Statistique et citoyenneté : les maths s'ouvrent sur le monde

Groupe « statistique et citoyenneté » de l'IREM Paris-Nord(*)

> « Si la notion de *vérité statistique* devenait familière à tous ceux qui parlent ou écrivent au sujet de questions où la vérité statistique est la seule vérité, bien des sophismes et bien des paradoxes seraient évités. »
>
> Émile Borel – *Le Hasard* – Alcan, 3e édition, 1914.

À un moment où se posent de façon aiguë les questions du rôle de l'enseignement des mathématiques, et pas seulement de la part de jeunes demandant « à quoi ça sert ? », de la définition d'une nécessaire « culture mathématique » et de la motivation des élèves, il nous semble pertinent de présenter quelques activités statistiques qui peuvent contribuer à une réponse. Les exemples présentés ici ont été expérimentés dans le cadre du lycée, mais ils pourront s'interpréter au collège dans l'esprit des « thèmes de convergence » et participeront, dans les futurs programmes de BEP et de Bac Pro, à une nécessaire initiation à l'aléatoire qui, en France, arrive encore trop tardivement dans le cursus scolaire.

Comme on le constatera, on fait de « vraies maths » dans ces activités et on y acquiert des connaissances en informatique. Au delà du traitement statistique, des compétences de calcul, d'interprétation graphique, de raisonnement scientifique, inscrites dans les programmes officiels, sont sollicitées.

Le choix d'exemples réalistes, en prise avec le monde et les questions de société, motive indéniablement les élèves, en particulier les « non matheux », permettant de mettre en évidence les implications de notre enseignement. Ce n'est pas que les exemples ludiques, comme les jets de pièces et de dés, soient sans intérêt ; bien au contraire, ils constituent un outil pédagogique essentiel à l'étude de l'aléatoire, mais ils ne peuvent, à eux seuls, en justifier les enjeux. Les études statistiques développées ici suscitent des interrogations, en décelant des « tendances » qui, sans trancher sur la causalité des phénomènes, suggèrent des lieux où aller regarder « ce qui a pu se passer ». Elles débouchent donc sur un dialogue qui, dans la classe, entre élèves et professeurs de mathématiques ou d'autres disciplines, et avec l'appui de sources documentaires telles qu'Internet, peut être analogue à celui que doivent mener de vrais utilisateurs de la statistique.

Enfin les activités présentées s'appuient sur une pratique des mathématiques « par l'expérience », facilitée par les moyens informatiques. Les mathématiques ont un aspect pratique et expérimental, en particulier pour leur enseignement, pour conjecturer bien sûr, mais aussi pour comprendre et parfois pour emporter la

(*) www-irem.univ-paris13.fr/

conviction. Voir un théorème à l'œuvre peut être plus convainquant et plus éclairant que lire sa démonstration, au moins dans un premier temps.

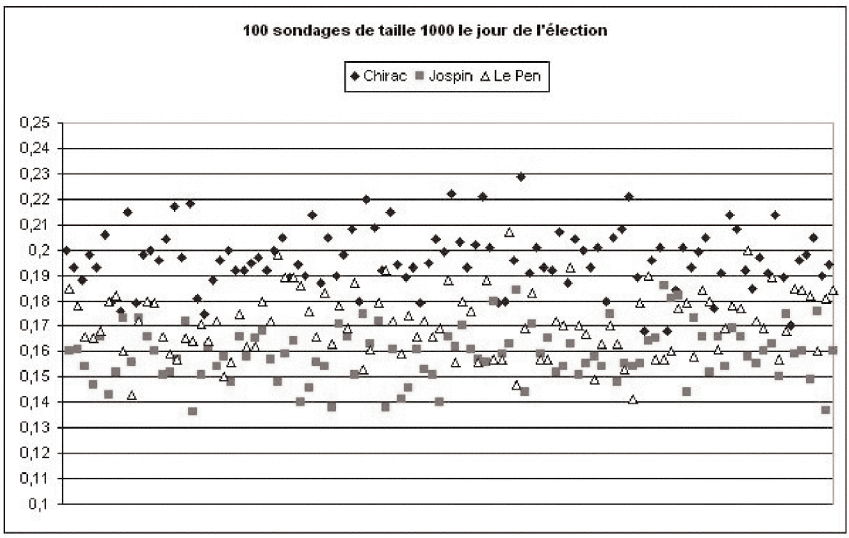
Fluctuations et sondages politiques en France

