



HAL
open science

La quantité ne fait pas la qualité

Nicolas Blanchard

► **To cite this version:**

| Nicolas Blanchard. La quantité ne fait pas la qualité. 2017. hal-01448427

HAL Id: hal-01448427

<https://hal.science/hal-01448427>

Preprint submitted on 1 Feb 2017

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Copyright

Chapitre VI : La quantité ne fait pas la qualité

Nicolas K. Blanchard

Il entre dans toutes les actions humaines plus de hasard que de décision.
André Gide

Il est plus que raisonnable de penser que les athéniens avaient foi en la qualité de leur système démocratique. Le mot foi est approprié, les institutions politiques et religieuses étant alors indissociables, et le tirage au sort – au delà de l'intuition d'équité – était au final un appel aux puissances supérieures. De même, les républiques italiennes se rendaient compte de l'utilité et de la fiabilité de cet outil, mais n'avaient aucun outil leur permettant de l'étudier – d'où l'incroyable complexité du système vénitien. Plusieurs siècles ont passé, et la question du tirage au sort revient sur le tapis, que ce soit sous la forme du sorteo – une personne au pouvoir choisie au hasard dans la population, ce qui semble dangereux – ou de la sortition – une personne élue par un collège électoral restreint, composé de membres choisis au hasard dans la population générale. On va ici analyser ce second système. Heureusement pour nous, les mathématiques ont fait d'immenses progrès par rapport à celles des grecs et des vénitiens. Nous pouvons donc répondre en toute rigueur à la question suivante, première à poser avant d'étudier en détail ce système :

Un échantillon de personnes prises au hasard dans une population peut-il être représentatif ?

Pour répondre à cette question nous avons besoin de quelques outils mathématiques, provenant de l'étude des probabilités, notamment pour le jeu de pile ou face. Il faut noter que les jeux de hasard, la politique et les probabilités partagent des liens très étroits¹. En effet, les premiers mathématiciens à formaliser ces questions – le lombard Gerolamo Cardano et, un siècle plus tard le français Blaise Pascal – voulaient avant tout comprendre comment parier de manière efficace et gagner aux jeux de dés². De même, la pratique actuelle du loto vient des premières loteries organisées dans les républiques italiennes (Milan et surtout Gênes). Les membres du grand conseil étaient alors tirés au sort, comme à Venise, et la population se mit à parier de l'argent sur les prochains gagnants. Vu l'engouement populaire, le gouvernement ne tarda pas à monétiser l'affaire en faisant des tirages au sort plus fréquent – sans composante politique³. Pour analyser la représentativité d'un échantillon de personnes, nous allons faire deux hypothèses simplificatrices sur lesquelles nous reviendront : la présence de seulement deux partis, et l'absence d'abstention. Nous considérons donc désormais une population où chaque personne veut voter pour l'un des deux candidats possibles. Nous allons aussi avoir besoin de plusieurs théorèmes mathématiques, dont un plutôt célèbre, *la loi des grands nombres*. Celle-ci s'énonce de manière très intuitive :

Théorème 1 : Si l'on joue à pile ou face à répétition, sans tricher, notre taux de victoire se rapproche automatiquement de 50%.

