Adéquation statistique à un modèle

Exemples dans le domaine de l'environnement

Philippe Dutarte(*)

« Quelque fois les phénomènes paraissent dépendre d'une cause régulière ; et cependant, ils ne sont que le résultat de ces causes irrégulières, variables et inconnues, auxquelles nous donnons le nom de hasard. C'est à l'analyse des probabilités à déterminer ***jusqu'à quel point*** *une cause régulière est probable en vertu de ces phénomènes, et à l'indiquer aux philosophes, comme objet digne de leurs recherches. »*

P. S. Laplace – *Leçons à l'École normale de l'An III* (1795).

Introduction

Pour la première fois, les programmes de seconde, premières et terminales ES et S, entrés en application en 2000, 2001 et 2002, font une place aux méthodes statistiques dites « inférentielles », au travers de sujets comme : fluctuation d'échantillonnage, sondages, adéquation de données à un modèle équiréparti. L'intention des concepteurs de ces programmes, fort louable, était clairement affichée : *« Former les élèves en statistique, c'est leur donner les moyens de développer une forme de pensée critique sans laquelle ils seront exclus du débat social et scientifique »* (projet de programme de terminale ES et S). La presse, et pas seulement la presse scientifique, se fait en effet régulièrement l'écho de sondages, évoque les notions de risques, d'un illusoire risque zéro, de « preuves statistiques » (qui ne sont pas des preuves à 100 %…), de modèles, de méthodes statistiques aidant à prendre des décisions dans un environnement incertain. Certains sociologues prétendent que nous sommes entrés dans la « société du risque », nécessitant une politique de « précaution », terrain qu'on ne saurait laisser aux seuls experts et groupes de pression.

Les sections de techniciens supérieurs ont été confrontées à l'enseignement de la statistique inférentielle en mathématiques dès 1988. L'industrie française, avec quelque retard, notamment par rapport aux États-Unis et au Japon, a en effet dû intégrer dans les années 1960/70 les méthodes statistiques de contrôle de qualité et de fiabilité. Face à la « déraisonnable » mais néanmoins spectaculaire efficacité de ces méthodes, appliquées à l'industrie japonaise et américaine, il en allait de la survie de l'industrie nationale. Dès 1988 sont mis en place des stages de formation pour les enseignants de mathématiques des sections de BTS concernées. À ce propos, Bernard Verlant, responsable de la Commission inter-IREM « Lycées techniques », remarque : « *Je m'occupe de la formation continue depuis 1977 et la statistique inférentielle est le seul thème de formation qui ait nécessité durant cette période un dispositif aussi étalé dans le temps. Les stages n'ont pas désempli jusqu'en 2000… Quant à l'accueil réservé au premier exercice du bac ES Métropole 2003* [adéquation à une loi équirépartie]*, il ne fait que confirmer la difficulté d'introduction de cette notion* ».

(*) Commission inter-IREM Lycées techniques.

Et pourtant, il serait souhaitable que la tentative des nouveaux programmes de seconde à terminale dans ce domaine ne soit pas un échec. Une éducation, assez précoce, des futurs citoyens aux méthodes d'inférence statistique est sans doute aussi nécessaire que délicate.

