L'ENSEIGNEMENT DE L'AFFAIRE WOBURN, SUITE
SUSCITER L'INTÉRÊT SANS TROMPER

Vincent COUALLIER[1], Léo GERVILLE-RÉACHE[2] et Gilles STOLTZ[3]

TITLE

Teaching the Woburn case, continued: raising interest without deceiving

RÉSUMÉ

L'affaire Woburn est une controverse de statistique et santé publique qui a donné lieu à la rédaction d'exercices au niveau de la classe de première scientifique du lycée. Nous défendions dans une communication orale, critiquée par Jeanne Fine dans le numéro précédent de cette revue, et défendons dans le présent libre propos le point de vue suivant. Montrer l'éclairage qu'apporte la statistique aux débats de société est une perspective séduisante. Dans le cas d'espèce, les énoncés d'exercices proposés par diverses sources officielles ou académiques souffrent toutefois de simplifications à outrance, qui donnent l'illusion qu'un calcul rapide à portée d'élèves de lycée peut permettre d'éclairer une affaire pourtant complexe et qui a suscité un débat nourri dans la communauté académique. Nous recommandons un nouvel énoncé pour s'intéresser à cette affaire, qui est formulé de manière plus prudente ; cependant, notre réelle recommandation est de remplacer toute résolution d'exercice par la lecture critique d'un document écrit par des professionnels de la statistique, replaçant notamment le court calcul probabiliste nécessaire pour obtenir la *P*-valeur au sein de considérations rigoureuses sur la méthodologie et la démarche statistiques.

Mots-clés : *statistique au lycée, données réelles, risques méthodologiques, niveau de preuve d'une étude.*

ABSTRACT

The Woburn case is a public-health and statistical controversy that inspired problem statements for teaching statistics at the pre-university level in France. Our point is the following; we developed it in an oral communication (criticized by Jeanne Fine in the previous volume of this journal) and will develop it further in the present article. It is of course desirable to show how useful statistics can be for society. But in the present case, the problem statements that we could find oversimplify the reality of a thorough statistical analysis and are misleading. Indeed, they seem to suggest that some short calculations understandable by pre-university students can shed light on a complex case that gave rise to an important controversy in the academic community. We thus recommend and provide a more cautious problem statement. However, our actual recommendation would be to go from a problem-solving approach to reading a document (written by professional statisticians) with a critical eye. The latter document would detail the short calculations leading to a *P*-value but most importantly it would frame them in a rigorous statistical agenda, following in particular standard methodological guidelines.

Keywords: *teaching statistics at a pre-university level, real data, methodological errors, levels of evidence in a study.*

Nous revenons dans cet article sur la communication orale effectuée par deux d'entre nous (Gerville-Réache et Couallier, 2014). Ainsi que le montre le libre propos publié par Jeanne Fine

[1]Université de Bordeaux, IMB UMR CNRS 5251, vincent.couallier@u-bordeaux2.fr
[2]Université de Bordeaux, IMB UMR CNRS 5251, leo.gerville-reache@u-bordeaux.fr
[3]HEC Paris, CNRS, Université Paris-Saclay, Jouy-en-Josas, stoltz@hec.fr

(2015) dans le numéro précédent de *Statistique et Enseignement*, certains de nos arguments n'avaient pas été développés de manière suffisante dans le résumé écrit afférent à la communication orale. Ce dernier commençait ainsi :

> Dans cette communication, nous revenons sur deux exercices emblématiques de l'enseignement de la statistique au lycée. Les affaires Woburn et Castaneda sont deux exemples où la justice a usé de la statistique pour se prononcer. Présente dans bons nombres de livres d'élèves de première et terminale, l'analyse de ces affaires utilise la simulation numérique et les intervalles de fluctuation pour traiter de la question. Malheureusement, les modélisations et les questions qu'elles posent sont plus complexes qu'il n'y paraît.

Le présent libre propos est l'occasion de préciser notre message, qui ne rejette bien évidemment pas la pertinence de l'illustration de la statistique sur des données réelles en général, mais veut mettre en garde sur celle-ci lorsqu'elle est mise en œuvre au niveau du lycée. En outre, au lieu de simplement soulever des difficultés comme dans la communication orale originelle, nous allons nous atteler à les résoudre, en proposant un énoncé d'exercice nous semblant à la fois rigoureux et accessible à ce niveau.

En effet, pour séduisante que soit la perspective de l'enseignement d'outils statistiques comme éléments de réponse à des questions de santé publique, de justice, de psychologie, il ne faut pas négliger l'existence de risques méthodologiques et de mécompréhension rédhibitoires, difficiles à éradiquer autrement que par des considérations au-delà des compétences requises pour un élève de lycée. Les solutions que nous proposons sont au moins que l'énoncé du problème guide les élèves dans la résolution du problème selon une méthodologie rigoureuse (formulation des hypothèses et indication du mode de recueil des données avant la collecte et le traitement de ces dernières), et idéalement, que l'on remplace ces exercices par des lectures critiques de documents. Ces derniers seraient rédigés par des professionnels, mettraient en œuvre de manière correcte l'intégralité de la démarche statistique sur le cas considéré, et seraient assortis de retours et mises en garde sur la démarche suivie et ses écueils.