Cette loi forme la base de l'étude des probabilités, et se généralise même : si la pièce est truquée et qu'on a 70% de chance de gagner à chaque lancer, alors notre taux de victoire se rapprochera de 70%. Cet outil n'est hélas pas suffisant pour étudier la représentativité d'un échantillon. Par contre, il nous permet de calculer facilement la probabilité qu'un échantillon soit entièrement composé de citoyens membres d'un certain parti. En effet, prenons comme exemple l'élection présidentielle de 2002, et simplifions les chiffres pour donner 80% à Jacques Chirac et 20% à Jean-Marie Le Pen⁴. En prenant un citoyen au hasard, on a une chance sur cinq d'avoir quelqu'un votant pour Le Pen. Si on prend un deuxième citoyen, on a à nouveau une chance sur cinq d'avoir quelqu'un votant pour Le Pen⁵. Comme l'avis du premier citoyen tiré au hasard ne dépend pas de celui du deuxième citoyen – en probabilités, on appelle cette propriété l'indépendance – la probabilité d'avoir deux électeurs d'affilée votant Le Pen est $\frac{1}{5} \times \frac{1}{5} = \frac{1}{5^2} = \frac{1}{25}$. De même, la probabilité de prendre trois personnes d'affilée votant pour Le Pen est égale à $\frac{1}{5} \times \frac{1}{5} \times \frac{1}{5} = \frac{1}{5^3} = \frac{1}{125}$, et ainsi de suite pour n'importe quel nombre d'électeurs d'affilée. Tomber sur un million d'électeurs de ce parti d'affilée est donc un événement hautement improbable, vu qu'il arrive avec probabilité $\frac{1}{5^{1000000}} \approx 10^{-698970}$, un zéro suivi d'une virgule, de 698970 zéros et d'un 1. L'entendement humain a du mal à se représenter des nombres aussi petits, donc on peut donner deux cas de figure ayant la même chance d'arriver. Tout d'abord, pour rester sur la thématique du vote, on peut la comparer

¹Et cela jusque dans les termes : aléatoire vient du mot latin "alea" signifiant "dé", hasard vient de l'arabe az-zahr, signifiant encore dé, et chance vient de "cadentia", la chute (des dés) en latin.

²Le fils de Cardano ne suivit apparemment pas ses conseils et ruina la famille avec ses dettes de jeu.

³Ce qui permit l'existence de ces loteries fut la création d'une entité impartiale chargée du tirage au sort (contrairement aux essais de loteries privées qui pouvaient facilement être truquées).

⁴Les résultats du scrutin étant 82,21% à 17,79%.

⁵Pour les personnes qui insistent sur la rigueur mathématique : vu la petite taille de l'échantillon on va considérer que la probabilité reste la même malgré les tirages successifs – en fait elle diminue encore plus. Nous reviendrons proprement là-dessus un peu plus tard.

à la probabilité qu'une urne remplie de bulletins soit détruite par une chute de météorite – sachant que quelques centaines de météorites tombent chaque année en France. Mais même cet événement reste beaucoup plus probable que le fait d'avoir un échantillon votant entièrement pour Le Pen. En effet, la probabilité de tomber sur un tel échantillon est du même ordre de grandeur que la probabilité que la majorité des urnes se retrouvent détruites par des météorites à la même heure lors d'une élection présidentielle⁶. Pour donner une autre situation improbable, on a plus de chance de gagner tous les tirages EuroMillions pendant 800 ans que d'avoir un échantillon intégralement frontiste. Enfin la probabilité d'avoir un échantillon où tout le monde vote pour Chirac est elle aussi extrêmement faible, similaire à la chance de gagner tous les tirages EuroMillions pendant un peu plus d'un siècle.

Tous ces calculs nous montrent que nos craintes précédentes sont infondées, voire même absurdes. Cependant, ce qui nous intéresse n'est pas de savoir la probabilité que cet échantillon soit unanimement frontiste, mais bien la probabilité que l'erreur soit suffisante pour changer le résultat de l'élection. En pratique, que le parti gagnant reçoive 52% ou 99% des voix ne change rien au fait que son candidat devienne président⁷. On doit donc calculer non pas la probabilité que l'échantillon soit entièrement composé d'un parti, mais la probabilité que l'erreur soit suffisamment grande pour changer le résultat de l'élection. Le problème est que nous ne connaissons pas à l'avance l'écart entre les deux partis. Heureusement, nous avons un outil, certes beaucoup plus complexe, capable de nous donner avec une très haute précision cette marge d'erreur : la borne de Chernoff⁸. En considérant une population votant pour les candidats A ou B, on a la formule suivante, hélas peu ragoûtante :

$$\Pr(X > y + z) \leq e^{-2y^2n}$$

Heureusement, une fois les termes explicités, cela devient un peu plus simple :

- X représente ici la proportion de personnes votant pour le candidat A dans notre échantillon.
- y représente la proportion de personnes votant pour le candidat A dans l'ensemble de la population.
- z correspond à la marge d'erreur considérée.
- n correspond au nombre de personnes dans l'échantillon.

