Sondons les sondages

Avner Bar-Hen

Kafémath 14 mars 2012

Présentation du secteur

- La statistique publique : 8000 employés dont 5800 à l'INSEE
  - Le Conseil national de l'information statistique(Cnis)
  - Le service statistique public(Insee et services statistiques ministériels)
  - L'Autorité de la statistique publique
- Un secteur privé qui ne connaît pas la crise
  - Près de 400 instituts d'étude de marché et d'opinion identifiés en France
  - Marché estimé de 2 milliards d'euros en 2010
  - Environ 12 000 personnes, hors enquêteurs
- L'opinion : une faible part de l'activité des instituts

Effectifs et chiffres d'affaires des principaux instituts de sondages

Instituts de sondages	**Nombre de salariés** (au 31/12/2009)	**Chiffre d'affaires en 2009** (en millions d'€)	**Part des sondages politiques dans le chiffre d'affaires*** (en %)
BVA	230	53	1 %
CSA	106	32	16 %
IFOP	159 (au 30/04/10)	35,2	20 à 25 %
IPSOS	574	98,7	1 %
LH2	95	19	3 à 5 %
Opinion Way	45	9,1	6 %
TNS-Sofres	559** (en 2006)	126** (au 31/12/2008)	NC
Viavoice	NC	0,8**	25 %

** données communiquées à vos rapporteurs par les instituts*
*** source Infogreffe*
NC = non communiqué à vos rapporteurs

« Une armée entière de critiques ne saurait arrêter les sondages » George Gallup, 1949

- La loi 77-708 du 19 juillet 1977 modifiée relative à la publication et à la diffusion de certains sondages d'opinion ne s'applique qu'aux sondages électoraux qui remplissent deux critères :
  - le sondage doit présenter un lien direct ou indirect avec un scrutin, à savoir "un référendum, une élection présidentielle ou l'une des élections réglementées par le code électoral ainsi qu'avec l'élection des représentants au Parlement européen"
  - les résultats du sondage doivent être soit publiés dans un support de presse écrite (y compris Internet), soit diffusés dans un organe de presse audiovisuelle.
- Selon le rapport Sueur-Portelli : un sondage se définit comme une opération visant à donner une indication quantitative des opinions, attitudes et comportements d'une population par l'interrogation d'un échantillon représentatif de celle-ci.

Représentativité

- Notion peu scientifique
- Souvent confondue avec le respect de certaines proportions (modèle réduit)

- Fluctuations d'échantillonnage : avec les mêmes probabilités d'inclusion, répéter q fois un sondage donnera q résultats différents
- Sans biais : si la moyenne des moyennes de tous les échantillons possibles est égale à la moyenne de la population (pas d'écart systématique)

Sondage aléatoire simple

- tirage équiprobable sans remise de n personnes parmi N ;
- estimation de la proportion d'intérêt p, variance et intervalle de confiance
- IC diminue avec n et augmente avec variance (ie proportionnel à $\frac{p(1-p)}{n}$ dans la plupart des modèles)

INTERVALLE DE CONFIANCE A 95% DE CHANCE

Et si l'effectif est...	Si le pourcentage trouvé est... 5 ou 95%	10 ou 90%	20 ou 80%	30 ou 70%	40 ou 60%	50%
50	6,2	8,5	11,3	13,0	13,9	14,1
100	4,4	6,0	8,0	9,2	9,8	10,0
200	3,1	4,2	5,7	6,5	6,9	7,1
250	2,8	3,8	5,1	5,8	6,2	6,3
300	2,5	3,5	4,6	5,3	5,7	5,8
350	2,3	3,2	4,3	4,9	5,2	5,3
400	2,2	3,0	4,0	4,6	4,9	5,0
450	2,1	2,8	3,8	4,3	4,6	4,7
500	1,9	2,7	3,6	4,1	4,4	4,5
600	1,8	2,4	3,3	3,7	4,0	4,1
700	1,6	2,3	3,0	3,5	3,7	3,8
800	1,5	2,1	2,8	3,2	3,5	3,5
900	1,4	2,0	2,6	3,0	3,2	3,3
1000	1,4	1,8	2,5	2,8	3,0	3,1
2000	1,0	1,3	1,8	2,1	2,2	2,2
4000	0,7	0,9	1,3	1,5	1,6	1,6
6000	0,6	0,8	1,1	1,3	1,4	1,4
10000	0,4	0,6	0,8	0,9	0,9	1,0

