À vos stats citoyens

Induction statistique et citoyenneté au lycée

Philippe Dutarte(*)

La statistique inductive (ou inférentielle) est née au XX^e siècle. Pour illustrer les enjeux actuels de son enseignement au lycée, on peut mettre en parallèle, ou dos à dos, les deux mathématiciens suivants. En 1940, André Weil, membre du groupe Bourbaki, affirme que « la statistique moderne paraît avoir enfin résolu le problème légendaire qui consistait, connaissant la longueur du navire et la durée de la traversée […], à calculer l'âge du capitaine ». À la même époque, aux États-Unis, les procédures statistiques séquentielles d'Abraham Wald étaient classées « secret défense ».

On développe dans cet atelier des exemples que l'on peut traiter en classe et qui montrent l'intérêt des méthodes de la statistique inductive dans de nombreuses questions de société. On est loin de « l'âge du capitaine ».

Fluctuations

Inquiétudes à Woburn (Massachussets)

Woburn est une petite ville industrielle du Massachusetts, au Nord-Est des États-Unis. Du milieu à la fin des années 1970, la communauté locale s'émeut d'un grand nombre de leucémies infantiles survenant dans certains quartiers de la ville. Le tableau suivant résume les données statistiques concernant les enfants de Woburn de moins de 15 ans, pour la période 1969-1979 (Sources : *Massachusetts Department of Public Health et Harvard University*).

Enfants entre 0 et 14 ans	Population de Woburn selon le recensement de 1970 n	Nombre de cas de leucémie infantile observés à Woburn entre 1969 et 1979	Fréquence des leucémies à Woburn f	Fréquence des leucémies aux États-Unis p
Garçons	5 969	9	0,001 51	0,000 52
Filles	5 779	3	0,000 52	0,000 38
Total	11 748	12	0,001 02	0,000 45

La question statistique qui se pose est de savoir si le hasard seul peut raisonnablement expliquer les fréquences observées à Woburn, considérées comme résultant d'un échantillon prélevé dans la population américaine. La population des États-Unis étant très grande par rapport à celle de Woburn, on peut considérer que l'échantillon résulte d'un tirage avec remise et simuler des tirages de taille n avec le tableur.

(*) Commission inter-IREM lycées technologiques. dutarte@club-internet.fr

Dans le cas des garçons, simulons sur le tableur 100 échantillons de taille $n = 5\,969$ prélevés dans une population où $p = 0{,}000\,52$. La simulation montre que plus de 95% des fluctuations aléatoires des valeurs de f s'effectuent dans l'intervalle [0 ; 0,001]. On ne peut donc pas raisonnablement attribuer au seul hasard le niveau très « significativement » élevé des leucémies infantiles observées chez les garçons à Woburn.
Pour ce qui est des filles, nous simulons de manière analogue sur le tableur 100 échantillons de taille $n = 5\,779$ prélevés dans une population où $p = 0{,}000\,38$. La simulation montre que 25% des échantillons simulés avec $p = 0{,}000\,38$ font apparaître une fréquence f supérieure ou égale à celle observée avec les données de Woburn. On peut donc penser que le taux de leucémies infantiles observé chez les filles à Woburn n'est pas « significativement » élevé. Le hasard pourrait l'expliquer. La taille de l'échantillon est en tout cas trop faible pour mettre en évidence ici un phénomène « anormal ».
Le taux anormalement élevé de leucémies infantiles chez les garçons à Woburn est officiellement confirmé par le Département de Santé Publique du Massachusetts en avril 1980. Les soupçons se portent alors sur la qualité de l'eau de la nappe phréatique qui, par des forages, alimente la ville. On découvre alors le syndrome du trichloréthylène.

Pesticides à Ufa (Russie)

L'exemple suivant est fondé sur une étude publiée par des chercheurs de l'Université de Montréal en 2002(1) à propos de l'influence des pesticides sur le rapport garçons/filles à la naissance. Cette étude a été menée dans la ville d'Ufa (fédération de Russie) auprès de 198 personnes (150 hommes et 48 femmes) ayant été exposés, dans une usine agrochimique active de 1961 à 1988, à des pesticides contenant de la dioxine.
Le rapport garçons/filles à la naissance pour la ville d'Ufa est $p = 0{,}512$ (soit environ l'habituel 105 garçons pour 100 filles). Sur la descendance des personnes exposées que l'on a étudiées, on observe $n = 227$ enfants dont 91 garçons et 136 filles, soit une fréquence observée de $f = 0{,}4$ garçons.
La question qui se pose est de savoir si l'écart observé par rapport à la valeur « normale » est significatif, compte-tenu de la taille n de l'échantillon.
On simule sur le tableur 1 000 échantillons de taille 227 prélevés avec remise dans une urne contenant des boules marquées 1 ou 0, les premières (figurant les garçons) étant en proportion $p = 0{,}512$ dans l'urne.
On peut constater qu'environ 95% des échantillons, et généralement davantage, fournissent une fréquence f comprise dans l'intervalle $\left[0{,}512-\frac{1}{\sqrt{227}}\,;0{,}512+\frac{1}{\sqrt{227}}\right]$, c'est-à-dire environ [0,45 ; 0,58]. On compte généralement moins de 50 points en dehors des limites fixées à 0,45 et 0,58. En