Par souci de clarté, nous nous concentrons sur les données de « l'affaire Woburn ». Celle-ci nous semble exemplaire des difficultés liées à l'usage de la statistique : comment constituer un dossier prouvant un problème de santé publique (lié à une pollution chimique de l'eau dite potable) *après* que la population signale des faits inhabituels (un taux inhabituellement élevé de cancers dans la population). De nombreuses polémiques entre statisticiens, publiées au long des années 1980 dans le prestigieux *Journal of the American Statistical Association*, témoignent de la difficulté de prouver rigoureusement dans ce contexte qu'il existe un problème de santé publique, dont la cause serait l'eau potable contaminée. Les questions que nous posions dans notre communication orale, à savoir

> Pourquoi Woburn ? Pourquoi la leucémie ? Pourquoi les garçons de moins de 15 ans ?

n'étaient que des raccourcis incitant à réfléchir aux difficultés méthodologiques rencontrées dans l'usage de la statistique inférentielle comme outil d'aide à la décision. Pour faire bref, les données réelles sans leurs méta-données (décrivant par qui, quand, et comment les données ont été recueillies, dans quel but et pour mettre à l'épreuve quelles hypothèses) ne valent rien. Ce n'est pas nous qui le disons, c'est Sir Ronald Fisher (1938) lui-même :

> *To consult the statistician after an experiment is finished is often merely to ask him to conduct a post-mortem examination. He can perhaps say what the experiment died of.* (Traduction libre : Appeler un statisticien après que l'expérience est terminée revient souvent à lui demander de mener une autopsie ; il pourra peut-être déterminer la cause de l'échec de l'expérience.)

Plan de notre propos — Selon nous, un énoncé d'exercice doit donc toujours procurer ces méta-données ; or, nous verrons à la partie 1 que ce n'est pas le cas pour ceux utilisés à notre connaissance dans l'enseignement secondaire pour traiter de l'affaire Woburn. La partie 2 détaillera ainsi l'écueil premier de tous ces énoncés : à les lire, l'analyse menée peut sembler issue d'un choix rétrospectif de répartition en sous-échantillons *ad hoc*. Elle reviendra également sur la nécessité de la comparabilité de ces sous-échantillons. La partie 3 embrassera un point de vue plus large et reprendra l'historique des débats formulés dans la communauté scientifique dans les années 1980 ; nous verrons également que selon la classification de la Haute autorité de santé de la France, les énoncés de la partie 1 correspondraient à des études de terrain ayant le plus bas niveau de qualité méthodologique. La partie 4 reviendra sur le danger et les limites des études par simulations recommandées par de nombreux énoncés de la partie 1.

Finalement, nous formulerons à la partie 5 une proposition d'énoncé d'exercice nous semblant méthodologiquement rigoureuse tout en respectant l'esprit (mais pas la lettre) de l'affaire Woburn. Nous recommanderons toutefois de faire lire aux élèves de manière critique un document mettant en œuvre une démarche statistique rigoureuse, plutôt que de la leur faire réaliser eux-mêmes. Enfin, une annexe dépassera le cas de Woburn et fournira des liens vers d'autres mises en garde sur le mauvais usage possible de la statistique inférentielle, montrant que bien au-delà du lycée, des erreurs méthodologiques sont possibles et même fréquentes.

Notre démarche est constructive et incite à faire réfléchir les élèves sur les questions de méthodologie statistique, qui ne nécessitent pas de connaissances mathématiques des méthodes statistiques au-delà de celles déjà aux programmes. En ce sens, nous armerons ainsi de futurs citoyens face aux chiffres et leur permettrons d'acquérir un sens critique vis-à-vis de la masse d'information reçue.

1. L'affaire Woburn (mal) résumée par des exercices

Nous commençons par rappeler quelques énoncés de l'affaire Woburn. Jeanne Fine (2015) s'appuie sur une version courte et condensée présentée par exemple par un document de *Ressources pour la classe de première générale et technologique* (Direction générale de l'enseignement scolaire, 2012, page 56) :

> Une petite ville des États-Unis, Woburn, a connu 9 cas de leucémie parmi les 5 969 garçons de moins de 15 ans sur la période 1969–1979. La fréquence des leucémies pour cette tranche d'âge aux États-Unis est égale à 0,00052 (source : *Massachussetts Department of Public Health*). Les autorités concluent qu'il n'y a rien d'étrange dans cette ville. Qu'en pensez-vous ?

Philippe Dutarte (2007) avait proposé une version plus riche, tant dans la description du contexte de l'affaire que dans celle des données relevées (elle précisait notamment les prévalences nationales de leucémies à la fois pour les garçons et les filles) :

> 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 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 recencement 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 Etats-Unis p
Garçons	5969	9	0,00151	0,00052
Filles	5779	3	0,00052	0,00038
Total	11748	12	0,00102	0,00045

> 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 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. [...]

Suit alors une étude par simulations pour le cas des garçons, permettant de s'intéresser (sans les nommer) aux P–valeurs issues d'une approche modèle ou d'une approche sondage, comme l'explique et le détaille Jeanne Fine (2015). Une troisième version, intermédiaire, est proposée par Fabrice Barache *et al.* (2011, page 247) :

> Dans les années 1970, les habitants de Woburn, petite ville industrielle au Nord-Est des Etats-Unis, surtout les garçons dans certains quartiers de la ville, sont touchés par un grand nombre de leucémies infantiles. Les familles accusent la présence de décharges et de polluants. Les autorités accusent le hasard. [...]

L'énoncé se poursuit avec l'indication des données pour les garçons uniquement et une étude par simulations.