Sondages

- Les méthodes d'échantillonnage aléatoire supposent l'existence d'une base de sondage à partir de laquelle on tire aléatoirement (mais avec probabilité connue) un échantillon sans biais dont la taille a été déterminée à la suite de considérations sur le niveau de précision souhaité.
- Or, pour la majorité des enquêtes d'opinion on ne dispose pas de base de sondage.
- On fait alors en sorte de construire un échantillon dont la structure corresponde à la structure de la population toute entière, selon certains critères que l'on a préalablement choisi
- On suppose que si l'échantillon reproduit fidèlement certaines caractéristiques de la population étudiée (et peut donc être considéré, par abus de langage, «représentatif»), alors il sera également à même de reproduire d'autres caractéristiques non contrôlées et/ou contrôlables qui constituent l'objet même de l'enquête

Variabilité pour a méthode des quotass

- La probabilité d'être touché varie également avec l'accessibilité des personnes à interroger
- La précision des estimateurs par quotas n'est pas calculable, puisque aucune probabilité d'inclusion n'est connue. Par contre, le fait de respecter des proportions fixés à l'avance limite la marge de manoeuvre laissée à l'aléa.
- On peut donc supposer que la variance d'un sondage par quotas est une grandeur plutôt faible dès lors que la variable d'intérêt est bien expliquée par les critères sur lesquels on a basé les quotas

Pourquoi continue-t-on ?

- Ce n'est pas parce que l'on ne connaît pas la précision d'une estimation que cette estimation est mauvaise.
- De façon empirique nous avons d'innombrables exemples de résultats issus d'échantillons par quotas fort comparables à ceux fournis par des échantillons aléatoires

Quotas classiques : sexe, âge et profession de la personne de référence

Différence entre les bases Insee du recensement et les listes électorales

Niveau de diplôme dans l'enquête emploi INSEE 1995 et dans trois bases de données SOFRES (en %).

	INSEE (1995)	Prés. 1995	Prés. 1995	Régionales 1998
Sans diplôme	22.3	11.7	8.2	7.8
CEP	17.7	17.8	15.7	12.1
BEPC	9.3	8.9	10.4	30.7
CAP BEP	23.7	26.5	23.2	..
BAC	11.4	11.5	14	16.3
BAC + 2	7.9	10.6	12.1	15.3
Supérieur	7.7	12.5	15	17.7
Sans Réponse	0	0.5	1.3	0.2

Inactifs : environ 30% des sondés

Niveau de diplôme dans l'enquête emploi INSEE 1995 et dans trois bases de données SOFRES (en %).

Inactifs	INSEE (1995)	Prés. 1995	Prés. 1995	Régionales 1998
Sans diplôme	32.5	20.1	10.4	13.5
CEP	31.3	36.3	30.7	24.4
BEPC	6.7	10.7	15.4	30.7
CAP BEP	13.5	14.3	14.3	..
BAC	7.8	6.9	10.9	10.3
BAC + 2	3.8	3.7	4.2	7.4
Supérieur	4.4	7.4	11.4	13.7
Sans Réponse	0	0.	6 2.7	0

Diplôme peu relié aux différences droite/gauche mais relié au FN par exemple

Mode de l'enquête

- Face à face
- Téléphone
- On line

Pas de méthode parfaite

Bilan d'appels fourni par l'IFOP pour la troisième vague du Baromètre Politique Français (décembre 2006)

