Rapport Technique No. GIDE-2017-01

Aller au-delà des marges d'erreur dans les sondages d'intentions de vote

par

Jean-Marc BERNARD

GIDE

<jeanmarc@gide.net>

GIDE

17 rue La Noue Bras de Fer
44200 Nantes, France

31 mars 2017

Aller au-delà des marges d'erreur dans les sondages d'intentions de vote

Jean-Marc BERNARD

GIDE

<jeanmarc@gide.net>

31 mars 2017

Résumé

Dans un sondage, à un pourcentage observé, $Fobs$, on associe une "marge d'erreur". L'incertitude que décrit cette marge d'erreur peut s'exprimer par un énoncé probabiliste simple sur le pourcentage réel sous-jacent, $Freel$, du type $Prob(Freel \in [Finf; Fsup]) = 0.95$. Nous montrons qu'il est possible d'aller au-delà, en s'intéressant à des énoncés probabilistes plus complexes, tels que : "Quelle est la probabilité que les intentions de vote pour Macron et Le Pen soient toutes deux supérieures à celles de tous les autres candidats ?".

Mots-clés : Sondage, Marge d'erreur, Inférence statistique, Inférence bayésienne, Probabilités.

1 Un sondage pré-électoral : conclusions usuelles

L'époque est riche en sondages sur les intentions de vote au premier tour des prochaines élections présidentielles du 23 avril 2017. En plus des pourcentages d'intentions de vote pour les divers candidats, on indique souvent la "marge d'erreur" associée à chacun de ces pourcentages. [*Note* : Par la suite, j'utiliserai le terme de "fréquence" plutôt que celui de "pourcentage".]

Je prendrai l'exemple des deux premiers sondages disponibles après la publication par le conseil constitutionnel des 11 candidats qualifiés (Wikipedia, 2017) :

- sondage IFOP-Fiducial, réalisé les 18-20 mars 2017, sur un échantillon de 1000 individus (en fait $n = 935$ inscrits sur les listes électorales) ;
- sondage ELABE, réalisé les 17-19 mars 2017, sur un échantillon de 3010 individus (en fait $n = 1995$ inscrits et qui expriment une intention de vote). .

Les fréquences observées (arrondies à 0.5% près), et les marges d'erreur associées (marges approchées au niveau de confiance 95%) pour ces deux sondages sont données dans les deux tableaux suivants. [*Note* : Les colonnes $MErr$ (marge d'erreur) et Eff (effectifs) ont été reconstituées à partir des données disponibles, et sans tenir compte d'éventuels redressements.]

Tableau 1 : Sondage IFOP-Fiducial (18–20 mars 2017)

Candidat	Code	$Fobs$	$MErr$	Eff
Arthaud	Art	1.0%	0.6%	9
Poutou	Pou	0.5%	0.5%	5
Mélenchon	Mel	11.5%	2.0%	107
Hamon	Ham	12.5%	2.1%	117
Macron	Mac	25.0%	2.8%	234
Lassalle	Las	0.5%	0.5%	5
Fillon	Fil	18.0%	2.4%	168
Dupont-Aignan	Dup	4.5%	0.8%	42
Le Pen	Pen	26.0%	2.7%	243
Asselineau	Ass	0.5%	0.5%	5
Cheminade	Che	0.0%	–	0
Total		100.0%		$n = 935$

Tableau 2 : Sondage ELABE (17–19 mars 2017)

Candidat	Code	$Fobs$	$MErr$	Eff
Arthaud	Art	0.5%	0.3%	10
Poutou	Pou	0.5%	0.3%	10
Mélenchon	Mel	13.0%	1.5%	259
Hamon	Ham	13.5%	1.5%	269
Macron	Mac	25.5%	1.9%	509
Lassalle	Las	1.0%	0.4%	20
Fillon	Fil	17.5%	1.7%	349
Dupont-Aignan	Dup	3.0%	0.7%	60
Le Pen	Pen	25.0%	1.9%	499
Asselineau	Ass	0.5%	0.3%	10
Cheminade	Che	0.0%	–	0
Total		100.0%		$n = 1995$

Avant de développer la proposition spécifique de cet article, il nous faut revenir sur cette notion de "marge d'erreur" et sur son interprétation probabiliste.

