Quelques remarques sur les pourcentages et les statistiques

A l'heure actuelle de nombreuses analyses s'appuient sur ce qu'il est convenu d'appeler des «chiffres». J'aimerais revenir sur certaines de ces analyses et sur l'utilisation des statistiques comme «photographies» de l'opinion publique et surtout déplacer le débat : ce n'est pas une querelle d'experts, mais un problème de société. Au centre de ce débat se trouve la culture scientifique générale qui peut être apportée par la formation, l'école et aussi transmise par les médias.

Le pourcentage comme photographie

La première notion que l'on rencontre est celle de pourcentage. Les médias véhiculent en grand nombre des pourcentages pour illustrer un propos ou donner une information. On est souvent choqué par tel ou tel pourcentage et, en effet, certains d'entre eux peuvent faire peur. La question se pose donc de savoir ce que représentent ces pourcentages et même, plus simplement, la confiance dont on peut les créditer.

Par définition même un pourcentage est le rapport de deux quantités, le rapport d'une partie sur un tout. C'est donc une notion relative. De plus la synthèse opérée par l'énoncé d'un tel nombre est évidemment simplificatrice et risque à la fois de masquer de vrais problèmes ou d'en stigmatiser à l'excès. Et plus la situation est complexe, plus il est dangereux de ne se référer qu'à un seul indicateur pour se forger une opinion. Le premier danger quand on entend un pourcentage est de croire qu'il constitue une probabilité qu'un certain événement ait lieu. Quand on entend par exemple : « ce lycée a un taux de réussite de 97% au bac » ou bien « en France 8% des femmes perdent leur virginité sous la contrainte » ou encore « 1 français sur 6 vote pour l'extrême droite », on peut se dire que mettre son enfant dans ce lycée augmente ses chances de réussite ou penser qu'en croisant une centaine de femmes ou d'adultes, s'en trouveront parmi elles et eux qui correspondent à la description faite. Ce serait aller bien vite en besogne ! Cela reviendrait à penser que la probabilité de l'événement décrit est une probabilité uniforme sur la population, autrement dit qu'elle est la même partout où on peut la mesurer : que ce soit en ville ou à la campagne, dans n'importe quelle région, dans un train ou dans la rue etc. Il n'en est rien !

En ce qui concerne la réussite au bac plusieurs cas peuvent se produire et chacune des situations que je vais décrire (et il en existe bien d'autres) appelle en fait une analyse complémentaire, chiffrée ou non. Le taux de réussite peut être global, toutes sections confondues. L'une d'elle pourrait aller bien en-deça de l'espérance promise. Le lycée pourrait très bien faire des classes de niveau, ainsi, même au sein d'une section donnée, neuf classes pourraient avoir un taux de réussite de 100% et une autre de 70%. Le lycée pourrait aussi faire de la sélection ou de l'écrémage, ainsi ne parviendraient au bac que les élèves dont le lycée estime qu'il est certain qu'ils l'obtiendront, les autres étant orientés ailleurs (par exemple vers une section n'existant pas dans le lycée, histoire de ménager toutes les susceptibilités). Enfin 50% des bacheliers pourraient avoir obtenu leur bac en rattrapage à l'oral ou encore avoir des difficultés sérieuses à l'université. La question que doit se poser un parent ou le système éducatif s'il a envie de connaître l'apport d'un lycée donné en termes quantifiés c'est ce que fait le lycée de mieux que les autres et comment : sont à ce titre intéressants des indicateurs comme le taux d'accès au bac (combien d'élèves entrant en seconde accèdent en classe de terminale), le taux de redoublement, le taux de réussite par section, notamment en comparaison avec les taux nationaux ou départementaux par section. Encore plus informatif est la « plus value » : l'idée est de décomposer la population d'un lycée en plusieurs catégories correspondant à divers critères (par exemple l'âge et la classe sociale), de calculer au sein du lycée le taux de réussite pour chacune de ces catégories séparément et de les comparer aux taux nationaux pour ces catégories. On pourra alors comparer le taux obtenu par le lycée en comparaison avec le taux pronostiqué à l'échelon national et on verra ainsi parfois des lycées réputés faire moins bien que ce que l'on pourrait attendre d'eux compte tenu de la classe sociale de ses élèves et de leur âge moyen. Enfin il ne faut pas perdre de vue qu'une donnée quantifiée ne remplace pas une analyse plus fine en termes notamment éducatifs ou sociaux : les projets d'établissement ou l'encadrement extra-scolaire doivent rester des critères importants dans l'évaluation du système éducatif.

