L'ESSENTIEL

- On croit tous bien maîtriser les statistiques, les pourcentages, les résultats de sondages...
- Pourtant, ils recèlent des pièges et des chausse-trappes qui sont parfois difficiles à repérer et à déjouer.
- Ces nombres sont aussi porteurs de paradoxes qui défient le bon sens et autorisent parfois des interprétations contradictoires... en apparence.
- Averti de ces dangers, on doit examiner avec une grande attention les données chiffrées pour éviter les arnaques.

L'AUTEUR

JEAN-PAUL DELAHAYE
est professeur émérite à l'université de Lille et chercheur au Centre de recherche en informatique, signal et automatique de Lille (Cristal).

Déjouer LES PIÈGES des statistiques

Les statistiques sont parées d'une aura scientifique qui les élève au rang de vérité absolue. Pourtant, les pièges qu'elles tendent sont nombreux. La subtilité de certains d'entre eux nous ravit.

Lucide, Mark Twain écrivait: «Les faits sont têtus, il est plus facile de s'arranger avec les statistiques.» Il ajoutait: «Il y a trois sortes de mensonges: les mensonges, les sacrés mensonges et les statistiques.» Le domaine des calculs et des raisonnements statistiques est plein de pièges, d'évidences trompeuses, et même d'arnaques: soyons sur nos gardes, car l'intuition est souvent mauvaise conseillère.

Dans *Statistiques. Méfiez-vous!*, le mathématicien et psychologue Nicolas Gauvrit, du laboratoire Cognitions humaine et artificielle, recense toutes les erreurs, astuces et idées fausses dont il faut être averti. Nous lui emprunterons plusieurs exemples. Commençons par les pourcentages que nous croyons maîtriser depuis l'école primaire. Vérifions avec quelques questions (ni machine ni crayon ne sont nécessaires).

Question 1. Si le prix de l'essence augmente de 25%, il reviendra à son prix initial après une baisse de: (*a*) 25%, (*b*) 20% ou (*c*) 16,67%?

Question 2. Deux produits A et B ont le même prix. Le prix de A augmente de 12%, puis baisse de 23%. Celui de B baisse de 23%, puis augmente de 12%. Au final, (*a*) A est plus cher que B, (*b*) B est plus cher que A ou (*c*) A et B valent à nouveau le même prix.

Question 3. On augmente votre salaire de 2% par an. En dix ans, il a augmenté de: (*a*) 21,90%, (*b*) 20%, ou (*c*) 18,62%?

Question 4. Un membre du gouvernement assure que «l'augmentation de la dette qui était de 15% l'année dernière a été ramenée à 14% cette année». L'opposition prétend pourtant que «le déficit qui était de 15 milliards d'euros

l'année dernière a encore augmenté cette année de plus d'un milliard d'euros ». L'un des deux ment-il (a) ou est-ce possible (b) ?

Question 5. Dans une université, à l'examen de la licence de biologie, les filles ont mieux réussi que les garçons et à l'examen de la licence de physique, les filles ont, là encore, mieux réussi que les garçons. Pourtant, en regroupant les résultats des deux licences, on découvre que les garçons ont mieux réussi que les filles. Y a-t-il eu une malversation sexiste ?

Pour répondre correctement aux trois premières questions, une perception multiplicative des pourcentages s'impose.

Quand le prix P de l'essence augmente de 25%, plutôt que d'additionner P à $25P/100$, multipliez P par 1,25 $(1 + 0{,}25)$ pour calculer le nouveau prix. Pour qu'il revienne à son prix initial, il faut le multiplier par l'inverse de 1,25, c'est-à-dire 0,80 soit $(1 - 0{,}20)$. Il faut donc baisser le prix de 20% pour le ramener à sa valeur précédente. La bonne réponse de la question 1 est donc (b).

Pour la question 2, la bonne réponse est (c), car le prix de A a été multiplié par 1,12 puis 0,77, ce qui est équivalent à une multiplication par 0,77 puis par 1,12, l'opération étant commutative. Plus généralement, si vous avez des augmentations et des baisses à calculer, l'ordre de calcul n'a pas d'importance pour le résultat final. En revanche, il ne faut pas simplifier en disant que +12% et +23% fait +35%!

