A PROPOS DE L'ENSEIGNEMENT DES AFFAIRES WOBURN ET CASTANEDA

Jeanne FINE[1]

1 Introduction

Les affaires Woburn et Castaneda, proposées par Philippe Dutarte [1], sont bien connues des professeurs de mathématiques du lycée pour répondre à des questions statistiques issues de données réelles à l'aide de modélisation et de simulation.

Lors des 46èmes Journées de Statistique de la SFdS, qui se sont tenues à Rennes du 2 au 6 juin 2014, Léo Gerville-Réache et Vincent Couallier [2] reviennent sur l'enseignement de ces affaires et concluent : « *L'analyse de ces deux cas ne correspond pas à ce que l'on peut attendre d'un élève de lycée et devrait être retirée des enseignements* ».

Dans ce libre-propos, je souhaite au contraire montrer que ces deux exemples ont toute leur place dans les programmes et que les arguments utilisés pour les critiquer sont incorrects. Les critiques sont intéressantes pour voir que deux approches de la modélisation d'un problème (approche modèle et approche sondage) peuvent faire émerger une situation apparemment paradoxale.

Les critiques ne portant pas sur la simulation, nous utiliserons le raisonnement probabiliste.

2 L'affaire Woburn

Pour l'affaire Woburn, on observe 9 leucémies de jeunes garçons, ce qui semble anormalement important pour une population de 5969 garçons de moins de 15 ans alors que la proportion de leucémies dans la population des garçons de moins de 15 ans des USA est 0.00052.

Pour vérifier si cette observation est anormalement élevée, *on suppose que les 5969 jeunes garçons sont soumis, indépendamment les uns des autres, à la probabilité 0.00052 de contracter une leucémie* (approche modèle). Cette probabilité, 0.00052, appelée « prévalence », est estimée par la proportion de leucémies dans la population. Le nombre N de leucémies sur un tel échantillon suit alors une loi binomiale de paramètres 5969 et 0.00052 et on cherche la probabilité que N soit supérieur ou égal à 9. Cette loi peut être approchée par une loi de Poisson de paramètre 3.10 (= 5969 x 0.00052) et la probabilité cherchée est égale à 0.0047.

Cette probabilité est extrêmement faible ce qui confirme la présomption d'un nombre anormalement élevé de leucémies à Woburn et ouvre la voie à de nouvelles recherches pouvant expliquer ce résultat.

[1] Statisticienne, jeanne.fine@gmail.com

Pour la modélisation, nous aurions pu *supposer extraire au hasard un échantillon de 5969 garçons de moins de 15 ans dans la population des garçons de moins de 15 ans des USA (dont une proportion d'individus égale à 0.00052 a une leucémie) et assimiler cet échantillon sans remise à un échantillon avec remise car la taille de la population est très vaste par rapport à la taille de l'échantillon* (approche sondage). C'est la présentation faite par Philippe Dutarte. Nous retrouvons alors la même modélisation que dans l'approche modèle, échantillon de 5969 variables aléatoires indépendantes de même loi de Bernoulli de paramètre 0.00052. Le nombre N de leucémies suit alors une loi binomiale de paramètres 5969 et 0.00052, qui peut être approchée par une loi de Poisson de paramètre 3.10, et la probabilité que N soit supérieur ou égal à 9 est donc encore 0.0047. Puisqu'il s'agit de tirage équiprobable d'échantillons de taille 5969, cette probabilité correspond à la proportion d'échantillons de taille 5969 présentant un nombre de leucémies supérieur ou égal à 9.

J'ai développé l'approche modèle et l'approche sondage en probabilités et statistique inférentielle en [3] et conclu l'article ainsi : « *Il ne s'agit pas de proposer de remplacer l'approche modèle par l'approche sondage mais de proposer différentes approches en fonction des données. Les deux approches sont complémentaires* ».

Les critiques apportées par Léo Gerville-Réache et Vincent Couallier me renforcent dans ma conviction de la nécessité de présenter les deux approches. Selon la façon dont on envisage un problème, les deux approches peuvent entrer en conflit. Il semble que, pour eux, l'approche modèle se réfère aux individus, ici les garçons de moins de 15 ans, alors que l'approche sondage se réfère aux groupes de 5969 garçons de moins de 15 ans. On a déjà montré en début de paragraphe que, selon que l'on utilise l'approche modèle ou l'approche sondage (en assimilant un tirage sans remise à un tirage avec remise car le taux de sondage est négligeable), on est finalement conduit à une même modélisation probabiliste.

