Présenter aux futur(e)s professeur(e)s une image positive de la statistique et de ses enjeux citoyens

Philippe Dutarte

Lycée Édouard Branly Créteil – IREM de Paris-Nord.

L'attitude de la communauté mathématique française à l'égard de la statistique et des probabilités n'a pas toujours été favorable. On peut par exemple citer les propos diamétralement opposés d'André Weil et d'Émile Borel. Alors que le premier ironisait (vers 1940) sur le fait 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…* », le second affirmait (vers 1910) que « *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.* »

Notre expérience de formateur à l'IREM de Paris-Nord nous a le plus souvent conduit à rencontrer des professeurs de mathématiques dont l'attitude face à la statistique était celle de la méfiance voire du dénigrement, sur le thème de la statistique comme « art suprême du mensonge », qui plus est, sous couvert de méthodes prétendues scientifiques. Dans ces conditions, on comprend avec quelle conviction ces professeurs enseignent les chapitres de statistique, considérés comme hors des mathématiques. Jean-Pierre Raoult, dans un récent article[12], affirme qu'on « *ne peut faire admettre par les enseignants de collège ou de lycée une matière d'enseignement si on ne les a pas au préalable convaincus de son statut de discipline scientifique et, éventuellement, de sa place comme sous-discipline par rapport à celle qui, à leurs yeux, fonde leur légitimité disciplinaire (en ce qui nous concerne ici, les mathématiques)* », ajoutant que pour beaucoup la statistique apparaît comme une « *technique rudimentaire, mystérieuse et suspecte* ».

Or il y a de bonnes raisons pour enseigner la statistique comme un ensemble de méthodes scientifiques (mathématiques) permettant d'accéder à la connaissance (une certaine connaissance à défaut d'une connaissance certaine), en particulier parce que les thèmes abordés en statistique motivent les élèves pour les mathématiques et participent fondamentalement à l'éducation du citoyen. Jean-Pierre Raoult cite dans son article plusieurs grands noms à l'appui de cette thèse, à commencer par Condorcet qui affirmait à propos de l'utilité des « mathématiques sociales » : « *on verra qu'aucun de nos intérêts individuels ou publics ne lui est étranger, qu'il n'en est aucun sur lequel elle ne nous donne des idées plus précises, des connaissances plus certaines* », ou Didier Dacunha-Castelle écrivant en 1996 dans « Chemins de l'aléatoire. Le hasard et le risque dans la société moderne » que « *le recours aux mathématiques du hasard donne un instrument pédagogique important qu'il faut utiliser pour porter en avant la réflexion sur les sondages, sur l'acte médical ou sur les risques du nucléaire* » ainsi que Jean-Pierre Kahane affirmant devant l'Académie des Sciences en 2000 : « *En France, à la différence d'autres pays européens, les citoyens n'ont pas une formation suffisante à la prise en compte du mode de pensée statistique* ».

[12] Jean-Pierre Raoult – « La statistique : une discipline mathématique à part entière ? » – APMEP Revue PLOT n° 18 – 2007.

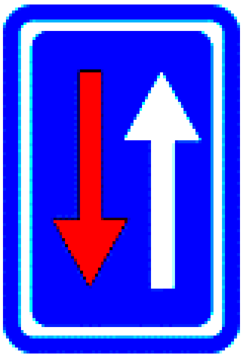
Il ne s'agit pas de feindre que les méthodes statistiques sont simples et sans danger (le développement de l'esprit critique est évidemment une priorité), mais il est certainement contre productif d'aborder la statistique sous l'angle du « méfiez-vous » avant d'être convaincu de son utilité et de son efficacité. Un ouvrage publié récemment par Nicolas Gauvrit[13] aux éditions Ellipse, se vend sous l'injonction « *Statistiques, méfiez-vous !* ». L'ouvrage est plutôt bien fait (quoique la majorité des exemples sont des constructions ad hoc), mais l'obsession des « pièges » et de la manipulation conforte dans ses opinions le lecteur ayant une image bien négative de la statistique. Ainsi, par exemple, le chapitre « *Statistique inférentielle : Qu'est-ce que c'est ?* » débute par (page 143) : « *Comme la statistique inférentielle est la partie la plus souvent utilisée des statistiques, mais aussi la plus dangereuse puisqu'elle permet de donner des conclusions affirmatives « prouvées » par l'expérience, il faut comprendre au moins l'idée générale qui se cache derrière elle pour en éviter les pièges.* », ou, trois pages plus loin, « *quel raisonnement nous cachent donc ceux qui affirment que la statistique a « démontré » que... ?* ». Oserons nous dire que la statistique ne sert pas qu'à tendre des pièges et que ces raisonnements prétendument cachés (s'agit-il d'un complot ?) commencent à s'enseigner (en théorie) en classe de seconde ?
Les activités présentées s'appuient sur une pratique des mathématiques « par l'expérience », facilitée par les moyens informatiques. 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.