La partie gauche de l'équation, $\Pr(X \geq y + z)$, représente donc la probabilité que la proportion de personnes votant A dans l'échantillon soit trop grande. Plus précisément, elle correspond à la probabilité que l'écart entre la proportion réelle et la proportion attendue de personnes votant pour A soit supérieure à la marge d'erreur prévue, z . La partie droite de l'équation donne une estimation sur la chance que ça arrive. Regardons cela sur un exemple avec des chiffres, en prenant :

- $y = 0.5$ (le parti A correspond à la moitié de la population),
- $z = 0.1$ (on regarde la probabilité qu'il y ait au moins 10% de marge d'erreur), et
- $n = 1000$ (il y a 1 000 personnes dans notre échantillon).

On cherche à trouver la probabilité, dans une population générale votant autant pour A que pour B, d'avoir un échantillon de 1 000 personnes comprenant au moins 60% de personnes votant pour A. En appliquant ces nombres dans la formule, on obtient ce qui suit :

$$\Pr(X > 0.6) \leq e^{-2(0.1)^2 \times 1000} \approx \frac{2}{1000000000}$$

Cela veut dire que la probabilité qu'au moins 60% de l'échantillon vote pour A est inférieure à $e^{-2(0.1)^2 \times 1000}$ – c'est à dire à deux chances sur un milliard. L'intérêt de cette formule est que la probabilité d'avoir une telle erreur descend extrêmement vite avec la taille de l'échantillon: pour un échantillon de 2000 personnes, la probabilité d'avoir 10% d'erreur chute à quatre chances sur un milliard de milliards. C'est donc l'outil qu'il nous faut pour analyser de manière suffisamment précise la représentativité d'un échantillon.

⁶Le calcul rapide se fait à partir du nombre d'urnes et des données sur les météorites de l'astronome Alan Harris.

⁷Même si ça affecte la politique menée par le candidat.

⁸Une lectrice très avertie – et probablement mathématicienne – dirait qu'on ne peut pas appliquer la borne de Chernoff, vu que nous n'avons pas prouvé l'indépendance. Elle se rendrait alors compte que corriger cette simplification ne ferait que diminuer encore plus la probabilité d'erreur – en complexifiant les calculs.

En mathématiques, la première étape de l'analyse d'un processus de ce type est généralement d'étudier le pire cas : dans quelle configuration le système a-t-il le plus de chances de commettre une erreur ? Ici, nous avons de la chance – pour l'analyse – vu que l'erreur est maximale quand les deux candidats se partagent équitablement la population, ayant chacun 50% des votes⁹. Pour montrer la fiabilité du système, il suffit donc de considérer le pire cas possible d'application : durant le deuxième tour de l'élection présidentielle de 1974. Cette élection est la plus serrée de l'histoire de la cinquième république, avec 50,81% des votes pour Valéry Giscard d'Estaing, et 49,19% pour François Mitterrand. Il se trouve que même dans ce contexte critique, un échantillon d'un million de votants aurait largement suffi pour être sûr d'élire le bon candidat. Par simple application de la formule, on peut trouver la probabilité que l'échantillon vote pour Mitterrand. Celle-ci aurait été inférieure à une chance sur 100 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 (un '1' suivi de 56 zéros¹⁰). Même avec un échantillon de 100 000 personnes, la probabilité d'erreur aurait été proche de un sur 1 000 000.