(1) *Sex Ratios of Children of Russian Pesticide Producers Exposed to Dioxin* – Revue *Environmental Health* Novembre 2002. Article disponible sur Internet

faisant F9, on visualise 1 000 nouveaux échantillons. On constatera qu'il est extrêmement rare d'observer une fréquence inférieure ou égale à 0,4 mais que cela se produit de temps à autre (moins d'une fois sur 1 000 en moyenne).
Il ressort de ces observations que le hasard seul peut très difficilement expliquer le faible taux de 40% de garçons parmi les enfants des personnes exposées aux pesticides. Ajoutons qu'à Seveso, en Italie, une diminution du nombre de naissances de garçons par rapport a celui des filles a également été observée chez les individus qui ont été de façon accidentelle fortement exposés aux dioxines en juillet 1976. Ceci ajoute du poids à l'alerte statistique que donnent les chercheurs de l'université de Montréal à propos de l'étude des enfants d'Ufa.

Classement des hôpitaux américains

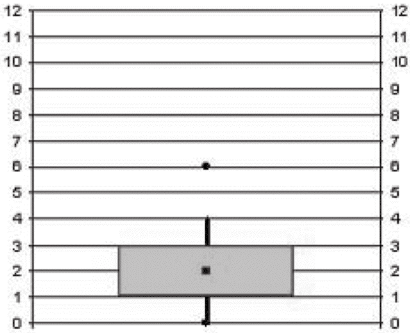
Les « clients » des hôpitaux américains peuvent accéder, sur le site www.healthgrades.com, aux notes attribuées à chaque hôpital des U.S.A. selon l'acte chirurgical envisagé.
Les notes vont de ★★★★★ quand l'établissement figure parmi les 10% les mieux notés, avec une « différence statistiquement significative » (selon la terminologie utilisée dans ce système de notation), à ★ lorsque l'établissement figure parmi les 10% plus mauvais, avec une « différence statistiquement significative ». Lorsque les résultats sont plus ou moins égaux à ceux attendus ou que la différence n'est pas statistiquement significative, la note attribuée est ★★★. Un établissement peut, malgré un bon score, n'avoir que la note ★★★. Pas sûr que le « consommateur » comprenne tout.

★★★★★	Figure parmi les 10% les mieux notés, avec une « différence statistiquement significative ».
★★★★	Figure parmi les 50% les mieux notés (mais pas les 10%) avec une « différence statistiquement significative ».
★★★	Les résultats sont plus ou moins égaux à ceux attendus ou la « différence n'est pas statistiquement significative ».
★★	Figure parmi les 50% les moins bien notés (mais pas les 10% plus mauvais) avec une « différence statistiquement significative ».
★	Figure parmi les 10% plus mauvais, avec une « différence statistiquement significative ».

La notion de « différence statistiquement significative » étant obscure, on se propose de l'expérimenter par simulation sur un tableur.

• Prenons un premier exemple avec un petit échantillon de taille 12.
Pour un certain type d'opération, deux petits hôpitaux ont réalisé 12 interventions :
– hôpital A : 1 décès sur 12, soit environ 8%.
– hôpital B : 3 décès sur 12, soit 25%.
On cherche à noter ces hôpitaux selon le barème annoncé.
On suppose que pour ce type d'opération la fréquence de décès est, dans les conditions habituelles, p = 17% (il pourra s'agir de la fréquence des décès pour ce type d'opération sur l'ensemble des hôpitaux américains).