Les exemples étudiés permettront d'envisager une réponse aux questions suivantes :

- La proportion d'élèves d'origine populaire à l'École polytechnique était de 21% en 1950 et de 7,8% en 1990. L'ascenseur social est-il en panne?
- La campagne de la Fédération française de cardiologie du printemps 2007affirme que « 80% des victimes d'infarctus sont des fumeurs ». L'information montre-t-elle que le risque d'infarctus est plus important pour un fumeur ?
- Une petite ville des États-Unis a connu 9 cas de leucémies chez de jeunes garçons en l'espace de 10 années. Doit-on en « accuser le hasard » ?
- Comment aux États-Unis, la statistique a-t-elle permis de mettre en évidence une discrimination conduisant à casser un jugement ?
- Dans un village du Canada, entre 1999 et 2003, sur 132 naissances, il n'y a eu que 46 garçons. Est-ce très étonnant ? Est-ce inquiétant (sachant que de nombreuses usines chimiques se trouvent à proximité) ?
- Le dernier sondage de 2002 ne prévoyait pas la présence de Jean-Marie Le Pen au second tour. Le dernier sondage CSA d'avril 2007 plaçait J-M Le Pen devant F. Bayrou. Les sondages sont-il scientifiques ?
- Le magazine *Le Point* du 13/09/2002 écrit : « Une crue de la Seine 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 ». Que faut-il en penser ?

[13] Nicolas Gauvrit – « Statistiques : méfiez-vous ! » - Ellipse 2007.

1 – Proportions et ascenseur social



L'exemple proposé ici a été cité par Jean-Louis Piednoir[14], en termes de proportions, pour illustrer le fait qu'il faut comparer ce qui est comparable et en particulier tenir compte de l'évolution de la constitution de la population sous-jacente. Le même exemple est beaucoup plus développé dans un ouvrage récent par Claudine Schwartz[15], en termes de probabilités conditionnelles, en particulier pour souligner le fait que la définition et les propriétés mathématiques d'un indicateur sont essentielles à connaître pour en comprendre le sens.

Chargé par le ministre Claude Allègre d'un rapport sur l'organisation des études supérieures, Jacques Attali, par ailleurs polytechnicien, citant une étude de Claude Thélot [16], concluait à une « dé-démocratisation » du recrutement des élites. Cette réflexion se fondait sur le tableau suivant :

Proportions d'élèves d'origine populaire	1951 - 1955	1989 - 1993
EP (École polytechnique)	21 %	7,8 %
ENA (École nationale d'administration)	18,3 %	6,1 %
ENS (Écoles normales supérieures)	23,9 %	6,1 %
HEC (Hautes études commerciales)	38,2 %	11,8 %
Grandes écoles (avec entrée sur concours)	29 %	8,6 %

D'après le tableau précédent, le verdict semble sans appel... l'ascenseur social est en panne. Mais il faut comparer ce qui est comparable et le tableau suivant, extrait de l'étude de Thélot, montre que la composition sociale française s'est considérablement modifiée entre ces deux époques.

Proportion de la population d'origine populaire	1951 - 1955	1989 - 1993
parmi les 20-24 ans	90,8 %	68,2 %

Il s'agit donc de tenir compte des deux tableaux pour se forger une opinion.

Énoncé (niveau 3ème – 2nde)

On dispose du tableau suivant, donnant, parmi les élèves reçus à l'École polytechnique, la proportion de ceux issus de classes « défavorisées » (classes notées *D*) et ceux issus de classes « favorisées » (classes notées *F*), pour les périodes 1950 et 1990.

Proportions d'élèves reçus à l'École polytechnique	1950	1990
d'origine défavorisée *D*	21 %	7,8 %
d'origine favorisée *F*	79 %	92,2 %

---

[14] Piednoir (Jean-louis) – Dutarte (Philippe) – *Enseigner la statistique au lycée : des enjeux aux méthodes* – IREM de Paris-Nord – 2001.

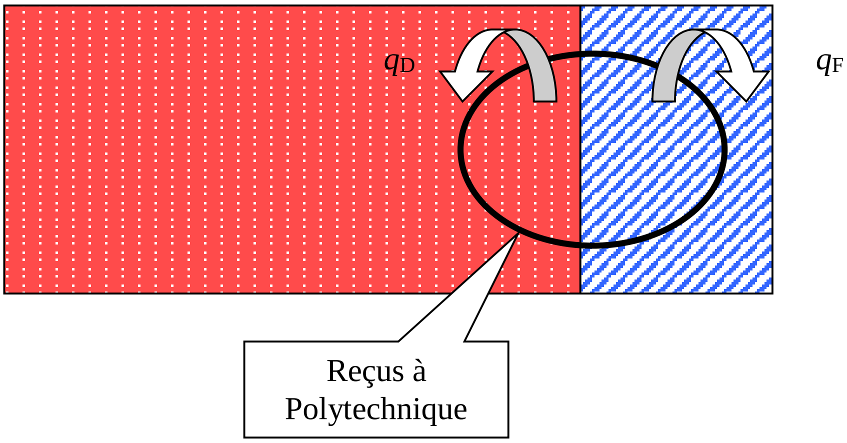
[15] Schwartz (Claudine) – *Pratiques de la statistique* – Vuibert 2006 – Chapitre 3 : *Des mots et des chiffres : probabilités conditionnelles, risques relatifs*. On consultera cet ouvrage pour mettre en perspective les activités proposées ici.

[16] Euriat et Thélot – *Le recrutement des élites scolaires depuis 40 ans* – Revue *Éducation et formation* juin 1995. Les chiffres cités ici proviennent de cet article.