2. Critique de l'étude intra-échantillon

Rappel de la bonne méthodologie statistique — Un point essentiel de la méthodologie statistique est qu'il faut formuler les hypothèses testées et le protocole de recueil des données avant

de mettre en œuvre ce dernier. En particulier, il est incorrect et parfois même manipulateur de choisir les hypothèses au vu des données. Un mal mineur est que cela entraînerait en général une division des *P*-valeurs par deux dans des cas où il s'agirait de déterminer si des déviations unilatérales (à la hausse ou à la baisse) sont significatives, alors même qu'*a priori*, avant tout recueil de données, la situation et le contexte n'amenaient pas à considérer uniquement des déviations dans un sens. Un mal plus grand consisterait à noter que tel ou tel sous-échantillon des données recueillies exhiberait des déviations significatives, alors que l'échantillon dans son intégralité ne présente pas de telles déviations. En effet, il est souvent possible de construire un tel sous-échantillon au vu des données ! Jeanne Fine (2015) souligne également ce point, mais sans concéder qu'il s'applique à l'affaire Woburn :

> [...] c'est à l'avance qu'il faut déterminer les tests statistiques qui vont être effectués. Il est incorrect, après observation des données, de faire des tests sur des données jugées *a posteriori* « aberrantes ».

Ainsi, en général, les sous-échantillons (appelés aussi sous-groupes) devraient être définis *a priori*[4], en fonction de connaissances extérieures et préalables à l'expérience statistique (issues dans le cas d'espèce de mécanismes biologiques bien documentés ou de précédents résultats statistiques, mais il n'y en avait pas). Il faut en revanche éviter de définir un sous-groupe à partir de caractéristiques mesurées au cours de l'essai, comme nous l'avons déjà souligné. En conclusion, il vaut donc mieux considérer les résultats des analyses en sous-groupes comme des hypothèses de travail pour des études futures plutôt que comme des preuves déjà établies.

Cette bonne méthodologie a-t-elle été mise en œuvre ? — Dans l'exemple de l'affaire Woburn, des données ont été recueillies et analysées, mais trop peu est dit dans les énoncés fournis ci-dessus sur le contexte de cette collecte, et sur les hypothèses qu'il était raisonnable d'avoir présentes à l'esprit avant de regarder les données du terrain. En particulier, pourquoi séparer garçons et filles, pourquoi se restreindre à une tranche d'âge 0–14 ans ? Ces catégorisations avaient-elles bien lieu d'être au préalable ?

Si une raison physiologique bien documentée permet d'affirmer que les leucémies touchent plus les garçons que les filles, il faut le préciser explicitement. Il semble que ce soit peut-être le cas au vu des prévalences nationales de leucémies indiquées dans le tableau de l'énoncé de Philippe Dutarte (2007), respectivement de 0,00052 pour les garçons et 0,00038 pour les filles. Cela justifierait alors la séparation en deux groupes.

De la même manière, comme un énoncé le suggère, « certains quartiers », ceux de l'Est, sont pensés par les habitants comme sources du phénomène. On peut alors en tenir compte, à condition de suivre une méthodologie y correspondant : collecter les données une fois ce découpage en sous-groupes bien défini géographiquement. Dans ce cas-là, il est légitime de

[4] Voir par exemple cette citation extraite d'un document de méthodologie édicté par la Conférence internationale sur l'harmonisation des exigences techniques pour l'enregistrement des médicaments à usage humain (*International conference on harmonisation of technical requirements for registration of pharmaceuticals for human use*, ICH, 1995) : "When extensive statistical analyses have been performed by the applicant, it is essential to consider the extent to which the analyses were planned prior to the availability of data and, if they were not, how bias was avoided in choosing the particular analysis used as a basis for conclusions. This is particularly important in the case of any subgroup analyses, because if such analyses are not preplanned they will ordinarily not provide an adequate basis for definitive conclusions."

considérer le sous-groupe des enfants vivant dans cette partie de la ville et buvant l'eau du robinet. Le rapport originel des autorités sanitaires (Parker et Rosen, 1981) suit effectivement cette voie, bien que la polémique vienne notamment de la possible confusion entre un facteur géographique et un facteur de consommation d'eau potable provenant de puits pollués. Il est à noter que ce même rapport sépare également garçons et filles.

En l'occurrence, la pollution à Worburn a bien été établie comme celle de deux puits desservant l'Est de la ville, l'objectif étant alors de prouver un sur-risque de cancer pour les consommateurs de l'eau polluée, indépendamment de la localisation géographique.

De la comparabilité des sous-groupes ainsi constitués — Nous précisons qu'en particulier, dans ce rapport, les arguments et controverses médicaux et statistiques d'attribution à la pollution de la différence de prévalence entre la zone Est et le reste de la ville de Woburn se sont intéressés à la comparabilité des groupes d'enfants malades et sains : il s'agit de montrer que ces deux sous-groupes ne se distinguaient que par leur lieu de vie et non par d'autres caractéristiques qui ont une influence sur le développement ou non d'une leucémie infantile. Ainsi, il s'agissait de mettre en évidence dans un premier temps une association entre une prévalence supérieure de leucémies infantiles et le lieu de vie, et peut-être établir avec des études complémentaires l'impact de l'exposition à l'eau polluée (variable qui n'était pas disponible dans les premières analyses).

Cela répond à une critique de Jeanne Fine (2015), qui ne voyait pas où la comparabilité de groupes pouvait intervenir dans un test de comparaison à une valeur de référence : pas dans le test lui-même, donc, mais dans les conséquences qu'on voudrait en tirer lorsque de tels tests sont mis en œuvre. Le lecteur intéressé par davantage de détails sur ce sujet pourra lire le rapport d'étape des autorités (*Massachusetts Department of Public Health*, 1997) ou se plonger dans le cas rédigé par Bair et Svitana (2008) et ses documents annexes (encore plus de rapports !).