Reprenons les différentes critiques que les auteurs ont apportées. La première est la suivante : « *on n'a pas des groupes comparables, fondement de l'expérimentation ; la randomisation des individus dans les groupes expérimentaux est une condition nécessaire de comparabilité statistique* ». Cette critique n'est pas recevable, car il n'est pas question ici de comparer deux groupes qui seraient soumis à des traitements différents. Si c'était le cas, il serait en effet nécessaire d'utiliser une assignation randomisée des individus à chacun des groupes et de vérifier l'homogénéité des groupes en début de traitement. L'objectif ici est de comparer une proportion de leucémies observée à Woburn 0.0015 (= 9/5969) à une valeur de référence 0.00052.

La deuxième critique s'appuie sur le fait que si l'on multiplie les échantillons de taille 5969, la probabilité d'en trouver au moins un pour lequel le nombre de leucémies est supérieur ou égal à 9 augmente. « *Les phénomènes rares peuvent être, paradoxalement, très fréquents* ». Ce commentaire fait référence à un problème célèbre illustré par l'exemple du singe dactylographe présenté en annexe. On montre en effet que, aussi faible soit la probabilité d'un événement attaché à une expérience, il se réalisera avec une probabilité aussi proche de 1 que l'on veut si on réalise l'expérience dans les mêmes conditions un nombre suffisant de fois. Mais que vient faire cet argument dans l'exemple de Woburn ? La réponse vient peut-être dans ce qui suit.

En effet, les auteurs ne multiplient pas à l'envi le nombre d'échantillons indépendants de taille 5969. Ils constituent, à partir d'une population de 20 millions de garçons de moins de 15 ans, 3350 groupes de taille 5969 deux à deux disjoints (il s'agit donc bien d'échantillons indépendants) et montrent que l'espérance du nombre de groupes, parmi 3350, ayant au moins

9 leucémies est 16, prouvant par-là que de tels groupes existent. Si l'on poursuit le raisonnement, on en déduit que la probabilité d'avoir choisi au hasard un tel groupe est estimée à 16 / 3350, soit la probabilité déjà calculée : 0.0047. Il existe bien sûr de tels groupes mais ils sont rares.

Plutôt que considérer 3350 échantillons indépendants, si l'on considère l'ensemble des échantillons de taille 5969 extraits au hasard dans une population dans laquelle la proportion de leucémies est de 0.00052, on sait, par l'approche sondage, qu'il existe des échantillons pour lesquels le nombre de leucémies est supérieur ou égal à 9, en proportion 0.0047. C'est parce que cette proportion est très faible que l'on conclut à un nombre anormalement élevé de leucémies à Woburn… tout en sachant que cette conclusion est fausse avec une probabilité de 0.0047.

Pour mieux faire comprendre l'argument, les auteurs font une analogie avec le problème des anniversaires : « *Si la probabilité qu'un de mes camarades ait la même date d'anniversaire que moi est faible, la probabilité qu'il existe deux élèves de ma classe ayant la même date d'anniversaire l'est beaucoup moins. Converti ici : ayant choisi une ville Woburn ayant 5969 jeunes garçons, la probabilité de trouver plus de 9 cas de leucémie dans cette ville est faible, mais si je considère qu'il existe plus de 500 villes de plus de 5969 jeunes garçons, la probabilité de trouver au moins une ville dont le nombre de cas de leucémies est supérieure à 9 ne l'est pas (90 % de chance) !* » Comme il ne s'agit pas du même problème, il est vraiment difficile de voir l'analogie.

La critique de fond semble être la suivante, en gras dans le texte : « *Il est statistiquement très difficile de répondre à une question sur la réalisation d'une variable aléatoire alors même que la question est posée parce que cette réalisation semble curieuse* ».

Il est au contraire tout à fait légitime, face à une situation étonnante, de vérifier à l'aide d'une modélisation simple si cette observation peut être due au hasard. En revanche, lors de la mise en place d'un protocole expérimental, pour comparer deux traitements médicaux par exemple, 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 ». Il semble que les auteurs confondent les deux situations.

Enfin, les auteurs posent les questions : « *Pourquoi Woburn ? Pourquoi la leucémie ? Pourquoi les garçons de moins de 15 ans ?* », ce qui semble être une façon de dire que les données réelles sont trop difficiles à traiter en classe.

3 Affaire Castaneda

Pour l'affaire Castaneda, les avocats de Partida, un condamné américain d'origine mexicaine, attaquent le jugement au motif que la désignation des jurés de ce comté est discriminante à l'égard des américains d'origine mexicaine.

Les critiques sont les suivantes :

« *Dans quelle mesure ce comté est un comté parmi d'autres ?* ». On retrouve l'idée de comparabilité des groupes.