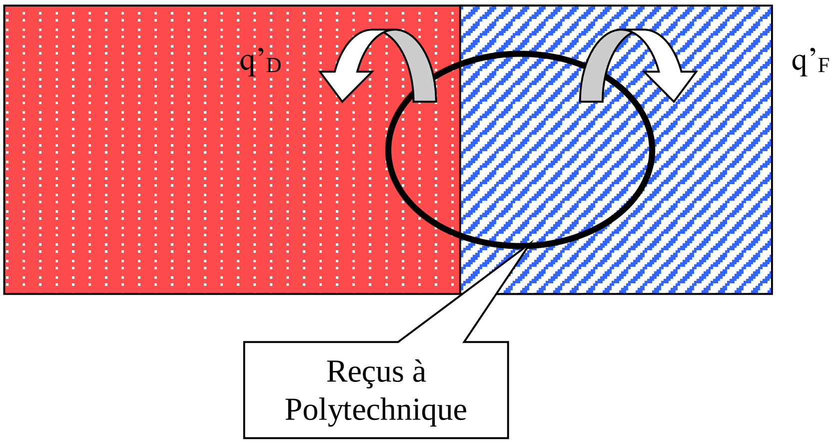
Pour étudier si la discrimination sociale est plus forte dans les années 1990 que dans les années 1950, il faut tenir compte de l'évolution de la composition de la société française entre ces deux périodes.

Proportions de la population des 20-24 ans	1950	1990
d'origine défavorisée D	90,8 %	68,2 %
d'origine favorisée F	9,2 %	31,8 %

1950 : population défavorisée | favorisée



1990 : population défavorisée | favorisée



1. Que signifie 21% dans le premier tableau ?
Que signifie 90,8% dans le second tableau ?
2. a) On note r le nombre de reçus à Polytechnique en 1950, exprimer en fonction de r le nombre de reçus à Polytechnique d'origine défavorisée en 1950.
b) On note p le nombre de jeunes de 20-24 ans en 1950.
Montrer que proportion q_D de reçus à Polytechnique parmi la population défavorisée en 1950 est $q_D = \dfrac{0{,}21 \times r}{0{,}908 \times p}$.
3. Donner, de même, l'expression de la proportion q_F de reçus à Polytechnique parmi la population favorisée en 1950.

4. On note $t = \dfrac{q_F}{q_D}$. Montrer que t □ 37.

On peut interpréter ce résultat en disant qu'en 1950 un jeune « favorisé » a 37 fois plus de chances d'entrer à Polytechnique qu'un jeune « défavorisé ».

5. Pour comparer avec la situation en 1990, on utilise les mêmes notations que ci-dessus, avec un ' . On considère donc $t' = \dfrac{q_F'}{q_D'}$.

Calculer t'.

Peut-on considérer que « l'ascenseur social » s'est amélioré entre 1950 et 1990 ?

6. On souhaite utiliser le tableur (selon l'image d'écran suivante) pour calculer les quotients t et t' dans les quatre autres cas de ce tableau :

Proportions d'élèves d'origine défavorisée (*D*)	1950	1990
ENA (École nationale d'administration)	18,3 %	6,1 %
ENS (Écoles normales supérieures)	23,9 %	6,1 %
HEC (Hautes études commerciales)	38,2 %	11,8 %
Grandes écoles (avec entrée sur concours)	29 %	8,6 %

On pourra organiser les calculs (ici pour l'École polytechnique) selon l'image d'écran, de sorte qu'il suffise de modifier le contenu des cellules B2 et C2.

	A	B	C	D
1	Proportion d'élèves reçus	1950	1990	
2	D	**0,21**	**0,078**	
3	F	0,79	0,922	
4				
5	Proportion des 20-24 ans	1950	1990	
6	D	0,908	0,682	
7	F	0,092	0,318	
8				
9		t	t'	
10		37,1283644	25,3509111	
11				

Interpréter les résultats obtenus.

Éléments de réponse

2. Le nombre de reçus d'origine *D* est $0{,}21$ □ r et le nombre de 20-24 ans d'origine *D* est 0,908 □ p.

3. On a $q_F = \dfrac{0{,}79 \times r}{0{,}092 \times p}$.

4. On a $q_D = \dfrac{0{,}21 \times r}{0{,}908 \times p}$. D'où $t = \dfrac{0{,}79}{0{,}092} \times \dfrac{0{,}908}{0{,}21}$ □ **37**.

5. On a $q_F' = \dfrac{0{,}922 \times r'}{0{,}318 \times p'}$ et $q_D' = \dfrac{0{,}078 \times r'}{0{,}682 \times p'}$. D'où $t' = \dfrac{0{,}922}{0{,}318} \times \dfrac{0{,}682}{0{,}078}$ □ **25**.

Pour ce qui concerne Polytechnique, on peut considérer que l'ascenseur social s'est (un peu) amélioré.

6. On peut entrer en B3 la formule =1-B2 puis la recopier vers la droite en C3.

On peut entrer en B10 la formule =(B3/B7)*(B6/B2) puis la recopier vers la droite en C10.

On obtient les résultats suivants (arrondis à l'unité).

	t	t'
ENA (École nationale d'administration)	44	33
ENS (Écoles normales supérieures)	31	33
HEC (Hautes études commerciales)	16	16
Grandes écoles (avec entrée sur concours)	24	23

On peut considérer globalement que les inégalités sociales de recrutement de ces écoles sont importantes et ont peu évolué entre 1950 et 1990.
Il faut cependant préciser qu'il ne s'agit ici que des grandes écoles, le cas du supérieur en général est différent, en particulier avec la création des BTS et des IUT.

2 – Proportions et risque : tabac et infarctus

 La campagne de la Fédération française de cardiologie du printemps 2007affirme que « 80% des victimes d'infarctus sont des fumeurs ». L'information montre-t-elle que le risque d'infarctus est plus important pour un fumeur ?

