Tangente

Bibliothèque Tangente
L'aventure mathématique

Les Statistiques

et leur décodage

EDITIONS POLE
HS n° 34
ISSN 0987-0806

Bibliothèque Tangente
L'aventure mathématique

Tangente Hors-série n° 34

Les Statistiques

et leur décodage

Sous la direction de Philippe Boulanger

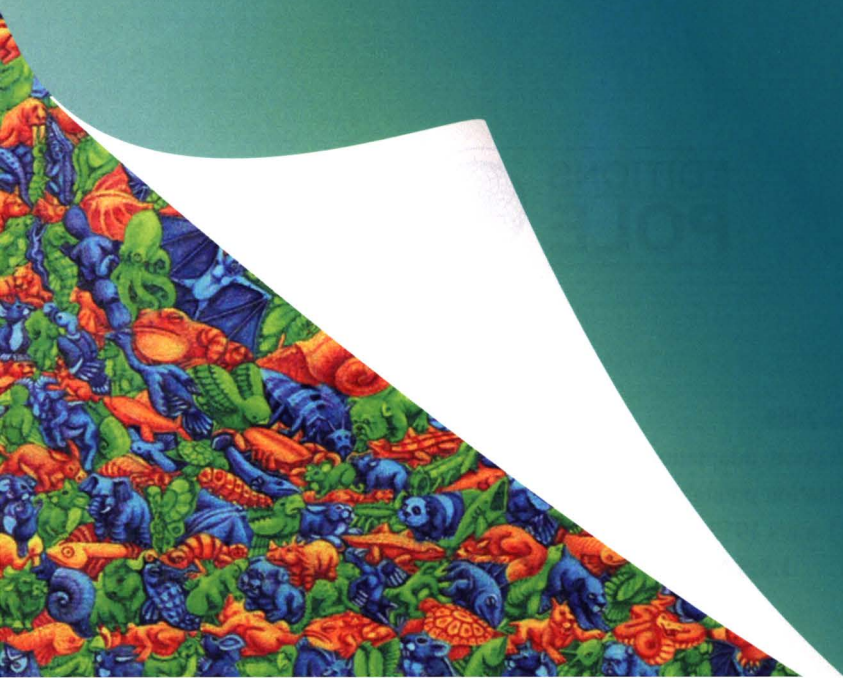
EDITIONS POLE

I.S.B.N. 9782848840963 **I.S.S.N. 0987-0806** **Commission paritaire 1011 K 80883**

Bibliothèque Tangente
L'aventure mathématique

Prochainement dans la Bibliothèque Tangente

Les transformations, de la géométrie à l'art.



EDITIONS POLE

Les Statistiques

Sommaire

(suite du sommaire au verso)

DOSSIER Le traitement des données

La statistique, c'est à la fois le recueil des données et leur assimilation. Le traitement des données, amélioré par des plans d'expériences et la segmentation, est analysé par des méthodes nouvelles.

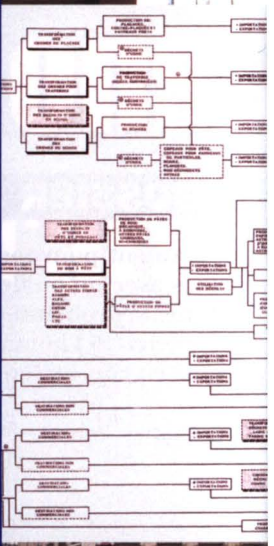
DOSSIER Interpréter les statistiques

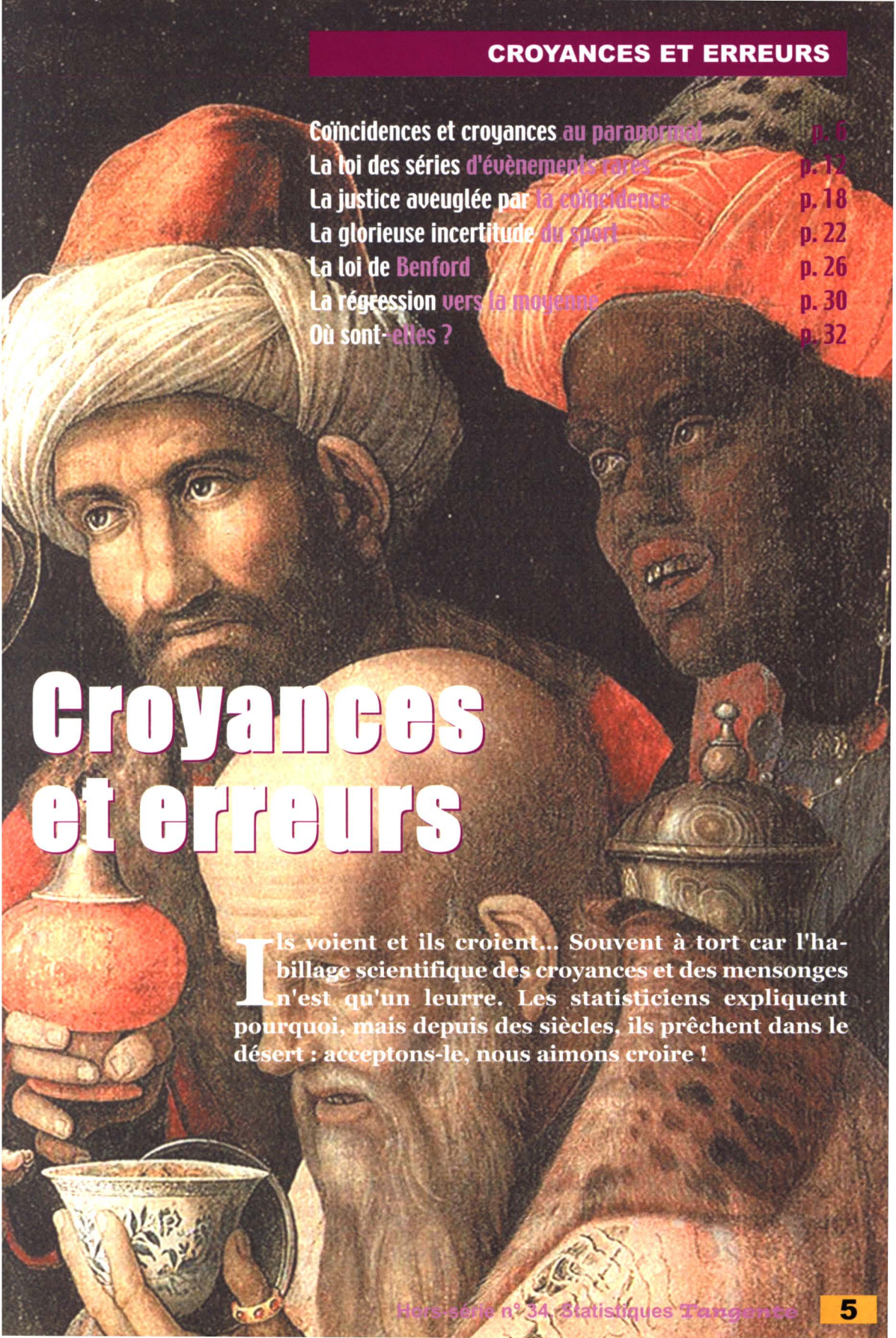
Tout au bord du précipice de l'erreur, les mathématiciens ordonnent leur environnement. Les écueils n'en sont plus quand ils sont reconnus.

DOSSIER Le choix collectif

La démocratie est fondée sur le quantitatif : la majorité a tout le pouvoir. Hélas le résultat du vote dépend grandement de la manière dont on compte les votes. Il est fréquent que la marge de succès d'un vote soit inférieure aux erreurs de dénombrement.

Jeux et problèmes





Croyances et erreurs

Ils voient et ils croient... Souvent à tort car l'habillage scientifique des croyances et des mensonges n'est qu'un leurre. Les statisticiens expliquent pourquoi, mais depuis des siècles, ils prêchent dans le désert : acceptons-le, nous aimons croire !

par N. Gauvrit et J.-P. Krivine

Coïncidences et croyances au paranormal

Nous aimons tous rechercher une causalité alors que le phénomène n'est qu'une coïncidence : c'est le revers de la médaille de notre quête des lois de la nature.

Nous vivons dans un monde surprenant. Se produisent sans cesse des événements étranges, des coïncidences invraisemblables qui finissent par convaincre certains que nous devons chercher une explication mystérieuse à décrypter.

En voyage à l'étranger, je croise mon ami Hervé que je n'avais pas vu depuis des années, et qui se trouve ici au même moment, dans le même hôtel. Ne faut-il pas y voir un signe ? Alors que je repense à Lydie dont je n'ai pas de nouvelles, le téléphone sonne : c'est elle. En effectuant quelques calculs numérologiques sur mon nom, je découvre qu'il est empli de 7, chiffre magique qui revient sans cesse. N'est-ce pas, là encore, un indice intrigant ?

N'y a-t-il pas, autour de nous, bien trop de coïncidences pour qu'on puisse les attribuer au seul hasard ?

Telle est, dans les grandes lignes, la question que se posent de nombreuses personnes tout à fait raisonnables. Et, bien sûr, elles sont tentées de répondre *« Oui, c'est trop étonnant pour qu'on puisse penser qu'il n'y a rien derrière »*. La télépathie expliquerait le coup de fil de Lydie, le destin que je croise Hervé et l'existence d'un sens mathématique caché à la récurrence obsédante du 7.

Le raisonnement que nous faisons spontanément est le suivant : si l'on se réfère au seul hasard, une telle avalanche d'événements rares semble improbable, presque impossible. C'est donc que le hasard n'est pas la seule explication. Comment se fait-il alors que les experts n'aient pas conclu à l'existence de la télépathie, des rêves prémonitoires, à la fiabilité de la numérologie ? Tout simplement parce que la méthode rigoureuse des statisticiens indique que les coïncidences que nous

La croyance au rapport de cause à effet est la superstition.

Ludwig Wittgenstein, Tractatus logico-philosophicus.

[1] Voir par exemple Nicolas Gauvrit (sous presse) *Le hasard, entre mathématiques et psychologie.* Belin/Pour la science.

Coïncidences. Nos représentations du hasard de Gérald Bronner (Vuibert).

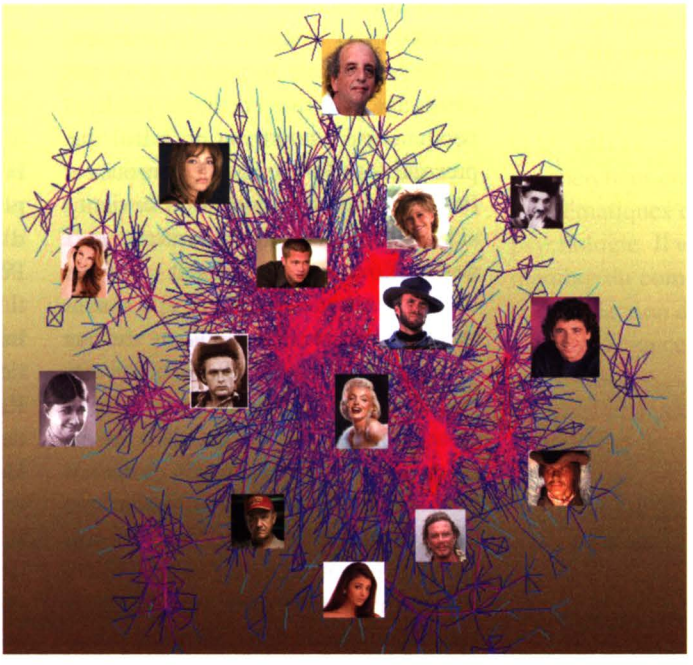
observons sont parfaitement conformes au hasard de la théorie des probabilités… Et l'explication de notre étonnement, aujourd'hui à peu près cernée, est le résultat de recherches que psychologues et mathématiciens ont produites ensemble[1].

Le monde est-il petit ?

Le premier constat dressé par les psychologues qui travaillent sur ces *probabilités subjectives* est que l'être humain est un piètre probabiliste. Nous passons notre temps à estimer des probabilités, à nous demander si tel ou tel événement est probable ou non (réussirai-je cet examen ?). Dès que la situation n'est plus rudimentaire, nos estimations deviennent aventureuses.

« Comme le monde est petit ! » nous exclamons-nous en découvrant que notre voisin partage avec nous un ami commun. Derrière cette expression se cache la conviction que ces rencontres d'amis d'amis sont trop fréquentes pour être le seul fait du hasard. Or, cette estimation intuitive n'est fondée sur rien et serait même contredite par les récents calculs des experts en « graphes ».

Un graphe est, mathématiquement, un ensemble de points (sommets) reliés par des traits (arêtes). Les sommets peuvent par exemple symboliser les personnes, et les traits les relations d'amitié (supposées symétriques). Ces « graphes de connaissances » ont certaines propriétés importantes. D'abord, ils sont localement denses. Autrement dit, on trouve beaucoup d'arêtes entre des points proches. Cela traduit le fait qu'on connaît mieux ses voisins que les gens à l'autre bout du monde. Ensuite, il existe quelques arêtes joignant des points très éloignés. Des graphes de ce type se nomment judicieusement des « petits mondes ». Des expériences par ordinateurs l'ont prouvé : bien que dans de tels graphes chaque sommet soit relié à très peu d'arêtes, il est très souvent facile de passer d'un sommet à un autre par un chemin court. Ainsi est résolue l'énigme de l'ami d'ami. La probabilité que nous imaginons est bien inférieure à la probabilité réelle, et il n'est pas si étonnant que cela de découvrir un ami commun avec son nouveau voisin… mais notre intuition nous trompe.

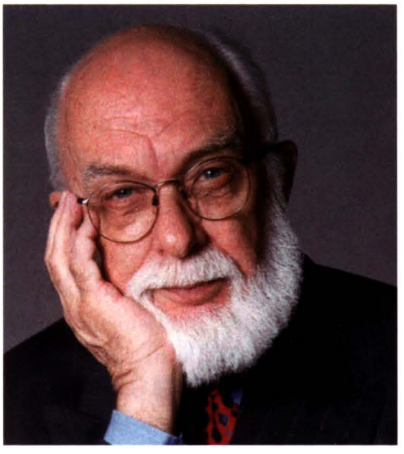


Graphe du « Petit monde holywoodien ». Les arêtes joignent des acteurs ayant… joué ensemble. La distance entre deux acteurs quelconques est en moyenne de 4 arêtes.

Des astres

L'illusionniste et pourfendeur de charlatans James Randi entre dans la salle de cours. Les étudiants sont invités à donner leurs dates de naissance et

quelques autres informations personnelles. Un astrologue, raconte James Randi à la séance suivante, a dressé un portrait détaillé de chacun, en tenant compte des informations fournies. On distribue les lettres cachetées, et chacun lit la description de sa personnalité. Surprise : presque tous, sceptiques et croyants, trouvent le portrait étonnamment ressemblant. James Randi demande alors aux étudiants d'échanger leurs portraits. Ils découvrent, abasourdis, qu'ils avaient tous le même texte ! Le paradoxe qui suscite ici l'étonnement est le suivant : si le portrait n'est pas déterminé par une méthode fiable et personnalisée, comment se fait-il que presque tout le monde s'y retrouve ? Posée autrement, la question devient : si la coïncidence entre le caractère et la description est seulement due au hasard, pourquoi se produit-elle si souvent ? La réponse probabiliste est que la coïncidence a une probabilité très forte... mais l'intuition commune nous susurre que tout ce qui est aléatoire est *équiprobable*. Autrement dit, ça devrait marcher une fois sur deux. Notons que c'est ce même « biais d'équiprobabilité » dont est victime (ou qu'utilise) Elizabeth Teissier dans sa « thèse de sociologie » quand elle défend la scientificité de l'astrologie en ces termes : *« Comme notre expérience nous avait donné des résultats très différents* [d'une chance sur deux] *(environ quatre prévisions sur cinq avérées), nous n'étions pas prête à laisser l'astrologie malmenée » (page 760).*



James Randi a offert un million de dollars à qui prouverait la transmission de pensées.

Des psychologues travaillant sur la perception du hasard se sont en effet rendu compte que, pour beaucoup de gens, une pièce truquée pour tomber un peu plus souvent sur pile que face n'est *pas aléatoire.* Et le résultat d'un lancer ne peut être, selon les personnes testées, attribué au hasard. Plus généralement, un présupposé que nous avons presque tous est que le hasard ne doit favoriser aucune issue par rapport à une autre, ce qui signifie que les résultats aléatoires sont « équiprobables ». C'est cette intuition fausse que les psychologues nomment *biais d'équiprobabilité*. On en atteint facilement les limites. Ainsi, si quelqu'un tire une carte au hasard par exemple, il a moins de chance de tomber sur une tête (valet, dame, roi, as) que sur une autre carte. Pourtant, aveuglés par le biais d'équiprobabilité, bien des gens pensent que, puisque la carte est aléatoire, elle a une chance sur deux d'être une tête. Vous pourrez ainsi étonner une bonne partie du public en « devinant » plus d'une fois sur deux, s'il s'agit d'une tête ou non…

Pour les mathématiciens en revanche, il n'y a aucune contradiction entre le hasard et un déséquilibre des réalisations possibles. Dans le cas de James Randi (son expérience a été reproduite de nombreuses fois, notamment par Henri Broch au laboratoire de zététique de Nice), ce qui fait fonctionner l'affaire est simplement que la description de la personnalité est générale et valorisante, et que chacun aura ainsi tendance à se l'approprier. Cette acceptation excessive d'une description floue est l'*effet Barnum*, qui s'ajoute en l'occurrence au *biais d'équiprobabilité*. Une phrase comme « vous avez tendance à penser d'abord à vous, mais vous savez aussi être très généreux avec vos amis » s'applique à tout le

monde ou presque, mais si nous imaginons qu'elle devrait s'appliquer à une personne sur deux seulement, nous serons étonné de la « coïncidence ». L'article qui suit dans ce numéro est consacré à l'effet Barnum.

La mauvaise probabilité

De biais d'équiprobabilité en effet Barnum, nous sommes peu fiables quand il s'agit de décider si un événement est étonnant (improbable) ou non. Mais il y a pire : nous ne cherchons pas toujours à estimer la bonne probabilité ! Quelle est la probabilité de trouver, dans une assemblée de 50 personnes, deux convives nés le même jour de l'année ? La plupart des gens à qui l'on pose la question font une estimation assez basse, de l'ordre de 15 ou 25%. Portant, il est facile de calculer cette probabilité. Le nombre de cas possible est 365^{50} et le nombre de cas correspondant à 50 dates de naissances *différentes* est de $365 \times \ldots \times 316$. Au total, la probabilité de trouver deux dates égales est donc $p=1-(365 \times \ldots \times 316)/365^{50}$, soit 97 %. Mais le raisonnement intuitif nous conduit à chercher, non la probabilité que deux dates coïncident, mais plutôt la probabilité qu'une date coïncide avec une autre, fixée à l'avance, ce qui est tout à fait différent, et donne naissance au « paradoxe des anniversaires ».

Ce paradoxe des anniversaires est tellement contre-intuitif que la psychogénéalogie, une discipline d'inspiration psychanalytique et numérologique, croit voir dans la coïncidence de certaines dates la preuve d'effets inconscients entre les générations (voir par exemple **http://www.psychogenealogie.name/**). Les arbres généalogiques utilisés par la psychogénéalogie comportent presque toujours une cinquantaine de dates, et parfois plus de cent. Comme nous l'avons vu, la probabilité d'une coïncidence est alors très élevée (plus de 99,99% pour 100 dates). Pourtant, le phénomène parut tellement extra-ordinaire qu'il a donné naissance à une théorie du « syndrome des anniversaires » pour l'expliquer… Et voici comment une erreur bien connue de l'intuition conduit à l'élaboration d'une croyance irrationnelle.

Conclusion

En ce début de XXI^{e} siècle, la croyance dans les phénomènes paranormaux est toujours bien présente. Guérisons miraculeuses, homéopathie, astrologie ou numérologie, et même les tables tournantes, ont encore la faveur d'une bonne partie de nos concitoyens. Toutes les couches de la population sont concernées. Le niveau d'instruction ne prémunit pas contre les croyances non fondées (mais certaines sont mieux admises dans certains milieux)[2]. Toutefois, il serait simpliste de réduire ces croyances à un esprit « prélogique », ou un penchant crédule particulier. En réalité, il existe beaucoup de « bonnes raisons de croire en des idées fausses ou des idées douteuses »[3,4]. Le fait que l'humain soit un mauvais statisticien, et que la statistique elle-même regorge de pièges, contribue à expliquer pourquoi les adeptes sont si nombreux, et aussi, pourquoi le commerce de l'étrange reste florissant.

N. G. & J.-P. K.

[2] Voir par exemple Daniel Boy (2002). Les Français et les parasciences : Vingt ans de mesures. *Revue Française de Sociologie, 43(1)*, 35-45.

[3] Raymond Boudon (1990), *L'art de se persuader des idées douteuses, fragiles ou fausses*, Fayard. Voir aussi la notion de rationalité limitée d'un agent chez H. A. Simon.

[4] *Peut-on tout faire dire aux nombres ?*, Numéro 278 de la revue *Science et pseudo-sciences*, Août 2007.

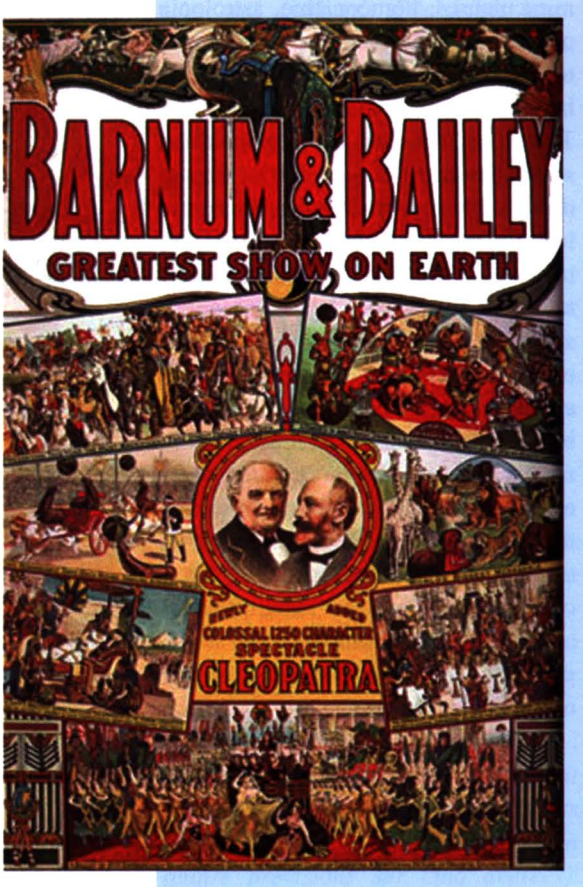
NICOLAS GAUVRIT
est chercheur en mathématiques et psychologie. Il est membre du comité de rédaction de la revue Science et Pseudo-sciences et auteur notamment de Statistiques, Méfiez-vous (Éditions Ellipses, 2007).

JEAN-PAUL KRIVINE
est rédacteur en chef de la revue Science et pseudo-sciences.

L'effet Barnum

Nous avons tous, un jour, lu l'horoscope de notre signe : «Il y a peut-être une vérité dans l'astrologie, avons-nous pensé, je lis un commentaire qui correspond bien à ma personnalité !» Un horoscope typique est le suivant :

«Le soleil brille dans votre ciel astral. Vous avez des qualités que vous n'exploitez pas suffisamment. Vous donnez l'apparence d'être sûr de vous,



mais vous l'êtes beaucoup moins que vous ne pouvez le prétendre. Vous vous targuez d'être un esprit indépendant et vous n'acceptez les arguments d'autrui que preuves à l'appui. Vous redoutez de révéler votre personnalité à autrui. Il vous arrive d'être extraverti, affable et aimable, mais à d'autres moments vous êtes misanthrope et bougon. Certains de vos souhaits tendent à être irréalistes.»

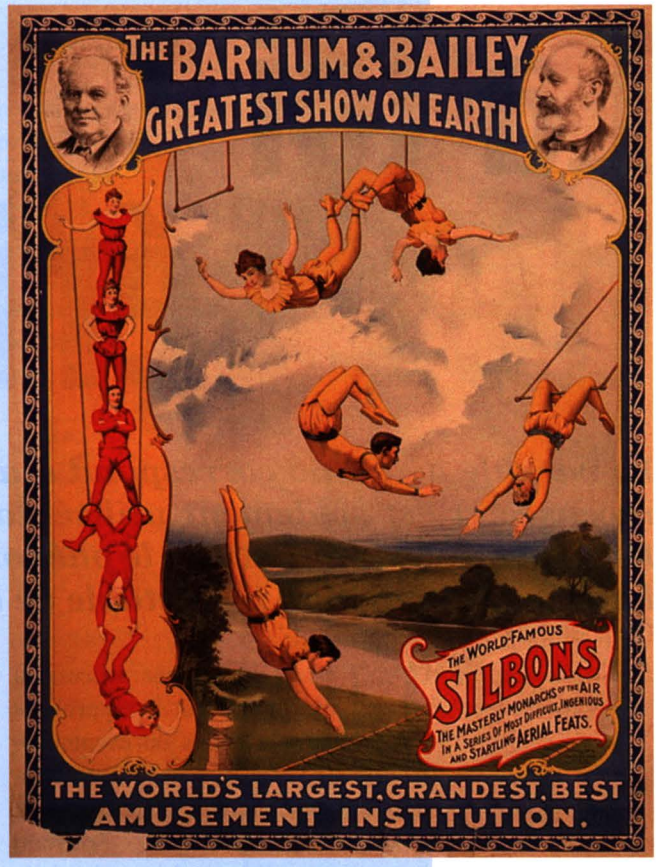
En croyant que ce commentaire dépend de votre signe, vous êtes victime de l'effet Barnum. Cet effet indique que chacun se reconnaît spontanément dans ce qu'il croit être une description de lui-même, alors que les quali-

ficatifs sont suffisamment généraux pour qu'ils s'appliquent à tout le monde.

Le psychologue Forer inventa un type de test en 1948 qui fut repris dans le monde entier et qui démontrait parfaitement l'effet Barnum. Forer faisait passer à ses étudiants un prétendu test de personnalité et quelles que soient leurs réponses, il leur présentait à tous, comme résultat du test, la **même** analyse de type général semblable à celle que vous venez de lire.

Puis Forer demandait à tous les sujets testés de noter la pertinence et la spécificité de l'évaluation de leur personnalité selon une grille de 0 à 5, la note 5 signifiant excellent, 4, signifiant bonne... La moyenne des notes attribuées fut, la première fois que le test fut donné, égale à 4,26 et dans les centaines d'autres essais, la moyenne ne s'écarta jamais beaucoup de 4,2 : la notation pour chacun était entre bonne et excellente !

Le succès de l'astrologie et de beaucoup d'analyses de la personnalité ne sont que des exploitations de l'effet Barnum. La conviction de l'exactitude du test amène les crédules à croire à la validité de la théorie qui donne de tels résultats. Même si l'analyse est juste et spécifique de la personnalité de l'individu et ne constitue pas le canular de Forer, l'effet Barnum jouera pour en faire reconnaître la pertinence, tout comme l'effet placebo agit sur l'efficacité d'un médicament.



Le nom de Barnum est associé à cet effet parce que Barnum, le fondateur du cirque moderne, prétendait qu'«À chaque minute il naît un gogo... ». Il se pourrait qu'à la naissance de ceux qui croient aux horoscopes le soleil de l'intelligence n'ait pas brillé dans leur thème astral.

Philippe Boulanger

par Gilles Dupin

La loi des séries

Le vocable de loi ne correspond à aucune loi de probabilité recensée et est seulement utilisé dans le langage commun. Est-il possible de quantifier le phénomène et d'estimer la probabilité d'apparition inéluctable de la série ?

Une série est une succession d'événements rapprochés dans le temps ou éventuellement dans l'espace dont la concentration est supérieure à celle que l'on observe habituellement. Souvent, l'apparition d'une série même de quelques événements consécutifs peu fréquents est si inhabituelle qu'elle est jugée extraordinaire.

On parle alors de loi des séries avec l'idée populaire que cette succession serait provoquée par une conjoncture propice, par une espèce de main invisible qui augmenterait la probabilité des événements ou les rendrait dépendants, de sorte que l'apparition de la série ne serait plus improbable.

Au niveau collectif, c'est le *fatum*, l'action divine, le malin qui montrerait aux mortels son pouvoir et l'avertirait par une série noire de catastrophes de mieux se comporter. Ce pourrait être aussi la conjoncture des astres qui favorise et rend possible une série d'événements rares.

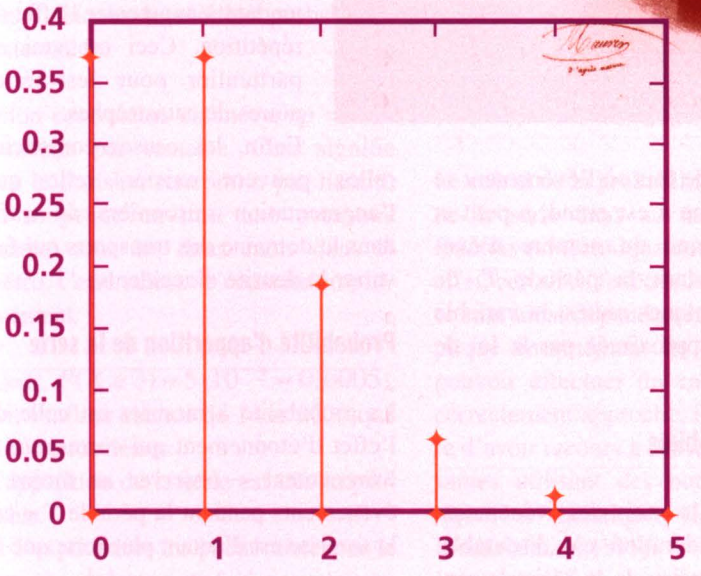
Au niveau individuel, c'est la bonne ou mauvaise passe, la disposition personnelle passagère qui provoque une accumulation d'événements heureux ou malheureux.

Même sans croyance particulière, l'intuition commune de la dépendance entre événements est forte. Ainsi vont les proverbes : « Jamais deux sans trois », « Un malheur n'arrive jamais seul ».

La loi des séries de la signalisation routière…

La loi de Poisson
(Loi des événements rares)

Au cours d'une période T, un événement arrive en moyenne λ fois. On appelle X la variable aléatoire déterminant le nombre de fois où l'événement se produit dans la période T. X prend des valeurs entières : 0, 1, 2, ... Cette variable aléatoire suit sous certaines conditions la loi de probabilité définie par $p(k) = P(X = k) = e^{-\lambda}\lambda^k/k!$ pour tout entier naturel k. Dans la formule, e est la base de l'exponentielle (2,718...), $k!$ est la factorielle de k, λ est un nombre réel strictement positif.



Loi de Poisson P(k) pour un paramètre λ égal à 1. La probabilité est en ordonnée, le nombre d'événements en abscisse. L'espérance d'une loi de Poisson est λ, sa variance est λ, son écart type est $\sqrt{\lambda}$

La question est : la répétition d'événements rares et indépendants est-elle si improbable que l'on doive pour l'expliquer introduire des phénomènes non rationnels ?

Formalisation du problème

Considérons un exemple classique d'une série inattendue de rencontres fortuites. Il est observé que dans une grande métropole comme Paris, un Parisien ne rencontre presque jamais par hasard une personne de son entourage dans la rue ou dans le métro. Et, ô sorcières malicieuses, il rencontre trois personnes connues dans le métro pendant la semaine d'Halloween 2007. Quelle est la probabilité d'apparition de cette série ?

Construisons un modèle probabiliste : Soit un type d'événement ayant la probabilité p, supposée constante, de se produire. Soit n le nombre d'événements (indépendants) susceptibles de survenir pendant une durée T. Soit

X le nombre de fois où l'événement se réalise. Comme n est grand, p petit et np, la moyenne du nombre d'événements pendant la période T, de l'ordre de quelques unités, la variable X peut être approximée par la loi de Poisson P_{np}.

Attention aux biais

Tout d'abord, le choix des événements pris en considération est discutable. L'effet de surprise de la série de rencontres s'est focalisé sur l'observation de rencontres dans le métro, mais il aurait été presque identique si les rencontres s'étaient produites dans un autre lieu. L'effet de surprise n'existe par essence qu'après observation et il vaudrait mieux définir une population d'événements *a priori* qui serait plus large que celle observée *a posteriori* : le nombre n augmenterait alors, ainsi que la probabilité calculée.

De même, la constatation de la série dépend de la finesse d'observation.

En particulier si l'observateur est superstitieux, il observera par le détail de nombreux événements, pourra les relier et en constituera une « série ». Un autre observateur moins minutieux passera à côté de cette « série ».

Dans le même domaine, lorsque l'observation n'est pas individuelle mais portée par la couverture médiatique universelle, l'information peut être biaisée par une transcription médiatique des événements par coups de projecteur tendant à augmenter l'effet de répétition. Ceci est vrai en particulier pour les séries noires de catastrophes.

Enfin, les causes conjoncturelles peuvent exister, telle que l'augmentation saisonnière du trafic dans le domaine des transports qui fait varier la densité d'accidents.

Probabilité d'apparition de la série

La probabilité à mesurer est celle de l'effet d'étonnement qui correspond à l'événement « observer au moins k événements pendant la période T » car la surprise est d'autant plus forte que k est grand : $P(X \geq k) = 1 - \sum_{i=0}^{k-1} P(X = i)$.

$P(X \geq k)$ est alors de valeur faible, conformément à l'intuition.

A contrario, si le produit np était grand, la probabilité serait élevée et la série surviendrait souvent. Elle serait jugée habituelle et normale et l'on ne parlerait pas de loi des séries à son sujet. Tel est le cas d'une série de rencontres dans un village provincial (p n'est pas petit).

Évaluons les valeurs de n et p : la population de personnes susceptibles d'être

rencontrées par hasard dans le métro est de l'ordre de grandeur du nombre d'habitants de la capitale, soit $N = 2$ millions. Notre Parisien connaît ou peut se souvenir de $m = 300$ personnes vivant à Paris. La probabilité que, lorsqu'il croise une personne, celle-ci lui soit connue vaut donc $p = \frac{m}{N} = 1{,}5 \cdot 10^{-4}$.

L'homme croise par le regard 200 personnes chaque jour dans le métro, soit $n = 1000$ personnes par semaine.

Même si le tirage n'est pas exactement avec remise, le taux de sondage $\frac{n}{N}$ est très faible et le modèle probabiliste reste valide.

Selon ces valeurs estimées, $np = 0{,}15$ personne par semaine. Ceci signifie que le Parisien rencontre, par hasard, en moyenne environ une personne connue tous les deux mois dans le métro. Cette valeur se constate empiriquement.

Alors, $P(X \geq 3) \approx 5 \cdot 10^{-4} \approx 0{,}0005$, soit l'infime chance de 1 sur 2000 que la série survienne.

L'apparition de la série est très improbable.

Si l'explication du phénomène par le hasard ne nous satisfait pas, doit-on alors s'en remettre à une explication supra-naturelle ?

Ce calcul est faux

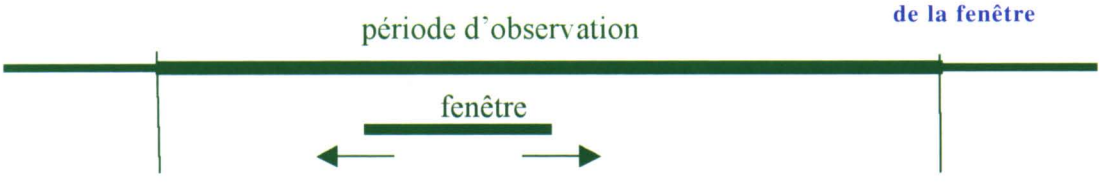
La solution calculée est en réalité une réponse fausse à un problème mal posé. En effet, l'effet d'étonnement n'aurait pas lieu uniquement pendant la semaine d'Halloween 2007 mais pendant n'importe quelle semaine. Il faut non pas s'intéresser à l'apparition de la série au cours de la semaine en question, mais mesurer la probabilité d'observer au moins 3 rencontres dans un intervalle de temps d'une semaine dont l'origine est quelconque dans une période d'observation plus large. Il s'agit donc de faire glisser la fenêtre à l'intérieur de la période d'observation et de prendre en compte toutes les fenêtres possibles.

La probabilité est alors plus élevée.

Statistiques de balayage

Le calcul de la probabilité d'apparition de la série dans au moins une des fenêtres est techniquement difficile. En effet, certaines fenêtres se chevauchent et sont ainsi dépendantes, ce qui complique énormément les calculs. Pour pouvoir effectuer un calcul exact ou correctement approché, il est nécessaire d'avoir recours à des méthodes puissantes utilisant des outils mathématiques et statistiques élaborés (Glaz, Balakrishnan, 1999).

Les probabilistes ont nommé « statistique de balayage » (*Scan statistics*) le calcul du nombre d'événements survenant dans une fenêtre prenant toutes les positions possibles dans une région donnée. Ces statistiques ont été développées par des chercheurs américains



La méfiante prudence envers les séries.

depuis les années 1960. Elles sont en particulier utilisées de plus en plus dans le domaine médical. Par exemple, les chercheurs tentent d'expliquer les causes communes de séries inhabituelles dans le cas du cancer ou de l'apparition de malformations de naissance ; les biologistes cherchent des séries de palindromes dans l'ADN comme indice à l'origine de la réplication du virus. D'une manière analogue, les ingénieurs en télécommunications dimensionnent la capacité des centres téléphoniques afin qu'ils puissent absorber les concentrations temporelles d'appels simultanés ; les experts de contrôle de qualité examinent les séries d'objets défectueux dans une chaîne de production (Langrand, 2005).

Des formules approchées, simples d'utilisation, donnent un résultat précis, notamment la formule de Naus (1982) : cette formule est fonction du rapport entre la durée d'observation et la longueur de la fenêtre notée L, du nombre moyen λ de cas et du nombre k de cas observés durant la durée w.

Dans notre exemple, nous prendrons une période d'observation L de 1 an ou de 5 ans, avec $\lambda = 0{,}15$ et $k = 3$. Sur une période d'observation de 1 an, la probabilité de rencontrer trois personnes connues est égale à 0,069, et sur une période d'observation de 5 ans, 0,302. Ces valeurs sont donc très différentes de celles calculées précédemment : elles ne sont pas suffisamment faibles pour considérer la série comme improbable.

La série a même une probabilité assez élevée de se produire au moins une fois durant cinq ans. Par conséquent, elle peut être expliquée par le seul phénomène de hasard. Point n'est donc besoin d'invoquer des éléments extérieurs !

Les séries noires de catastrophes

Si l'exemple des rencontres, certes réaliste, a été construit pour ses vertus pédagogiques, d'autres cas réels concernant les séries ont fait l'objet de calculs utilisant la statistique de balayage, et tout particulièrement les séries noires de catastrophes qui sont largement exploitées par les médias pour leur caractère spectaculaire et dramatique.

Citons la série noire du mois d'août 2005 dans le monde des transports aériens (Janvresse, De la Rue, 2005) :

Une série noire de 5 catastrophes aériennes a eu lieu à travers le monde sur une courte durée de 22 jours. La moyenne empirique relevée en 10 ans fait état de 14,7 accidents par an de ce type, soit 0,88 accident tous les 22 jours. La densité d'accidents observée près de 6 fois supérieure à la moyenne a donné lieu à débats sur la sécurité aérienne.

Sur la fenêtre de 22 jours la probabilité d'avoir plus de 5 accidents est de 2 millièmes, mais dans une fenêtre quelconque à l'intérieur d'une période d'observation de 1 an la probabilité est de 0,114 et sur 5 ans, 0,466.

Le calcul sur la seule fenêtre pendant laquelle se sont produites les séries

donne une probabilité très faible. Mais, pour un observateur professionnel travaillant sur le long terme, l'apparition de la série est probable, avec environ une chance sur deux de se produire, sans présumer une augmentation de la densité temporelle d'accidents.
De l'observation de chaque série, il n'est alors pas possible d'en déduire une baisse du niveau de sécurité aérienne.

Soyons prudents

Il faut être très prudent dans la définition et la mesure de l'effet d'étonnement consécutif à l'apparition d'événements rares et de séries. Par son aspect subjectif, l'univers des événements rares et son cardinal sont difficiles à cerner. Il en va de même de la mise en relation des événements pour la constitution d'une série.

Mais plus déformant encore est le point de vue de l'observateur concernant la période d'observation. Puisqu'il y a série, c'est qu'il existe une concentration temporelle d'événements supérieure à la moyenne et formant un agrégat parmi un ensemble plus vaste. Il faut donc bien calculer la probabilité sur l'une des fenêtres possibles. Et la probabilité devient alors beaucoup plus élevée pour atteindre des valeurs non négligeables et mesurables à l'échelle du temps.
La loi des séries n'existe pas bien sûr dans la vie courante, sauf dans le cas d'une dépendance claire entre événements. Les calculs mathématiques montrent qu'il est assez probable qu'une série apparaisse « un jour ».
Finalement, c'est peut-être de cette façon qu'il faudrait comprendre l'expression « loi des séries ». À l'intérieur d'une période d'observation plus longue, une série a une probabilité non nulle, voire élevée, d'apparaître.

Alors pourquoi notre intuition est-elle mise à défaut, comme cela se produit assez souvent en probabilité, et refuse-t-elle la constitution d'agrégats par le seul phénomène aléatoire ?
Tout d'abord par nature, l'homme se déplace à un pas cadencé et régulier. Il préfère la fluidité aux à-coups, l'homogénéité à la concentration.
Puis, l'homme vit avec la périodicité des phénomènes célestes naturels qui ponctuent le temps : les rythmes nycthéméral, circadien, lunaire.

De surcroît, notre intuition est leurrée par un monde environnant « faussement uniforme ». Les observations que nous rencontrons dans la vie courante et que nous assimilons à des données aléatoires ne le sont pas totalement. Les arbres d'une forêt s'espacent régulièrement pour partager les ressources naturelles, lumière et eau (Delahaye, 2005). Les voyageurs dans une voiture de métro se répartissent également de manière régulière afin de ne pas être trop proches de leurs voisins et non de manière purement aléatoire. Ces forces d'attraction ou de répulsion concourent à un optimum collectif et à une répartition différente de celle obtenue par l'aléa.
Ainsi, contrairement à l'intuition, les événements survenant de manière aléatoire peuvent se répartir en agrégats plutôt que d'une manière régulière.

G.D.

Gilles Dupin est diplomé de l'École des Mines de Saint-Étienne, chargé de la maîtrise des risques à la RATP.

Bibliographie

• **Janvresse Elise, de la Rue Thierry,** *La loi des séries, hasard ou fatalité ?, Le Pommier Paris, 2007*

par É. Janvresse & T. de la Rue

La justice aveuglée par la coïncidence

Si un événement est rare, la justice pense que ce ne peut être un hasard et condamne quelquefois un innocent. Ces erreurs judiciaires résultent d'une méconnaissance des statistiques.

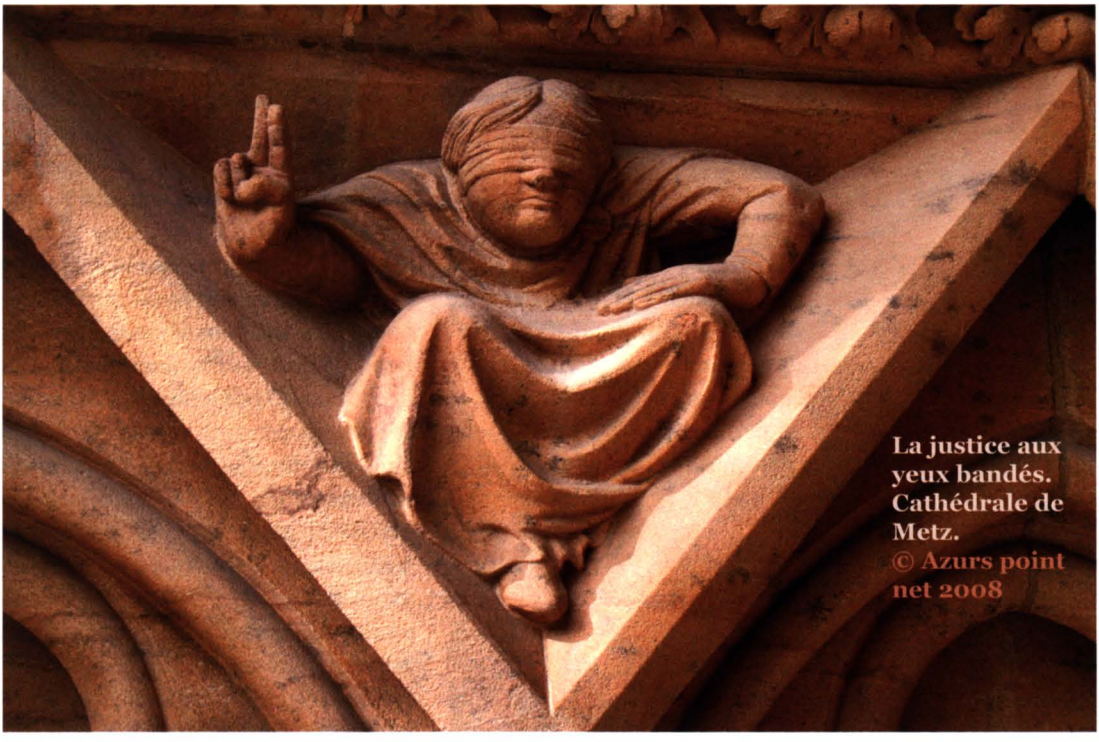
Un malheur n'arrive jamais seul... et quand une série de catastrophes survient, on est tenté de désigner un coupable. Mais si tout cela n'était que pure coïncidence, et que le vrai responsable était le hasard ? Pour le traquer et identifier ses effets, il est nécessaire de faire appel aux meilleurs experts : les statisticiens. Confier ce travail à un amateur peut conduire à des conséquences dramatiques, comme en témoigne l'affaire Sally Clark.

Il entre dans toutes les actions humaines plus de hasard que de décision.
André Gide (1869-1951)

En 1996, Sally et Steve Clark perdent leur fils Christopher, âgé de quelques semaines et victime de la mort subite du nourrisson (MSN). Treize mois plus tard, leur second fils Harry décède dans des circonstances similaires. Comme « la foudre ne frappe jamais deux fois au même endroit », les parents sont soupçonnés d'avoir tué les deux enfants et Sally Clark, inculpée de meurtres, est emprisonnée. Malgré l'absence de preuves matérielles, elle est jugée coupable en 1999 sur la base du calcul avancé par le pédiatre Sir Roy Meadow. Selon lui, la probabilité que deux enfants d'un même couple meurent de la mort subite du nourrisson serait de 1 sur 73 millions.

D'où vient ce chiffre ? D'après les statistiques relevées, le risque d'une mort subite au Royaume-Uni pour un enfant vivant dans une famille aisée, non-fumeur et dont la mère a plus de 26 ans - le cas de la famille Clark - est de 1 sur 8543. Meadow en conclut que le risque de deux morts subites consécutives est de $(1/8543)^2$, soit 1 sur 73 millions. Comme il y a environ 700 000 naissances par an au Royaume-Uni, Meadow souligna qu'une telle coïncidence ne devrait arriver qu'une fois par siècle !

Vous jouez aux dés et votre adversaire sort 5 fois de suite le double six ! Concluez-vous qu'il triche ?



La justice aux yeux bandés. Cathédrale de Metz.

Convaincus par ce chiffre les jurés condamnèrent Sally à la prison à perpétuité.

Erreurs de calculs

Or, le raisonnement qui a conduit Sally Clark en prison est entaché de plusieurs erreurs majeures, immédiatement relevées par la *Royal Statistical Society* [5].
La première objection concerne le chiffre lui-même : le risque retenu est de 1/8543, alors que dans l'ensemble de la population du Royaume-Uni, la probabilité d'une MSN est de 1/1300. Cette probabilité de 1/8543 concerne une sous-population (aisée, non-fumeurs, *etc.*) définie en ne gardant chez les Clark que des caractéristiques diminuant le risque. Mais si on souhaite utiliser l'information spécifique à la famille Clark, on devrait aussi tenir compte de facteurs qui font augmenter le risque, comme le fait que les deux enfants étaient des garçons, statistiquement deux fois plus exposés.

Meadow commet une erreur beaucoup plus grave en élevant 1/8543 au carré : en effet, cela suppose que le risque d'une seconde mort subite est indépendant du fait qu'un premier enfant en ait déjà été victime. Or une analyse plus fine des statistiques montre qu'une mort subite est 5 à 10 fois plus risquée dans une famille ayant déjà subi un tel drame : des facteurs génétiques ou environnementaux prédisposent certaines familles. En reprenant un risque de base de 1/1300, et en tenant compte de la corrélation à l'intérieur d'une même famille, on obtient une probabilité de deux MSN successives de l'ordre de $(1/1300) \times (1/130) = 1/169000$. Ce chiffre semble beaucoup plus proche de la réalité que celui avancé par Meadow, puisque un à deux cas de double MSN sont malheureusement constatés chaque année au Royaume-Uni.
Ce risque de 1 sur 73 millions, sérieusement remis en cause, est en outre trompeur s'il est utilisé seul.

La justice est aveugle, mais doit-elle être sourde aux vérités statistiques ?

L'erreur du procureur

En effet, il est fort possible que le jury soit tombé dans le piège de ce qui est connu sous le terme anglais de *Prosecutor's fallacy* : les jurés ont-ils cru, à tort, à partir des affirmations de Meadow, que la probabilité que Sally Clark soit innocente était de 1 sur 73 millions ? C'est une erreur de raisonnement courante, qui reviendrait à croire que le gagnant du loto de la semaine dernière a très certainement triché, puisqu'il n'avait qu'une chance sur 14 millions de trouver les six bons numéros ! Mais est-il plus probable qu'un individu réussisse à tricher au loto ou qu'il y ait un gagnant de premier rang honnête parmi les millions de joueurs ?

De façon générale, quand on constate un événement surprenant, il faut pour l'expliquer confronter toutes les causes possibles en comparant leurs vraisemblances. Dans l'affaire Sally Clark, est-il plus vraisemblable d'observer un double meurtre d'enfants qu'une double MSN ?

On peut formaliser ce raisonnement en utilisant la formule de Bayes, qui lie la probabilité P d'innocence, sachant qu'on a observé deux décès, à la probabilité d'observer 2 décès sachant qu'aucun acte criminel n'a été commis.

$$P(\text{Innocence}|2\text{ décès}) = P(2\text{ décès} \mid \text{Innocence}) \times P(\text{Innocence})/P(2\text{ décès})$$

Le risque de 1 sur 73 millions avancé par Meadow serait la valeur de P (2 décès | Innocence). On voit clairement par la formule de Bayes que l'on ne peut estimer la probabilité que Sally Clark soit innocente sachant que ses deux enfants sont décédés que si l'on connaît aussi le rapport P(Innocence)/P(2 décès) : il s'agit de la probabilité qu'une mère ne tue pas ses enfants (proche de 1) divisée par la probabilité qu'une famille déplore deux décès successifs (heureusement proche de zéro).

Ce quotient est difficile à estimer de manière précise, mais on peut penser qu'il est grand : il n'y a bien sûr que peu de statistiques sur les doubles meurtres d'enfants, toutefois on peut noter que les caractéristiques de la famille Clark qui faisaient baisser le risque de MSN diminuent aussi la probabilité de meurtres. À partir de la formule de Bayes et des travaux du mathématicien Ray Hill [1], Helen Joyce [3] obtient que la probabilité d'innocence de Sally Clark doit au moins être de 2/3. Évidemment cela ne prouve pas l'innocence de Sally Clark, mais, au Royaume-Uni comme en France, l'accusé est présumé innocent et l'accusation doit prouver sa culpabilité.

Devant la faiblesse des arguments ayant conduit à la condamnation, de nombreuses voix se sont levées et ont poussé à la révision du procès. Sally fut libérée en deuxième appel en 2003, mais ne se remit jamais vraiment de ces épreuves, et décéda en mars 2007.

Le cas de Sally Clark n'est pas isolé. Les magistrats essaient naturellement de baser leur jugement sur des raisonnements qui paraissent solides et objectifs, et il est facile de donner cette apparence à des chiffres issus d'un calcul plus ou moins compliqué. En mars 2003, c'est une infirmière néerlandaise qui en fit la douloureuse expérience. Lucia de Berk fut conduite devant les tribunaux parce que des coïncidences troublantes avaient été observées entre ses horaires de garde et les décès ou incidents majeurs survenus dans les services où elle travaillait [6]. Comme pour Sally Clark, c'est un calcul de probabilité assez fantaisiste qui fut

l'argument majeur de sa condamnation à perpétuité. Un "expert" avança devant le tribunal que la probabilité que le hasard seul soit responsable des coïncidences était de 1 sur 342 millions. Non seulement ce calcul a été invalidé par des statisticiens [4], mais il y a tout lieu de croire qu'à nouveau ce chiffre astronomique ait été mal interprété par le jury.