2 Inférence statistique : de l'échantillon à la population

2.1 Echantillon vs. Population

Concentrons nous sur l'analyse d'un seul sondage (IFOP-Fiducial). Si le sondage avait pu être poursuivi sans contrainte sur 100000, 1M, puis 44.8M d'individus (la population des votants potentiels, au 1er mars 2016, source INSEE), on obtiendrait les fréquences réelles, $Freel$, des candidats, dont les $Fobs$ sont le reflet certes, mais un reflet nécessairement incertain du fait d'un échantillonnage limité à n individus. Le problème est alors le suivant : ce sont les $Fobs$ sur

l'échantillon qu'on recueille, qu'on observe, alors qu'on est en fait intéressé par les $Freel$ sur la population.

[*Note* : Qu'il s'agisse de $Fobs$ ou de $Freel$, il s'agit toujours de fréquences d'intentions de vote à un instant / moment déterminé. Les fréquences $Freel$ seront sans doute différentes à un autre moment, et ne sont pas non plus les fréquences de vote effectif le jour du scrutin.]

2.2 Méthodes d'Inférence statistique

Les méthodes d'inférence statistique permettent justement d'établir un lien entre les $Fobs$ et les $Freel$, un lien non de certitude mais basé sur des raisonnements probabilistes. Ces méthodes permettent ainsi d'aboutir à des conclusions sur les $Freel$, avec un certain degré de probabilité. [Attention : on aboutit donc à des probabilités portant sur des fréquences ! Ne pas confondre, car les deux entités sont des nombres entre 0 et 1 et peuvent s'exprimer comme des pourcentages.]

[*Note* : Sans entrer dans les détails, l'hypothèse sous-jacente à ces méthodes est celle d'un échantillonnage au hasard dans la population. Mais il convient d'être prudent, car d'éventuels biais de sélection de l'échantillon ne donnent en fait "accès" qu'à une sous-population de la population générale.]

2.3 Interprétation probabiliste des marges d'erreur

Les marges d'erreur indiquées précédemment, sont en fait calculées avec une certaine probabilité - on dit aussi "garantie" ou "niveau de confiance" -, et peuvent s'interpréter de la façon suivante,

$$Prob(Fobs - MErr < Freel < Fobs + MErr) = 0.95, \tag{1}$$

soit pour le candidat Macron, pour qui on trouve $Fobs = 25.0\%$ et $MErr = 2.8\%$,

$$Prob(25.0\% - 2.8\% < Freel < 25.0\% + 2.8\%) = 0.95, \tag{2}$$

i.e. on a une garantie de 0.95=95% que $Freel$ soit dans l'intervalle $[22.2\%; 27.8\%]$.

2.4 Diverses approches pour l'inférence statistique

Il existe en fait plusieurs théories de l'inférence statistique, (inférence fréquentiste et inférence bayésienne, pour les plus connues), et chacune d'elle comporte des points de choix. Et selon la théorie qu'on adopte et les choix qu'on effectue, les conclusions probabilistes peuvent différer. De façon rassurante, dès que n est suffisamment grand et que chaque $Fobs$ n'est pas trop extrême (proche soit de 0%, soit de 100%), ces différences sont petites et les diverses approches conduisent en fait à des résultats convergents (cf. Bernard, 1996). Dans ce cas, on peut alors parler de "la marge d'erreur" (à une garantie/confiance donnée) sans ambiguité (au moins de façon approchée).

[*Note* : Selon la théorie adoptée, l'interprétation à donner aux probabilités varie : les probabilités peuvent être vues comme des fréquences de conclusion exacte, si on analysait un

grand nombre de sondages donnant des résultats identiques ; ou bien comme des degrés de croyance induits par un sondage unique, pouvant s'exprimer en termes de paris cohérents et équilibrés.]

2.5 Calcul des marges d'erreur

Pour une fréquence observée $Fobs$, obtenue sur n individus, et une garantie G, la formule approchée usuelle (approximation par une distribution normale) de la marge d'erreur $MErr$ est :

$$MErr \quad = \quad z_G \times \sqrt{\frac{Fobs(1-Fobs)}{n}}, \tag{3}$$

où $z_G = 1.645\,;\, 1.960\,;\, 2.575$ (selon que $G = 0.90\,;\, 0.95\,;\, 0.99$). Par exemple, pour $Fobs$=0.50, $n = 1000$ et $G = 0.95$ soit $z_G = 1.960$, on retrouve la classique marge d'erreur de 3% (en fait $MErr = 3.1\%$).