Total	83997	
Pas de réponse	18251	21.7
Occupé	1461	1.7
Disque France Télécom (Faux Numéro)	4708	5.6
Composition interrompue	960	1.1
Répondeur	13099	15.6
Fax/Modem	530	0.6
Autres	1292	1.5
ABANDON du fait de l'interviewé	1353	1.6
Entrevue complétée	**5240**	**6.2**
HORS QUOTA AVEC RAPPEL	1552	1.8
HORS QUOTA SANS RAPPEL	839	1.0
RAPPELER PLUS TARD	10914	13.0
(INTRO) Ca décroche	71	0.1
REFUS (sans autre indication)	14151	16.8
REFUS (de sondage en général)	6805	8.1
REFUS (lié au commanditaire de l'étude)	39	0.2
REFUS (lié à la durée du questionnaire)	1342	1.6
HORS CIBLE - Numéro de société	196	0.2
HORS CIBLE - Nationalité	471	0.6
HORS CIBLE - Non inscrit	623	0.7

Fausses déclarations

- Spirale du silence structurelle. Par exemple la tendance à dissimuler le vote pour des formations extrémistes.
- Spirale du silence conjoncturelle. Elle regrouperait des phénomènes propres à la dynamique d'une campagne électorale et concernerait les phénomènes de domination, de légitimation et donc de surestimation d'un parti ou d'un candidat attribuable à la conjoncture politique spécifique de cette élection.

On ne dispose pas de moyens sûrs de corriger ce phénomène conjoncturel puisqu'il n'y a pas, par définition, de valeur structurelle ou historique qui permettrait le redressement.

Incertitude

- Incertitude sur le vote (abstention et choix du candidat)
- Comment prendre en compte les choix politiques incertains ? Faut-il les éliminer d'emblée et ne fonder sa prévision que sur les seuls « certains d'aller voter et sûrs de leur choix ».

On ne retrouve pas, dans les sondages pré ou post-électoraux, des pourcentages de non-votants potentiels correspondant au taux d'abstentionnisme réellement observé.

Ceux qui renoncent finalement à exercer leur droit de vote se répartissent-ils à peu près à proportion des différentes forces politiques de sorte que leur absence n'affecte pas les équilibres politiques, qui ont été mesurés en prenant en compte leur intention de vote ? Il n'y pas de réponse claire à cette question.

Redressement

- Redressement : Opération qui consiste à pondérer les données, c'est-à-dire à modifier le poids des individus, qui initialement vaut 1, en fonction de critères socio-démographiques ou politiques.
  - Quotas socio-démographiques
  - Reconstitution des votes antérieurs

Fragmentation de l'offre politique,
volatilité des intentions de vote.
Lien entre votes et catégorie sociale ?

Filtrage

- Filtrage : deux catégories de questions sont généralement posées, la première portant sur la probabilité de participer au scrutin, la seconde sur la fermeté du choix partisan qui a été déclaré.