Pour la question 3, dix augmentations consécutives de 2% correspondent à une multiplication par $(1{,}02)^{10}$. Sans faire de calcul, on sait que l'augmentation totale est supérieure à 20%, car $(1 + x)^n > 1 + nx$. La bonne réponse est (a).

Les deux dernières questions du test sont plus subtiles. Les deux affirmations de la question 4 peuvent être vraies simultanément. Les 15 milliards d'euros du déficit de l'année dernière correspondent à 15% de la dette initiale (d'il y a deux ans). Celle-ci était donc de 100 milliards d'euros. L'année dernière, la dette est ainsi passée de 100 milliards à 115 milliards. Si, comme l'indique >

la première affirmation, l'augmentation de la dette, c'est-à-dire le déficit, a été de 14%, cette année, l'augmentation a donc atteint 14% de 115 milliards, soit 16,1 milliards. C'est bien conforme à la deuxième affirmation selon laquelle le déficit a augmenté de plus d'un milliard. Les deux affirmations sont parfaitement compatibles: l'augmentation de la dette peut diminuer en pourcentage chaque année en même temps qu'elle s'accroît en valeur absolue.

LES SIMPSON À L'ÉCOLE

La question 5 correspond à une situation remarquable qui froisse l'intuition et peut fausser l'analyse de certains chiffres réels. En effet, il est tout à fait possible que les filles réussissent mieux que les garçons en licence de biologie et mieux en licence de physique, et que, globalement, les garçons réussissent mieux que les filles!

Prenons l'exemple (fictif) d'une situation où 100 candidates filles et 100 candidats garçons se présentent aux examens des licences de biologie ou de physique.

	PHYSIQUE		BIOLOGIE		CUMUL	
	G	F	G	F	G	F
RÉUSSITE	80	10	4	50	84	60
ÉCHEC	10	0	6	40	16	40
TOTAL	90	10	10	90	100	100

Les filles réussissent mieux en physique puisque 100% ont obtenu leur diplôme contre seulement 88% des garçons. De même, en biologie, 55,5% des filles et seulement 40% des garçons réussissent. Pourtant, globalement, 84% des garçons ont un diplôme, contre 60% chez les filles. Comment l'expliquer?

Les filles sont plus nombreuses en licence de biologie et les garçons plus nombreux en licence de physique. Or le taux de réussite est meilleur en physique qu'en biologie. Les filles tentent donc en moyenne un examen plus difficile que les garçons. Ceux-ci ne gagnent, au total, que parce qu'ils optent pour la facilité! Quatre-vingt-dix filles sur 100 tentent la licence de biologie qui a un taux de réussite de 55% alors que 90 garçons sur 100 tentent l'examen de physique qui a un taux de réussite de 90%. Cette situation est nommée paradoxe de Simpson ou effet de Yule-Simpson. Elle a été décrite par Edward Simpson en 1951... et par George Yule en 1903.

Cet effet peut avoir d'ennuyeuses conséquences, comme le montre une étude comparée sur l'efficacité de deux traitements différents contre les calculs rénaux, menée en 1986 par C. Charig, D. Webb, S. Payne et O. Wickham.

Globalement le traitement A avait conduit à 273 succès (78%) sur 350 cas alors que le traitement B sur 350 cas en donnait 289, soit 83%. Le traitement B semblait donc meilleur que A. Pourtant en y regardant de plus près, on constatait que dans le cas des petits calculs rénaux, le traitement A était meilleur que le traitement B: A obtenait 80 succès sur 87, soit 93%, contre 234 succès sur 270, soit 87%, pour le traitement B. Et il en allait de même pour les gros calculs rénaux où A obtenait 73% de réussites (192 succès sur 263) alors que B n'en obtenait que 69% (55 succès sur 80)!