Énoncé (niveau 3ème – 2nde)

La Fédération française de cardiologie affiche l'information ci-contre dans une campagne de presse.

1. Parmi les victimes d'infarctus ayant moins de 45 ans, quelle est la proportion de non-fumeurs ?
2. Pourquoi l'information donnée permet de penser que fumer augmente le risque d'infarctus ?
3. On peut estimer qu'en France, parmi les moins de 45 ans, il y a environ 40% de fumeurs.
a) On note n le nombre de cas d'infarctus observés chez les moins de 45 ans.
Exprimer en fonction de n le nombre d'infarctus parmi les fumeurs.
b) On note p le nombre de personnes de moins de 45 ans.
Montrer que proportion q_F d'infarctus parmi les fumeurs est $q_F = \dfrac{0{,}80 \times n}{0{,}40 \times p}$.
4. Donner, de même, l'expression de la proportion $q_{\overline{F}}$ d'infarctus parmi les non-fumeurs.

5. Montrer que $\dfrac{q_F}{q_{\overline{F}}} \approx 6$.

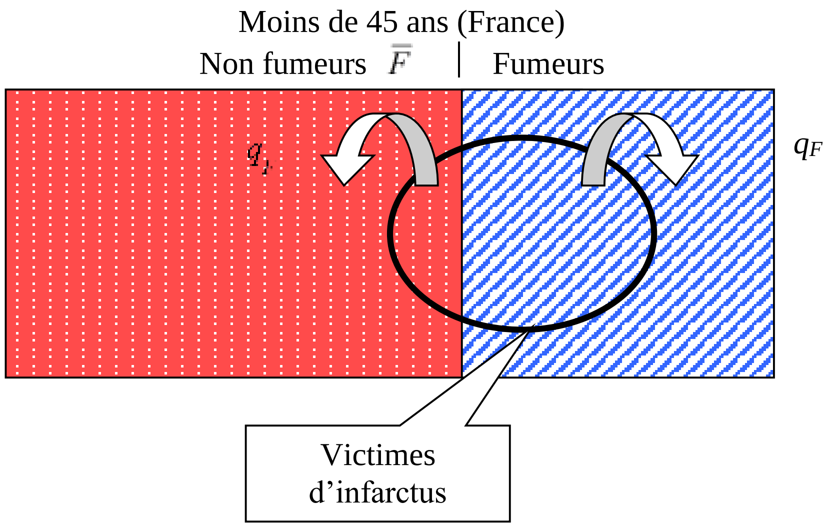
On peut interpréter ce résultat en disant que pour les moins de 45 ans, un fumeur a 6 fois plus de risques d'avoir un infarctus qu'un non-fumeur.

Sources :

Fédération française de cardiologie – www.tabac.gouv.fr – INSEE.

Pour la proportion des fumeurs en France, les estimations sont variables : aux alentours de 34% de la population. Le site tabac.gouv.fr affirme, à partir des données de l'INSEE, que la proportion de fumeurs dans la classe d'âge 18-25 ans était de 48% en 2000, 40% en 2003 et est remontée en 2005-2006.

Éléments de réponse



1. 20%.
2. On sait qu'il y a moins de 80% de fumeurs. On en déduit que les fumeurs sont sur-représentés parmi les cas d'infarctus.
3. $q_F = \dfrac{0{,}8 \times n}{0{,}4 \times p}$.
4. $q_{\overline{F}} = \dfrac{0{,}2 \times n}{0{,}6 \times p}$.
5. $\dfrac{q_F}{q_{\overline{F}}} = \dfrac{0{,}8 \times 0{,}6}{0{,}4 \times 0{,}2} = 6$.

2 – Inquiétudes à Woburn : jusqu'où croire au « hasard » ?

□ Une petite ville des États-Unis a connu 9 cas de leucémies chez de jeunes garçons en l'espace de 10 années. Doit-on en « accuser le hasard » ?

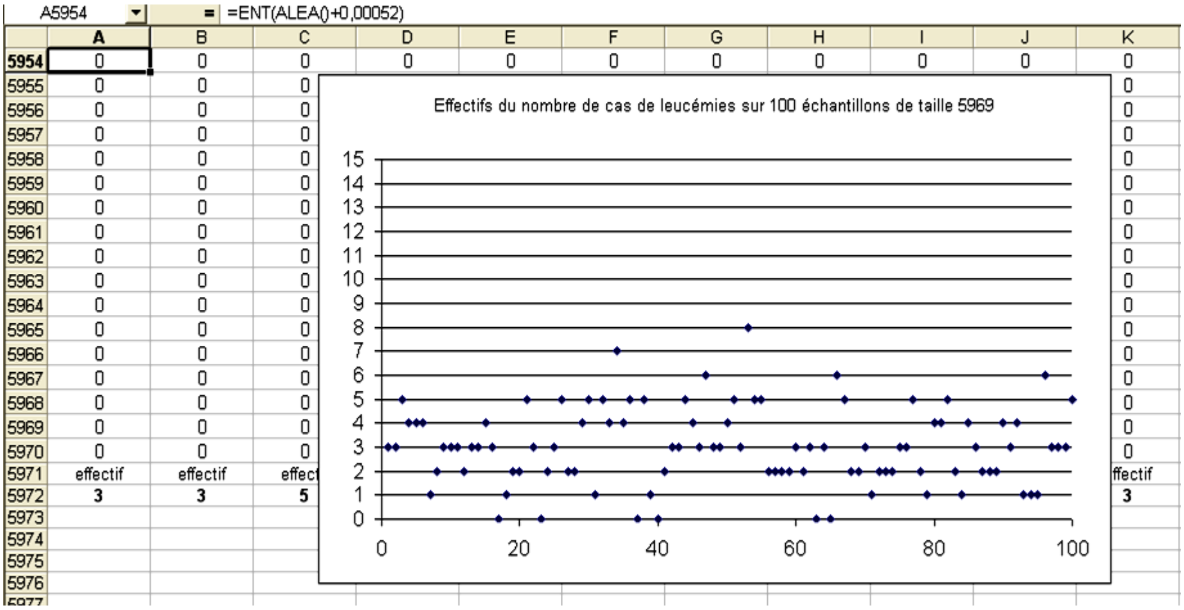
Cet exemple montre l'importance des enjeux de la méthode statistique.

