La fourchette d'un sondage

Jacques Istas

La fourchette d'un sondage

Le 10 février 2009, par **Jacques Istas**
Professeur, Université de Grenoble (**page web**)

Combien de personnes pour un bon sondage ?

COMMENT calcule-t-on la fourchette d'un sondage ?
Nous allons regarder le cas d'une élection où deux candidates Ségolène et Martine s'affrontent. Nous voulons savoir combien de personnes il faut sonder pour pouvoir évaluer la popularité de nos deux candidates. Nous allons nous placer en situation idéale :

- les sondés disent effectivement pour qui ils veulent voter [1] ;
- les sondés sont indépendants entre eux [2] ;
- les sondés représentent bien la population totale. [3].
- les sondeurs font honnêtement leur travail. [4]

Modélisons cette situation. Nous allons regarder la méthode avec tirage au sort, qui est différente de la méthode des quotas.

Habituellement, on donne la cote de popularité en %, par exemple 53 %. Ici, nous ramenons cette cote entre 0 et 1 et 53 % devient 0,53. Notons p la cote de popularité de Martine ; celle de Ségolène vaut $1-p$ [5]. Arbitrairement, nous attribuons la valeur « 0 » à la candidate Ségolène et la valeur « 1 » à la candidate Martine. Notons X_i le vote du ième sondé. X_i vaut donc 0 ou 1. Ici intervient un point clé de la modélisation. Nous supposons que le ième sondé a été tiré au hasard au sein de la population. Il y a a priori une probabilité p qu'il soit en faveur de Martine et $1-p$ en faveur de Ségolène. Nous allons donc faire comme si X_i était aléatoire, et que la probabilité que X_i vaille 1 soit p, et que la probabilité que X_i vaille 0 soit $1-p$. Les probabilistes disent que X_i suit une loi de Bernoulli de paramètre p. Une dernière remarque, que l'on peut omettre en première lecture : pour simplifier les choses, nous supposons que la population totale d'électeurs est grande par rapport au nombre de sondés, de sorte que nous puissions négliger les questions de tirage avec ou sans remise de sondés.

Nous sondons n personnes, en supposant que les conditions idéales précédentes sont vérifiées. Le nombre de sondés se déclarant pour Martine vaut donc $S_n = X_1 + X_2 + \ldots + X_n$. Une somme de n variables de Bernoulli indépendantes et de même paramètre p s'appelle une binomiale de paramètres n et p. S_n est donc une binomiale de paramètres n et p et vérifie

$$Proba[S_n = k] = C_n^k p^k (1-p)^{n-k},$$

où

$$C_n^k = \frac{n!}{k!(n-k)!}.$$

$n! = n(n-1)(n-2)...3x2$ est la factorielle de n et C_n^k les coefficients binomiaux.

Estimons maintenant la popularité p de Martine. L'estimateur naturel est la fréquence :

$$p_n = \frac{S_n}{n}$$

Nous voulons maintenant savoir si p_n est proche de p. Intuitivement, nous sentons bien que plus n est grand, plus p_n sera proche de p. Et c'est effectivement ce qu'affirme un théorème essentiel en Probabilités : la loi des grands nombres [6]. Cette loi des grands nombres nous affirme que p_n va être d'autant plus proche de p que n est grand. Mais cette loi des grand nombres ne quantifie pas l'écart, ou marge d'erreur, entre p_n et p.

Tentons un calcul direct pour mesurer cet écart. Dire que l'écart entre p_n et p est inférieur à d revient à dire que S_n est compris entre $np - nd$ et $np + nd$. Tentons de calculer la probabilité de cet événement. Il faut faire la somme des $C_n^k p^k (1-p)^{n-k}$ pour tous les entiers k compris entre $np - nd$ et $np + nd$. Sortons notre calculette. Dès que n est plus grand que 70, le calcul de $n!$ dépasse les possibilités de la calculette et évaluer C_n^k devient impossible. Calculer p^k pour k grand sera également impossible.

Il nous faut donc contourner ces difficultés. Pour cela, nous allons utiliser un autre théorème essentiel en Probabilités, le théorème de la limite centrale. Si vous voulez vérifier expérimentalement ce théorème de la limite centrale, cliquez **ici**.Ce théorème de la limite centrale nous affirme que, plus n est grand, plus $\sqrt{n}(p_n - p)$ ressemble à une variable gaussienne de moyenne 0 et de variance $p(1-p)$. Pas de panique si vous ne connaissez pas la gaussienne, il suffit de savoir qu'elle est bien connue, et qu'elle est **tabulée** depuis longtemps. Quand on regarde une table de la loi gaussienne, on se rend compte que seule la gaussienne de moyenne 0 et de variance 1 est tabulée. Pourquoi ? Parce que, lorsque l'on divise une gaussienne de moyenne 0 et de variance σ^2 par σ, on obtient une gaussienne de moyenne 0 et de variance 1. Nous avons donc remplacé la binomiale dont les paramètres étaient trop gros pour être gérés par une gaussienne, qui est « tractable » comme on dit en bon franglais. Ainsi, nous approximons la probabilité que $\sqrt{n}(p_n - p)$ soit en valeur absolue plus petit qu'une valeur e par la probabilité que la valeur absolue d'une gaussienne de moyenne 0 et de variance $p(1-p)$ soit plus petite que e