Un autre cas réel de paradoxe de Simpson a été rapporté en 1975 par P. Bickel, E. Hammel et J. O'Connell à propos de l'admission des étudiants dans les diverses facultés de l'université de Berkeley, aux États-Unis. L'analyse globale des données indiquait un biais en faveur des garçons qui, comme dans l'exemple fictif précédent, réussissaient mieux que les filles. Pourtant l'étude détaillée des admissions, faculté par faculté, ne montrait rien de tel: les facultés favorisant les garçons n'étant pas plus nombreuses que celles favorisant les filles.

L'explication du phénomène était semblable à celle de notre exemple: les filles postulaient dans des facultés plus difficiles en moyenne que celles que tentaient les garçons. Doit-on conclure qu'on peut faire dire une chose et son contraire aux statistiques? Risque-t-on toujours de rencontrer de telles situations? La réponse est non: dans l'agrégation des données qui engendre le paradoxe de Simpson, on ne mélange pas des données correspondant à des effectifs égaux pour les sous-cas. Si l'on agrège le résultat des examens de 100 filles passant la biologie, 100 filles passant la physique, 100 garçons passant la biologie, 100 garçons passant la physique, alors le paradoxe de Simpson disparaît. Si l'on veut obtenir des conclusions sensées, l'agrégation des résultats doit respecter certaines règles d'homogénéité.

Notons que le paradoxe de Simpson se généralise en prenant plus de deux catégories d'étudiants. Des données, selon qu'on les regarde d'une façon ou d'une autre, peuvent conduire à des classements exactement inversés. Poursuivons par quelques exemples proposés par Nicolas Gauvrit qui montrent que des paradoxes analogues à celui de Simpson sont plus fréquents qu'on ne l'imagine.

Bienvenue au sein de l'entreprise Marchive dont la situation salariale se résume ainsi:

		OUVRIERS	CADRES
2016	SALAIRE	200 €	2000 €
	EFFECTIF	1000	100
2017	SALAIRE	180 €	1800 €
	EFFECTIF	600	500

Un conflit oppose les syndicats et le patron. Les premiers disent: «Les salaires des ouvriers et ceux des cadres ont baissé cette année de 10%.» Le patron répond: «Nos calculs indiquent que le salaire moyen dans l'entreprise a augmenté. Il est passé de 363,64 euros par semaine

CITATIONS

«La statistique est la première des sciences inexactes.»

Edmond et Jules de Goncourt

«Fêter les anniversaires est bon pour la santé. Les statistiques montrent que les personnes qui en fêtent le plus deviennent les plus vieilles.»

Den Hartog

«Les statistiques, c'est comme le bikini. Ce qu'elles révèlent est suggestif. Ce qu'elles dissimulent est essentiel.»

Aaron Levenstein

«Dans toute statistique, l'inexactitude du nombre est compensée par la précision des décimales.»

Alfred Sauvy

«La mort d'un homme est une tragédie. La mort d'un million d'hommes est une statistique.»

Joseph Staline

«Tout comme certaines sciences occultes, les statistiques possèdent leur propre jargon, volontairement mis au point pour dérouter les non-initiés.»

G. O. Ashley

«Les statistiques ont une particularité majeure: elles ne sont jamais les mêmes selon qu'elles sont avancées par un homme de droite ou par un homme de gauche.»

Jacques Mailhot

LES PARADOXES DE L'ESPÉRANCE DE VIE

Le sort des nouveau-nés se répercute sur l'espérance de vie de tous. Pour calculer celle d'un ensemble d'individus, la méthode la plus naturelle serait d'attendre qu'ils soient tous morts, puis de calculer l'âge moyen de leur mort. Dans la pratique, son calcul utilise une autre méthode, rationnelle, mais susceptible d'engendrer incompréhensions et paradoxes.