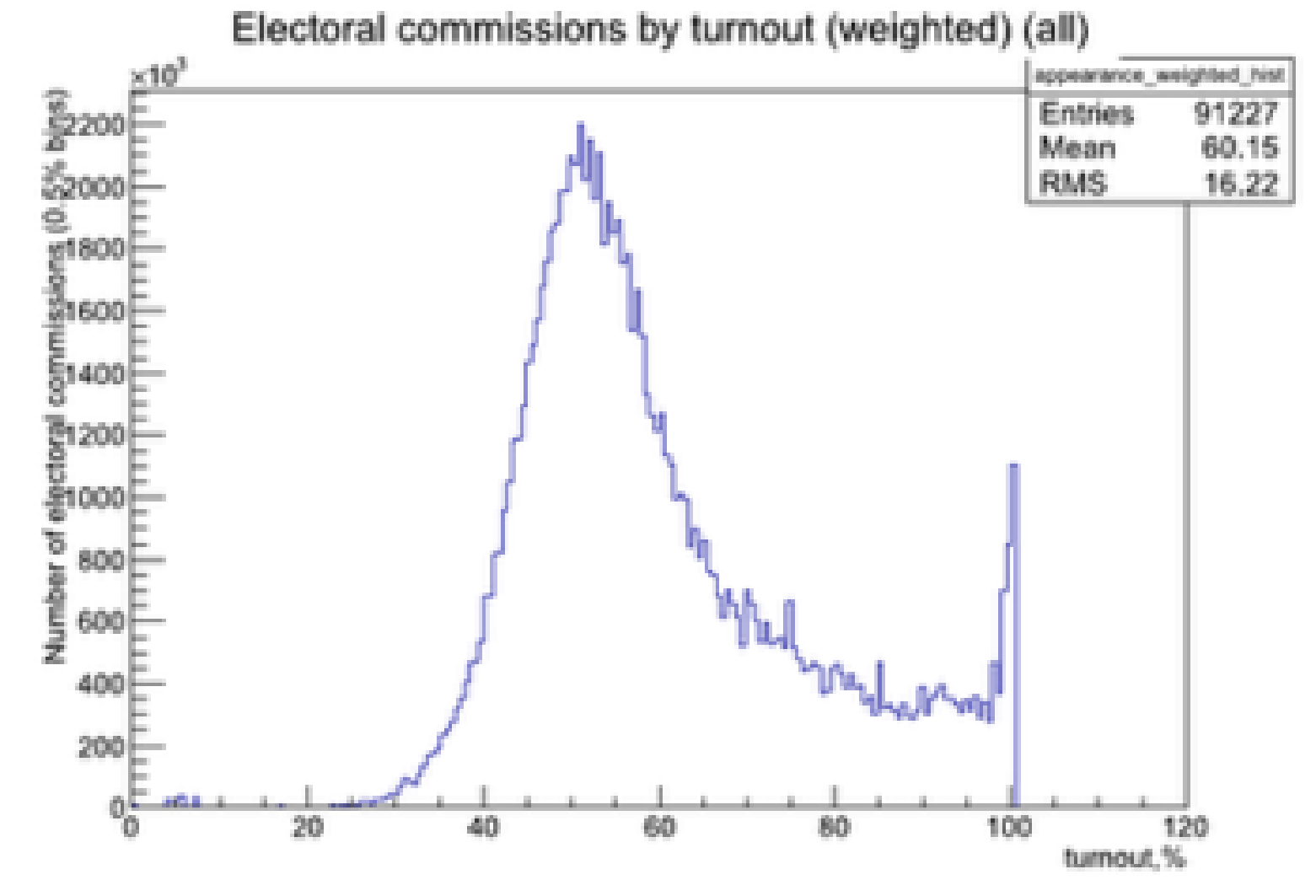
Les pondérations effectuées sur des reconstitutions de vote différentes donnent parfois des résultats très éloignés.

Panacher ces différents paramètres pour estimer l'état des forces politiques au moment du sondage demeure par conséquent un art difficile.