De la difficulté à traiter des données réelles en classe — Ainsi, la question « Pourquoi les garçons de moins de 15 ans ? » originellement posée dans la communication orale de Gerville-Réache et Couallier (2014) résume notre sentiment : introduire les élèves (et leurs professeurs) à une démarche statistique méthodologiquement rigoureuse est délicat et nécessite du temps. Ici en l'occurrence, il faut justifier soigneusement et au préalable les sous-groupes à constituer : filles-garçons ou Est-reste de la ville, ce qu'aucun énoncé de la partie 1 ne fait assez explicitement (seul l'énoncé de Philippe Dutarte, 2007, le fait implicitement pour les groupes filles-garçons).

Nous proposons en conclusion de ce libre propos un énoncé selon nous plus satisfaisant que ceux proposés en ouverture car, à l'aide d'éléments de contexte, il amène justement à formuler des hypothèses précises et des découpages en groupes (fondés sur la géographie et non le sexe, pour changer, mais peu importe), hypothèses qu'il s'agit ensuite d'invalider ou de conserver au vu de données explicitement recueillies dans un second temps. Cela évite l'écueil de formuler des hypothèses à tester à partir des données, ce qui est une tentation naturelle mais coupable. Cet énoncé se conclut par l'indication de démarches statistiques ultérieures à effectuer : vérification de la comparabilité des groupes, renforcement d'une association entre deux variables en la preuve d'un lien de causalité.

Toutefois, au vu d'un programme de mathématiques déjà chargé, contenant en particulier l'étude de fondements du calcul des probabilités et de la notion d'intervalle de confiance, on

peut légitimement s'interroger s'il convient d'aller plus loin que ce programme officiel et s'intéresser, même implicitement, aux tests d'hypothèses et à la notion de *P*-valeur.

Il est à noter que même la résolution par intervalles de confiance du caractère significatif ou non d'une déviation est délicat, car il faut choisir la forme (unilatère ou bilatère) des intervalles en jeu, et cela ne doit se faire qu'*a priori* et au vu des éléments de contexte, pas en fonction des données à traiter.

3. Controverses scientifiques et formulations prudentes

Nous donnons dans cette partie un bref aperçu des controverses scientifiques ayant trait à l'affaire Woburn. Elles sont à garder en toile de fond pour éviter l'écriture d'énoncés trop péremptoires, comme celui de la Direction générale de l'enseignement scolaire (2012), et pour, au contraire, recourir à des formulations prudentes et montrant qu'un débat est lancé au vu des données disponibles. Le contenu de cette partie (et de la suivante) ne s'adresse pas directement ni aux élèves du lycée ni même à leurs enseignants, mais plutôt aux rédacteurs d'énoncés d'exercices destinés à être utilisés par eux.

Gradation du niveau de preuve d'une étude statistique — Nous commençons par placer la discussion dans un contexte général, relatif au niveau de preuve : comment assurer et garantir suffisamment de rigueur scientifique à une étude statistique pour que la découverte qui en découle puisse avoir le statut de preuve scientifique, au moins dans la science expérimentale ? (Dans notre cas, les découvertes putatives sont l'effet toxique d'un polluant dans l'eau potable, ou encore la sur-prévalence de leucémie dans une ville donnée pendant une période de 10 ans.)

Il existe de nombreux biais et problèmes méthodologiques ou mathématiques bien connus qui mettent à mal cette tentative de rigueur, ce qui a conduit les chercheurs et responsables scientifiques à tenter de normaliser la situation en définissant soit des recommandations de bonnes pratiques (détaillées ci-dessous), soit des incitations à changer les politiques scientifiques et de publication des éditeurs (voir l'annexe de cet article).

La Haute Autorité de Santé (2013) a publié un état des lieux des recommandations de bonnes pratiques en études cliniques et épidémiologiques. Il existe dans de nombreux pays des systèmes permettant de qualifier la capacité d'une étude à répondre à une question posée en la positionnant sur une échelle de valeurs selon l'adéquation du protocole à la question posée, l'existence ou non de biais importants dans la réalisation, l'adéquation des méthodes statistiques aux objectifs de l'étude, la puissance de l'étude et en particulier la taille de l'échantillon. Il est donc intéressant de savoir que les praticiens des statistiques ont tenté de hiérarchiser le niveau de preuve apporté par une étude ; la classification retenue par la Haute Autorité de Santé (2013) est reproduite à la figure 1.

Or, les énoncés de la partie 1 correspondent tous à des études observationnelles rétrospectives, sans contrôle de l'échantillon : des études de niveau 4 dans la classification ci-dessus, le plus bas niveau dans l'échelle des recommandations. Il est dommage d'offrir en exemple (en modèle ?) aux élèves une telle étude ; c'est pourquoi, à la partie 5, dans notre proposition d'énoncé, nous essaierons de faire monter le niveau de preuve scientifique en faisant intervenir le statisticien dès la conception de la collecte des données.

Niveau de preuve scientifique fourni par la littérature	Grade des recommandations
Niveau 1 • Essais comparatifs randomisés de forte puissance • Méta-analyse d'essais randomisés • Analyse de décision basée sur des études bien menées	**A** Preuve scientifique établie
Niveau 2 • Essais comparatifs randomisés de faibles puissance • Etudes comparatives non randomisées bien menées • Etudes de cohorte	**B** Présomption scientifique
Niveau 3 • Etudes Cas-Témoins	**C** Faible niveau de preuve
Niveau 4 • Etudes comparatives comportant des biais importants • Etudes rétrospectives • Séries de cas	

FIGURE 1 – *Gradation du niveau de preuve scientifique retenue par la Haute Autorité de Santé.*

Existence de nombreuses études statistiques sur Woburn — Depuis les années 1970, et au-delà de la simple comparaison d'une fréquence observée dans un sous-groupe à une valeur de référence nationale, de nombreuses autres études ont été menées, avec parfois des données différentes ou mises à jour, et en tout cas, selon des protocoles variés.