Pour calculer l'espérance de vie en 2016, on imagine une population fictive d'individus qui naîtraient tous en 2016 et qui, chaque année de leur future vie, auraient une probabilité de mourir égale à celle constatée en 2016 pour cette tranche d'âge. On imagine par exemple 100000 individus qui naissent en 2016, si 1,3% des enfants ayant entre 0 et 1 an sont morts en 2016, on considère que, dans notre population fictive, il en sera de même. Ensuite, pour les individus de la population fictive qui franchissent leur premier anniversaire (98,7%), on considère qu'ils mourront au cours de leur seconde année dans la même proportion que les enfants qui, en 2016, avaient entre 1 et 2 ans. Et ainsi de suite. L'âge moyen du décès de tous les individus de cette population fictive est par définition l'espérance de vie à la naissance pour l'année 2016. Pour simplifier les calculs, on considère que ceux qui meurent dans leur première année meurent à 0,5 an, que ceux qui meurent dans leur seconde année meurent à 1,5 an, etc.

L'espérance de vie à l'âge de 30 ans se calcule de la même façon en se donnant au départ une population fictive d'individus de 30 ans qui ensuite mourront année après année en se conformant aux chiffres de mortalité constatés en 2016, etc. Ces espérances de vie pour 2016 dépendent donc des conditions de mortalité de l'année 2016 et de nulle autre. Elles ne donnent pas comme on le croit naïvement la durée de vie moyenne des gens vivants en 2016, car les taux de mortalité par âge évolueront dans l'avenir. Plusieurs paradoxes résultent de ce mode de calcul.

Le massacre des innocents

Si un massacre des innocents éradique tous les bébés de moins de 1 an le jour de Noël 2016 sans tuer personne d'autre, alors l'espérance de vie à la naissance en 2016 sera de 0,5 an, car les individus fictifs envisagés par le calcul seront tous morts dès leur première année et seront donc comptabilisés comme vivant 0,5 an.

La bombe atomique

En 2016, l'espérance de vie en France était d'environ 79 ans pour les hommes et 85 pour les femmes. Cela n'aura rien de faux même si en 2017 une bombe atomique tue tous les Français et que la moyenne de la durée de vie des Français ayant vécu en 2016 est donc en réalité d'environ 41 ans!

Le médicament de un an

Le paradoxe du médicament de un an est encore plus frappant. Imaginons qu'un nouveau médicament-miracle empêche totalement de mourir dans l'année qui suit son absorption, sauf ceux de la classe d'âge la plus élevée (laquelle mourra dans l'année), mais qu'il n'ait aucun effet au-delà. Une seule prise du médicament est efficace; imaginons aussi que tous les Français aient pris le médicament le 1er janvier 2016 et donc qu'aucun ne soit mort en 2016 sauf les individus de 114 ans (l'âge du plus vieux des Français, Honorine Rondello). Alors l'espérance de vie en 2016 sera exactement de 114 ans, bien qu'en réalité, la vie de chaque Français aura été prolongée au plus de une année! En effet, avec nos hypothèses, le taux de mortalité par classe d'âge est de 0% quelle que soit la classe d'âge, sauf pour la dernière, et donc tous les individus de la population fictive qui sert au calcul de l'espérance de vie pour l'année 2016 atteindront l'âge maximal constaté en France, âge auquel ils décéderont tous. L'espérance de vie à la naissance ou à n'importe quel âge est donc de 114 ans exactement pour les données de 2016... et redevient normale dès 2017. La méthode de calcul de l'espérance de vie n'a rien d'absurde, mais il est bon de ne pas lui attribuer plus de sens qu'elle n'en a.

> à 916,34 euros, ce qui correspond à une augmentation de 152%. Pourtant, à nouveau personne ne ment. Comment est-ce possible?

Le tableau répond: le salaire hebdomadaire des ouvriers, passant de 200 euros à 180 euros, a baissé de 10%. Celui des cadres, passant de 2000 à 1800 euros, a lui aussi diminué de 10%. Les syndicats ont donc raison d'affirmer que les salaires ont baissé de 10%.

De son côté, le patron ne triche pas! Le salaire versé aux 1100 employés de l'entreprise en 2006 était chaque semaine de 400000 euros ($1000 \times 200 + 100 \times 2000$), soit 363,64 euros par employé. En 2007, le salaire versé aux 1100 employés (l'effectif global est inchangé) s'est élevé à 1008000 euros ($180 \times 600 + 1800 \times 500$), soit 916,34 euros par employé. Le salaire moyen a donc bien augmenté de 152%.