À propos de statistique et de citoyenneté, une des premières idées venant à l'esprit, en particulier en période d'élections, est de faire réfléchir les élèves sur les sondages électoraux pour montrer la qualité toute relative de l'information fournie. L'attitude de l'opinion vis-à-vis des sondages est souvent sans nuance : on leur prête des pouvoirs de prédiction qu'ils n'ont pas (en omettant souvent de fournir les « fourchettes ») et (ou) on déclare qu'ils se trompent 9 fois sur 10. Nous avons plusieurs fois proposé à des élèves de seconde de travailler sur les résultats du premier tour de l'élection présidentielle de 2002, qui avait particulièrement marqué les esprits, suscitant à chaque fois intérêt et débats dans la classe.

On peut mettre en parallèle le dernier sondage publié par BVA et effectué le vendredi 19/04/02 sur 1 000 électeurs (Jacques Chirac 19 %, Lionel Jospin 18 %, Jean-Marie Le Pen 14 %) avec le résultat du premier tour deux jours plus tard (Jacques Chirac 19,88 %, Lionel Jospin 16,18 %, Jean-Marie Le Pen 16,86 %) et poser la question « le sondage est-il faux ? ». Si le sondage avait la prétention de prévoir exactement les résultats, ou ne serait-ce que l'ordre des candidats, il serait bien sûr faux. Mais cette prétention n'est pas scientifique, comme nous l'apprend l'observation des fluctuations d'échantillonnage. Dans un sondage d'intention de vote, tout se passe comme si on tirait au sort les 1 000 électeurs. En fait, essentiellement parce qu'un véritable tirage au hasard revient trop cher (il est très coûteux de devoir interroger précisément la personne qui a été tirée au sort et pas une autre), on a recours à d'autres méthodes, comme celle des quotas, mais avec une qualité équivalente au tirage totalement aléatoire. On peut donc dire que dans la situation précédente, simuler un sondage « bien fait », c'est-à-dire en supposant que les intentions de vote deux jours avant correspondent au vote réel (ce qui est loin d'être le cas : indécis, intentions de votes non avouées, …), consiste à faire tourner 1 000 fois une roue de loterie partagée en quatre secteurs de 20 % (Jacques Chirac), 16 % (Lionel Jospin), 17 % (Jean-Marie Le Pen) et 47 % (autres candidats) correspondant, en arrondissant, à l'état de l'opinion le jour de l'élection. La simulation d'une telle roue de loterie est facile à mettre en place avec un tableur. Le générateur de nombres aléatoires fournit un nombre « au hasard » dans l'intervalle [0, 1[[(1)]. Il suffit alors de partager cet intervalle proportionnellement aux pourcentages des secteurs de la roue de loterie, ce qui peut être demandé aux élèves.

Ainsi, l'instruction =ALEA() entrée en cellule A1 et l'instruction =SI(A1<=0,2;"Chirac";SI(A1<=0,36;"Jospin";SI(A1<=0,53;"LePen";"Autre"))) entrée en cellule B1 simule le sondage d'un électeur. Il suffit de recopier ces instructions 1 000 fois vers le bas pour simuler un sondage de 1 000 personnes puis 100 fois vers la droite pour simuler 100 sondages.

(*) Voir Parzysz (Bernard) - *Quelques questions à propos des tables et des générateurs aléatoires*, in B. Chaput & M. Henry (éd.) *Statistique au lycée* vol. 1, 181-199. Éd. APMEP-ADIREM 2005.