La marge d'erreur est d'autant plus petite que : *(i)* G est petit, *(ii)* $Fobs$ est extrème, *i.e.* proche de 0 ou 1, *(iii)* n est grand. Et inversement. Ceci est clairement visible dans les tableaux 1 et 2. Pour chaque sondage, où n est fixé, les marges d'erreur sont plus petites sur les fréquences faibles. Entre les deux sondages, on voit que, à fréquence observée identique, les marges d'erreur sont plus petites pour le sondage ELABE ($n = 1995$) que pour le sondage IFOP-Fiducial ($n = 935$), d'un facteur $\sqrt{2} = 1.4$ environ.

3 Probabilité d'un énoncé complexe : approche bayésienne

L'approche bayésienne de l'inférence autorise en fait à calculer la probabilité de n'importe quel énoncé d'intérêt, aussi complexe soit il. La marge d'erreur calculée usuellement se traduit par un énoncé simple, relatif à une seule fréquence, on l'a vu précédemment. Mais, il est en fait possible d'aller au-delà, et de considérer des énoncés portant sur plusieurs fréquences simultanément (voir Bernard, 1991 & 1998).

3.1 Qui serait au second tour ?

On constate par exemple que $Fobs[Mac]$ et $Fobs[Pen]$ dépassent les fréquences $Fobs$ de tous les autres candidats, dans chacun des deux sondages, ce qu'on peut exprimer par :

$$Min(Fobs[Mac], Fobs[Pen]) > Max(autres \;\; Fobs) \tag{4}$$

Cette propriété observée des données, en termes des $Fobs$, peut-elle être généralisée à la population entière ? Est-elle vraie aussi pour les $Freel$ sous-jacents ? La réponse consiste simplement à calculer la probabilité de l'énoncé correspondant sur les fréquences réelles :

$$\begin{aligned} &Prob\left(Min(Freel[Mac], Freel[Pen]) > Max(autres\ \ Freel)\right) \\ &\quad = 0.9993 \quad \text{(IFOP)} \\ &\quad = 1.0000 \quad \text{(ELABE)} \end{aligned} \tag{5}$$

Le passage de l'échantillon observé à la population est simple : sur l'échantillon on a des certitudes, telle ou telle propriété est vérifiée ou non ; sur la population, on ne peut avoir qu'une probabilité plus ou moins forte que la propriété analogue soit vérifiée. Ici, on a une bonne garantie, 0.9993 (IFOP) ou 1.0000 (ELABE), que la propriété soit vérifiée sur la population.

En notant que $Fobs[Mac]$ et $Fobs[Pen]$ sont toutes deux supérieures à 24.0%, on peut vouloir aller encore au-delà en considérant un énoncé plus fort :

$$Min(Fobs[Mac], Fobs[Pen]) > 24\% > Max(autres\ \ Fobs)$$

A nouveau, la réponse sera simplement la probabilité de l'énoncé sur la population correspondant :

$$\begin{aligned} &Prob\left(Min(Freel[Mac], Freel[Pen]) > 24\% > Max(autres\ \ Freel)\right) \\ &\quad = 0.6319 \quad \text{(IFOP)} \\ &\quad = 0.7635 \quad \text{(ELABE)} \end{aligned} \tag{6}$$

Ce nouvel énoncé est plus fort, a une portée plus grande, que le premier énoncé considéré. Ceci se traduit par une garantie plus faible, on passe de $Prob = 0.9993$ à seulement $Prob = 0.6319$ pour IFOP, et de $Prob = 1.0000$ à seulement $Prob = 0.7635$ pour ELABE. Si on peut dire que Macron et Le Pen sont au-dessus des autres, rien ne permet d'affirmer qu'ils dépassent chacun 24% de surcroît.

Une remarque s'impose ici : il n'y a pas de "vraie" probabilité. Chaque probabilité indiquée dans ce texte n'est que la traduction d'un certain état de connaissance. A chaque fois, il s'agit des connaissances provenant d'un seul sondage, en ne tenant aucun compte d'éventuelles informations extérieures à ce sondage. Ceci étant dit, si les deux sondages étaient jugés homogènes (*i.e.* ayant les mêmes $Freel$ sous-jacents), il serait possible d'en agréger les résultats bruts et d'en dériver d'autres probabilités, fondées quant à elles sur un état de connaissance plus fort ($n = 935 + 1995$).