L'arnaque, car il y en a une, est que les effectifs par catégorie ont changé entre 2016 et 2017. Il en résulte que la baisse du salaire dans chaque catégorie est compensée par l'augmentation du nombre des cadres qui sont mieux rétribués. Une baisse de 10% du salaire de chaque catégorie de personnel est parfaitement compatible avec une augmentation du salaire moyen des employés.

Cette situation n'est pas tant fictive que cela, car chaque année l'État français insiste sur les chiffres de la masse salariale des fonctionnaires et sur le salaire moyen d'un fonctionnaire qui évoluent plus favorablement que le point d'indice utilisé pour payer les salaires des fonctionnaires. Ce point d'indice détermine, à grade fixé, le salaire d'un fonctionnaire et bien sûr c'est à lui que les syndicats préfèrent se fier. L'augmentation de l'âge moyen des fonctionnaires – due en particulier à un fort recrutement dans l'enseignement dans les années 1970 et 1980 – entraîne mécaniquement une augmentation du grade moyen (comme dans l'entreprise fictive de l'exemple indiqué plus haut) et conduit donc à une divergence entre les chiffres portant sur le salaire moyen d'un fonctionnaire et ceux portant sur le point d'indice, que ministres et syndicats utilisent de la façon qui les arrange le mieux.

Toujours à propos des fonctionnaires, voyons un autre paradoxe apparent. On peut affirmer sans se contredire que: (a) le salaire moyen dans la fonction publique est supérieur à celui dans le privé; (b) la majorité des fonctionnaires gagneraient plus s'ils partaient dans le privé. L'explication réside à nouveau dans la répartition des emplois du secteur public et du secteur privé: dans le premier, le nombre d'emplois qualifiés est plus important, ce qui a pour effet d'augmenter le salaire moyen du secteur public, sans pour autant contredire qu'à diplôme égal on gagne moins dans le public que dans le privé.

La statistique utilise des indices aux définitions précises qui résument en un seul nombre une masse parfois considérable de chiffres. Ces indices sont inévitables: sans eux, on ne pourrait synthétiser des tableaux de données. Cependant les indices des statisticiens tendent des pièges et peuvent engendrer des résultats totalement absurdes. Soit que celui qui les examine les comprenne imparfaitement, soit qu'il prenne mal en compte des situations particulières.

L'espérance de vie est un indice piège (*voir l'encadré page précédente*). Le seuil de pauvreté tel que le définit l'Insee est tout aussi troublant. Par définition, le seuil de pauvreté dans un pays est la moitié du revenu médian, c'est-à-dire la moitié du revenu X tel qu'il y a autant de gens gagnant plus de X que de gens gagnant moins dans le pays. L'indice de pauvreté est le pourcentage de gens vivant sous le seuil de pauvreté.

Si maintenant quelqu'un vous annonce que dans tel pays 60% des gens vivent sous le seuil de pauvreté, c'est une ânerie. Par définition, ce n'est pas possible: la moitié (à quelques unités près) des gens ont un revenu inférieur au revenu médian, et nécessairement moins de la moitié ont un revenu inférieur à la moitié du revenu médian.

La conséquence dramatique est que même dans un pays où tout le monde mourrait de famine, le nombre de pauvres resterait inférieur à 50%, alors qu'à l'opposé, dans un pays composé uniquement de milliardaires, il pourrait tout à fait y avoir 45% de gens vivant sous le seuil de pauvreté. La pauvreté et la richesse sont des concepts relatifs, certes, mais tout de même!