On peut enfin aller plus loin, en faisant abstraction du candidat. Séparons simplement la population en deux catégories sociales, selon un critère quelconque (qui peut être l'âge, le genre, le fait d'aimer ou pas les choux de Bruxelles à la menthe). On veut estimer la probabilité qu'une des deux catégories soit mal représentée (c'est-à-dire qu'il y ait une forte différence entre la proportion de personnes de cette catégorie dans l'échantillon et parmi la population dans son ensemble). Dans ce cas, le théorème précédent nous permet de montrer un autre résultat :

Théorème 2 : Pour toute catégorie sociale, la probabilité qu'elle soit représentée dans notre échantillon avec une erreur de plus de 0.5% est inférieure à une chance sur mille milliards de milliards (1 suivi de 21 zéros).

C'est une conséquence directe du théorème précédent. En effet, il suffit de remplacer "voter pour un candidat" par "être membre de la catégorie sociale considérée". On sépare à nouveau la population en deux et la formule s'applique. Et cela a une conséquence importante : même en prenant 5000 groupes sociaux différents (en faisant des catégories pour les tendances politiques, les classes sociales, le genre, la religion, le statut marital, les goûts musicaux...), la probabilité d'avoir ne serait-ce qu'un groupe mal représenté reste infime (de l'ordre d'un sur un milliard de milliards)¹¹. Une amélioration possible de la sortition est souvent proposée, en y introduisant la méthode des quotas (prendre 500 000 hommes et 500 000 femmes au hasard par exemple). Il est important de se rendre compte que cette méthode est strictement pire : on améliore peut-être (et encore) la parité, mais le taux d'erreurs dans toutes les autres catégories augmente. Et la représentativité de l'échantillon diminue d'autant plus qu'il y a de quotas à respecter. Garder le hasard pur et sans manipulation est la seule manière permettant de satisfaire non seulement les critères visibles (comme la parité homme/femme) mais aussi tous les critères invisibles (assurant la représentation équitable des personnes à l'intersection de plusieurs catégories spécifiques, comme les hommes protestants ouvriers de moins de 35 ans).

On voit que, mathématiquement, un échantillon d'un million de personnes paraît donc tout à fait représentatif de la population générale. Il faut cependant faire attention car nous avons au début fait deux hypothèses simplificatrices, ce cas d'étude n'est donc pas réaliste. La première hypothèse, sur l'existence de seulement deux partis, peut être rejetée facilement. En effet, prenons un système avec un nombre quelconque de partis et considérons un de ces partis en particulier. Dans ce cas on peut séparer la population en deux catégories : ceux qui votent pour le parti choisi et les autres. Nous venons de créer deux catégories sociales, et le théorème 2 s'applique donc. Cela étant vrai pour chacun des partis, ils seront tous représentés de manière équitable. On peut éliminer la première hypothèse vu que, peu importe le nombre de partis, un échantillon sera toujours représentatif de ces partis. La deuxième hypothèse – concernant l'absence d'abstention – est plus difficile à éliminer car elle vise un comportement social et non pas une réalité mathématique, et nous y reviendrons bientôt. A part l'abstention, deux questions subsistent encore :

- Introduire du hasard dans le système ne met-il pas en danger la pureté démocratique ?

⁹Si un candidat a une petite minorité, on dispose d'une variante de la borne de Chernoff donnant des probabilités d'erreur encore plus faibles.

¹⁰Ce nombre se prononce "cent nonillions", ou cent milliards de milliards de milliards de milliards de milliards de milliards.

¹¹Pour les férus de mathématiques, cela se démontre en utilisant l'inégalité de Boole (aussi appelée union bound).

- Si notre démarche mathématique est correcte, comment expliquer les performances désastreuses des sondages ?

