Université Lille1 **Département de Mathématiques** **M.Fradon**

Collège Carnot, Lille **F.Laithier, S.Daniel** **2017-2018**

Maths en Jean

Le coté clair des forces obscures

Grave question, qui engage l'avenir de l'humanité :

Si, sur chaque planète de la galaxie, une proportion croissante de jedis est du coté obscur de la force, est-il possible que la proportion de ceux qui sont du coté clair augmente dans la galaxie ?

Règle du jeu : Pas de jedi "hors sol" qui ne serait d'aucune planète. Mais les jedis changent de planète, meurent, de nouveaux jedis apparaissent, etc. Et pas de jedi "clignotant" qui change de coté tout le temps. Certains peuvent basculent du coté obscur, ou revenir du coté clair, mais quand on fait le décompte on sait qui est dans quel camp.

La même question, à l'usage de ceux qui n'aiment pas Star Wars :

Politique et polémique

Dans un pays que nous ne nommerons pas, le gouvernement lance un "grand projet pour l'enseignement" : construction d'écoles, nouveaux programmes, scolarisation des enfants de tous les milieux, recrutement d'enseignants, mieux formés, mieux payés, pas plus de 24 élèves par classe, financement de matériel, voyages scolaires pour tous, etc, etc. Et, pour constater les progrès fulgurants attendus, évaluation tous les ans de tous les enfants. On calcule le taux de réussite, c'est-à-dire la proportion d'enfants qui savent faire ce qui leur a été enseigné pendant l'année.

10 ans après, le taux de réussite a bien diminué !

L'opposition exige l'arrêt immédiat du grand projet pour l'enseignement, accusé de faire baisser le niveau des élèves. Le gouvernement se défend en publiant des statistiques qui montrent que, chez les enfants de milieux défavorisés, le taux de réussite a bien augmenté. "Le grand projet pour l'enseignement est un facteur de justice sociale..." annonce un communiqué ministériel.

La presse en déduit que le grand projet fait du nivellement par le bas : "Le gouvernement reconnait que la progression des uns s'est faite au détriment des autres. Arrêtons le gâchis !".

Le gouvernement rétropédale à la hâte en calculant le taux de réussite des enfants de milieux favorisés. Il a aussi augmenté. "Le grand projet fait progresser tous les enfants..." lit-on dans le communiqué suivant.

L'opposition hurle à la supercherie et à la manipulation des données : "Si le niveau avait augmenté dans toutes les catégories d'enfants, il n'aurait pas baissé au total. C'est mathématique !"

Est-ce mathématique ?

A part conseiller à tout le monde de se calmer, que peut-on faire pour éclaircir les choses ? Peut-on déterminer si le "grand projet", oui ou non, améliore le niveau des élèves ?

Encore une autre version de la question des jedis, et pourquoi elle engage l'avenir de l'humanité :

Darwin annonce-t-il la fin du monde ?

La loi de la sélection naturelle de Darwin constate que :

Au sein d'une population donnée, la proportion d'individus porteurs d'une caractéristique avantageuse pour leur survie augmente au fil des générations.

Mais qu'est-ce qu'un individu, qu'est-ce qu'une population ?

Question d'échelle : on peut considérer Lille comme une population d'humains, mais chaque humain est aussi une population de cellules, et du point de vue des mitochondries c'est la cellule qui est une population. On peut même considérer la France comme une population de communes dont Lille est un individu !

A tous les étages, il y a des conflits d'échelle : des caractéristiques avantageuses pour l'individu mais nuisibles pour la population, ou l'inverse. Une cellule qui s'affranchit des mécanismes limitant sa prolifération (cellule cancéreuse), ou un lillois qui ne paye pas ses impôts, sont des exemples de conflit d'échelle.