ÊTRE PAUVRE EN COCAGNE

Nicolas Gauvrit imagine la petite histoire suivante qui montre par l'absurde combien l'utilisation de l'indice de pauvreté est dangereuse. Au pays de Cocagne, une pomme coûte un millième de sol. Un logement en coûte 5. On mange bien pour 0,2 sol. On vit dans le confort pour 100 sols par mois, et comme un nabab pour 200 sols par mois. Il y a dans ce pays deux types de travailleurs: les ouvriers qui ont un revenu de 1000 sols par mois, et les penseurs qui ont un revenu mensuel de 3000 sols. Il y a autant d'ouvriers que de penseurs, exactement. Tous vivent en harmonie et dans une opulence que les pays voisins jalousent. Seule exception à la règle des deux salaires: le président du pays touche, pour sa part, une rémunération de 2800 sols. Le revenu médian est alors de 2800 sols, comme vous le confirmerait tout statisticien. La moitié du revenu médian est donc de 1400 sols, et la moitié de la population (sans compter le président) touche une rétribution inférieure. L'indice de pauvreté est donc de 50% et c'est la valeur la plus élevée possible de cet indice.

Mais un beau jour de mai 2068, le président accepte de baisser ses indemnités, car, dit-il «il n'y a pas de raison pour que je réclame plus qu'un ouvrier! Je suis un travailleur comme les autres». Tenant compte de ses responsabilités, il s'accorde toutefois une petite rallonge et se déclare satisfait de son nouveau revenu de 1400 sols mensuels.

CITATIONS

«Le loto, c'est un impôt sur les gens qui ne comprennent pas les statistiques.»

Anonyme

«Il est statistiquement prouvé que sur dix personnes atteintes de bronchite, une seule va chez son médecin et les neuf autres dans une salle de spectacle.»

Anonyme

«En moyenne chaque personne possède un testicule.»

Anonyme

«À la question: "Faites-vous encore confiance aux sondages?", 64% des Français répondent OUI et 59% répondent NON.»

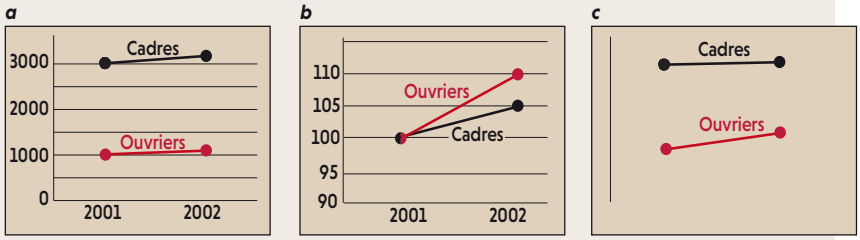
Philippe Geluck

L'ART DU DESSIN

Il y a mille façons de projeter graphiquement des données, chacune permettant de mettre en avant un fait, d'en masquer un autre, ou de donner à croire autre chose que ce que les données indiquent. L'exemple suivant a d'abord été proposé dans la revue *The Economist*. Entre 2001 et 2002, les salaires des cadres et des ouvriers ont évolué ainsi :

	SALAIRES MENSUELS MOYENS	
	2001	2002
OUVRIERS	1000 €	1100 €
CADRES	3000 €	3150 €

En moyenne, les cadres et les ouvriers gagnent plus en 2002 qu'en 2001. En valeurs brutes, l'augmentation des cadres (150 euros) est supérieure à celle des ouvriers (100 euros). Cependant, la situation est inversée en valeur relative : 5 % pour les cadres contre 10 % pour les ouvriers. Ces points de vue différents ont leurs équivalents graphiques, respectivement en *a* et en *b*.
Le commentaire naturel de la représentation *a* est : « Les ouvriers et les cadres ont été augmentés. Les cadres recevaient plus, et reçoivent encore plus que les ouvriers. L'écart entre les salaires empire. »
La représentation *b* s'attache aux augmentations relatives. Elle montre que les cadres ont été moins augmentés en pourcentage que les ouvriers, mais elle ne renseigne pas sur les éventuelles modifications de rémunération en valeur absolue. Cette courbe donne l'impression que les cadres sont moins gâtés que les ouvriers…
Peut-on représenter les évolutions relatives et les différences entre les professions ? Oui, en ayant recours à une échelle logarithmique en ordonnée *(c)*. Avec cette graduation, une même différence de hauteur correspond à un même coefficient multiplicateur. La lecture de ce graphique suggère cette fois : « Les cadres recevaient et reçoivent encore plus que les ouvriers, mais les inégalités se réduisent. »
Ces trois représentations étant toutes parfaitement honnêtes, on peut commenter l'évolution des salaires selon sa préférence.