L'exercice que l'on vient de faire pour les lycées pourrait être fait dans les deux autres exemples cités. Ce qu'il faut garder à l'esprit c'est qu'un pourcentage brut n'est rien d'autre qu'une moyenne. Cette moyenne décrit une situation dans son ensemble, de façon statique (on parle d'indicateur de position). Un autre, tout aussi intéressant, est la médiane. Il s'agit de la valeur située au milieu d'une plage de donnée lorsqu'on les range dans un ordre croissant. Par exemple si l'on observe la vitesse moyenne sur une autoroute et que l'on trouve 110km/h, on peut être rassuré. Mais si l'on rajoute que la vitesse médiane est 130km/h, autrement dit qu'un conducteur sur deux dépasse la limitation de vitesse, on peut être effrayé. Et pourtant ces deux informations ne sont pas incompatibles. Il est enfin d'autres indicateurs (écart-type, quartiles, déciles) qui décrivent la façon dont se répartissent les données autour de la moyenne : ce sont les indicateurs de dispersion. On peut les relier à la notion de barre d'erreur en statistiques et nous y reviendrons.

Le pourcentage comme révélateur d'évolution

Après avoir écarté le danger de croire qu'un pourcentage décrit une probabilité uniforme que l'on peut donc expérimenter autour de soi, la seconde tentation bien naturelle avec une information chiffrée est de vouloir s'en servir pour faire des comparaisons. Or, comme on l'a dit un pourcentage est une donnée relative. Ainsi la comparaison de pourcentages issus de situations différentes (c'est-à-dire non relatives à la même situation globale, au même « tout ») n'a aucun sens a priori. Prenons cette fois-ci l'exemple des sondages. Parler de progression dans les sondages est déjà problématique, nous y reviendrons, mais comparer des scores à diverses élections est totalement creux. C'est oublier des données très importantes comme le taux d'abstention et surtout la taille de la population. Pour bien faire il faudrait se rapporter à une population idéale et constante pour qui on pourrait parler d'évolution dans le choix de vote. C'est bien entendu impossible si l'on pense aux personnes qui sont décédées entre deux élections, aux jeunes qui ont acquis le droit de vote, à ceux qui ont été radiés des listes ou au contraire ont pensé à s'y inscrire. Et il n'y a pas de vraies raisons de penser que toutes ces personnes expriment

des opinions en moyenne identiques à celles des autres électeurs. De toute manière il faut déjà songer à préciser l'ensemble de personnes que l'on souhaite décrire. Quand on dit « 1 français sur 6 vote pour l'extrême droite », on est déjà dans l'erreur car le mot « français » n'est pas pris dans son acception usuelle mais est un raccourci pour « français inscrit sur les listes électorales ayant exprimé son vote », ce qui est un ensemble bien plus petit de personnes et qui se démarque bien évidemment de l'ensemble des personnes que l'on peut croiser dans la rue quand on se promène quelque part en France.

Si l'on veut tenter de se rapporter à une mesure commune, il faut diviser le nombre de suffrages obtenus par un candidat non pas par le nombre de suffrages exprimés mais **par le nombre d'inscrits**. On peut en effet espérer que ces inscrits reflètent toujours un peu de la même façon l'ensemble des électeurs potentiels (i.e. ceux qui ont le droit de vote, et non pas ni les français, ni les personnes résidant en France). Pour le premier tour du 21 avril 2002, on obtient alors les scores suivants, à comparer à ceux du 23 avril 1995 :

	1er tour 1995		**1er tour 2002**	
	Inscrits	Exprimés	Inscrits	Exprimés
	39 992 912	76.2 %	41 196 339	69.2 %
	Suffrages	Pourcentage	Suffrages	Pourcentage
J. Chirac	6 348 375	15.9 %	5 666 440	13.7 %
L. Jospin	7 097 786	17.8 %	4 610 749	11.2 %
J-M. Le Pen	4 570 838	11.4 %	4 805 307	11.7 %
P. De Villiers	1 443 186	3.6 %		
B. Mégret			667 123	1.6 %