Woburn est une petite ville industrielle du Massachusetts, au Nord-Est des Etats-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. Les familles se lancent alors dans l'exploration des causes et constatent la présence de décharges et de friches industrielles ainsi que l'existence de polluants. Dans un premier temps, les experts gouvernementaux concluent qu'il n'y a rien d'étrange. Mais les familles s'obstinent et saisissent leurs propres experts. Une étude statistique montre qu'il se passe sans doute quelque chose « d'étrange ».

Le tableau suivant résume les données statistiques concernant les garçons de moins de 15 ans, pour la période 1969-1979 (Source : *Massachusetts Department of Public Health*).

Population des garçons de moins de 15 ans à Woburn selon le recensement de 1970 : n	Nombre de cas de leucémie infantile observés à Woburn entre 1969 et 1979	Fréquence des leucémies aux Etats-Unis p
5969	9	0,00052

La question statistique qui se pose est de savoir si le hasard seul peut raisonnablement expliquer le nombre de leucémies observées chez les jeunes garçons de Woburn, considérés comme résultant d'un échantillon prélevé dans la population américaine.
La population des Etats-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.
On peut aisément simuler sur le tableur 100 échantillons de taille n = 5969 prélevés au hasard dans une population où p = 0,00052 en utilisant l'instruction :
=ENT(ALEA()+0,00052) .
Sur chaque échantillon, en faisant la somme, on obtient le nombre de cas observés.
On peut représenter sur un graphique les 100 résultats observés sur les échantillons simulés.



Les simulations montrent que le nombre de cas observés à Woburn (9 cas) est extrêmement rare (moins de 1 % des simulations), si l'on ne considère que le hasard comme explication.
On ne peut donc pas raisonnablement attribuer au seul hasard le niveau très « significativement » élevé des leucémies infantiles observées chez les garçons de Woburn.
Ce taux anormalement élevé de leucémies 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 ainsi le syndrome du trichloréthylène.

3 – Justice et discrimination : « preuve » statistique

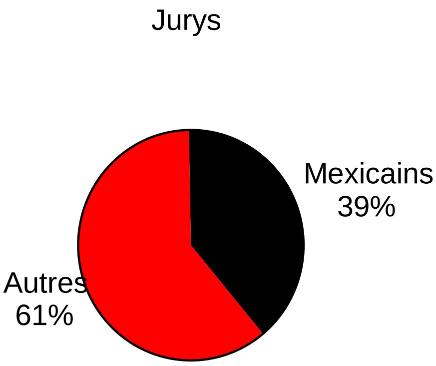
 Comment aux États-Unis, la statistique a-t-elle permis de mettre en évidence une discrimination conduisant à casser un jugement ?

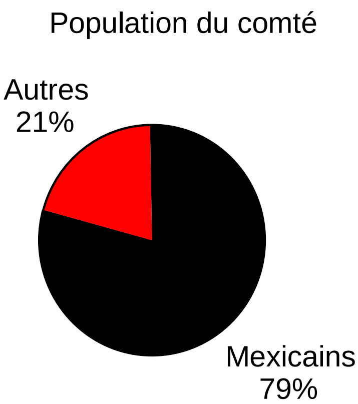
En fin de chapitre de statistique, la situation suivante a été proposée à des élèves de seconde. Après une recherche en salle informatique (50 minutes), un travail d’argumentation est demandé sous forme de devoir à la maison.
La grande liberté donnée dans les consignes (voir l’énoncé ci-dessous) a quelque peu déstabilisé les élèves (« Qu’est-ce qu’on doit faire ? ») mais a favorisé une grande diversité des réponses et l’utilisation d’un langage personnel. La correction a associé la professeure de français qui, après avoir visé le travail des élèves, a pu faire le parallèle avec les techniques d’argumentation en littérature (la seule différence véritable étant qu’ici les exemples sont des graphiques et des calculs au lieu de citations d’un texte).