$$Proba[\sqrt{n}|p_n - p| \leq e] \sim Proba[|U_{p(1-p)}| \leq e],$$

où $U_{p(1-p)}$ est une gaussienne de moyenne 0 et de variance $p(1-p)$. Nous nous ramenons ensuite à une gaussienne U_1 de moyenne 0 et de variance 1

$$Proba[|U_{p(1-p)}| \leq e] = Proba[|U_1| \leq e/\sqrt{p(1-p)}]$$

Un problème demeure néanmoins. La variance $p(1-p)$ de la gaussienne est évidemment inconnue puisqu'elle dépend du paramètre p que nous cherchons à estimer ! Un calcul rapide montre que le polynôme $p(1-p)$ est maximal en $p = 1/2$ et vaut $1/4$. Comme nous pressentons que le scrutin sera serré, nous ne perdons pas grand chose en majorant la variance

$p(1-p)$ par $1/4$. Au final, nous faisons donc l'approximation suivante

$$Proba[\sqrt{n}|p_n - p| \leq e] \sim Proba[|U_1| \leq 2e],$$

Nous devons maintenant aborder un autre point important : un sondage ne peut pas être sûr ! Il faut accepter l'idée de se tromper ! Il n'y a pas de règle en la matière pour fixer un seuil. Faisons un calcul avec un risque de se tromper de 5 % [7] Quelle fourchette (=marge d'erreur) voulons-nous pour le sondage ? Egalement 5 % ? Allons-y pour 5 % ! Nous lisons dans une table que la probabilité qu'une gaussienne de moyenne 0 et de variance 1 soit plus petite en valeur absolue que (approximativement) 2 vaut 0,05. Nous sommes maintenant capables de trouver le nombre n de sondés qu'il nous faut pour avoir une fourchette de 5 % avec un risque de se tromper de 5 %. Il suffit en effet de résoudre

$$\frac{1}{\sqrt{n}} = 0,05,$$

soit $n = 400$. Résumons la situation. Supposons que la cote soit estimée à 42% sur la base de 400 sondés. Nous affirmons donc, avec une probabilité de 95%, que la vraie cote se trouve entre 37% et 47%. Si nous avions voulu une fourchette à 1 %, toujours avec un risque de se tromper de 5 %, nous aurions à résoudre

$$\frac{1}{\sqrt{n}} = 0,01,$$

soit $n = 10000$. Imaginons que la cote estimée soit maintenant de 43%. Nous affirmons alors, avec une probabilité de 95% (ce risque n'a pas changé), que la vraie cote se trouve entre 42% et 44%.

Nous sommes donc capables de calculer la fourchette d'un sondage. Qu'avons-nous appris d'autre ?

- Contrairement à une idée répandue, le nombre de sondés ne dépend pas de la population totale. Ceci explique pourquoi il est facile, le soir du second tour de l'élection présidentielle, de dire qui est le candidat gagnant dès 20 heures. En revanche, dans le cas du second tour d'élections législatives, les résultats ne sont pas tous connus à 20 heures, car ils nécessitent d'avoir sondé dans les 577 circonscriptions.
- La fourchette du sondage ne diminue qu'avec la racine carrée du nombre de sondés. La fonction racine carrée croît lentement, et cela coûte très cher aux instituts de sondage !

Pour continuer

- Supposons que Martine ait une cote de popularité estimée à 51 % avec une fourchette de 3 %, soit un intervalle 48 %-54 %. Martine se moque en fait de sa popularité, elle veut simplement savoir si elle sera élue, c'est-à-dire savoir si sa cote est supérieure à 50 %. Peut-elle avoir confiance dans ce 48 %-54 % ?
- Une semaine plus tard, un deuxième sondage, effectué dans les mêmes conditions, donne Martine à 52%. Sa cote a-t-elle réellement augmentée ?

P.S. :

Pour en savoir plus : *article (difficile) sur les* ***sondages*** *d'IdM 2006*

Notes

[▲1] Ce qui n'est pas toujours vrai en pratique, pensons par exemple à l'électorat du Front National qui avait tendance à se censurer vis-à-vis des sondeurs.

[▲2] Difficile à savoir ! La méthode choisie pour tirer des sondés peut contenir une dépendance, bien cachée, entre sondés.

[▲3] Selon l'heure, le lieu, le type de contact (téléphone, mail), on peut sonder plus de femmes, de chômeurs, de personnes agées, ... Il est par exemple très dur de joindre en semaine un étudiant qui ne vit pas chez ses parents sur un téléphone fixe. Il faut également vérifier que les sondés iront voter, ne changeront pas d'avis ...

[▲4] En particulier, les sondeurs « de base » sont grassement payés et ne sont pas tentés de remplir eux-mêmes les fiches ; les responsables des sondeurs ne réajustent pas les résultats du sondage en fonction de leur perception personnelle de l'opinion. Voir à ce sujet ce qu'en disent les **intéressés**.

[▲5] Nous avons écarté les indécis.

[▲6] Rien à voir avec la pseudo-loi des séries, qui n'est qu'une fumisterie !

[▲7] Ce qui signifie qu'en moyenne un sondage sur vingt sera faux.

Crédits images

Pour citer cet article : **Jacques Istas**, **La fourchette d'un sondage**. *Images des Mathématiques*, CNRS, 2009. En ligne, URL : **http://images.math.cnrs.fr/La-fourchette-d-un-sondage.html**