Il ne faut donc pas hâtivement conclure à la progression de l'électorat pour le Front National : il est globalement constant. Sur les trois dernières élections présidentielles (celle-ci incluse), l'électorat a progressé de moins de 500 000 personnes, tandis que les inscrits ont augmenté d'un peu plus de 3 millions. Par conséquent **la proportion l'électorat frontiste est essentiellement constante** (de l'ordre de 11.5 %). Ce sont par contre les autres formations (et notamment le PS) qui, elles, ont perdu massivement des électeurs. Bien entendu le résultat du premier tour peut faire réfléchir dans plusieurs directions. L'une d'elles est la façon dont fonctionne une élection présidentielle : vote-t-on pour soutenir un candidat et ses idées même si l'on pense qu'il n'a aucune chance de passer au second tour ou vote-t-on pour désigner les deux personnes qui seront présentent au second tour, indépendamment du reste ? Les résultats quantifiés montrent que la question se pose réellement mais pour autant ces « chiffres » ne répondent pas à la question ! Une autre direction de réflexion est sociale. Comment peut-on décrire la situation et ce qui a pu être exprimé lors du vote ? Il faut alors, comme on l'a déjà esquissé pour les résultats au bac, chercher d'autres indicateurs. On peut découper l'électorat en classes sociales, en classes d'âge, en classes professionnelles, en origines géographiques etc. C'est là un travail lourd et minutieux, dont on peut voir une esquisse dans l'article d'Emmanuel Todd paru dans Le Monde entre les deux tours de la présidentielle.

Mais pour aller plus loin, on voit bien que la question posée lors d'une telle élection est suffisamment complexe pour qu'on ne puisse pas lire dans son résultat un message clair. Il en est en fait de même pour toute information mêlant à la fois des données quantifiées et des données qualitatives. C'est une tentation simplificatrice que de vouloir tout décrire par un nombre mais cela conduit à de nombreuses erreurs et on peut parfois soupçonner ceux qui véhiculent de telles données de savoir qu'ils cachent ainsi une partie du problème, parfois la plus épineuse. On touche ici à la légitimité du pourcentage : avant de lui donner un quelconque crédit il faut savoir comment il a été obtenu, quelles simplifications ont été faites, comment ont été traitées les données non chiffrées et si, de ce fait, le résultat a tout simplement plus de sens que d'ajouter torchons et serviettes … C'est pourtant là monnaie courante dans le domaine économique où l'on chiffre tout. Ainsi pour comparer divers projets d'urbanisation, on chiffrera le mécontentement des habitants, lui donnant par exemple le coût d'un mur anti-bruit. Mais a-t-on réellement évalué les problèmes posés en faisant cela ? Ce débat dépasse le cadre de cet article. Si je l'évoque c'est pour que chacun prenne conscience de la part qu'il a prendre dans l'utilisation des données chiffrées dans la société : tant ceux qui les produisent, que ceux qui les véhiculent ou les utilisent ou encore que ceux qui les « subissent ». Chacun doit mener sa part de réflexion et, même si toute vérité est bonne à dire, il ne faut pas oublier qu'un mensonge par omission reste un mensonge et qu'une réalité partiellement décrite, par manque d'information ou de méthode, par un pourcentage est bien plus proche d'un mensonge que d'une vérité.