D'une certaine façon tous ces sondages sont corrects et l'observation du nuage de points correspondant aux fréquences des trois candidats sur 100 sondages suffit à prendre conscience des fluctuations dues au hasard.
On peut faire examiner la proportion des sondages plaçant, à tort, Lionel Jospin devant Jean-Marie Le Pen, ou faire étudier l'amplitude des fluctuations (grosso modo, dans plus de 95 % des cas, de plus ou moins $\frac{1}{\sqrt{1\,000}}$ autour de la fréquence dans l'opinion, dans le cas d'un sondage purement aléatoire) pour aller vers la notion de « fourchette » de sondage qui figure comme thème d'étude dans les programmes de seconde.
Il est ainsi formateur de montrer qu'une étude statistique peut permettre de déceler la possibilité de ce qui a priori apparaît à beaucoup comme complètement imprévu, par exemple ici la présence de Jean-Marie Le Pen au second tour. L'étude des fluctuations d'échantillonnage montre que la seule variabilité « naturelle » incite à la prudence dans la situation du premier tour de 2002, alors que d'autres éléments doivent être pris en compte comme les nombreux biais existant dans le contexte des sondages d'intention de vote tels qu'ils sont pratiqués, comme le choix des personnes interrogées par la méthode des quotas, la dissimulation des intentions réelles ou la grande versatilité d'une partie de l'opinion.

Étude du sex-ratio

Le sex-ratio est le rapport du nombre de garçons au nombre de filles à la naissance. Ce rapport n'est pas exactement de 1 et ce fut une des premières découvertes statistiques du XVII^{e} siècle de constater qu'il est habituellement, et avec une remarquable constance, de 105 garçons pour 100 filles, du moins pour un échantillon

suffisamment important et sauf circonstances exceptionnelles (guerres, famines, …), sans d'ailleurs que l'on ait d'explication biologique à cette constatation. L'observation de l'actualité nous a amené à traiter deux exemples, l'un en Chine témoignant d'une triste sélection des naissances, l'autre au Canada, à propos de problèmes sans doute liés à l'environnement.

Trop de garçons en Chine

Dans un article du *Washington Post*, on apprend que dans le village de Xicun, dans les montagnes du sud de la province de Guangxi en Chine, il est né, en 2000, vingt enfants, parmi lesquels seize garçons.
Peut-on considérer cette proportion comme « naturelle » ? Question que l'on peut reformuler ainsi : est-il rare d'observer une fréquence $f = \frac{16}{20} = 0,8$ sur un échantillon de taille 20 prélevé dans une urne bicolore où la proportion correspondant à f est $p = 0,512$ (en l'occurrence la proportion correspondant à 105 garçons pour 100 filles) ?
Compte tenu de la petite taille de l'échantillon, on aura recours à la simulation (on ne peut utiliser en seconde les formules de fluctuation s'appuyant sur une approximation par une loi normale, en terminale S on pourrait en revanche avoir recours à la loi binomiale). On peut simuler sur le tableur un grand nombre de tels échantillons et observer que le phénomène $f = 80$ % (ou $f \geq 80$ %) lorsque $p = 51,2$ % est effectivement très rare, malgré la petite taille de l'échantillon.
L'observation effectuée dans ce village chinois est donc « statistiquement anormale ». L'explication est à chercher, par exemple, avec le professeur d'histoire-géographie ou de sciences économiques et sociales.

Trop de filles au Canada

L'énoncé suivant a été donné en devoir surveillé dans une classe de seconde en fin de chapitre de statistique.
Les données proviennent d'une étude effectuée au Canada et montrant une différence (très) significative du sex-ratio à la naissance (déficit de garçons) sur une population exposée à une pollution chimique. Dans ce cas particulier, l'inquiétude provient du fait que, bien que ces industries canadiennes respectent les normes, une exposition prolongée à de faibles doses de polluants pourrait avoir un impact sanitaire mesurable.

Énoncé du devoir :

Le « sex-ratio » est le rapport du nombre de garçons à celui des filles à la naissance. Il est habituellement de 105 garçons pour 100 filles.