D'après la loi de Darwin, si une caractéristique induit un conflit d'échelle, soit les individus porteurs vont progressivement disparaitre (caractéristique nuisible à leur survie) soit les populations qui les contiennent vont progressivement disparaitre (caractéristique nuisible à la population en tant qu'individu à l'échelle au-dessus). Bref, comme il y a des conflits d'échelle absolument partout, toute vie va disparaitre sur la Terre...

On aimerait qu'il y ait un bug dans cette histoire. Mais y en a-t-il un ? Et où ?

Université Lille1 **Département de Mathématiques** **M.Fradon**

Collège Carnot, Lille **F.Laithier, S.Daniel** **2017-2018**

Maths en Jean

Inversion de Simpson : le piège des comparaisons de proportions

Spoiler alert : Comme l'an dernier, ceci est pour les enseignants seulement. Ca contient des "solutions". Ne pas montrer aux élèves ! Ne pas transmettre à l'Asso Maths en Jeans car elle met tout sur Internet ! Et peut-être ne pas lire tout de suite, pour ne pas se gâcher le plaisir. Comme vous voulez ;-)

Et aussi : Les idées et valeurs numériques qui suivent sont extraites du génial article de Jean-Paul Delahaye dans "Pour la Science" en 2013. Il contient entre autres une jolie représentation graphique du problème, et d'autres applications que celles que je mentionne.

1 Simpson, une catastrophe pour les stats

Deux questions à poser à froid au premier matheux qui nous tombe sous la main (j'ai essayé sur moi, sur l'équipe de proba-stats, etc) :

— On teste un médicament contre un placebo, sur un groupe de volontaires.
Dans l'ensemble, le médicament guérit une plus grande proportion de gens que le placébo.
Le test doit-il conclure que le médicament est bénéfique pour une majorité de gens ? Ou est-il possible que ces mêmes données montrent que le médicament nuit à la guérison ? Par exemple, que sur les femmes testées, le médicament soit moins performant que le placébo, et sur les hommes testés aussi ?
A cette question, tout le monde répond que le test montre que le médicament est bénéfique, au moins pour les hommes *ou* les femmes.

— Pour a, b, c, d et a', b', c', d' dans $\mathbb{N}^*$, est-il vrai que

$$\frac{a}{b} > \frac{c}{d} \quad \text{et} \quad \frac{a'}{b'} > \frac{c'}{d'} \qquad \text{implique} \qquad \frac{a+a'}{b+b'} > \frac{c+c'}{d+d'} \qquad ?$$

Tous les matheux répondent "non, ça se saurait !".

Et pourtant c'est la même question, et les deux réponses sont exactement contradictoires ! ! !

Il suffit de noter b le nombres de femmes qui prennent le placébo et d le nombre de celles qui prennent le médicament, a le nombre de guérisons parmi celles sous placébo et c le nombre de guérisons parmi celles sous médicament. Faire de même avec a', b', c', d' pour les hommes.

$\frac{a}{b} > \frac{c}{d}$ signifie alors que le médicament guérit moins que le placébo chez les femmes, idem avec $\frac{a'}{b'} > \frac{c'}{d'}$ chez les hommes. Dire qu'on peut avoir $\frac{a+a'}{b+b'} < \frac{c+c'}{d+d'}$ est équivalent à dire qu'en regroupant hommes et femmes on peut constater que le médicament guérit une plus grande proportion de gens et croire qu'il est meilleur.

La réponse juste est la deuxième. Il peut se produire que

$$\frac{a}{b} > \frac{c}{d} \quad \text{et} \quad \frac{a'}{b'} > \frac{c'}{d'} \qquad \text{et pourtant} \qquad \frac{a+a'}{b+b'} < \frac{c+c'}{d+d'}$$

J'appellerai cette situation une inversion de Simpson. Le terme officiel est "paradoxe de Simpson" mais ça m'ennuie d'appeler paradoxe le simple fait qu'une affirmation fausse ait des contre-exemples. Ce n'est pas paradoxal. C'est normal. Même si au niveau des applications des maths ce truc parfaitement normal est extrémement génant.