Le pourcentage comme support de décision

Une autre tentation est de se servir de deux données chiffrées pour comparer deux phénomènes distincts et décider d'un moindre mal. Le premier danger dans cette direction est d'être amené à chiffrer des phénomènes non quantifiables puis de pratiquer un amalgame sans tenir compte des spécificités de chacun. Ainsi on propose, sûrement pour plus très longtemps, aux femmes enceintes ayant entre 35 et 38 ans de pratiquer une prise de sang afin de déterminer leur taux de risque de porter un bébé trisomique. Ce risque est évalué par diverses méthodes ayant toutes des avantages et des inconvénients. Une fois l'évaluation accomplie, le taux est comparé au taux de risque de fausse couche lors d'une amniocentèse (1/250). Si le risque détecté lors de la prise de sang est grand, on pratique une amniocentèse pour déterminer de façon sûre si le bébé est trisomique ou non. On fait face ici à tous les problèmes soulevés précédemment : le mode de calcul du pourcentage est contestable, le risque lors de l'amniocentèse décrit une réalité hétérogène et il est en fait bien plus grand dans certains cas bien identifiés par les spécialistes alors qu'il est presque nul pour le reste des femmes, enfin la comparaison entre ces deux taux n'a pas de sens.

En conclusion ce test est une sorte de bouclier qui permet de se mettre à l'abri derrière un processus décisionnel automatique, ce qui est la négation même de la diversité humaine. Par ailleurs c'est chercher une légitimité grâce à un outil mathématique

alors que celui-ci ne garantit en rien la conclusion à laquelle il sert. De façon explicite, en statistique on peut mettre en place des tests afin de valider des hypothèses, par exemple « le risque de trisomie est plus grand que le risque de fausse couche ». De façon intrinsèque un tel test fournira une réponse du genre « cette hypothèse est vraie avec une certitude de 95% ». On voit bien les limites de cette assertion mais on n'imagine peut-être pas que la réponse de ce test n'est pas contradictoire avec la réponse : « le risque de fausse couche est plus grand que le risque de trisomie avec une certitude de 95% ».

La raison en est que, dans un test comme le précédent, on teste en fait une hypothèse par opposition à une hypothèse de référence et même si l'hypothèse testée est en fait la négation de l'hypothèse de référence, il n'y a pas de symétrie dans le rôle joué par les deux hypothèses. On touche ici du doigt la notion de barre d'erreur qui est liée à la caractéristique de dispersion du phénomène étudié, caractéristique très souvent occultée dans les descriptifs.

Prenons l'exemple des sondages d'opinion en politique : aucun statisticien ne peut prétendre, avec la taille des échantillons proposés, avoir une précision de l'ordre de 0.1 % dans les pourcentages qu'il estime. D'une part cette précision est de l'ordre d'au moins 1 %, d'autre part il faut prendre l'habitude de donner des fourchettes pour les estimations et non des nombres bruts. Le point important dans cette description du problème est le suivant : **on ne peut pas prédire un classement** avec un sondage. On peut juste décrire une tendance comme par exemple : les trois candidats en tête seront Chirac, Jospin et Le Pen. Il est illusoire d'espérer mieux. Ce que je dis ici est valable pour toutes les estimations statistiques, comme par exemple les classements faits par tant de journaux sur les lycées français : un tel classement n'a aucun sens, tout au plus peut-on distinguer 4 ou 5 grandes catégories de lycées. Et c'est tout.
Quand bien même on pourrait dire Chirac devrait obtenir entre 17% et 23% des voix, Jospin entre 15% et 21%, Le Pen entre 11% et 17%, cela n'entraîne absolument pas que l'ordre d'arrivée lors de l'élection sera l'ordre des moyennes ou des fourchettes hautes. Avec de telles prédiction il y a fort à parier qu'aucun classement entre les trois candidats ne serait rejeté par un test d'hypothèse au sens décrit précédemment. Autrement dit tout institut de sondage pourrait annoncer un classement, aléatoirement, et être de bonne foi, mathématiquement. La preuve en a été donnée de façon claire lors du premier tour de la présidentielle !
Je le dis et le répète : aucun statisticien ne peut en conscience prétendre pouvoir prédire un classement, c'est un problème structurel lié à la méthode d'évaluation.

L'utilisation des sondages
Il faut donc apprendre à parler avec réserve de ces sondages, non pas parce que les sondeurs se trompent mais parce que le phénomène qu'ils décrivent est statistique, c'est-à-dire, par nature, aléatoire. Cette mise en garde concerne tous ceux qui usent et abusent des sondages, qu'ils soient politiques, journalistes ou électeurs.