1) La fréquence habituelle des garçons est $p = \frac{105}{105+100} \approx 0,512$. Si cette proportion est exactement respectée, combien de garçons et de filles a-t-on sur 1 000 naissances ?
2) Si l'on prélève des échantillons aléatoires de taille n « assez grande » (disons au

moins 30) dans une population où la fréquence étudiée est $p = 0{,}512$, dans plus de 95 % des échantillons, la fréquence f observée sera comprise dans l'intervalle $\left[0{,}512-\frac{1}{\sqrt{n}}, 0{,}512+\frac{1}{\sqrt{n}}\right]$.

Si une fréquence f n'appartient pas à cet intervalle, on dira que f présente une « différence significative » au niveau 0,95 avec $p = 0{,}512$.

a) Dans la réserve indienne d'Aamjiwnaag, située au Canada, il est né entre 1999 et 2003, $n = 132$ enfants dont 46 garçons.

Que vaut la fréquence f des garçons pour cette période à Aamjiwnaag (arrondir à 10^{-3}) ?

b) La fréquence des garçons observée à Aamjiwnaag pour la période 1999-2003 présente-t-elle une « différence significative » au niveau 0,95 avec $p = 0{,}512$ (justifier par un calcul) ?

Sources : Science et Vie février 2006 – Environnemental Health Perspectives octobre 2005 (article en ligne).

Éléments de réponse :

1. Si la proportion $p = 0{,}512$ est exactement respectée, on doit avoir, sur 1 000 naissances, 512 garçons et 488 filles.

2. a) La fréquence de garçons f observée à Aamjiwnaag est $f = 46/132 = 0{,}348$.

b) La différence est significative au niveau 0,95 lorsque la fréquence observée n'appartient pas à l'intervalle [0,424 ; 0,599].

La fréquence 0,348 n'appartient pas à cet intervalle donc la valeur observée présente une différence significative au niveau 0,95 avec $p = 0{,}512$.

Donné en devoir surveillé, cet exercice n'y déchaîne naturellement pas des passions. Toutefois les élèves en rendant leurs copies ont posé la question inhabituelle de savoir si c'était une « histoire vraie » et un intérêt soutenu a accompagné la correction.

Tel que l'énoncé est présenté ici, il soulève immanquablement la question « Que peut-on tirer comme conclusion de ces statistiques ? ».

Une première version de l'énoncé ajoutait comme précision que la réserve d'Aamjiwnaag est située « à proximité de nombreuses industries chimiques ». Cela laissait entendre qu'il y avait une relation de causalité entre la pollution chimique et le « défaut » observé du sex-ratio, causalité qui n'est pas établie par l'étude statistique. À la question « Que peut-on tirer comme conclusion ? », on peut seulement ici répondre que « cette étude pose question ». La statistique donne l'alerte, ce qui est déjà beaucoup. Le fait que la réserve soit située au cœur d'industries chimiques devient un élément troublant sur lequel on doit enquêter.

De façon générale, c'est la notion de « preuve statistique » qui est ici en jeu. Il ne s'agit pas d'une « preuve » au sens habituel mais d'un élément probant, d'une présomption. Plutôt que de parler de « preuve statistique », les anglo-saxons disent plus justement *piece of evidence*.

D'autres explications possibles du déséquilibre du sex-ratio pourraient être liées au mode de vie de ces indiens ou à leur patrimoine génétique. Une étude statistique comparative a été menée sur des indiens de la même tribu vivant dans un autre environnement et a (dé)montré que ce n'était (sans doute) pas le cas. En revanche l'influence de certains produits chimiques sur le sex-ratio a été établie « statistiquement » par d'autres études.
Une recherche sur Internet (entrer « Aamjiwnaag » dans un moteur de recherche) permettra d'avoir d'autres éléments sur ce dossier qui a fait polémique au Canada.

Discrimination raciale au Texas

Le travail suivant s'est déroulé en terminale S (mais une activité sur le même thème a également été expérimentée en classe de seconde). Il s'agit de l'étude d'un véritable document juridique où les arguments avancés sont d'ordre mathématique.
Un accusé d'origine mexicaine, condamné pour vol et tentative de viol dans un comté du sud du Texas attaqua le jugement sous le motif que la désignation des jurés dans l'État du Texas était discriminatoire pour les Américains d'origine mexicaine. Son argument était que ceux-ci n'étaient pas suffisamment représentés dans les jurys populaires. On demande aux élèves de terminale de lire le document juridique puis de répondre à quelques questions.