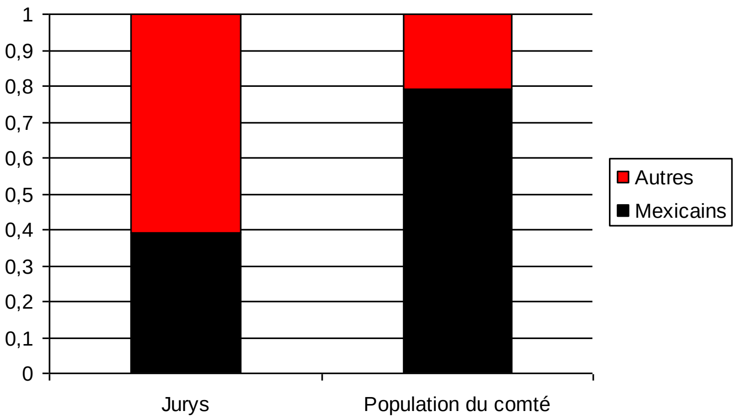
Énoncé distribué aux élèves

L’affaire Castaneda contre Partida.
En Novembre 1976 dans un comté du sud du Texas, Rodrigo Partida était condamné à huit ans de prison.
Il attaqua ce jugement au motif que la désignation des jurés de ce comté était discriminante à l’égard des Américains d’origine mexicaine. Alors que 79,1% de la population du comté était d’origine mexicaine, sur les 870 personnes convoqués pour être jurés lors des 11 années précédentes, il n’y eût que 339 personnes d’origine mexicaine.
Produisez votre expertise statistique.
Devant la Cour Suprême, un expert statisticien produisit des arguments pour convaincre du bien fondé de la requête de l’accusé (les juges votèrent à 5 contre 4 en faveur de la requête).
En vous situant dans le rôle de cet expert, produisez à votre tour des calculs, des raisonnements, des graphiques... pour montrer que le hasard ne peut pas « raisonnablement » expliquer à lui seul la sous-représentation des américains d’origine mexicaine dans les jurys de ce comté.
 Vous commencez ce travail en binômes en utilisant les documents disponibles, la calculatrice, le tableur.
 Vous terminez la rédaction (arguments en français, calculs, graphiques...) en devoir individuel, à la maison.

Une première partie du travail sur tableur a consisté en une analyse descriptive des données conduisant les uns ou les autres à produire des tableaux croisés, des histogrammes ou des camemberts.



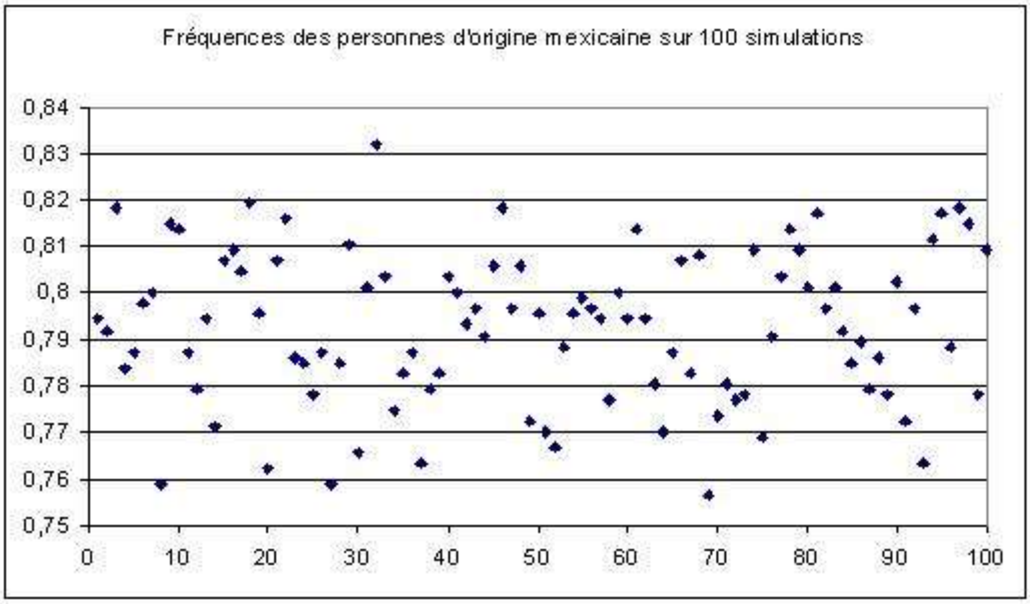




La seconde partie du travail devait consister à utiliser les moyens informatiques pour montrer que l'écart (important) observé ne peut raisonnablement pas s'expliquer par le seul hasard.
Pour cela, on peut simuler 870 tirages au sort dans la population du comté en utilisant la formule =ENT(ALEA()+0,791) . Certains ont ensuite raisonnés en effectifs, d'autres en fréquences de personnes d'origine mexicaine.
On peut vérifier que plus de 95 % des simulations fournissent une fréquence de personnes d'origine mexicaine comprise dans l'intervalle :

$[\, p - \frac{1}{\sqrt{n}} ,\ p + \frac{1}{\sqrt{n}} \,]$ c'est-à-dire $[0{,}791 - \frac{1}{\sqrt{870}}\ ;\ 0{,}791 + \frac{1}{\sqrt{870}}\,]$

c'est-à-dire environ [0,76 ; 0,82].
Voici un exemple de 100 simulations en fréquences des personnes d'origines mexicaines sur les échantillons de taille 870 :



La fréquence observée des personnes d'origine mexicaine dans les jurys est 0,39 qui est très éloignée de l'intervalle [0,76 ; 0,82]. Jamais les simulations n'ont permis d'observer un résultat aussi bas. Ceci permet de dire que le hasard n'est sans doute pas responsable de la sous représentation des personnes d'origine mexicaine dans les jurys.
Reste à rechercher les causes de cette sous représentation (l'une des raisons est la nécessité d'une bonne connaissance de l'anglais écrit et parlé).