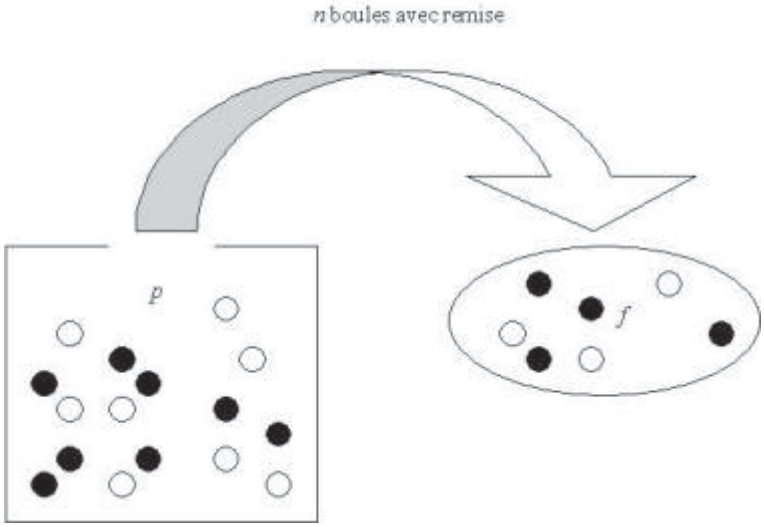
Nous développerons dans cet atelier le thème de « l'adéquation statistique à un modèle », pour reprendre la terminologie des programmes de terminale (ES et S), c'est-à-dire des « tests statistiques ». C'est un peu l'aboutissement, dans le domaine de la statistique inférentielle, des thèmes nouveaux mis aux programmes des lycées. La « méthode statistique », avec ses spécificités par rapport aux habituelles démarches mathématiques, s'y déploie pleinement et sa présentation aux élèves ne va pas de soi.

Nous resterons à un niveau élémentaire, la plupart du temps accessible aux élèves de terminale (et donc réutilisable en classe), l'essentiel (et le « risque » de contresens) se situant davantage dans la démarche que dans les démonstrations mathématiques. Pour « coller » au thème des journées d'Orléans, « Mathématiques et environnement », les exemples seront choisis dans ce domaine.

Trois confusions fréquentes

Dans le cadre des thèmes de statistique inférentielle des programmes des lycées, trois confusions fréquentes, et plus graves qu'il n'y paraît, surviennent souvent chez les élèves, bien sûr, mais surtout chez les professeurs débutant dans ce domaine (nous en parlons d'autant plus facilement que nous en avons fait partie), dans les manuels de terminale, et même dans certains sujets d'examen ! Ce n'est peut-être pas des plus pédagogiques, mais commençons par voir ce qu'il faut éviter.

Pour faire simple, considérons une urne « de Bernoulli » contenant une proportion p de boules blanches, dont on extrait (tirage avec remise) n boules, la proportion de boules blanches dans le tirage (ou échantillon) étant notée f.



Trois thèmes sont envisagés au lycée : fluctuation des échantillons, estimation (« fourchette de sondage ») et test d'hypothèse (dans le cas de l'équidistribution seulement, c'est-à-dire ici, $p = 1/2$).

- Si p est connu, on peut dire : dans plus de 95% des tirages

$$f \in \left[p - \frac{1}{\sqrt{n}}, p + \frac{1}{\sqrt{n}} \right]$$

(intervalle de fluctuation de 95% des échantillons, nommé aussi intervalle de probabilité).

- Si p est inconnu, mais que l'on procède à un tirage donnant une valeur de f, on peut dire : dans plus de 95% des tirages on affirmera à juste titre :

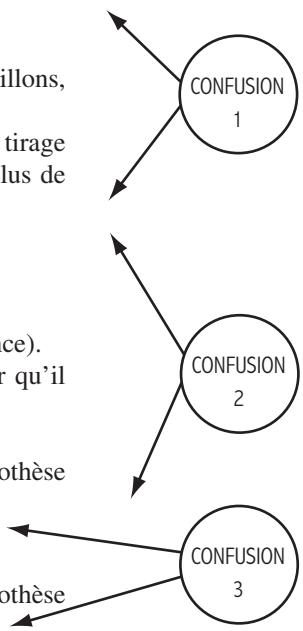
$$p \in \left[f - \frac{1}{\sqrt{n}}, f + \frac{1}{\sqrt{n}} \right]$$

(fourchette de sondage ou intervalle de confiance).

- p est inconnu mais on a des raisons de penser qu'il vaut peut-être 1/2, on possède une valeur de f.