Attendu de la Cour Suprême des États-Unis (affaire Castaneda contre Partida) :
« Si les jurés étaient tirés au hasard dans l'ensemble de la population, le nombre d'américains mexicains dans l'échantillon pourrait alors être modélisé par une **distribution binomiale**… Étant donné que **79,1** % de la population est mexico-américaine, le nombre attendu d'américains mexicains parmi les **870** personnes convoquées en tant que grands jurés pendant la période de 11 ans est approximativement **688**. Le nombre observé est **339**. Bien sûr, dans n'importe quel tirage considéré, une certaine fluctuation par rapport au nombre attendu est prévisible. Le point essentiel, cependant, est que le modèle statistique montre que les résultats d'un tirage au sort tombent vraisemblablement dans le voisinage de la valeur attendue… La mesure des fluctuations prévues par rapport à la valeur attendue est l'**écart type**, défini pour la distribution binomiale comme la racine carrée de la taille de l'échantillon (ici 870) fois la probabilité de sélectionner un américain mexicain (ici 0,791) fois la probabilité de sélectionner un non américain mexicain (ici 0,209)… Ainsi, dans ce cas, l'écart type est approximativement de **12**. En règle générale pour de si grands échantillons, si la différence entre la valeur attendue et le nombre observé est plus grand que deux ou trois écarts types, alors l'hypothèse que le tirage du jury était au hasard serait suspecte à un spécialiste des sciences humaines. Les données sur 11 années reflètent ici une différence d'environ **29** écarts types. Un calcul détaillé révèle qu'un éloignement aussi important de la valeur attendue se produirait avec moins d'**une chance sur 10^{140}**. »

Questions :
1. Définir la variable aléatoire qui, dans cette situation, suit une loi binomiale.
2. Quels sont les paramètres de la loi binomiale ?

3. À quel calcul correspond la valeur 688 ?
4. Effectuer le calcul de l'écart type.
5. À quoi correspond la « différence de 29 écarts types » ?
6. À quel événement correspond la probabilité 10^{-140} ?
7. La constitution des jurys est-elle totalement aléatoire ?

Éléments de réponse :
1. Il s'agit de la variable aléatoire X qui à 870 personnes tirées au sort dans la population (celle-ci étant très importante, on peut assimiler ces tirages à des tirages avec remise) associe le nombre de personnes d'origine mexicaine.
2. Les paramètres de la loi binomiale sont $n = 870$ et $p = 0{,}791$.
3. La valeur 688 correspond au calcul de l'espérance : $np \approx 688{,}17$.
4. On a $\sigma = \sqrt{870 \times 0{,}791 \times 0{,}209} \approx 11{,}99 \approx 12$.
5. La quantité 29 écarts types correspond à la différence entre la valeur observée de la variable aléatoire, c'est-à-dire 339, et son espérance, c'est-à-dire 688.
6. La probabilité 10^{-140} correspond à celle d'observer une valeur distante de l'espérance d'au moins 29 écarts types , c'est-à-dire : $P(X \leq 351) + P(X \geq 1\,025) = P(X \leq 351)$ puisque X prend ses valeurs entre 0 et 870.
7. La probabilité précédente est beaucoup trop faible pour considérer que la variable aléatoire binomiale introduite ici modélise correctement la situation. On doit donc douter du fait que les jurys sont constitués par un tirage au sort totalement aléatoire.

Il est à noter que lors de la discussion qui a suivi la correction, il est apparu que l'observation d'un éloignement, par rapport à la proportion de mexico-américains dans la population de l'état, qui n'a qu'une probabilité de 10^{-140} d'être atteint ou dépassé, ne semble pas immédiatement « humainement impossible » à un certain nombre d'élèves.