Par ailleurs, il est un phénomène bien connu des physiciens : **toute mesure perturbe le phénomène observé**. Ce que je veux dire, c'est qu'un sondage n'est pas neutre, par le fait même qu'il a été fait et communiqué à tous, il change la donne, il fait évoluer les opinions. Et ce d'une façon qui n'est pas aisément modélisable, donc non prédictible par des moyens statistiques. Ce phénomène est bien connu et on peut penser que ceux qui mentent en répondant à un sondage réagissent instinctivement contre ce phénomène. Ce type de mensonge peut être détecté par les sondeurs (c'est le principe des coefficients multiplicateurs pour ajuster l'échantillon sondé), mais il est d'autres phénomènes, en aval, qu'il est impossible de prédire. Enfin, dernier artefact produit par cette utilisation : l'attitude même des politiques. Certains sont tellement prisonniers de cette évaluation qu'ils ne se positionnent presque que par rapport à elle, d'où des commentaires laissant penser qu'une gifle aurait pour effet de faire monter un candidat dans les sondages ! Et pourtant c'est un exercice de niveau terminale de montrer que cette augmentation n'est probablement due qu'à une fluctuation d'échantillon, autrement dit qu'elle n'est très probablement due qu'à la différence entre les deux échantillons d'électeurs qui ont répondu aux deux sondages incriminés.

De l'importance du vrai
En tant que mathématicien je pense que l'information est une bonne chose. Mais cette phrase n'a que peu de sens : une information partielle ou mal exploitée est pire que l'absence d'information, surtout si elle est présentée comme information complète et véritable. Je suis, en conséquence, pour l'interdiction des sondages au moins 1 mois avant les élections, de façon à garantir l'existence d'un **vrai** débat d'idées suivi d'une **vraie** élection. A quoi servent donc toutes ces fausses élections que sont les sondages ? A faire croire que la vraie n'est rien d'autre qu'une fausse de plus et qu'on n'a même plus besoin d'y participer ?

J'aimerais, pour clore, rebondir sur ce dernier point. Si les sondages ont un tel impact c'est qu'ils sont au centre du débat. C'est la chose sur laquelle on glose quand on a rien à dire. N'est-on pas en droit d'attendre des hommes politiques, des politologues, des médias etc. autre chose ? Des analyses, des perspectives historiques, des débats d'idées etc. J'ai entendu beaucoup accuser la problématique sécuritaire comme principal responsable du vote FN. Et par ailleurs nombreux se sont étonnés que des villages exempts de toute violence ou de toute immigration se soient laissés emporter par la vague frontiste. Voire même une partie de la population immigrée. Je pense, pour ma part, que le vote FN a été peu influencé par cela : une fois encore il n'a pas progressé de façon significative. Par contre la violence a progressé partout et surtout dans les villages tranquilles : la violence est sur le poste de télévision et personne, surtout pas ceux qui ne la côtoient pas encore dans leur village, ne veut qu'elle en sorte. Se pose-t-on la question, avant de montrer autant d'horreurs à la télévision, de savoir qui les regarde et surtout comment il prend la chose ? C'est une chose d'informer, c'en est une autre de montrer. Est-ce informer que se contenter de montrer ? N'y a-t-il rien d'autre à dire, à analyser, à décrire ? Se contenter d'une description de surface n'est pas analyser. Il y a dans cette boulimie de faits, un réel problème des temps modernes : le traitement de l'information et son exploitation. Que dit-on **vraiment** quand on donne une information à la télévision, à la radio ou dans un journal ? Et, surtout,

comment cela est-il **interprété**, car c'est bien là la finalité de ceux qui informent, à savoir, que ceux qu'ils informent aient compris ce qu'ils voulaient dire.

De tels enjeux ne peuvent être affrontés sereinement que grâce à une formation adéquate et une culture qui les met en perspective. Je dis cela tant pour ceux qui donnent l'information que pour ceux qui la reçoivent. C'est pourquoi je crois que l'enjeu majeur pour notre société, à tous les niveaux, est dans l'éducation et qu'elle est, ainsi qu'elle le fut le siècle passé, la clef de l'intégration et de l'ascension sociale. Pour cela il faut que l'école soit adaptée à la société et surtout qu'elle puisse garantir l'insertion dans la vie active, et donc qu'elle en ait les moyens structurels.