  – Si $f \notin \left[\frac{1}{2} - \frac{1}{\sqrt{n}}, \frac{1}{2} + \frac{1}{\sqrt{n}} \right]$, on rejette l'hypothèse $p = 1/2$ avec un risque de 5%.

  – Si $f \in \left[\frac{1}{2} - \frac{1}{\sqrt{n}}, \frac{1}{2} + \frac{1}{\sqrt{n}} \right]$, on accepte l'hypothèse $p = 1/2$ avec un risque inconnu.



La *première confusion* consiste à pousser la symétrie apparaissant dans la formulation de l'intervalle de fluctuation et de la fourchette de sondage, jusqu'à confondre les deux. Notons que si l'intervalle de fluctuation est au programme « obligatoire » de seconde, la fourchette de sondage n'est qu'un thème d'étude « facultatif ». Dans le premier cas, on connaît p (information énorme) et un raisonnement purement déductif, même s'il est probabiliste, et admis en seconde, permet d'obtenir l'intervalle de fluctuation, que l'on peut nommer intervalle de probabilité. Le second cas, celui de l'estimation, est beaucoup plus délicat et de nature profondément différente. On ne connaît qu'une seule valeur de f (c'est-à-dire pas grand chose) et c'est un raisonnement de nature inductive qui permet de donner un intervalle « de confiance ». Ce n'est pas pour rien que l'on emploie ici cette expression « confiance » (et pas ailleurs) au lieu de « probabilité ». Si l'on dit, par exemple, « un intervalle de confiance à 95 % pour p est [0,3 ; 0,5] », posez-vous la question « où est la variable aléatoire ? ». Une phrase telle que « dans 95 % des tirages $p \in$ [0,3 ; 0,5] » ne veut rien dire : p est ou n'est pas dans cet intervalle. Nous n'insisterons pas davantage là-dessus, ce n'est pas le thème de cet atelier.

La *seconde confusion* consiste à confondre estimation et test. Elle se traduit parfois dans le vocabulaire : on parle de « confiance » pour un test, ou de « fourchette » pour un intervalle d'acceptation de l'hypothèse testée, un sujet de BTS 2003 demande un « intervalle de confiance au seuil de risque 10 % ». En mathématiques, la rigueur commençant par le vocabulaire, évitons de tout mélanger. Un test est plus facile à comprendre que l'estimation. La construction de l'intervalle d'acceptation de l'hypothèse testée est déductive (c'est l'intervalle de fluctuation lorsque l'hypothèse est vraie). La difficulté réside dans la notion de risque.

C'est la *troisième confusion*, parmi celles pointées ici, le risque n'est clair que lorsque l'on rejette l'hypothèse testée (on va détailler cela dans les exemples qui suivent). Une expression telle que « peut-on accepter au risque de 10 % l'hypothèse selon laquelle il y a équiprobabilité » (Bac ES 2003) est bien dangereuse.

Jouez à faire la chasse à ces trois « confusions » dans les manuels scolaires, vous serez surpris de votre moisson.

Modélisation et tests

De la connaissance des fluctuations des échantillons aléatoires (extraits d'une population supposée connue), la méthode statistique infère les qualités vraisemblables d'une population inconnue dont un échantillon est issu, aidant ainsi à la prise de décision. Dans les procédures d'estimation, type fourchette de sondage, on n'a pas d'*a priori* sur les qualités que devrait posséder cette population. À l'inverse, bien souvent (dans les contrôles de qualité par exemple), on a à l'esprit une norme que devrait vérifier la population. On procède alors à un test statistique. On formule une hypothèse (appelée « hypothèse nulle ») dont on suppose qu'elle est vérifiée par la population. Il s'agit de voir si les résultats observés sur l'échantillon sont, aux fluctuations près, compatibles avec cette hypothèse, ou si une différence « significative » avec les résultats attendus rend improbable l'hypothèse nulle.