Faut-il alors bannir les statistiques des tribunaux ? Évidemment non ! Le danger serait que les jurés tirent des conclusions erronées, en sous-estimant la fréquence des coïncidences : notre intuition concernant les effets du hasard est souvent mise en défaut (voir [2]).

Les expertises médicales sont le ressort de spécialistes et il est étonnant que les analyses statistiques ne soient pas confiées à des statisticiens. Il est difficile de reconnaître si une série d'événements est due à un concours de circonstances, ou si elle témoigne d'une cause cachée à identifier. C'est une des questions type auxquelles les statisticiens doivent répondre. La traiter sans une approche rigoureuse mène à de nouvelles catastrophes.

Et si c'était cela, la loi des séries ?

E. J. & T. R.

Cette demoiselle a gagné 50 millions de livres à la loterie : la probabilité qu'elle gagne était-elle si faible qu'il faut en conclure qu'elle a triché ?

Références

[1] Ray Hill : "*Multiple sudden infant deaths, A coincidence or beyond coincidence?*", Paediatric and Perinatal Epidemiology, 18 (2004), 320-326. http://www.cse.salford.ac.uk/staff/RHill/ppe_5601.pdf

[2] Élise Janvresse, Thierry de la Rue. "La loi des séries : hasard ou fatalité ?" Les Petites Pommes du Savoir 98, Éd. Le Pommier, 2007.

[3] Helen Joyce "*Beyond Reasonable Doubt*" http://plus.maths.org/issue21/features/clark/

[4] Ronald Meester, Marieke Collins, Richard Gill, Michiel van Lambalgen : "*On the (ab)use of statistics in the legal case against the nurse Lucia de B.*" http://arxiv.org/abs/math/0607340

[5] Royal Statistical Society http://www.rss.org.uk/

[6] Lucia de Berk sur **Wikipedia** http://en.wikipedia.org/wiki/Lucia_de_Berk

La glorieuse incertitude du sport

... est mesurée. Un tirage aléatoire des résultats des matchs de football et les statistiques du championnat montrent que la qualité des équipes ne prime pas toujours sur le hasard.

L'équipe qui gagne le championnat se targue d'être la meilleure. Est-ce vrai?

Si la première équipe du championnat gagne avec un point de plus que la deuxième, on peut douter de sa supériorité réelle : nombreux sont les cas où, pendant la saison, elle a marqué un but «par un hasard heureux», à la suite d'un rebond inattendu, d'une faute inhabituelle du gardien de but ou d'une erreur d'arbitrage. Ainsi, il s'en est

1. Comparaison des résultats de simulations numériques et du championnat de France

1	73	21	10	7
2	63	18	9	11
3	60	19	3	16
4	58	17	7	14
5	55	15	10	13
6	55	15	10	13
7	53	15	8	15
8	52	15	7	16
9	52	14	10	14
10	52	14	10	14
11	51	13	12	13
12	50	13	11	14
13	50	13	11	14
14	49	13	10	15
15	48	12	12	14
16	48	13	9	16
17	47	11	14	13
18	45	10	15	13
19	41	10	11	17
20	40	11	7	20

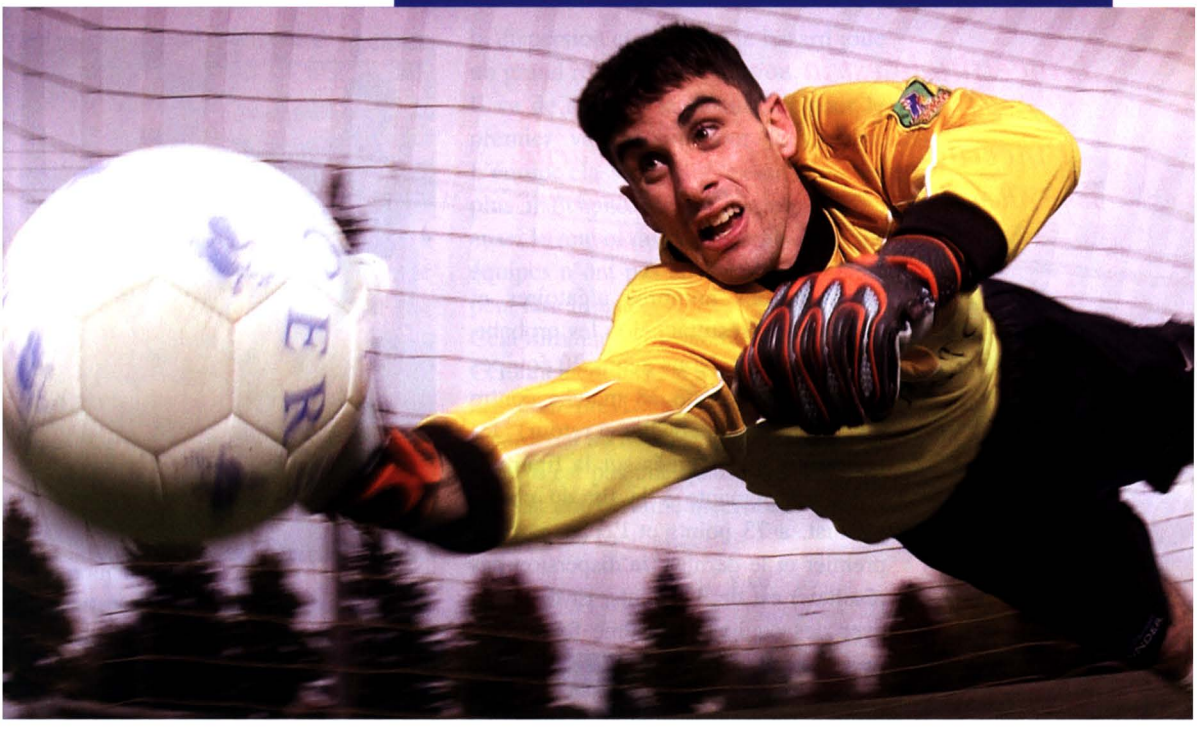
Championnat français simulé
Résultats obtenus par tirage aléatoire
Équipes d'égale force

1	77	23	8	7
2	66	19	9	10
3	64	17	13	8
4	62	17	11	10
5	62	19	5	14
6	56	14	14	10
7	55	15	10	13
8	54	13	15	10
9	53	15	8	15
10	49	13	10	15
11	48	12	12	14
12	48	13	9	16
13	47	11	14	13
14	47	11	14	13
15	46	13	7	18
16	44	11	11	16
17	41	10	11	17
18	39	10	9	19
19	39	10	9	19
20	37	8	13	17

Championnat français simulé
Résultats obtenus par tirage aléatoire
Équipe rouge plus forte

1	Lyon	79	24	7	7
2	Bordeaux	75	22	9	7
3	Marseille	62	17	11	1
4	Nancy	60	15	15	8
5	Saint-Étienne	58	16	10	12
6	Rennes	58	16	1	12
7	Lille	57	13	18	7
8	Nice	55	13	16	9
9	Le Mans	53	14	11	13
10	Lorient	52	12	16	10
11	Caen	51	13	12	13
12	Monaco	47	13	8	17
13	Valenciennes	45	12	9	17
14	Sochaux	44	10	14	14
15	Auxerre	44	12	8	18
16	PSG	43	10	13	15
17	Toulouse	42	9	15	14
18	Lens	40	9	13	16
19	Strasbourg	35	9	8	21
20	Metz	24	5	9	24

Championnat de France 2008



fallu d'un rien que l'équipe classée deuxième ne gagne le championnat. Inversement, si une équipe gagne le championnat avec un grand nombre de points d'avance, elle semble mériter sa victoire en prouvant ainsi qu'elle est la meilleure. Les mathématiques nous aident à distinguer la part de chance et à déterminer la probabilité que l'équipe gagnante est la meilleure.

Imaginons un championnat où toutes les équipes se valent et la probabilité qu'une équipe gagne un match est la même pour toutes les équipes. Les écarts entre les équipes seront déterminés par le hasard. Quel sera le classement final?

Simulation d'un championnat égalitaire

Nous allouons des probabilités à chaque résultat. Il y a environ un match nul tous les quatre matchs dans les championnats professionnels, aussi prendrons-nous la probabilité d'un match nul égale à 1/4. La probabilité qu'un match se termine par une victoire de l'une des équipes est donc égale à 3/4 et comme les équipes sont de même force, chacune d'entre elles a trois chances sur huit de gagner. Un match gagné rapporte trois points, un match nul, un point, et un match perdu aucun.

Nous allons maintenant faire «jouer» les équipes pendant une saison ; pour cela, nous utilisons un programme qui

L'incertitude est la seule certitude qui existe avec le fait que vivre avec l'incertitude est notre seule sécurité.
John Allen Paulos (professeur de mathématiques à l'Université Temple de Philadelphie)

engendre des nombres aléatoires et décidons du résultat selon les probabilités indiquées. Nous avons 20 équipes qui jouent chacune 38 matchs contre les 19 autres (match aller et retour) ; les résultats sont indiqués sur la figure 1. L'équipe gagnante à la fin du championnat, a 73 points et l'écart entre le premier et le dernier (la dispersion des résultats) est de 33 points. Et tout cela avec des équipes de même force ! La moyenne des points, calculée selon les probabilités est de 34($3 \times 3/8 + 1 \times 1/4$), soit 46,7.

Il est patent que des différences de force entre équipes augmentent la dispersion des résultats, mais nous voyons déjà que la contribution du hasard est notable. Lyon, qui a gagné le championnat l'année dernière, avait 79 points et l'écart avec le dernier, Metz, était de 33 points. La question est reposée : l'équipe de Lyon a-t-elle gagnée parce qu'elle était la meilleure équipe, ou était-ce un hasard? Il est clair que le hasard joue un rôle, mais dans quelle mesure ?

La force supérieure peut être inopérante

Ajoutons une équipe dont la probabilité de victoire est supérieure. Nous conservons la probabilité de match nul égale à 1/4 ; l'équipe meilleure, que nous dénommerons Parenlyon, a 3 chances sur 5 de gagner ses matchs, ce qui fait neuf chances sur vingt ($3/5 \times 3/4$) de gagner contre chacune des 18 autres équipes. Dans cette instance, l'espérance mathématique, c'est-à-dire le score moyen à la fin de la saison du score de Parenlyon, est égale à $36/3 \times 9/20 + 1 \times 1/4$), soit 57,6 points contre 47,90 pour les autres équipes. Lors d'une simulation, Parenlyon a fait un meilleur résultat que la moyenne espérée et marqué 64 points, mais, en dépit de cela, n'a pas gagné le championnat : elle a terminé troisième, derrière une équipe ayant totalisé 77 points ! D'autres simulations, avec les mêmes probabilités, donnent des résultats différents, et parfois Parenlyon gagne. Les mathématiques montrent que, souvent, la meilleure équipe ne gagne pas!

Pourquoi ces injustices? À cause des règles du football : au contraire des sports comme le basket ou le rugby, les scores d'un match de football sont très bas, et cela affecte les résultats des rencontres.

Injustice due aux faibles scores

Quand il y a peu de buts, l'équipe la plus faible a quand même des chances de gagner. Supposons que Parenlyon marque en moyenne deux fois plus de

buts que son adversaire Auxaco (moyenne calculée sur les matchs précédents entre les deux équipes). Si le résultat du match est 1–0, score assez fréquent dans le football professionnel, la probabilité que Parenlyon ait marqué le but victorieux est 2/3, mais Auxaco a 1/3 de chance de gagner en marquant le but. L'équipe la moins forte a une bonne chance de gagner.

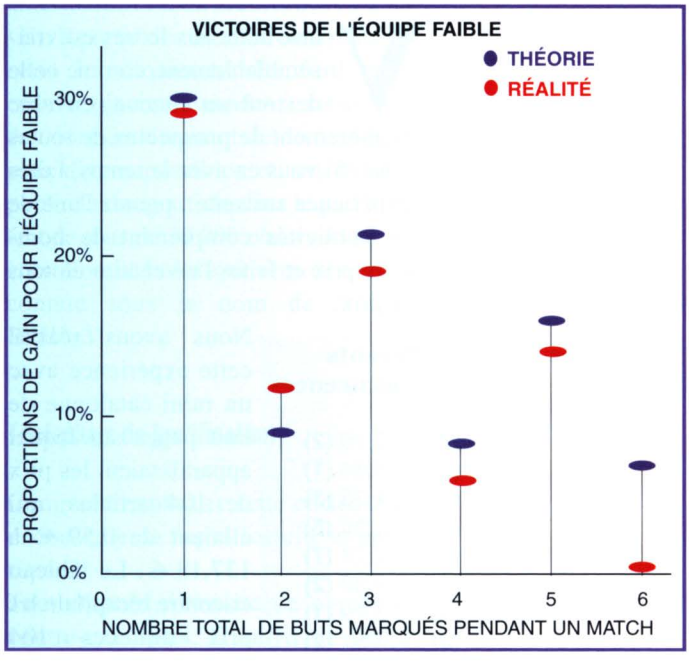
Quand le nombre de buts augmente, les chances de victoire de la moins bonne équipe diminuent. Dans le cas d'une partie où trois buts sont marqués, Auxaco n'a qu'une chance sur quatre de vaincre Parenlyon et 15 chances sur 100 pour une partie à neuf buts (toutefois les parties avec ce grand nombre de buts marqués sont plus rares). Donc, les faibles scores du football augmentent la part du hasard dans les résultats…

Qu'en est-il dans la statistique des vrais championnats ? Nous avons étudié, dans le championnat anglais, les matchs entre les cinq premières équipes et les cinq dernières pendant quatre ans. En moyenne, les meilleures équipes battaient les moins bonnes 7 fois sur 10 (et perdaient 3 fois sur 10). Nous utiliserons ces chiffres pour calculer l'espérance de gain par l'équipe la plus faible et pour la comparer avec les résultats réels des matchs sur les quatre saisons. Les résultats sont représentés sur la figure 2. Les résultats observés sont en bon accord avec la théorie et les équipes les plus faibles gagnent plus souvent les matchs à petit nombre de buts que les matchs à gros scores. Les « creux » pour les matchs à nombre pair de buts, 2, 4, 6…, sont dus au fait que l'écart doit être de deux buts au moins : l'équipe la plus faible gagne donc moins souvent quand le nombre de buts marqués au cours du match est pair. Ainsi les règles du football créent la dispersion et font que le hasard joue un grand rôle dans la décision.

Il va de soi que le calcul fait pour le premier vaut aussi pour le dernier. L'équipe classée dernière peut avoir été plus malchanceuse qu'incompétente : aussi le renvoi des entraîneurs dont les équipes n'ont pas eu de bons résultats peut être injuste.

Cela diminue-t-il l'intérêt du football ? Certainement pas : si le résultat des parties était prévisible, plus personne ne regarderait les matchs : la justice dans le sport n'est pas une qualité primordiale et c'est une certaine incertitude qui fait l'intérêt du football.

J.W.



2. Les victoires de l'équipe faible, en théorie et en réalité.

Bibliographie

- **Wesson John,**
La glorieuse incertitude du football, in Pour la Science, n° 301, novembre 2002.
- **Wesson John,**
La science du football, Belin-Pour la Science, 2004.

par Michel Criton

Une curiosité : la loi de Benford

Contre toute attente, le premier chiffre significatif des nombres utilisés dans la vie courante n'obéit pas à une répartition équiprobable entre les chiffres de 1 à 9.

Votre boîte aux lettres est vraisemblablement, comme celle de tout un chacun, envahie régulièrement de prospectus de toutes sortes. Si vous en avez le temps, faites l'expérience suivante : prenez l'une de ces publicités comprenant de nombreux prix et faites l'inventaire de tous ces prix.

les 46 prix différents (en euros) et leurs occurences

0,59 (1)	0,74 (2)	0,76 (2)
0,89 (3)	1,05 (4)	1,14 (1)
1,20 (8)	1,29 (1)	1,35 (3)
1,52 (7)	1,66 (2)	1,96 (5)
2,05 (1)	2,11 (1)	2,27 (7)
2,40 (2)	2,72 (1)	2,88 (2)
3,03 (5)	3,04 (3)	3,79 (1)
3,94 (2)	4,55 (4)	4,57 (2)
5,32 (1)	5,33 (1)	6,02 (1)
6,08 (1)	6,09 (1)	7,60 (8)
9,13 (1)	10,59 (5)	11,43 (1)
12,04 (1)	12,18 (1)	13,56 (1)
13,70 (1)	14,48 (1)	15,09 (1)
15,22 (2)	17,53 (1)	18,14 (1)
19,66 (1)	22,85 (1)	30,47 (1)
	137,18 (1)	

Nous avons réalisé cette expérience avec un mini-catalogue de huit pages sur lequel apparaissaient les prix de 104 articles, qui allaient de 0,59 € à 137,18 €. Le tableau ci-contre récapitule les prix de ces 104 articles, en donnant, entre parenthèses, l'occurence de chaque prix. Si l'on étudie la fréquence d'apparition du premier chiffre significatif de ces nombres, on obtient le tableau ci-après où la fréquence d'apparition du "1" est pour le moins frappante.

1er chiffre	occurences	fréquence
1	49	47,1 %
2	15	14,4 %
3	12	11,5 %
4	6	5,8 %
5	3	2,9 %
6	3	2,9 %
7	12	11,5 %
8	3	2,9 %
9	1	1 %
Total	109	100 %

Vous penserez peut-être qu'il s'agit là d'un simple hasard et qu'un changement d'unité pourrait faire apparaître la prééminence d'un autre chiffre que le "1". Nous avons donc, toujours à titre d'expérience, étudié les 46 prix encore donnés en francs. L'encadré et le tableau ci-après donnent la répartition des premiers chiffres significatifs non nuls.

La position du "1" est encore privilégiée, puisque celui-ci apparaît comme premier chiffre dans 36,5 % des prix !

les 46 prix différents (en francs) et leurs occurences

3,90 (1)	4.90 (2)	5 (2)
5,90 (3)	0.90 (4)	7,58 (1)
7,90 (8)	0.50 (1)	8,98 (3)
10 (7)	10.90 (2)	12,98 (5)
13,50 (1)	13.90 (1)	14.80 (7)
15,90 (2)	17.90 (1)	18,98 (2)
19,90 (5)	20 (3)	24.90 (1)
25.90 (2)	29,90 (4)	30 (2)
34,90 (1)	35 (1)	39,59 (1)
39.90 (1)	39,95 (1)	49,90 (8)
59,90 (1)	89,50 (5)	75 (1)
79 (1)	79.90 (1)	89 (1)
89.90 (1)	95 (1)	99 (1)
99,90 (2)	115 (1)	119 (1)
129 (1)	149.90 (1)	199,90 (1)
	899.98 (1)	

1er chiffre	occurences	fréquence
1	38	36,5 %
2	10	9,6 %
3	8	7,7 %
4	10	9,6 %
5	6	5,8 %
6	9	8,7 %
7	12	11,5 %
8	7	6,7 %
9	4	3,8 %
Total	104	100 %

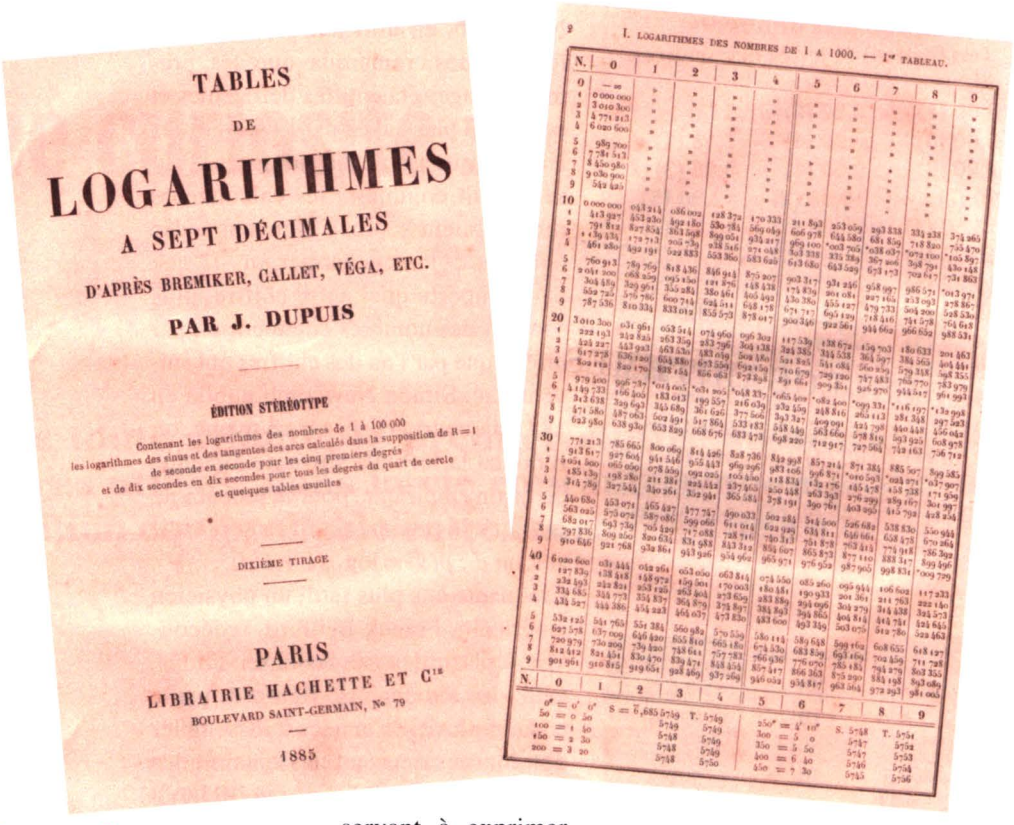
Cette prévalence du "1" sur les autres chiffres a été remarquée il y a plus d'un siècle par l'astronome et mathématicien américain Simon Newcomb.

La vérité dans les tables de logarithmes

À cette époque, tous les calculs devaient se faire "à la main", et les tables de logarithmes, qui permettaient de remplacer les multiplications par des additions et les divisions par des soustractions, étaient d'un usage quotidien pour tous ceux qui avaient beaucoup de longs calculs à effectuer. Newcomb, en utilisant des tables de logarithmes, remarqua que les premières pages étaient très défraîchies et beaucoup plus salies par les doigts des lecteurs que les pages suivantes. Tout se passait comme si les calculateurs rencontraient plus souvent des nombres commençant par un "1" que par n'importe quel autre chiffre, plus souvent des nombres commençant par un "2" que par l'un des chiffres qui suivent, etc. Simon Newcomb publia en 1881 un court article sur ce thème, où il donna la formule suivante, obtenue de façon empirique, pour la probabilité $p(d)$ que le premier chiffre significatif soit un d : $p(d) = \log_{10}(1 + 1/d)$.
Cinquante ans plus tard, un physicien américain, Franck Benford, redécouvrit, à partir des mêmes observations, cette loi étonnante. Benford passe ensuite plusieurs années à rassembler des données émanant de domaines les plus divers possibles et, en 1938. il publie un long article dans lequel il fait état de plus de 20 000 observations qui corroborent la loi qui sera désormais connue sous le nom de "loi de Benford".

Tentatives de justification

Comment expliquer cette curieuse loi découverte de façon fortuite et empirique ?
Tout d'abord, il existe de nombreux contre-exemples. Il suffit de penser aux numéros de téléphone d'une région donnée, ou bien aux numéros de Sécurité Sociale français (ici, c'est le "2" qui l'emporte d'une courte tête, les femmes étant un peu plus nombreuses que les hommes, mais le "1" demeure bien placé !). Mais dans la plupart des cas, les nombres que nous utilisons



TABLES
DE
LOGARITHMES
A SEPT DÉCIMALES
D'APRÈS BREMIKER, CALLET, VÉGA, ETC.
PAR J. DUPUIS
ÉDITION STÉRÉOTYPE
Contenant les logarithmes des nombres de 1 à 100 000
les logarithmes des sinus et des tangentes des arcs calculés dans la supposition de R = 1
de seconde en seconde pour les cinq premiers degrés
et de dix secondes en dix secondes pour tous les degrés du quart de cercle
et quelques tables usuelles
DIXIÈME TIRAGE
PARIS
LIBRAIRIE HACHETTE ET Cie
BOULEVARD SAINT-GERMAIN, No 79
1885

I. LOGARITHMES DES NOMBRES DE 1 A 1000. — 1re TABLEAU.

La page de titre d'une table de logs à 7 décimales de 1885 (c'est en examinant des tables de logarithmes que Benford fit une remarque qui le conduira à établir sa loi) et la première page de la même table.

servant à exprimer des quantités (des nombres d'objets, des poids, des prix, des longueurs, des durées, *etc.*). Ces quantités peuvent prendre des valeurs comprises entre des valeurs extrêmes, avec des fréquences déterminées pour chaque valeur possible.

Si une quantité s'exprime avec des nombres entiers compris entre 1 à 99. tous les nombres étant équiprobables, on aura toutes les chances d'obtenir une répartition "équilibrée" du premier chiffre significatif entre les neuf chiffres de 1 à 9. On aboutirait à la même conclusion si les nombres variaient de 1 à 999 ou de 1 à 9999, ou plus généralement de 1 à 10^n-1. Mais, d'une part, dans une série statistique, tous les nombres ne sont généralement pas équiprobables. D'autre part (et surtout), la valeur maximale d'une série statistique n'est qu'exceptionnellement un nombre de la forme 10^n-1. Or, si la série comprend tous les nombres de 1 à N, des que N n'est pas de cette forme, on n'a plus équirépartition des chiffres de 1 à 9. Prenons un exemple simple pour illustrer notre propos. Sur toute facture d'un garage concernant des travaux effectués sur une automobile figure le kilométrage affiché au compteur de l'automobile. Si l'on effectuait un relevé de ces kilométrages apparaissant sur toutes les factures éditées pendant une année dans un pays ou une région donnée, l'effet Benford serait largement confirmé. Pourquoi ? Là encore, c'est l'étendue de la série statistique qui intervient. La plupart des automobiles ont un "kilométrage de vie" compris entre 100 000 et 200 000 km. Un petit pourcentage

Une page de pub d'un prospectus kitsch où le 1 est prépondérant.

dépasse les 200 000 km et un plus faible encore les 300 000. Dans le "bruit de fond numérique", les "effets Benford" vont toujours dans le même sens et favorisent toujours les premiers chiffres du système décimal : parfois les quatre premiers, d'autre fois les trois premiers ou les deux premiers, d'autres fois encore, seulement le "1". Mais, au bout du compte, tous effets cumulés, on obtient une fréquence décroissante de 1 à 9.

Une application a récemment été trouvée à la loi de Benford : la détection des fraudes. En effet, les fraudeurs, qui fabriquent de toutes pièces de fausses données, ne la respectent généralement pas, sans le savoir bien sûr, et une anomalie à la loi de Benford dans un ensemble de données est souvent l'indice de données falsifiées.

M. C.

par Philippe Boulanger

La régression vers la moyenne

Par le seul fait du hasard, les extrêmes d'une distribution ont tendance à revenir vers la moyenne.

La statistique n'est pas qu'un empilement de formules, c'est aussi l'observation de faits dont le statisticien expert tire des lois. L'une d'elles est la régression vers la moyenne. Grâce à cette loi, nous saurons préciser les causalités vraies et les causalités factices et soulignerons une erreur que nous faisons quotidiennement.

Quelques exemples

L'entraîneur d'une équipe sportive, de football par exemple, a de mauvais résultats et il est limogé. Avec la même équipe, le nouvel entraîneur fait remonter, par ses victoires, l'équipe dans le classement. La conclusion semble s'imposer, il est meilleur que son prédécesseur. Mais curieusement, l'ancien entraîneur, qui s'occupe maintenant d'une nouvelle équipe, a de bien meilleurs résultats qu'avec son ancienne équipe. Qu'en conclure ?

Rester immobile ne sert à rien. Il faut choisir entre progresser ou régresser. Allons donc de l'avant et le sourire aux lèvres.

Baden-Powell (1857-1941)

Voyons le cas d'étudiants qui réussissent leur partiel du mois d'avril. Ils ont de bonnes notes, ce qu'ils attribuent à leur travail, ou à leur intelligence s'ils ont été paresseux. Hélas, ces mêmes étudiants ont, en moyenne, de moins bons résultats aux examens du mois de juin. Les professeurs qui expliquent tout, c'est leur fonction, affirment : « Ces étudiants se sont reposés sur leurs lauriers, alors que ceux qui avaient eu de moins bons résultats au partiel ont fait un effort pour se remettre à niveau, et cela a payé. »

La faille du raisonnement, dans le cas des entraîneurs comme dans celui des élèves, résulte de notre volonté de trouver une cause unique à des phénomènes complexes.

L'entraîneur a pu être sanctionné pour incompétence, alors qu'il n'était que malchanceux : nous savons que ce n'est pas toujours le meilleur qui gagne (voir l'article sur *La glorieuse incertitude du sport*). Les résultats aux examens dépendent de nombreux facteurs, la chance d'avoir révisé la question la veille ou la tristesse d'une impasse, la proximité d'un voisin brillant, la plus ou moins grande mansuétude du correcteur…

Quand il y a de nombreux facteurs indépendants, les résultats tendent à se répartir selon une courbe en cloche dite de Gauss, et les valeurs extrêmes tendent à revenir vers la moyenne, seulement parce que les facteurs du hasard ne se reproduisent pas à l'identique à chaque fois.

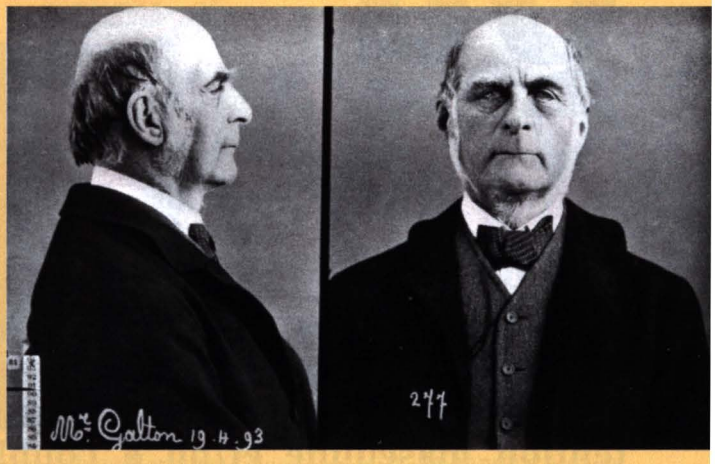
Galton

Le phénomène a été observé par Galton, qui a remarqué que la taille des enfants de parents très grands était inférieure à celle de leurs parents, mais encore supérieure à la moyenne ; en revanche la taille des enfants de parents très petits était plus grande que celle de leurs parents…

Si les variables X et Y ont pour écart type S_X et S_Y et r pour coefficient de corrélation linéaire, alors la pente ρ de la droite de régression, calculée par la méthode des moindres carrés, est égale à $r\,(S_X / S_Y)$ (voir les définitions des termes dans l'article *Robustesse et régressions linéaires robustes*).

Le psychologue Daniel Kahneman relate un *Eureka* ayant trait à la régression vers la moyenne. « Je voulais persuader les instructeurs de pilotage que les compliments étaient plus efficaces que les reproches quant à l'amélioration des performances des apprentis pilotes. À la fin de mon exposé un des instructeurs se leva et relata son expérience inverse : « En de multiples occasions j'ai complimenté les pilotes de la bonne exécution de leur manœuvre et ils ont fait pire la fois suivante. Inversement, j'avais réprimandé les cadets de leur mauvaise prestation et ils ont fait mieux la fois suivante. Donc ne dites pas que les compliments marchent mieux que les remontrances. »

Sir Francis Galton (1822–1911), inventeur de la corrélation et de la régression vers la moyenne.
Question : quelle était la taille probable des parents de Galton ?



J'étais hilare : j'avais compris qu'en moyenne nous sommes punis pour avoir récompensé ceux qui ont réussi et nous sommes récompensés pour avoir puni ceux dont les performances étaient médiocres…

J'ai alors organisé une compétition où chacun lançait une pièce derrière son dos vers une cible et évidemment ceux qui réussissaient bien lors d'une épreuve avaient des résultats plus mauvais lors de l'épreuve suivante, qu'on les complimente ou non. »

Certaines argumentations prônent l'efficacité médicale de *perlimpimpinates* : il arrive souvent que les traitements soient pris en période de crise et la rémission n'a pas été due aux traitements pris mais aux fluctuations naturelles. Je pense là à l'homéopathie et autres sornettes dépourvues de fondement scientifique, donc dépourvues de fondement tout court.

P. B.

Cette lauréate d'un prix de beauté (Miss Wassilia 1984) méritait-elle son titre ? Quel effet la récompense aura-t-elle sur sa carrière et sur la survie des ours polaires ?

par François Lavallou

Où sont-elles ?

Selon les sondages, les hommes prétendent avoir connu bibliquement plus de partenaires que les femmes. On attribue, avec une pointe de malice, cette rupture de symétrie à la prétention masculine et/ou à l'oubli féminin. N'y a-t-il pas d'autres explications ?

Une étude effectuée en juillet 2007 sur la sexualité en Europe et aux États-Unis met en évidence une disparité généralisée entre le nombre de partenaires sexuel(le)s que déclarent avoir eu(e)s les hommes (N_H) et celui annoncé par les femmes (N_F). Ainsi, le rapport N_H/N_F varie de 2,5 pour les Espagnols (N_H=11,5) à un peu moins de 1,3 pour les Américains (N_H=13,8), en passant à 1,6 pour les Français (N_H=13,6).

Les médias réagirent à ce manque de symétrie en estimant que le nombre des relations entre partenaires devait être égaux, en moyenne. Sans contester la qualité du sondage, on s'empressa alors de sortir des poncifs pour expliquer doctement cette différence (diagramme 1). Les hommes sont vantards et les femmes, plus sincères, ont la mémoire sélective : on ne compte que quand on aime !

Sans remettre en cause ce discours

Les relations hommes-femmes sont symbolisées par des flèches, chacune reliant un homme à une femme :

psychologiquement correct, prenons prétexte de cette « étude » pour aborder quelques éléments de réflexion sur la méthodologie statistique.

Biais statistiques

L'intitulé même de la question : « *Combien avez-vous eu de partenaires sexuels au cours de votre vie, même s'il s'agit de partenaires d'un soir ?* » appelle quelques remarques.

Le partenaire à beau être qualifié de sexuel, il est asexué, laissant libre court à toute expérimentation. Pour la simplicité de l'analyse des relations homme-femme, nous éliminerons de notre échantillonnage (diagramme 2) certains éléments tels les zoophiles, pour la non-participation de leurs partenaires au sondage, et les populations homosexuelles, car pour un modèle uniforme, l'homo gène.

Admirons au passage la mémoire *ex abrupto* des plus vaillants des sondés et regrettons qu'une vie semble se terminer au moment du sondage !

Bien sûr, les spécialistes pondèrent les réponses pour qu'elles correspondent à celles d'un échantillon représentatif. Mais que penser du modèle de population choisi quand les sondés, dans notre cas, ont entre 16 et 64 ans : la vie, même sexuelle, est-elle finie après 2^6 ans ? À une époque où l'espérance de vie des femmes dépasse de dix ans celle des hommes et le poids économique du troisième âge augmente, on

le nombre moyen de relations sexuelles par homme est égal au nombre de flèches divisé par le nombre d'hommes, le nombre moyen de relations sexuelles par femme est égal au même nombre de flèches divisé par le nombre de femmes. Comme il y a autant de femmes que d'hommes ces deux nombres moyens sont égaux.

En ce qu'ils ont de commun, les deux sexes sont égaux ; en ce qu'ils ont de différent, ils ne sont pas comparables.

Jean-Jacques Rousseau (1712-1778)

crée un biais indéniable en ignorant gigolos et gérontophiles !

Une cause d'erreur naturelle est l'hypothèse implicite d'uniformité de la population. Quel que soit le traitement, une population trop hétérogène ne pourra être représentée par un faible nombre de sondés. Dans le cas d'une ruche humaine avec une « reine » plus ou moins nymphomane, le diagramme 3, représentatif de la population d'une ruche, peut donner, pour l'énorme proportion de 50% de sondés féminins, $N_F=2$ ou $N_F=0$ alors que nous aurons toujours $N_H=1$.

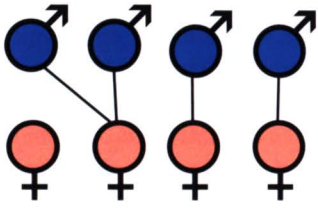
Autre explication possible, la non-fermeture du système. Imaginons que nombre de sondés aient pratiqué de fructueuses expériences linguistiques avec des jeunes filles aux pair(e)s, Vénus de pays où le soleil, lui, ne se couche pas en été. Ce flux sortant de partenaires déséquilibre le bilan.

Ignorant les nymphomanes, les péripatéticiennes (improbables disciples d'Aristote), les veuves argentées donc joyeuses ou les étudiantes samoyèdes, comment peut-on quantifier l'erreur, que l'on ne présente d'ailleurs pas, d'un sondage qui prétend être une mesure ?

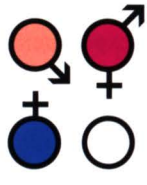
Méfions-nous donc des solutions « évidentes » : *nous sommes souvent abusés par les mythes, et il est difficile de le constater.*

F. L.

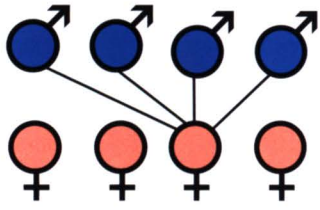
Diagrammes



1 : *Cas quelconque* ($N_H=N_F=1$)



2 : *Cas de populations exclues*



3 : *Cas nymphomane* ($N_H=N_F=1$)

Recueil des données

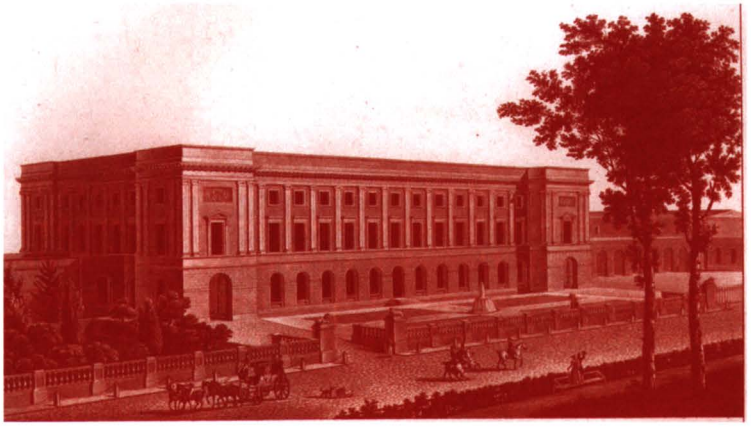
Comment pouvons-nous résumer un ensemble de données avec un ou deux chiffres ? Dès le début de la statistique, le problème s'est posé avec « l'homme moyen » de Quételet. La représentativité des chiffres est une donnée majeure en économie, et si l'homme moyen existait, il faudrait le mettre sous cloche, au Pavillon de Sèvres.

par Norbert Verdier

Quételet
ou l'homme moyen

Pour une population donnée comment trouver un « bon » représentant ? La question est en filigrane des travaux d'Adolphe Quételet : « Un homme moyen » peut-il représenter un groupe d'hommes ? Petit voyage autour du parcours d'un homme « au-dessus de la moyenne » ...

Lambert Adolphe Jacques Quételet est né le 22 février 1796 à Gand. La Belgique d'alors n'est pas la Belgique d'aujourd'hui. Gand est sous administration française. Le père de Quételet est un Picard installé depuis peu dans cette ville. En 1814-1815, les puissances européennes décident de rattacher la Belgique aux Pays-Bas sous la souveraineté du roi Guillaume 1er d'Orange. Le 4 octobre 1830, la Belgique proclame son « indépendance ». La France fait un geste en refusant d'octroyer la couronne (proposée par Bruxelles) à l'un des fils de Louis-Philippe. L'Angleterre impose, en 1831, son candidat Léopold de Saxe-Cobourg. Devenu Léopold Ier (1790-1865), il épouse Louise-Marie d'Orléans, fille de Louis-Philippe. Pendant que la Belgique se structure, Quételet achève ses études. En 1819, il soutient sa thèse (de géométrie) et entre dès 1820 à l'Académie royale de Bruxelles, dont il devient secrétaire perpétuel en 1834. À Bruxelles, il enseigne à l'Athénée les mathématiques élémentaires, la physique expérimentale, l'astronomie, les probabilités, le calcul dif-



Académie Royale de Bruxelles

férentiel et intégral, la géométrie analytique supérieure en mettant au point avec son ami Germinal Dandelin des théorèmes sur les coniques qu'il est courant de désigner aujourd'hui sous l'appellation de « théorèmes belges ». En 1824, Quételet épouse Cécile Virginie Curtet, la fille d'un médecin français établi à Bruxelles. Institutionnellement et socialement installé, Quételet développe son réseau de sociabilité savante, un réseau qui dépasse rapidement les frontières de Bruxelles et du royaume des Pays-Bas.

Sa correspondance

En 1825, il fonde avec son maître J.G. Garnier la *Correspondance mathématique et physique,* le premier périodique consacré aux mathématiques et à la physique au royaume des Pays-Bas.

Le timbre belge en hommage à Quételet (1796 -1874)

ÉPOQUES (1)		RÉSULTATS.
Des naissances.	*De la conception.*	
Janvier.....	Avril.....	1,0403
Février.....	Mai.....	1,1570
Mars.....	Juin.....	1,0991
Avril.....	Juillet.....	1,0790
Mai.....	Août.....	0,9893
Juin.....	Septembre.....	0,9559
Juillet.....	Octobre.....	0,9012
Août.....	Novembre.....	0,9033
Septembre.....	Décembre.....	0,9401
Octobre.....	Janvier.....	0,9492
Novembre.....	Février.....	0,9679
Décembre.....	Mars.....	1,0175

Le premier article de statistique de Quételet (1825) : l'homme moyen conçoit en mai...

La médiocrité est la moyenne à son plus bas niveau.

Albert Brie, Sociologue canadien (1925-)

Dans le *prospectus* de lancement, les auteurs justifient leur démarche éditoriale. Ils partent du constat que dans le royaume n'existe aucun « recueil consacré aux sciences Mathématiques et Physiques », un recueil « qui permette à ceux qui les cultivent, d'établir entre eux un commerce scientifique, et qui, entre autres avantages, offre celui de garantir à chacun la propriété et la prompte publicité des résultats de ses recherches. »

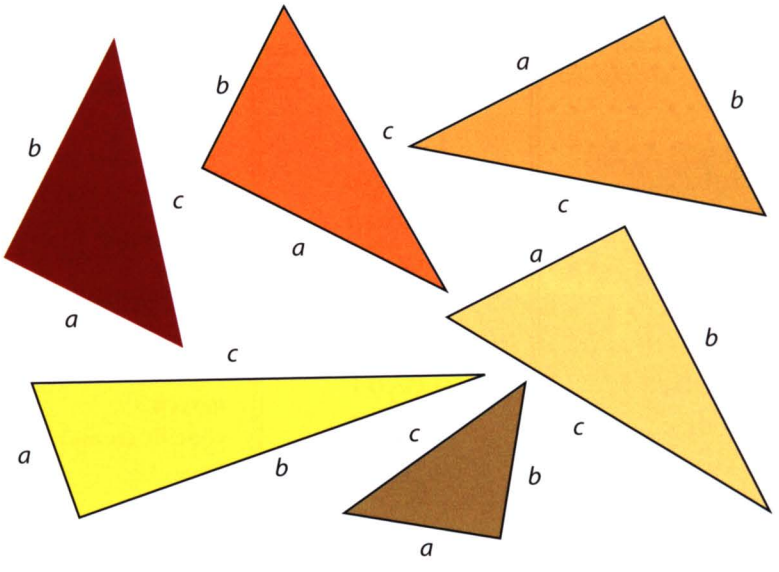
Dès le premier tome, Garnier annonce dans une note de bas de page « l'importance des recherches de notre collaborateur M. Quételet. » Une rubrique statistique est lancée. Quételet l'anime en publiant des études sur la natalité, la mortalité, *etc.* Certains détracteurs réagissent. Passe encore d'étudier des collections d'objets, mais manipuler des hommes – créatures de Dieu – comme les vulgaires points d'un plan ? La rubrique statistique est intégrée dans la division des mathématiques appliquées et résiste à la séparation entre les deux fondateurs. Garnier et Quételet n'ont pas les mêmes visions sur leur projet éditorial commun : Garnier veut que la *Correspondance* soit un journal pour les élèves du royaume des Pays-Bas ; Quételet veut en faire un recueil savant de dimension internationale. Le duo rompt et la *Correspondance* devient la *Correspondance de Quételet*. Malgré les difficultés éditoriales, elle tient jusqu'à la fin des années trente.

Et son homme moyen

En relevant les mensurations de conscrits français et en analysant celles de 5.000 soldats écossais reprises en 1817 dans la revue *Edinburgh Medical Journal*, Adolphe Quételet applique les lois des probabilités aux données biométriques de l'homme, comme le poids, la taille, le périmètre thoracique, devenant ainsi un des fondateurs de l'anthropométrie et de la biostatistique. En correspondance avec les savants de son temps (Laplace, Fourier, Poisson, *etc.*) qu'il avait rencontrés à l'Observatoire de Paris, Quételet développe ses conceptions des statistiques dans sa *Correspondance mathématique et physique* et synthétise ses résultats dans son ouvrage de 1835 intitulé *Sur l'homme et le développement de ses facultés ; Essai d'une physique sociale*. En 1869 paraît *Physique sociale*, une seconde édition remaniée ; elle constitue en fait une réédition de la plupart des œuvres statistiques de Quételet.

Quételet élabore, tout au long de son œuvre statistique, une

Le triangle rectangle moyen n'est pas rectangle...



véritable « mécanique sociale » centrée autour de la notion d' « homme moyen » : l'homme moyen d'une population donnée est, selon Quételet, un individu dont les caractéristiques physiologiques sont chacune égale à la moyenne des caractéristiques physiologiques des autres individus de la population.

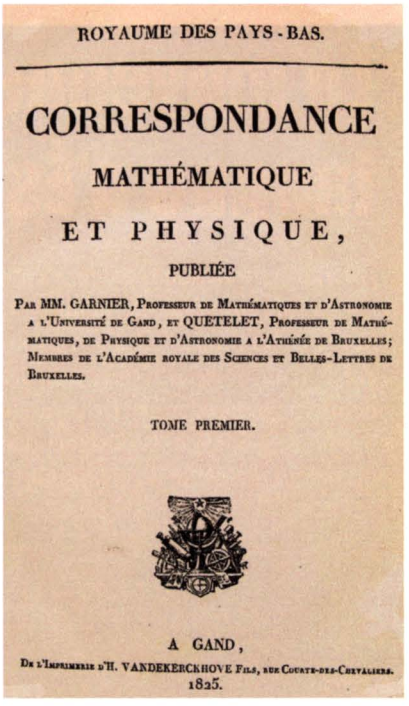
Pour mieux comprendre la notion et les oppositions, comme celles de Cournot, à Quételet qui en résulteront, oublions les conscrits français et écossais pour nous focaliser sur une population de n triangles rectangles. Chaque triangle de longueurs respectives a_i, b_i et c_i vérifie la sacro-sainte relation de Pythagore : $a_i^2 = b_i^2 + c_i^2$.

Le « triangle moyen de Quételet » a pour longueurs $A = \frac{1}{n}\sum_{i=1}^{n} a_i$,

$B = \frac{1}{n}\sum_{i=1}^{n} b_i$ et $C = \frac{1}{n}\sum_{i=1}^{n} c_i$ et,

puisque le carré de la somme n'est pas égal à la somme des carrés, le « triangle moyen » n'est même pas rectangle ! Un drôle de représentant ! Maurice Fréchet a montré au XX^{e} siècle que cette fabrication artificielle d'un individu moyen (qui n'existe pas !) est concevable si chacune des quantités utilisées pour le calcul est peu dispersée autour de sa moyenne (comme c'était le cas pour la taille et le poids des conscrits de Quételet), selon la distribution statistique « en cloche » connue sous le nom de loi de Laplace-Gauss.

D'autres polémiques jaillirent des travaux de Quételet : appliquées initialement aux qualités physiques, leurs applications aux qualités morales et intellectuelles n'étaient-elles pas dangereuses ? Ainsi, un problème social se pose : si le nombre annuel d'homicides


ROYAUME DES PAYS-BAS.

CORRESPONDANCE

MATHÉMATIQUE

ET PHYSIQUE,

PUBLIÉE

PAR MM. GARNIER, PROFESSEUR DE MATHÉMATIQUES ET D'ASTRONOMIE A L'UNIVERSITÉ DE GAND, ET QUETELET, PROFESSEUR DE MATHÉMATIQUES, DE PHYSIQUE ET D'ASTRONOMIE A L'ATHÉNÉE DE BRUXELLES; MEMBRES DE L'ACADÉMIE ROYALE DES SCIENCES ET BELLES-LETTRES DE BRUXELLES.

TOME PREMIER.

A GAND,

DE L'IMPRIMERIE D'H. VANDEKERCKHOVE FILS, RUE COURTE-DES-CHEVALIERS.

1825.

Le premier tome de la Correspondance mathématique et physique en 1825.

en Belgique est constant, quelle est la responsabilité d'un individu et de la collectivité ? N'est-ce pas l'organisation sociale qui est fautive ? Marx utilisa une partie des travaux de Quételet pour développer ses conceptions politiques. Retenons de Quételet ses tentatives d'utilisation des statistiques pour mieux appréhender les phénomènes sociaux.

N.V.

Bibliographie

• **Elkhadem, Hossam.,** *« Histoire de la correspondance mathématique et physique d'après les lettres de Jean-Guillaume Garnier et Adolphe Quételet », Bulletin de la classe des lettres et des sciences morales et politiques, 5 ème série, Tome LXIV, t. 10-11, (1978), 316-366.*

Sitographie

• **Cf.** *http://www.statbel.fgov.be/info/quetelet_fr.asp#3*

La moyenne, langue d'Ésope

La moyenne prétend résumer en un chiffre un ensemble de données, ce qui est ambitieux. L'espoir qu'il existait un homme moyen est déçu depuis Quételet, et l'expression « Français moyen » désigne quelqu'un qui n'a guère de qualités apparentes. Il faudrait préciser en quoi cet homme est moyen, sa taille, son poids, sa vitesse sur 100 mètres, son talent aux échecs…

La moyenne d'un ensemble d'éléments généralement utilisée est la moyenne arithmétique (la somme des valeurs des éléments divisée par le nombre des éléments), mais il en existe d'autres, notamment la médiane et le mode.

La médiane est la valeur pour laquelle il y a autant d'éléments de valeur inférieure que d'éléments de valeur supérieure. Elle donne parfois une meilleure indication, comme nous allons le voir. Supposons que nous ayons 11 timbres, les neuf premiers valant respectivement 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 euros et un

La même montagne de vos vacances sous un ciel clair (le jour de votre départ) et sous un ciel plombé (le jour où vous voulez l'escalader)…

timbre rarissime qui vaut 1000 euros. La moyenne est 1055/11, soit environ 95,9 euros. Il est évident que si vous tirez un timbre au hasard de cet ensemble, vous avez plus de chances de tomber sur un des dix premiers timbres de valeur inférieure à 10 euros et une chance sur 10 de titrer un timbre de plus de 95,9 euros… Toutefois la médiane, de valeur 5, semble mieux représenter la valeur « moyenne ».

Le nombre qui représente le mieux l'ensemble des valeurs est sujet à discussion, mais il est certain que la moyenne arithmétique est trop influencée par les extrêmes. Les hommes politiques en jouent. Le mode est la valeur la plus fréquente, mais souvent il n'existe pas… comme ici.

La confusion entre moyenne et médiane est criante quand on s'aperçoit avec tristesse que la moitié des élèves d'une classe ont des capacités inférieures à la moyenne. On devrait dire inférieure à la médiane. Cependant, si cette constatation avec la médiane est mathématiquement correcte, elle n'en est pas moins triste.

Être ou avoir été…

La moyenne peut ne pas représenter une situation vécue. Lors des dernières vacances à la montagne, j'avais remarqué qu'il pleuvait ou que le plafond était très bas, comme on dit, et que le ciel était couvert de nuages : la couverture nuageuse était totale, environ un jour sur deux (été pourri). Il fallait s'en accommoder : on ne peut pas être et avoir été. Heureusement un jour sur deux aussi (quand la météo avait dit le contraire), il a fait très beau et il n'y avait quasiment aucun nuage dans le ciel. Donc, en moyenne, la couverture nuageuse était de 50%, mais cela ne représentait aucune situation concrète.