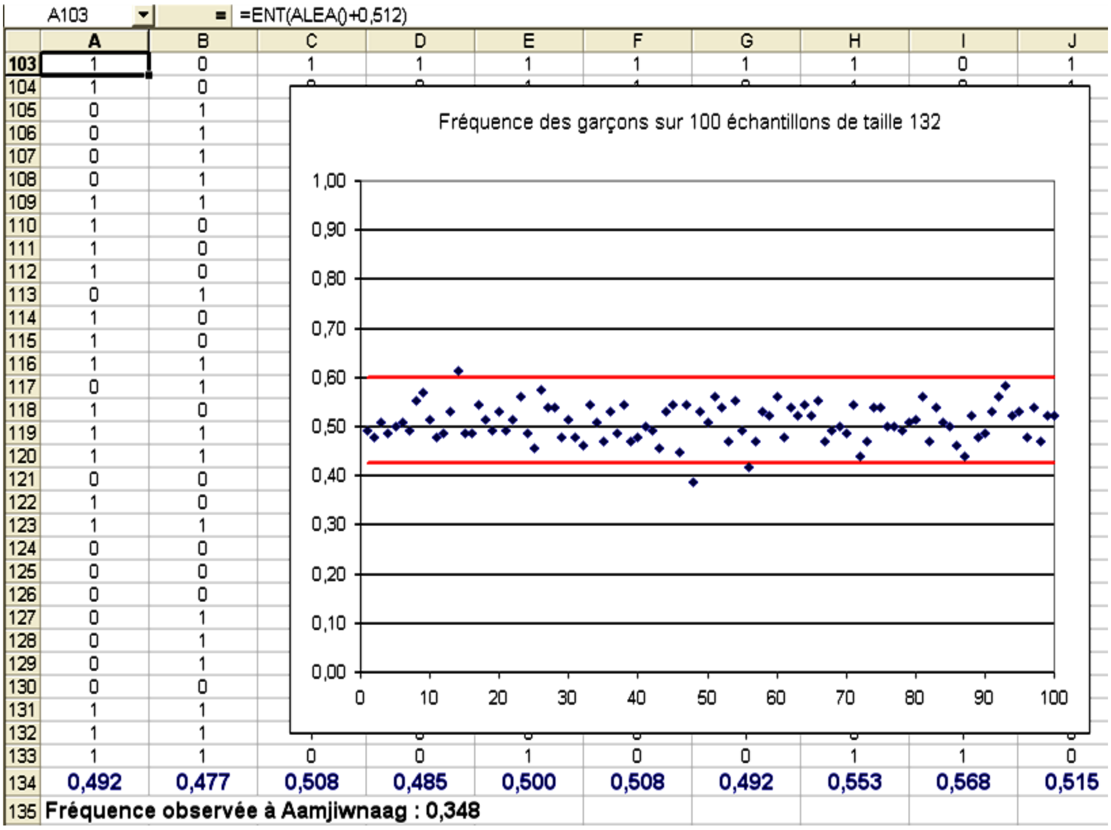
4 – Pollution et sex-ratio : « alerte » statistique

 Dans un village du Canada, entre 1999 et 2003, sur 132 naissances, il n'y a eu que 46 garçons. Est-ce très étonnant ? Est-ce inquiétant (sachant que de nombreuses usines chimiques se trouvent à proximité) ?

Le « sex-ratio » est le rapport du nombre de garçons à celui des filles à la naissance. Sa remarquable stabilité est une des premières découvertes de la statistique. Il est habituellement de 105 garçons pour 100 filles, soit une fréquence de garçons de

$p = \dfrac{105}{105+100}$ 0,512.

Dans la réserve indienne d'Aamjiwnaag, située au Canada, il est né entre 1999 et 2003, 132 enfants dont 46 garçons (Sources : *Science et Vie* février 2006 – *Environnemental Health Perspectives* octobre 2005 – article en ligne). Cela donne une fréquence observée des garçons valant $f = \dfrac{46}{132}$ 0,348. C'est bien peu, mais dans l'absolu cela ne veut rien dire si l'on n'a pas étudié les fluctuations des échantillons de taille 132 sous l'hypothèse d'une fréquence « normale » valant $p = 0{,}512$.



Cette étude sur le tableur sera l'occasion d'expérimenter la « loi » selon laquelle, si la taille des échantillons n'est pas trop petite et la fréquence p pas trop faible plus de 95 % des échantillons fournissent une fréquence comprise dans l'intervalle $[0{,}512 - \frac{1}{\sqrt{n}}\ ;\ 0{,}512 + \frac{1}{\sqrt{n}}]$ c'est-à-dire [0,424 ; 0,599], qui correspond, en quelque sorte, à l'intervalle de « variabilité naturelle » des échantillons de taille $n = 132$.

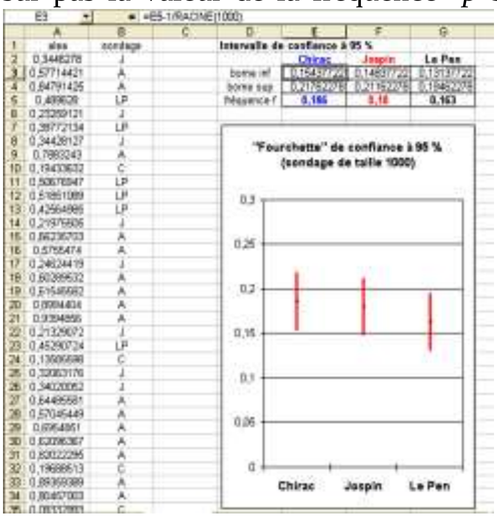
On constate que la fréquence des garçons observée, soit 0,348 , n'appartient pas à l'intervalle de fluctuation de plus de 95 % des échantillon sous l'hypothèse $p = 0{,}512$. On dit que la fréquence observée présente une « différence significative » au niveau 0,95.