Le test retenu par les programmes de terminale ES et S est celui de l'adéquation à une loi équirépartie, dont l'exemple type est le test d'un dé. La question est de « savoir » si un dé est truqué d'après l'observation d'une centaine de lancers, par exemple. L'hypothèse nulle est celle d'un modèle équidistribué (chaque face apparaît avec une probabilité 1/6). La connaissance des fluctuations des résultats de 100 lancers selon ce modèle permet de déterminer des limites au delà desquelles il est peu vraisemblable d'avoir affaire à un dé non truqué. Bien sûr, quand on dit « savoir », ce n'est pas une connaissance à 100 % (pour cela il ne faut pas utiliser une méthode statistique mais passer le dé aux rayons X par exemple), mais l'intérêt de la méthode consiste à évaluer les risques d'erreur. Cependant, si le risque de rejeter à tort l'hypothèse nulle est facile à quantifier, celui d'acceptation à tort de l'hypothèse nulle est moins évident, comme on le précise par la suite.

Encore une fois, tout n'est pas aussi simple qu'il n'y paraît à première vue. Raison de plus pour éduquer nos futurs citoyens à cette notion de risque lors d'une prise de décision selon une procédure statistique. Ces procédures sont en effet de plus en plus utilisées et la gestion des risques est une conséquence nécessaire de notre organisation économique et sociale. Par ailleurs, la procédure du test statistique est souvent liée à celle de la validation d'un modèle et donc à la modélisation. Les scientifiques ne sont pas les seuls à modéliser : nombre de décisions sur des enjeux

de société se prennent également sur la base de modélisations.

Voici quelques exemples simples (voire simplifiés) pouvant en grande partie faire l'objet d'exercices en terminale (la fréquente interdisciplinarité des situations fait que la statistique inférentielle trouvera sa place dans nombre de projets type T.P.E.).

Un exemple pour comprendre la méthode : le Q.C.M

L'expérience en classe (terminale S et BTS) montre qu'un excellent exemple pour faire comprendre à nos élèves la notion de risque lors d'un test statistique est celui où le test en question est un examen scolaire (on s'appuie sur le vécu !). La présentation suivante utilise la loi binomiale, qui est connue en terminale S.

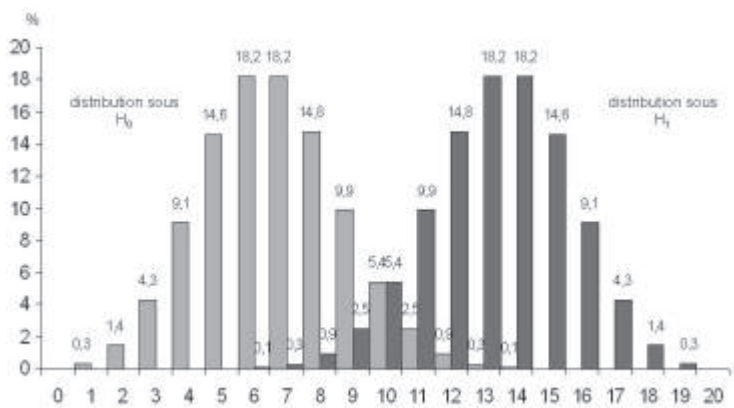
Un professeur construit un Q.C.M. de 20 questions indépendantes, proposant trois réponses possibles à chaque question, une seule réponse étant exacte. Il souhaite par cet examen éliminer environ 95 % des élèves n'ayant rien appris et répondant au hasard. La détermination du nombre de bonnes réponses pour être admis à l'examen revient à la construction d'un test statistique. On désigne par p la probabilité pour qu'un élève donné réponde bien à une question portant sur le programme de ce Q.C.M.

L'hypothèse nulle H_0 est « $p = 1/3$ » (l'élève répond au hasard). L'hypothèse alternative, notée H_1, sera « $p > 1/3$ » (lorsque l'élève ne répond pas au hasard).