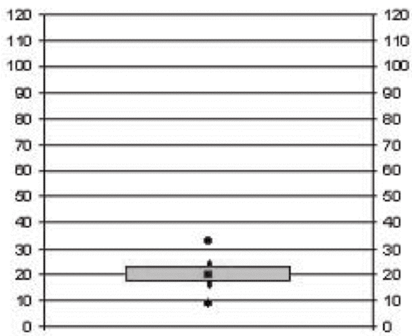
On simule le nombre de décès dans de petits hôpitaux réalisant 12 opérations de ce type par an, selon cette fréquence « normale ».
On obtient un graphique tel que celui montré ci-contre.
On peut, en appuyant sur la touche F9, refaire des simulations sur 1 000 hôpitaux.
L'intervalle interquartile le plus souvent observé est [1, 3] (voir le graphique). Il s'agit des résultats les plus fréquents : dans 50 % des cas, lorsqu'on opère avec $p = 0{,}17$, il y a 1, 2 ou 3 décès sur 12 opérations effectuées.



1000 simulations de 12 opérations

Les résultats des hôpitaux A et B se situent dans cet intervalle interquartile. On attribue donc aux hôpitaux A et B la même note ★★★ car les différences observées ne sont pas « significatives » compte tenu du petit nombre d'interventions, le « hasard » seul peut suffire à expliquer cette différence.

• Prenons un second exemple avec un grand échantillon de taille 120.
Pour le même type d'opération chirurgicale, deux hôpitaux de grande dimension ont réalisé 120 interventions et ont obtenus des pourcentages de décès analogues à ceux des hôpitaux A et B :
– hôpital C : 10 décès sur 120, soit environ 8%.
– hôpital D : 30 décès sur 120, soit 25%.
Nous allons simuler le nombre de décès dans des hôpitaux réalisant 120 opérations de ce type par an, selon la fréquence théorique de décès $p = 17\%$. Les résultats obtenus sont du type de ceux de la figure ci-contre.



1000 simulations de 120 opérations

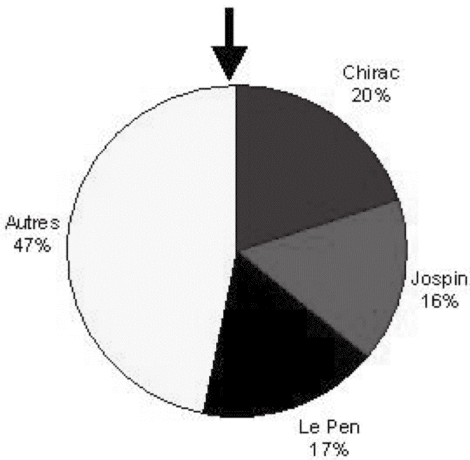
En faisant plusieurs fois F9, on se convainc de la faible incidence des fluctuations d'échantillonnage.
L'intervalle interquartile le plus souvent observé est [17, 23] (voir graphique) et les résultats des hôpitaux C et D , respectivement 10 et 30, se situent à l'extérieur des moustaches des premier et neuvième déciles. On peut donc attribuer à l'hôpital C la note ★★★★★ et à l'hôpital D la note ★.

Estimation

Présidentielle 2002

Lors du premier tour des élections présidentielles, le dernier sondage publié par l'institut B.V.A., effectué sur 1 000 électeurs le vendredi 19/04/02, prévoyait : Jacques Chirac 19 %, Lionel Jospin 18 %, Jean-Marie Le Pen 14 %.

La surprise a été grande le dimanche 21/04/02 au vu des résultats, puisque Jean-Marie Le Pen figurait au second tour : Jacques Chirac 19,88 %, Lionel Jospin 16,18 %, Jean-Marie Le Pen 16,86 %.
On peut rappeler aux élèves de seconde que la formule des fourchettes de sondage à plus de 95 % de confiance, calculée à partir d'une fréquence f obtenue sur un échantillon aléatoire de taille 1 000 est $\left[f-\frac{1}{\sqrt{1\,000}}\,;f+\frac{1}{\sqrt{1\,000}}\right]$, leur demander de calculer les trois fourchettes à partir du sondage B.V.A. et de les représenter sur un graphique. On peut alors poser la question suivante, en se basant sur ces fourchettes, peut-on « prévoir » l'ordre des candidats au premier tour de l'élection ?
À partir du dernier sondage B.V.A., le calcul des fourchettes donne environ les estimations suivantes :
Jacques Chirac : [15,8% ; 22,2%] ; Lionel Jospin : [14,8% ; 21,2%] ; Jean-Marie Le Pen : [10,8% ; 17,2%].
Comme on le constate, ces fourchettes incitent à la prudence. Elles ne permettent pas de prévoir l'ordre des candidats. Elles n'excluent même pas que le troisième dans le sondage soit le premier dans les urnes.
Même un sondage de taille 1 000 le jour de l'élection, « sorti des urnes », aurait été assez indécis, compte tenu des résultats particulièrement serrés des candidats en deuxième et troisième position. On peut utilement expérimenter de tels sondages, par simulation, sur le tableur.
Le jour de l'élection, la structure de l'électorat correspond au « camembert » ci-contre.
La simulation d'un sondage aléatoire sur 1 000 personnes consiste à faire tourner 1 000 fois cette roue de loterie. On peut faire la simulation sur le tableur et observer de nombreux sondages de taille 1 000. On s'aperçoit que ces sondages ont beaucoup de difficultés pour distinguer les candidats en deuxième et troisième position, que les fluctuations des résultats sont importantes et que parmi eux figurent des sondages analogues à celui de B.V.A. le vendredi précédent le premier tour de 2002.