Car l'inversion de Simpson est un **gros** problème !

On a des techniques statistiques très performantes pour déterminer si une proportion est supérieure à une autre. Mais cette non-conservation des inégalités quand on regroupe des populations dit carrément que la constatation qu'une proportion est supérieure à une autre ne permet pas de tirer de conclusions solides.

En fait, il y a bien pire : une même étude peut montrer que le médicament est moins bon que le placébo chez les hommes *et* chez les femmes, mais meilleur que le placébo chez les moins de 40 ans *et* chez les plus de 40 ans !

Et encore pire :
Le médicament moins bon que le placébo chez les hommes *et* chez les femmes, mais meilleur que le placébo chez les moins de 40 ans *et* chez les plus de 40 ans peut être, selon les cas, moins bon ou meilleur que le placébo chez les hommes de moins de 40 ans, chez les femmes de moins de 40 ans, etc. Autrement dit, il peut se produire n'importe quoi pour les quatres sous-catégories possibles.
J-P. Delahaye donne l'exemple suivant :

	femmes	hommes	total
jeunes	$36,11\% = \frac{13}{36} > \frac{3}{9} = 33,33\%$	$61,53\% = \frac{8}{13} < \frac{21}{34} = 61,76\%$	$42,85\% = \frac{21}{49} < \frac{24}{43} = 55,81\%$
vieux	$27,27\% = \frac{9}{33} < \frac{4}{14} = 28,57\%$	$66,66\% = \frac{8}{12} > \frac{19}{32} = 59,37\%$	$37,77\% = \frac{17}{45} < \frac{23}{46} = 50\%$
total	$31,88\% = \frac{22}{69} > \frac{7}{23} = 30,43\%$	$64\% = \frac{16}{25} > \frac{40}{66} = 60,60\%$	$40,42\% = \frac{38}{94} < \frac{47}{89} = 52,80\%$

et les mêmes totaux peuvent provenir d'inégalités inverses :

	femmes	hommes	total
jeunes	$33,33\% = \frac{12}{36} < \frac{3}{8} = 37,50\%$	$69,23\% = \frac{9}{13} > \frac{21}{35} = 60\%$	$42,85\% = \frac{21}{49} < \frac{24}{43} = 55,81\%$
vieux	$30,30\% = \frac{10}{33} > \frac{4}{15} = 26,66\%$	$58,33\% = \frac{7}{12} < \frac{19}{31} = 61,29\%$	$37,77\% = \frac{17}{45} < \frac{23}{46} = 50\%$
total	$31,88\% = \frac{22}{69} > \frac{7}{23} = 30,43\%$	$64\% = \frac{16}{25} > \frac{40}{66} = 60,60\%$	$40,42\% = \frac{38}{94} < \frac{47}{89} = 52,80\%$

Les collègues statisticiens avec qui j'ai parlé du Simpson ont qualifié la chose de "vraie s*******e" et franchement c'en est une. Il n'empêche pas de faire de l'estimation statistique de proportions. Il dit juste que ça ne sert à rien d'en faire, pour les comparaisons de qualité, les mesures d'utilité, l'observation de progressions, et quantités d'autres cadres où on est habitué à utiliser des proportions !

2 Un problème de dénominateurs

Le problème des jedis est le même que celui du suivi du taux de réussite des enfants, avec juste la différence qu'il peut y avoir beaucoup de planètes à jedis alors qu'il n'y a que deux catégories d'enfants dans ce débat politique (favorisés et défavorisés).

On devine ce qui peut induire une inversion de Simpson dans le cas d'un investissement massif dans l'enseignement qui cherche, entre autres, à scolariser davantage ou à maintenir en scolarité plus longue des catégorie d'enfants qui n'y étaient pas auparavant : la proportion plus élevée d'enfants défavorisés, ayant un taux de réussite plus faible, peut induire une baisse du taux de réussite global alors même que le taux de réussite augment dans chaque catégorie.