Sous l'hypothèse nulle, le nombre de bonnes réponses correspond à la réalisation d'une variable aléatoire suivant la loi binomiale de paramètres $n = 20$ et $p = 1/3$ (répétition de 20 épreuves aléatoires indépendantes à deux issues possibles, succès ou échec, avec une probabilité de succès 1/3). Or la distribution de cette loi (cf. fig. 1) montre que la probabilité d'avoir strictement moins de 11 réponses exactes est environ 96 %. Le professeur décidera donc que pour être reçu à l'examen, on doit avoir au moins 11 bonnes réponses.

Le risque de rejeter à tort l'hypothèse H_0 est environ 4 %. Ce risque correspond au fait de recevoir à l'examen un élève qui répond au hasard. C'est un risque que les élèves sont prêts à prendre ! (en fait, c'est le risque du professeur).

Les élèves perçoivent bien qu'il existe un autre risque : celui d'échouer à l'examen alors qu'on ne le mérite pas (« la faute à pas de chance »). C'est le risque consistant à accepter à tort l'hypothèse nulle H_0.



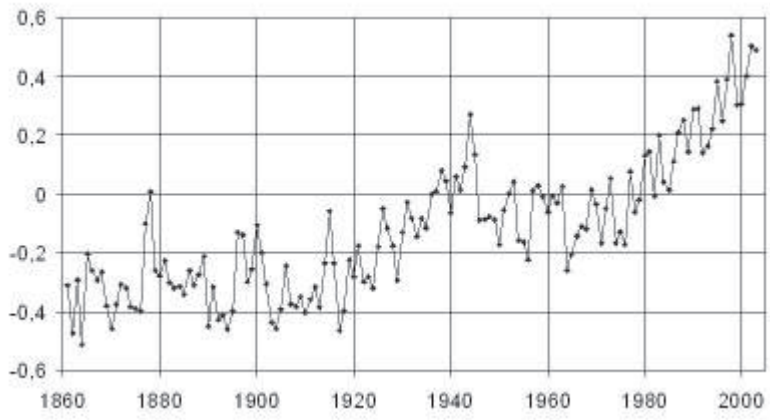
Ce second risque est plus difficile à évaluer, puisqu'il dépend des connaissances du candidat. Supposons que celui-ci ait une probabilité $p = 2/3$ de répondre correctement à chaque question. Pour ce candidat, le second risque correspond à la probabilité d'avoir strictement moins de 11 succès selon la loi binomiale de paramètres $n = 20$ et $p = 2/3$, soit un risque d'être recalé à tort d'environ 9 %.

On peut avoir le sentiment que dans le calcul de la « barre d'acceptation », on suspecte *a priori* le candidat d'être coupable de répondre au hasard. Cette impression est exacte. Dans la construction d'un test, l'hypothèse nulle est privilégiée dans la mesure où on ne la rejettera que si les observations sont vraiment trop peu compatibles.

Cette procédure du test statistique présente bien des défauts et peut sembler discutable (historiquement, elle fut très discutée). Élaborée dans les années 1930 par Jerzy Neyman (1894-1981) et Egon Pearson (1895-1980), elle s'est imposée, par son efficacité, comme une démarche de décision en milieu aléatoire universellement admise.

Adéquation à un modèle équiréparti : exemple du réchauffement de la planète

Les données suivantes, concernant la température moyenne du globe (un sujet « sensible »), sont celles du Groupe d'experts intergouvernemental sur l'évolution du climat, dépendant des Nations Unies (*Intergouvenmental Panel on Climate Change*, dont le site est www.ipcc.ch). Elles sont disponibles sur le site du *Climatic Research Unit* en Grande-Bretagne (www.cru.uea.ac.uk/cru/data/temperature).