La pratique des sondages politiques

En France, à la différence semble-t-il des États-Unis, la méthode des sondages aléatoires est peu utilisée et les instituts ont souvent recours, pour leurs enquêtes d'opinion, à la méthode des quotas. Dans cette méthode, on exploite la structure connue de la population (par exemple grâce au recensement) pour reproduire la même structure dans l'échantillon. On choisit pour cela certains caractères de la population, que l'on pense devoir être en rapport avec l'enquête menée, comme le sexe, l'âge, la catégorie socioprofessionnelle, le type de commune, … Ces caractères sont nommés variables de contrôle. Si l'on connaît la distribution de la population selon ces variables de contrôle, on obtient ainsi des quotas qui devront être respectés par les enquêteurs.

« Avec la méthode des quotas, il n'existe pas de loi mathématique permettant de déterminer la marge d'erreur d'un sondage », explique Jean-François Doridot, directeur du département opinion d'Ipsos[(2)], « en pratique toutefois, on considère que la marge d'erreur des sondages par quotas est égale, voire inférieure à celle des sondages aléatoires. » Des études ont cependant montré que cette méthode avait tendance à sous-représenter les travailleurs de l'industrie, les personnes les moins instruites ou ayant peu d'activités sociales… On peut douter de l'affirmation précédente selon laquelle la marge d'erreur par la méthode des quotas est égale « voire inférieure » à celle des sondages aléatoires. Le hasard est encore, pour éviter les biais, le meilleur allié du statisticien.

De plus, des difficultés spécifiques aux sondages politiques (ou aux enquêtes d'opinion) tiennent non plus aux problèmes de biais affectant la constitution de l'échantillon, mais aux réponses des sondés : abstentionnistes répugnant à avouer qu'ils n'ont pas l'intention de voter, indécision jusqu'au dernier moment, sympathisants d'extrême droite hésitant à afficher leurs opinions… À la lumière des élections précédentes, des coefficients rectificatifs sont alors appliqués, faisant ainsi du sondage politique davantage un art alchimique qu'une science.

Pour « redresser » les résultats bruts des sondages effectués lors de la campagne présidentielle de 2002, on a pratiqué les « pondérations » suivantes. Lionel Jospin, lorsqu'il recueille 26 à 27 % en données brutes, est crédité de 22 % après pondération, de même Jacques Chirac passerait de 30 % en brut à 27 % en pondéré, ou Jean-Marie Le Pen de 4 % à 8 % en pondéré (chiffres cités par Philippe Méchet de la Sofres). Ces pondérations sont établies à partir de plusieurs élections antérieures et de questions posées par le sondeur et permettant de mesurer le « degré de certitude » du choix de l'électeur.

Adéquation

Dans une revue technique de juin 2003[(3)], des « spécialistes », ingénieurs et hydrologues, insistent sur le fait « qu'informer les citoyens sur les risques d'inondation par des messages clairs et compréhensibles est un enjeu social et économique fort mais complexe ». Ils affirment que différentes notions sous-jacentes à ces questions sont « difficilement assimilables sans la maîtrise de concepts mathématiques minimaux ». Voici donc, à n'en pas douter, une formation mathématique utile au citoyen. Les auteurs de l'article ont noté qu'en contexte aléatoire, en l'occurrence les risques de crue, plusieurs biais psychologiques

(2) Journal *Le Monde* du 17/03/02.