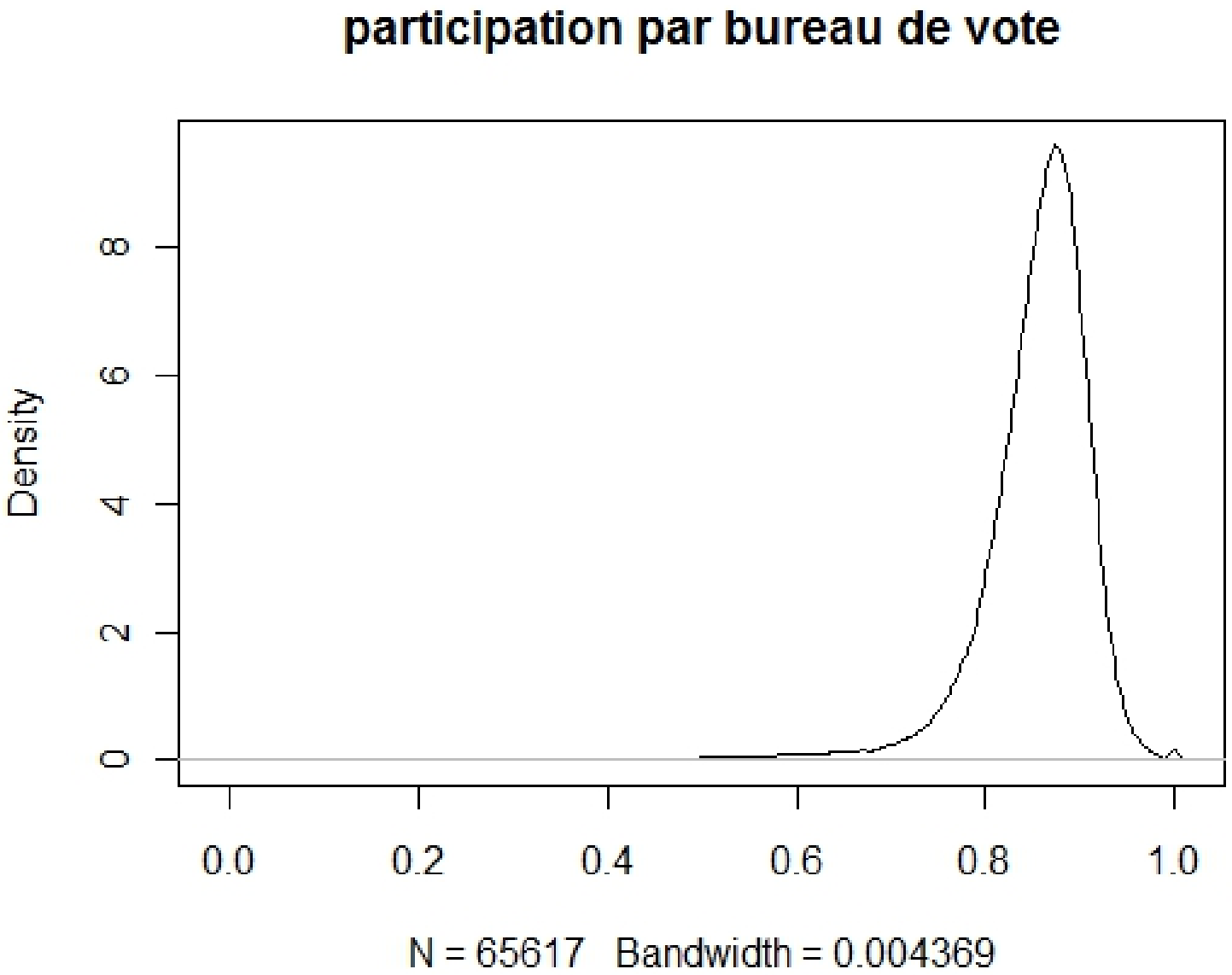
Mais dans l'établissement de ce résultat le sondeur est aussi très probablement influencé