Le revenu médian en Cocagne passe alors illico à 1400 sols, et le demi-revenu médian à 700 sols, et par un miracle statistique à couper le souffle, l'indice de pauvreté passe aussitôt, en mai 2068, à 0, le minimum possible. »

Créé suite à une polémique sur l'indice de pauvreté, le BIP 40 ou baromètre des inégalités et de la pauvreté est un indicateur synthétique des inégalités et de la pauvreté. Cet indice a été proposé par le Réseau d'alerte sur les inégalités en 2002 et il élimine certains des problèmes mentionnés ci-dessus. Toutefois, il est assez complexe, si bien qu'il est délicat d'en comprendre le sens et qu'on ne peut garantir qu'il évitera toutes les absurdités des indices plus simples.

UNE AFFAIRE DE FAMILLE

Parmi tous les pièges que détaille le livre de Nicolas Gauvrit, l'un d'eux est assez subtil et mérite une attention particulière, nous le nommerons le paradoxe du nombre moyen d'enfants. Une enquête exhaustive menée dans une ville lointaine indique que les familles ayant des enfants de moins de 18 ans se répartissent de la manière suivante : 10% de familles à 1 enfant, 50% à 2 enfants, 30% à 3 enfants, 10% à 4 enfants. Le nombre moyen d'enfants par famille (parmi celles qui ont des enfants) est donc de (10 + 100 + 90 + 40)/100 = 2,4.

Pour contrôler cette statistique, les autorités administratives procèdent à un sondage. On interroge 1000 enfants de moins de 18 ans soigneusement pris au hasard et on leur demande combien il y a d'enfants dans leur famille, eux compris. En faisant la moyenne des réponses, on obtient… 2,68 ! Cela semble absurde. On recommence donc le sondage en interrogeant cette fois 10000 enfants, on trouve maintenant 2,67. Un troisième sondage sur 100000 enfants donne 2,668 à nouveau. Pourquoi cet écart si important avec les 2,4 de la statistique qui prenait en compte toutes les familles ayant des enfants ?

La réponse tient dans le fait qu'en interrogeant des enfants au hasard, vous interrogerez 4 fois plus d'enfants des familles à 4 enfants que vous n'en interrogerez dans les familles à 1 enfant, ce qui fausse la moyenne. S'il y a 1000 familles, il y aura 100 enfants uniques, 1000 enfants appartenant à une famille de 2 enfants, 900 enfants appartenant à une famille de 3 enfants, 400 enfants appartenant à une famille de 4 enfants. Au total, les réponses données par ces 2400 enfants conduiront au résultat de 2,666… enfants par famille.

Les sondages opérés n'évaluent pas le nombre moyen d'enfants d'une famille prise au hasard, mais le nombre moyen d'enfants qu'on trouve dans la famille d'un enfant pris au hasard. « Prendre une famille au hasard » et « Prendre un enfant au hasard » n'est pas la même chose.

Nous le savons depuis Condorcet pour les votes, il est bien difficile de synthétiser un ensemble de nombres en un seul. Nous espérons que les pièges de la statistique, des représentations graphiques, des indices synthétiques, des sondages présentés ici – et ceux que vous trouverez dans le livre de Nicolas Gauvrit – vous aideront à mieux comprendre les vérités cachées derrière l'inquiétant et fluctuant monde des nombres. ■

BIBLIOGRAPHIE

N. GAUVRIT, Statistiques. Méfiez-vous, *Éditions Ellipses, 2014.*

I. EKELAND, Statistiques incroyables, *Pour la Science* n° 334, p. 6, août 2005.

P. BICKEL *ET AL*, Sex Bias in Graduate Admissions: Data from Berkeley, *Science*, vol. 187, n° 4175, pp.398-404, 1975.

G. YULE, Notes on the Theory of Association of Attributes in Statistics, *Biometrika*, vol. 2, n°2, pp.121-134, 1903.