Prenons le cas caricatural où le taux de réussite passe de 60% à 70% chez les plus favorisés et de 20% à 30% chez les autres, mais avec un taux d'enfants favorisés parmi les élèves qui tombe de 75% à 25%, comme ça peut être le cas quand un pays du tiers-monde organise l'enseignement obligatoire (ou en France en 1880...).

— taux de réussite avant : $\dfrac{60}{100}\times\dfrac{3}{4}+\dfrac{20}{100}\times\dfrac{1}{4}=50\%$

— taux de réussite après : $\dfrac{70}{100}\times\dfrac{1}{4}+\dfrac{30}{100}\times\dfrac{3}{4}=40\%$

La démocratisation massive a fait baisser l'indicateur global de réussite alors même que la réussite augmente dans tous les milieux.

Dans cette histoire, la hausse du niveau des élèves favorisés est une bonne choses, la hausse du niveau des élèves défavorisés est aussi une bonne chose et le fait que plus d'enfants défavorisés deviennent des élèves est encore une bonne chose. Mais l'indicateur final, la variation du taux de réussite global, peut être interprété de façon négative par les gens qui ne sont pas prêts à le décortiquer soigneusement.

Accessoirement, cette présentation politique du Simpson est librement adaptée d'une situation qui s'est réellement produite, aux Etats-Unis.

On remarque que ce qui permet l'inversion de Simpson, c'est le déséquilibre dans les effectifs, i.e. les dénominateurs b et b' (nombre d'élèves favorisés avant et nombre d'élèves défavorisés avant) et d et d' (idem après) dans l'écriture en fractions.

Si ce déséquilibre n'existe pas, si $b=b'$ et $d=d'$, l'inversion du sens des inégalités est impossible parce qu'on est ramené à une vraie somme de fractions

$$\frac{a}{b}>\frac{c}{d}\quad\text{et}\quad\frac{a'}{b'}>\frac{c'}{d'}\qquad\text{implique}\qquad\frac{a+a'}{b+b'}=\frac{1}{2}\left(\frac{a}{b}+\frac{a'}{b'}\right)>\frac{c+c'}{d+d'}=\frac{1}{2}\left(\frac{c}{d}+\frac{c'}{d'}\right)$$

On y voit plus clair dans le problème du Simpson en se libérant des nombres entiers. Notons

$$\alpha=\frac{a}{b}>\frac{c}{d}=\gamma\quad\text{et}\quad\alpha'=\frac{a'}{b'}>\frac{c'}{d'}=\gamma'$$

L'inversion

$$\frac{a+a'}{b+b'}=\frac{a}{b}\times\frac{b}{b+b'}+\frac{a'}{b'}\times\frac{b'}{b+b'}\quad<\quad\frac{c+c'}{d+d'}=\frac{c}{d}\times\frac{d}{d+d'}+\frac{c'}{d'}\times\frac{d'}{d+d'}$$

peut se produire quand il existe des coefficients p et q dans $]0;1[$ tels que

$$\alpha\times p+\alpha'\times(1-p)\quad<\quad\gamma\times q+\gamma'\times(1-q)$$

Quand p varie dans $]0;1[$ la combinaison convexe $\alpha\times p+\alpha'\times(1-p)$ parcourt tout l'intervalle entre α et α'. Idem pour $\gamma\times q+\gamma'\times(1-q)$ qui parcourt tout l'intervalle entre γ et γ'.

L'inversion de Simpson est donc a priori possible dès lors que l'intervalle $\left]\frac{a}{b}\ ;\ \frac{a'}{b'}\right[$ ou $\left]\frac{a'}{b'}\ ;\ \frac{a}{b}\right[$ (on ne sait pas dans quel sens il faut l'écrire) est d'intersection non-vide avec l'intervalle $\left]\frac{c}{d}\ ;\ \frac{c'}{d'}\right[$ ou $\left]\frac{c'}{d'}\ ;\ \frac{c}{d}\right[$. En fait, le vrai Simpson avec des nombres entiers ne se produit que quand l'intersection entre les deux intervalles contient deux nombres correspondants à des p et q qui sont eux-mêmes des rapports d'entiers, ce qui la rend quand même assez rare.