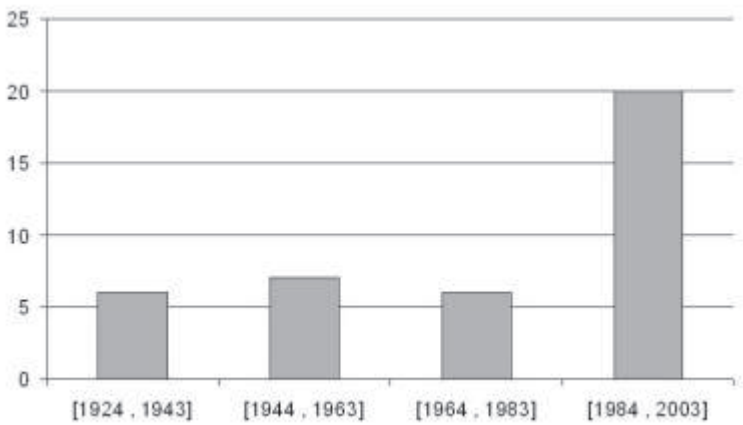


Si l'on considère les 40 années « chaudes » sur la période 1861-2003, une seule apparaît avant 1930, en 1878 (sur la figure 4 l'échelle des ordonnées est en degrés Celsius, le 0 correspondant à une climatologie moyenne sur la période 1961-1990). Il n'est guère utile de procéder à un test statistique pour savoir si la répartition des 40 années les plus chaudes peut être considérée comme totalement aléatoire sur la période 1861-2003. Cependant, comme la figure suggère un changement qualitatif à partir des années 1920-1930, nous allons étudier s'il y a une accélération

significative sur la deuxième période, en utilisant l'adéquation à un modèle équiréparti aux programmes de terminale ES et S.

On choisit d'étudier la période [1924, 2003] et de la partager en quatre intervalles d'égale longueur (20 années) dans lesquels on considère la distribution des années les plus chaudes (il y en a 39).

Classe nº i	Années	Effectifs observés x_i	Fréquences observées f_i
1	[1924 , 1943]	6	0,154
2	[1944 , 1963]	7	0,179
3	[1964 , 1983]	6	0,154
4	[1984 , 2003]	20	0,513



Supposer que la répartition des 39 années « chaudes » sur cette période de 80 ans se fait « au hasard » revient à dire que l'on observe ici le résultat d'un tirage simultané de 39 boules dans une urne contenant 80 boules marquées du millésime de chaque année. De façon plus précise, utilisant quatre couleurs, on aurait peint chaque boule de la couleur associée à la période de 20 ans à laquelle elle correspond. Supposer que les années les plus chaudes sont le fruit du hasard implique que l'on prenne simultanément 39 boules dans cette urne quadricolore équirépartie. Étudions l'hypothèse selon laquelle la distribution de ces 39 années s'effectue au hasard selon un modèle équiréparti.

L'écart entre les fréquences observées et la fréquence 1/4 théorique peut être quantifié par

$$d^2_{\text{obs}} = \sum_{i=1}^{4} \left(f_i - \frac{1}{4} \right)^2 .$$

On observe

$$100 \times d^2_{\text{obs}} \approx 9,3.$$

On simule, sous l'hypothèse d'équirépartition, la distribution de 39 années entre 4 classes (tirages simultanés dans l'urne équirépartie) puis on calcule la valeur $100 \times d_{\text{obs}}^2$ obtenue sur cette simulation. On peut effectuer 1 000 simulations pour avoir un bon aperçu des fluctuations de $100 \times d_{\text{obs}}^2$ sous l'hypothèse d'équirépartition. Les simulations montrent que plus de 99 % des valeurs de $100 \times d_{\text{obs}}^2$ obtenues sous l'hypothèse d'équirépartition sont inférieures à 4.

Au risque de 1 %, on rejette donc le modèle équiréparti pour expliquer la répartition des 39 années les plus chaudes dans les quatre classes. La différence observée est « significative ».

Telle est la présentation que l'on peut faire en terminale.