Pour la première question, la réponse se fait en deux temps. Le recours au hasard est souvent perçu comme une défaite, et un comportement irrationnel. Olivier Dowlen a introduit récemment une notion très utile, celle d'a-rationnalité, qui peut s'expliquer sur un exemple très simple, le jeu de pierre-feuille-ciseaux. Supposons que vous vouliez jouer avec une championne du jeu, experte en psychologie humaine¹². Cette personne est capable de deviner n'importe quelle stratégie que vous puissiez concevoir. Elle ne peut cependant rien faire si vous décidez de jouer au hasard (par exemple en ayant lancé des dés avant votre match pour décider quoi jouer). Cette a-rationnalité vous permet d'éviter de perdre face à un adversaire ayant des moyens supérieurs. De même, pour la sortition, cela permet d'éviter un certain nombre de phénomènes de corruption, vu que même en achetant de nombreuses voix, la personne voulant acheter l'élection n'est pas sûre d'y arriver. Enfin, l'argument de pureté démocratique ne tient pas. En effet, si nous étions sûrs que le décompte était parfaitement correct lors d'une élection, nous pourrions utiliser cet argument de pureté. Ce n'est hélas pas le cas, comme l'a montré une étude récente de Chantal Enguehard et Jean-Didier Graton. Les taux d'erreurs – dans les bureaux avec urnes et décompte en présence de plusieurs personnes – sont proches de 0.2%, avec près de 10% des bureaux de vote concernés (l'erreur ne portant généralement que sur quelques bulletins)¹³. L'erreur moyenne causée par le recours à la sortition étant de l'ordre de 0.1%, ces craintes semblent infondées.

La deuxième question est beaucoup plus intéressante, et est liée à un phénomène relativement difficile à étudier. Le problème des sondages vient du fait que la majorité de la population n'a pas envie d'y répondre, et que cela touche certaines catégories sociales en particulier (les réponses des financiers et des SDFs sont particulièrement rares). De plus, ces sondages ne se font de toute manière que sur des échantillons de petite taille (un millier de personnes le plus souvent, correspondant déjà à des marges de plusieurs pourcents).

Pour remédier à cela, les organismes de sondages utilisent la méthode des quotas. Cette méthode consiste à séparer la population en un certain nombre de catégories, et à prendre un échantillon tel que la proportion de personnes dans chaque catégorie corresponde à la proportion parmi la population générale. C'est certes mieux que rien (surtout avec des taux de réponses souvent entre 5% et 10%), mais cela introduit de très nombreuses failles, notamment parce qu'on suppose généralement qu'à l'intérieur de ces catégories les opinions sont homogènes. Vu que les données sur la proportion de personnes dans ces catégories (parmi l'ensemble de la population) sont imprécises, on ajoute encore une autre source d'erreur.

Si nous voulons que le choix d'un représentant politique soit laissé à un collège électoral restreint (changeant naturellement à chaque élection), nous devons étudier un facteur qu'on avait mis de côté jusqu'ici : l'abstention. Si le taux de participation est élevé dans les élections présidentielles (aux alentours de 80%), il est faible – et diminue – pour les législatives (55%) les européennes (40%). Cependant, il est ridicule pour les sondages, comme dit auparavant, et malgré la méthode des quotas, cela introduit suffisamment d'erreur pour diminuer fortement leur utilité. Si la qualité des sondages a baissé depuis quelques décennies, c'est aussi lié à ce taux de réponse : il était proche de 60% au milieu du 20ème siècle, et baisse continûment, (encore plus depuis l'apparition des sondages en ligne). Nous avons donc montré que l'échantillon est représentatif, mais cela ne veut pas dire qu'il votera de la même manière que la population dans son ensemble. En effet, plusieurs effets interviennent, dont trois paraissent majeurs :

- Il y a un lien entre taille du corps électoral et abstention : plus il y a de votants, plus l'abstention est forte. Contrairement à quelqu'un répondant à un sondage (qui n'a aucun

¹²Il y a plusieurs personnes sur Terre qui gagnent la plupart de leurs parties, et un certain nombre de champions, sur ce qui semble être un jeu de hasard mais se trouve être au final une compétition de psychologie pratique. La partie la plus chère de l'histoire fut jouée entre deux maisons d'enchères – Christie's et Sotheby's – pour avoir le privilège de vendre la collection privée d'une entreprise japonaise – contenant entre autres des toiles de Cézanne, Picasso et Van Gogh – et récupérer une commission de plusieurs millions.