3 Heureusement qu'il y a Simpson !

Le troisième problème, celui avec Darwin, est lié aux efforts des biologistes pour comprendre comment, malgré la présence de conflits d'échelle dans tout le monde vivant, nous existons, et le reste du vivant aussi.

Car la loi de Darwin pose un problème : tout trait défavorable à l'individu qui le porte est destiné à devenir de moins en moins fréquent dans la population au fil de la sélection naturelle. Si ce trait est nécessaire à la survie à long terme de la population (c'est la définition même de conflit d'échelle) alors l'espèce semble mal partie... sauf si le vivant est capable d'utiliser une inversion de Simpsons pour faire en sorte que ce qui décroit dans chaque population augmente dans la réunion de toutes les populations.

Et c'est le cas !

L'expérience a été réalisée en vrai, et ça marche, du moins sur des unicellulaires (sur les clans d'hommes de Cro-magnon, il est plus difficile de faire l'expérience, mais le fait que nous soyons toujours là est peut-être un bon indice :-)). Si on reprend la situation proposée par Delahaye :

— une population contient 75% d'"altruistes" et 25% de "non-altruistes" qui profitent du travail des altruistes sans se fatiguer et ont donc des chances de survie plus grandes que les altruistes
— une autre population contient 50% d'altruistes
— et une troisième population est à 25% d'altruistes

Chez les unicellulaires, l'"altruisme" est le fait de produire une substance dont la présence dans le milieu est utile à tous, producteurs et non-producteurs, alors que sa fabrication n'épuise bien sûr que ceux qui la produisent. Dans cet exemple, si les trois populations sont de tailles équivalentes, la proportion globale d'altruistes est de 50%.

Imaginons que la sélection naturelle fait tomber la proportion d'altruistes :

— à 70% dans la première population
— à 45% dans la deuxième population
— et à 20% dans la troisième

La proportion d'altruistes dans l'espèce (réunion des trois populations) va baisser si les trois populations ne changent pas de taille.

Mais puisque (conflit d'échelle) l'altruisme est un trait individuel qui ne profite pas à l'individu mais est utile à la population, il est normal que les variations de taille des trois populations soient liées à la proportion d'altruistes qu'elles contiennent :

— si la première population triple en effectif
— la deuxième population double
— et la troisième stagne

la proportion finale d'altruiste sur l'effectif total est de

$$\frac{3 \times 0,7 + 2 \times 0,45 + 1 \times 0,20}{6} \simeq 53,3\% \quad > \quad 50\%$$

La loi de Darwin est respectée puisque tous les altruistes voient diminuer autour d'eux la proportion de leurs semblables, mais l'espèce peut survivre. Et à l'échelon supérieur, celui où chaque population est un individu, la loi de Darwin est également respectée avec de plus forte chance de survie pour les populations à plus fort taux d'altruistes.

Dans l'affaire des jedis, il est parfaitement possible que la proportion de ceux qui sont du coté clair augmente dans la galaxie tout en diminuant sur chaque planète. Et ce problème est bien une grave question qui engage l'avenir de l'humanité : si l'inversion de Simpson n'existait pas, la vie n'existerait pas non plus !

Petite question pour finir :

Peut-on avoir un véritable double Simpson, c'est-à-dire une situation où le sens des inégalités change quand on sépare la population en deux catégories, puis change de nouveau quand on sépare les deux catégories en 4 sous-catégories ? Delahaye n'en donne pas. Je n'en connais pas. Mais il serait rigolo d'essayer d'en construire.