C'est le cas par exemple des études cas-témoin, séparant les individus selon la variable étudiée (cas de leucémie ou non), entre d'une part l'échantillon formé des enfants malades et d'autre part, un échantillon issu d'un tirage aléatoire contrôlé d'enfants sains ; l'objet de l'étude est de déterminer quel(le)s caractéristique(s), comme le lieu de résidence, l'âge, etc., diffèrent significativement entre les deux échantillons. Kevin Costas *et al.* (2002) effectuent une telle étude.

L'article fondateur des études académiques sur Woburn est toutefois celui de Steven Lagakos *et al.* (1986). Il essaie notamment de comparer les risques de leucémie à Woburn selon l'accès à l'eau polluée ou non : selon la quantité d'eau polluée ingérée par la mère pendant la grossesse, selon la quantité d'eau polluée bue par l'enfant lui-même. Cet article est notamment fondé sur une enquête téléphonique effectuée en 1982 pour analyser les facteurs, dont l'absorption d'eau des puits pollués, potentiellement liés à la survenue d'un problème de santé (leucémie, mort périnatale, anomalie congénitale, etc.) chez les enfants habitant Woburn pendant la période 1960–1982. De nombreuses critiques ont été apportées à cette approche : pas moins de cinq articles de commentaires ont été publiés dans le même volume du *Journal of the American Statistical Association*. Ces détracteurs, Brian MacMahon *et al.* (1986), soulignaient notamment la faiblesse méthodologique de l'enquête téléphonique, ses biais de sélection et de réminiscence (les familles touchées par la maladie retrouvent plus facilement des causes possibles) ; les faibles tailles d'échantillons disponibles pour ajuster un modèle statistique complexe ; le glissement prématuré d'une association significative entre pollution et maladie vers un lien de causalité (alors qu'à cet égard, neuf critères définis par Bradford Hill en 1965 sont requis pour réaliser le

glissement). Enfin, une commentatrice rappelait l'existence autre part dans le monde de *clusters* de leucémies inexpliqués.

L'affaire Woburn a officiellement été close par un autre rapport du *Massachusetts Department of Public Health* (1997), long de 138 pages, après le premier rapport de Parker et Rosen (1981). Les conclusions ont rejeté la mise en évidence d'une association entre le risque de leucémie et la contamination environnementale par des produits chimiques, malgré l'existence d'une concentration statistiquement significative de leucémies dans des zones particulières de la ville.

Ce très bref aperçu des controverses scientifiques doit permettre maintenant de mieux juger les présentations du cas Woburn rappelées à la partie 1. Le premier énoncé, notamment, semble un peu court, comme si un lycéen allait à l'aide de ses connaissances de probabilités et statistique résoudre une controverse difficile à lui seul : « Les autorités concluent qu'il n'y a rien d'étrange dans cette ville. Qu'en pensez-vous ? » Formuler un énoncé aussi peu disert sur le contexte de l'étude et sur les implications (tout à fait faibles !) du calcul mené est pour nous une réelle source d'illusions à venir et de confusion.

4. Pensons à une étude inter-échantillons (par simulations)

Ce qui a précédé a voulu mettre en lumière la question « Pourquoi les garçons de moins de 15 ans ? » issue de notre communication orale originelle (Gerville-Réache et Couallier, 2014), en fixant la maladie (la leucémie) et le lieu (Woburn).

Mais cette communication avait également pour objet les questions « Pourquoi Woburn ? Pourquoi la leucémie ? ». Elles avaient été suscitées de manière indirecte par l'énoncé de Philippe Dutarte (2007), qui invitait finalement à simuler d'autres comtés que celui de Woburn, en supposant que chacun de leurs garçons soit atteint d'une leucémie avec une probabilité égale à la fréquence de l'affection au niveau fédéral. (De telles simulations peuvent relever au choix, comme l'explique Jeanne Fine, 2015, d'une approche sondage ou d'une approche modèle, ces deux approches étant complémentaires et ne se contredisant pas.) Ces simulations permettaient d'avoir une estimation de la *P*-valeur associée aux données des garçons de Woburn pour le test de comparaison à la valeur de référence nationale, pour une hypothèse alternative de déviation unilatérale vers une prévalence locale plus forte.

Retour sur l'argument du singe dactylographe — Cependant, il nous semble que des éléments de contexte peuvent (doivent ?) perturber l'interprétation des résultats de ces simulations aux yeux d'un lycéen ; nous sommes conscients que les arguments qui suivent à cet égard sont toutefois bien plus discutables que ceux évoqués dans les parties précédentes, relatifs à la méthodologie statistique à suivre.

Imaginons en effet le cas limite de ces simulations, pour un grand nombre de comtés : il serait équivalent à une surveillance fédérale simultanée de l'ensemble des comtés des Etats-Unis pour diverses affections, sous une hypothèse de prévalences géographiques uniformes. Au vu du nombre de couples comtés-affections considérés, des faux positifs signalant des déviations locales statistiquement significatives pour une maladie donnée par rapport à la prévalence fédérale sont inévitables et uniquement imputables au « hasard » (voilà où résidait l'argument du

singe dactylographe). Des simulations permettent de mettre nettement en évidence ce phénomène et c'est peut-être une explication aux *clusters* de leucémies inexpliqués par ailleurs, dont Brian MacMahon *et al.* (1986) notaient l'existence.