L'histoire des deux statisticiens qui tirent sur un ennemi, l'un trop à droite et l'autre trop à gauche et qui sont très contents du résultat, car la moyenne est juste sur la cible, est du même type.

Moyenne géométrique plus appropriée

La factorielle d'un nombre N est égale au produit des nombres de 1 à N : $N ! = N \times (N-1) \times \ldots \times 3 \times 2 \times 1$. Une quantité aléatoire d'un intérêt certain est la factoïdelle : pour la calculer on tire au sort des valeurs entre 1 et N, et l'on s'arrête lorsque l'on tombe sur 1. On fait alors le produit des valeurs obtenues. Pour calculer la factoïdelle de 6 on lance un dé jusqu'à ce que l'on tire 1. Nous voulons calculer la moyenne des différentes valeurs de la factoïdelle et pour cela nous faisons plusieurs séries de tirages. Horreur, la valeur de cette moyenne augmente avec le nombre de tirages (elle ne converge pas). En revanche la moyenne géométrique égale à la racine $N^{ième}$ des valeurs de N tirages converge bien vers la racine nième de la factorielle de N.

La moyenne est versatile : bien fol est qui s'y fie.

Philippe Boulanger

par Marie-José Pestel

Du recensement au sondage

Les recensements et les statistiques sont nés en concomitance avec l'État. Le pouvoir a toujours eu besoin de connaître ses forces et ses faiblesses. Le sondage, ultime avatar du recensement, est devenu l'outil indispensable des puissances politiques, médiatiques et économiques.

C'est dans un texte administratif de Colbert que le mot "statistique", du latin «statisticum», qui a trait à l'état, apparaît pour la première fois.

Le besoin «statistique», c'est-à-dire l'activité humaine de recueil de données chiffrées, remonte à la plus haute antiquité. L'idée de recensement, ou de liste d'inventaire, apparaît de façon naturelle dans l'histoire des états, fort désireux de connaître des éléments de leur puissance : population, potentiel militaire, richesses...

De Sumer aux Incas

Les premiers recensements semblent remonter à la civilisation sumérienne, de 5000 à 2000 ans avant notre ère ; on a retrouvé des listes d'hommes et de biens inscrits sur des tablettes d'argile. Le relevé des personnes et des biens a lieu régulièrement en Mésopotamie en 3000 ans avant J.-C. L'Égypte paraît avoir été la première nation à organiser des recensements systématiques de population, au moins depuis l'an 2900 av. J.-C., mais aussi à institutionnaliser des recensements à finalité fiscale. Elle semble avoir la primeur du principe de la déclaration obligatoire : en effet, sous le règne du pharaon Amasis II, au V^{e} siècle avant notre ère, tout individu était tenu de déclarer ses sources de revenu et son activité. Tout manquement à cette règle était punissable de mort. Il est impossible de dresser la liste des premières tentatives de dénombrement exhaustif ; notons simplement qu'ils étaient le fait d'états forts (Japon, Rome, Inca, Hindous possédaient un système administratif puissant).

L'état dans tous ses états

À partir du XIII^{e} siècle, les données deviennent plus nombreuses grâce à la prolifération des rôles fiscaux. Le plus célèbre est, en France, «l'état des paroisses et des feux des baillages et sénéchaussées de France» dressé en 1328.

Le XIV^{e} siècle voit le début des enregistrements des actes civils ; l'obligation de tenir des registres de naissances

date de François I^{er} puis, sous Henri III, les mariages et les décès doivent aussi être enregistrés. Les progrès fondamentaux de la statistique vont apparaître lors de la seconde moitié du XVII^{e} siècle, avec le besoin que ressentent les monarques de connaître et d'expliquer les phénomènes économiques. Apparaît alors la nécessité de faire des estimations et des prévisions, d'autant plus que l'apparition du calcul des probabilités permet de justifier le remplacement d'une connaissance exhaustive par une extrapolation fondée sur l'examen d'une partie de la population.

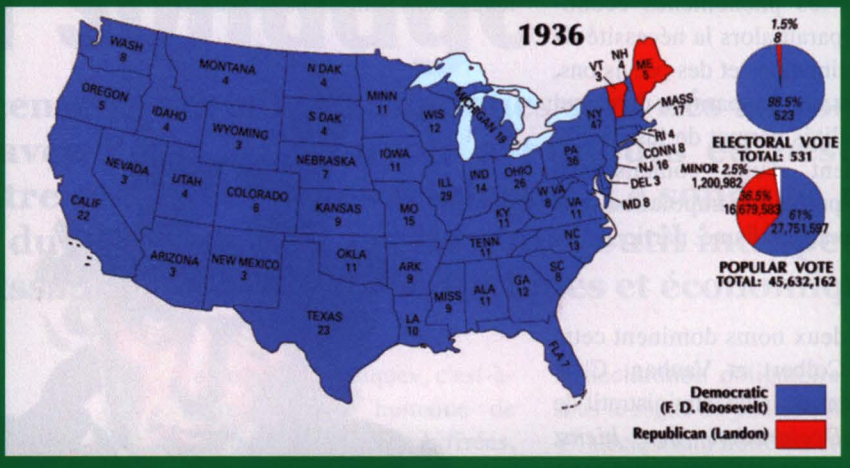
En France deux noms dominent cette période : Colbert et Vauban. C'est d'ailleurs dans un texte administratif de Colbert *«Déclaration des biens, charges, dettes et statistiques des communautés de la généralité de Bourgogne»* que le mot «statistique», du latin «statisticum», qui a trait à l'état, apparaît pour la première fois.
Le marquis de Vauban s'intéresse à la connaissance du chiffre, que ce soit par des recensements ou des enquêtes ; par exemple il préconise l'utilisation d'échantillons de terres arables pour estimer au mieux les capacités agricoles.

Les premiers outils mathématiques

Au XVIII^{e} siècle, des esprits éclairés réclament la création d'un *Bureau pour recueillir les divers dénombrements* ; d'autres, comme Saint Simon, veulent mettre un frein aux recensements jugés trop onéreux, peu précis et «monstrueux» et pensent à mettre en place diverses techniques d'extrapolation ou de multiplicateur. Par exemple, Pierre-Simon de Laplace propose à l'Académie des Sciences, en 1783, une procédure de détermination du multiplicateur consistant à réaliser une enquête portant sur un million d'habitants répartis sur plusieurs régions et de calculer l'incertitude de l'estimateur *-l'erreur à craindre-* en supposant que les régions sont tirées au hasard.

Au début du XIX^{e} siècle, tout semble être bien en place pour permettre un développement de la méthodologie des relevés partiels : la pratique démographique, la progression du calcul des probabilités, la création par Lucien Bonaparte d'un Bureau de Statistique.

Cependant, il faudra encore plus de 100 ans pour que l'intérêt des sondages soit reconnu et un peu plus encore pour établir une théorie fondée sur l'aléa-



États-Unis, 1936. Deux candidats s'affrontent pour l'élection présidentielle américaine : Franklin Roosevelt, le président sortant, et Landon. On s'interroge : qui remportera ?

Le magazine *Literary Digest*, pour donner la réponse, ne lésine pas : il interroge **deux millions d'électeurs** sur leurs intentions de vote, et conclut, sûr de ses méthodes : **Landon va l'emporter**. George Gallup, de son côté, avec de petits moyens, n'interroge que **3000 personnes**, et prédit publiquement, chiffres à l'appui, la **réélection de Roosevelt**. Sa prévision paraissait insignifiante, et pourtant... c'est bien Roosevelt qui a été élu cette année-là ! Pourquoi l'échantillon de loin le plus petit a-t-il été le plus fiable ? *Literary Digest* avait trouvé pratique d'interroger son échantillon par téléphone, et ne s'était donc adressé qu'aux abonnés du téléphone. Or, en 1936, qui avait le téléphone, même aux États-Unis ? La bourgeoisie, plus riche, plus cultivée, celle qui vivait en ville... et cela faussa le résultat, qui ne correspondait pas au vote de l'Américain moyen. **Représentatif ou pas ? Telle est la question !** Les statisticiens diront que l'échantillon de Look était **biaisé**. Gallup, lui, par contre, avait eu l'astuce de choisir un échantillon qui reflétait bien la composition de la population américaine : son échantillon était **représentatif**. Mais au-delà de l'anecdote, cet épisode marquait la victoire d'un principe qui, aujourd'hui, est bien ancré dans le contexte social de notre époque : celui de la validité d'un sondage.

De la même façon que l'on peut apprécier un papier peint à partir d'un échantillon, **l'on peut appréhender les caractéristiques d'une population à partir d'un petit groupe d'individus bien sélectionnés.**

Si le papier peint est uni, aucun problème pour se faire une idée de l'ensemble à partir d'une partie, mais si c'est un papier à motifs ? Il faut alors définir des règles strictes pour déterminer la partie qui sera représentative du tout.

toire.

L'ère moderne

L'Institut International de la Statistique (IIS), créé le 24 juin 1885, sera le cadre pendant plus de vingt ans du débat sur la représentativité d'un résultat. Plusieurs personnalités s'y affrontent, proposant des méthodes diverses et variées : contrôle *a posteriori*, méthode des quotas, procédure de

rééchantillonnage, utilisation d'un test significatif.

Après 1925, le débat n'est plus sur «faut-il échantillonner ou non ?» mais sur «comment tirer l'échantillon ?».

11 y a les partisans du *random sampling* (la prise au hasard de l'IIS) et ceux de la *purposive selection* (le choix judicieux). Le vrai berceau des sondages d'opinion se trouve aux États-Unis, à l'occasion des couvertures de presse des élections présidentielles. Depuis 1824, on y fait des enquêtes préélectorales par consultation individuelle portant sur des échantillons de tailles très élevées, mais sans aucun critère de représentativité. Une date cruciale pour l'histoire de l'échantillonnage est le 3 novembre 1936, jour de la publication des résultats de l'élection présidentielle. Alors que le *Literary Digest* a prédit l'élection de Landon en utilisant un échantillonnage de plus de 2 000 000 de personnes, c'est F. D. Roosevelt qui est élu, ce qui avait été pronostiqué par trois "sondages" réalisés indépendamment. L'histoire a surtout retenu le nom de George Gallup (voir encadré) qui avait créé son propre institut de sondages en 1935 et qui utilisait une méthode par choix judicieux : tirage d'un petit nombre de personnes avec contrôles par quotas, une «Amérique en microcosme» en quelque sorte. En dépit de sa grande taille, l'échantillonnage du *Literary Digest* s'est trouvé biaisé et dorénavant les sondages représentatifs vont fleurir. En 1938 sont crées les premiers instituts d'études de l'opinion en Grande-Bretagne et en France. Jean Stœtzel, professeur de sociologie à la Sorbonne, fonde l'Institut Français d'Opinion Publique (IFOP), il importe les méthodes de Gallup et invente le mot «sondage» «dans un esprit de recherche scientifique, à la fois pour étudier au jour le jour les faits d'opinion, et pour analyser les conditions sociologiques de ce phénomène».

Le 27 avril 1946, une loi crée l'Institut National de la Statistique et des Études Économiques (INSEE) pour «entreprendre à la demande du gouvernement et des administrations publiques et, éventuellement de personnes physiques ou morales de droit privé, des recherches et des études sur les questions statistiques et économiques».

M.-J. P.

F. D. Roosevelt

G. M. Landon

par E. Busser & G. Cohen

La méthode des quotas

La méthode des quotas combine déterminisme et hasard. Elle est au cœur des études de marché et des sondages d'opinion. Cette méthode essentiellement empirique a prouvé depuis plus d'un demi-siècle sa redoutable efficacité.

La composition de l'échantillon doit refléter celle de la population entière.

Bonjour monsieur, accepteriez-vous de répondre à un sondage ? C'est sur les bandes dessinées.

Vous êtes pressé, mais l'enquêtrice est sympathique, et si vous étiez à sa place, vous n'aimeriez pas qu'on vous tourne le dos. Et puis, vous aimez bien les BD. Alors, vous vous installez à côté d'elle sur l'inconfortable escalier auprès duquel elle vous a abordé et vous répondez à ses questions. Ces dernières commencent par des renseignements sur vous. Âge, profession...

- Situation de famille ?
- Marié, deux enfants.

Là, elle vous regarde d'un air désolé, et vous avoue franchement que vous ne l'intéressez plus ! Vous pouvez poursuivre votre chemin, manant ! Dites donc, c'est pour un sondage ou pour passer la soirée que vous m'avez abordé ? Elle s'excuse : l'échantillon qu'elle devait interroger est défini selon la méthode des quotas, et elle a déjà son quota d'hommes mariés.

La méthode des quotas ? Son nom vous est pourtant familier : c'est la méthode utilisée pour les études de marché et les sondages d'opinion.

Elle combine déterminisme et hasard : déterminisme dans le choix contrôlé des caractéristiques de la population, hasard quant à la sélection des sujets.

Elle consiste à construire un échantillon de telle sorte qu'y soient respectées les mêmes proportions que dans la population entière pour les variables supposées fortement liées à l'objet du sondage. Ces proportions sont appelées quotas.

Grâce aux recensements de population et à des mises à jour régulières, on connaît de façon précise certaines caractéristiques de la population française : la répartition hommes-femmes, pour chaque sexe la répartition par âge, la proportion d'ouvriers, d'employés, de cadres, de retraités, la proportion des Français vivant dans une grande ou une petite ville, ou à la campagne. Il s'agit donc, dans l'échantillon constitué, de respecter ces caractéristiques. On donnera donc à l'enquêteur des quotas à res-

pecter, et le calcul se fera selon le modèle de la page suivante.

Seulement, avec cette méthode, il n'est pas possible de calculer l'incertitude du résultat obtenu.

Le calcul des probabilités n'est ici d'aucun secours. C'est l'expérience seule qui a prouvé l'efficacité de la méthode, quand elle est utilisée dans de bonnes conditions. On sait maintenant, grâce à cette expérience, qu'il faut par exemple interroger un millier de personnes environ pour obtenir des résultats satisfaisants sur la population française.

Échantillons non biaisés

La taille de l'échantillon seule ne fait pas sa qualité ; il peut y avoir d'autres causes d'erreurs d'échantillonnage. Les principales sont le biais et la dispersion.
En fonction de l'objet du sondage, on définit un certain nombre de variables supposées fortemant liées à cet objet : l'âge, le sexe, la catégorie socioprofessionnelle, la zone d'habitation, etc. On cherche évidemment à obtenir un échantillon pour lequel les informations sont les plus proches possibles de celles (inaccessibles) de l'ensemble de la population.
❍ Le biais d'un échantillon est caractérisé par un important écart à la moyenne sur l'une de ces variables par rapport à l'ensemble de la population. Imaginez par exemple un sondage national sur les bandes dessinées où l'on interrroge un groupe comportant 80% de personnes âgées. Le résultat risque d'être bien peu représentatif, et pour cause ! La moyenne d'âge de l'échantillon est nettement supérieure à celle de la population française.
L'écart à la moyenne, on le voit, est important, et c'est à partir de sa valeur que l'on mesure le biais de l'échantillon. Il en est de même si l'échantillon compte 90% de femmes. Cette fois, la variable étudiée peut se ramener à une fonction qui vaut 0 si l'individu interrogé est un homme, et 1 si c'est une femme. On sait que la population française compte un peu plus de femmes que d'hommes. La moyenne de cette fonction est donc un peu supérieure à 0,5. Celle de notre échantillon est 0,9.

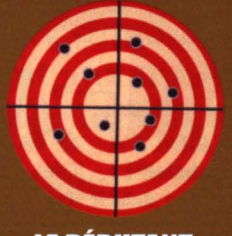
Le centre de la cible est la moyenne du statisticien, et le joueur de fléchettes, comme le sondage, peut avoir divers défauts ou qualités :

LE MYOPE

Son tir est groupé, non dispersé, mais biaisé.

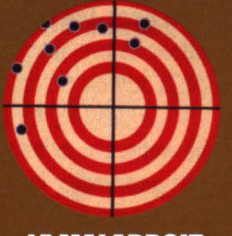
LE TIREUR D'ELITE

C'est un champion, et son tir a toutes les qualités requises par un bon échantillon : ni biaisé, ni dispersé.



LE DÉBUTANT

Il manque de pratique, c'est sûr ! Son tir a une mauvaise dispersion, mais il est non biaisé, les impacts étant répartis autour de la moyenne.



LE MALADROIT

Celui-là ferait mieux de jouer aux dominos, c'est moins dangereux ! Son tir est biaisé et dispersé à la fois.

Le biais est considérable !

❍ Un autre indicateur est également important : la dispersion, c'est-à-dire, pour un échantillon donné, l'étendue des valeurs prises autour de la moyenne. Ainsi, si vous n'interrogez que des enfants et des vieillards, la moyenne d'âge de votre échantillon s'approchera peut-être de celle de l'ensemble de la population, mais la moyenne des écarts à la moyenne, elle, sera très supérieure.

Là encore, votre échantillon n'est pas représentatif, dans la mesure où l'opinion de la tranche des adultes n'ayant pas atteint le troisième âge n'est pas prise en compte.

Pour éviter biais et dispersion, il faudra prendre d'énormes précautions dans le choix de son échantillon. On verra que ces précautions sont plutôt difficiles à respecter, au point que le mieux est de s'en remettre... au hasard, c'est-à-dire de tirer au sort l'échantillon dans l'ensemble de la population. Mais il s'avérera que même tirer au sort n'est pas si simple !

E. B. & G. C. D'après "Comptes de la Vie ordinaire", éd. O. Jacob.

La méthode des quotas : le calcul.

On veut tester sur 200 personnes le marché potentiel d'un nouvel hebdomadaire régional. D'après les résultats d'une précédente enquête, on sait que l'âge, le sexe et le milieu social sont les principaux facteurs influençant la réponse des sujets. Ce sont, d'autre part, des facteurs aisément accessibles pour la constitution d'un échantillon. On va donc faire en sorte que l'échantillon soit le reflet de la population pour ces facteurs. Le dernier recensement fait dans la région donne :

	Caractères	Effectif	%	Échantillon
Sexe	Masculin	333	46,3	93
	Féminin	387	53,7	107
Âge	16-24 ans	117	16,2	32
	25-44 ans	270	37,5	75
	45-64 ans	243	33,8	68
	65 ans et plus	90	12,5	25
Milieu social	Lycéens-Étudiant	92	12,8	26
	Patrons	106	14,7	29
	Cadres supérieurs	52	7,2	14
	Cadres moyens	135	18,8	38
	Ouvriers	234	32,5	65
	Retraités	101	14,0	28
	Total	720		200

L'échantillon de **200 personnes** doit contenir, comme la population, 46,3% d'hommes et 53,7% de femmes, donc 200 x 46,3% = 92,6 soit **93 hommes**, et 200 x 53,7% = 107,4 soit **107 femmes**, dont la répartition se fait selon les résultats de la dernière colonne du tableau ci-dessus.

Dans la pratique, on enverra par exemple 10 enquêteurs sur le terrain, avec pour consigne de rechercher chacun 20 personnes, dont :

- 9 hommes et 11 femmes,
- 3 personnes de 16 à 24 ans, sept de 25 à 44 ans, 7 aussi de 45 à 64 ans, et 3 de 65 ans et plus,
- 3 lycéens ou étudiants, 3 patrons, un cadre supérieur, 4 cadres moyens, 6 ouvriers et 3 retraités.

Statistiques de comptoir (1)

➔ **Précision illusoire.** Un géologue affirmait que le cours d'une rivière avait changé il y a 2 000 004 ans. Admirez la précision ! Interrogé, il avoua qu'il avait pris le chiffre de 2 000 000 dans une publication qui datait de quatre ans. Dans toute statistique, l'inexactitude du nombre est compensée par la précision des décimales, disait Alfred Sauvy.

➔ **Précis, mais non expliqué.** Il y a quelques décennies un article mentionnait des statistiques en Union soviétique sur le coefficient de statut social. Les physiciens avait 7,64, les pilotes d'avion 7,62, les mathématiciens 7,34 et les géologues seulement 7,32. Rien n'était dit sur la manière dont ces coefficients étaient mesurés.

➔ **Corrélation légendaire.** Périodiquement les démographes relient l'augmentation du nombre des cigognes en Alsace à la recrudescence de la natalité, confirmant que ce sont bien les cigognes qui apportent les bébés. Une telle corrélation ne reflète qu'une cause commune : la population des villes augmente et donc, le nombre de nids de cigognes et les naissances.

➔ **Auto-contradictions.** Il en existe de plusieurs types. L'une est que 95% des sondés pensent que les chiffres des sondages ne signifient rien. L'autre, de l'humoriste Robert Benchley, est que la population est divisée en deux types d'individus : ceux qui croient que la population peut être divisée en deux types d'individus… et les autres.

➔ **Statistiques inutiles.** Dans 95% des occasions où ils n'ont rien à dire, 99% des commentateurs sportifs donnent des statistiques : ainsi, au cours de la finale 1998 France-Brésil : « Zidane a été en possession de la balle 17 fois en première mi-temps. Zizou avec ses 33 ans son 1,85 m et ses 78 kg se situe au dessus de la moyenne de l'équipe de France, respectivement 29 ans, 1,81 m et 75,09 kg. »

➔ **Interprétation douteuse.** Les chiffres de la mortalité chez les personnes de plus de 80 ans suivent strictement l'âge. Chaque année, il y a plus de personnes qui meurent à 80 ans qu'à 90. Conclusion hâtive : plus vous vieillissez, moins vous avez de risque de mourir. Pour interpréter justement, regardez la pyramide de population et raisonnez en valeur relative.

➔ **Théories bizarres.** Un « guide spirituel » américain nommé Wilbur Volivia arguait : « Le Soleil n'est éloigné que de 3240 miles et n'a que 32 miles de diamètre. Imaginez les étés glacials que nous aurions s'il était à 93 millions de miles comme « ils » essaient de nous le faire croire.

➔ **Statistiques policières.** Les statistiques de la délinquance de voie publique sont fondées sur les faits constatés, dépôts de plainte ou contrôles policiers. Tout changement de la terminologie décrivant les actes de délinquance (l'inclusion des injures par exemple) ou toute recrudescence des contrôles augmentent automatiquement la délinquance.

par E. Busser et G. Cohen

Le panier de la ménagère

Avez-vous vu un enquêteur de l'INSEE se rendre dans un lieu de vie pour s'en payer une tranche ? « Une tranche de vie s'il vous plaît ! - J'en ai de toutes sortes : vie professionnelle, vie familiale... et de toutes qualités ! Une vie de chien vaut beaucoup moins cher qu'une vie de château... - Mettez-m'en une de chaque. » Rentré chez lui en sirotant un verre d'eau-de-vie, l'enquêteur fait consciencieusement la moyenne des prix. C'est ainsi qu'est publié chaque mois l'indice du coût de la vie !

Cette fable surréaliste a pour but de vous sensibiliser au fait que ce que l'on appelle *coût de la vie* est une notion bien vague dont on peut faire toutes sortes d'interprétations. Pour mesurer cette grandeur insaisissable, on utilise un « thermomètre » appelé indice des prix à la consommation.

Plus généralement, un indice sert à mesurer une variation relative, entre deux situations, d'une grandeur *simple* – définie par la donnée d'un seul nombre, correspondant à un seul bien ou à une seule donnée économique – ou *complexe*.

Indices élémentaires

Par exemple, le prix d'un carnet de métro de seconde classe – la première classe existait encore il y a quelques années – plein tarif à Paris, grandeur *simple*, va donner lieu à un indice *élémentaire* qui va décrire son évolution dans le temps.

Cet indice vaudra, par convention, 100, à sa "date de référence".

Dans le tableau précédent figurent deux indices pour le prix du carnet de métro, IP_{70}, prenant le 1/1/70 comme

Date	Prix (€)	IP70	IP80
1/1/70	0,91	**100**	40
1/1/72	1,22	133,3	53,3
1/1/76	1,37	150	60
1/1/80	2,29	250	**100**
1/1/90	5,34	583,4	233,3
1/1/96	6,71	733,3	293,3
1/1/02	9,30	1022	408,8
1/1/03	9,60	1054,9	422
1/1/04	10	1098,9	439,6
1/1/05	10,50	1153,8	461,5
1/1/06	10,70	1175,8	470,3
1/1/07	10,90	1197,8	479,1
1/1/08	11,10	1219,8	487,9
1/1/09	11,40	1252,7	501,1

date de référence, et IP_{80}, prenant pour référence le 1/1/80 (il vaut donc 100 à cette date). Les deux listes IP_{70} et IP_{80} sont proportionnelles.

Plus généralement, l'indice à une date donnée t sera obtenu en écrivant le rapport du prix à la date t sur le prix à la date de référence.
Entre le 1/1/70 et le 1/1/72, l'indice IP_{70} est passé de $IP_{70}(70) = 100$ à $IP_{70}(72) = 133{,}3$. L'augmentation a été de 33,3%. De même, une régression de l'indice de 100 à 80 aurait signifié une baisse de 20%.
Mais entre le 1/1/72 et le 1/1/76, l'indice 70 est passé de $IP_{70}(72) = 133{,}3$ à $IP_{70}(76) = 150$. L'augmentation de 16,7 points d'indice ne signifie pas augmentation de 16,7%. En pourcentage, l'augmentation a été de $16{,}7/133{,}3 = 12{,}5\%$.
Le prix du carnet de métro est une grandeur simple qui donne lieu aux indices de prix IP_n. Le nombre de carnets vendus par habitant dans l'année en est une autre, qui peut donner lieu à d'autres indices élémentaires, les indices de quantité IQ_n, relatifs à l'année de référence n. En faisant le produit du prix du carnet par le nombre, on obtient une troisième grandeur simple, la valeur de la consommation moyenne d'un Parisien en carnets de métro. Elle donnera lieu à des indices de valeur IV_n.

Indices complexes

Les grandeurs évoquées jusqu'ici sont *simples*. À un instant donné, elles ne prennent qu'une valeur. Mais attention ! Le prix d'un paquet de 250g de café est déjà une grandeur *complexe* ! Il varie selon la marque, la qualité, le conditionnement, le lieu où il est vendu…

Pour le mesurer, il faudrait idéalement faire la moyenne des prix de tous les paquets de café vendus sur tout le territoire français. Rendez-vous compte ! S'il y a trois millions de paquets de café vendus à une date donnée, cela représente une addition de trois millions de prix, total que l'on divise par trois millions pour obtenir le prix moyen ! Autant dire que c'est impossible !

C'est alors que l'on utilise la statistique ! On choisit un *échantillon représentatif* sur lequel on effectue un sondage. Quelques marques, quelques qualités, quelques points de vente : robusta et arabica, *Grand Arôme Diamant* et *Petit Noir de Grand'papa*, Vraichon Paris et Hyper-Dia Trifouillis… Encore faut-il prendre ses précautions ! Il se vend beacoup moins de *Grand Arôme Diamant* que de *Petit Noir de Grand'papa*, moins de paquets en petits commerces qu'en grandes surfaces, *etc.*

On procède par sondages (les *enquêtes de consommation*) pour déterminer ces proportions. On mettra alors dans notre *échantillon* plus de paquets *Petit Noir de Grand'papa* que de *Grand Arôme Diamant*, *etc.*
Autre précaution indispensable : garder le secret des points de vente et des marques choisies pour que personne ne puisse être tenté de fausser l'indice du prix du café ! Le gouvernement n'est pas le dernier, dans l'histoire de la Quatrième République, à avoir tenté une telle falsification !

On fait alors la moyenne de tous les prix de l'échantillon. Ouf ! Voici déterminé le prix moyen du café à

l'instant t, $P(t)$, qui nous permet de construire un indice *synthétique* du café en fixant une date de référence.

Je vois d'ici votre objection : que se passerait-il si l'on choisissait d'autres marques et d'autres points de vente ? On obtiendrait un autre indice, probablement différent du premier ! Comment déterminer le *meilleur*, celui qui se rapproche le plus de la moyenne des trois millions de paquets vendus ?

C'est la statistique qui permet, à travers ces mesures et une fonction appelée écart-type*, de déterminer la fiabilité d'un échantillonnage ! On mesure ainsi qu'il y a une probabilité de 95% que l'erreur commise sur le calcul d'un indice comme le prix du café reste inférieure à 0,3%.

Dernière remarque : l'enquête a également permis de calculer la consommation moyenne $Q(t)$ en paquets de café à la date t. Mais attention ! La valeur moyenne consommée n'est pas le produit $P(t) \times Q(t)$!

Le panier du gorille

Des paquets de café, cela s'additionne ! Le nombre de paquets de café est une grandeur *sommable*. Mais on ne peut pas additionner des bananes et des noix de coco !

Pourtant, sur la planète des singes, ce sont Messieurs les gorilles qui font les courses, et le *panier du gorille* n'est composé que de bananes et de noix de coco. Les primates statisticiens appellent, comme leurs homologues humains, *ménage* toute communauté vivant sous le même toit. Un *ménage* peut donc être composé d'un singe célibataire, d'un couple, ou même d'une famille complète avec enfants et grands-parents !

La consommation moyenne d'un ménage singe à la date t consiste en $Q(t)$ bananes payées au prix $P(t)$ – valeur de la consommation : $V(t)$ –, et $Q'(t)$ noix de coco payées au prix $P'(t)$ – valeur de la consommation : $V'(t)$. Le prix de la banane et de la noix de coco est, à un instant donné, le même sur l'ensemble de la planète des singes. Le problème consiste donc, en prenant pour référence la date 0, à partir des deux indices élémentaires de valeurs IV_0 et IV'_0, à construire un indice synthétique IW_0 représentant l'évolution de la valeur du *panier du gorille*.

La synthèse va s'opérer en faisant une moyenne *pondérée* des deux indices.
Une moyenne, c'est naturel !
Pondérée, car on ne peut pas donner la même importance aux bananes, consommées en masse, et aux noix de coco, consommées moins fréquemment, notamment dans le but de se

désaltérer !
Pondérer, c'est trouver deux *coefficients*, C_0 et C_0', appelés encore *poids*, de telle sorte que l'on puisse définir l'indice synthétique comme la combinaison :

$$IW_0 = C_0 \times IV_0 + C'_0 \times IV'_0.$$

Pour trouver une valeur acceptable de C_0 et C_0', on considère la valeur de la consommation moyenne d'un ménage :

$$W(t) = P(t) \times Q(t) + P'(t) \times Q'(t) = V(t) + V'(t),$$

ce qui permet de calculer l'indice IW_0 à la date t :

$$\mathrm{IW}_0(t) = 100 \times \frac{\mathrm{V}(t) + \mathrm{V}'(t)}{\mathrm{V}(0) + \mathrm{V}'(0)}$$

$$\mathrm{IW}_0(t) =$$

Tout se passe comme si l'on faisait une moyenne pondérée des deux indices élémentaires IV_0 et IV'_0 à l'aide de coefficients C_0 et C'_0. La formule n'est pas si effrayante qu'elle en a l'air, mais vous pouvez faire l'impasse dessus. Les *poids* C_0 et C'_0 *y* représentent les valeurs relatives de la consommation de chaque produit à l'instant 0.

$$\mathrm{IW}_0(t) = \frac{\mathrm{V}(0)}{\mathrm{V}(0) + \mathrm{V}'(0)} \times 100 \times \frac{\mathrm{V}(t)}{\mathrm{V}(0)} + \frac{\mathrm{V}'(0)}{\mathrm{V}(0) + \mathrm{V}'(0)} \times 100 \times \frac{\mathrm{V}'(t)}{\mathrm{V}'(0)}$$

$$C_0 \times IV_0 + C_0' \times IV_0'$$

On peut généraliser ce calcul à des indices *synthétiques*. Si le prix de la banane et de la noix de coco varient selon le point de vente ou la qualité, leurs indices de prix, de quantité ou de valeurs sont eux-mêmes *synthétiques*. L'indice résultant (appelé indice de Laspeyres, voir ci-contre) des prix, des quantités ou des valeurs sera encore obtenu comme la moyenne pondérée des deux indices correspondants, avec les mêmes coefficients de pondération, C_0 et C'_0, valeurs relatives à l'instant 0.

L'indice des prix à la consommation

En février 1993 a eu lieu une petite révolution passée inaperçue aux yeux du grand public : l'INSEE (Institut National des Statistiques et des Études Économiques) a mis en place un nouvel indice des prix à la consommation.

Un tel indice est obtenu de la même façon que les indices synthétiques examinés dans les pages précédentes. Mais, destiné à servir de référence dans la mesure de l'évolution du coût de la vie, il se doit d'être établi avec une grande rigueur et de couvrir un maximum de produits consommés par les Français.

D'un point de vue technique, un tel indice s'apparente à un indice chaîne de Laspeyres*, avec, pour date de référence, l'année 2008 (année du dernier recensement de la population).

D'un point de vue pratique, il faut examiner, outre la date, les références prises pour le calcul de cet indice :

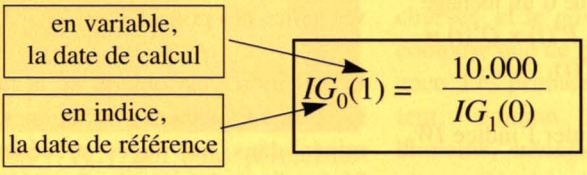
☛ *La population*
Jusqu'à la fin de 1992, pour des raisons politiques liées aux souhaits des syndicats, la population de référence était constituée par les *ménages* urbains dont le chef de famille est ouvrier ou employé.
La rénovation de 1993 a consacré le

Propriétés des indices élémentaires

Dans cette page, les dates seront notées 0, 1, 2 ou simplement t ou n. La grandeur «simple» G pourra être en particulier P, Q ou V= P × Q. Les indices élémentaires qui en résultent seront notés IG_t (indices de cette grandeur en prenant la date t =0, 1 ou 2 comme référence).

La réversibilité

Pour une grandeur simple *G*, les indices *IG* vérifient :



Ce résultat se montre très simplement, dans la mesure où

$$IG_0(1) = 100 \times \frac{G(1)}{G(0)} \text{ et } IG_1(0) = 100 \times \frac{G(0)}{G(1)}$$

La transitivité

Pour une grandeur simple *G*, les indices *IG* vérifient :

$$\mathbf{100 \times IG_0\,(2) = IG_0(1) \times IG_1(2)}$$

formule utilisée le plus souvent sous la forme :

$$\mathbf{IG_1(2) = 100 \times \frac{IG_0(2)}{IG_0(1)}}$$

Là encore, le résultat se montre simplement, puisque

$$IG_1(2) = 100 \times \frac{G(2)}{G(1)} = 100 \times \frac{100 \times \frac{G(2)}{G(0)}}{100 \times \frac{G(1)}{G(0)}} = 100 \times \frac{IG_0(2)}{IG_0(1)}$$

La factorité

Si des grandeurs simples *P*, *Q* et *V* vérifient la relation $V = P \times Q$, alors leurs indices vérifient :

$$\mathbf{100 \times IV_t = IP_t \times IQ_t}$$

En effet,

$$\mathrm{IV_t\,(n)} = 100 \times \frac{\mathrm{P(n) \times Q(n)}}{\mathrm{P(t) \times Q(t)}} = \frac{(100 \times \frac{\mathrm{P(n)}}{\mathrm{P(t)}}) \times (100 \times \frac{\mathrm{Q(n)}}{\mathrm{Q(t)}})}{100} = \frac{\mathrm{IP_t\,(n) \times IQ_t\,(n)}}{100}$$

calcul de deux indices :
– un indice publié portant sur l'ensemble des ménages et (presque) l'ensemble des postes budgétaires des familles,
– un indice non publié, mais servant à la revalorisation du SMIC, portant sur la consommation hors tabac des ménages urbains dont le chef est employé ou ouvrier.

☛ *Le panier de la ménagère*
Les différentes générations d'indices sont en général désignées par le contenu du *panier de consommation* pris en compte dans son calcul. Ainsi, l'indice inauguré en 1993 portera-t-il le nom d'*indice des 265 postes*, succédant ainsi à un *indice des 296 postes*. Il ne faut cependant pas croire que sa couverture a diminué, bien au contraire ! Seuls des regroupements de postes expliquent cette diminution.

☛ *Les coefficients de pondération*
Obtenus à partir d'enquêtes *en taille réelle* remises à jour par des sondages réguliers, ils sont recalculés chaque année pour tenir compte de l'évolution de la consommation.

Plus de 92% des biens et services consommés sont maintenant représentés. 145 000 séries de prix élémentaires sont agrégées pour calculer l'indice. Seuls certains services comme les assurances, les établissements scolaires ou hospitaliers privés, ou les jeux de hasard ne sont pas comptabilisés pour des raisons conceptuelles ou organisationnelles.

L'indice des prix à la consommation, base 1998, constitue la septième génération d'indice. Il couvre l'ensemble de la population et du territoire et se compose aujourd'hui de 305 postes repartis en 161 groupes.

Les 265 postes de 1993

Voici les principaux postes pris en compte pour le calcul de l'indice des prix à la consommation publié par l'INSEE. La valeur indiquée est celle au 01/02/1997 de l'indice 1990 (valeur 100 en 1990)

Alimentation, boissons et tabacs
- Produits alimentaires : 107, 2
- Boissons alcoolisées : 115, 7
- Boissons non alcoolisées : 112, 6
- Tabacs : 196, 5

Habillement et chaussures
- Habillement : 108
- Chaussures : 106, 6

Frais d'habitation
- Logement et eau : 130, 1
- Chauffage, éclairage : 109

Entretien de la maison
- Meubles, tapis, revêtements de sol : 113
- Gros appareils ménagers : 97, 4
- Textiles divers : 114, 9
- Verrerie, vaisselle… : 121,3

Santé
- Produits pharmaceutiques : 104, 2
- Appareils thérapeutiques : 118,5
- Médecins et auxiliaires médicaux : 110, 1

Transports
- Achats de véhicules : 105, 1
- Utilisation de véhicules : 127
- Services de transports : 119, 6
- Communications : 99, 6

Loisirs et éducation
- Appareils et accessoires de loisirs : 99, 8
- Loisirs, spectacles, culture : 120, 3
- Livres, journaux et périodiques : 118
- Enseignement : 123, 5
- Restaurants, cafés, hôtels : 123, 1
- Voyages organisés : 117
- Soins et produits personnels : 118, 2
- Autres articles personnels : 103, 8

Divers
- Autres biens et services : 118, 8
- Services financiers : 123, 6

Ces défauts de couverture pourraient en grande partie disparaître dans les années à venir, sous l'impulsion d'une réglementation communautaire qui devrait imposer une harmonisation entre les modes de calcul des différents indices européens.

Indices de Laspeyres et de Paasche

Pour agréger des indices de grandeurs sommables (prix, quantités), on opère leur moyenne arithmétique pondérée.*
Les coefficients de pondération dépendent de ce qui est appelé la structure de la valeur, c'est-à-dire que chaque coefficient correspond à la valeur relative du produit concerné. Mais à quel moment choisir cette structure ?

La structure de la valeur à l'instant *t*

Dans le cas de deux produits comme la banane et la noix de coco, les coefficients de pondération C_t et C'_t à l'instant t vaudront :

$$C_t = \frac{P(t) \times Q(t)}{P(t) \times Q(t) + P'(t) \times Q'(t)} \text{ et } C'_t = \frac{P'(t) \times Q'(t)}{P(t) \times Q(t) + P'(t) \times Q'(t)}$$

Indice de Laspeyres

Si la structure est choisie à l'instant 0 de référence des indices, l'indice synthétique construit est un indice de Laspeyres L_0. Il obéit à la relation :

$$L_0 = C_0 \times I_0 + C'_0 \times I'_0.$$

Indice de Paasche

Si la structure est considérée à l'instant où est calculé l'indice, l'indice synthétique construit est un indice de Paasche $\prod$.

On aura cette fois : $\prod_0 = C_t \times I_0 + C'_t \times I'_0.$

Propriétés des indices complexes

Plusieurs relations intéressantes lient ces deux indices. S'ils ne bénéficient pas des propriétés des indices simples, comme la factorité, on montre en revanche aisément les égalités :

$$IV_0 = LP_0 \times \prod Q_0 = \prod P_0 \times LQ_0$$

• *L'indice des valeurs est égal au produit de l'indice de Laspeyres des prix par l'indice de Paasche des quantités, ainsi qu'au produit de l'indice de Laspeyres des quantités par l'indice de Paasche des prix.*

• Les indices de Laspeyres et de Paasche ne sont pas réversibles.

Autrement dit, $L_0(1) \times L_1(0) \neq 10\,000$ et $\prod_0(1) \times \prod_1(0) \neq 10\,000$.

L'indice de Fisher, égal à la racine carrée du produit des indices de Laspeyres et de Paasche, a été inventé pour rétablir cette propriété.

• Les indices de Laspeyres et de Paasche ne sont pas non plus transitifs.

Autrement dit, $L_0(2) \neq L_0(1) \times L_1(2)$ et $\prod_0(2) \neq \prod_0(1) \times \prod_1(2)$

S'il y a augmentation de 10% entre les périodes 0 et 1, puis augmentation de 10% entre les périodes 1 et 2, il n'y a pas en général augmentation de 21% entre les périodes 0 et 2 ! Les indices "de chaîne" CL pallient cette lacune.

On aura : $CL_0(n) = CL_0(1) \times CL_1(2) \times CL_2(3) \times \ldots \times CL_{n-1}(n)$.

Ce sont eux qui sont utilisés en particulier pour l'indexation.

La vie est plus chère à l'étranger

Les indices construits dans les pages précédentes observaient l'évolution d'une grandeur avec le temps. Mais on peut également étudier l'évolution d'une grandeur en fonction de la situation géographique.

Ainsi, il est fréquent d'être amené à comparer les prix à la consommation entre plusieurs pays. Pour une grandeur G, on va donc construire par exemple les indices IG_A et IG_F, en désignant l'Allemagne par A et la France par F.
Si G est une grandeur simple, ces indices sont élémentaires, donc en particulier réversibles.
On aura donc :

$$IG_A(F) \times IG_F(A) = 10\,000.$$

Si la banane, compte tenu du taux de change, est plus chère de 25% en Allemagne qu'en France, elle sera moins chère de 20% en France qu'en Allemagne ! $80 \times 125 = 10\,000$.

Mais lorsque l'on a affaire à des indices synthétiques, il en va tout autrement ! Il n'est pas rare, en comparant par exemple les indices des prix à la consommation, d'arriver à la conclusion que la vie est plus chère en France qu'en Allemagne, mais aussi plus chère en Allemagne qu'en France !

Pour expliquer ce paradoxe apparent, imaginez que ces deux pays soient situés sur la planète des singes. Les habitudes alimentaires des singes gaulois et de leurs cousins germains sont différentes ! Les données sont résumées dans le tableau ci-dessous :

Allemagne :
Consommation :
1 noix de coco pour 10 bananes
Prix :
Noix de coco : 3,4 DM,
Banane : 0,5 DM

France :
Consommation :
1 noix de coco pour 1 banane.
Prix :
Noix de coco : 8 FF,
Banane : 2 FF

Dernier renseignement, le change, une difficulté qui a disparu avec l'euro : un deutschmark vaut 3,25 FF. On peut alors évaluer le *panier* français :
$1 \times 2 + 1 \times 8 = 10$ FF en France
$1 \times 0{,}5 + 1 \times 3{,}4 = 3{,}9$ DM en Allemagne
ce qui correspond avec le change à 12,675 FF. L'indice de valeur correspondant vaut :

$$IV_F(A) = 126{,}75 > 100$$

La vie est plus chère en Allemagne pour un Français !

Le *panier* allemand vaut :
$10 \times 0{,}5 + 1 \times 3{,}4 = 8{,}4$ DM en Allemagne
$10 \times 2 + 1 \times 8 = 28$ FF en France
ce qui correspond avec le change à 8,62 DM.

$$IV_A(F) = 102{,}56 > 100$$

La vie est plus chère en France pour un Allemand !

Heureux Alsaciens qui peuvent acheter leurs noix de coco en France et leurs bananes en Allemagne !

E. B. et G. C.

par Jean Lefort

Échantillonnages et interprétations

Selon que le test est destructif ou non, il faut procéder à des modes opératoires adaptés et en évaluer la signification.

Si le test destructif vérifiant la fabrication d'un médicament est onéreux, il ne faut le faire que sur un nombre raisonnable d'échantillons pour faire passer la pilule.

Imaginons une usine de produits pharmaceutiques qui fabrique un certain médicament. La marge de tolérance sur la composition de ce médicament est très faible et les différents appareils qui interviennent dans le processus de fabrication doivent être très fiables ; or ils s'usent et malgré les réajustements réguliers que préconise le concepteur des appareils une panne ou un dérèglement inhabituel peuvent survenir à tout moment. Il est donc nécessaire de vérifier la qualité du médicament en sortie de fabrication.

Malheureusement ces vérifications sont destructrices en ce sens que le produit analysé sera ensuite impropre à la consommation et il est donc impossible de tester tous les médicaments qui sortent de la chaîne. En conséquence on procède à des sondages. On va, par exemple, vérifier un médicament sur mille. Tant que la vérification donne un résultat positif, la fabrication continue sinon on l'arrête et on procède aux réglages nécessaires. Le problème est alors de savoir ce que l'on fait des mille médicaments précédents celui qui s'est révélé défectueux.

Tout dépend du coût de fabrication des médicaments et du coût du test de vérification. Si le coût de fabrication est très faible par rapport au coût du test, on pourra

tout simplement détruire les mille médicaments précédents. Si le médicament vaut plus, on précisera l'endroit à partir duquel la ou les machines se sont déréglées en effectuant un test tous les cent médicaments (on pourrait aussi utiliser la dichotomie, c'est-à-dire prendre le 500 ième, puis si le test est positif le 250 ième) ce qui implique 9 tests supplémentaires et si, toujours à titre d'exemple, le 7ème test conduit au rejet du médicament on détruira tous les médicaments à partir du 600ème. Mais on peut aussi envisager de faire un test tous les vingt médicaments dans la centaine entre 600 et 700 si le prix très élevé du médicament rend cette opération rentable.

Tester pour classer

Il existe un autre cas où les tests de contrôle ne sont pas destructifs. Considérons, par exemple, la fabrication de mécanismes de montres de prestige. Ces mécanismes qui vont permettre toute une série de fonctions (trotteuse, calendrier, alarme, chronomètre,...) seront testés en fin de chaîne pour en contrôler la rigueur. Ceux de ces mécanismes qui passeront tous les tests seront montés sur les montres de grand prix et vendus à un tarif en conséquence, les autres seront classés en différentes catégories selon le type de défaut et montés dans des montres de moindre valeur, celle-ci décroissant avec le nombre de défauts relevés. C'est ainsi que si vous inspectez le mécanisme de certaines montres vous y trouverez des fonctions qui n'ont pas été branchées : ces fonctions avaient été prévues pour une montre élaborée, mais, à la sortie de chaîne, des tests ont révélé quelques anomalies sur cette fonction. L'intérêt de cette procédure

La statistique est la première des sciences inexactes.

Edmond et Jules de Goncourt

est de ne pas sacrifier complètement les éléments qui ne sont pas conformes à la norme requise. Il est évident que le prix de chaque mécanisme dépend de la rigueur des normes auxquels il satisfait et qu'une étude statistique de ce qui se passe en bout de chaîne est nécessaire pour déterminer le prix de chaque catégorie.

Prévoir le résultat du prochain test

Dans chacun de ces deux cas (et des cas voisins) l'industriel ne se contente pas de faire des statistiques en bout de chaîne pour savoir à quel prix il vendra son produit fini. Il cherchera à améliorer sa production en analysant soigneusement les causes de pannes ou de dérèglements des machines qui interviennent dans la fabrication et en y remédiant. Cela lui permet d'améliorer sa productivité, ou, dans certains cas, son taux de retour d'appareils sous garantie (ou d'allonger la durée de garantie) et ainsi de faire plus efficacement face à la concurrence. Dans ce but, il évalue le taux probable de panne à venir, c'est-à-dire sur le prochain appareil ou sur le prochain lot de produits livrés sur le marché. Cette évaluation s'effectue en portant en abscisse le nombre d'objets fabriqués depuis la mise en route de la chaîne et en ordonnée le taux de réussite. On obtient de cette façon un nuage de points que l'on cherche à approcher par une courbe (ce n'est pas une droite mais une courbe asymptotique à 100%) correspondant à un modèle admis dans la profession. C'est ainsi qu'en 1998, lors du lancement de la 100ème fusée

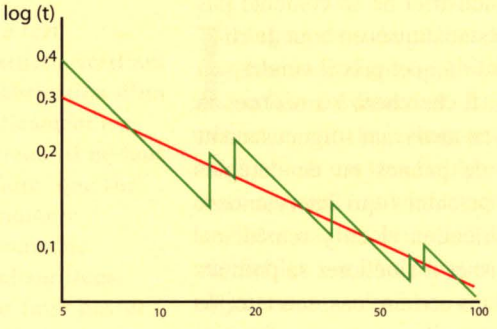
Ariane, on a pu annoncer un taux de réussite de 96,2% alors que seuls 93 lancers avaient été couronnés de succès Dans ce cas, le taux influe de façon non négligeable sur le coût de l'assurance tant du lanceur (qui peut retomber sur des habitations) que du satellite.

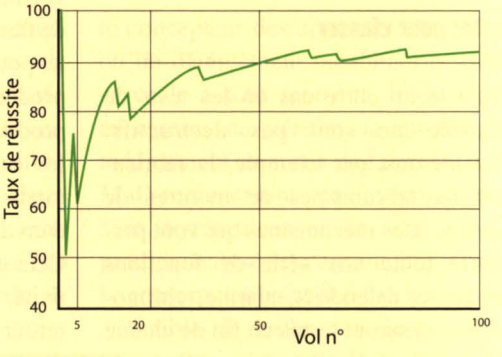
Quand on vous présente le résultat d'un test, il serait fastidieux de demander les conditions de prise du test, mais en cas de suspicion, interrogeons-nous…

J.L.

L'amélioration des statistiques

Les succès passés et les réparations de défaillances qui avaient entraînés ces échecs, améliorent la probabilité de réussite des prochains lancements.
En 1998 avait lieu le lancement (avec succès) du centième Ariane.





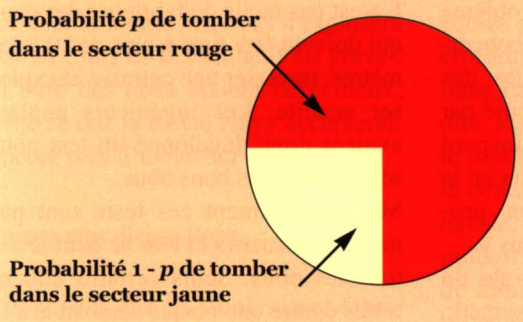
Si on note $e(n)$ le nombre d'échec après le n-ième lancer, le taux d'échec est $t(n) = \frac{e(n)}{n}$ dont on peut raisonnablement supposer qu'il décroît vers 0. On choisit de modéliser cette fonction t par la fonction $f(n) = \alpha\, n^{\beta}$ avec $\beta < 0$. En passant aux logarithmes cela revient à approcher la courbe $\ln(t(n))$ par la droite $\beta \ln(n) + \ln(\alpha)$, ce qui se fait à l'aide de la droite de régression. Pour éliminer les problèmes liés aux premiers lancements, on prend $n \geq 5$. Cela permet de calculer β. Pour évaluer le risque sur le prochain lancement, on considère naturellement la quantité $e(n+1) - e(n)$ que l'on remplace par son approximation $\alpha\,(n+1)^{\beta+1} - \alpha\, n^{\beta+1}$ à laquelle on subsitue son développement limité à l'ordre 1, $\alpha\,(\beta+1)n^{\beta} \approx (\beta+1)t(n)$.
En appliquant cette méthode pour n variant de 5 à 100, on trouve $\beta \approx -0{,}464$ et avec $t(100) = 0{,}07$, il vient $e(n+1) - e(n) \approx 0{,}038$ soit bien une probabilité de réussite de 96,2 %.

Avez-vous déjà volé dans un supermarché ?

La question est embarrassante, et peu de personnes accepteront d'y répondre. Les "sondeurs" -dont l'imagination est grande- vont redoubler d'astuce pour obtenir cependant des réponses fiables. C'est en 1965 que le statisticien Stanley L. WARNER (décédé en 1992) propose l'une de ces astuces : "*Randomized method*". L'enquêteur propose à chaque personne interrogée une carte avec les deux affirmations :

1) J'ai déjà volé dans un supermarché
2) Je n'ai encore jamais volé

En même temps, il fait tirer au sort à la personne interrogée, par exemple à l'aide d'une roulette, l'affirmation qu'il devra déclarer "vraie" ou "fausse". On évitera bien sûr un tirage équiprobable ($p = 1/2$), qui ne fournirait aucune indication. La personne donnera donc "oui" ou "non" comme réponse, sans que l'enquêteur sache à quelle question elle a répondu.