Si au lieu d'un tirage simultané (sans remise) dans une urne équirépartie, on fait tourner une roue de loterie partagée en quatre quartiers égaux (ce qui équivaut à un tirage avec remise), on peut montrer (mais c'est hors programme de terminale) que la variable aléatoire correspondant aux réalisations de $4 \times 39 \times d_{\text{obs}}^2$ sous l'hypothèse d'équirépartition suit approximativement la loi du khi-deux à 3 degrés de liberté. On vérifie alors que la probabilité d'obtenir $100 \times d_{\text{obs}}^2 \approx 9,3$ sous cette hypothèse est 0,002. Cela signifie que la répartition (6, 7, 6, 20) observée avec les années les plus chaudes est très peu vraisemblable comme issue d'un tirage avec remise dans une urne équirépartie, et encore moins vraisemblable comme provenant d'un tirage simultané (les écarts des tirages simultanés avec l'équirépartition ont tendance à être moindres que ceux des tirages avec remise). Cela laisse moins de 0,2 % de chances à une explication totalement aléatoire de la répartition des années les plus chaudes sur la période 1924-2003. Il reste bien peu de place au hasard.

Un test fréquemment pratiqué est celui de comparaison de deux moyennes (« test de différence significative »). Celui-ci faisant intervenir la loi normale, il ne sera vu des élèves qu'après le bac (dans bien des domaines, en particulier la médecine). On constatera que le raisonnement est analogue. Effectuons ce test sur les moyennes des températures observées avant et après 1930.

Les observations fournissent les résultats suivants :

Années	moyenne	écart type	effectif
[1861, 1929]	- 0,299	0,11	69
[1930, 2003]	0,045	0,18	74

Faisons l'hypothèse nulle selon laquelle les deux échantillons de taille 69 et 74 sont extraits aléatoirement de populations ayant la même moyenne. Le théorème limite central montre que la variable aléatoire D faisant correspondre à deux tels tirages la différence d des moyennes observées suit approximativement une loi normale de moyenne nulle et d'écart type $\sqrt{\frac{0,11^2}{69} + \frac{0,18^2}{74}}$. Or cette distribution de probabilité est telle que dans 99 % des cas D prend une valeur comprise entre −0,065 et 0,065.

La différence observée étant 0,344 on rejette l'hypothèse nulle au risque de 1 % (en fait le risque est bien plus faible). La différence est bien significative.

Conclusion

Nous espérons, par les exemples précédents (libre à vous d'en exploiter d'autres, grâce en particulier aux ressources d'Internet), avoir montré l'intérêt de l'enseignement des méthodes de statistique inférentielle pour l'éducation de nos futurs citoyens. Vous aurez noté au passage qu'il s'agit de situations motivantes (il y a bien d'autres catastrophes possibles à envisager, mais aussi des exemples plus positifs, dans l'étude de traitements médicaux par exemple) où l'on exploite des outils mathématiques parfois variés, ainsi que les possibilités de l'informatique. Cependant, la compréhension de la méthode statistique (où l'on n'est sûr de rien) nécessiterait un apprentissage assez précoce (dès la seconde) et pas trop superficiel, ainsi qu'une formation conséquente des professeurs. Les réductions d'horaires, les diminutions des budgets de formation ne vont pas, hélas, dans le sens de cette préoccupation et cela risque (non quantifié) de compromettre « la bonne intention », dans ce domaine, des nouveaux programmes de lycée.

Bibliographie

CALLON (Michel). *Agir dans un monde incertain. Essai sur la démocratie technique.* Le Seuil, 2001.

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

IREM PARIS-NORD. *Le nouveau programme de Statistique et Probabilité au lycée*. Villetaneuse, 2003.

LOMBORG (Bjorn). *L'écologiste sceptique*. Le Cherche Midi, 2004.

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

REEVES (Hubert). *Mal de Terre*. Seuil, 2003.

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), SAINT-PIERRE (Geneviève). *Statistique et probabilités. BTS*. Éd. Foucher.

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

Le Web pour les données statistiques.