Mais le point négligé par l'ensemble des énoncés reproduits ci-dessus (y compris des commentaires afférents lorsqu'il y en avait), c'est que bien sûr, l'étude menée par les experts ne s'est pas intéressée à une seule affection, mais à plusieurs (autres cancers, comme ceux du rein, du foie ou de la vessie, fausses couches, problèmes oculaires, etc.), et que des déviations significatives ont été observées pour plusieurs affections (mais pas toutes). C'est la coïncidence en un même lieu qui est signe d'un phénomène non uniquement attribuable au hasard. Là encore on ne retrouve pas cette complexité dans les énoncés d'exercices cités, qui se focalisent sur une seule affection et ne mentionnent pas que d'autres maladies sont étudiées en parallèle.

Retour sur « la communauté locale s'émeut » — L'énoncé de Philippe Dutarte (2007) contient l'indication que « la communauté locale s'émeut d'un grand nombre de leucémies infantiles survenant dans certains quartiers de la ville ». Nous avons déjà expliqué dans les deux parties précédentes que les études rétrospectives ne pouvaient donner lieu à l'établissement de preuves statistiques. L'observation d'un nombre anormalement élevé de cas d'une maladie par la population d'une aire géographique restreinte donnée est en un sens vouée à se produire, selon l'argument de singe dactylographe évoqué ci-dessus, lorsque l'on considère un grand ensemble découpé en de nombreuses telles petites aires. Des études approfondies sont donc nécessaires pour déterminer si l'excès de cas observé est statistiquement significatif ou non.

L'actualité nous fournit régulièrement des exemples de ce type. Le dernier en date en France correspond à une suspicion d'agrégat de cancers pédiatriques dans une commune viticole de Gironde, Peignac, liée, peut-être, à des épandages de pesticides sur les vignes situées à proximité de l'école communale. Un rapport statistique et épidémiologique a été écrit à ce sujet par les pouvoirs publics (Agence régionale de santé Aquitaine et Institut de veille sanitaire, 2013). Il contient notamment cette phrase :

> Néanmoins, cette étude étant réalisée *a posteriori* (après l'observation d'un excès de cas), il n'est pas statiquement possible de tester la présence de cet excès.

C'est pourquoi, dans notre proposition d'énoncé ci-dessous, nous parlons d'une inquiétude de la population non pas liée à des maladies déjà déclarées en nombres plus grands qu'attendus, mais causée par la découverte d'un phénomène de pollution dont on peut se dire qu'il risque d'avoir une influence sur la santé.

5. Recommandation d'énoncé et conclusion

Il découle des parties précédentes que les énoncés d'exercices considérés, en simplifiant la réalité du traitement statistique à outrance et par leur rédaction certes suggestive, donnent l'impression trompeuse qu'un petit calcul de niveau terminale scientifique permet de mettre en évidence un problème de santé publique. Nous voulons corriger cet effet malheureux en proposant un nouvel énoncé. Il s'agit d'un énoncé de « science-fiction » au sens où nous tâchons de rester dans l'esprit de la controverse originelle, quitte à en perdre la lettre : le déroulé exact des

plaintes, réponses, enquêtes, expertises et contre-expertises. On rappelle qu'on pourra avoir une idée de la complexité de l'affaire en lisant le cas rédigé par Bair et Svitana (2008), qui retrace notamment le déroulé juridique de l'affaire.

Enoncé corrigeant les critiques soulevées — L'énoncé ci-dessous corrige de nombreux points évoqués dans les parties précédentes et réalise selon nous un bon compromis entre rigueur et accessibilité. Le plus important à nos yeux est de faire formuler aux élèves les hypothèses et le mode de recueil des données avant de préciser ces dernières, car telle est la bonne méthodologie statistique (ainsi que nous l'avons discuté en partie 2). En outre, nous soulignons que d'autres affections que la leucémie sont étudiées et soulignons sans la décrire la complexité de l'étude globale à effectuer (afin de tenir compte de la nécessité d'une étude de comparabilité des sous-groupes et des griefs soulevés à la partie 4).

Pour changer et offrir de nouvelles données à ceux qui sont intéressés par ce cas, nous nous intéressons plutôt à une répartition géographique que sexuée des cas.

> *L'énoncé qui suit est inspiré de faits réels mais n'embrasse pas toute leur complexité, il se veut simplement une introduction à l'utilisation d'arguments statistiques dans les questions publiques : santé et justice.*
>
> Dans les années 1970, quelques habitants de Woburn, petite ville industrielle au Nord-Est des Etats-Unis, se plaignent des impacts potentiels de la pollution sur leur santé après que des produits toxiques ont été découverts dans certains puits d'eau potable de l'Est de la ville. Devant l'inaction des autorités locales, un groupe de huit familles d'un quartier de l'Est de la ville veut faire effectuer ses propres expertises, notamment des expertises statistiques sur la prévalence de différents cancers.
>
> 1. Quelles données les experts vont-ils recueillir selon vous, quelles vérifications veulent-ils effectuer ? Les réponses à ces questions doivent vous amener ensuite à formuler des hypothèses statistiques, à conserver ou à invalider au vu des données.
>
> *L'enseignant ne donnera la suite de l'énoncé qu'une fois que les élèves auront répondu à cette première question.*
>
> Le recueil des données pour un des cancers étudiés, en l'occurrence, les leucémies infantiles (survenant chez les enfants d'au plus 14 ans) est le suivant ; on précise, par zone de la ville, le nombre total d'enfants et celui des cas observés :
>
> | Zone | Total enfants | Nombre leucémies |
> |---|---|---|
> | Est | 1 707 | 6 |
> | Reste | 10 041 | 6 |
> | Total | 11 748 | 12 |
>
> La prévalence fédérale des leucémies infantiles est de 45 cas pour 100 000 enfants.
>
> 2. Confrontez les hypothèses formulées à ces données : montrez que le hasard seul ne peut raisonnablement expliquer les nombres de cas de leucémies observés dans la zone Est de la ville.