Elections Russes

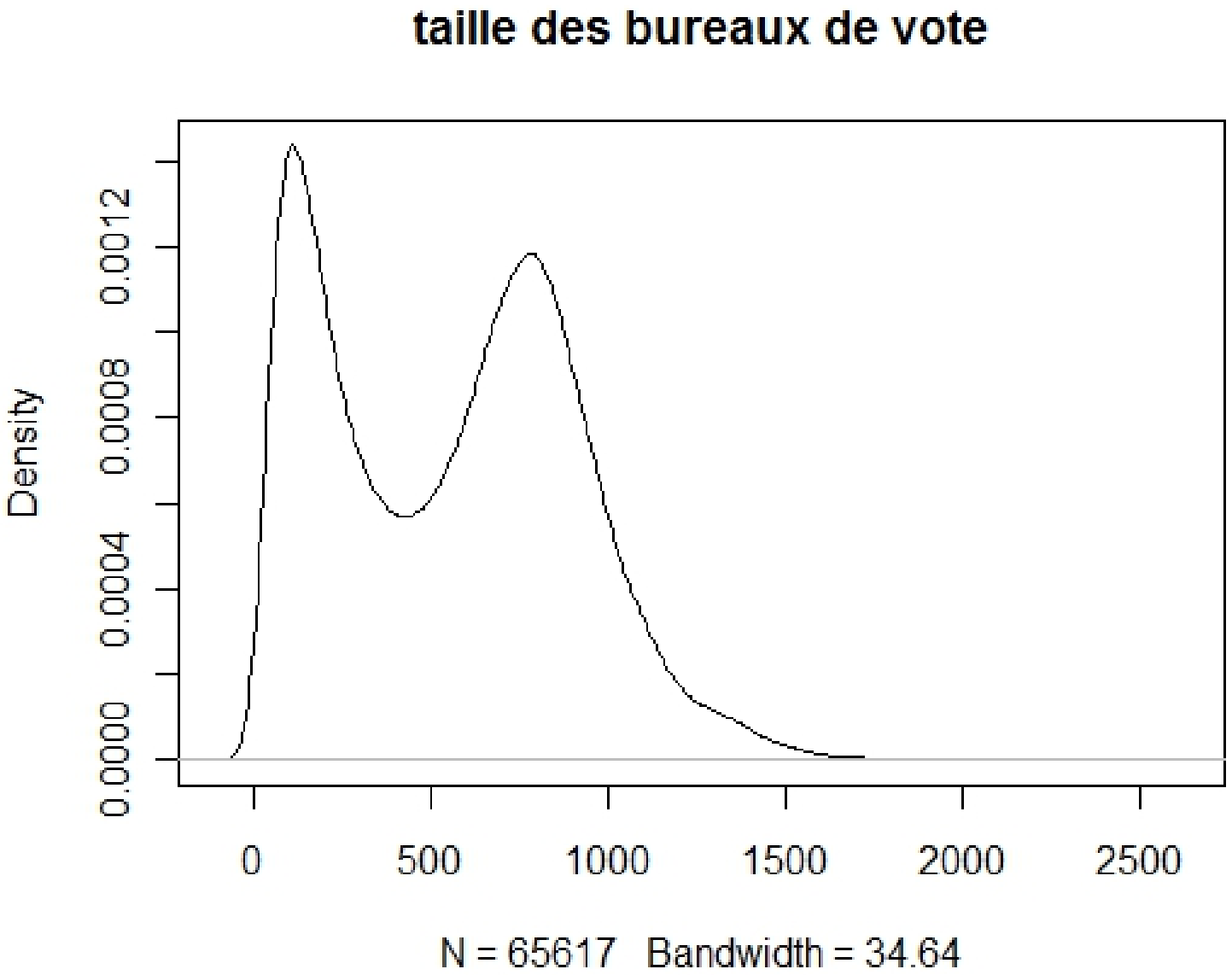
Elections Russes



Elections Françaises : participations



Elections Françaises : taille des bureaux de vote



Modèle binomial versus modèle de Markov

On considère en général un schéma de Bernouilli pour les élections : chaque personne a une probabilité p de voter pour un candidat. Les sondés sont indépendants

Schéma de Markov : L'urne contient initialement une boule blanche et une noire. On tire une boule au hasard, on note la couleur, on la remet dans l'urne et on rajoute une boule de la couleur de la boule tirée.

Après deux essais, nous avons pu sortir soit deux boules noires, soit deux boules blanches soit une boule noire et une boule blanche. Un calcul de combinatoire simple montre que ces trois combinaisons sont équiprobables.

Il est facile de montrer par récurrence qu'à l'issue de N tirages, on peut avoir $0, 1, \ldots, N$ boules blanches et tous ces événements sont équiprobables.

Bibliographie

- Rapport Sueur-Portelli
- Cours de Gilbert Saporta
- Image des mathématiques
  Elections russes et françaises
  Les sondages sont-ils devenus fous
- Séminaire "Le Bon usage des sondages d'intentions de vote"
  de la Société Française de Statistique le 19 mars à l'IHP