A 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. 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 permettra d'avoir d'autres éléments sur ce dossier (qui a fait polémique au Canada).

5 – Sondages

□ Le dernier sondage de 2002 ne prévoyait pas la présence de Jean-Marie Le Pen au second tour.
Le dernier sondage CSA d'avril 2007 plaçait J-M Le Pen devant F. Bayrou.
Les sondages sont-il scientifiques ?

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.
On peut mettre en parallèle le dernier sondage publié par BVA et effectué sur 1000 électeurs :
Jacques Chirac **19 %**, Lionel Jospin **18 %**, Jean-Marie Le Pen **14 %**,
avec le résultat du premier tour :
Jacques Chirac **19,88 %**, Lionel Jospin **16,18 %**, Jean-Marie Le Pen **16,86 %**,
et poser la question « le sondage est-il faux ? ».
On a vu que si la fréquence dans la population est p, plus de 95 % des échantillons aléatoires de taille n fournissent une fréquence dans l'intervalle $[\, p - \frac{1}{\sqrt{n}} \,,\ p + \frac{1}{\sqrt{n}} \,]$.
Quand on effectue un sondage, on ne connaît bien sûr pas la valeur de la fréquence p sur la population, on ne connaît que la fréquence f obtenue sur l'échantillon. On peut alors calculer la fourchette $[\, f - \frac{1}{\sqrt{n}} \,,\ f + \frac{1}{\sqrt{n}} \,]$ et affirmer que ce procédé est fiable à plus de 95 % c'est-à-dire que plus de 95 % des fourchettes ainsi calculées recouvrent effectivement la valeur p que l'on veut estimer.
Il est fort instructif d'expérimenter ces fourchettes dans la situation du premier tour de l'élection de 2002.
On peut représenter les fourchettes des trois candidats en utilisant le « diagramme boursier » du tableur. Il faut pour cela avoir, dans cet ordre, la borne inférieure, la borne supérieure et le centre de la fourchette.



Sur l'image d'écran ci-contre, on a d'abord entré en E5 la formule
=NB.SI(B2:B1001;"C")/1000 puis en E3 la formule
=E5-1/RACINE(1000) et en E4 la formule =E5+1/RACINE(1000) .

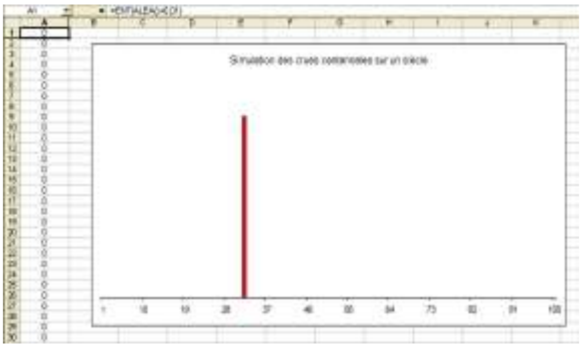
Le sondage suivant, effectué avant le premier tour de 2007, pose davantage de questions, en particulier quant aux méthodes de « redressement » utilisées.
Dernier sondage C.S.A. , effectué sur 1002 électeurs le vendredi 20/04/07 :
François Bayrou 16 % – Jean-Marie Le Pen 16,5 % .
Résultats du premier tour le dimanche 22/04/07 :
François Bayrou 18,57 % – Jean-Marie Le Pen 10,44 % .

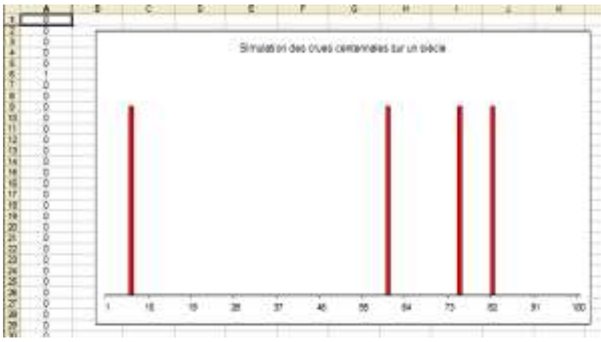
6 – Crues et notion de probabilité

 Le magazine *Le Point* du 13/09/2002 écrit : « Une crue de la Seine 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 ». Que faut-il en penser ?

Dans une revue technique de juin 2003[17], 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* ». Lorsqu'on parle de crue centennale, cela signifie que chaque année elle a une chance sur 100 de se produire (indépendamment des années précédentes). Il faut se rendre compte que sur l'échelle d'un siècle, cela peut donner des situations très variables.

Il est utile au citoyen d'expérimenter les fluctuations d'échantillons de taille 100 (un siècle). Ce qui est très facile sur le tableur (ou une calculatrice).
On peut ensuite reconsidérer la déclaration du journaliste du *Point*.





[17] 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.

Bibliographie succincte

- CHAPUS (Brigitte) & HENRY (Michel) éditeurs – Commission Inter-Irem Statistique et probabilités - APMEP – Brochure n° 156 – *Statistique au lycée* - volume 1 et 2 (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.

- SCHWARTZ (Claudine) – *Pratiques de la statistique* – Vuibert 2006.

Internet

- www-irem.univ-paris13.fr/

  A partir du site de l'IREM de Paris-Nord, on peut accéder aux pages du groupe « statistique et citoyenneté ».
  Ou, plus directement,
  http://dutarte.club.fr/Sitestat/default.htm
- www.statistix.fr/

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