C'est le calcul qui va permettre de venir à bout des réticences du public interrogé. Appelons la proportion de ceux qui ont déjà volé. Le "sondé" répond "vrai" dans deux cas :

- Il tire l'affirmation 1 (probabilité p) et iI a effectivement volé (probabilité π).
- Il tire l'affirmation 2 (probabilité $1 - p$) et iI n'a jamais volé (probabilité $1 - \pi$).

La probabilité d'obtenir la réponse "vrai" est donc :

$$m = p\pi + (1 - p)(1 - \pi).$$

On dispose, non pas de m, mais d'une estimation de m : c'est la proportion de réponses "vraies" obtenues, et on connaît p, donné par la fabrication de la roulette. Un simple calcul donnera une estimation de la proportion des personnes ayant déjà volé dans un magasin :

$$\pi = \frac{m + p - 1}{2p - 1}$$

Le statisticien peut prouver que, plus p est voisin de 0 ou de 1, meilleure est la précision, mais... attention à ne pas éveiller les soupçons !

Bibliographie

L'analyse des données. Jean-Marie Bouroche et Gilbert Saporta, Que sais-je ? numéro 1854, *Presses Universitaires de France (PUF)*, Paris, 125 pages, 1980.

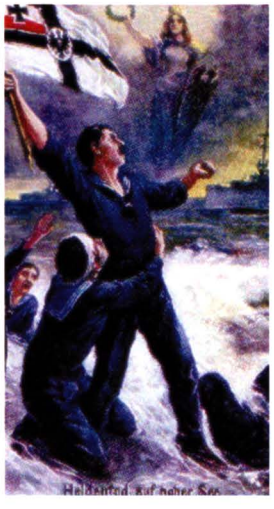
Probabilités, analyse des données et statistiques. Gilbert Saporta, *Éditions Technip*, Paris, 193 pages (656 pages dans la nouvelle édition de 2006), 1980.

Méthodes statistiques pour données qualitatives. Jean-Jacques Droesbecke, Michel Lejeune et Gilbert Saporta, *Éditions Technip*, Paris, 276 pages, 2005.

par Philippe Boulanger

Faire parler la poudre

L'interprétation des statistiques des tests destructifs et la bataille du Jutland : les erreurs du commandement militaire britannique.



Carte postale glorifiant l'héroïsme des marins allemands : au fond une walkyrie s'apprête à les conduire au Walhalla.

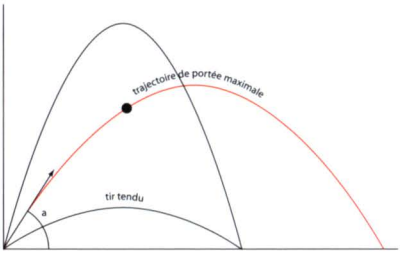
Les stratégies militaires ont été la source de problèmes mathématiques. Le problème de recherche opérationnelle concernant les affectations optimales des troupes sur divers fronts a été posé par Napoléon au géomètre Gaspard Monge qui le résolut et le publia en le transformant pudiquement en problème de « déblais et de remblais ».

En 1916, pour la bataille navale du Jutland au large des côtes du Danemark, la marine anglaise avait mis au point des canons dont la portée atteignait quatorze kilomètres. Pour atteindre de telles distances, les trajectoires devaient être moins tendues. Les obus arrivaient donc sur leur objectif avec un angle plus grand, proche de 45 degrés, pour avoir la portée maximale et l'impact était moins violent que pour des impacts au voisinage de 90 degrés ; aussi certains obus n'explosaient-ils pas. Comment pouvait-on sélectionner les bons obus, c'est-à-dire ceux suffisamment sensibles ?

Un test coûteux

Il n'est pas facile de fabriquer des obus qui doivent être tirés à plus de 10 kilomètres, traverser une cuirasse et exploser ensuite. Les ingénieurs anglais avaient donc développé un test pour sélectionner les bons obus.

Malheureusement ces tests sont par nature destructifs et l'on ne peut tester tous les obus. Seuls certains étaient testés contre une plaque semblable à la cuirasse d'un navire ennemi, à une vitesse et un angle correspondant aux conditions réelles d'utilisation. La portée de la trajectoire parabolique des obus pour une vitesse initiale donnée dépend de l'angle a de la trajectoire. Pour des



distances inférieures à la portée maximale, les trajectoires peuvent être « tendues » et l'angle d'impact sur la coque plus près de la perpendiculaire que pour la trajectoire de portée maximale où l'angle est proche de 45 degrés.

Les testeurs prenaient un lot de 100 obus et en essayaient un : si l'obus explosait, le lot était accepté, sinon les ingénieurs testaient un autre obus du même lot. Si l'obus explosait le lot était accepté ; si aucun des deux obus n'était efficace, le lot était rejeté.

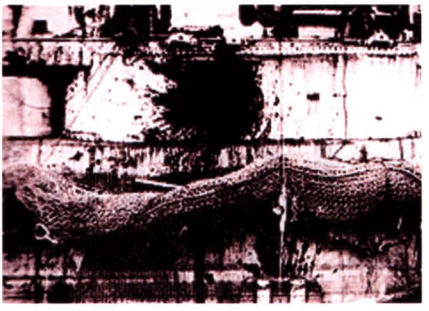
Supposons que dans un lot de 100 obus, seul un obus sur deux soit correctement fabriqué et donc en mesure d'exploser. Si nous prenons deux obus de ce lot, nous avons quatre possibilités : les deux obus sont bons ; seul le premier obus est bon ; seul le second est bon ; les deux obus sont mauvais.

Chacune de ces possibilités a une probabilité proche de 1/4. Il y a environ trois chances sur quatre d'accepter un lot dont la moitié des obus n'explosera pas au cours de la bataille. Plus grave : si 84% des obus étaient défectueux, avec ce test la Royal Navy accepterait encore près d'un lot sur deux.

Encore plus dispendieux...

Supposons que les obus soient fabriqués par lots de 100 et que la proportion moyenne d'obus défectueux puisse être ramenée à 10%. Nous sommes cependant prêts à accepter des lots qui ont jusqu'à 20% d'obus défectueux.

Nous testons n obus d'un lot de 100, et si l'un de ces n obus est défectueux, nous rejetons le lot. On montre que pour n'avoir qu'une chance sur 10 d'accepter un lot comportant 20% de mauvais obus ou plus, il faut tester $n = 10$ obus par lots, ce qui est coûteux. Dans ce cas, nous rejetterons tout de même 65% des lots dans lesquels la proportion d'obus défectueux est de 10 %. C'est du gaspillage. Il faudrait plutôt améliorer les procédures de production d'obus. Toutefois la Marine anglaise continua son procédé de tests statistiques fondé sur ce principe jusqu'en 1944.



Le croiseur allemand Seydlitz en cale sèche après l'engagement (trou d'obus visible).

La bataille du Jutland est considérée comme une victoire anglaise car après l'affrontement les navires allemands ne quittèrent plus leur port d'attache et se consacrèrent à la lutte sous-marine. Pourtant l'affrontement avait coûté 14 bâtiments aux Britanniques et 11 aux Allemands. Les Britanniques, commandés par l'amiral Beatty, avaient perdu trois croiseurs de bataille (*Queen Mary*, *Indefatigable*, *Invincible*). Les Allemands, un seul.

Les croiseurs de bataille anglais ont été coulés par les projectiles ennemis : le croiseur de bataille est un cuirassé de ligne rapide et pour obtenir cette rapidité, les Britanniques devaient réduire le poids de l'armement ou le poids de la défense (en matière de construction navale, tout se réduit à une question de poids). Comme il n'était pas possible de diminuer l'artillerie, les militaires britanniques ont diminué la surface des parties cuirassées et l'épaisseur des plaques. Ils pensaient qu'en raison de leur vitesse, ces croiseurs pourraient imposer la « distance de combat ». Hélas, la distance de combat ne pouvant être imposée que par la portée des canons, une partie de l'escadre de l'amiral Beatty a été détruite.

Les Britanniques avaient un moyen trop onéreux pour tester leurs obus et ne se protégeaient pas contre ces mêmes obus !

P. B.

Bibliographie
The pleasure of counting
T. W. Körner
Cambridge University Press 1996

http://pagesperso-orange.fr/grande.guerre/jutland.html

par Jean Lefort

L'élimination des biais

Les tactiques d'échantillonnage doivent inclure des considérations psychologiques pour cerner la vérité des réponses.

Tant qu'il s'agit d'objets physiques, il est relativement facile de faire des statistiques. Il n'en est pas de même lorsqu'on cherche à faire des enquêtes d'opinion ou de comportement social. Poser une question qui met en cause ce que les gens considèrent comme un tabou ou comme faisant partie de leur intimité conduit au mieux à un refus de répondre au pire à une réponse aléatoire ou convenue. C'est le cas dans notre civilisation pour les enquêtes relative à la sexualité. Ainsi celle qui a été menée en France par l'INSERM, l'INED et l'ANRS au cours de l'année 2007 a conduit au résultat que les femmes déclarent, en moyenne, beaucoup moins de partenaires sexuels que les hommes, respectivement 4,4 pour les femmes et 11,6 pour les hommes au cours de leur vie sexuelle. Les deux résultats devraient pourtant être sensiblement égaux, ce qui traduit manifestement un biais dont sont parfaitement conscients les auteurs de l'étude ; mensonge des uns, vantardise des autres, restriction mentale de certains…

Pour éviter ce type de biais, l'anonymat ne suffit pas car l'expérience prouve que, même dans ce cas, les réponses ne reflètent pas la vérité.

Répondre aléatoirement à la question

Il existe une autre technique qui diminue le biais. Imaginons que l'on veuille enquêter sur la proportion de personnes qui volent dans les grandes surfaces. La question est : « Avez-vous volé dans un magasin au cours des douze derniers mois ? ». On comprend que la personne interrogée hésite à répondre « oui » de peur d'être dénoncée. On lui propose alors une deuxième question très anodine du type : « Avez-vous pris des vacances à plus de 100 kilomètres de chez vous au cours des douze derniers mois ? » L'enquêteur va demander à la personne de jouer à pile ou face (hors de sa présence) et de répondre à la première question s'il obtient pile et à la deuxième s'il obtient face. Il est clair que l'enquêteur est incapable de savoir si le « oui » ou le « non » est relatif à la première ou la deuxième question. Mais la magie des statistiques permet de récupérer la réponse moyenne à la seule première question si l'on connaît le taux de réponse affirmative à la deuxième. Or ce taux est facile à connaître à l'aide d'une enquête indépendante.

Imaginons qu'il soit de 60 % soit 0,6. Nous cherchons à connaître le taux x de réponse à la première question connaissant le taux t de réponse à la double question. Nous avons donc l'arbre ci-contre :

Il est alors clair que

$0{,}5\,x + 0{,}5 \times 0{,}6 = t$

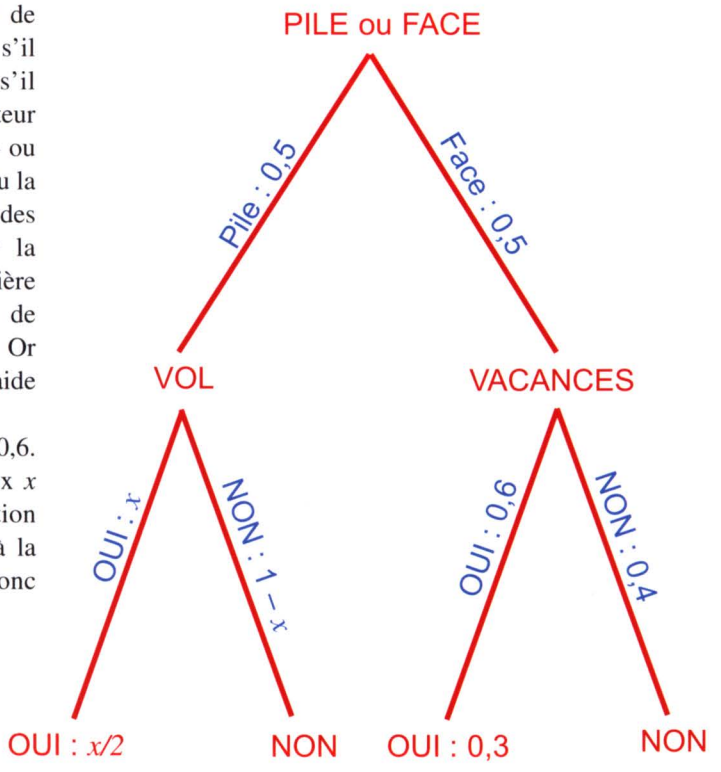
soit $x = 2t - 0{,}6$.

La marge d'incertitude est plus difficile à évaluer et si la procédure permet d'éviter une grande partie du biais, rien n'indique qu'il est entièrement supprimé. En fait, il ne l'est pas entièrement : l'esprit humain est ainsi fait qu'il est fort difficile d'atteindre l'intimité de chacun, et c'est sans doute heureux.

OUI est-il opposé à NON ?

D'un point de vue sociologique, la réponse "oui" et la réponse "non" ne sont pas symétriques. Une enquête assez ancienne le montre à l'évidence. Elle a été menée aux États-Unis il y a une quarantaine d'années :

À la question : "Pensez-vous que les U.S.A. doivent *autoriser* les discours publics contre la démocratie ?" il y eut 21 % de oui, 62% de non, 17 % de sans opinion.



Publicité pour le placebo : ce serait le médicament universel le plus efficace. Pour être la cure idéale, ce fabricant doit obtenir un produit absolument pur sans aucun principe actif. Si le placebo est autant utilisé en recherche clinique et dans la vie courante, son fabricant doit faire fortune... mais il n'obtiendra pas l'autorisation de mise sur le marché de son produit car il faudrait pour cela comparer son effet à un autre placebo.

À la question: "Pensez-vous que les U.S.A. doivent *interdire* les discours publics contre la démocratie ?" il y eut 39 % de non, 46 % de oui, 15 % de sans opinion.

Même si d'un point de vue strictement logique le contraire d' « interdire » n'est pas « autoriser », on sait que c'est le cas dans l'esprit des gens. Aussi la différence est-elle éloquente. Cependant, dans le cas présent, la personne interrogée est relativement hors du débat. Mais si on l'implique personnellement dans la question les réponses changent du tout au tout. Ainsi à la question :

"Doit-on recourir à la force contre les auteurs d'une prise d'otages ?" Les partisans de la fermeté sont environ 50% d'un sondage à l'autre ; mais que l'on vienne à mentionner la possibilité qu'un proche soit parmi les otages et ne les voilà plus que moitié moins ! De plus les réponses varient beaucoup en fonction de l'actualité (prise d'otages récente s'étant bien ou mal terminée).

Le double aveugle

C'est en médecine que l'on utilise une autre procédure pour tenter d'éviter les biais psychologique. Quand une molécule dont la recherche a montré qu'elle pouvait être efficace dans le traitement d'une maladie a été brevetée, testée *in vitro* puis *in vivo* sur des animaux de laboratoire, il faut passer aux essais sur l'homme. On distingue habituellement quatre phases dont trois avant la mise sur le marché. La première sur quelques dizaines de personnes saines pour connaître les vitesses de réaction, la deuxième sur quelques centaines permet de comparer différents dosages et d'étudier les effets secondaires. C'est au cours de la troisième phase qui cherche à mettre en évidence les effets indésirables et la réelle efficacité du médicament que l'on utilise la procédure en double aveugle. Quelques milliers de patients présentant la pathologie sur laquelle le médicament est censé agir sont divisés en plusieurs groupes dont un recevra un placebo (c'est-à-dire quelque chose qui ressemble extérieurement au médicament mais qui ne contient aucun produit actif). Cependant, ni les malades, ni le personnel soignant (médecins, infirmiers…) ne sait qui reçoit quoi. Si on utilise une telle procédure c'est qu'on connaît l'importance du psychisme dans le traitement des maladies : un malade persuadé de guérir, guérit mieux qu'un malade qui pense que la médecine ne peut rien pour lui. Or ce psychisme peut être sollicité soit par le médecin qui va encourager le malade, soit tout simplement par le simple fait de prendre un médicament quel qu'il soit.

Après la mise sur le marché, débute la phase quatre sur toute la population à laquelle est administré le médicament, phase qui permet parfois de mettre en évidence des effets non relevés auparavant.

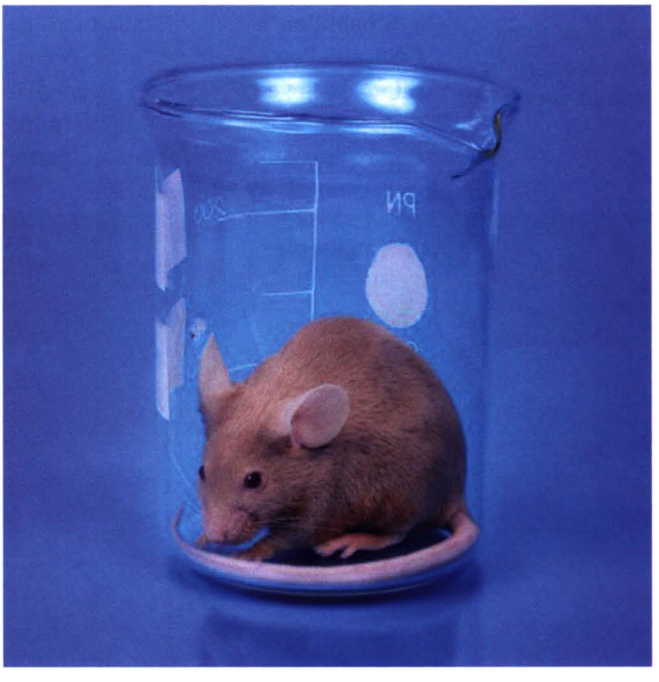
Favoriser le biais

Il y a cependant des organismes qui cherchent à favoriser le biais. Ce sont les groupes de pressions, politiques ou commerciales. Ils poseront des questions telles que la réponse y soit contenue ou qui oblige à répondre dans un sens. Un exemple en est fourni par les pro-nucléaires qui commanditent des enquêtes où l'on pose comme question : « Jugez-vous que la maîtrise de l'énergie solaire justifie que l'on y consacre de larges crédits budgétaires, EN PLUS du programme nucléaire ? » Difficile de répondre autrement que « oui ». Et même si les questions sont neutres, la présentation des résultats peut facilement être biaisée. Il suffit d'écouter les différentes personnalités politiques au lendemain d'élections. Les résultats sont là et ne souffrent en général aucune contestation (tout au moins en France) et l'on voit les différents partis donner des interprétations quasi opposées à ces mêmes résultats ! Ainsi lors de récentes élections municipales, il a suffi de mettre en exergue telle ou telle catégorie de communes (plus de 100 000 habitants, plus de 20 000 ou…) pour minimiser les pertes et maximiser les gains apparents ! Il faut donc aller chercher l'information brute et être capable de la traiter (ce qui n'est pas toujours possible) ou bien trouver des organismes indépendants qui délivrent des commentaires plus nuancés. Il y a deux sortes de statistiques : celles que vous examinez et celles que vous fabriquez…

Pour ou contre les statistiques

Alors ? Pour ou contre l'usage intensif des statistiques ? 98% des statistiques sont fausses prétend l'humoriste. En fait comme toute invention humaine, tout dépend de l'usage que l'on en fait, soit pour mentir, soit pour prévoir, soit pour avertir,… Il faut donc être capable d'analyser soigneusement ce qui est présenté. Cela n'est possible que si l'on peut accéder aux données brutes, aux questions exactes posées, et si, à partir de là, on travaille avec rigueur. Ce n'est pas évident, il y faut de l'expérience et une solide formation tant en mathématique statistique qu'en sociologie. Il est regrettable que trop souvent la présentation de résultats statistiques soit faite par des personnes qui n'ont pas cette formation que ce soit des hommes politiques ou des journalistes. Or une bonne démocratie ne va pas sans une bonne information des citoyens.

J.L.



par Daniel Temam

La valeur des sondages

L'intervalle de confiance donne une idée de la précision des sondages aléatoires et quelle valeur leur accorder. Cet intervalle dépend du nombre de personnes interrogées.

Les trois premières anglaises rencontrées sont rousses : pouvons-nous en déduire qu'elles le sont toutes?

Combien de Français se connectent-ils tous les jours à l'Internet ? Faut-il poser la question à tous les Français ? Cette opération serait coûteuse et on se contente aujourd'hui d'interroger un « échantillon » d'un petit nombre de Français. Pour constituer notre échantillon, on tire au hasard dans la population un nombre prédéterminé d'individus qui sont interrogés ; parmi eux une proportion $\hat{p}$ surfent quotidiennement. Cette proportion $\hat{p}$ est une estimation de la proportion p des Français qui surfent chaque jour.
Évidemment, sauf hasard particulièrement heureux, cette estimation n'est pas tout à fait exacte. Comment estimer sa précision ?

Les physiciens qui effectuent une mesure l'assortissent généralement d'un intervalle dans lequel ils pensent pouvoir affirmer, après avoir examiné toutes les sources d'erreurs possibles, que la vraie valeur se trouve. Il est impossible de faire la même chose dans le cas des sondages: quelle que soit la proportion trouvée dans l'échantillon, la proportion pour l'ensemble de la population peut être quelconque. Une proportion $\hat{p}$ dans l'échantillon de 90 % est compatible avec une proportion réelle p de 1 %, bien qu'une telle configuration soit très peu probable.

Les statisticiens doivent se contenter d'un intervalle où ils sont « à peu près sûrs » que se trouve la valeur exacte p et ils ont introduit la notion d'intervalle de confiance. L'idée est de choisir, avant le tirage, une taille d'échantillon telle que, pour 95 % des échantillons possibles, l'intervalle contienne la vraie valeur p de la proportion. À un échantillon où une proportion $\hat{p}$ d'individus ont la caractéristique considérée (ici regardent Internet) sera associé l'intervalle :

$$[\,\hat{p} - 1{,}96\sqrt{\hat{p}\,(1-\hat{p})/n}\,,\ \hat{p} + 1{,}96\sqrt{\hat{p}\,(1-\hat{p})/n}\,]$$

où n est le nombre d'individus de l'échantillon.

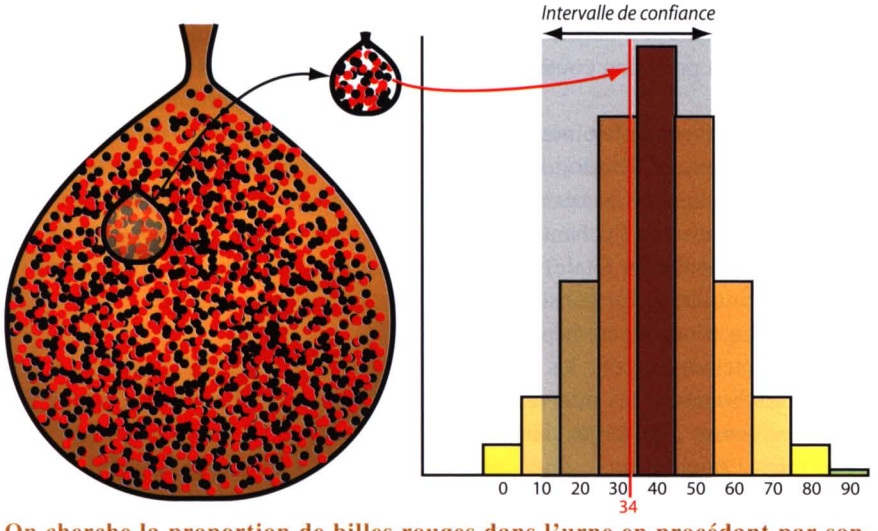
On démontre en effet que, pour 95 % des échantillons choisis au hasard, l'intervalle ainsi défini contiendra la bonne valeur p de la proportion recherchée.

Les sondages ne sont qu'un moyen pour évaluer l'opinion publique. Quand un président ou tout autre dirigeant examine les résultats des sondages, il examine l'opinion des gens. Toute autre interprétation est absurde. ***George Gallup***

Pour une valeur donnée de n, la taille de l'intervalle est maximale (cas défavorable) si $\hat{p}$ est égale à 50 %. Si n est égal à 1000, la taille est ainsi de 6 %, soit ± 3 %. Si $\hat{p}$ est égale à 10 %, la taille est alors de ± 2 %.

Le coût de la précision

La formule de l'intervalle de confiance est instructive. Tout d'abord, la taille de l'échantillon intervient par sa racine carrée : pour diviser l'intervalle de



On cherche la proportion de billes rouges dans l'urne en procédant par sondage : on tire au hasard un échantillon de 100 billes où l'on trouve 34 billes rouges. Si l'on dénombrait toutes les billes et non seulement un échantillon, on mesurerait 40% de billes rouges. Cette proportion est inconnue, mais si elle était connue, on pourrait déterminer la fréquence de billes rouges dans différents tirages (ici notée de 10 en 10). L'intervalle de confiance détermine que dans 95 % des échantillons, la vraie valeur de la proportion est contenue dans l'intervalle indiqué.

confiance par deux, il faut multiplier par quatre la taille de l'échantillon. La recherche de la précision coûte cher.

Par ailleurs, la formule fait intervenir le nombre absolu d'individus de l'échantillon, mais pas le taux de sondage (taille relative de l'échantillon par rapport à la population totale) contrairement à l'intuition. En conséquence, si l'on veut estimer séparément la proportion d'internautes pour les femmes et pour les hommes, les intervalles de confiance seront plus larges dans chacune des populations que pour l'ensemble de la population des deux sexes. Comme il y a à peu près autant de femmes que d'hommes, l'intervalle de confiance sera $\sqrt{2}$ fois plus grand. Quand on subdivise la population en catégories plus fines, par catégories socioprofessionnelles par exemple, la taille de l'intervalle de confiance augmente d'autant plus que l'effectif de la catégorie est faible. Pour que les résultats restent significatifs pour des sous-populations peu nombreuses, il faut augmenter la taille de l'échantillon.

Une présentation délicate

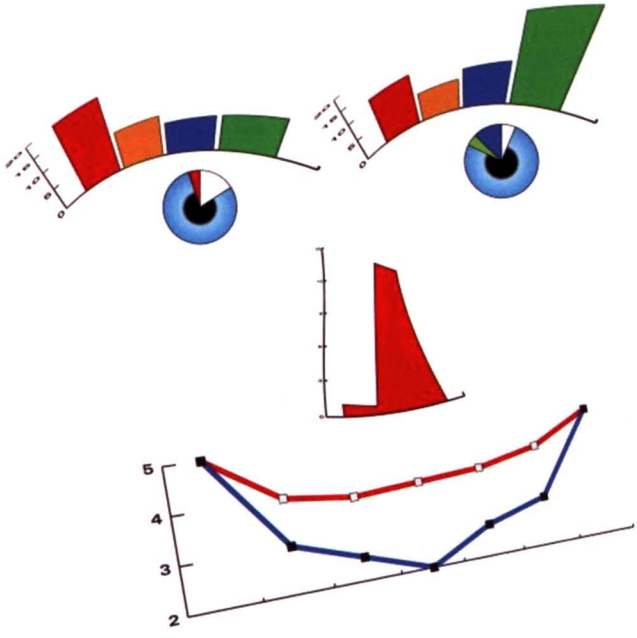
Une fois le tirage effectué, si l'échantillon contient 1 000 personnes, et si par exemple $\hat{p}$ égale à 30 % le résultat sera donné sous la forme : « 30 % des Français surfent tous les jours sur l'Internet ; l'intervalle de confiance à 95 % est de ± 3 % ».

Le résultat est souvent donné sous une forme plus compréhensible : « La proportion de Français qui se connectent tous les jours à l'Internet a 95 % de chances d'appartenir à l'intervalle 30 % ± 3 % ». Cela n'est pas tout à fait exact car p n'est pas une variable aléatoire : p appartient ou non à l'intervalle, mais on ne le sait pas puisque l'on ne connaît pas sa valeur. Tout ce que l'on peut dire, c'est que l'on s'est donné *a priori* 95 % de chances pour que p appartienne à l'intervalle de confiance.

Nous avons introduit le seuil de 95 % des échantillons possibles. Cette valeur est arbitraire. Il faut choisir un seuil assez élevé, voisin de 100 %, et il est naturel de prendre un chiffre rond (en base 10 !). Mais, bien entendu, il est possible de retenir un seuil plus élevé, par exemple 99 % si on veut une plus grande marge de sécurité. L'intervalle sera alors plus large.

Nous avons pris jusqu'à présent l'exemple d'un tirage aléatoire simple

où tous les individus ont la même probabilité d'être tirés. Les statisticiens ont mis au point des méthodes de tirage plus complexes qui réduisent la taille des intervalles de confiance pour une taille d'échantillon donnée. Leur principe est de répartir la population en sous-populations dont le comportement est plus homogène que celui de l'ensemble de la population. Ainsi, la population rurale est souvent distinguée de la population urbaine, celle-ci étant elle-même partagée selon la taille de l'agglomération de résidence.



D'autres sources d'erreur

La notion d'intervalle de confiance ignore toutes les autres causes d'erreur. Parmi celles-ci, la principale est qu'une partie des personnes de l'échantillon refusent de répondre, ou ne peuvent pas être jointes. Il faut alors essayer d'imaginer quelles auraient été leurs réponses les plus probables, compte tenu de leurs caractéristiques connues. Pour résoudre ce problème, les statisticiens ont élaboré des méthodes complexes, mais qui ne sont pas très fiables. S'ajoutent d'autres sources d'erreurs : les personnes interrogées ne comprennent pas bien les questions posées, leurs réponses ne sont pas toujours correctement retranscrites, *etc.*

Notons que cette théorie des intervalles de confiance ne concerne que les sondages aléatoires. Elle ne s'applique pas aux sondages utilisant la méthode des quotas, ce qui est le cas de pratiquement tous les sondages politiques où il n'est pas possible de calculer des intervalles de confiance.

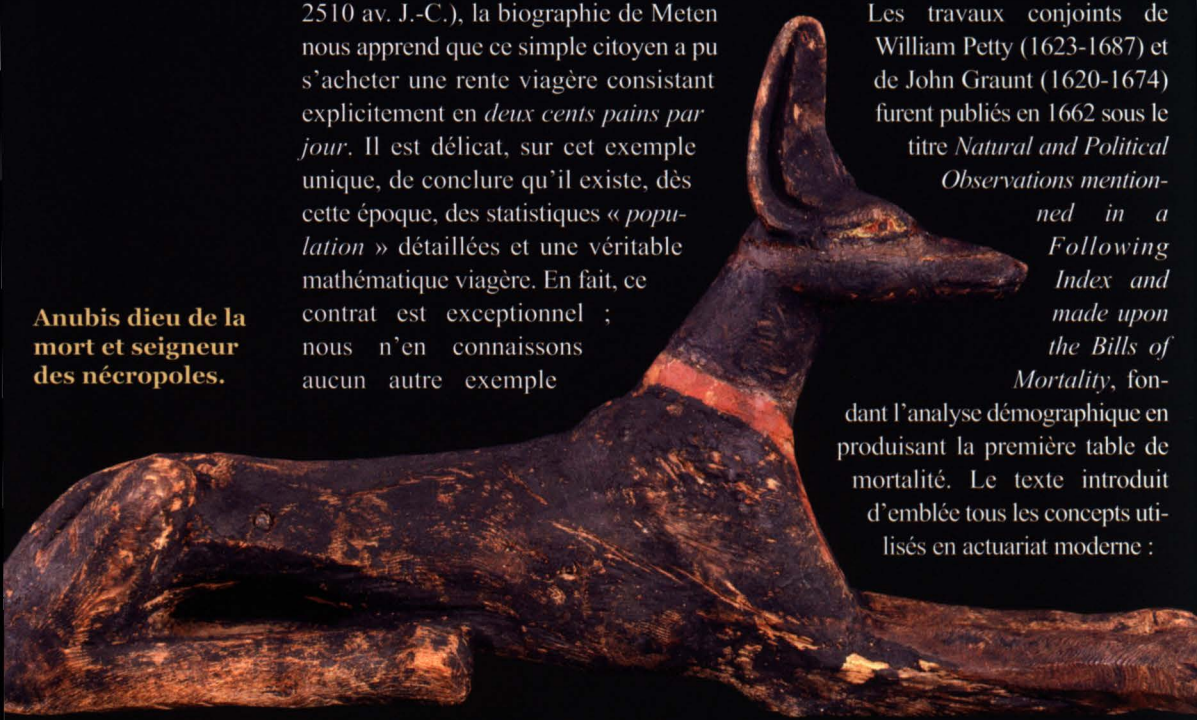
D.T.

par Daniel Justens

Tables de mortalité et pyramides des âges

Les contrats de rente en cas de survie sont plus vieux qu'on ne pense : déjà, dans l'antiquité, les Égyptiens pouvaient y recourir sans pour autant que ces derniers aient élaboré des... pyramides des âges. Aujourd'hui, il existe des modèles très fins de tables de mortalité permettant aux assureurs d'établir rationnellement les primes d'assurances vie.

L'Égypte pharaonique est la première civilisation pour laquelle on a retrouvé les traces d'un recensement (2e dynastie : 2925-2700 av. J.-C.). Mais il y a plus étonnant : sous la 4e dynastie (2625-2510 av. J.-C.), la biographie de Meten nous apprend que ce simple citoyen a pu s'acheter une rente viagère consistant explicitement en *deux cents pains par jour*. Il est délicat, sur cet exemple unique, de conclure qu'il existe, dès cette époque, des statistiques « *population* » détaillées et une véritable mathématique viagère. En fait, ce contrat est exceptionnel ; nous n'en connaissons aucun autre exemple dans l'Antiquité et il faudra attendre 4000 ans pour voir apparaître les premières études démographiques sérieuses.



Anubis dieu de la mort et seigneur des nécropoles.

Premières tables de mortalité

Les travaux conjoints de William Petty (1623-1687) et de John Graunt (1620-1674) furent publiés en 1662 sous le titre *Natural and Political Observations mentioned in a Following Index and made upon the Bills of Mortality*, fondant l'analyse démographique en produisant la première table de mortalité. Le texte introduit d'emblée tous les concepts utilisés en actuariat moderne :

Edmund Halley

7 . 8 9 . . 14 . 18 . 21 . 27 . 28 . . 35 .
11 . 11 . 6 . 5½ . 2 . 3½ 5 6 4½ 6½ 9 . 8 . 7 . 7 .
36 . 42 . 45 49 54 . 55 . 56 . 63
8 . 9½ 8 . 9 . 7 . 7 . 10 11 . 9 . 9 . 10 . 12
70 71 . 72 77 81 84 . 90 91 .
9½ 14 9 . 11 9½ 6 . 7 . 3 . 4 . 2 . 1 . 1 . 1 .
98 . 99 . 100.
0 . $\frac{1}{5}$. $\frac{3}{5}$

La table de Halley représente, sur deux lignes, la mortalité des habitants de Breslau en 1693. Elle indique le nombre de décès (seconde ligne) selon l'âge (première ligne). Lorsque seul est connu le nombre de décès d'une tranche d'âge (par exemple 22 décès entre 10 et 13 ans) alors Halley note la moyenne par âge (22 divisé par 4, soit 5 1/2).

Puisque nous avons trouvé que sur 100 conceptions prises au départ, à peu près 36 n'atteignent pas l'âge de 6 ans, et que peut-être une seule survit à 76 ans, ayant sept décennies entre 6 et 76 ans, nous avons recherché six moyennes proportionnelles entre 64, ceux qui sont encore vivants à 6 ans, et l'unique survivant à 76 ans. (...) De là, il s'ensuit que sur 100 personnes conçues, il en reste,

au bout de six années pleines : 64
au bout de 16 ans : 40
au bout de 26 ans : 25
au bout de 36 ans : 16
au bout de 46 ans : 10
au bout de 56 ans : 6
au bout de 66 ans : 3
au bout de 76 à 80 ans : 0.

Plusieurs méthodes de reconstitution de ces valeurs ont été proposées, fondées sur des séries géométriques de raison 0,63, ou 0,625 mais c'est Hervé Le Bras qui nous en livre l'explication. Le chiffre initial de 64 représente six multiplications successives par 2. Or la *duplicatio* (multiplication par 2) et la *manducatio*, (division par 2) étaient considérées au XVII^e^ siècle comme des opérations au même titre que nos quatre opérations usuelles. Pour prendre 64 % d'un nombre donné, il suffit de le multiplier six fois de suite par 2 et de le diviser par 100, en supprimant les deux derniers chiffres. Si l'on opère ainsi, et si l'on arrondit par suppression de la partie décimale comme on le faisait alors, on reconstitue *presque* la table des *Observations*. La technique repose sur des arrondis et des multiplications simples : la première table de mortalité fut donc une série géométrique décroissante calculée de manière approximative.

La voie était tracée. Au cours des XVII^e^ et XVIII^e^ siècles, de nombreuses tables structurées de façon identique furent éditées, comme celles, en 1693, du célèbre astronome anglais Edmund Halley (1656-1742), l'homme de la comète, concernant la ville de Breslau (Wroclaw, en Pologne). Mais il fallut attendre les travaux de Benjamin Gompertz (1779-1865) pour voir naître le premier vrai modèle.

Enfin des mathématiques !

Graunt et Petty ont ouvert la voie en décrivant les mortalités observées par la donnée du nombre de survivants à

certains âges pour 100 naissances. C'est ce point de vue qui a été formalisé et généralisé. Notons S_0 un effectif à la naissance (âge « 0 »). (Pour Graunt, $S_0 = 100$ mais, aujourd'hui, on choisit généralement $S_0 = 10^5$ ou 10^6). En notant p_x la *probabilité* pour un individu d'âge x d'être encore en vie un an plus tard et q_x la probabilité complémentaire, on décrit l'évolution des « survivants » à tout âge naturel x par :

$$S_{x+1} = p_x \cdot S_x = (1 - q_x)\, S_x.$$

C'est cet ensemble de valeurs S_x que l'on appelle aujourd'hui *table de mortalité*. Toutes les tarifications de contrats d'assurances « vie » sont fondées sur cette suite de nombres.

Remarquons que la table ainsi construite se contente de représenter la survie à court terme d'une population à un moment donné en fonction de l'âge de ses individus et qu'elle ne décrit en aucune façon l'évolution d'une population de sa naissance à sa disparition !

Pour arriver à un modèle, considérons la population S_x comme une fonction continue et différentiable de l'âge (réel positif) x. Calculons le nombre de décès par unité de temps entre les âges x et $x + \Delta x$, à savoir $(S_x - S_{x+\Delta x})/\Delta x$. On appelle *taux instantané de décès* à l'âge x, le taux de décès par unité de temps au voisinage de x. On vérifie que ce taux de décès vaut :

$$\mu_x = \lim_{\Delta x \to 0} 1/S_x \cdot (S_x - S_{x+\Delta x})/\Delta x$$
$$= -\, d/dx[\ln(S_x)]$$

Les tables de mortalité actuelles se fondent sur une modélisation interprétable de ce taux instantané. Benjamin Gompertz constata (en 1825), outre une énorme mortalité infantile suivie d'une décroissance lors de l'adolescence, que les taux de décès des adultes croissaient exponentiellement avec l'âge. Il posa donc ($20 \leqslant x \leqslant 80$) :

$$\mu_x = \alpha c^x.$$

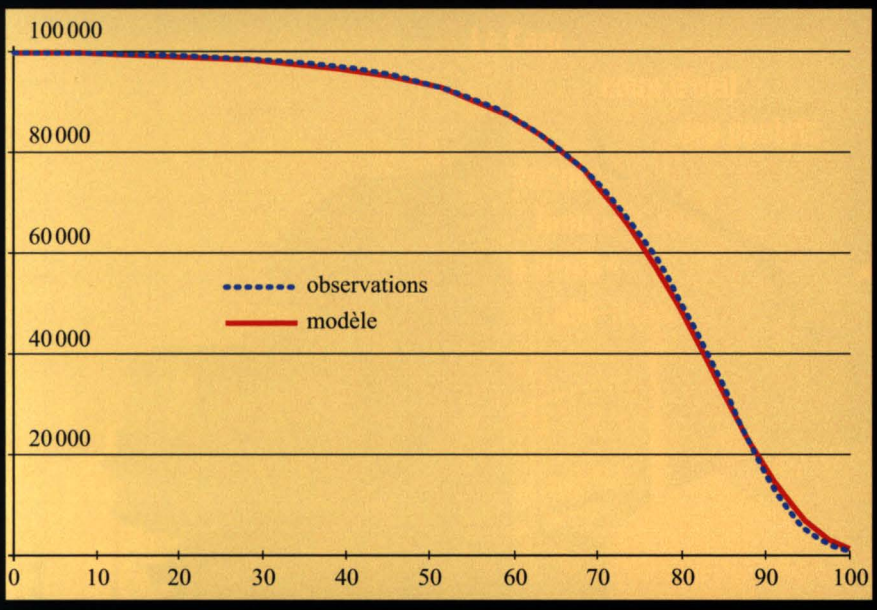
En 1860, William Makeham observa que l'adjonction d'une constante additive améliorait considérablement l'adéquation du modèle aux observations. Il proposa une représentation basée sur l'hypothèse d'un risque accidentel constant quel que soit l'âge et d'un risque exponentiel lié au vieillissement :

$$\mu_x = A + \alpha c^x.$$

A représente ici ce risque de décès accidentel ($A > 0$) supposé constant pour toute catégorie d'âge, α quantifie le risque initial lié à la population considérée ($\alpha > 0$) et c le coefficient d'aggravation du taux de décès par année ($c > 1$).

Observer des taux de mortalité

Pour quantifier des taux de mortalité par catégories d'âges, il faut scinder la population étudiée en classes d'individus homogènes. Étudier les hommes ou les femmes de 40 ans pendant une année, ne peut se faire qu'en deux ans… de façon à observer chaque individu entre deux dates d'anniversaires successives. Les tables publiées font la moyenne de deux ou trois observations de ce type.



Le modèle de Makeham reproduit très exactement les données des taux de mortalité.

Comme nous l'avons vu plus haut, on passe facilement des q_x observés à la *table* S_x. Pour proposer un modèle théorique, il convient d'exprimer S_x en fonction des paramètres introduits dans la représentation de μ_x et de calibrer le modèle, ce qui n'est pas toujours évident, les méthodes d'ajustement étant multiples et les critères d'optimisation arbitraires. Pour la base de données *insee* 2003-2005, nous comparons ici les observations et le modèle de Makeham (voir figure ci-dessus).

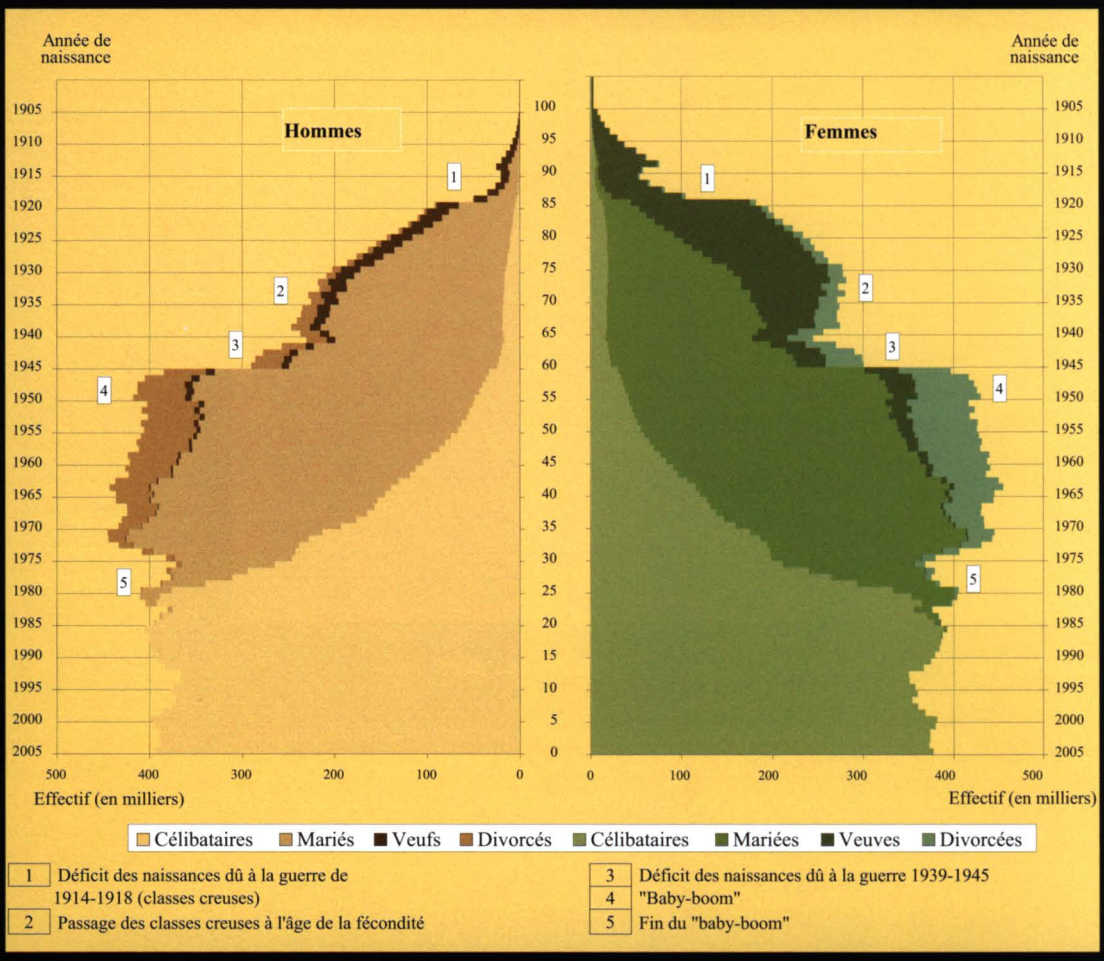
Les taux de mortalité évoluent évidemment au cours du temps. On convient de noter $q_{x,t}$ les *probabilités de décès dans l'année* des individus d'âge x à l'instant t. En pratique, ces taux sont recalculés tous les 2 ou 3 ans mais les compagnies d'assurances utilisent des tables corrigées, tenant compte de l'évolution de la mortalité et surestimant systématiquement le risque couvert, ce qui leur permet de ne revoir leurs tarifs que tous les 10 ou 15 ans. Les clivages les plus fréquents tiennent compte du sexe et du type de couverture : vie ou décès.

Pyramides des âges

Proposons à présent une photographie, un instantané, d'une population à un moment déterminé, en comptant pour chaque âge naturel le nombre d'individus vivants. Voici le graphique des observations en France au 1er janvier 2008 disponibles sur le site de l'INSEE.

On remarque les effets de la première et deuxième guerres mondiales qui se traduisent par un différentiel important au niveau des naissances. Nous avons présenté ci-avant le nombre de personnes vivantes (ordonnées) en fonction de l'âge (abscisses), qui est la représentation mathématique classique d'une fonction. Mais ceci est rarement le cas en matière de pyramides des âges pour lesquelles on opte généralement pour un graphisme plus conforme au vocabulaire usuel car « pyramidal ». La même base de données prend l'allure de la figure page suivante.

Nous avons constaté qu'une table de mortalité ne pouvait se construire que relativement à un intervalle de temps alors que la pyramide des âges était un



Source : INSEE

Répartition de la population totale par sexe, âge et état matrimonial au 1er janvier 2006.

instantané. Mais cette photographie de la population évolue elle aussi au cours du temps. Mathématisons ce passage de l'art de Nadar à celui des frères Lumière.

Notons $P(x, t)$ le nombre d'individus de la population observée d'âge x à l'instant t. L'état de cette fonction de x à l'instant $t+1$ dépend de sa valeur en $x-1$ à l'instant t, tout en tenant compte, d'une part des décès (intervention de la table de mortalité), d'autre part des mouvements de population (immigration et émigration).

En première approximation (les tables de mortalité traduisent une tendance sur plusieurs années et ne représentent pas exactement les taux de décès entre t et $t+1$), on a :

$$P(x, t+1) = P(x-1, t)\, p_{x,t} + I(x, t) - E(x, t).$$

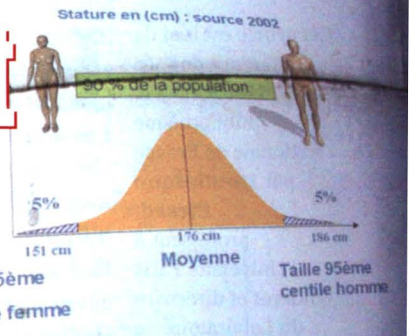
Expression dans laquelle nous avons noté $I(x, t)$ le nombre d'immigrés d'âge x entre les instants t et $t+1$ et $E(x, t)$ le nombre d'émigrés du même âge pour la même période.

L'évolution de la pyramide des âges au cours du temps dépend donc d'un grand nombre de facteurs indépendants : nombre des naissances, évolution de la mortalité, limitation ou intensification des mouvements de population.

Bien malin qui peut prévoir son évolution future.

D. J.

Le traitement des données



La statistique, c'est à la fois le recueil des données et leur assimilation. Le traitement des données, amélioré par des plans d'expériences et la segmentation, est analysé par des méthodes nouvelles, Monte-carlo, test du khi-deux, utilisation des travaux de Le Cam, puis présenté par les visualisations qui soulignent les points marquants.

par Gaël Octavia

Lucien Le Cam : exhaustivité et expérience statistique

Dans un article publié en 1964, Lucien Le Cam, l'un des fondateurs des statistiques modernes, définissait l'exhaustivité approchée d'une expérience statistique. Cette notion réduisant la quantité d'information nécessaire à la connaissance d'un phénomène, s'applique dans de nombreux domaines, notamment en physique, en stockage de données, en finance.

Cet article est issu de la conférence donnée le 19 mars 2008 à la Bibliothèque Nationale de France par **Dominique Picard**, professeur à l'Université Paris Diderot et directrice du Laboratoire de Probabilités et Modèles Aléatoires, dans le cadre du cycle *Un texte, un mathématicien* organisé par la SMF et la BnF.

Les statistiques sont un domaine apparu assez tardivement dans l'histoire des mathématiques. Les fondements des mathématiques peuvent être situés au III^e^ siècle avant J.-C., avec Euclide (– 325-– 265). Les probabilités arrivent environ 2000 ans plus tard, en 1654, avec Pascal (1623-1662) et Fermat (1601-1665), puis Jakob Bernoulli (1654-1705). La loi des grands nombres est énoncée en 1713. Les origines de la statistique, remontent à la fin du XVII^e^ et au début du XVIII^e^ siècle, avec l'écrivain et mathématicien écossais John Arbuthnot (1667-1735), qui avait fait des statistiques sur le sexe des bébés (et attribué à la « Divine providence » le plus grand nombre de garçons). Citons comme autre grands noms Thomas Bayes (1702-1761), Abraham de Moivre (1667 - 1754) ou encore Pierre Simon de Laplace (1749 - 1827).

En statistique, on suit une démarche qui fait l'aller-retour entre la réalité et les mathématiques : on part de données que l'on observe, puis on va vers les mathématiques, et enfin on retourne vers l'application aux données. Les objectifs d'un statisticien sont les suivants : prendre des décisions (pour cela il devra définir des paramètres, des dépendances, valider des hypothèses), réduire des données, transférer les savoir-faire d'une expérience facile à une expérience plus difficile.

Sondages, économie, médecine… : les statistiques font partie de notre vie quotidienne !