> Ainsi, des investigations plus poussées ont été menées et de nombreuses expertises et contre-expertises ont eu lieu pour savoir s'il fallait attribuer la sur-prévalence des leucémies et d'autres affections à la contamination de l'eau potable. L'affaire judiciaire Woburn avait commencé et durera de nombreuses années.
>
> *L'enseignant pourra alors expliquer brièvement la nécessité de pouvoir comparer les groupes géographiques en tous les termes et variables sauf, précisément, la localisation géographique, afin d'être en mesure d'associer la sur-prévalence à cette localisation. Dans un deuxième temps, la cause de l'association pourra alors être recherchée.*

Véritable recommandation : lecture critique — L'énoncé recommandé ci-dessus nous semble à la fois correct du point de vue de la méthodologie statistique et abordable au niveau lycée, tout du moins sous la direction d'un enseignant quelque peu conscient des tenants et aboutissants de la démarche statistique des tests d'hypothèses, et de ses écueils. Bien évidemment, la méthodologie statistique, sa bonne application ou les écueils à éviter, ne sont pas à enseigner en tant que tels : les élèves n'ont pas à être formés en profondeur sur ce sujet.

Au vu de l'énoncé proposé, chacun sera juge de l'opportunité ou pas d'enseigner les probabilités et la statistique à partir de données réelles en contexte.

Notre avis personnel, cependant, est le suivant. La mise en œuvre d'une telle démarche statistique et le traitement de données réelles requièrent selon nous une bonne pratique et ne peuvent s'apprendre dans les manuels, contrairement par exemple aux fondements du calcul des probabilités et même peut-être à ceux de l'estimation par intervalles de confiance. Ainsi, nous pensons que le traitement en classe de l'affaire Woburn devrait se réduire à la lecture critique d'un document mêlant calculs de probabilités (détermination de P-valeurs, éventuellement approximation de leur valeur par simulations à défaut d'une détermination exacte) et surtout, retours et mises en garde quant à la démarche statistique suivie.

Lecture critique d'un document bien rédigé (par des professionnels) plutôt que mise en œuvre personnelle par les élèves : d'aucuns pourront trouver qu'il est dommage de s'arrêter en si bon chemin, mais ne pas instiller d'idée fausse ou excessivement simplificatrice quant à la bonne détermination de la significativité d'un phénomène est selon nous à ce prix.

A. Annexe : autres illustrations de risques méthodologiques

Dans cette annexe, nous voulons dépasser le cas Woburn et présenter quelques exemples mettant en garde contre un usage non méthodologiquement contrôlé de la statistique inférentielle. La conclusion sera que même après le lycée et même dans le monde professionnel des chercheurs, des chausse-trappes méthodologiques sont à craindre. Cela renforcera notre conclusion qu'au niveau du lycée, il serait plus profitable de se borner à une lecture critique de documents bien construits et méthodologiquement exacts.

La statistique inférentielle comme preuve et « argument d'autorité » — Le recours à la statistique inférentielle est devenue un élément de preuve incontournable dans de nombreuses disciplines scientifiques ou pour des applications pratiques devant être validées par des autorités supérieures. Ainsi, la preuve de l'effet d'un médicament, de la liaison entre deux phénomènes,

la recherche et la mise en évidence d'associations présupposées (en psychologie, en sciences sociales, en santé publique, en économie, en finance) passent désormais par l'analyse statistique de données et très souvent par la modélisation de celles-ci. Cela confère aux tests statistiques et aux *P*-valeurs associées une toute-puissance qui peut se révéler bien dangereuse si des vérifications élémentaires de méthodologie ne sont pas réalisées.

Pour rester dans le cadre judiciaire, on peut ainsi commencer par mentionner une conférence filmée de Peter Donnelly (2005) reprenant en particulier certaines critiques sur l'usage du calcul des probabilités et de la statistique dans les cours de justice.

La psychologie et les neurosciences sont deux domaines scientifiques dont les études ont fait l'objet elles-mêmes d'études. Les questions posées et objections soulevées tournent autour de la reproductibilité des résultats. Ainsi, l'étude menée par le groupe de chercheurs *Open Science Collaboration* (2015) indique que sur 100 études publiées dans trois revues scientifiques majeures en psychologie, 97 comprenaient la mise en évidence d'un résultat statistiquement significatif, fondée sur un test statistique et des données de *P*-valeur associée inférieure à 5%. La reproduction des mêmes protocoles n'a pu aboutir à ces résultats positifs que dans 36% des cas. Ce problème de non-reproductibilité se pose dans de nombreux autres domaines scientifiques, comme le souligne Pierre Barthélémy (2013) dans un article sur le *blog* « Passeur de sciences » du *Monde*. C'est pourquoi Valen E. Johnson (2013) suggère, par une étude adoptant un point de vue bayésien, que les seuils de significativité des *P*-valeurs soient drastiquement abaissés à des niveaux autour de 0.5%, comme cela est d'ailleurs la pratique en sciences physiques.

Un seuil inadéquat n'est sans doute pas la seule explication ou le seul remède à la non-reproductibilité des études. Trois grandes familles de causes peuvent être avancées : premièrement, John Ioannidis (2005) met en avant un biais de publication conduisant à ce que les résultats positifs soient soumis à publication mais pas les résultats négatifs.