¹³On peut comparer cela aux taux d'erreurs des machines de vote (proche de 0.6%), qui restent heureusement rares en France, et au taux moyen lors d'élections d'autres pays, dépassant généralement les 0.5%, preuve de la fiabilité (relative) de nos institutions.

pouvoir), la personne se retrouvant membre de l'échantillon ressentira de manière amplifiée le poids de son devoir civique¹⁴. Un collège électoral restreint donnerait donc une forte responsabilité aux votants, qui s'abstiendrait sans doute moins. Une multitude d'études montre que les taux d'abstentions sont les plus forts parmi les centristes et ceux ayant peu d'opinions politiques, une baisse de l'abstentionnisme diminuerait donc la puissance des partis extrémistes¹⁵ – que cela soit un bien ou un mal.

- Il y a aussi un lien entre pouvoir politique et investissement temporel, lié à un effet connu en psychologie : l'ignorance rationnelle. Les citoyens ressentent de plus en plus que leurs choix n'auront pas d'impact et nécessitent un grand investissement temporel – nécessaire pour se renseigner en détail sur les différents points critiques. Dans ces conditions, leur intérêt rationnel peut être simplement d'ignorer le sujet. Ce coût élevé (en temps) et l'effet négligeable de son choix font que la citoyenne moyenne va s'abstenir, voter blanc ou au hasard, sans s'être renseignée¹⁶. Un collège électoral réduit ne change pas le coût de l'investissement temporel, mais il change drastiquement l'impact du votant, augmentant donc probablement son intérêt.
- Il faut contraster à cela l'aspect social du vote : il n'est pas rare que les citoyens aillent voter en groupe (famille ou amis), et avant même le vote, la perspective du devoir civique pousse à une délibération citoyenne en petits groupes. Cela disparaîtrait sans doute avec la sortition, et pourrait mener à un désengagement total de la population pour les questions politiques. Cette éventualité n'est pas nécessairement négative mais paraît dangereuse en tout état de cause.

Au delà du risque lié au désengagement politique, un deuxième danger nous guette, beaucoup plus présent : la corruption. En effet, une fois les membres de l'échantillon choisis, ces derniers – qui ont désormais un pouvoir politique non négligeable, surtout avec des petits échantillons – peuvent vendre leur voix. Pour empêcher tout risque de corruption, les Athéniens avaient une méthode simple : les personnes tirées au sort allaient immédiatement se réunir pour débattre puis voter sur le problème à l'ordre du jour, sans aucune intervention possible d'un membre extérieur, et sans pause. Ne serait-ce que pour des questions logistiques, cela serait impossible aujourd'hui, et il faut donc trouver une parade. Heureusement, celle-ci existe, et nous vient d'avancées technologiques très récentes¹⁷.

¹⁴Il n'y aurait pas "d'effet du témoin" et de dilution de la responsabilité : le citoyen ne peut pas compter sur les autres pour faire son devoir civique.

¹⁵Aux élections régionales de 2015, 31% des personnes non affiliées à un parti ont voté, par rapport à 56% au FN et 62% à l'UDI – qui avait été créé peu de temps auparavant, d'où sa bonne performance.

¹⁶Il y a aussi une corrélation étudiée entre polarisation du votant et temps passé à s'informer. Cela ne veut pas dire que s'informer radicalise les citoyens, mais peut-être qu'être extrémiste pousse les citoyens à s'informer – ou à avaler de la propagande.

¹⁷Nous utilisons le terme technologique au sens large, la technologie en question correspondant principalement à un développement mathématique étant tout à fait compatible avec l'usage des bulletins papiers.