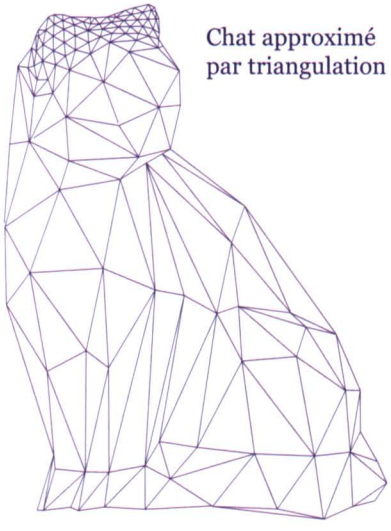
En décembre 1959, Lucien Le Cam écrit *Sufficiency and approximate sufficiency* (Exhaustivité et exhaustivité approchée). L'article n'est publié que 5 ans plus tard, en 1964 dans *Annals of Mathematical Statistics* (vol. 35, n° 4,

On risque autant à croire trop qu'à croire trop peu.
Denis Diderot (1713 - 1784)

Lucien Le Cam (1924 - 2000)

D'origine paysanne, Lucien Le Cam est né en 1924, dans la Creuse. Il connaîtra un parcours sinueux. Tenté par une vocation religieuse, il commence par rentrer au Séminaire de Limoges, qu'il quitte au bout de 24 heures. Ensuite, du fait d'une procédure spéciale en vigueur sous l'Occupation, il ne peut se présenter à l'Ecole Polytechnique. Il passe enfin sa Licence ès sciences en 1945 à Paris. Là, il a sa nuit de Pascal mathématique et, plus spécifiquement, statistique. Il travaille d'abord pendant 5 ans comme statisticien appliqué à EDF.
C'est en 1950, lors d'un séminaire, qu'il rencontre le mathématicien et statisticien américain Jerzy Neyman (1894 - 1981). Celui-ci l'invite pour un an à Berkeley. En effet, la statistique s'est plutôt développée dans les pays anglo-saxons. Le Cam arrive donc à Berkeley avec l'intention d'y séjourner un an. En réalité, il y restera 50 ans. Il y poursuit son parcours universitaire : *lecturer* en 1950, *graduate student* en 1951, PH. D en 1952.
En 1953, il est *assistant professor* et a un premier étudiant en thèse. En 1960, il est *full professor of statistics*. En 1973, il est titulaire de la Chaire de statistiques et de mathématiques. Il aura 40 étudiants en thèse et a 290 descendants à ce jour. Le Cam est l'un des fondateurs des statistiques modernes.

pp. 1419-1455), les experts ayant jugé l'article original de Le Cam trop difficile. Cet article montre que l'on peut approximer des expériences statistiques par des expériences statistiques élémentaires (de même que l'on peut approximer une courbe localement par des segments de droites, ou une figure géométrique compliquée à l'aide de triangles).



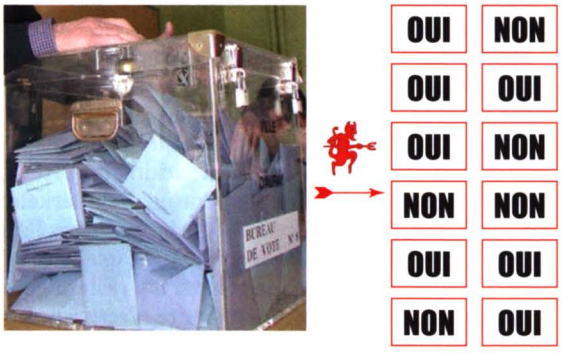
Chat approximé par triangulation

Le Cam définit donc ce qu'est une expérience statistique élémentaire, ainsi qu'une distance sur l'ensemble des expériences statistiques (voir encadré *Quelques définitions et formules*).

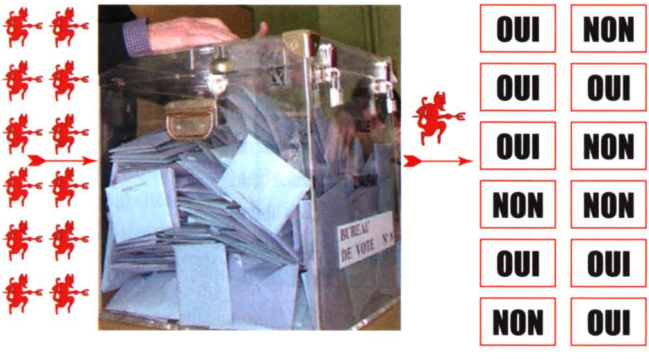
L'expérience statistique

Pour comprendre l'idée de Le Cam, définissons ce qu'est une expérience statistique. Cette notion est introduite par Wald en 1939 et par Blakwell en 1950. Une expérience statistique ε est un objet mathématique : c'est un ensemble de probabilités.
$\varepsilon = \{P_\theta, \theta \text{ appartenant à } \Theta\}$, où Θ est un ensemble de paramètres (caractéristiques de cette expérience) et P_θ est une probabilité sur Θ associée au paramètre θ.
Prenons l'exemple d'un sondage. Lors d'un référendum, les électeurs peuvent voter « oui » ou « non » et l'espace Θ est composé d'un seul élément θ qui vaut 0 (pour « non ») ou 1 (pour « oui »). On effectue un sondage pour prévoir les

résultats, c'est-à-dire la proportion (inconnue) d'individus votant « oui ». On peut voir la loi de probabilité P_θ comme un diable tirant les bulletins de l'urne.



On définit alors la **donnée** χ comme une **réalisation aléatoire** tirée suivant la loi P_θ inconnue. Dans le cas de notre sondage, χ sera un ensemble de réponses « oui » ou « non » qui sont les expressions de vote de n individus choisis indépendamment : $\chi = \{\chi_1, \chi_2, \ldots, \chi_n\}$ où chaque χ_i vaut soit 0 soit 1. Le problème du probabiliste, c'est que plusieurs lois de probabilité peuvent tirer une même donnée.



Face à une donnée, on a donc deux incertitudes. L'une est liée au fait que la donnée est aléatoire. L'autre est liée au fait qu'on ne connaît pas la probabilité qui a tiré cet événement (est-ce une loi uniforme, une loi de Dirac, … ?).

Le statisticien essayera de deviner cette loi de probabilité à partir de l'observation des données. Dans le cas du sondage, P_θ est la loi de n variables de Bernoulli indépendantes.

Pour $n = 4$, si on tire une donnée

$$\chi = \{0, 0, 0, 1\},$$

cette loi s'écrit

$$P_\theta(\chi = \{0, 0, 0, 1\}) = (1 - \theta)^3\, \theta.$$

Décision, risque et exhaustivité

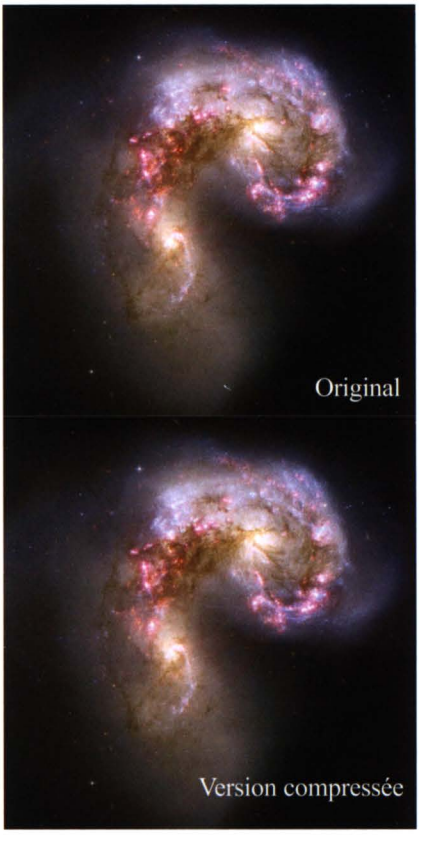
Une expérience étant donnée, on l'utilise pour prendre une *décision* (que l'on note $d(\chi)$, χ étant la donnée aléatoire définie précédemment) dont on mesure le *risque*. On peut définir qu'une expérience est meilleure ou moins bonne qu'une autre du point de vue du risque. Par exemple, dans une classe, les élèves sont identifiés par leurs noms et prénoms. On peut vouloir les identifier par leurs seules initiales, ce qui représente une économie en termes de stockage de données. Cependant, cela conduit à une perte d'information trop importante, puisque plusieurs élèves peuvent avoir les mêmes initiales. Cette expérience sera jugée moins bonne (plus restreinte) du point de vue du risque.

De manière générale, une sous-expérience d'une expérience donnée est moins bonne du point de vue du risque. Mais elle peut aussi lui être *équivalente*. C'est le cas lorsque, pour reprendre l'exemple du sondage, au lieu de stocker les « oui » et les « non », on décide de ne compter que les « oui ».

Une sous-expérience (notée ε^Y) est équivalente à l'expérience ε si, à partir de ε^Y, on peut reconstruire une expérience qui n'est pas ε mais qui aura le même comportement que ε du point de vue des décisions et de leurs risques.

On peut parler d'*exhaustivité* lorsque l'équivalence est vérifiée, qu'il n'y a aucune perte d'information. Le Cam introduit quant à lui l'idée d'*exhaustivité approchée*, dont on peut avoir l'in-

tuition en regardant une photo et sa version compressée.
L'image compressée ne représente que 10 % de l'originale en termes de place, de quantité d'information. Cependant, cette information est « suffisante » dans le sens où la photo compressée est tout à fait lisible et exploitable.



Le théorème que Le Cam a démontré dans l'article de 1964 lie la notion d'exhaustivité approchée aux comportements des risques dans les expériences en question.

Normalité asymptotique locale

Le Cam définit au début des années 1960 la notion de *normalité asymptotique locale* : « La notion asymptotique ici traduit deux idées : La première traduit le fait que l'information amenée par l'observation est suffisante pour produire des estimations assez précises des paramètres du modèle. La deuxième traduit le fait que dans le voisinage des « valeurs plausibles » pour ces paramètres, la famille de probabilités peut être approximées assez finement par une expérience gaussienne de nature plus simple. »
Il y a donc une théorie de la régularité sur les expériences statistiques : si une expérience est assez régulière, elle vérifie cette propriété de normalité asymptotique locale, c'est-à-dire que localement, la loi se comporte comme une gaussienne.

Big Bang et stockage des données

Le résultat de Le Cam a pour conséquences directes ou indirectes le codage, la compression de données, la théorie de l'information, l'estimation fonctionnelle, la notion de sparsité (si l'on s'y prend bien, beaucoup de phénomènes peuvent s'exprimer avec peu de paramètres), le bootstrap (rééchantillonnage)… La normalité asymptotique locale se vérifie dans des cas concrets. Par exemple, des données financières suivent en général une loi compliquée, mais elles sont exploitables localement car approchées par des gaussiennes. Autre domaine : la cosmologie, où un exemple d'application est celui du bruit de fond cosmologique, qui est une radiation fossile provenant du Big Bang et qui renseigne les physiciens sur l'Univers. Le fait de pouvoir dire si ces données sont gaussiennes est très utile. Les statistiques en médecine et en médecine légale, l'analyse du signal et de l'image, les logiciels de notation musicale en sont aussi des champs d'application.

G. O.

La Fondation Sciences Mathématiques de Paris

Le Laboratoire de Probabilités et Modèles Aléatoires fait partie des 9 laboratoires parisiens fédérés par la *Fondation Sciences Mathématiques de Paris*, un réseau d'excellence qui regroupe la plus grosse concentration de mathématiciens au monde, avec plus de 1000 chercheurs parmi lesquels 4 Médaillés Fields, 14 Académiciens et des lauréats, chaque année, de prix nationaux et internationaux. La Fondation initie et finance des programmes d'envergure internationale : bourses, chaire d'excellence, positions post-doctorales, invitations de chercheurs... Elle a trois objectifs essentiels : faire de Paris le pôle le plus attractif pour l'élite mondiale des étudiants et enseignants-chercheurs en sciences mathématiques, développer les collaborations entre la recherche mathématique et le monde économique et industriel, et enfin favoriser l'intérêt général pour les mathématiques.
Pour en savoir plus sur la Fondation : **www.sciences-maths-paris.fr**

fondation SCIENCES MATHÉMATIQUES de PARIS

par Philippe Boulanger

Comprendre d'un coup d'oeil

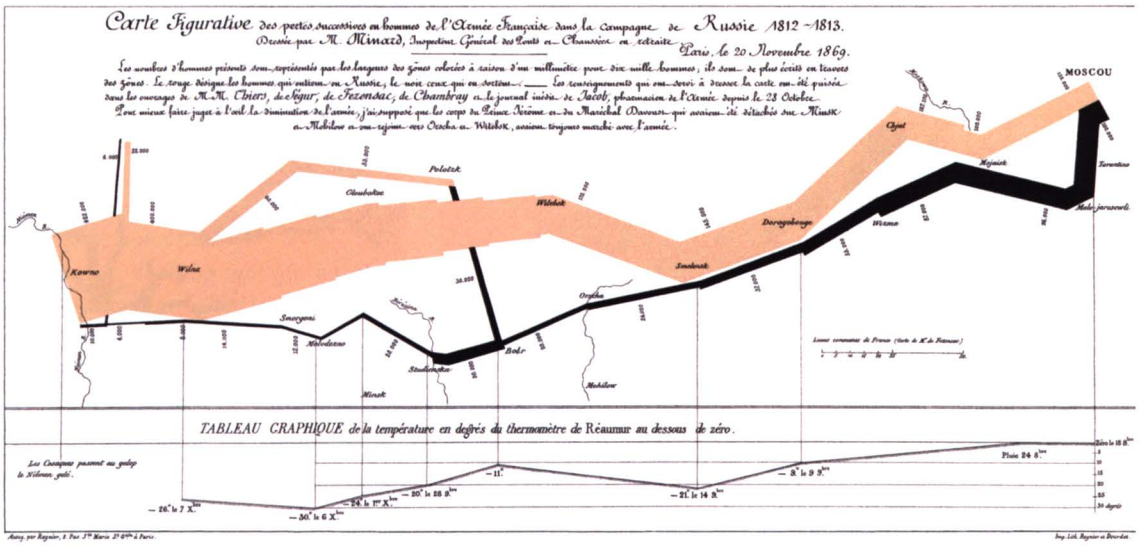
Charles Minard (1781-1870), polytechnicien de talent (si, si...), a inventé des moyens frappants pour représenter les données, encore utilisés aujourd'hui.

"Un bon croquis vaut mieux qu'un long discours." Napoléon Bonaparte (1769-1821)

Le monde moderne engendre une énorme quantité de chiffres, économiques, scientifiques, sportifs, qu'il nous faut assimiler. Or les tableaux de chiffres ne « parlent » guère. Les représentations graphiques portent plus à la maturation des idées. C'est cet examen des données qui peut confirmer nos intuitions et vérifier des lois ou faire apparaître de nouvelles régularités de la nature et des comportements humains. Les graphiques peuvent aussi dégager des erreurs qui auraient faussé notre jugement.

L'idée d'une bonne figuration graphique n'est pas neuve : en 1861, Charles Minard (1781-1870) traça une représentation graphique de l'effectif de la Grande Armée lors de la campagne de Russie de 1812. Pour cela il s'appuya sur les récits des mémorialistes, russes ou français : Adolphe Thiers (1797-1877), le Général De Ségur (1780-1873), le Duc de Montesquiou-Fézensac (1784-1867), le Marquis de Chambray (1783-1848) et le pharmacien Paul-Iréné Jacob (1782-1855). Le tableau de Minard défie les longs discours de l'historien par sa brutale éloquence.

Charles Minard est le fils d'un officier de la gendarmerie et d'une intendante d'un collège de Dijon. À quinze ans, il est admis à École Polytechnique et ses professeurs sont prestigieux, notamment Fourier et Legendre. En 1800, il poursuit ses études dans une école d'application, l'École nationale des Ponts et Chaussées, la première école à former des ingénieurs chargés de la construction de ports, routes, canaux et, plus tard, de lignes de chemins de fer en France. Toute sa carrière (1803-1851) se déroulera aux Ponts et Chaussées, tout d'abord en tant qu'ingénieur de terrain, plus tard comme



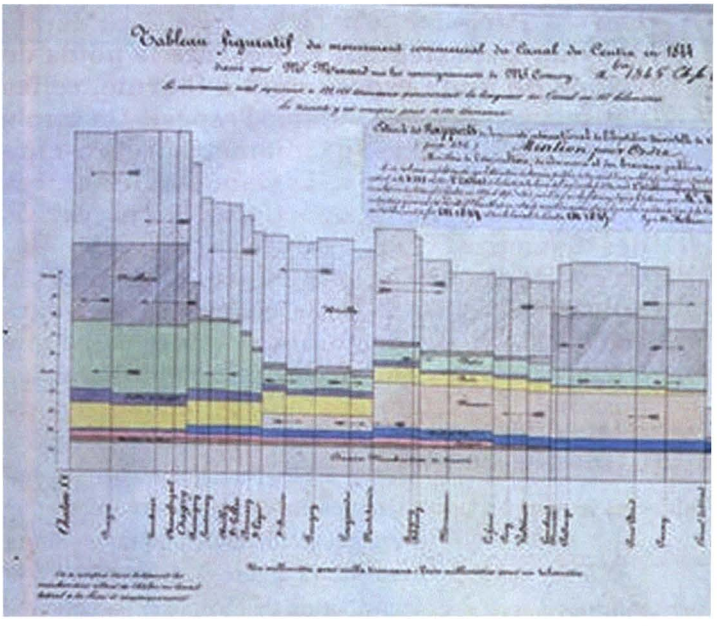
1. La Grande Armée, forte de 422 000 hommes, est anéantie, surtout par le froid : les températures, en degrés Réaumur, sont indiquées le long du périple. Seuls 10 000 hommes revinrent à Paris. Minard représente l'effectif de l'armée par un trait plus ou moins épais.

formateur dans le domaine de la navigation intérieure et en construction de chemins de fer. Même après sa retraite forcée en 1851, Minard continue d'occuper son fauteuil au conseil d'administration des Annales des Ponts et Chaussées. Sa production de nouvelles formes et thèmes graphiques double en dix années et se poursuit jusqu'à sa mort à l'âge de quatre-vingt dix ans.

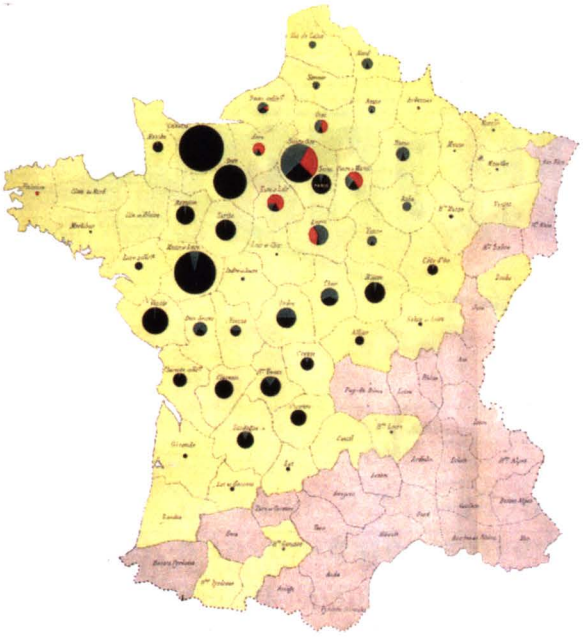
Le but de Charles Minard est d'attirer l'attention sur des relations qui ne sont pas évidentes quand il faut calculer de tête.

Si la campagne de Russie est la réalisation la plus connue de Minard, il fait œuvre de pionnier dans bien d'autres représentations.

Minard invente une nouvelle forme de graphique divisé en barres où l'épaisseur des barres est proportionnelle à la distance parcourue et la hauteur de leurs subdivisions proportionnelle au nombre des passagers ou au tonnage de chaque type de marchandises. Par conséquent, l'aire correspond à l'importance du trafic.



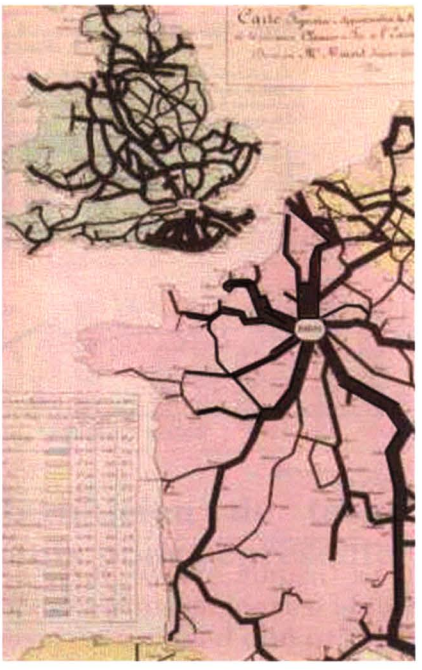
2. Mouvements commerciaux le long du canal du Centre en 1844. Chaque rectangle correspond à la distance entre deux destinations, chaque couleur à un type de marchandise transportée, la hauteur de la barre étant un tonnage.

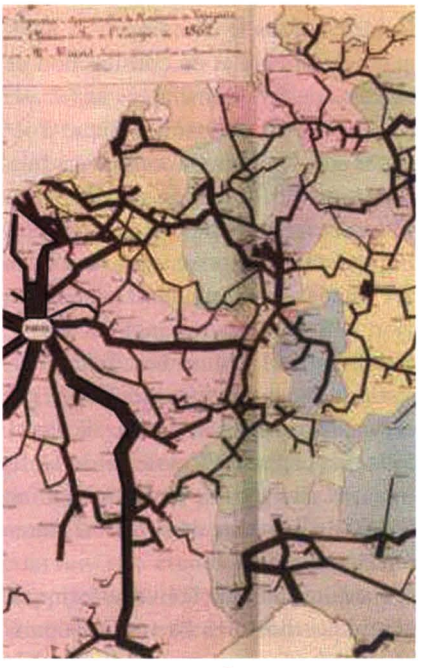


3. Quantités de viandes de boucherie envoyées sur pied par les départements et consommateurs à Paris (1858). La surface d'un cercle dans un Département représente le poids de viande de toute espèce qu'il a fournie, celles des secteurs colorés indique l'espèce. Un cercle de six millimètres de diamètre représente 555 000 kilogrammes de viande et les autres cercles des poids proportionnels aux carrés des diamètres.
La couleur noire indique le bœuf et la vache, le secteur bleu la viande de veau et les secteurs rouges la viande de mouton. On remarque que les lenteurs des communications annulent les possibles contributions des départements du Sud.

Minard est le premier à utiliser des diagrammes en « camembert » dans une carte géographico-économique. Sur la carte de la figure 3, il représente la contribution de chaque département à l'approvisionnement en viande de Paris.

Minard innove en figurant l'intensité d'une grandeur, le trafic de voyageur ou le commerce d'un bien, par l'épaisseur du trait, comme le trafic ferro-



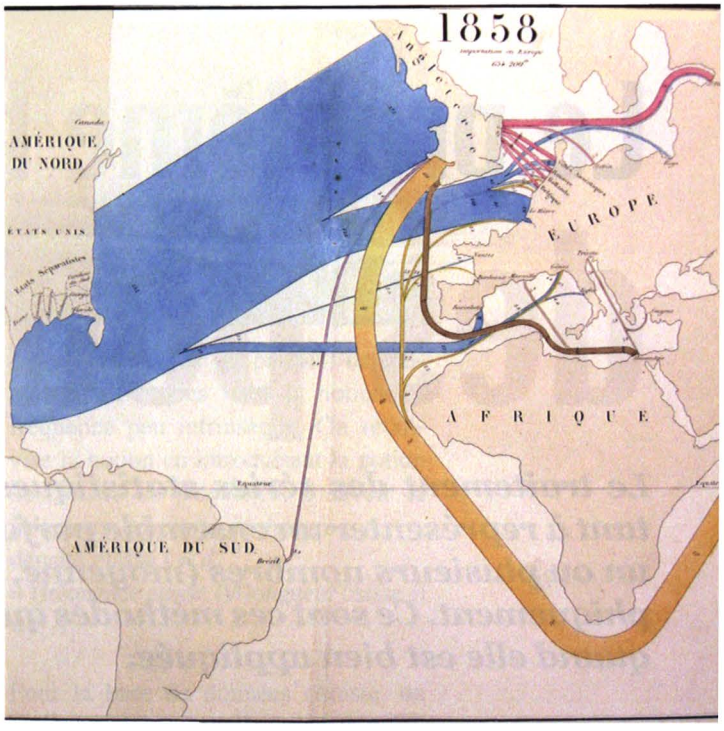


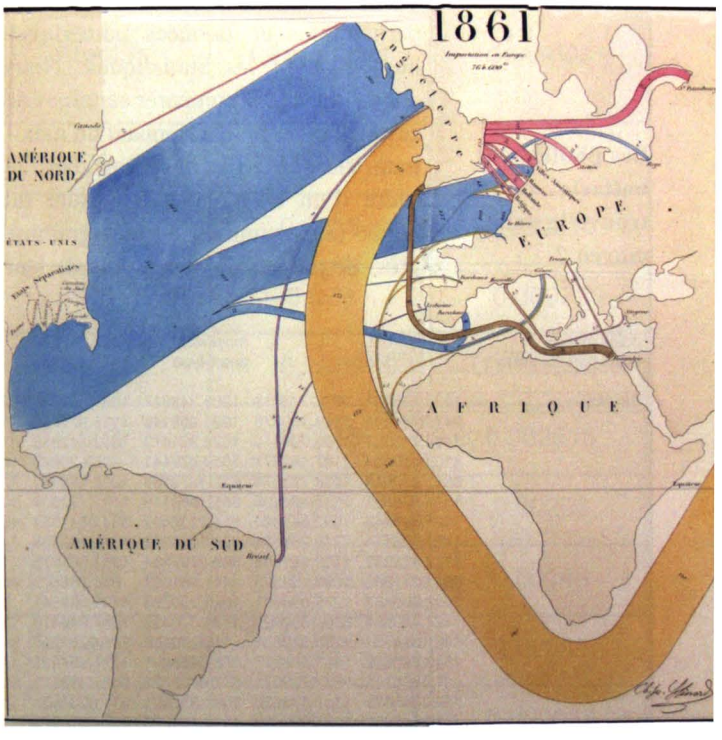
4. Mouvement des voyageurs sur les principaux chemins de fer de l'Europe en 1862.

viaire ou les approvisionnements anglais en coton (figure 4 et 5).

En 1861, les travaux de Minard sont présentés à Napoléon III (un honneur surprenant pour un ingénieur de modeste extraction), qui les apprécie et félicite leur auteur.
Les lecteurs de *Tangente* qui habitent Paris peuvent consulter les documents à la bibliothèque des Ponts.

P. B.





5. L'effet de la politique sur les échanges internationaux est illustré sur cette carte de Minard qui représente le commerce du coton en Europe en 1856 et de nouveau en 1862, peu après le déclenchement de la guerre de Sécession aux États-Unis : le blocus de Grant sur les exportations de coton brut du Sud, destiné à ruiner les Confédérés, a permis l'essor de ce commerce avec l'Inde.

Bibliographie
M. Friendly, Visions and Re-Visions of Charles Joseph Minard. Journal of Educational and Behavioral Statistics, 27(1) : 31-51, 2002.
Edward Tufte, The Visual Display of Quantitative Information, 2001.
http://www.19e.org/personnages/france/M/minard/1.htmhttp://www.math.yorku.ca/SCS/Gallery/minbib.html

par Daniel Justens

Le problème horrifique des poulets

Le traitement des séries statistiques est une gageure consistant à représenter un ensemble parfois énorme de données par un ou plusieurs nombres (moyenne, variance, ...) ou bien graphiquement. Ce sont ces méthodes que donnent la statistique ... quand elle est bien appliquée.

Les poulets en batterie : où se trouve le poulet moyen ?

La statistique regroupe l'ensemble des méthodes permettant de traiter certaines suites particulières de données numériques homogènes : *les* statistiques. Nous allons illustrer et comparer certaines de ces méthodes en leur donnant un maximum de sens. Plaçons-nous dans le cadre d'un exemple : la mesure du poids (en grammes au décigramme près) de poulets d'élevage lors de leur abattage après 8 semaines d'engraissement. Le contexte agronomique n'est pas limitatif.

Notons x_1, x_2, ... x_n les mesures effectuées dans un ordre quelconque. Pour une exploitation avicole de taille raisonnable, n est, par livraison à l'abattage, de l'ordre de 2000. Chacune de ces livraisons est un échantillon. On peut mettre ces résultats dans un tableau (illisible) sur 10 colonnes de 200 lignes ! Comment traiter, résumer et interpréter une telle masse de données ?

poids moyen :	1085		moyenne	1085,532649	
écart-type du poids	145		écart-type	147,3221964	
Simulation	1039,106458	987,7745113	1093,343817	1061,805433	923,6521999
	947,3784994	1344,160376	1026,958149	1156,977796	1190,696333
	1104,716337	1195,794444	1020,391479	1002,342515	1120,668371
	810,6469584	1187,192577	1059,879441	1339,26873	1050,727763
	951,7940942	1034,723079	912,3573106	1209,886649	1215,328337
	1516,876327	1217,494908	885,7669149	1190,722609	1013,520336
	1121,346443	1092,545024	824,2206409	884,8841354	1046,269739
	1121,405174	1219,958684	926,6098072	1046,432218	1155,679901
	1172,123232	1102,261905	889,2792834	1203,681975	1102,899696
	992,0221363	1093,198893	1117,961687	938,868525	1001,358339
	1002,993487	1275,039641	1023,230763	1191,069541	989,306727
	1124,409551	859,1659525	1134,370425	1087,346419	1232,044133
	896,1384133	1235,259835	1169,60486	714,9261258	998,3365168
	1049,646652	1447,668277	858,5456891	1211,567852	1194,388566
	821,5806703	957,6729921	977,6400121	902,4383319	1015,245605
	724,6083675	1187,873638	1091,830901	941,0234838	1181,655724
	1129,148352	1053,033038	1419,494693	997,6910971	1116,095054

Classement et présentation clairs des données

Commençons par réunir les valeurs *proches* (regroupement par *classes*), considérées comme étant du même ordre de grandeur et construisons un tableau simple sur base de

règles arbitraires, mais raisonnables : choix du nombre de classes, de leurs longueurs (pas nécessairement égales), des points de subdivision interclasse et, enfin, d'intervalles ouverts à droite et fermés à gauche (pour qu'une observation n'appartienne qu'à une seule classe). On compte le nombre d'observations de chaque classe et on calcule sa fréquence (nombre d'observations de la classe divisé par nombre total d'observations). Mais le choix de classes arbitraires (et parfois) de longueurs différentes rend la notion de fréquence peu intrinsèque. On relativise la notion en introduisant la notion de *densité de fréquence*, à savoir la *fréquence par unité de mesure* :

densité classe $i = d_i$

= (fréquence classe i)/(longueur classe i)

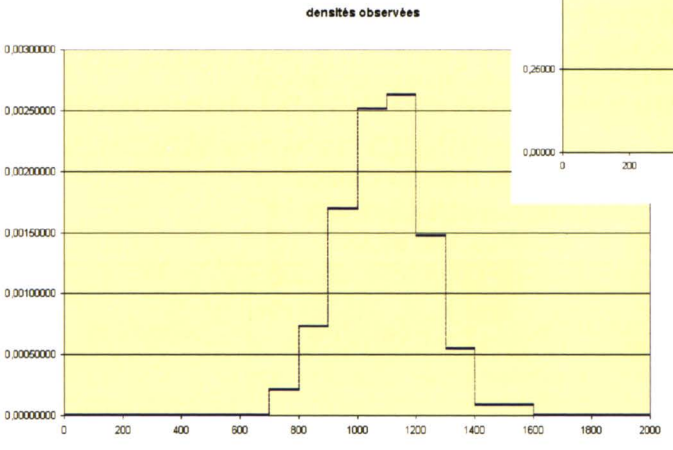
$= f_i / l_i$

Pour la base de données choisie, on obtient un tableau déjà plus lisible :

borne inférieure	borne supérieure	effectifs	fréquences	densités
[500	700[	3	0,0015	0,0000075
[700	800[	43	0,0215	0,000215
[800	900[	146	0,073	0,00073
[900	1000[	339	0,1695	0,001695
[1000	1100[	503	0,2515	0,002515
[1100	1200[	526	0,263	0,00263
[1200	1300[	296	0,148	0,00148
[1300	1400[	110	0,055	0,00055
[1400	1600[	34	0,017	0,000085

La notion de *densité observée* peut se transformer en fonction : on estime la fréquence par unité de mesure au voisinage de chaque valeur x appartenant à la classe C_i par la densité de cette classe : $f(x) = d_i$ *si et seulement si* x *appartient à* C_i

Ceci conduit, graphiquement, à la représentation d'une fonction en escaliers présentant globalement la dispersion des observations :



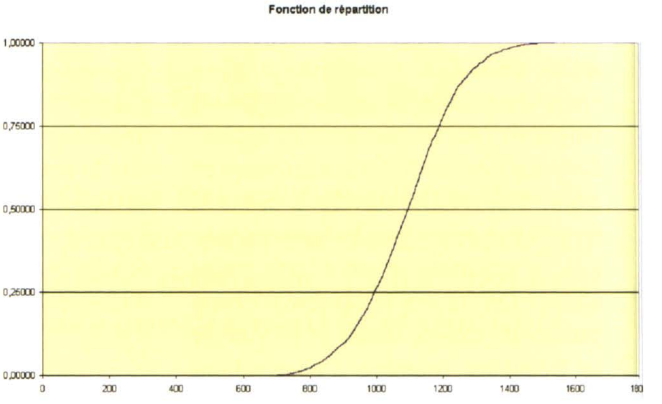
On vérifie que la fréquence d'un intervalle quelconque $[a, b[$ est donnée par l'intégrale de la fonction densité sur cet intervalle : une intégrale élémentaire de fonction en escaliers. Mais, le regroupement fait perdre une partie de l'information (qu'en est-il de la dispersion des observations dans chaque classe ?). On classe alors les observations par ordre croissant. La base de données non décroissantes est notée

$x_{(1)}, x_{(2)}, \ldots x_{(n)}$ avec $x_{(i)} \leq x_{(i+1)}$.

En associant à chaque observation une fréquence uniforme $(1/n)$, on construit la *fonction de répartition observée* (également en escaliers mais cela se voit moins) en définissant :

$F(x)$ = fréquence des observations $< x$

Graphiquement et pour la même base de données, cette fonction est :



Les notations f et F ne sont pas arbitraires, en basant F sur le tableau de valeurs regroupées (et non plus sur les observations), et en considérant des densités uniformes par classes, on peut vérifier que la fonction F ainsi construite est bien une primitive de f.

Résumé des données : paramètres de centralité et de dispersion

Voilà nos données regroupées et (re)présentées. Peut-on les résumer au moyen d'une constante ? Remplacer une suite de nombres variables $x_1, x_2, \ldots x_n$ par un nombre (que nous notons provisoirement a) induit une nouvelle statistique composée des erreurs $e_i = x_i - a$. Pour un résumé « a » bien choisi, central, certaines de ces erreurs sont positives, d'autres négatives. Le « bon » choix pour a doit cor-

respondre à un certain critère de minimum global des erreurs. Il convient donc d'une part de rendre toutes ces erreurs positives, d'autre part de les agréger.

Moyenne, variance et intervalles de confiance

Ici encore intervient l'arbitraire. Les travaux de Gauss publiés au début du XIX^{e} siècle ont imposé une manière unique de procéder à laquelle une alternative crédible ne fut proposée qu'en fin de XX^{e} siècle (l'abandon du critère de Gauss a ouvert les portes de la *statistique robuste*, voir l'article de Valérie Henry dans ce numéro).

Il faut dire que la méthode est efficace, simple à mettre en œuvre et qu'elle garantit existence et unicité de la solution *a*. Pour Gauss, l'agrégat le plus opportun est constitué par la *somme des carrés des erreurs* :

$$ET(a) = \sum_{i=1}^{n} \left[x_i - a \right]^2$$

Cette fonction (parabole) de *a* possède un minimum unique obtenu après annulation de la dérivée première.

$$a = \bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

Cette valeur porte le nom de *moyenne arithmétique observée* d'une série statistique. Dans le cas de notre exemple, la moyenne vaut 1090,3 grammes qui, on le voit sur le graphique des densités, est bien une valeur centrale. Au passage, le lecteur qui a détaillé les calculs aura remarqué que la somme des carrés des erreurs est minimale lorsque la somme des erreurs est nulle. Signalons que la présence de quelques observations très éloignées des autres altère très fort ce paramètre, exerçant sur lui une espèce d'attraction qui est une conséquence de la mise au carré des erreurs dans l'agrégat de Gauss. On dit que la moyenne est sensible aux valeurs extrêmes. Ce manque de robustesse du résumé *moyenne* est l'un de ses principaux défauts. Le minimum de l'erreur totale est

$$ET_{min} = \sum_{i=1}^{n} \left[x_i - \bar{x} \right]^2$$

Pour obtenir une mesure de la *dispersion* des observations, on peut répartir l'erreur totale minimale sur le nombre d'observations ; on arrive à la notion classique de *variance* :

$$s^2 = \frac{1}{n} \sum_{i=1}^{n} \left[x_i - \bar{x} \right]^2$$

On peut représenter moyenne et variance d'un échantillon sur le graphe de *ET(a)*.

Mais les erreurs relatives à la moyenne ne sont pas indépendantes. La statistique des e_i retenue ne contient que $(n-1)$ renseignements indépendants. On peut décider de répartir l'erreur totale sur ces $(n-1)$ valeurs pour introduire une *variance non biaisée* :

$$s'^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left[x_i - \bar{x} \right]^2$$

En moyenne, cette variance « échantillon » est bien égale à la variance de la population. L'unité de mesure de la *variance* et de la *variance non biaisée* est le carré de l'unité initiale. Pour retrouver l'unité de départ (dans notre exemple le gramme), on extrait la racine carrée de la variance pour définir les écarts-types :

$$s = \sqrt{s^2} \qquad s' = \sqrt{s'^2}$$

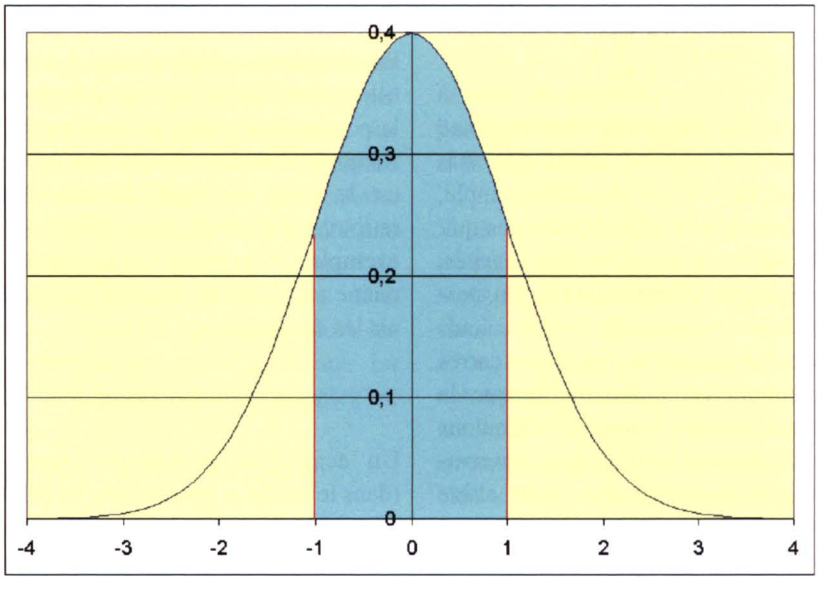
Un écart-type se calcule aisément (dans le cas de notre exemple on trouve s' = 143,4, mais la précision des mesures étant le gramme, on peut rete-

nir s' = 143). Mais quelle est la signification du nombre obtenu ? Clairement : un écart-type est une mesure de la *dispersion* des observations. Mais il demeure bien difficile à interpréter, étant la résultante de quatre opérateurs successifs, à savoir la *racine* de la *moyenne* des *carrés* des *erreurs*. De plus, la répartition des observations à gauche et à droite de la moyenne peut être fortement asymétrique. Le paramètre ne rend absolument pas compte de cette éventualité, réalisant une sorte de *moyenne de la dispersion*.

Son succès est dû à Gauss. Le théorème-central-limite démontre que tout phénomène complexe résultant de la somme d'un grand nombre de causes indépendantes est distribué de manière unique : la distribution normale ou distribution de Gauss. Dans ce cas particulier, l'écart-type prend tout son sens. Il permet de construire des intervalles de fréquences théoriques calculables *a priori* au moyen des seuls moyenne et écart-type. On fait donc très souvent l'hypothèse simplificatrice de normalité des mesures observées (très souvent sans raison objective). La distribution gaussienne se définit sur l'ensemble des réels à partir de sa densité

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-1}{2}\left(\frac{x-m}{\sigma}\right)^2}$$

expression dans laquelle m représente la moyenne de la variable et σ son écart-type. L'allure générale de la courbe (symétrique par rapport à $x=m$) est bien connue et possède un maximum en $x = m$ qui est la valeur de densité maximale, et deux points d'inflexion en $x = m - \sigma$ et $x = m + \sigma$. Ces derniers séparent les observations rares (concavité vers le haut), situées bien à gauche ou bien à droite de la moyenne (à plus d'un écart-type), des observations fréquentes (concavité vers le bas) constituées des valeurs centrales. Ceci est illustré sur le graphique suivant représentant une loi normale dite *standard* c'est-à-dire de moyenne nulle et d'écart-type 1. Les points d'inflexion sont visibles aux points d'abscisses –1 et 1.



La densité normale n'est pas intégrable au moyen de fonctions habituelles. Mais ceci n'est qu'une question de point de vue. On peut très bien introduire la fonction

$$\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi\sigma}} e^{\frac{-1}{2}\left(\frac{t-m}{\sigma}\right)^2} dt$$

qui n'est pas plus difficile à tabuler qu'un sinus ! Avec cette notation, la probabilité d'observer une valeur comprise entre a et b ($a<b$) est donnée par $\Phi(b) - \Phi(a)$.

On vérifie par calcul numérique approché que l'intervalle de *concavité négative*, regroupant les valeurs les plus fréquentes est de probabilité 68.26 %. Les valeurs « petites », inférieures à m - σ, sont assez rares (probabilité 15.87 %), tout comme comme les observations « plutôt grandes », supérieures à m + σ.

On considère néanmoins que la fiabilité de l'intervalle central est insuffisante : il ne couvre en gros que les 2/3 des observations. On décide de fixer une fiabilité standard et de calculer à quel intervalle centré en m, symétrique, celle-ci correspond. On vérifie que l'intervalle

$$I_{0.95} = [m - 1.96\,\sigma;\ m + 1.96\,\sigma]$$

est de fiabilité 95 % et que la probabilité de

$$I_{0.99} = [m - 2.56\,\sigma\ ;\ m + 2.56\,\sigma]$$

est de 99 %. C'est à ce dernier intervalle que l'on se réfère dans le langage courant : on dit qu'une mesure x est *significativement différente* de la tendance attendue m lorsque x est extérieur à $I_{0.99}$, c'est-à-dire lorsque $|x-m| > 2.56\,\sigma$.

Voici les chiffres communiqués par les services de la statistique et intéressant la période comprise entre le 2 juillet et le 4 septembre : 545 285 ; 6 282 826 ; 1 285 938 743,601 ; 601 ; 602 ; 603 ; 604 ; 605 ;106 ; 206 ; 306 ; 406 ; 506 ; 983 ; 882 ; 780 ; 680 ; 579.
Nous ne savons pas à quoi se rapportent ces chiffres, mais nous sommes heureux de les communiquer à nos lecteurs qui auront ainsi toute latitude de les adapter suivant leur goût ou leur appréciation.
Pierre Dac

Médiane, quartiles et boîtes à moustaches

L'agrégat de Gauss n'est pas le seul. Pour rendre les erreurs positives, on peut aussi en prendre la valeur absolue. En conservant l'idée d'agrégat par sommation on arrive à une autre erreur totale :

$$ET^*(a) = \sum_{i=1}^{n} |x_i - a|$$

La fonction valeur absolue n'étant pas dérivable en son minimum, il faut raisonner autrement. On procède au classement des observations par ordre croissant comme nous l'avons fait plus haut pour la construction de la fonction de répartition. On peut écrire :

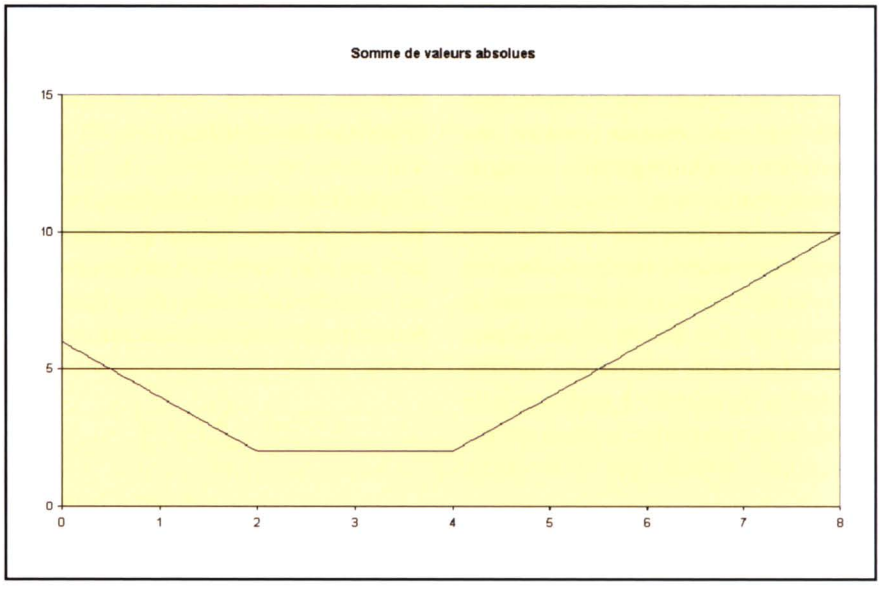
$$ET^*(a) = \sum_{i=1}^{n} |x_{(i)} - a|$$

On regroupe les termes deux par deux en considérant une suite d'intervalles emboîtés : on réunit le premier et le dernier, le deuxième et l'avant dernier

et ainsi de suite. Il convient de traiter différemment les cas n pair et impair. Dans le cas pair, la somme devient :

$$ET^*(a) = \sum_{i=1}^{n/2}\left[\left|x_{(i)} - a\right| + \left|x_{(n-i+1)} - a\right|\right]$$

On étudie mathématiquement la fonction $[|x-a|+|x-b|]$ $(a < b)$ pour constater qu'elle est affine par morceaux : une première demi-droite de pente (-2) jusqu'en x = a, un segment horizontal (valeur $(b-a)$) entre a et b et une demi droite de pente 2 après $x = b$. Le minimum est donc réalisé pour tout point de l'intervalle $[a, b]$. Le graphique qui suit illustre cette propriété ($a = 2$ et $b = 4$) :



Les intervalles étant emboîtés, on constate que le minimum global de ET* est réalisé pour toute valeur de l'intervalle $\left[x_{(n/2)}, x_{(n/2+1)}\right]$

Par convention, on choisit le centre de cet intervalle pour résumé, que l'on appelle *médiane* :

$$x = \frac{x_{(n/2)} + x_{(n/2+1)}}{2}$$

Pour n impair, l'erreur totale comprend un terme isolé, relatif à la valeur *centrale* des observations classées par ordre non décroissant, et constitué d'une valeur absolue. Pour obtenir un minimum global, il suffit d'annuler cette valeur et de poser

$$x = x_{\left(\frac{n+1}{2}\right)}$$

Dans le cas de l'exemple, la médiane vaut 1093,4. Elle est proche de la moyenne. On vérifie que la médiane est moins sensible aux valeurs extrêmes que la moyenne. C'est un paramètre plus *robuste*. On peut introduire une mesure de dispersion, l'écart-médian, constitué de l'erreur totale minimale répartie sur le nombre d'observations en posant :

$$EM = \frac{1}{n}\sum_{i=1}^{n}\left|x_{(i)} - x\right|$$

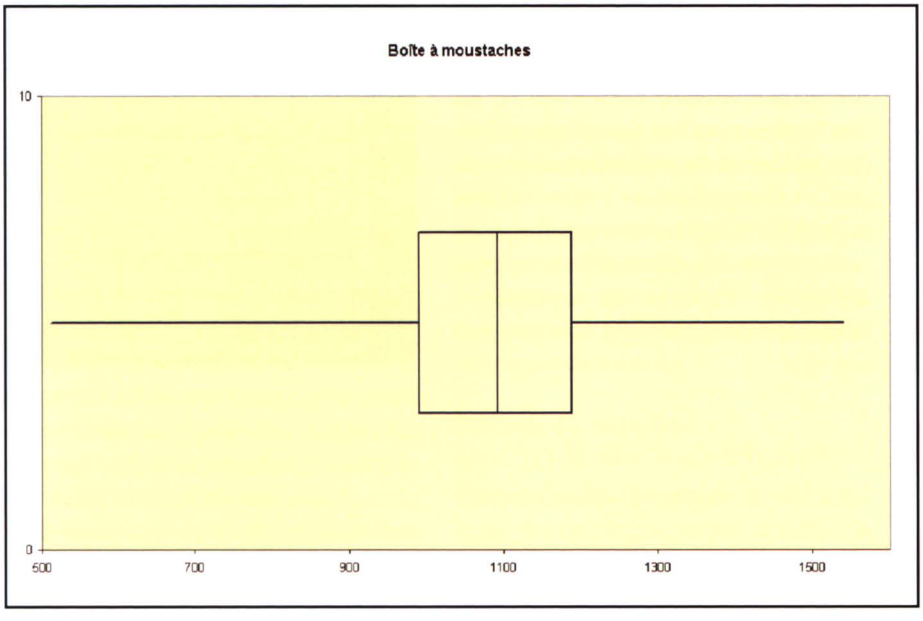
Dans notre exemple, EM = 114,6. Encore une fois comment interpréter cette valeur ? Les travaux de Tucker en analyse exploratoire de données, il y a une trentaine d'années, suggèrent une

manière plus parlante et plus simple de procéder : la construction d'une boîte à moustaches !
Après calcul de la médiane d'une série, on sépare cette dernière en deux demi-séries ordonnées, complétées chacune par la médiane. Dans le cas pair, les nouvelles demi-séries comprennent chacune $(n/2 +1)$ valeurs, et, dans le cas impair, $(n+1)/2$. On refait le même travail de calcul de médiane pour chaque demi-série, construisant ainsi pour celles des valeurs plus petites un *premier quartile pragmatique* et, pour la deuxième série un *troisième quartile pragmatique*.

On trace une *boîte centrale*, allant du premier au troisième quartile, comprenant (en gros) la moitié des observations. On complète la boîte en signalant la valeur médiane par une verticale et en lui ajoutant deux moustaches allant de l'observation la plus petite au premier quartile et du troisième quartile à l'observation la plus grande.

La présentation claire et le résumé intelligent d'une base de données importante est une gageure. Comment fournir un maximum d'informations de manière simple et concise ? Les différentes méthodes que nous avons esquissées tentent de répondre à cette question. Chacune a ses qualités. Chacune a ses défauts. À l'utilisateur de choisir dans cet arsenal, la méthode convenant le mieux à ses besoins. Et à lui de la suivre honnêtement.

D.J.



par Alain Zalmansky

Plan d'expérience

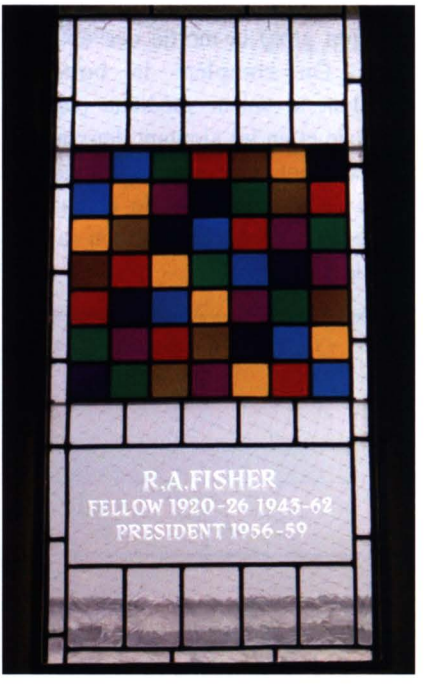
La mise au point de plans d'expérience a permis de définir les paramètres importants d'un phénomène avec un nombre limité d'expériences.

De la chimie à l'agronomie, de la physique à la gastronomie, toutes les sciences expérimentales sont amenées à chercher à établir des liens de causes à effets et pour cela doivent procéder à une série d'expériences selon un protocole bien défini. En 1627, Francis Bacon faisait déjà macérer des grains de blé dans neuf solutions différentes afin d'étudier leur effet sur la rapidité de germination.

Ronald Aylmer Fisher (1890-1962)

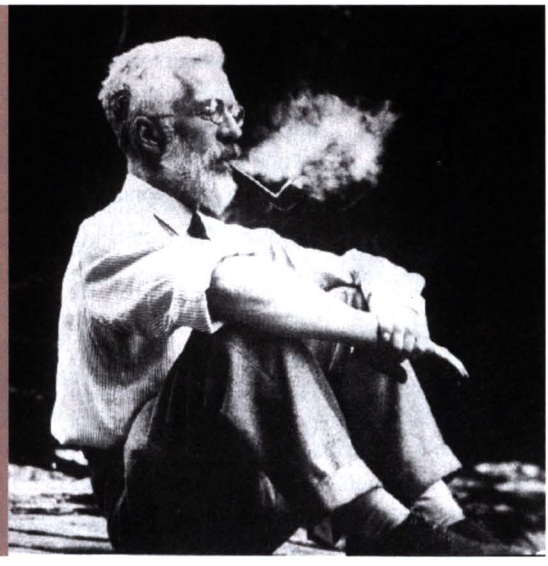
Sir Ronald Fisher (1890-1962), un des fondateurs de la statistique inductive moderne, introduira la notion de plan d'expérience, lors d'une recherche d'augmentation de rendements agricoles mettant en jeux type d'engrais, variétés de traitement, méthodes de cultures et composition des sols.

