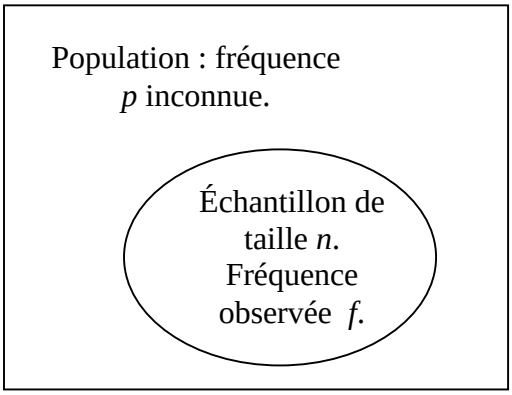
## Lexique des sondages... et méthodologie

On aurait pu intituler ce paragraphe, selon un titre paru dans la presse :

« Dans les cuisines des sondeurs ».

### *Sondage aléatoire simple*

C'est le type de « sondage » actuellement 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 (on effectue des sondages aléatoires simples dans l'industrie, pour le contrôle de qualité par exemple).

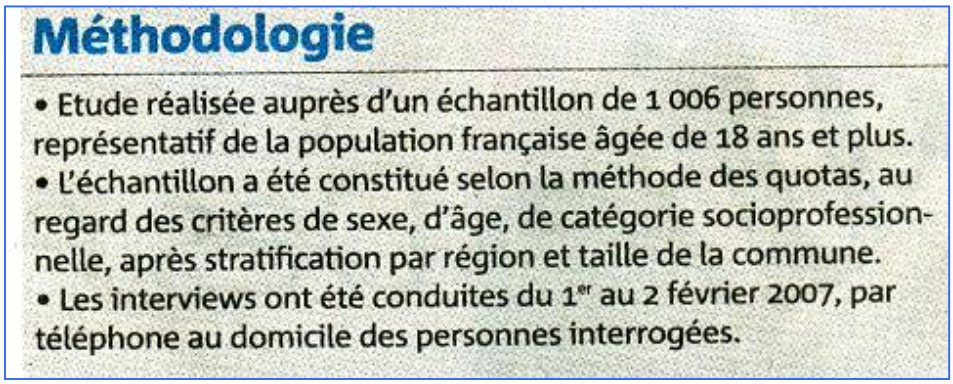



Il s'agit de tirer au hasard $n$ éléments dans une population où la fréquence $p$ d'un caractère est inconnue (par exemple le pourcentage $p$ d'électeurs en faveur d'un candidat).

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, on a un exemple ci-dessous).
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 « Fourchettes pour la cuisine des sondages »** des activités suivantes.

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 « Non réponses »** des activités suivantes.

### *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 « Éviter les fausses déclarations grâce au hasard »** des activités suivantes.

### *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. Comme on le constate, c’est une technique mathématique assez élémentaire...

☺ Voir l’**exercice 4 « Où l’on suspecte la méthode des quotas »** des activités suivantes.