Les avis des élèves sur l'activité sont partagés. Plusieurs élèves ont trouvé l'activité « plus intéressante que ce que l'on fait d'habitude », parce qu'il s'agit de la « vraie vie ». Une élève a même spontanément affirmé que « s'il y avait eu plus souvent ce type d'exercice en mathématiques, elle se serait peut-être davantage intéressée à la matière ». D'autres élèves, cependant, ont trouvé cette activité « plus difficile » que les exercices classiques et à la question « Que penseriez-vous de ce type d'exercice en maths au bac S (analyse d'un texte scientifique comme cela se fait en physique) ? » ils ont répondu qu'ils ne voulaient pas de ce type d'exercice à l'examen.

Les données étudiées constituent une « preuve statistique » du fait que la constitution des jurys n'est pas totalement aléatoire, c'est-à-dire que ceux-ci ne sont pas représentatifs du point de vue du caractère hispanique de la population. Les calculs précédents montrent qu'on ne peut pas considérer que les jurys résultent d'un tirage au sort où chaque élément de la population a les mêmes chances d'être choisi. Mais c'est tout ce que le statisticien peut dire et, en particulier, il n'a pas à se prononcer sur les causes et à porter des accusations de discrimination raciale, c'est le rôle du juge. L'étude statistique précédente doit inciter à enquêter sur les conditions de constitution des jurys. On constatera alors que pour être juré on doit maîtriser la

langue anglaise (écrite et parlée), ce qui n'est pas le cas de la majorité de la population d'origine hispanique, que pendant les 11 années correspondant à l'étude, la proportion des hispaniques dans la population a évolué, et que la proportion d'hispaniques dans les jurys a également évolué au cours de ces 11 années.
Une autre activité possible, en particulier en seconde sur ce thème (en liaison avec l'ECJS), pourrait consister en un exercice d'argumentation, les élèves dans le rôle de l'avocat devant s'appuyer sur les données mathématiques pour présenter le dossier.

Pour aller plus loin

Le groupe « Statistique et citoyenneté » de l'IREM Paris-Nord s'est donné les objectifs suivants. À partir de situations concrètes, ayant une forte résonance sociologique :
– montrer l'utilité d'une formation mathématique pour « décrypter » le monde moderne ;
– développer l'intérêt pour les mathématiques par des activités ayant une signification forte et favorisant l'interdisciplinarité ;
– privilégier l'autonomie des élèves à mettre en place une démarche « scientifique » dans l'analyse d'une situation : formulation d'hypothèse, « construction » d'un « modèle », expérimentation (simulations, …), conclusions ;
– initier à l'aléatoire et aux notions de risques et de confiance.

Si ces thèmes vous intéressent, vous pouvez consulter nos publications et nos pages web. Différentes activités sont présentées, accompagnées de témoignages dans les classes, et ne demandent qu'à s'enrichir de vos commentaires et expériences.

• www-irem.univ-paris13.fr/
À partir du site de l'IREM de Paris-Nord, on peut accéder aux pages du groupe « statistique et citoyenneté ».

• www.statistix.fr/
Le site « Statistix », coordonné par Claudine Schwartz, présente de nombreuses ressources pour les enseignants, avec souvent une dimension inter-disciplinaire.

• CHAPUT (Brigitte) & HENRY (Michel) éditeurs - Commission Inter-Irem Statistique et probabilités - APMEP - Brochure nº 156 - *Statistique au lycée* - volumes 1 et 2 (ce dernier volume étant à paraître).
• DOWEK (Gilles) - *Peut-on croire les sondages ?* - Le Pommier 2002.
• DUTARTE (Philippe), *L'induction statistique au lycée illustrée par le tableur*, Didier 2005.
• PIEDNOIR (Jean-Louis), DUTARTE (Philippe), *Enseigner la statistique au lycée : des enjeux aux méthodes*, IREM Paris-Nord 2003. Diffusé par l'APMEP sous le nº 820.
• SCHWARTZ (Claudine) - *Pratiques de la statistique* - Vuibert 2006. Diffusé par l'APMEP sous le nº 940.