« *Le processus de sélection d'un juré n'est pas uniquement basé sur un tirage aléatoire dans la population. Par exemple, il y a un âge minimum, la nécessité d'une bonne connaissance de l'anglais écrit et parlé…* ». Il est malgré tout possible, dans un premier temps, de modéliser avec un tirage aléatoire, en sachant que le modèle simplifie la réalité.

« *On peut également se demander quel processus les avocats de Partida ont utilisé pour choisir de regarder la proportion de personnes d'origine mexicaine dans les jurys du comté ? Ont-ils choisi une caractéristique du condamné au hasard parmi l'ensemble (très grand) des caractéristiques du condamné ou ont-ils exploré un grand ensemble de ces caractéristiques et choisi celle dont l'écart à l'aléatoire théorique était le plus grand ? Ce que la statistique nous dit c'est qu'en cherchant un écart "significatif" à l'aléatoire, on finit toujours par en trouver !* ». Comme s'il fallait chercher très loin la possibilité d'une discrimination vis-à-vis de la population d'origine mexicaine.

4 Conclusion

Les critiques de Léo Gerville-Réache et Vincent Couaillier sont intéressantes pour voir que les deux approches, approche modèle et approche sondage, peuvent faire émerger une situation apparemment paradoxale. Les bases de la théorie des sondages (échantillonnage en population finie, estimation et inférence) devraient être enseignées en complément de l'approche classique de la statistique inférentielle (modélisation d'expériences aléatoires indépendantes) pour mieux appréhender les différents aspects d'un problème. Quand les deux approches peuvent être utilisées, elles conduisent heureusement aux mêmes résultats.

En conclusion, ces deux affaires, Woburn et Castaneda, qui ont utilisé des arguments statistiques pour leur résolution, sont très intéressantes à présenter au lycée. C'est à partir d'une modélisation simple et de simulations que les élèves peuvent apporter une première réponse. Bien sûr, la réponse n'est pas définitive et peut donner lieu à de nouvelles questions.

Références

[1] Dutarte, P. (2007), *Présenter aux futur(e)s professeur(e)s une image positive de la statistique et ses enjeux citoyens*, http://dutarte.perso.neuf.fr/statistique/corfem.htm

[2] Gerville-Reache, L. et V. Couaillier (2014), L'enseignement des affaires Woburn et Castaneda, http://papersjds14.sfds.asso.fr/submission_41.pdf

[3] Fine, J. (2010), Probabilités et statistique inférentielle. Approche modèle versus approche sondage, *Statistique et Enseignement*, **1**(2), 5-21.

Annexe : Le singe dactylographe

Énoncé (à partir d'un texte d'Émile Borel, 1909)

Un singe tape au hasard 25 caractères de suite sur un clavier de 50 touches (correspondant aux lettres, aux chiffres et autres caractères… blanc, de ponctuation, accentués).

1) Quelle est la probabilité qu'il écrive la phrase suivante (de 25 caractères, y compris le blanc et le point d'exclamation) ?

le hasard mène le monde !

2) Soit $N = 50^{25}$. Il répète N fois de suite dans les mêmes conditions l'expérience précédente. Quelle est la probabilité qu'il écrive au moins une fois la phrase ci-dessus ?
3) Combien de fois doit-il répéter l'expérience pour que la probabilité qu'il écrive au moins une fois la phrase ci-dessus soit supérieure ou égale à 0.999 ?

Solution

1) La probabilité qu'il écrive la phrase est $\left(\frac{1}{50}\right)^{25}$ (il faut qu'il tape 25 fois de suite le caractère correct).
2) La probabilité qu'il écrive la phrase au moins une fois sur 50^{25} expériences indépendantes est
   $1 - \left(1 - \frac{1}{N}\right)^{N}$ avec $N = 50^{25}$, soit approximativement $1 - e^{-1}$, environ 0.63.
3) On cherche n tel que $1 - \left(1 - \frac{1}{N}\right)^{n} \geq 0.999$ avec $N = 50^{25}$.
   Si on pose $x = n/N$, cela revient à chercher x tel que $1 - e^{-x} \geq 1 - 0.001$, soit $x \geq \ln(1000)$ (environ 6.9).
   Si le singe répète l'expérience 6.9 fois 50^{25}, il est quasiment sûr qu'il tapera au moins une fois la phrase "le hasard mène le monde !".

Morale de l'histoire : aussi faible que soit la probabilité d'un événement attaché à une expérience, il se réalisera avec une probabilité aussi proche de 1 que l'on veut si on réalise l'expérience dans les mêmes conditions un nombre suffisant de fois.

Cet argument est utilisé dans le débat sur l'existence de la vie sur une planète hors du système solaire. Le problème est d'évaluer la probabilité qu'il y ait de la vie sur une planète et le nombre de planètes dans l'univers !