3.2 Dans quel ordre seraient les 3 principaux candidats ?

Comme autre illustration, on peut s'intéresser à l'ordre des intentions de vote pour les trois principaux candidats. Quelle est la probabilité d'avoir l'ordre "$Mac > Pen > Fil$" ? Et d'avoir les autres ordres possibles ? Le tableau ci-dessous donne la réponse a cette question :

Tableau 3 : Ordre des 3 principaux candidats

	IFOP	ELABE
Ordre	Prob	Prob
$Mac > Pen > Fil$	0.3406	0.6239
$Mac > Fil > Pen$	0.0001	0.0000
$Pen > Mac > Fil$	0.6588	0.3761
$Pen > Fil > Mac$	0.0006	0.0000
$Fil > Mac > Pen$	0.0000	0.0000
$Fil > Pen > Mac$	0.0000	0.0000
TOTAL	1.0000	1.0000

Pour aucun des deux sondages, un des 6 ordres possibles n'atteint une garantie suffisante. Ceci est essentiellement du à la proximité des fréquences de Macron et Le Pen. Il n'est pas possible de trancher sur l'ordre entre ces deux candidats. Par contre la probabilité que Fillon soit derrière Macron et Le Pen est forte :

$$\begin{aligned} Prob &= 0.3406 + 0.6588 = 0.9994 \quad \text{(pour l'IFOP), et} \\ &= 0.6239 + 0.3761 = 1.0000 \quad \text{(pour ELABE).} \end{aligned}$$

On retrouve ici une probabilité quasiment égale à celle trouvée en 3.1. Ce n'est que la conséquence d'énoncés eux-mêmes très proches, compte-tenu des données observées.

3.3 Qui l'emporterait à gauche ?

Enfin, une question d'importance — pour le futur des partis de la gauche et du parti socialiste en particulier — est celle de l'ordre entre les candidats Hamon et Mélenchon. Dans chacun des deux sondages, on trouve :

$$Fobs[Ham] > Fobs[Mel]$$

Pour les fréquences réelles, on obtient les probabilités :

$$\begin{aligned} &Prob(Freel[Ham] > Freel[Mel]) \\ &\quad = 0.7476 \quad \text{(IFOP)} \\ &\quad = 0.6681 \quad \text{(ELABE)} \end{aligned} \tag{7}$$

Pour aucun des sondages, la garantie n'est suffisante pour parvenir à une conclusion suffisamment étayée.

[*Note générale* : Nous avons implicitement considéré qu'il fallait au minimum une garantie de 0.95=95% pour pouvoir conclure. C'est une valeur souvent utilisée, mais que rien n'impose. On peut vouloir prendre davantage de risques et se contenter de 0.90=90%, ou au contraire rechercher une quasi-certitude avec une garantie plus forte, de 0.99 ou 0.999.]

3.4 Logiciel BayCat

Tous les probabilités indiquées dans cet article ont été calculées à l'aide du logiciel BayCat-5.0 (Bernard, 2017). Ce logiciel implémente les méthodes d'inférence bayésienne pour des données catégorisées, et permet notamment le calcul des probabilités associées à tous les énoncés logiques (vrai ou faux sur les données observées) possibles.

[*Remarque théorique* : Nous avons évoqué à la section 2.4 les deux approches principales de l'inférence statistique, inférence fréquentiste et inférence bayésienne. Pour les énoncés simples — comme le calcul de la marge d'erreur sur une fréquence — ces deux approches conduisent à des résultats très proches, et même asymptotiquement identiques. Pour les énoncés plus complexes que nous avons considérés ici, seule l'approche bayésienne peut être adoptée de façon générale. Elle repose sur l'utilisation de distributions de Dirichlet, avec le choix d'une distribution initiale uniforme.]

Références

Bernard J.-M. (1991), "Introduction a l'inférence bayesienne ; Illustration pour l'inférence bayesienne et prédictive sur les fréquences", In Rouanet H., Lecoutre M.-P., Bert M.-C., Lecoutre B., Bernard J.-M., *L'inférence statistique dans la démarche du chercheur*, Bern : Peter Lang, pp. 121–153.

Bernard J.-M. (1996), "Bayesian Interpretation of Frequentist Procedures for a Bernoulli Process", *The American Statistician*, Vol. 50 no. 1, pp. 7–13.

Bernard J.-M. (1998), "Bayesian Inference for Categorized Data", In Rouanet H. et al., *New Ways in Statistical Methodology, From Significance Tests to Bayesian Inference*, Berne : Peter Lang, pp. 159–226.

Bernard J.-M. (2017), Logiciel BayCat Version 5.0.

Wikipedia (2017) : `https://fr.wikipedia.org/wiki/` , Phrase-clé : "Liste de sondages sur l'élection présidentielle française de 2017".