(3) Revue *Ingénieries – eau, agriculture, territoires* n° 34 juin 2003, article intitulé *Risque d'inondation : une notion probabiliste complexe pour le citoyen* de N. Gendreau, F. Grelot, R. Garçon et D. Duband.

interfèrent dans la compréhension des situations comme « une réelle difficulté à appréhender des événements à faible probabilité ». Est cité à l'appui de cette constatation un extrait du journal *Le Point* du 13/09/2002, où il est écrit à propos de la Seine : « une crue comparable à celle de 1910 se produit en moyenne tous les cent ans et la probabilité d'une telle catastrophe augmente d'année en année ». Le fait est qu'on se rapproche du centenaire, 2010, faut-il en déduire que le « risque » augmente ? Par définition une crue « centennale » a une probabilité $p = 1/100$ de se produire pour une année donnée. C'est le cas pour l'année prochaine, et le « risque », s'il est vrai que le « modèle crue centennale » est vrai, est le même pour l'année à venir que ce qu'il était en 1911. Ce qui n'est ni plus ni moins rassurant. En revanche, il est vrai de dire que plus la période d'étude est longue, plus le risque est grand. C'est la répétition d'épreuves indépendantes. Le risque que la crue centennale se produise dans les 20 ans à venir est plus grand que celui qu'elle se produise dans les 10 ans à venir. Dans ces situations, le futur est généralement indépendant du passé (c'est ce qui est implicite ici) et ce n'est pas facile à appréhender. Nos spécialistes des inondations concluent ainsi leur article :
« Il est particulièrement difficile de prendre conscience des situations à risque, et notamment en matière d'inondation pour lequel il existe une croyance collective que les aménagements, entre autre, permettent de maîtriser la situation. Or il s'avère qu'il reste toujours un risque résiduel, rarement affiché, souvent tu. Il est indispensable qu'un minimum d'information soit donné au citoyen, du moins au riverain. En effet le citoyen peut légitimement demander la nature des choix politiques qui définissent par défaut ou volontairement le niveau de risque choisi. De même, le riverain est en droit de connaître le risque auquel il est exposé.
[…] Informer des risques reste un enjeu fort de société. C'est une tâche ardue mais qui nécessite une vrai formalisation et un vrai effort pédagogique. Ce n'est que grâce à ces actions que nous pourrons plus facilement comprendre, et par là accepter les catastrophes qui surviendront, et réagir en acteur averti. »

Pics d'ozone

On dispose de l'historique suivant donnant le nombre de jours de dépassement du « niveau d'information et de recommandations » (180 micro-grammes par m^3) du polluant O3 (Ozone) pour la « zone rurale Nord et Est de Paris » (source : www.airparif.asso.fr).

Année	1999	2000	2001	2002	2003
Nb de jours	5	1	11	4	21

Le hasard seul pourrait-il expliquer cette répartition ?
Il s'agit d'abord, par simulation sur tableur, d'examiner le hasard à l'œuvre. C'est-à-dire de générer 42 valeurs selon un modèle équiprobable en cinq classes et de calculer la distance au carré qui sépare la répartition observée de l'équirépartition théorique : $\mathrm{d}^2_{\mathrm{obs}} = \sum_{i=1}^{5}\left(f_i - \frac{1}{5}\right)^2$.

Les simulations montrent que dans 90 % des cas le modèle équiparti conduit à $d^2_{obs} \leq 0,04$.

On adopte la règle de décision suivante :

• Si $d^2_{obs} > 0,04$ on rejette le modèle équiréparti avec un risque de 10%.

• Si $d^2_{obs} \leq 0,04$ on accepte le modèle équiréparti (sans quantification du risque sur cette décision).

Mettons en œuvre ce test sur les données des pics d'ozone.

On a $d^2_{obs} = \left(\frac{5}{42}-\frac{1}{5}\right)^2 + \ldots + \left(\frac{21}{42}-\frac{1}{5}\right)^2 \approx 0,142..$

On dépasse largement la limite de 0,04, on rejette donc le modèle équiréparti, avec un risque très inférieur à 10 %. Le hasard n'est pas ici une explication suffisante (on s'en doutait un peu…).

Pour en savoir plus

– DUTARTE (Philippe) – *L'induction statistique au lycée illustrée par le tableur* – Didier 2005.

– GIRARD (Jean-Claude) et HENRY (Michel) – L'inférence statistique. Deux exemples d'application du calcul des probabilités : estimations et tests d'hypothèses – *Enseigner les probabilités au lycée* – Commission inter-IREM Statistique et Probabilité 1997.

– PIEDNOIR (Jean-Louis) et DUTARTE (Philippe) – *Enseigner la Statistique au lycée : des enjeux aux méthodes* – IREM PARIS-NORD 2001.

– ROBERT (Claudine) – *Contes et décomptes de la statistique. Une initiation par l'exemple* – Éd. Vuibert 2003.

– SAPORTA (Gilbert) – *Probabilités, analyse des données et statistiques* – Éd. Technip.

– VERLANT (Bernard) et SAINT-PIERRE (Geneviève) – *Statistique et probabilités – BTS* – Éd. Foucher.

– WONNACOTT et WONNACOTT – *Statistique* – Éd. Économica.