Consulter un statisticien après la fin d'une expérience, c'est lui demander un examen post mortem. Il peut quelquefois dire pourquoi l'expérience est morte.
Ronald Fisher



Vitrail de Caius College, de l'Université de Cambridge, en l'honneur de Ronald Fisher, avec un carré latin qu'il utilise dans ses plans d'expérience. On remarquera que les cases sont colorées en 7 couleurs différentes de façon qu'il y ait une case de chaque couleur et une seule sur chaque ligne et chaque colonne.

Ronald Fisher en 1956.
Côté tabac, il donne le mauvais exemple, et en bon fumeur impénitent, a argumenté que les corrélations entre le cancer et le fait de fumer n'étaient pas convaincantes. Il avait tort. Des exceptions statistiques. Ronald Fisher s'est beaucoup intéressé à la théorie de l'Évolution et on l'a surnommé le second Darwin. Une de ses phrases favorites était : « La sélection naturelle est un mécanisme engendrant l'improbable à son plus haut degré. »



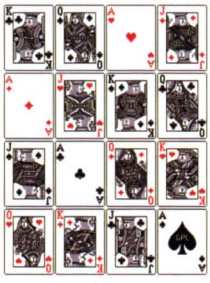
L'étude d'un phénomène peut, le plus souvent, être schématisé de la manière suivante : on s'intéresse à une grandeur, Y, nommé par la suite *réponse*, qui dépend d'un grand nombre de variables, X_1, X_2, ..., X_n, que nous dénommerons par la suite *facteurs*.
La modélisation mathématique consiste à trouver une fonction f telle que

$$Y = f(X_1, X_2, \ldots, X_n).$$

Une méthode classique d'étude « toutes choses égales par ailleurs » consiste à mesurer la réponse Y pour plusieurs valeurs de la variable X_i tout en laissant fixe la valeur des $(n-1)$ autres variables. On poursuit alors cette méthode pour chacune des variables. Ainsi, par exemple, si nous avons 4 variables et si l'on décide de donner 5 valeurs expérimentales à chacune d'elles, nous sommes conduits à effectuer $5^4 = 625$ expériences.
Ce nombre élevé dépasse les limites de faisabilité tant en coût qu'en durée. Il faut donc réduire le nombre d'expériences à effectuer sans pour autant perdre sur la qualité des résultats recherchés.
Il fallait donc que ces recherches soient organisées, d'où l'utilisation de « plans d'expériences » dont la grande novation était de proposer une *expérimentation factorielle*, c'est-à-dire dans laquelle tous les facteurs varient simultanément, avec traitement des résultats à l'aide de structures mathématiques comme la régression linéaire multiple, l'analyse de variance, la combinatoire, les carrés latins ou gréco-latins.
L'avènement de l'informatique permet maintenant la confection et l'utilisation de plans d'expériences conformes à un cahier des charges précises et conduit à une modélisation des phénomènes étudiés, grâce à des outils mathématiques sophistiqués. Loin de nuire à la rigueur des conclusions, la réduction à des expériences bien choisies, permet d'aboutir, contrairement à ce que l'on pouvait imaginer intuitivement, à des résultats plus précis, plus fiables et mieux exploitables. La chimie, l'agronomie et la médecine sont les plus gros utilisateurs de cette technique de travail.

Méthode du carré latin

Un cas d'analyse limitée de la variance à triple entrée est fréquemment utilisé en pratique dans la mesure où il ne nécessite qu'un budget d'études assez limité. Il porte le nom de carré latin.



Un carré latin est une grille carrée de n lignes et n colonnes où les n^2 cases contiennent n lettres, nombres ou symboles, qui apparaissent une et une seule fois sur chaque ligne et dans chaque colonne. Chacune des lignes ou colonnes est constituée par la permutation des n éléments. La table de Pythagore et le sudoku sont deux exemples de carrés latins.

Dans un plan d'expérience en carré latin, les trois facteurs contrôlés vont intervenir avec un même nombre k de niveaux chacun (kA = kB = kC). Le plan d'expérience peut être représenté par un carré divisé en kA colonnes correspondant aux kA niveaux du facteur A et kB lignes (kB = kA) correspondant aux kB niveaux du facteur B. À toute case du carré Aa Bb est associé un niveau Cc du facteur C, de façon que sur chaque ligne et sur chaque colonne du carré apparaisse une seule fois chacun des kC niveaux du facteur C.

Le croustillant du pain

Ainsi on voudrait étudier le « croustillant » par trop irrégulier d'un pain en fonction de divers facteurs de fabrication - farine, machine à pain, température en fin de cuisson. On fabrique donc 25 pains portant sur 5 qualités de farine (repérées 1, 2, 3, 4, 5), 5 machines à pain (repérées A, B, C, D, E), 5 températures de fin de cuisson (repérées a, b, c, d, e). L'association de ces paramètres est effectuée conformément au tableau ci-après de telle sorte que le croisement des différents facteurs soit bien assuré.
On aboutit à la conception d'un plan d'expérience du type :

Farines	Machines à pain				
	A	B	C	D	E
a	1	3	5	4	2
b	5	4	2	1	3
c	2	1	3	5	4
d	4	2	1	3	5
e	3	5	4	2	1

Températures de fin de cuisson (dorage)

Les lignes correspondent à la température de fin de cuisson, les colonnes à la machine à pain, les qualités de farine sont indiquées dans les cases, de façon à ce que chaque farine apparaisse une fois et une seule dans chaque ligne et dans chaque colonne (d'où le nom de méthode du carré latin).

Dureté	Machines à pain				
	A	B	C	D	E
a	164	169	170	171	172,5
b	169	172	174,5	170	170
c	173	170,5	166	172	169
d	169	170,5	166	166	168
e	166	174	173	174	169,5

Températures de fin de cuisson (dorage)

Toutes choses égales par ailleurs, et après refroidissement, on effectue une mesure que nous appellerons dureté, caractéristique du « croustillant » de chaque pain obtenu.
Sans entrer dans les détails classiques d'exploitation d'un tel tableau (changement de variable, calcul de variations, quotients, seuils de confiance…), on peut dresser un tableau récapitulatif :

Variation	Somme des carrés	Degré de liberté	Quotients
due à la température	155,8	4	$v_1 = 38,9$
due à la machine	101,4	4	$v_2 = 25,4$
due à la farine	403,1	4	$v_3 = 100,7$
résiduelle	116,3	12	$v_R = 9,7$
totale	776,6		

On déduira, entre autres des rapports v_1/v_R, v_2/v_R et v_3/v_R, qu'on doit rejeter l'hypothèse de l'influence du facteur « machine à pain ». Le facteur farine est influent et l'on conseillera d'être attentif en priorité au facteur « température » pour conserver obtenir un croustillant et la constance de ce dernier.

A. Z.

Nous sommes tous fichés !

Les fichiers de données personnelles

Le fichier de police Edvige, créé par le décret 2008-632 du 27 juin 2008, succède au FRG (Fichier des renseignements généraux) et permet un traitement automatisé de données à caractère personnel. Ces six lettres sibyllines signifient en fait Exploitation documentaire et valorisation de l'information générale. De nombreuses voix se sont élevées contre la nature des données qui peuvent y être renseignées (orientation sexuelles d'un individu, sa santé, ses handicaps, origines raciales ou opinions politiques...). Face à la pression des associations, Edvige a été remplacé par le non moins critiqué Edvirsp (Exploitation documentaire et valorisation de l'information relative à la sécurité publique). Ce dernier a également pour vocation à ficher toutes les personnes « dont l'activité individuelle ou collective indique qu'elles peuvent porter atteinte à la sécurité publique ».

Des fichiers contestés

Comme Edvige ou Edvirsp, Cristina (Centralisation du renseignement intérieur pour la sécurité du territoire et des intérêts nationaux) est sur la sellette. Classé secret-défense, il est relatif au terrorisme et à l'espionnage.

D'autres fichiers sont, eux, illégaux. Par exemple celui de la batellerie. Mais également le logiciel Ardoise (Application de recueil de la documentation opérationnelle et d'information statistique sur enquêtes), qui succède au Logiciel de rédaction des procédures (LRP). Selon la CNIL (Commission nationale de l'informatique et des libertés), *« de telles applications (...) ne peuvent être créées que par un décret en Conseil d'État après avis de la CNIL »*. Ardoise doit à terme alimenter la future base de données Ariane (Application de rapprochement, d'identification et d'analyse pour les enquêteurs).

Le ministère de l'Intérieur (sous la responsabilité des Archives nationales) réfléchit à l'avenir de ces informations. Une vaste opération de nettoyage des fichiers a débuté, qui devrait voir 60 millions de fiches détruites car obsolètes, erronées, périmées ou illégales.

Actuellement en France, au moins quarante-cinq fichiers de gendarmerie et de police sont en activité. Faisons le compte : 60 millions d'entrées dans le FAR (Fichier alphabétique de renseignement), registre de la gendarmerie dans lequel figure tout conflit de voisinage ou toute possession d'un animal dangereux. 5 millions de noms dans le fichier policier Stic (Système de traitement des infractions constatées), et 3 millions dans le Judex (système judiciaire d'exploitation et de documentation, équivalent du Stic pour la gendarmerie). 2,5 millions d'entrées figurent dans Edvige, actuellement suspendu. 300 000 noms se trouvent dans le FPR (Fichier des personnes recherchées), et 52 000 dans le fichier de la batellerie, qui recense des mariniers, leurs employés, leur famille, leur bateau... Malgré les inévitables redondances, on pourrait croire que nous sommes tous bel et bien fichés !

Au-delà des chiffres, les fichiers interpellent. *« On tend à amalgamer des individus représentant un risque potentiel pour l'État, et ceux dont l'activité s'avère indispensable à son fonctionnement dans un cadre démocratique »*, résume Pierre Piazza, universitaire spécialiste des techniques d'identification policières.

par P. Dossantos-Uzarralde, H.-P. Jacquet, F. Lavallou

La méthode Monte-Carlo

L'ordinateur a permis un formidable essor des techniques numériques : la simulation statistique prend le pas sur l'expérimentation et la méthode Monte-Carlo y a une place de choix.

Le nom de la *Méthode Monte-Carlo*, proposé par les scientifiques du projet Manhattan, fait allusion aux jeux de hasard pratiqués à Monaco, plus précisément à la roulette qui produit les meilleures suites de nombres aléatoires. Le récit traditionnel de l'histoire du projet Manhattan réserve les premiers rôles à de grands noms de la physique nucléaire : Fermi, von Neumann, Ulam, Metropolis. Leurs travaux étaient liés à la simulation directe de diffusion aléatoire des neutrons dans les matériaux fissiles. Le mérite de l'invention est le plus souvent attribué à Stanislaw Ulam, un mathématicien polonais qui voulait calculer la probabilité de gagner une partie de solitaire. Il a développé de nouveaux algorithmes et transformé des problèmes non aléatoires en problèmes aléatoires afin de les résoudre par le biais de simulations statistiques.

Tout montre que Dieu est un vrai joueur, et que l'Univers est un grand casino où les dés sont jetés et où la roulette tourne à tout moment. Stephen Hawking (cosmologiste)

L'approche Monte-Carlo consiste à déterminer la solution en simulant directement le problème initial par la génération de nombres aléatoires. Dans un problème déterministe, si l'état du système est parfaitement défini, son comportement est prédictible. Il en est ainsi pour un système de particules obéissant aux lois de la mécanique classique. On peut cependant traiter certains paramètres comme des variables aléatoires, transformant ainsi le problème déterministe en un problème probabiliste que l'on résout numériquement. Cette approche est intéressante lorsque le nombre de paramètres est excessivement grand.

Intégrations statistiques

Pour comprendre la méthode Monte-Carlo, nous nous restreindrons à une seule dimension : « Comment calculer l'intégrale entre *a* et *b* de la fonction *f* ? »

Une intégrale est, approximativement, proportionnelle à la somme des valeurs de la fonction prises sur des points régulièrement répartis dans le domaine d'intégration. Remplaçons « régulièrement répartis » par « répartis selon une loi de probabilité uniforme », et nous avons notre première simulation de Monte-Carlo. Le calcul se fait alors en créant n nombres aléatoires ξ_i compris entre a et b et distribués de façon telle que toutes les valeurs comprises entre a et b sont équiprobables. À chaque valeur du nombre tiré ξ_i on associe la valeur de la fonction $f(\xi_i)$ que l'on multiplie par $(b–a)/n$. L'estimation de la valeur de l'intégrale est :

$$I = \lim_{n\to\infty} \frac{b-a}{n} \sum_{i=1}^{n} f(\xi_i)$$

$$\cong \frac{b-a}{n} \sum_{i=1}^{n} f(\xi_i)$$

L'intérêt principal de la méthode est de calculer des intégrales de grandes dimensions.

En comparant les performances d'une méthode d'intégration numérique déterministe et celles de la méthode de Monte-Carlo pour des dimensions de plus en plus grandes, il se trouvera une dimension au-delà de laquelle la méthode de Monte Carlo donnera, le plus souvent, une valeur plus proche de la valeur exacte de l'intégrale que celle obtenue par l'approximation déterministe. Néanmoins, il existe un prix à payer pour ces avantages. En raison de la nature aléatoire de l'échantillonnage de $f(x)$, des simulations de Monte Carlo sur un même problème, dans des conditions identiques, produiront des valeurs différentes qui suivront une loi de probabilité ayant une certaine variance (**moyenne des carrés des écarts à la moyenne**).

Là où une méthode déterministe produit une approximation, une simulation de Monte Carlo fournit une estimation. La différence, pour être subtile, n'en est pas moins importante : il est souvent possible de trouver une borne supérieure à l'erreur d'une approximation, ce qui n'est généralement pas le cas pour une estimation. Malgré tout, une conséquence de la loi des grands nombres indique que pour un seuil ε donné, la probabilité que cette erreur soit supérieure à ε peut être rendue aussi petite que l'on veut en augmentant le nombre de tirages. Inconvénient majeur : les simulations de Monte Carlo sont coûteuses en temps calcul.

Ce type de méthode, utilisée ici pour une illustration pédagogique du concept « d'intégration », porte le nom de *crude Monte Carlo estimator.*

Avec cette méthode, l'erreur décroît comme la racine carrée de la taille n du nombre de réalisations (les ξ_i dans le cas de notre exemple) et ne dépend pas de la dimensionnalité du problème. *A contrario*, la plupart des techniques déterministes d'intégration sont victimes de la « malédiction de la dimensionnalité » : quand on les généralise à des problèmes à dimensions multiples, le coût informatique augmente exponentiellement avec la dimension de l'intégrale. La méthode de Monte Carlo ne souffre pas de ce syndrome.

L'augmentation du nombre de réalisations conduisant à des temps calculs

prohibitifs, la diminution de l'erreur standard nécessite des techniques de réduction de variance.

Réduction de variance

Nous allons illustrer l'efficacité d'une de ces techniques sur un calcul de transport neutronique, particulièrement consommateur de temps calcul. Lors d'une expérimentation de fusion thermonucléaire, des neutrons de forte énergie sont émis dans toutes les directions. Le destin de ces particules libres de toute charge est de rebondir, tels des boules de billard, sur les noyaux des atomes, perdant de l'énergie à chaque interaction pour finir inéluctablement par être capturé à une énergie proche de l'agitation thermique du milieu. Le noyau cible absorbe un de ces neutrons voyageurs en formant un isotope qui,

L'aiguille sur des lattes de parquet

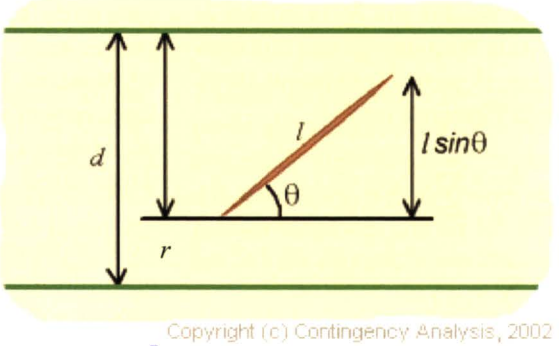
Sur un parquet formé de planches parallèles et équidistantes, on jette des aiguilles au hasard, et on cherche la probabilité pour qu'une aiguille tombe à cheval sur deux planches. Soit d la largeur des planches, et l la longueur de l'aiguille. On suppose l<d. On note r la distance entre le centre de l'aiguille et la jointure de deux planches la plus proche, et θ l'angle de l'aiguille avec la jointure. L'aiguille croise une jointure si et seulement si sinθ/2 > r.

On suppose, ce qui est raisonnable, que les jets sont tels que r et θ suivent des lois uniformes sur respectivement [0, d/2] et [0, π/2], la probabilité de tomber sur une rainure du parquet est alors :

$$\frac{\int_0^{\frac{\pi}{2}} \frac{l}{2}\sin\theta\, d\theta}{\frac{\pi}{2}\frac{d}{2}} = \frac{2l}{\pi d}$$

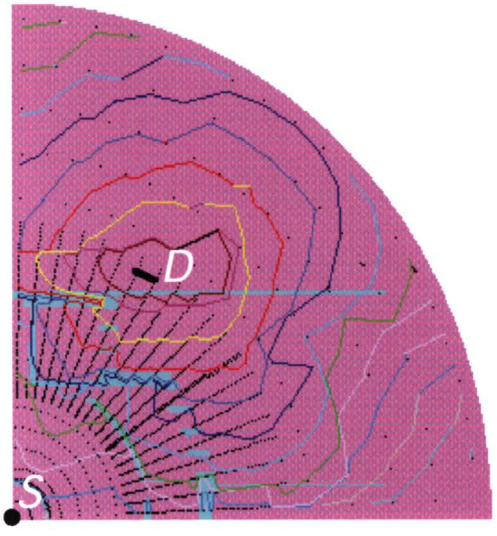
Par un grand nombre d'expériences, on estime de façon empirique cette probabilité, et on en déduit une estimation du nombre π.

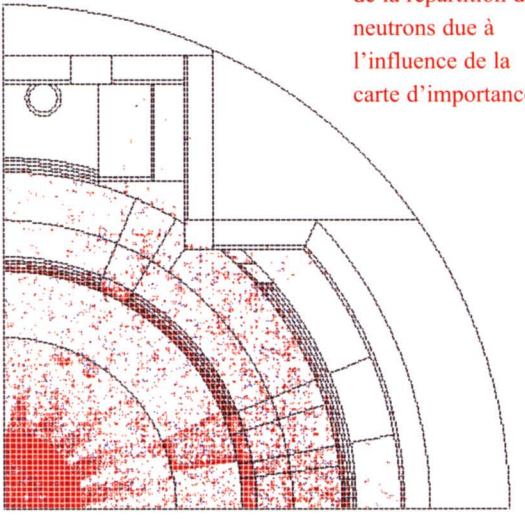
Ce jeu, étudié par Georges-Louis Leclerc, comte de Buffon en 1733 a été une des premières simulations de calcul d'une valeur.



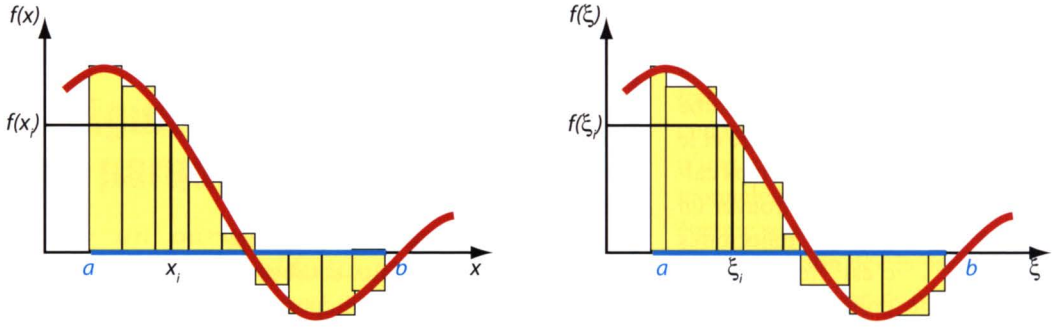
Le comte de Buffon (1707-1788)

souvent, est instable. Cette radioactivité induite est estimée en tout point de l'édifice expérimental en simulant numériquement le cheminement des neutrons incidents. Un grand nombre de ces neutrons virtuels est absorbé avant d'arriver au point de mesure et le calcul de leur histoire est du temps calcul perdu. Il s'agit donc de trouver un moyen de privilégier les trajectoires « utiles » : c'est le rôle de l'échantillonnage préférentiel, *importance sampling* en anglais. Cette méthode de réduction de variance consiste dans son principe à privilégier les régions où la fonction à mesurer possède des valeurs élevées en les suréchantillonant. Pour notre exemple, nous aurons à choisir une distribution, la fonction d'importance, qui s'écartera de la loi de distribution uniforme dans les régions de fort flux neutronique pour diminuer la variance sans pour autant modifier le résultat de notre calcul. On montre mathématiquement (voir encadré) qu'une fonction d'importance proportionnelle au flux de neutrons annule la variance. Dans la pratique, ce flux, qui est l'objet de notre étude, nous est inconnu et toute la difficulté est d'établir cette fonction à l'aide des caractéristiques physiques de l'environnement de la source. À partir de cette fonction d'importance, on construit une carte d'importance de telle sorte que les particules aient tendance à suivre les trajectoires orthogonales aux courbes de niveaux, les lignes d'équi-importance. La carte d'importance « guide » donc statistiquement les particules comme un GPS aide le conducteur pressé à éviter les bouchons. On élimine ainsi les chemins les plus improbables pour, à nombre de particules initiales donné, augmenter la statistique du calcul. Cette expertise que constitue l'établissement d'une carte d'importance peut être vue comme la démarche classique pour tout problème inverse d'introduire de l'information *a priori*. « Le vent n'est favorable que pour celui qui sait où il va » disait déjà Sénèque (-4, +65).





Simulation du transport de neutrons dans une expérimentation du Laser Méga Joule (CEA DAM).
À gauche : carte d'importance pour la source S et le détecteur D avec visualisation des lignes d'équi-importance.
À droite : étape d'un calcul de transport. On note la non-uniformité de la répartition des neutrons due à l'influence de la carte d'importance.



Calcul numérique de la valeur d'une intégrale (à gauche), où l'on sépare l'intervalle d'intégration en valeurs égales et estimation (à droite) de la même intégrale par la méthode de Monte-Carlo où les valeurs de l'abscisse où sont évalués la fonction sont choisies au hasard.

Réduction de variance

La valeur moyenne $\bar{f}$ d'une fonction f sur un domaine D est l'espérance $E_P(f)$ de cette fonction pour la densité de probabilité uniforme p : $\bar{f} = E_P(f) = \int_D f^2(u)\, p(u) du$. La variance de cette estimation est : $V_P(f) = \int_D f^2(u)\, p(u) du - \bar{f}$. Nous cherchons une loi de probabilité q telle que $E_q(f) = E_P(f)$ et $V_q(f) < V_P(f)$. Un simple jeu d'écriture nous donne : $E_P(f) = \int_D f(u)\, p(u) du = \int_D \frac{f(u)\, p(u)}{q(u)}\, q(u) du = E_q\left[\frac{f \cdot p}{q}\right]$ pour toute fonction $q > 0$ telle que $\int_D q(u) du = 1$.

Nous avons donc à choisir la densité q telle que $= V_q\left[\frac{f \cdot p}{q}\right] < V_P(f)$, c'est-à-dire :

$$\int_D \frac{f^2(u)\, p^2(u)}{q^2(u)}\, q(u) du = \int_D f^2(u)\, \frac{p(u)}{q(u)}\, p(u) du < \int_D f^2(u)\, p(u) du.$$

Il suffit de suréchantillonner ($q > p$) pour les fortes valeurs de la fonction pour remplir notre objectif. L'optimum serait atteint pour une distribution q proportionnelle à la fonction $f \cdot p$ puisqu'alors $q = \frac{f \cdot p}{\bar{f}}$ et $V_q\left[\frac{f \cdot p}{q}\right] = V_q\left[\bar{f}\right] = 0$.

Pseudo-aléatoire ou quasi-aléatoire ?

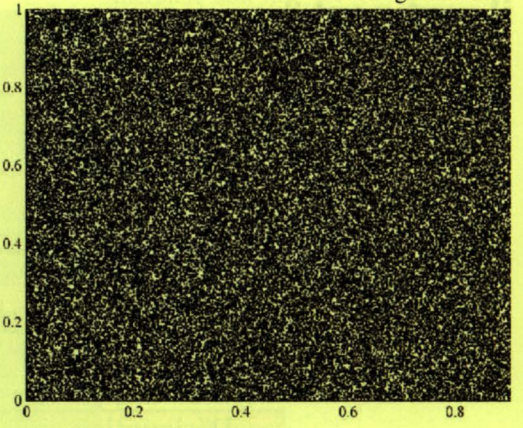
Pour toute simulation Monte Carlo, il est crucial de savoir générer une suite de nombres aléatoires de loi donnée et, inversement, de pouvoir juger si une suite de nombres donnée est une réalisation « acceptable » de la loi demandée.

***Stricto sensu*, il est impossible de générer des suites aléatoires, mais on s'en rapproche en construisant des suites vérifiant tous les théorèmes connus du calcul des probabilités. Ces suites, dites *pseudo-aléatoires*, sont généralistes et on peut les utiliser pour tout problème avec la certitude qu'elles fournissent, lentement mais sûrement, la solution recherchée. Une autre façon de procéder consiste à renoncer au caractère «aléatoire» des tirages et à tirer des points de façon «plus ordonnée». On parle alors de méthode de « quasi-Monte Carlo » qui utilise des suites dont la discrépance, écart à la non uniformité, est faible.**

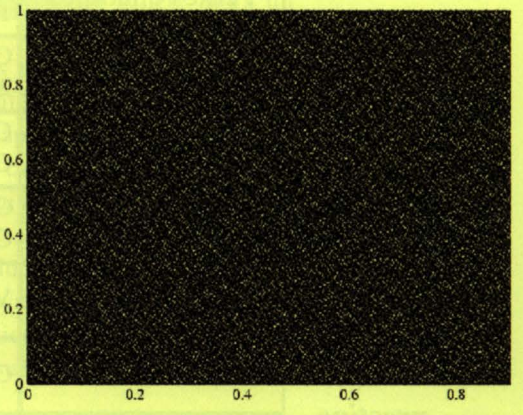
La figure suivante montre la simulation de 100 000 tirages de deux variables aléatoires indépendantes uniformément répartis dans le carré $[0, 1]^2$ par le générateur de la librairie GNU utilisée pour programmer en C (en haut) et par la suite déterministe de Van der Corput (en bas).

La méthode de Monte Carlo, par la simplicité de sa mise en œuvre, est un outil précieux pour la compréhension de ces phénomènes. Elle bénéficie de l'augmentation des capacités de calcul des nouveaux ordinateurs auxquels elle est particulièrement bien adaptée.

Tirage aléatoire



Suite de Van der Corput



par Hervé Lehning

Segments à vendre

Traditionnellement, le marché automobile est divisé en segments correspondant à la taille des voitures. Ce critère est-il pertinent ?

Pour mieux atteindre les clients potentiels, les vendeurs découpent les marchés en segments. Par exemple, voici les segments automobiles les plus souvent utilisés en France :

Désignation	Exemple
Petites citadines	C1, Smart, Twingo, Panda, 107, Lupo
Citadines	C2, C3, Clio, Punto, Fiesta, 207, Polo, Corsa, Classe A
Compactes	C4, Focus, 307, Mégane, Golf
Familiales	C5, Laguna, Passat, 407, Classe C
Routières	Velsatis, 607, Classe E
Limousines	Classe S
Tous terrains et véhicules de loisirs	4007, Defender, Cayenne, Touareg, Samuraï

Un vendeur de voitures ne vend pas seulement des automobiles, il vend un vecteur de prestige.

Dans l'idéal, chaque segment correspond à une clientèle, et donc à une politique commerciale spécifique.

Arguments de vente

Les arguments ne sont pas les mêmes pour vendre une petite citadine, un tout terrain, ou une limousine. Ils correspondent aux motivations des clients, qui doivent donc être analysées. Cette mesure ne peut être faite qu'au moyen de statistiques, donc de sondages. Les motivations réelles des clients ne sont pas faciles à cerner. Ils donneront toujours une réponse plus « rationnelle » que la réalité.

Qui avouera désirer acheter une Velsatis ou une classe E pour faire étalage d'une certaine réussite ? Ou faire croire à cette réussite ! Il préférera vanter les équipements de la voiture, sa vitesse de pointe, ou d'autres critères. Chaque segment doit ainsi s'adapter aux motivations véritables de sa clientèle, celles-ci ne sont pas toutes de nature quantitative, elles sont aussi de nature qualitative. Ceci ne signifie pas que les voitures les plus chères soient de meilleures qualités. C'est bien souvent le contraire qui est vrai !

Entre la Twingo et la Rolls-Royce, les acheteurs ne se fondent pas sur des arguments de Qualité/Prix, mais sur des impressions d'images associées au prix.

Segmentation

Alors, comment regrouper les voitures dans ces segments ? L'idée traditionnelle est saugrenue : selon leur taille ! Est-ce vraiment le bon critère ? Le plaisir et la vanité ne se mesurent pas en centimètres ! L'image est plus importante. La segmentation a légèrement évolué de ce fait, mais il reste de type « rationalisant ». On tient davantage compte de la forme de la carrosserie. Pourtant, le client voulant prétendre à un certain statut social à travers sa voiture ne cherche pas forcément une automobile longue. L'essentiel est qu'elle soit associée à une idée de luxe. La politique commerciale doit lui montrer le statut social qu'il peut atteindre grâce à elle.

On retrouve le même phénomène dans la vente de café. En le présentant comme un produit de luxe, le prix n'a plus guère d'importance. Le client achète alors un standing, une image, plus qu'un certain nombre de doses de café. On ne compare pas le prix d'une capsule d'un café de prestige au prix du café en poudre, même de qualité équivalente. Cela n'a aucun sens.

Pour conclure, la segmentation des produits devrait se faire davantage sur les motivations potentielles des acheteurs que sur des mesures soi-disant objectives.

H. L.

par Benoît Rittaud

Petits pois et khi-deux

Gregor Mendel découvrit les lois de la génétique à l'aide de petits pois. Avec les données dont il disposait, pouvait-il légitimement déduire ces lois, ou l'ignorance des statistiques a-t-elle servi, par chance, l'auteur de la découverte ?

Prenez deux sortes de petits pois : des jaunes et des verts. Arrangez-vous pour les faire se reproduire sans mélanger les deux races, jusqu'à être certain que les plants dont vous disposez soient "purs" : jamais de vilain petit pois vert chez les petits pois jaunes, jamais de jaune parmi les verts. Une fois la certitude acquise que les deux races sont bien vierges de tout mélange, faites une nouvelle génération à partir de mariages mixtes entre les plants de petits pois, et vous ferez la même constatation que Mendel au dix-neuvième siècle : cette nouvelle génération est constituée de rejetons tous jaunes.

La race des jaunes l'a-t-elle définitivement emporté ? Regardons ce qui se passe à la seconde génération, obtenue par mariages des plants de nos petits pois jaunes : des verts réapparaissent. Sur les 556 petits pois recensés, Mendel en compte 416 jaunes et 140 verts, soit 74,82% contre 25,18%. D'où la conclusion qu'en tire le père de la génétique : la couleur qui apparaît comme dominante à la première génération (le jaune, qui fait l'unanimité) reste dominante dans une proportion de trois sur quatre à la seconde génération. L'explication de cette proportion 3/4 constitue les lois de la génétique aujourd'hui classiques, brièvement rappelées en encadré.

Le test du khi-deux

Considérer que 74,82 % reflète une proportion de 3/4 est une approximation nécessaire pour relier les résultats de l'expérience à la théorie que constituent les lois de Mendel. Dans quelle mesure cette approximation est-elle légitime ? Doit-on penser que l'écart entre 74,82 % et 75 % est suffisamment faible pour pouvoir le négliger, ou y a-t-il un risque qu'il recèle en réalité une complication cachée dans les phénomènes de la génétique ?

Les statistiques disposent de plusieurs manières de tester les données pour analyser les écarts entre les résultats empiriques et la théorie que l'on veut valider (ou infirmer). L'une d'elles est couramment appelée le test du khi-deux.

Si f_{jaune} et f_{vert} représentent la proportion observée de petits pois jaunes et verts (f pour fréquence) et p_{jaune} et p_{vert} la proportion théorique donnée par les lois de Mendel, alors le test consiste d'abord à calculer la valeur suivante (le signe χ est la lettre grecque khi, d'où le nom du test) :

$$\chi^2 = 2\left[\frac{\left(f_{jaune} - p_{jaune}\right)^2}{p_{jaune}} + \frac{\left(f_{vert} - p_{vert}\right)^2}{p_{vert}}\right].$$

En toute généralité, lorsqu'il s'agit de valider une hypothèse selon laquelle n caractères possibles doivent chacun théoriquement apparaître dans une proportion p_i (pour i variant de 1 à n) et qu'une expérience produit chacun de ces caractères avec la fréquence f_i la valeur à calculer est la suivante :

$$\chi^2 = n\sum_{i=1}^{n}\frac{\left(f_i - p_i\right)^2}{p_i}.$$

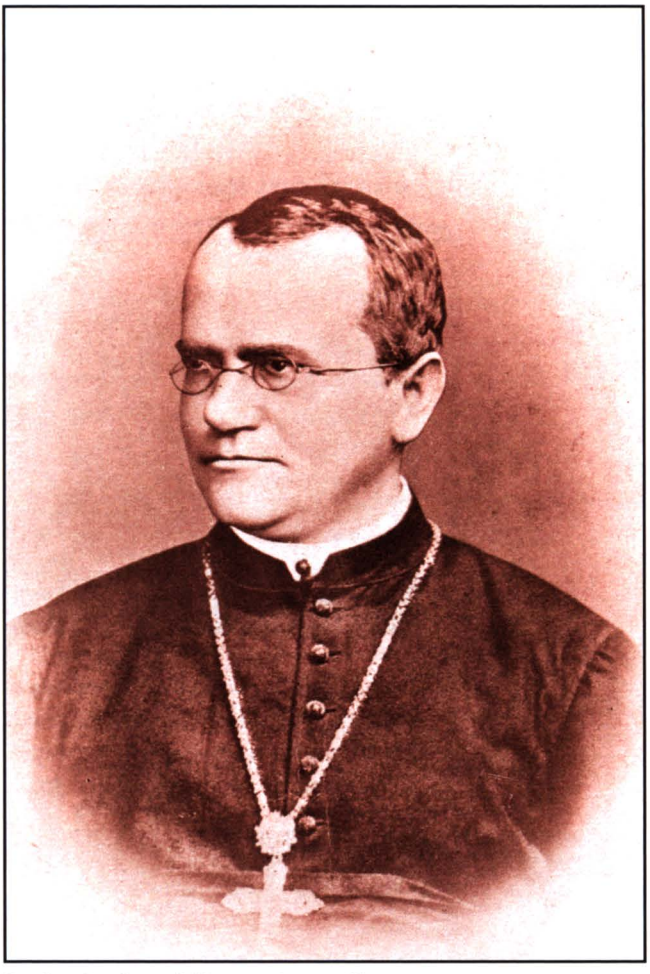
La valeur χ^2 d'autant plus petite que les fréquences empiriques f_i sont proches des proportions théoriques. Le fait de diviser par les proportions p_i a pour utile effet de "*redresser*" l'importance accordée aux écarts concernant des proportions petites : la différence entre deux quantités petites est toujours faible (surtout quand, comme dans le calcul de χ^2 on élève cette différence au carré), mais peut être significative d'un point de vue théorique (comme entre 1/1000 et 1/2000, par exemple).

Plus la valeur de χ^2 est petite, plus le test est favorable. La multiplication par n dans la formule a ainsi pour effet de rendre compte du fait qu'un écart à la proportion théorique est "plus grave" lorsqu'il porte sur des données nombreuses que lorsqu'il a lieu à une échelle réduite.

Et Mendel dans tout ça ? La valeur de χ^2 pour les données expérimentales est de l'ordre du dix-millième : c'est suffisamment petit pour que les conclusions de Mendel soient acceptables d'un point de vue statistique. Précisons qu'en fait, les petits pois du père de la génétique n'avaient pas que la couleur comme différence : ils étaient jaunes (J) ou verts (V), et ronds (R) ou anguleux (A). Il y en eu 315 RJ, 101 AJ, 108 RV et 32 AV.

Le χ^2 qu'il convient de calculer est donc celui de la seconde formule, plus générale (et qui confirme, lui aussi, les conclusions de Mendel).

N'allons d'ailleurs pas dire que tout cela n'a finalement servi à rien au motif que le test de khi-deux apporte une conclusion dont on se doutait : si l'expérience portait



Grégor Mendel, moine et savant, découvrit les lois de l'hérédité en croisant des petits pois (Pflanzen-Hybriden, 1865).

sur quelques centaines de milliers de petits pois et que la proportion de jaune soit encore de 74,82 %, l'écart à la proportion théorique aurait été beaucoup plus suspect (ce que le test de khi-deux aurait indiqué), alors qu'une vision "de loin" de ces nouvelles données ne le laisserait pas penser.

Il y a beaucoup d'exemples d'expériences historiques qui, si elles avaient été mieux faites ou plus rigoureusement interprétées, auraient finalement empêché la découverte (c'est la cas de la loi de Coulomb, par exemple). Celle de Mendel, au moins du point de vue statistique, n'est pas dans ce cas... si tant est qu'on considère que quelques centaines de petits pois constituent une échelle suffisante pour tirer des conclusions.

B. R.

Les lois de la génétique

Tous les êtres vivants possèdent des caractères héréditaires, exprimés à partir des gènes (portés par l'ADN, comme nous le savons aujourd'hui). Chaque gène exlste en deux exemplaires (on parle de gènes allèles), apportés par chacun des parents. Chez les petits pois de Mendel, par exemple, le gène qui porte le caractère "couleur" est soit "jaune", soit " vert ". Si les deux gènes qu'apportent les parents portent la même couleur, alors c'est cette couleur qui sera celle du petit pois : c'est le cas des races "pures" du début. Lorsque les deux couleurs sont représentées dans les gènes, l'une des deux (toujours la même) prend le dessus : elle est dominante. Cest la couleur jaune qui est dominante chez les petits pois : ceux de la première génération obtenus par croisement, qui possèdent un gène porteur de chaque couleur, sont donc logiquement tous jaunes. Pour chacun de leurs descendants, en revanche, plusieurs choses peuvent se produire : il peut prendre de ses deux parents le gène de la couleur verte (probabilité : 1/4), ou de la couleur jaune (probabilité : 1/4), ou bien des gènes des deux sortes (probabilité 1/2). Puisque le jaune est dominant, seuls ceux d'entre eux qui ont les deux gènes porteurs de la couleur verte sont verts : ils sont statistiquement un quart dans ce cas.

Interpréter les statistiques

Tout au bord du précipice de l'erreur, les mathématiciens ordonnent leur environnement. Les écueils n'en sont plus quand ils sont reconnus et les statistiques aident les sciences naturelles à classer les êtres et les choses en fonction de leurs caractéristiques.

par Philippe Boulanger

Le paradoxe de Simpson

Le rassemblement de résultats donne des résultats surprenants. Ce qui est bon pour chaque partie d'un tout peut être mauvais pour le tout...

«Je voulais tester l'efficacité d'un médicament, le Perlimpimpinate» gémissait Kasiro, consultant de la société *Placentis*, sur une maladie nouvelle, la zymosis de l'écrivain. Nous avions choisi un premier échantillon de mille cent personnes atteintes de zymosis, auxquelles nous avons donné, soit du Perlimpimpinate, soit un placebo. Puis nous avons dénombré le nombre de personnes guéries.

	Perlimpimpinate	Placebo
Personnes non guéries	100	200
Personnes guéries	300	500

Sur 400 personnes à qui l'on a donné du Perlimpimpinate, 300 ont été guéries, soit 75% des sujets testés, et sur les 700 personnes à qui l'on a prescrit du placebo, seules 500 ont été guéries, soit 71%. Le Perlimpimpinate était plus efficace que le placebo. Et *Placentis* aurait pu le commercialiser. Mais, les dirigeants de Placentis ont été pris d'un doute : la zymosis de l'écrivain ne serait-elle pas plus grave à la campagne? Une nouvelle série de tests a été pratiquée sur mille cent ruraux, avec les résultats suivants :

	Perlimpimpinate	Placebo
Personnes non guéries	500	300
Personnes guéries	200	100

Sur 700 écrivains traités avec le perlimpimpinate, seules 200 sont guéries soit 29%. Avec le placebo, 100 personnes sur 400 sont hors du danger zymosinien, soit 25% : comme ce pourcentage était inférieur à 29%, la supériorité du perlimpimpinate sur le placebo était, là encore, avérée.

Tout va très bien jusqu'à ce que l'on regroupe les deux tableaux, addition-

nant les guérisons et les échecs avec le Perlimpimpinate et le placebo. Là, voyez le désastre :

	Perlimpimpinate	Placebo
Personnes non guéries	600	500
Personnes guéries	500	600

Vous pouvez vérifier les chiffres : sur les 1100 zymosiniens qui avaient pris le Perlimpimpinate seuls 500 avaient guéri, soit 45%, alors qu'avec le placebo, 600 zymosiniens étaient hors de péril, soit 55%.

«C'est invraisemblable, éructait Kasiro, alors que les proportions de guérisons avec le Perlimpimpinate sont supérieurs dans les deux échantillons, lorsqu'on les réunit, ils deviennent inférieurs à celui du placebo.»

Le paradoxe examiné par le mathématicien Simpson il y a une cinquantaine d'années, apparaît quand deux variables ne sont pas très corrélées, ici la guérison et la prise de Perlimpimpinate.

Mal corrélées signifie que la relation de cause à effet n'est pas nette. Les guérisons de la zymosis étaient probablement dues à autre chose qu'à la prise de Perlimpimpinate ou de placebo.

L'utilisation d'un test « médicament contre placebo » remonte au Moyen Age, où l'on essayait diverses décoctions contre le scorbut qui faisait des ravages. Plomb, arséniates, queue de rat séchée, tisane d'orties, tout y passait, mais malheureusement les apothicaires du temps prirent comme placebo... le jus de citron. Bien évidemment, aucune de leurs concoctions n'était supérieure au placebo, car le placebo était le seul produit actif.

Combien de médicaments ont été indûment autorisés grâce au paradoxe de Simpson, avec des effets moins heureux ?

P. B.

par Philippe Boulanger

L'estimateur des prouesses futures

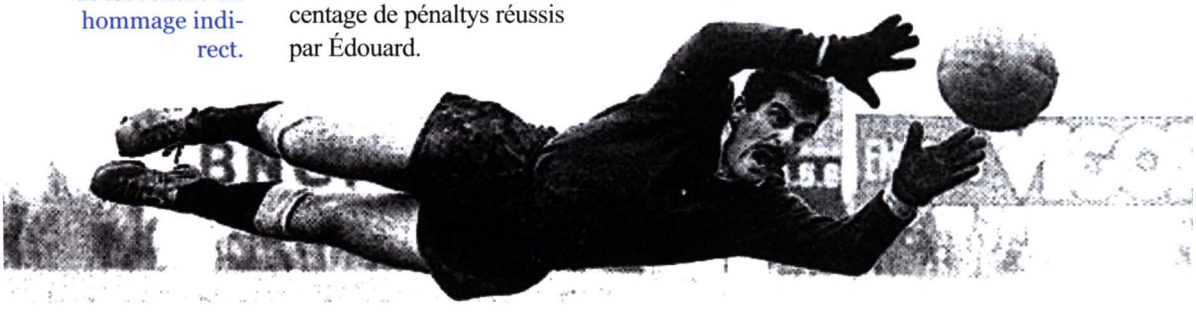
La moyenne des résultats antérieurs n'est pas la meilleure estimation de la performance future, c'est le sulfureux estimateur de Stein.

Edouard Guinotte va tirer un pénalty, René Vignal est dans les buts. Vous voulez évaluer les chances de succès de notre ami Édouard. Il n'a tiré qu'un pénalty dans sa carrière contre Vignal et l'a réussi. Allons-nous penser que sa probabilité de rentrer le prochain pénalty est la moyenne des succès passés, c'est-à-dire 1 ? Ce serait trop optimiste ! Ne faudrait-il pas mieux prendre en compte les pénaltys qu'Édouard a rentrés contre d'autres gardien de but ? Ou bien la proportion réussie de penaltys par les autres joueurs de son équipe, *L'étoile flamboyante de Boulogne* ? Ou bien le pourcentage de pénaltys encaissés par Vignal au cours de sa carrière ? Il est certain que ces diverses mesures diminueront le pourcentage de pénaltys réussis par Édouard.

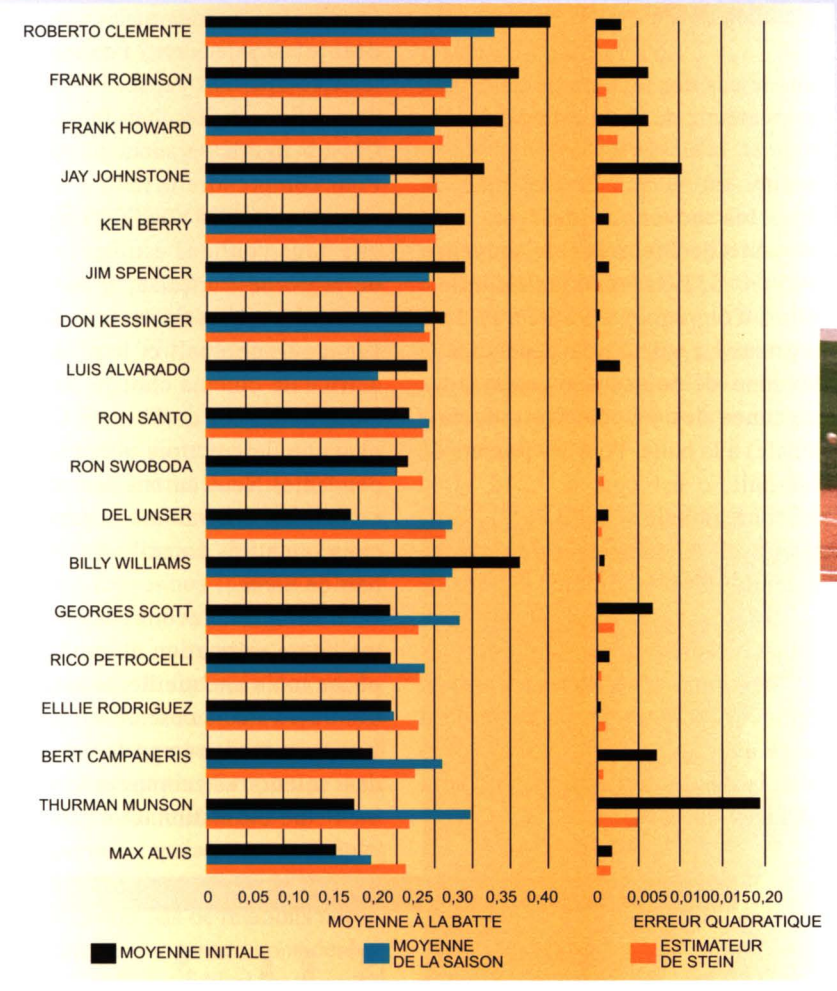
L'estimateur de Stein

Le résultat mathématique énoncé par C. Stein est en frappante contradiction avec une croyance répandue, celle que la probabilité d'un événement futur est égale à la moyenne des évènements passés. Les exemples précédents montrent l'esprit de la méthode du statisticien C. Stein.

L'exemple utilisé par Charles Stein (né en 1920) pour illustrer sa méthode a trait au base-ball. Les règles du jeu sont inutilement compliquées pour un amateur de football, mais il nous suffira de savoir qu'un paramètre important est le pourcentage de réussite à la batte d'un joueur. Avec sa batte, il tente de frapper la balle envoyée par le lanceur.



1• René Vignal, le gardien volant a été après la guerre le gardien mythique de l'équipe de France. Ce texte est une occasion de lui rendre un hommage indirect.



2• Comparaison de l'estimateur de Stein et de la moyenne pour l'évaluation d'évènements futurs : la moyenne initiale est mesurée pour 45 essais et indiquée en noir. L'estimateur de Stein de chaque joueur est en rouge et la moyenne de la saison (9 fois plus de données) est indiquée en bleu. Les erreurs quadratiques pour chaque joueur sont les carrés d'une part de la différence entre la valeur de l'estimateur de Stein et de la moyenne de la saison (en rouge) et d'autre part de la différence entre la valeur tirée de la moyenne initiale et la moyenne de la saison (en noir). La somme des erreurs quadratiques commises avec l'estimateur de Stein est inférieure à celle des valeurs prises à partir de la valeur moyenne initiale. L'estimateur de Stein est préférable.

S'il réussit à frapper 8 fois la balle sur 20 lancers son pourcentage de réussite (sa « capacité à la batte ») est 8/20, soit 0,40.

Or si l'on nous demandait quelle est la proportion de coups qu'il va réussir lors des 100 prochaines tentatives, nous répondrions 40. Eh bien, ce serait moins bon que l'estimateur de Stein qui évalue le nombre de coups réussis à 29,4 soit proche de 30. Sur la figure 2, nous avons représenté les capacités de joueurs après une fraction de la saison, évalué les moyennes et montré que l'estimateur de Stein donne de meilleurs résultats d'évaluation de la capacité à la batte de l'ensemble des joueurs pendant toute la saison.

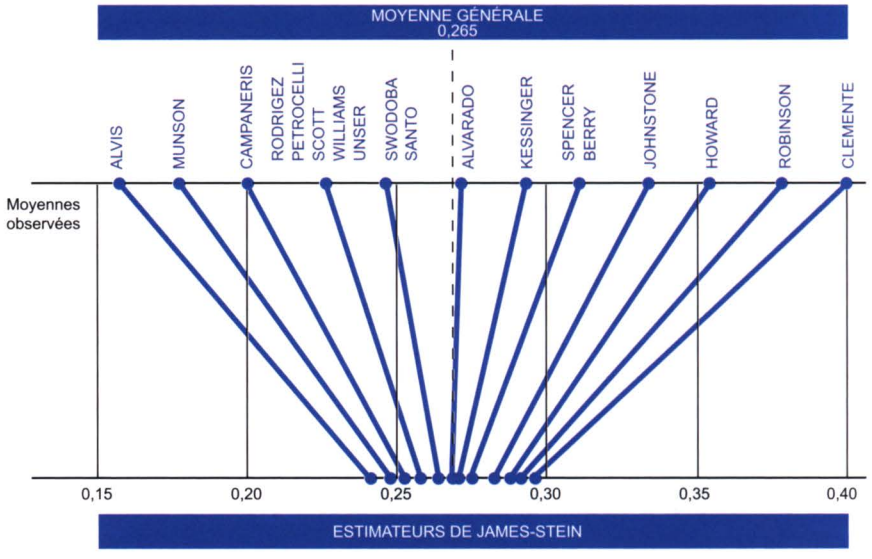
L'équation de Stein

Dans le cas des joueurs de base-ball, l'estimateur z de Stein est égal à

$$m + c(y - m),$$

formule où m est la moyenne de toutes les moyennes mesurées, c la constante de Stein et y la moyenne observée. Si l'on prend la valeur de c égale à 0 on retrouve la valeur m de la moyenne. La valeur de c dépend de la moyenne de toutes les valeurs des moyennes des joueurs (la moyenne globale) à la batte. Pour les joueurs de base-ball, c est égal à 0,212 et la moyenne globale m est 0,265 ; l'estimateur de Clemente n'est pas sa moyenne 0,400, mais 0,265 + 0,212(0,4 – 0,265) soit 0,294. Les valeurs des estimateurs de Stein (parfois dénommés de James-Stein en l'honneur du mathématicien Willard James avec qui Charles Stein développa la théorie des estimateurs) sont indiquées sur la figure 3.