Deuxièmement, comme nous l'avons déjà développé, des erreurs méthodologiques, comme des analyses en sous-groupes indues car non fondées sur des groupes constitués à l'avance, ou d'autres biais de sélection de l'échantillon peuvent compromettre la qualité d'une étude. (Voir à ce sujet le travail du consortium STROBE[5] pour l'amélioration de la qualité de publication des études observationnelles en épidémiologie.)

Ces erreurs méthodologiques sont sans doute plus fréquentes qu'on ne le croit, et des erreurs bien plus flagrantes, liées à l'utilisation de tests statistiques incorrects pour le but recherché, sont présentes dans la littérature. Ainsi en neurosciences, une étude menée par Sander Nieuwenhuis *et al.* (2011) montre que la moitié des articles de recherche publiés récemment dans plusieurs revues prestigieuses n'effectuent pas correctement le test de comparaison de moyennes de deux échantillons indépendants (et procèdent par deux tests séparés de comparaison à une valeur de référence).

Même conclusion que précédemment — Ainsi, si même des chercheurs confirmés peuvent commettre des erreurs méthodologiques voire appliquer de mauvaises procédures statistiques, peut-on raisonnablement demander à des lycéens ou même à leurs enseignants d'avoir suffisamment de recul face à des données réelles et de savoir les (faire) traiter en autonomie ? Ne vaut-il pas mieux leur demander, comme nous le préconisons, de (faire) effectuer une lecture critique d'un document bien écrit par des professionnels de la statistique ?

[5] http://www.strobe-statement.org/

Références

[1] Agence régionale de santé Aquitaine et Institut de veille sanitaire (2013), Investigation d'une suspicion d'agrégat de cancers pédiatriques dans une commune viticole de Gironde, lien vers le rapport, publié en juin 2013.

[2] Bair, S. et K. Svitana (2008), *webpage* "Science in the courtroom – the Woburn toxic trial", http://serc.carleton.edu/woburn, page consultée le 17 juillet 2015.

[3] Barache, F., R. Bauer, S. Barache, et S. Bauer (2011), *100% exos Maths 1re S*, Hatier.

[4] Barthélémy, P. (2013), Une étude ébranle un pan de la méthode scientifique, billet posté le 13 novembre 2013 sur le *blog* « Passeur de sciences » du *Monde*, lien vers le billet.

[5] Costas, K., R. S. Knorr, et S. K. Condon (2002), A case-control study of childhood leukemia in Woburn, Massachusetts : the relationship between leukemia incidence and exposure to public drinking water, *Science of the Total Environment*, **300**(1–3), 23–35.

[6] Direction générale de l'enseignement scolaire (2012), Statistiques et probabilités, in *Ressources pour la classe de première générale et technologique*, lien vers le document.

[7] Donnelly, P. (2005), How juries are fooled by statistics – filmed July 2005 at TEDGlobal 2005, lien vers la vidéo.

[8] Dutarte, P. (2007), Inquiétudes à Woburn, in Commission inter-IREM Lycées technologiques, éditeur, *Statistique et citoyenneté : le citoyen face au chiffre*, pages 33–55, lien vers le document.

[9] Fine, J. (2015), A propos de l'enseignement des affaires Woburn et Castaneda, *Statistique et Enseignement*, **6**(1), 45–49.

[10] Fisher, R. (1938), Presidential address to the first Indian statistical congress, *Sankhyā*, **4**, 14–17.

[11] Gerville-Réache, L. et V. Couallier (2014), L'enseignement des affaires Woburn et Castaneda, communication orale aux 46e Journées de Statistique, lien vers un résumé.

[12] Haute Autorité de Santé (2013), Niveau de preuve et gradation des recommandations de bonne pratique – État des lieux, lien vers le rapport, validé en avril 2013.

[13] ICH (1995), Structure and content of clinical study reports – E3, lien vers les recommandations édictées.

[14] Ioannidis, J. P. (2005), Why most published research findings are false, *PLOS Medicine*, **2**(8), e124, DOI :10.1371/journal.pmed.0020124.

[15] Johnson, V. E. (2013), Revised standards for statistical evidence, *Proceedings of the National Academy of Sciences of the United States of America*, **110**(48), 19313–19317.

[16] Lagakos, S. W., B. J. Wessen, et M. Zelen (1986), An analysis of contaminated well water and health effects in Woburn, Massachusetts, *Journal of the American Statistical Association*, **81**(395), 583–596.

[17] MacMahon, B., R. L. Prentice, W. J. Rogan, S. H. Swan, J. M. Robins, et A. S. Whittemore (1986), Comments on the article "An analysis of contaminated well water and health effects in Woburn, Massachusetts" by Lagakos, Wessen, and Zelen, *Journal of the American Statistical Association*, **81**(395), 597–610.

[18] *Massachusetts Department of Public Health* (1997), Woburn childhood leukemia follow-up study – Volume I: Analyses, lien vers une copie du rapport.

[19] Nieuwenhuis, S., B. U. Forstmann, et E.-J. Wagenmakers (2011), Erroneous analyses of interactions in neuroscience : A problem of significance, *Nature Neuroscience*, **14**(9), 1105–1007.

[20] *Open Science Collaboration* (2015), Estimating the reproducibility of psychological science, *Science*, **349**(6251), DOI :10.1126/science.aac4716, lien vers l'article.

[21] Parker, G. S. et S. L. Rosen (1981), Woburn: cancer incidence and environmental hazards, 1969–1978, Massachusetts Department of Public Health, lien vers le rapport.