Quel est le rapport entre les capacités de deux joueurs distincts ? Pourquoi faut-il tenir compte des capacités d'un autre pour estimer ses résultats futurs ?
C'est là tout le paradoxe de Stein.
Pour comprendre la nature peu commune du paradoxe de Stein supposons que nous voulions estimer le produit de la pêche française, le nombre de spectateurs de la finale de la coupe de France de football et le poids d'une actrice de cinéma choisie au hasard. Nous avons les courbes de Gauss de chacune de ces trois quantités indépendantes.Nous aurons une meilleure estimation des valeurs futures de ces trois quantités en utilisant l'estimateur de Stein…
Bien sûr nous n'avons pas obtenu une meilleure estimation du produit de la pêche mais une meilleure estimation des trois valeurs ensemble.
Les erreurs trop importantes sur une des valeurs est compensée par une meilleure estimation d'un autre.



3• Les estimateurs de Stein pour les 18 joueurs de base-ball ont été calculés en rapprochant les moyennes individuelles à la batte d'une moyenne des moyennes individuelles (la « moyenne globale », ici 0,265). Selon Stein, les capacités réelles à la batte sont plus regroupées que les moyennes préliminaires pouvaient le suggérer.

La moyenne arithmétique est un estimateur valable quand il n'y a qu'une variable à considérer, mais dès que le nombre de variables dépasse 2, l'estimateur de Stein est préférable.

Le facteur de rapprochement

Il faut maintenant déterminer la valeur du facteur c de rapprochement. Cette valeur est déterminée par l'équation

$$c = 1 - (k-3)s^2 / (\mathrm{S}\,(y-m)^2)$$

où k désigne le nombre de moyennes inconnues, s^2 est le carré de l'écart-type pour la distribution considérée et $\mathrm{S}\,(y-m)^2$ la somme des carrés des écarts des moyennes individuelles y par rapport à la moyenne globale m. Le nombre de moyennes estimées influe aussi sur le facteur c de rapprochement par le facteur $k-3$. S'il y a beaucoup de moyennes l'équation rend le rapprochement plus fort puisqu'il est alors peu vraisemblable que les variations observées représentent de simples fluctuations dues au hasard. Il n'y a paradoxalement pas de relations nécessaires entre les moyennes considérées, qui peuvent être aussi bien des réussites de pénalty que le nombre de spectateurs à Roland Garros. Le procédé de Stein est d'autant meilleur que les quantités inconnues sont voisines de la moyenne globale.

Les estimateurs de Stein sont utilisés dans les nombreux domaines économiques où les investissements dépendent des prévisions réalistes, par exemple l'emplacement des hôpitaux, des usines ou des magasins de stockage. La théorie de l'estimation a peu à peu perdu son caractère paradoxal et de multiples ponts ont été établis avec les mathématiques des probabilités. Les travaux sur ce type d'estimateur se poursuivent aujourd'hui, notamment sur leur adéquation fine aux problèmes posés.

On est travaillé par les choses, autant qu'on travaille sur elles.

Isabelle Adjani

P. B.

Bibliographie

Le paradoxe de Stein, **Bradley Efron et Carl Morris**, *Pour la Science*, novembre 1977.

http://en.wikipedia.org/wiki/James-Stein_estimator

http://projecteuclid.org/DPubS?service=UI&version=1.0&verb=Display&handle=euclid.ss/1063994981 pour utiliser le pdf d'un chantre de l'estimateur de Stein, Bradley Efron. Cette autobiographie (en anglais) décrit bien l'évolution des statistiques modernes.

par Jacques Lubczanski

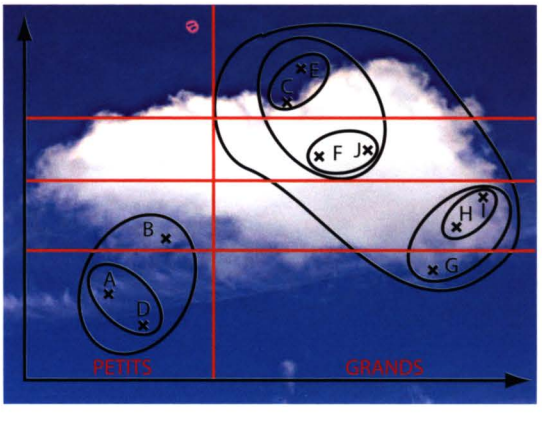
Le classement des données

La taille et le poids des individus d'une population étant donnés, comment y classer les petits gros ? Le problème, loin d'être futile, se pose en zoologie pour classer les individus selon leurs caractéristiques.

Les résultats des mesures se présentent souvent sous la forme d'un nuage de points. Par exemple la taille et le poids de n individus sont représentés dans un plan. Pour regrouper les individus en classes, dont l'une d'elles sera notre ensemble des « petits gros », nous allons mettre en pratique le vieux dicton : « Qui se ressemble s'assemble ».

Assemblons les points qui se ressemblent. Le problème prend donc la tournure suivante : quel critère mathématique choisir pour évaluer la ressemblance des individus, c'est-à-dire des points du nuage ? Comment utiliser ce critère pour regrouper les points en classes ?

Et comment interpréter les résultats obtenus à la lumière du problème posé ? Nous sommes là au cœur de la branche des mathématiques dénommée l'analyse des données. Mais rassurez-vous, il n'est pas question ici de tout dire sur le sujet ! Faisons plutôt un petit bout de chemin, une promenade touristique dans ce domaine souvent, et à tort, méconnu.



Nuages de points regroupés par classes

	A	B	C	D	E	F	G	H	I	J
A	0	26	113	13	181	106	169	229	298	169
B		0	53	25	89	32	65	101	148	77
C			0	146	10	13	100	100	125	34
D				0	208	113	130	194	261	164
E					0	17	90	74	85	20
F						0	41	45	68	9
G							0	8	25	26
H								0	5	18
I									0	29

Quel critère mathématique choisir pour évaluer la ressemblance entre deux points ?
Le critère le plus naturel consiste à mesurer la distance entre deux points, puis entre les classes d'une répartition. Il faudra enfin mesurer la cohésion interne et la rendre optimale. Cela est encore un peu vague, mais nous préciserons cette notion en examinant les défauts de la méthode employée.
Il s'agira avant tout de méthodes pratiques : dans tous les cas, la répartition de la population en classes sera le résultat d'un algorithme. En effet, si toutes les méthodes sont très simples en principe, leur mise en œuvre s'accompagne d'une quantité considérable de calculs, confiée bien sûr à un ordinateur, d'où la nécessité d'un algorithme. Il est même possible que les ordinateurs ne soient pas assez puissants, et que l'on doive se contenter d'une solution approchée.
Le but de notre méthode est d'établir une distribution hiérarchique des points du nuage : la population sera divisée en classes, les classes en sous-classes, les sous-classes en sous-sous-classes *etc.*, jusqu'aux classes réduites aux individus.
Un exemple célèbre du genre est la classification des zoologistes (vertébrés, invertébrés, ...).

L'outil

L'outil est d'abord la distance entre deux points et entre deux classes. Si C_1 et C_2 sont deux classes, la distance $d(C_1, C_2)$ est la plus petite distance possible entre un point de C_1 et un point de C_2. Une classe peut être composée d'un seul élément.
On part de classes réduites à un point, et à chaque étape, on fusionne deux classes ou plus, celles qui sont le plus

On oublie, dans le classement des grands événements ayant marqué le millénaire, d'inclure la vogue des classements.
Jean Dion (journaliste québécois)

rapprochées. D'où une nouvelle répartition.
Dans notre exemple, nous aurons dix points A, B, C, D, E, F, G, H, I, J à regrouper, mais les exemples pratiques peuvent inclure des millions de points.
On dresse d'abord le tableau des distances entre les points. Si l'effectif de la population est n, il faut $n(n-1)/2$ distances. Et c'est tout.
Les distances sont indiquées dans le tableau.

	A	B	C	D	E	F	G	HI	J
A	0	26	113	13	181	106	169	229	169
B		0	53	25	89	32	65	101	77
C			0	146	10	13	100	100	34
D				0	208	113	130	194	164
E					0	17	90	74	20
F						0	41	45	9
G							0	8	26
HI								0	18

Première étape
La distance entre le point H et le point I, égale à 5, est la plus petite distance entre deux points. On fusionne les points H et I pour former une classe, la classe HI. On continue en fusionnant les distances de classes : on remplace dans le tableau les colonnes H et I par un colonne HI, en prenant le plus petit des deux nombres des cases contiguës.

Le reste du tableau ne change pas (ici la colonne HI est celle de H, car I est plus éloigné que H de tous les autres points).

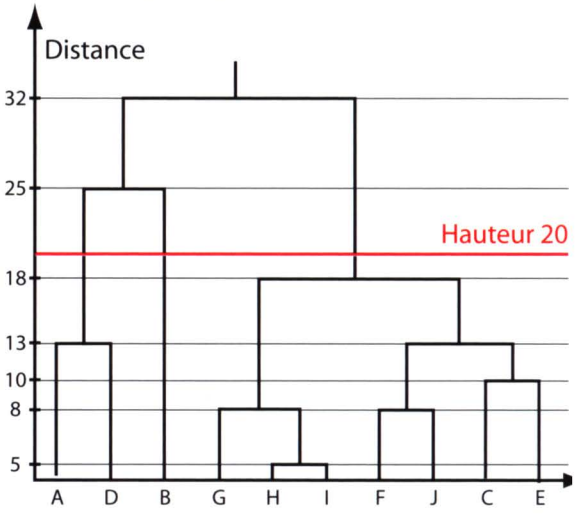
	A	B	C	D	E	F	GHI	J
A	0	26	113	13	181	106	169	169
B		0	53	25	89	32	65	77
C			0	146	10	13	100	34
D				0	208	113	130	164
E					0	17	74	20
F						0	41	9
GHI							0	18

Deuxième étape
On cherche à nouveau le plus petit nombre du tableau, c'est 8, la distance entre G et HI. On fusionne alors G et HI en GHI. Le tableau rétrécit encore d'une ligne et d'une colonne.

Troisième étape : le plus petit nombre du tableau est 9 entre F et J. On fusionne F et J.
Quatrième étape : le plus petit nombre du tableau est 10 : on fusionne C et E.
Cinquième étape : le plus petit nombre du tableau est 13, qui figure deux fois : on fusionne A et D d'une part, C et F d'autre part.
En pratique, il est inutile de récrire les tableaux à chaque étape, il suffit de suivre les nombres du tableau, du plus petit au plus grand : après 13 vient 17, distance entre F et E, mais F et E sont déjà dans la même classe et l'on passe au suivant.
C'est 18 = d (J,H) : on fusionne GHI et FJCE.
Puis 25 = d(B,D) et aussi d(G,I) : G et I sont déjà ensemble, on fusionne B et AD.
Il ne reste alors plus que deux classes, ABD et GHIFJCE : on les fusionne.

La classification en dendogramme

On dresse alors un arbre, dénommé dendogramme, un dessin qui représente la classification de la population en classes et sous-classes…
En coupant l'arbre à une hauteur quelconque, il tombe des branches ! On obtient la répartition en classes correspondant à cette hauteur-seuil. Par exemple si l'on scie à la hauteur 20, il tombe trois branches, AD, B et GHIFJCE.



Interprétation des résultats

La classification obtenue est indiquée sur la figure de droite. Les lignes de niveau de la taille (droites verticales) et du poids (droites horizontales) permettent alors d'interpréter chaque fourche

du dendogramme. On a les « petits gros » ou plutôt, le petit gros, B !

L'inconvénient de la méthode est d'opposer B, petit gros et G, grand maigre, alors que G est plus proche de B que de C.

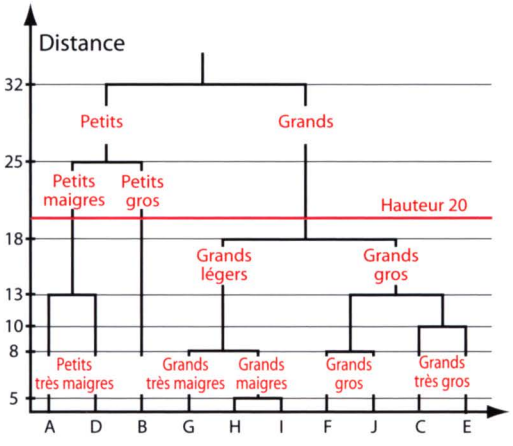
On attribue ce défaut au choix de la distance entre classes qui favorise la fusion de deux classes dès qu'elles possèdent des points proches : le risque est alors de retrouver, dans une classe, des points très éloignés. On peut remédier à cela avec d'autres définitions de la distance, mais toutes les distances ont leurs inconvénients.

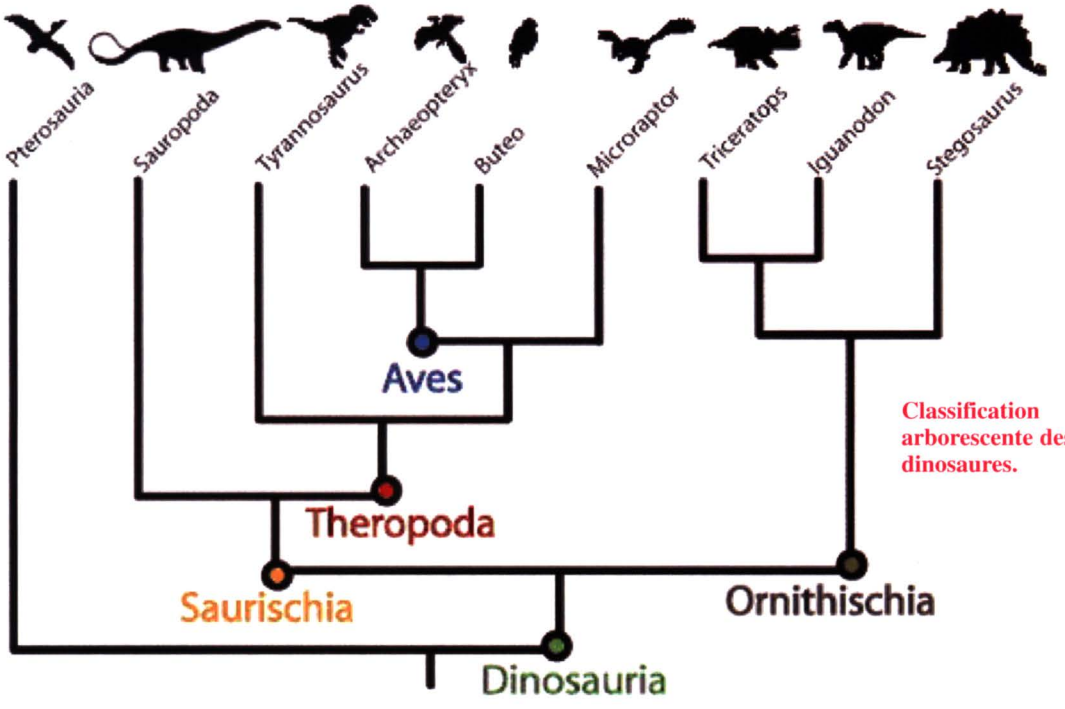
Les paléontologues classent ainsi les « dinosaures ». La querelle actuelle est de savoir si la distance doit être mesurée par les ressemblances comme on l'a fait longtemps, ou par les différences, comme conseillé par les cladistes.

J.L.

Références

Jacques Lubczanski, *Comment réussir le triangle quelconque ... et douze autres friandises !*, Cedic 1986.





Classification arborescente des dinosaures.

par Valérie Henry

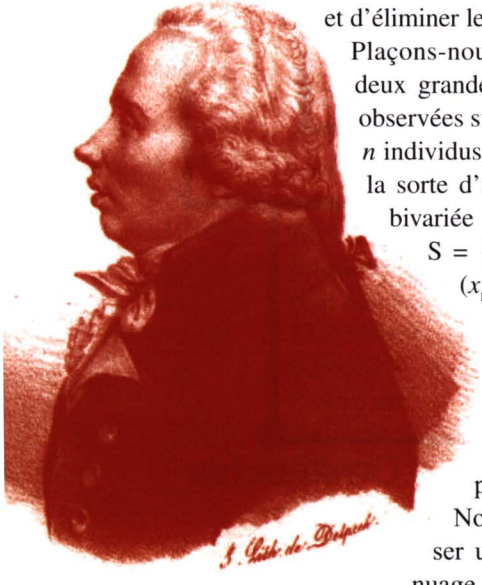
Robustesse

et régressions linéaires robustes

La méthode des moindres carrés, qui a deux siècles, peut être rendue plus robuste c'est-à-dire moins sensible aux valeurs extrêmes : la droite obtenue représente donc mieux les données.

Il est fréquent de vouloir considérer simultanément plusieurs variables sur un même ensemble d'individus ; un problème concret consiste alors à rechercher des relations fonctionnelles traduisant d'une certaine manière les observations réelles afin d'interpoler les résultats, éventuellement les extrapoler et d'éliminer les erreurs de mesure.

Adrien-Marie Legendre (1752-1833),



Plaçons-nous dans le cas de deux grandeurs, notées X et Y, observées sur une population de n individus : nous disposons de la sorte d'une série statistique bivariée (à deux variables), $S = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$, qui peut être représentée dans un plan muni d'un repère cartésien par un nuage de points $P_i(x_i, y_i)$ pour $i = 1, \ldots, n$.

Nous souhaitons réaliser un ajustement de ce nuage, c'est-à-dire déterminer une fonction f dont le graphe « approche au mieux » les points observés. L'expression « approche au mieux » pouvant prendre de nombreuses significations mathématiques, il existe plusieurs types d'ajustement suivant les données observées. Deux aspects sont à considérer dans une telle démarche : 1) la forme du nuage qui sera un premier facteur déterminant dans le choix du type d'ajustement envisagé ; 2) le critère utilisé pour traduire formellement le fait que la fonction d'ajustement doit représenter « au mieux » l'ensemble des données.

Méthode des moindres carrés

Ici, nous n'envisagerons que le cas d'un ajustement **linéaire**, c'est-à-dire à l'aide d'une fonction f affine définie par $f(x) = a\,x + b$, pour des paramètres a et b à calculer de manière que la droite d'équation $y = a\,x + b$ (où a désigne donc le coefficient directeur de la

droite et b son ordonnée à l'origine) représente « au mieux » les points du nuage des points observés.

Une méthode classique et courante est celle des **moindres carrés** ; elle est déjà ancienne puisqu'elle fut élaborée tout au début du XIX^{e} siècle et indépendamment par Legendre et Gauss à propos de travaux portant sur l'étude de trajectoires de corps célestes. Il s'agit de rechercher les paramètres a et b qui rendent aussi petite que possible la somme des carrés des résidus, un résidu étant l'écart entre une valeur observée y_i et la valeur correspondante $\hat{y}_i = a\,x_i + b$ estimée par la droite ; techniquement, il s'agit de minimiser la fonction S, en les deux variables réelles a et b, définie par

$$S(a,b) = \sum_{i=1}^{n} (y_i - ax_i - b)^2 .$$

Cette méthode revient à supposer que la fonction f retenue est affine avec une erreur aléatoire e : $y = a\,x + b + e$; la variable aléatoire e, prenant les valeurs $e_i = y_i - \hat{y}_i$, est supposée de moyenne estimée nulle et de variance minimale.

On peut montrer (voir encart en fin d'article) que la droite obtenue par cette méthode, et appelée la **droite de régression par la méthode des moindres carrés**, passe par le barycentre $\overline{P}(\overline{x}, \overline{y})$ du nuage des points P_i, où $\overline{x} = \dfrac{1}{n}\sum_{i=1}^{n} x_i$ et $\overline{y} = \dfrac{1}{n}\sum_{i=1}^{n} y_i$. De plus, le coefficient directeur a de cette droite est donné par la formule suivante : $a = \dfrac{\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n} (x_i - \overline{x})^2}$.

Remarquons que a s'écrit encore sous la forme $a = \dfrac{s_{xy}}{s_x^2}$, où le numérateur s_{xy} désigne la covariance de la série S bivariée, tandis que le dénominateur s_x^2 est la variance marginale de la série simple des x_i.

Les avantages de cet ajustement sont importants :

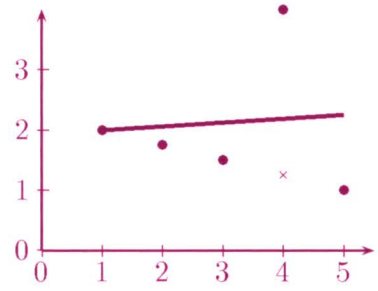
- les paramètres de la fonction d'ajustement sont univoquement déterminés grâce aux formules données ci-dessus ;
- la droite obtenue est un objet facile à concevoir et à représenter ;
- le critère utilisé s'illustre géométriquement de façon aisée : on minimise en fait un agrégat des « écarts verticaux » entre les points de la droite et ceux du nuage et l'on obtient de la sorte une droite qui « traverse au mieux » le nuage tout en passant par le barycentre de ce dernier.

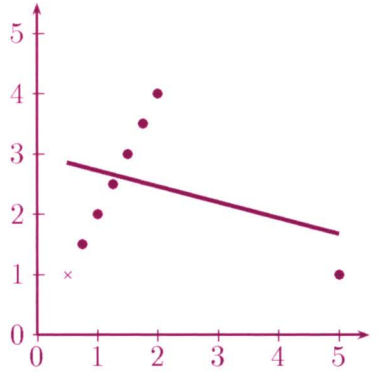
Karl Frederich Gauss (1777-1855)

Néanmoins, les résultats fournis par la méthode des moindres carrés, peuvent être fortement influencés par des données aberrantes ou atypiques : on dit alors que la droite des moindres carrés n'est guère « robuste ». Illustrons ce propos à l'aide d'exemples très simples.

La figure 1 montre ce que livre la méthode des moindres carrés (en abrégé et en anglais LS pour *Least Square*) lorsque, en partant de points parfaitement alignés, l'ensemble des données subit une contamination par un point aberrant en Y : cela correspond à un point qui se trouve « éloigné verticalement » de la droite initiale. La figure 2 illustre l'effet sur la droite de régression d'un point aberrant en X ou point se trouvant « éloigné horizontalement » du nuage de départ.

1. Effet d'une contamination par un point aberrant en Y





2. Effet d'une contamination par un point aberrant en X

La prise en compte des points aberrants est dès lors un problème important. La statistique « classique » est consciente du problème depuis longtemps et y remédie en tentant de repérer ces points particuliers, puis éventuellement de les supprimer de l'analyse. L'orientation prise par la statistique « robuste » est de développer des techniques d'analyse qui diminuent l'impact de points aberrants sur les résultats obtenus.

Nous donnons ici deux alternatives au critère de la minimisation de la somme des carrés des résidus qui prévaut dans la méthode LS.

Méthode des moindres écarts absolus

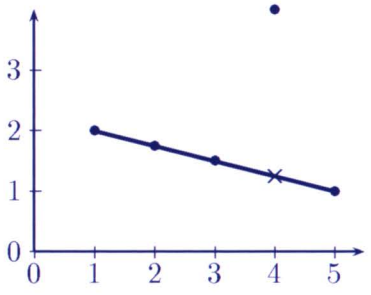
La première alternative consiste à retenir comme critère la minimisation de la somme des valeurs absolues des résidus.

Pour la série $S = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$, considérée ci-dessus, on est donc amené à minimiser la fonction, en les variables a et b, définie par

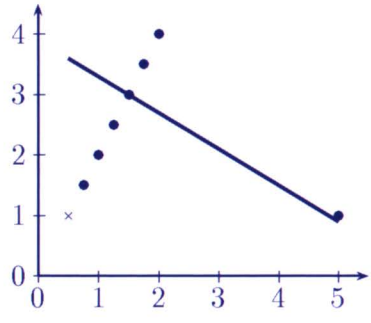
$$S_1(a,b) = \sum_{i=1}^{n} | y_i - ax_i - b |$$

Contrairement au cas de la méthode des moindres carrés, cette minimisation est un problème mathématique difficile, qui ne peut s'effectuer qu'à l'aide de l'outil informatique par le développement d'algorithmes adéquats. Cette méthode permet de supprimer l'effet d'une contamination par un point aberrant en Y.

Malheureusement, elle ne résout pas le problème de la contamination en X, ainsi qu'en attestent les exemples suivants.



3. Point aberrant en Y



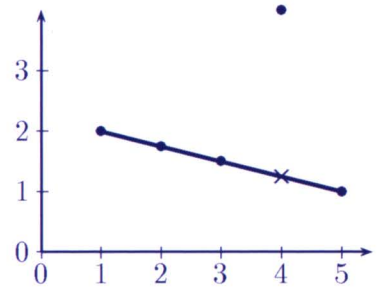
4. Point aberrant en X

Méthode des moindres carrés médians

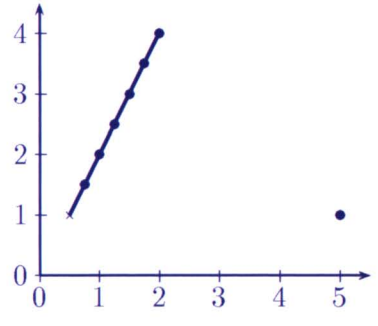
La solution proposée pour améliorer les résultats obtenus par la méthode des moindres écarts absolus est de revenir aux carrés des résidus mais de remplacer, dans le critère de minimisation, la somme par la médiane. Il s'agit donc de minimiser la médiane de la série statistique suivante :

$\{(y_1 - ax_1 - b)^2, (y_2 - ax_2 - b)^2, \ldots, (y_n - ax_n - b)^2\}$

De nouveau, cette méthode nécessite le support de l'outil informatique pour être appliquée. Elle permet d'atténuer fortement, voire d'éliminer, la contamination par des points aberrants en X. C'est ce qu'illustrent les deux graphiques ci-dessous où on observe la droite de régression obtenue par la méthode des moindres carrés médians pour les deux mêmes nuages de points que ci-dessus.



5. Point aberrant en Y



6. Point aberrant en X

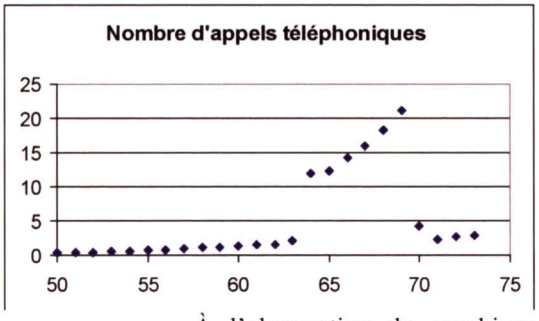
T'échappes à la police, pas aux statistiques. ***Jean-Jacques Goldman, Paroles de la chanson Des vies***

Un exemple en guise de conclusion

Considérons un exemple concret pour illustrer nos propos. Les données ci-dessous sont issues d'une référence classique en statistique robuste, à savoir le livre *Robust Regression and Outlier Detection*, par P.J. Rousseeuw et A.M.Leroy, publié chez John Wiley & Sons en 1987.
Lors d'une étude consacrée au recensement du nombre d'appels téléphoniques internationaux donnés en Belgique pendant la période 1950 – 1973, les données suivantes ont été enregistrées où X désigne l'année d'observation et Y le nombre d'appels enregistrés au cours de l'année correspondante :

X	Y	X	Y	X	Y
50	0.44	58	1.06	66	14.2
51	0.47	59	1.2	67	15.9
52	0.47	60	1.35	68	18.2
53	0.59	61	1.49	69	21.2
54	0.66	62	1.61	70	4.3
55	0.73	63	2.12	71	2.4
56	0.81	64	11.9	72	2.7
57	0.88	65	12.4	73	2.9

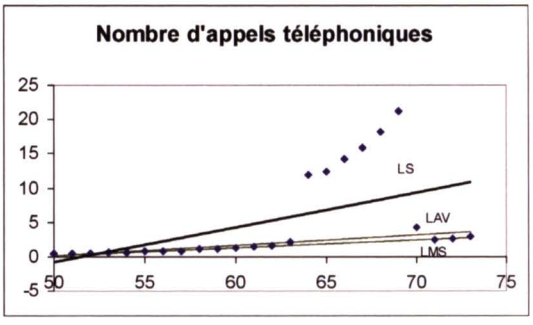
Elles sont représentées ci-dessous :



7. Nombres d'appels téléphoniques internationaux de 1950 à 1973

À l'observation du graphique, nous remarquons immédiatement une forte contamination des données correspondant aux années 1964 à 1969. Une enquête plus poussée explique cette contamination : elle est due à un changement dans le système d'enregistrement durant ces années. En effet, durant ces six années, le relevé a été effectué non pas sur le nombre d'appels, mais bien sur le nombre total de minutes de ces appels, ce qui explique les valeurs importantes prises par ces données par rapport aux années précédentes et suivantes. Notons que les changements d'enregistrement s'étant produits au cours des années 1963 et 1970, celles-ci ont également été contaminées, mais dans une moindre mesure.

Pour cet exemple, pour lequel les valeurs aberrantes peuvent être facilement détectées à la lueur du nuage de points, les trois droites de régression évoquées ci-dessus sont représentées sur le graphique suivant :



8. Nombre d'appels téléphoniques internationaux de 1950 à 1973 et droites de régression.

Nous remarquons immédiatement le manque de robustesse de la droite des moindres carrés (en abrégé et en anglais : LS, pour *Least Squares*) qui est très attirée par les observations contaminées. Les points aberrants étant principalement éloignés verticalement, la droite des moindres écarts absolus (en abrégé et en anglais : *LAV* pour *Least Absolute Values*) est relativement fiable : elle n'est que peu attirée par les valeurs aberrantes. La droite des moindres carrés médians (en abrégé et en anglais : *LMS* pour *Least Median Squares*) n'est quant à elle pra-

Bel exemple d'augmentation utile d'appels téléphoniques : au mois d'août 1914, la petite ville d'Étain (département de la Meuse) a subi deux bombardements. La ville fut bientôt en flammes. Le bureau de poste était resté confié à la garde d'une jeune employée. Pendant que les obus pleuvaient sur la ville, elle se tenait dans son bureau, téléphonant de quart d'heure en quart d'heure à Verdun pour rendre compte de ce qui se passait.

Droite de régression

$$S(a,b)=\sum_{i=1}^{n}\left[(y_i-\bar{y})-a(x_i-\bar{x})+(\bar{y}-a\bar{x}-b)\right]^2$$

$$=ns_y^{\,2}+na^2s_x^{\,2}+n(\bar{y}-a\bar{x}-b)^2-2ans_{xy}$$

$$=n\left[\left(as_x-\frac{s_{xy}}{s_x}\right)^2+(\bar{y}-a\bar{x}-b)^2+s_y^{\,2}-\left(\frac{s_{xy}}{s_x}\right)^2\right]$$

Les valeurs de a et b minimisant cette expression correspondent à celles qui annulent les deux premiers termes du dernier crochet, soit $a=\dfrac{s_{xy}}{s_x^{\,2}}$ et $b=\bar{y}-a\bar{x}$.

tiquement pas affectée par la présence de points atypiques ; elle fournit des résidus très faibles pour les observations non contaminées et des résidus très importants pour les points aberrants, ce qui permet évidemment une détection facile de ces derniers.
Les équations des différentes droites s'écrivent :

Droite des	Équation
moindres carrés	$y = 0.5x - 26$
moindres carrés absolus	$y = 0.153x - 7.519$
moindres carrés médians	$y = 0.1155x - 5.6175$

Pour information, nous pouvons également comparer les valeurs des minima obtenus à partir des différents critères : Conformément aux informations apportées par la figure 8, nous observons des résultats très proches pour les deux dernières droites et l' « isolement » de la droite des moindres carrés par rapport aux deux autres.

À minimiser / Droite des	$\sum_{i=1}^{n}(y_i-ax_i-b)^2$	$\sum_{i=1}^{n}\lvert y_i-ax_i-b\rvert$	Médiane des $(y_i-ax_i-b)^2$
moindres carrés	695.44	101.887	11.91
moindres carrés absolus	1069.167316	84.4	0.10433
moindres carrés médians	1165.397921	84.994	0.007396

Ces considérations veulent montrer que pour réaliser une régression linéaire à partir de données concrètes, il convient d'être prudent. En effet, même si la droite des moindres carrés possède des propriétés et des applications qu'on ne peut négliger, il apparaît qu'elle est très sensible à la présence de valeurs aberrantes. Lorsque c'est possible, il est recommandé de rechercher, en plus de celle-ci, une droite obtenue par une technique robuste. Si les résultats fournis par les deux méthodes sont cohérents, alors la droite des moindres carrés ainsi que ses propriétés peuvent être exploitées sans crainte. Par contre, lorsque les résultats diffèrent sensiblement, il est raisonnable d'affiner l'étude et notamment d'analyser les observations mises en évidence par des procédures robustes.

V.H.

par Nicolas Delerue

Petits poissons

et particules élémentaires

Quand les pêcheurs reviennent bredouille, peut-on en déduire qu'il n'y a pas de poissons ? Et quand on ne trouve pas ce que l'on cherche, faut-il conclure qu'il n'y a rien à trouver ou bien que l'on a mal cherché ? Pour répondre à de telles questions, les statisticiens ont mis au point des méthodes d'évaluation de la qualité d'une recherche qui s'appliquent même à la physique des particules.

Voilà près d'un mois qu'ils sillonnaient en vain le lac Léman à la recherche de la perche bleu rarissime (perca azur rarissimus). Le professeur en était pourtant sûr : d'après ses calculs, il devait y avoir des perches bleu rarissime dans le lac Léman.

Malheureusement, il n'avait aucune idée de la profondeur à laquelle elles se trouvaient. Chaque jour, après avoir remonté les filets, ils avaient examiné minutieusement chacun des poissons attrapés dans l'espoir de confirmer la prédiction de la théorie, au total 218 467 poissons, et chaque fois leurs filets semblaient ne contenir que des perches vertes communes.

À une exception toutefois : deux poissons possédaient des caractéristiques qui pouvaient les identifier comme perches bleu rarissime. Mais il arrive que les perches vertes communes deviennent bleues sous l'action du soleil. Et comme la probabilité d'une telle mutation est de l'ordre de un pour cent mille, ces deux perches (sur plus de 200 000) candidates à l'appellation "bleu rarissime" pouvaient donc s'avérer de vulgaires perches vertes communes. Ils les avaient tout de même envoyées à un laboratoire pour des examens génétiques approfondis qui prendraient de longs mois. Leur campagne de pêche terminée, le professeur se lamentait qu'elle n'avait servi à rien et que les perches bleu rarissime n'existaient peut-être pas.

La pêche aux statistiques

Clara, l'une des chercheuses du groupe, passionnée de statistiques, vint le rassurer : "Mais non. même si nous n'avons pas trouvé la perche bleu rarissime, tout notre travail n'a pas été inutile. En utilisant les résultats de notre campagne de pêche, nous pouvons déterminer la densité maximale de perche bleu rarissime en fonction de la profondeur à laquelle nous avons péché. Grâce à ces renseignements, vous pourrez probablement améliorer votre modèle.

- Comment cela ? demanda le professeur.

- Tout d'abord, il faut comparer le modèle avec nos données pour chaque profondeur à laquelle nous avons péché. Nous obtiendrons ainsi une fonction de similarité, *likelihood function* comme disent les Anglo-Saxons. Cette fonction de similarité permet de se faire une idée de la validité du modèle au vu des données expérimentales. Dans notre cas, seules deux perches ressemblent aux prédictions, et la fonction de similarité indiquera peu de similarité, mais si nous avions plus de candidats, c'est grâce à elle que nous pourrions évaluer la validité du modèle.

Ensuite, nous allons faire varier la densité de perche bleu rarissime vivant à une profondeur donnée jusqu'à ce que 95 % de nos données soient compatibles avec celles prédites par votre modèle, cette densité sera ce que nous appellerons notre limite à 95 % de niveau de confiance.

- Mais comment estimer exactement les données prédites par le modèle ?

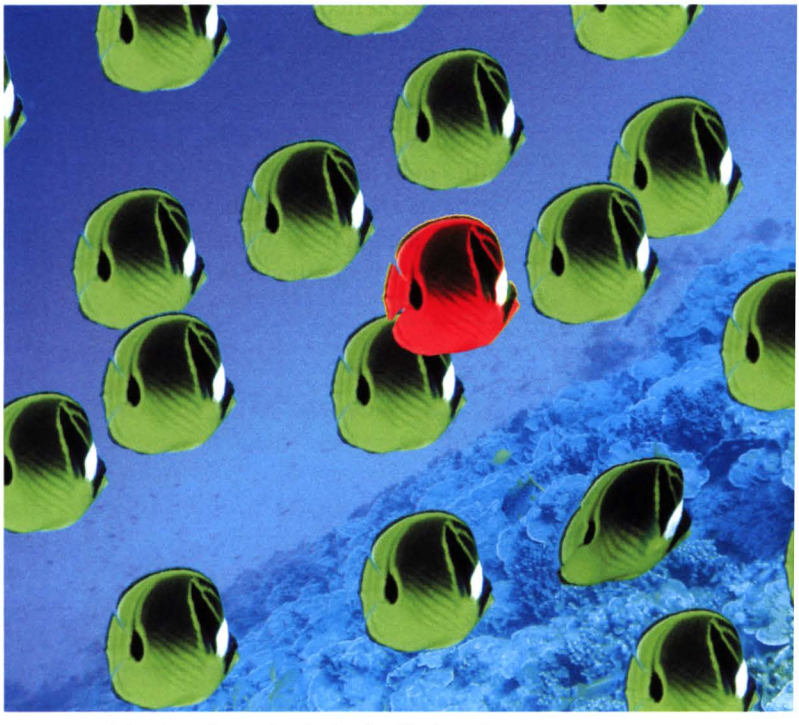
- Nous pourrions écrire un programme de simulation informatique utilisant votre modèle. Ce programme nous prédirait alors quel aurait dû être le résultat de notre pêche...

- Cela va demander un peu de travail mais n'a pas l'air trop compliqué.

- En fait, il y a une petite difficulté supplémentaire : il existe de nombreuses définitions de la limite de confiance et les statisticiens qui travaillent sur la question n'ont jamais été capables de se départager.

- Comment cela, toutes ces définitions ne sont-elles pas équivalentes ?

- Dans la plupart des cas, elles le sont, mais dans des cas très particuliers, certaines méthodes donnent des résultats incohérents. Par exemple, si nous avions trouvé moins de perches candidates à l'appellation "perche bleu rarissime" que votre prédiction basée sur le nombre de perches vertes communes ayant déteint, certaines méthodes de calcul de la limite de confiance auraient pu nous donner une densité maximale négative, ce qui n'a, bien entendu, pas de sens. À cela s'ajoutent des problèmes plus philosophiques sur la manière de calculer les

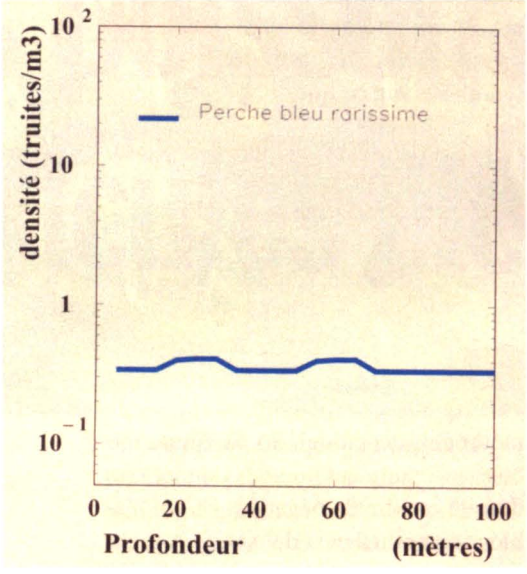


Comment peut-on filtrer les informations parasitées par le bruit de fond ? Les méthodes statistiques de calcul des limites apportent des solutions et ont connu ces dernières années de très nombreux développements, essentiellement grâce à la demande des physiciens des particules qui en sont les principaux utilisateurs.

probabilités qui ont donné naissance à deux écoles différentes : les "bayésiens" se basent sur le théorème de Bayes pour calculer leurs probabilités ; or ce théorème nécessite une connaissance a priori du système, ce que refusent les partisans de l'autre école dite classique ou fréquentiste.

- Qu'entendent-ils par connaissance a priori ?

- Pour donner un exemple très simplifié, si vous demandez la probabilité qu'il pleuve à un partisan de la méthode classique, il utilisera les dernières statistiques disponibles pour répondre à la question. Un bayésien va d'abord regarder le ciel et, s'il y a des nuages, il augmentera, fort de cette information, la probabilité a priori. Il est évident que, dans cet exemple, c'est la méthode bayésienne qui est la plus sensée, mais dans le cas de la recherche scientifique d'objets inconnus, il est moins évident que l'on puisse faire la moindre hypothèse sur ce que l'on cherche, d'où cette scission des statisticiens en deux écoles.



Courbe de limite de densité des truites bleu rarissime en fonction de la profondeur. Tous les points au-dessus de la courbe sont exclus avec un taux de confiance de 95 % par la campagne de pêche du professeur et de Clara (données fictives.)

- Mais quelle méthode choisir alors ?

- C'est à nous de décider en fonction de nos hypothèses et de nos convictions personnelles vis-à-vis du théorème de Bayes. L'important est de bien préciser la méthode utilisée.

- Alors allons-y !"

Un argument d'une certaine profondeur

Quelques semaines plus tard, le programme de simulation était prêt et le professeur alla trouver Clara pour lui montrer la courbe de limite.

"Clara, il y a quelque chose que je ne comprends pas dans cette courbe: à quoi correspondent les deux bosses que nous voyons ?

- Vous souvenez-vous que vous aviez deux perches candidates ? À la profondeur où nous avons trouvé ces deux perches, il subsiste donc un léger doute sur l'existence possible de perches à cette profondeur, donc nous ne pouvons pas exclure toutes les densités que nous excluons aux autres profondeurs."

Quelques heures plus tard, Clara vit le professeur débouler en trombe dans son bureau : "J'ai une idée des raisons de l'échec de notre première campagne !

- Ah bon! Et à quoi est-il dû ?

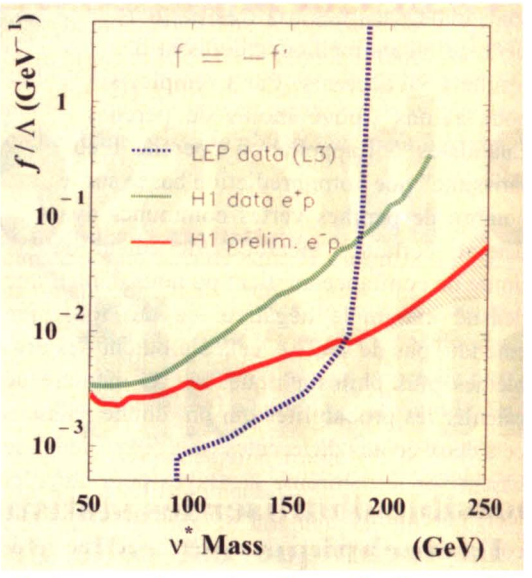
- Les perches bleues rarissimes mangent des algues bleues qui sont elles aussi très rares or je viens de lire les travaux d'un collègue qui excluent la présence de ces algues dans des densités suffisantes au moins jusqu'à 125 mètres de profondeur ! Donc nous aurions dû chercher à plus de 125

mètres de profondeur...

- Vous comprenez que notre première campagne de pêche n'a pas été totalement inutile maintenant ? Grâce aux limites que nous avons mises, d'autre chercheurs pourront, s'il les lisent, affiner leurs modèles et leurs recherches."

La pêche aux particules

Laissons-là Clara et le professeur en leur souhaitant une bonne pêche pour s'intéresser à ce qui se passe à quelques kilomètres de là en banlieue de Genève. Là, au CERN (Centre Européen de Recherche Nucléaire), se trouve l'un des accélérateurs de particules les plus puissants du monde, le LHC. Les physiciens des particules qui y travaillent tentent, entre autres, de trouver de nouvelles particules prédites par différents modèles. Elles sont affublées de noms poétiques ou étranges : boson de Higgs, particules super-symétriques, particules composites, leptoquarks etc. La plupart de ces traques ne débouchent pas sur la découverte d'une nouvelle particule, les physiciens calculent alors des limites sur les propriétés de la particule recherchée et ces limites aident les auteurs du modèle à le perfectionner et à le préciser jusqu'au jour où cette particule est enfin découverte. Eux aussi sont confrontés à ce qu'ils appellent des problèmes de bruit de fond, comme l'étaient les perches vertes communes qui avaient déteint pour le professeur. Après la découverte de la particule, il faut en mesurer les propriétés et distinguer les valeurs possibles des valeurs moins possibles. Cela se fait en calculant des limites, et là, pour les valeurs supérieures et les valeurs inférieures possibles, on parle d'intervalles de confiance.



Limites présentées lors de la conférence ICHEP 2000 à Osaka en juillet 2000. La figure montre une limite sur la manière dont une particule hypothétique appelée neutrino excité (une particule du modèle composite) réagit avec son entourage en fonction de sa masse, pour différentes expériences.

Un bel exemple d'interactions entre les sciences, puisque les méthodes statistiques de calcul des limites ont connu ces dernières années de très nombreux développements, essentiellement grâce à la demande des physiciens des particules qui en sont les principaux utilisateurs.

N. D.

Nicolas Delerue est enseignant-chercheur a l'Universite d'Oxford dans le departement de Physique des Particules

Bibliographie (en anglais) :
Workshop on confidence limits, F. James, L. Lyons et Y. Perrin, CERN Yellow report, 30 may 2000, CERN Genève (disponible gratuitement sur le site du CERN : http://www.cern.ch)

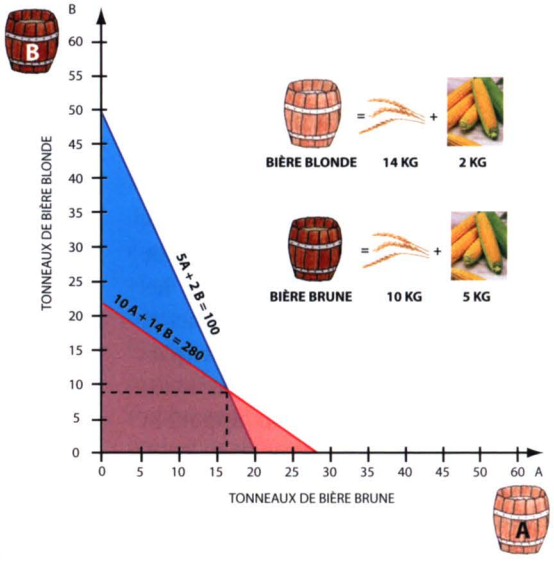
par Philippe Boulanger

Le brasseur satisfait

Il est possible d'utiliser les statistiques pour optimiser son profit. La technique est celle de l'optimisation linéaire, expliquée sur l'exemple du brasseur. Elle s'applique à de nombreuses industries.

« Le profit est bénédiction quand il n'est pas volé »
William Shakespeare (1564-1616)

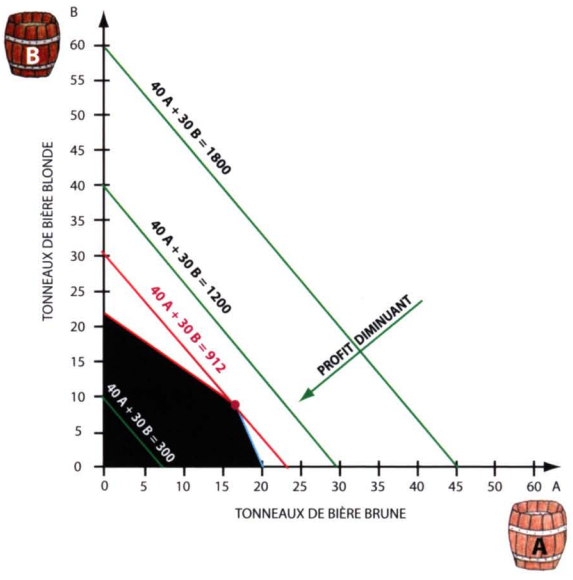
L'optimisation des données est fort utile au commerçant. Prenons le cas d'un brasseur qui vendait de la bière brune et de la bière blonde. Grâce à des statistiques accumulées pendant des années, il savait que la bière brune lui rapportait 40 euros par tonneau et que le bénéfice pour la bière blonde n'était que 30 euros par tonneau. Le brasseur était limité par ses stocks de maïs et d'orge. Il lui fallait 5 kilogrammes de maïs et 10 kilogrammes d'orge pour faire un tonneau de bière brune ; pour un tonneau de bière blonde il consommait 2 kilogrammes de maïs et 14 kilogrammes d'orge. Comment pouvait-il optimiser son bénéfice, sachant qu'il avait en stock 100 kilogrammes de



1. Les quantités en tonneaux de bière brune et de bière blonde sont en dessous des droites bleues et rouges. La zone permise est en violet et la quantité maximale obtenue sans reste est au coin de cette zone.

maïs et 280 kilogrammes d'orge? S'il fabriquait de la bière brune uniquement, il était limité par la quantité de maïs et ne pouvait faire que 20 tonneaux, il lui restait 80 kilos d'orge inutilisés, et son bénéfice était de 800 euros. S'il produisait seulement de la bière blonde, il était limité par la quantité d'orge et brassait 20 tonneaux avec un bénéfice limité à 600 euros et il lui restait 60 kilos de maïs inutilisés.

Là, l'optimisation aide. Nous utiliserons un graphique : sur un axe nous indiquons le nombre A de tonneaux de bière brune fabriquée, sur l'axe perpendiculaire, le nombre B de tonneaux de bière blonde. Pour faire A tonneaux de bière brune et B de bière blonde il faut 5A plus 2B kilos de maïs. Comme notre ami brasseur n'a que 100 kilos de maïs, traçons sur notre graphique la droite 5 A + 2 B égal à 100. Cette droite délimite une zone bleue. En chacun de points de cette zone, il y a suffisamment de maïs pour faire les quantités de bière brune et de bière blonde indiquées par les coordonnées d'un point. Nous traçons similairement la droite 10 A + 14 B = 280, qui délimite une seconde zone rouge où le brasseur a suffisamment d'orge.
Ces deux droites se coupent en un point de coordonnées correspondant à



2. Les profits constants sont situés sur des droites parallèles et le profit maximal est celui dont la droite passe par le coin de la zone permise (en noir).

8 tonneaux de bière blonde et 16,8 tonneaux de bière brune… C'est là que le brasseur utilise entièrement son stock de maïs et d'orge et que son bénéfice est maximal : 912 euros.

Comment sait-on que le bénéfice est maximal ? Nous allons le démontrer : nous avons vu que le tonneau de bière brune rapporte 40 euros et le tonneau de bière blonde, 30 euros. Donc le bénéfice pour A tonneau de bière brune et B tonneau de bière blonde est de 40 A + 30 B. Les droites correspondant à un bénéfice constant sont des droites représentées sur la figure 2. Le bénéfice est d'autant plus grand que la distance au point d'origine est plus grande. Ainsi le profit maximal se trouve au coin formé par la droite rouge et la droite bleue.

Cette méthode, dénommée la méthode du simplexe, inventée il y a une cinquantaine d'années, par le mathématicien Georges Dantzig, a d'innombrables applications industrielles. Dans le cas de la brasserie, on la modifie à volonté pour adapter la production en fonction des stocks de maïs et de houblon, et dans les optimisations industrielles on a souvent affaire à des milliers de contraintes. Les domaines sont alors délimités par des hyperplans, mais la méthode reste identique. Les optimisations de ce type représentent pour les compagnies d'aviation et les distributeurs de produits pétroliers quelques dixièmes de leurs frais de calcul sur ordinateur.

P. B.

Statistiques de comptoir (2)

→ **La bombe protectrice.** Un avion sur 1000 000 est sujet à un attentat par bombe. Voyager avec votre propre bombe diminue donc votre risque car la probabilité qu'un avion ait deux bombes est quasiment nulle. C'est le sophisme des variables indépendantes : la présence d'une bombe ne modifie pas la probabilité qu'il y en ait une autre.

→ **Conclusion hâtive.** 90% des accidents de voiture surviennent à moins de 5 kilomètres du domicile : faut-il sortir aussi rapidement que possible de cette zone à risque ? Et y rouler au-dessus de la vitesse limite, ce qui augmente le risque d'accidents ? En réalité, la statistique ne fait que montrer que, la plus grande partie du temps, vous conduisez près de chez vous…

→ **Biais fatal.** Le magasine américain *Literary Digest* avait organisé, en 1936, une enquête pour savoir qui serait élu président. Une grande majorité des lecteurs indiquèrent Landon, alors que Roosevelt fut élu avec une marge confortable. Pourquoi ? Parce que les lecteurs étaient tous de la même classe sociale fortunée et anti-Roosevelt. Le journal fit faillite après cette catastrophe. *O tempora, o mores.*

→ **Paternité contestable.** Qui a écrit la phrase : There are three kinds of lies: lies, damned lies, and statistics ? (Il y a trois sortes de mensonges, les mensonges, les sacrés mensonges et les statistiques). On l'attribue quelquefois à Mark Twain, qui cite, en 1907, l'homme d'État anglais Benjamin Disraeli.
La phrase aurait aussi été l'œuvre de Henry Du Pré Labouchère (1831-1912) ou de Leonard Courtney en 1895. En 1894, un médecin M. Price, rapporte qu'il a lu dans le *Philadelphia County Medical Society* la phrase proverbiale : « Il existe trois sortes de tromperies "Les mensonges, les sacrés mensonges et les statistiques".» Si la phrase est proverbiale, c'est que son auteur est encore antérieur.

→ **Crime,** que de mensonges commet-on en ton nom ! Un journal de 1995 a publié : « Depuis 1950, le nombre d'enfants américains tués par arme à feu a doublé chaque année ». Vous auriez bien tort de croire cette affirmation. Si, en 1950, un seul enfant aux États-Unis avait été abattu par balle, en doublant le résultat 44 fois, vous obtiendriez un total de 35000 milliards d'enfants victimes !

→ **Comparer le comparable.** Un fervent trop zélé de la pilule contraceptive affirmait que le nombre des accidents médicaux liés à l'aspirine faisait de la pilule un principe actif extraordinairement sûr. En réalité, la personne comparait les accidents liés à un usage normal de la pilule à une prise excessive d'aspirine. À vouloir trop prouver…

par Hervé Lehning

L'art de tricher

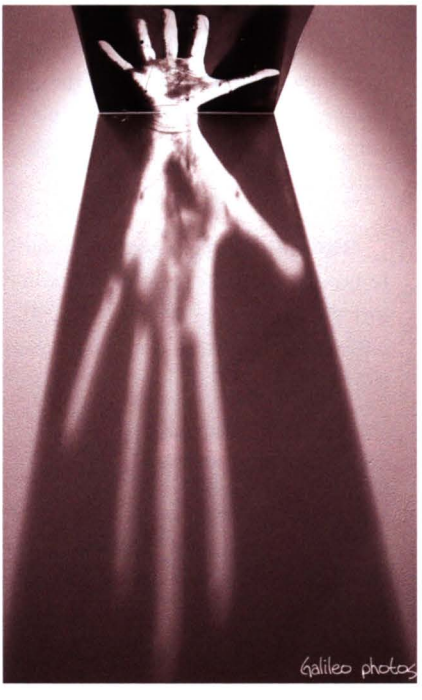
Quatrième carré d'as consécutif pour votre adversaire. Il triche, vous en êtes sûr ! Comment le coincer ? Cette situation se retrouve dans la vie courante. Comment débusquer les tricheurs des sondages ?

Comment détecter les tricheurs ?

Il y a cent façons de tricher, mais il n'y a guère que trois sortes de tricheurs. Tout d'abord, il y a le joueur qui triche - qui ne triche que parce qu'il joue. Qui le fait sans méthode, sans préméditation, d'une manière presque inconsciente, involontaire, et dont on sent très bien qu'il est parfaitement honnête en dehors du jeu. Il y a l'homme qui joue incorrectement parce qu'il est incorrect d'un bout à l'autre de la vie - et qui doit penser que ce n'est vraiment pas le moment de l'être. Enfin, il y a le tricheur de profession, conscient et organisé.

Les mémoires d'un tricheur

Dans ce texte, Sacha Guitry nous décrit trois sortes de tricheurs. Comment les détecter sans les prendre sur le fait ? Un peu de psychologie suffit dans les deux premiers cas. Le vrai professionnel est d'une autre nature. Les mathématiques sont alors un outil essentiel.



Un enquêteur nommé Bidon

Un institut de sondage emploie des centaines d'enquêteurs. Avec la méthode des quotas, ceux-ci doivent trouver un certain nombre de représentants de plusieurs catégories. Imaginez que l'enquêteur Bidon ait du mal à finir son travail. Il lui faut encore dix personnes dans le groupe des fameuses ménagères de moins de 50 ans, si chères aux instituts de sondages. Pour gagner du temps et donc de l'argent, la tentation est grande de remplir les questionnaires lui-même.

Que peut faire son employeur pour éviter cette tromperie ? La première méthode est extra mathématique : lui faire peur ! Les sondeurs procèdent donc à des contrôles en demandant les coordonnées des sondés. La mesure n'est guère efficace. Une bien meilleure méthode est de tester les résultats des enquêteurs. S'ils trichent beaucoup, cela se voit très vite : ils sont trop réguliers. Trop de personnes se situent dans la moyenne. Si la tricherie se situe sur peu de sondés, elle n'aura guère d'influence sur le résultat final.

L'art de tricher

Imaginez un exemple simple : la taille moyenne des hommes adultes en France est de 175,7 cm. Vous devez fournir des statistiques portant sur 100 hommes. Que fait le tricheur maladroit ? La réponse est simple, il obtient exactement la moyenne de 175,7 ce qui est très peu probable. S'il veut un résultat vraisemblable, il ne doit pas procéder ainsi. Tout d'abord, la moyenne ne suffit pas. La dispersion est essentielle. Cette dispersion est mesurée par l'écart-type. Quel est le pourcentage de Français de chaque taille possible ? Ici

encore, ne vous collez pas trop à la moyenne, et n'en décollez pas de façon trop systématique. Rien de plus difficile à simuler que le hasard.

Simulation du hasard

Pour cela, faites confiance aux vrais professionnels, aux mathématiciens. Ils ne vous tromperont pas ! Imaginez que vous vouliez répartir 1 000 sondés dans 6 groupes. Sur la population totale, les pourcentages de chacun de ces groupes sont 5, 20, 25, 25, 20, 5 pourcents.

Si vous voulez passer pour un escroc doublé d'un imbécile, proposez la distribution suivante :

50	200	250	250	200	50

Nous y trouvons les pourcentages exacts de la population entière. C'est invraisemblable. Il est plus habile de les modifier un peu, par exemple :

49	198	254	246	205	48

C'est mieux bien sûr mais loin de la perfection. N'essayez pas davantage dans cette direction, vous n'y arriverez jamais !

Le tricheur à l'as de carreau, Georges de La Tour (1593 – 1 652)

Alors comment faire pour tricher intelligemment ? Utilisez un générateur de nombres pseudo-aléatoires entre 1 et 100 comme en fournissent les principaux langages de programmation et même les tableurs. Leur nom vient de l'anglais, essayez « random » ou « rand ». Si vous tirez un nombre entre 1 et 5, vous le comptez dans le premier groupe. Si vous tirez un nombre entre 6 et 25, vous le comptez dans le second, et ainsi de suite. Vous obtiendrez ainsi une distribution vraisemblable. Voici ce que nous trouvons en l'appliquant dix fois :

41	209	240	263	199	48
38	220	253	224	220	45
51	176	243	274	205	51
53	192	270	240	201	44
53	185	240	259	208	55
59	189	238	255	203	56
49	195	249	241	216	50
46	189	252	260	191	62
54	209	263	222	213	39
54	192	242	285	184	43

Une excellente contrefaçon

Pourquoi faire confiance à ces suites de nombres pseudo-aléatoires ? Il s'agit en fait de suites déterministes offrant certaines caractéristiques du hasard. En voici une. Prenez l'heure actuelle en secondes, soit X, puis calculez le reste du produit de 16 807 par X dans la division par 2 147 483 647. C'est votre premier nombre. Appelez-le X et recommencez la procédure précédente. La suite de nombres entre 0 et 2 147 483 646 obtenues passe tous les tests statistiques usuels, utilisés pour contrôler le hasard. Pourtant, cette suite est complètement déterminée par son premier élément !

H. L.

Le choix collectif

La démocratie est fondée sur le quantitatif : la majorité a tout le pouvoir. Hélas le résultat du vote dépend grandement de la manière dont on compte les votes, d'où les mécontentements et les triomphes exagérés. Il est fréquent que la marge de succès d'un vote est inférieure aux erreurs de dénombrement.

par Philippe Boulanger

Le rêve du système de vote parfait

Selon la manière de compter les résultats des votes, on peut faire gagner n'importe qui !

Il faudrait que nous ayons un système électoral juste, souhaitent tous les électeurs, c'est-à-dire qui reflète au mieux leurs choix. Hélas cela est impossible. Le résultat du vote dépend du système utilisé et il n'existe pas de système de vote parfait.

Voyons cela avec l'élection de la plus belle actrice de tous les temps le vote étant fait à l'échelle de la France. Nous avons cinq candidates, Marylin Monroe, Angelina Jolie, Brigitte Bardot, Emmanuelle Béart et Claudia Cardinale. Les électeurs se répartissent en six catégories de choix : ainsi 7,2 millions d'électeurs placent en premier choix Marylin Monroe, en deuxième choix Emmanuelle Béart, en troisième choix Claudia Cardinale, en quatrième choix, Brigitte Bardot, et en cinquième choix, Angelina Jolie. 4,8 millions d'électeurs placent au premier rang Angelina Jolie, en deuxième rang Claudia Cardinale etc.

Votants (en millions)						
	7,2	4,8	4	3,6	1,6	0,8
1er choix	M. Monroe	A. Jolie	B. Bardot	E. Béart	C. Cardinale	C. Cardinale
2e choix	E. Béart	C. Cardinale	A. Jolie	B. Bardot	A. Jolie	B. Bardot
3e choix	C. Cardinale	E. Béart	C. Cardinale	C. Cardinale	E. Béart	E. Béart
4e choix	B. Bardot	B. Bardot	E. Béart	A. Jolie	B. Bardot	A. Jolie
5e choix	A. Jolie	M. Monroe	M. Monroe	M. Monroe	M. Monroe	M. Monroe

Examinons un premier système électoral, le plus simple, une élection à un

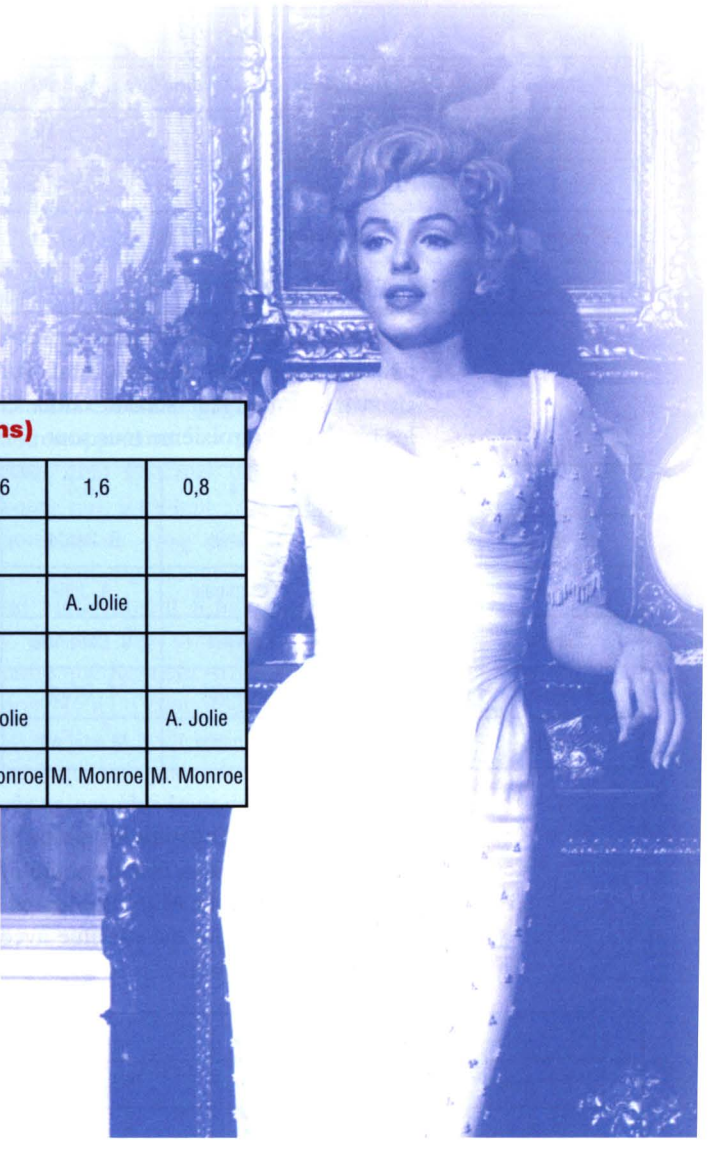
tour : chaque électeur vote pour son premier choix. Marylin Monroe est élue avec 7,2 millions de voix.

Le système est-il juste ? 14,8 millions de personnes préféraient une autre candidate à Marylin Monroe …. Marylin Monroe n'est élue que par moins d'un tiers des votants. N'est-ce pas choquant ?
Appliquons à cette élection frivole le système électoral présidentiel actuel à deux tours où Marylin Monroe et Angelina Jolie sont les deux candidates qui ont eu le plus de voix au premier tour. Alors les électeurs des quatre autres groupes votent, au second tour, selon leur ordre de choix.

Votants (en millions)						
	7,2	4,8	4	3,6	1,6	0,8
1er choix	M. Monroe	A. Jolie				
2e choix			A. Jolie		A. Jolie	
3e choix						
4e choix				A. Jolie		A. Jolie
5e choix			M. Monroe	M. Monroe	M. Monroe	M. Monroe

Comme les électeurs des quatre derniers groupes préfèrent Angelina Jolie à Marylin Monroe, Angelina Jolie est élue avec l'écrasante majorité de 14,8 millions d'électeurs contre 7,2 à Marylin Monroe.
Quel que soit le système électoral, pourrions-nous dire, le résultat sera toujours une de ces deux candidates, Marylin Monroe ou Angelina Jolie. Ce n'est pas si injuste, pourrions-nous penser…
« Que nenni, lui répondons-nous ! Prenons un système à plusieurs tours où, à chaque tour, on élimine la candidate qui a le moins de voix. Ce système correspond au dicton selon lequel les votants éliminent plus qu'ils ne choisissent. Au premier tour, Claudia

L'important n'est pas comment on vote, mais comment on compte les votes.
Pierre Tougne



Cardinale, qui n'obtient que 2, 4 millions de voix, est éliminée. »
Au second tour, 1,6 million des voix de Claudia Cardinale se reportent sur Angelina Jolie, leur second choix, et 0,8 million sur Brigitte Bardot. Aussi, les résultats du second tour sont :

	7,2	6,4	4,8	3,6
1er choix	M. Monroe	A. Jolie	B. Bardot	E. Béart
2e choix	E. Béart	C. Cardinale	A. Jolie	B. Bardot
3e choix	C. Cardinale	E. Béart	C. Cardinale	C. Cardinale
4e choix	B. Bardot	B. Bardot	E. Béart	A. Jolie
5e choix	A. Jolie	M. Monroe	M. Monroe	M. Monroe

À ce tour, Emmanuelle Béart est éliminée, et les votes se reportent sur Brigitte Bardot, leur second choix et les résultats du troisième tour sont :

	7,2	6,4	8,4
1er choix	M. Monroe	A. Jolie	B. Bardot
2e choix	E. Béart	C. Cardinale	A. Jolie
3e choix	C. Cardinale	E. Béart	C. Cardinale
4e choix	B. Bardot	B. Bardot	E. Béart
5e choix	A. Jolie	M. Monroe	M. Monroe

Au tour suivant, Angelina Jolie est éliminée, et comme ses électeurs préfèrent Brigitte Bardot à Marylin Monroe, c'est Brigitte Bardot qui est élue avec 14,8 millions de voix.

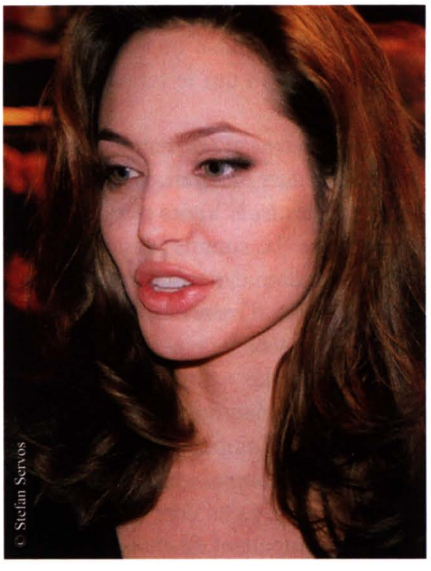
	7,2	14,8
1er choix	M. Monroe	B. Bardot
2e choix	E. Béart	A. Jolie
3e choix	C. Cardinale	C. Cardinale
4e choix	B. Bardot	E. Béart
5e choix	A. Jolie	M. Monroe

Et si nous utilisions le système du mathématicien Borda (1733-1799), où chaque votant attribue 5 points au premier choix, 4 au deuxième choix, 3 au troisième, *etc.* ? Avec ce système à un tour qui serait élue ?

Faisons les comptes :

Votants (en millions)						
	7,2	4,8	4	3,6	1,6	0,8
1er choix	M. Monroe	A. Jolie	B. Bardot	E. Béart	C. Cardinale	C. Cardinale
2e choix	E. Béart	C. Cardinale	A. Jolie	B. Bardot	A. Jolie	B. Bardot
3e choix	C. Cardinale	E. Béart	C. Cardinale	C. Cardinale	E. Béart	E. Béart
4e choix	B. Bardot	B. Bardot	E. Béart	A. Jolie	B. Bardot	A. Jolie
5e choix	A. Jolie	M. Monroe	M. Monroe	M. Monroe	M. Monroe	M. Monroe

Le vainqueur est Emmanuelle Béart, qui obtient, en suivant les colonnes de gauche à droite :
$(7{,}2 \times 4) + (4{,}8 \times 3) + (4 \times 2) + (3{,}6 \times 5) + (1{,}6 \times 3) + (0{,}8 \times 3) = 76{,}4$ millions de points.
Les autres obtiennent moins de points, c'est hallucinant.
La seule candidate qui n'ait pas gagné avec un des quatre systèmes électoraux que nous venons d'examiner est Claudia Cardinale : nous allons la faire gagner avec un dernier système, le système de Condorcet (1743-1794)… Dans celui-ci nous opposons chaque candidate à toutes les autres, et nous comptons celle qui a le plus de victoires.


© Stefan Servos

Claudia Cardinale gagne contre Angelina Jolie : elle obtient 7,2 + 3,6 + 1,6 + 0,8, soit 13,2 millions de voix, alors qu'Angelina Jolie n'obtient que 4,8 + 4 égale 8,8 millions de voix. Et Claudia Cardinale gagne contre tous les autres opposants.
Avec cinq systèmes de décompte des votes qui semblent également justes, nous avons cinq résultats différents. Ceux qui choisissent le système électoral déterminent l'heureux gagnant. C'est sur des considérations de cet ordre que Kenneth Arrow (né en 1921), prix Nobel d'économie en 1972, prouva qu'il n'y avait pas de système électoral qui soit juste. Arrow traitait bien sûr un cas plus important, les élections présidentielles et en concluait que la démocratie parfaite est un rêve impossible…

P.B.

par Philippe Boulanger

Tout le monde est content

Il existe des systèmes de choix qui satisfont tous les participants !

Dans la ville de Samarkand, trois candidats se présentent au poste de Sultan. L'élection à Samarkand a deux particularités. D'abord une règle d'or : aucun des trois vizirs qui élisent le Vizir n'ont le droit de se présenter, ce qui empêche les ambitions génératrices d'intrigues. Ensuite, dans ce mode d'élection, les trois vizirs sont toujours contents du résultat ! Impossible, s'exclame le téléspectateur aussi attentif que dubitatif ! Nous allons voir que cela est pourtant vrai.

Le sultan Iz Verygood

Donc les trois Vizirs ont le choix entre les trois candidats que nous dénommerons, comme de coutume, A, B, et C. Pour montrer sa préférence, chaque Vizir indique, pour chacun des candidats, la somme en sequins dont il est prêt à voir diminuer son salaire mensuel si ce candidat, qu'il apprécie, est élu, et la somme qu'il désire recevoir en dédommagement si un candidat, en qui il n'a pas confiance, est désigné.

	A	B	C
Vizir 1	+ 1000	– 2000	+ 1000
Vizir 2	+ 1000	– 6000	+ 5000
Vizir 3	– 1000	+ 2000	– 1000
	+ 1000	– 2000	+ 5000

Ainsi le premier Vizir exige 1000 sequins de plus si le candidat A, qu'il n'aime pas, est élu Sultan ; il est prêt à abandonner 2000 sequins de son salaire si le candidat B est élu, ce que l'on désigne par – 2000. En revanche il veut 1000 sequins de plus mensuellement si le candidat C est nommé Sultan, car il pense que ce serait un mauvais Sultan qui le ferait travailler inutilement.

Les deux autres Vizirs manifestent similairement leurs préférences. Notons que pour que les trois Vizirs aient la même importance la somme des demandes d'augmentation et des offres de diminution de chacun d'entre eux doit être égale à zéro. À chacun de répartir les sommes selon cette règle.

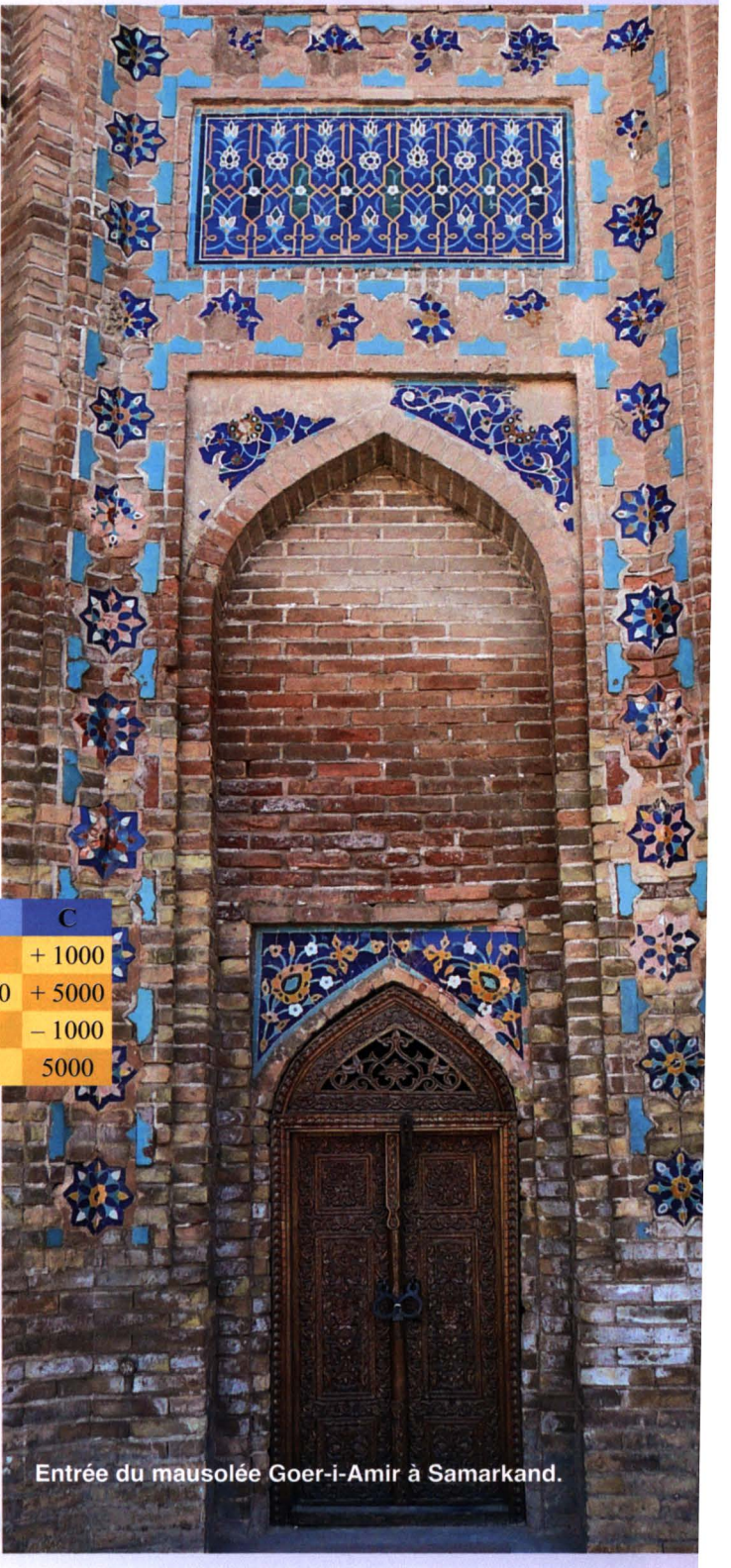
Comment choisit-on alors le meilleur Sultan? En additionnant, dans chaque colonne, les sommes qui se rapportent à sa candidature. Celui dont le total négatif est le plus important est choisi.

Ce qui est bien normal : si le chiffre de ce candidat est négatif, c'est qu'il y a beaucoup de Vizirs qui l'estiment (et qui acceptent des diminutions importantes de leur allocation) et qu'il y a peu de Vizirs qui souhaitent une augmentation de leurs salaires s'il est élu.
Dans notre exemple, c'est le candidat B qui est élu Sultan, car il correspond à une diminution totale des salaires de 6000 sequins, alors que la somme est + 1000 pour le candidat A et + 5000 pour le candidat C.
Maintenant, et c'est là que l'ingéniosité des lois de Samarkand est particulièrement admirable, cette somme de 6000 sequins est redistribuée à chacun équitablement. On peut donner par exemple 2000 sequins de dédommagement supplémentaire au Vizir 3 qui sera ravi, n'imposer qu'une diminution de salaire que de 4000 sequins au Vizir 2 au lieu des 6000 qu'il proposait : il sera heureux. Et de laisser le salaire du Vizir 1 inchangé, au lieu de la perte de 2000 sequins qu'il proposait.

	A	B	C
Vizir 1	+ 1000	– 2000 + 2000 = 0	+ 1000
Vizir 2	+ 1000	– 6000 + 2000 = – 4000	+ 5000
Vizir 3	– 1000	2000 + 2000 = 4000	– 1000
Totaux	+ 1000	– 6000 à distribuer	5000

Ainsi cette méthode d'élection fait que chaque électeur a plus qu'il n'était prêt à abandonner ou qu'il voulait recevoir en dédommagement : les finances publiques ne sont pas pénalisées par cette méthode. Aussi les élections à Samarkand sont-elles, contrairement à celles d'autres pays moins civilisés, source de joie générale parmi les électeurs. Pourquoi ne pas élire ainsi nos dirigeants, s'interrogeons-nous ?

P. B.



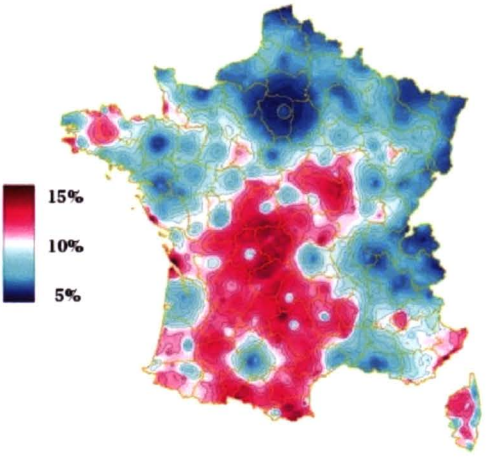
Entrée du mausolée Goer-i-Amir à Samarkand.

par Hervé Le Bras

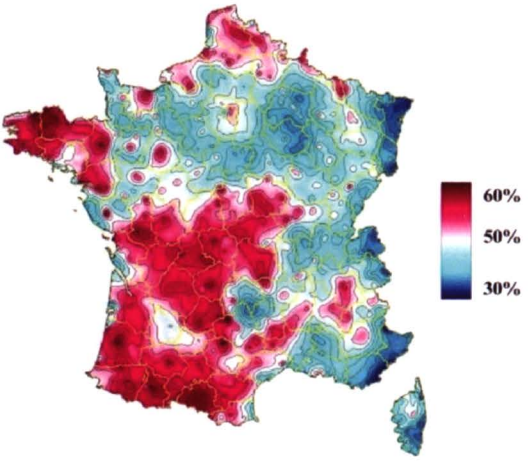
Election, piège à... tromperie écologique

Lors de comparaison des votes il faut se méfier d'un paramètre sous-jacent qui les fausse. L'agrégation de résultats amène les biais.

Les statistiques électorales ont le vent en poupe. Dès le lendemain de chaque élection, le ministère de l'intérieur met en ligne les résultats décomptés dans les 36 565 communes françaises. Avant les élections, les instituts spécialisés multiplient les sondages sur les intentions de vote avec des catégories de plus en plus finement ciblées. À terme, l'accumulation de ces connaissances géographiques et socio-culturelles ne rendrait-elle pas inutile un vote devenu prévisible ?



% de personnes âgées de + de 75 ans En 1999



% votes Royal Au 2e tour de la Présidentielle 2007

© Anne97432

La réponse est négative heureusement. Le secret de l'isoloir ne sera pas percé en raison particulièrement d'une « tromperie écologique » pour reprendre le terme forgé par William Robinson en 1950.

La dernière élection présidentielle en donne un bon exemple. Sur les deux cartes suivantes réalisées à partir des résultats de toutes les communes, on voit la proportion de personnes âgées de plus de 75 ans et la proportion de votes en faveur de Ségolène Royal au second tour. À première vue, les deux cartes se ressemblent beaucoup. Le coefficient de corrélation entre les deux est d'ailleurs élevé (0,56 à l'échelle des départements). Certes quelques désaccords subsistent : l'extrême nord et la région toulousaine sont jeunes et ségolénistes, la côte d'Azur est âgée et sarkoziste, mais pour le reste, un grand sud-ouest, âgé, vote à gauche tandis que le nord de la Loire et l'est du Rhône, plus jeunes votent à droite. L'âge venant, se découvrirait-on de gauche ? Devenus dépendants des autres pour leur retraite et leurs soins, les âgés reporteraient-ils leur confiance en direction de l'État, de la collectivité et de la solidarité au lieu de cultiver l'individualisme?

Avant de transformer la coïncidence des deux cartes en théorie causale, consultons les sondages effectués à la sortie des bureaux de vote de cette

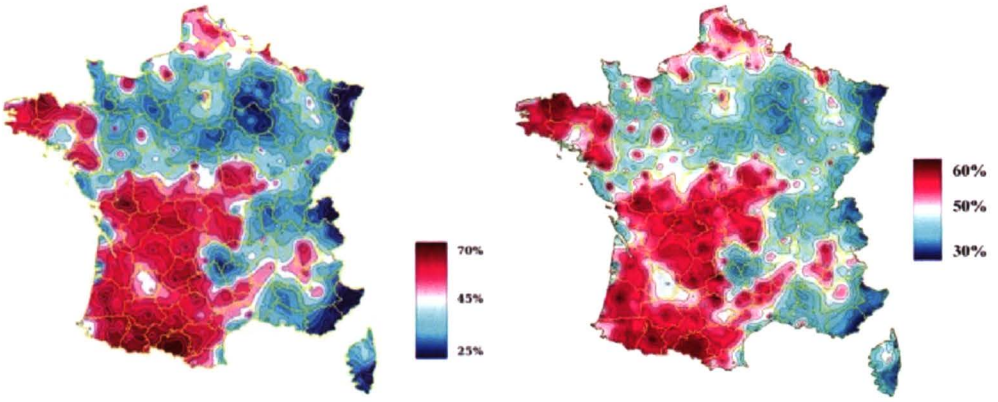
Durant les élections de 1956, un électeur apostropha le candidat Adlai E. Stevenson : « Sénateur, vous avez pour vous toutes les personnes intelligentes ». Stevenson répondit : « Ce n'est pas suffisant, nous avons besoin d'une majorité ! »

même élection présidentielle (sondage LH2). Ils indiquent que 52% des moins de 35 ans ont voté Royal, 52% des personnes de 35 à 65 ans aussi, mais seulement 28% des plus de 65 ans. Ce sont donc les personnes âgées qui ont assuré le succès de Sarkozy, soit l'inverse de ce que laissait supposer la ressemblance des deux cartes. Qui croire ? Le sondage ou la carte ? En fait les deux résultats sont justes, mais séparément. Dans le Sud-Ouest comme à l'Est, les personnes âgées votent plus à droite que les jeunes mais les unes et les autres votent plus à gauche dans le sud-ouest. Un petit exemple va aider à le comprendre. Supposons que la répartition des électeurs (pour 100) selon l'âge et la région soit la suivante :

Âge / région	Nord et Est	Sud-ouest
> 65 ans	10	10
< 65 ans	57	23

Supposons que les votes en faveur de Royal soient dans chacun des quatre groupes :

Âge / région	Nord et Est	Sud-ouest
> 65 ans	20%	40%
< 65 ans	50%	60%



Partage des voix entre Royal et Sarkozy aux deux tours de 2007

Par simple regroupement, on peut vérifier que 30% des personnes de plus de 65 ans votent Royal et 53% des moins de 65 ans, mais que 54% des votes du Sud-ouest sont en faveur de Royal contre 45,5% de ceux du Nord et de l'Est, ce que disaient à la fois le sondage et les cartes : là où l'on trouve le plus de personnes âgées, le vote pour Royal est le plus élevé, mais, les personnes âgées votent globalement moins pour Royal que les jeunes.

Dès lors, faut-il renoncer à comparer des cartes à cause de ce risque de tromperie écologique ? Une troisième carte montre que non. Elle représente le partage des voix entre Royal et Sarkozy au premier tour quand on ne tient compte que de ces deux votes (par exemple si Royal a atteint 20%, Sarkozy, 30% et les autres candidats 50%, Royal obtient 40% des voix qui se sont portées sur elle ou sur Sarkozy seulement). Cette fois-ci, la coïncidence avec la carte des résultats de Royal au second tour est presque parfaite (corrélation de 0,78 à l'échelle départementale) et oriente vers une interprétation intéressante : dès le premier tour, les jeux étaient faits. Là où Royal était forte, elle a récupéré l'électorat des autres candidats et, au contraire, là où Sarkozy dominait, c'est lui qui a bénéficié de la majorité des reports. Il est donc relativement vain de gloser sur le report des électeurs des autres candidats tels Bayrou ou Besancenot, un par un. Ou bien leur préférence pour le second tour était forgée dès avant le premier tour, ou bien la pression de leur entourage s'est exercée de manière homogène, ce qui a mécaniquement renforcé le résultat du premier tour. Le secret de l'urne n'empêche pas la mécanique des votes.

H.L.B. (EHESS, INED)

par Sylvain Merlus

Assurance-auto :

les citadins pénalisés

Les assureurs doivent résister à la concurrence tout en faisant des bénéfices. Pour cela, ils ont besoin d'estimer aussi précisément que possible les coûts qu'entraîneront pour eux les contrats qu'ils proposent, en fonction des caractéristiques des conducteurs et de leurs véhicules. L'appel aux méthodes de régression leur permet d'affiner ce calcul.

La concurrence empêche de demander aux clients présentant les plus faibles risques une solidarité vis-à-vis du reste des assurés.

Le secteur de l'assurance automobile représente un chiffre d'affaires annuel de 800 milliards de francs. C'est un des secteurs économiques qui fait le plus appel aux outils statistiques, car les assureurs disposent de bases de données très riches, qu'ils utilisent pour établir leurs tarifs.

Un contrat d'assurance est un contrat passé entre un assuré et un assureur : le premier paie en début d'année une somme d'argent, appelée la prime ; en contrepartie, le second rembourse les dommages subis ou provoqués par l'assuré, à l'occasion de sinistres survenus dans l'année. Le cycle financier de l'assurance est inversé par rapport aux activités industrielles et de services classiques. En effet, l'industriel qui produit un bien quelconque connaît son prix de revient. Il sait donc à quel prix minimal il doit vendre ce bien pour dégager un profit. Ce n'est pas le cas pour l'assureur, qui doit déterminer le prix de sa prestation avant de la réaliser. Il doit donc, pour vendre celle-ci au meilleur prix, essayer d'évaluer au mieux, à l'aide de techniques statistiques, la fréquence et le coût des sinistres qui surviendront. Le tarif proposé par l'assureur doit être suffisamment haut pour qu'il ne perde pas d'argent sur les contrats qu'il vend, mais il ne doit pas être trop haut, sinon l'assureur perdra ses clients qui iront s'adresser à la concurrence.

La notion de bénéfice ou de perte en assurance ne peut s'appréhender que sur l'ensemble des contrats. Le principe de base de l'assurance est la "mutualisation des risques".

La majorité des assurés paie une prime faible, ce qui permet à l'assureur de régler les sinistres importants qui concernent un nombre réduit de personnes (chaque année, environ un assuré sur 15 a un accident). La charge de sinistre est par nature très aléatoire, aussi est-ce le coût moyen des sinistres par assuré que l'assureur cherche à évaluer.

Une différenciation nécessaire

L'assureur pourrait proposer un tarif unique, calculé à partir d'un coût moyen obtenu en divisant le total des coûts des sinistres qu'il a réglés l'année passée par le nombre de contrats. Pourtant, les comportements des assurés en termes de risque sont très différents. Les citadins ont par exemple plus d'accidents que les habitants des campagnes. L'assureur qui demanderait le même prix aux premiers et aux seconds ferait payer un peu plus que le coût réel aux assurés des campagnes et un peu moins aux assurés des villes. Si un second assureur décidait de n'assurer que les habitants des campagnes, son coût moyen par assuré serait plus faible, il pourrait donc proposer un tarif plus intéressant. Les clients du premier assureur habitant à la campagne le quitteraient pour son concurrent. Le premier assureur n'attirerait plus que les citadins, qui paient pourtant une somme inférieure à ce qu'ils coûtent. Il perdrait de l'argent, puisque ses tarifs ne seraient plus suffisants.

Cet exemple montre que l'assureur doit proposer à chacun le "juste tarif". Il doit discriminer au mieux les comportements différents face au risque, et doit éviter les "subventions croisées" qui consistent à proposer un tarif plus élevé à une catégorie d'assurés pour pouvoir proposer un tarif moins élevé à une autre catégorie tout en maintenant l'équilibre financier global. Pour différencier le tarif en fonction des caractéristiques des assurés, l'assureur pourrait simplement affiner l'approche proposée plus haut : il pourrait créer des catégories de population ayant des caractéristiques identiques, puis calculer la charge moyenne par assuré pour chacune de ces catégories et proposer à chacun un tarif correspondant à ses caractéristiques. Ainsi, si l'on suppose que les coûts des sinistres sont principalement déterminés par l'urbanisation et par la taille du véhicule, l'assureur calculerait le coût moyen pour quatre catégories : ville-petite voiture, campagne-petite voiture, ville-grosse voiture, campagne-grosse voiture. Une telle approche n'est pourtant pas utilisable en pratique. En effet, les assureurs construisent leur tarification en fonction d'un nombre très important de variables : les unes concernent l'assuré : son âge, son sexe, sa profession, son lieu de résidence, l'ancienneté de son permis de conduire ; les autres variables sont relatives au véhicule :

modèle, puissance fiscale, ancienneté. Pour tenir compte directement de toutes ces variables, il faudrait créer une multitude de catégories. Il n'y aurait alors qu'un nombre trop réduit d'assurés par classes, souvent les classes seraient vides, et les estimations de coût moyen ne seraient pas robustes, car elles dépendraient trop de ce qui se serait passé l'année précédente. Ainsi, les assurés qui auraient le malheur d'avoir des caractéristiques

La régression linéaire

Soient deux variables x et y. La variable x est observable (on peut facilement généraliser cette approche au cas où l'on dispose de plusieurs variables observables, mais pour la commodité des calculs, nous présenterons le modèle avec une seule variable explicative). La variable y n'est en revanche pas observable pour tous les individus. On cherche à prévoir pour ces individus la valeur de y la plus probable en fonction de la valeur de x. On suppose que la variable x a une influence linéaire sur y, donc qu'il existe deux nombres réels a et b tels que :

$$y = a\,x + b.$$

Si l'on dispose de l'observation de x et y sur deux individus, on peut déterminer exactement a et b tels que :

$$y_1 = a\,x_1 + b$$
$$y_2 = a\,x_2 + b.$$

Si par contre l'on dispose des valeurs de x et y sur un plus grand nombre d'individus, on ne pourra plus avoir l'égalité : $y_i = a\,x_i + b$ quel que soit i, à moins que tous les points $M_i\,(x_i, y_i)$ soient sur une même droite, ce qui est très peu probable. On doit donc transformer l'écriture précédente et introduire une variable u, appelée "perturbation", définie telle que l'on ait, quel que soit i :

$$y_i = a\,x_i + b + u_i\,.$$

On souhaite que les y_i soient les plus proches possibles des $a\,x_i + b$. Pour cela, on choisit les valeurs de a et b qui minimisent

$$\sum_{i=1}^{n} u_i^2 = \sum_{i=1}^{n} \left(y_i - ax_i - b\right)^2.$$

Le minimum de cette fonction de deux variables $(a\,;\,b)$ est atteint aux points d'annulation des dérivées par rapport à a et b, et l'on obtient aisément :

$$a = \frac{\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} x_i y_i}{\left(\sum_{i=1}^{n} x_i\right)^2 - \sum_{i=1}^{n} x_i^2}, \qquad b = \sum_{i=1}^{n} y_i - a \sum_{i=1}^{n} x_i$$

Une fois a et b ainsi estimés, on peut, pour tous les individus dont on ne connaît pas y_i, estimer cette variable a priori par $a\,x_i + b$. Ainsi, si l'on suppose que seul le poids d'un véhicule détermine le coût moyen des sinistres que le conducteur occasionnera, on peut calculer les a et b correspondants, et prévoir le coût moyen de l'assuré qui se présente pour souscrire un contrat. Ce modèle proposé à titre d'exemple où seul le poids du véhicule intervient ne donnerait bien évidemment pas de résultats satisfaisants. Aussi faut-il poser au départ un modèle suffisamment riche, avec plusieurs variables, pour pouvoir discriminer efficacement les classes de risque.

similaires à celles d'un individu ayant eu un gros sinistre l'année précédente se verraient proposer une prime très élevée, alors que leurs caractéristiques ne les prédisposent pas nécessairement à avoir plus de sinistres que d'autres.

Un modèle de régression

L'assureur doit donc utiliser des techniques statistiques plus élaborées pour évaluer a priori le coût de l'assurance. L'assureur utilise souvent des techniques de régression. Il s'agit de prévoir une grandeur non observable (le coût futur des sinistres) en fonction d'autres caractéristiques que l'on connaît, et ce compte tenu des coûts et des caractéristiques observés dans le passé. Le coût est lié aux caractéristiques par un modèle, qui impose par exemple au coût de dépendre, de manière linéaire, de la taille du véhicule. On recherche le coefficient qui rend le mieux compte de cette dépendance, à partir de l'observation des données passées. Dans de tels modèles, toutes les variables explicatives jugées pertinentes sont introduites.

Supposons par exemple que l'assureur observe n variables, x_1 , ..., x_n, et cherche à déterminer le coût y qui, selon lui, dépend linéairement de x_1, ..., x_n. Il cherche des coefficients c_0, c_1, ..., c_n tels que :
$y = c_0 + c_1 x_1 + c_2 x_2 + ... + c_n x_n + u$.
y est en moyenne égal à :
$c_0 + c_1 x_1 + c_2 x_2 + ... + c_n x_n$, mais la dépendance n'est pas parfaite. C'est pourquoi l'on ajoute une perturbation u que l'on cherche à minimiser (voir encadré).
L'assureur évalue ainsi les valeurs de c_0, c_1, ..., c_n en fonction des données de

l'année ou des années passées pour lesquelles il dispose des variables relatives aux assurés (x_i) et aux coûts des sinistres (y) qu'ils ont occasionnés. S'il démarre son activité, il achètera une base de données à un autre assureur, ou se basera sur des enquêtes réalisées sur des échantillons représentatifs de la population qu'il cherche à assurer. Quoiqu'il en soit, ce n'est qu'après quelques années d'activité qu'il disposera de données en quantité suffisante pour proposer une tarification robuste.
Lorsqu'une personne souhaite faire assurer son véhicule, elle communique à l'assureur l'ensemble des caractéristiques x_i qu'il lui demande. L'assureur calcule alors le coût associé,
$c_0 + c_1 x_1 + c_2 x_2 + ... + c_n x_n$
et propose un tarif.

Le travail de l'actuaire

Nous avons ici pris l'exemple, le plus simple, de la régression linéaire. Pourtant, ce modèle n'est pas forcément le meilleur. On peut "forcer" le lien linéaire en transformant les différentes variables, par exemple en prenant leur logarithme ou leur racine carrée. Il existe beaucoup d'autres types de régressions qui n'imposent pas de manière aussi simpliste ce lien linéaire, et qui sont donc plus réalistes. Ce sont ces modèles de régression plus complexes qu'utilisent les assureurs. Ce travail d'analyse des risques et de leurs coûts constitue le métier de l'actuaire. Il utilise le maximum d'informations contenues dans les données statistiques dont il dispose pour prévoir *a priori* le comportement de risque des conducteurs, et établit les critères techniques de la tarification.
Aujourd'hui, le risque automobile est très connu, bien appréhendé. Le travail statistique de l'actuaire se concentre plus sur l'étude d'autres types de contrats ou d'autres garanties aujourd'hui moins répandues, mais que les assureurs veulent développer.

S. M.

Niveau de difficulté

- ❍ très facile
- ✔ facile
- ✔✔ pas facile
- ✔✔✔ difficile
- ✔✔✔✔ très difficile

Déjouer les pièges des statistiques

HS3401 – Histoires de moyennes ✔

Dans un groupe constitué de dessinateurs et de journalistes, la moyenne d'âge est de 35 ans. **Sachant que les dessinateurs ont 30 ans en moyenne et les journalistes 48 ans en moyenne, quel est le rapport entre le nombre de dessinateurs et celui des journalistes ?**

HS3402 – Les anniversaires ✔✔

Je suis invité à une soirée chez une amie mathématicienne. Lorsque j'arrive chez elle, celle-ci m'accueille par ces mots :
« Grâce à votre arrivée, nous sommes maintenant suffisamment nombreux pour que la probabilité qu'au moins deux personnes présentes aient la même date d'anniversaire soit supérieure à 1/2. Je néglige bien entendu la possibilité que quelqu'un soit né un 29 février. ».
Combien de personnes sont-elles présentes, au total ?
Quelle est la probabilité pour qu'au moins une des personnes présentes ait la même date d'anniversaire que moi ?

HS3403 – Le sondage ✔

Lors d'un sondage sur la sécurité routière, 90 % des personnes interrogées ont répondu qu'elles pensent conduire mieux que la moyenne des conducteurs, et les 10 % restants qu'elles pensent conduire moins bien que la moyenne des conducteurs.
Se peut-il que leurs appréciations soit toutes correctes ?

HS3404 – Les patrons sont sympas ✔

Votre patron, satisfait de vos services, vous propose une augmentation régulière de votre salaire. Mais il vous donne le choix entre :
* ou bien une augmentation de 1000 euros chaque année sur votre salaire annuel,
* ou bien une augmentation de 300 euros tous les six mois sur votre salaire de six mois.

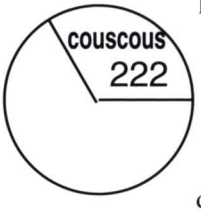
Que choisissez-vous ?

HS3405 – Statistiques ✔✔

Après les demi-finales du Championnat des jeux mathématiques et logiques, lors d'une étude statistique portant sur 1000 bulletin-réponse provenant exclusivement des 3 catégories « Haute compétition », « Grand Public » et « Lycéens », Alain, qui a étudié les 1000 bulletins, signale à Alex, afin qu'il effectue les calculs, qu'à la question 14 :
il y a exactement 10 % des « Haute compétition » qui ont répondu faux ;
exactement 50 % des « Lycéens » qui ont répondu faux ;
et exactement 40 % des « Grand Public » qui ont répondu faux.
Alex fait les calculs et trouve le nombre de bulletins parmi les 1000 qui ont faux à la question 14. Par souci de vérification, Alain fait lui aussi les calculs, et il trouve un résultat double de celui d'Alex avant de s'apercevoir qu'il a interverti les pourcentages des catégories « Haute compétition » et « Grand Public ».
Combien y avait-il de bulletins de la catégorie « Lycéens » ?

HS3406 – Les trois diagrammes ✔✔

Thomas est directeur d'une chaîne de restauration rapide qui propose trois plats tous les jours : un couscous, un poisson et un plat végétarien. Chacun des restaurants vient d'envoyer un diagramme circulaire donnant la répartition des ventes des trois plats proposés. Étrangement, les trois diagrammes comportent tous un secteur angulaire de 120°, et pour chacun des trois restaurants, on peut lire : 222 couscous et 114 poissons. Pourtant les nombres de plats végétariens vendus sont tous différents.



Combien, à eux trois, ces restaurants ont-ils vendu de plats végétariens ?

HS3407 – Le contrôle ✔✔

Un contrôle a eu lieu dans une classe. On sait qu'au moins les deux tiers des questions de ce contrôle étaient difficiles : pour chacune de ces questions difficiles, au moins les deux tiers des élèves n'ont pas su répondre. On sait aussi qu'au moins les deux tiers des élèves ont bien réussi le contrôle : chacun d'eux a su répondre à au moins deux tiers des questions.

a) Est-ce possible ?

b) La réponse à la question précédente serait-elle la même si l'on remplaçait partout deux tiers par trois quarts ?

c) La réponse à la première question serait-elle la même si l'on remplaçait partout deux tiers par sept dixièmes ?

HS3408 - Biodiversité ✔✔

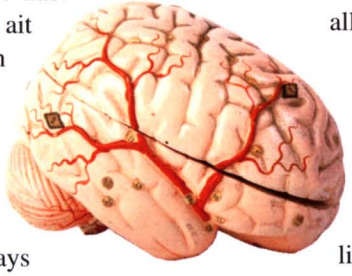
Toutes les espèces de plantes en Russie ont été numérotées par les nombres de 2 à 20 000 (chaque nombre est utilisé une et une seule fois). Pour chaque paire d'espèces différentes, le plus grand commun diviseur de leurs numéros a été retenu, tandis que les numéros eux-mêmes ont été perdus (suite à une défaillance dans l'ordinateur).

Peut-on rétablir les numéros de toutes les espèces ?

HS3409 - Q.I. contre Q ✔✔✔

Un groupe de psychologues a élaboré un test qui attribue à chaque personne un nombre Q qui mesure ses capacités intellectuelles (plus Q est grand, plus les capacités sont élevées).

Supposons que chacun des habitants de deux pays A et B ait obtenu son nombre Q. On prend alors pour le niveau intellectuel de chaque pays la moyenne arithmétique des nombres Q de ses habitants.



Un groupe d'habitants du pays A a émigré dans le pays B.

a) Est-il possible que le niveau intellectuel des deux pays ait augmenté ?

Après cela, un groupe d'habitants du pays B (parmi lesquels il peut y avoir des anciens émigrés de A), émigre dans le pays A.

b) Est-il possible que le niveau intellectuel des deux pays augmente de nouveau ?

c) Un groupe d'habitants a émigré du pays A dans le pays B et un autre groupe du pays B dans le pays C. Cela a fait augmenter le niveau intellectuel de chacun des trois pays. Ensuite les flots migratoires ont changé de direction : un groupe d'habitants a émigré de C dans B et un autre de B dans A. Les agences d'information des trois pays affirment que le niveau intellectuel de chaque pays a augmenté encore plus après cette deuxième migration.

Est-ce que c'est possible (si oui, comment, si non, pourquoi) ?

On suppose qu'entre les migrations le quotient intellectuel Q de chaque personne ne change pas et que personne ne meurt et personne ne naît.

HS3410 - L'ascenseur ✔✔✔

Aristide habite un immeuble de plus de 3 étages, de moins de 25 étages, sans sous-sol, et possédant un unique ascenseur. On suppose que les allées et venues sont telles que l'appareil, lorsqu'il est à l'arrêt, a une chance sur deux d'être au rez-de-chaussée, et des probabilités égales d'être au premier étage, au deuxième étage, au troisième, ...

Lorsqu'Aristide sort de son appartement, et qu'il appelle l'ascenseur, alors que celui-ci est à l'arrêt, l'appareil parcourt en moyenne exactement deux fois plus de distance que lorsqu'on l'appelle du rez-de-chaussée ou du premier.

À quel étage Aristide habite-t-il ?

HS3401 - Désignons par d le nombre de dessinateurs et par j le nombre de journalistes. On a $30d + 48j = 35(d + j)$, d'où $13j = 5d$. On en déduit que $d / j =$ **13 / 5**.

HS3402 - Il est plus simple de calculer la probabilité pour que n personnes aient des dates d'anniversaire toutes différentes. Cette probabilité est égale à $\frac{A^n_{365}}{365^n}$.

On observe que cette probabilité, qui vaut 1 pour n = 1, décroît quand n augmente, et devient inférieure à 1/2 lorsque n atteint 23. Si 23 personnes sont présentes, la probabilité pour qu'au moins deux d'entre elles aient la même date d'anniversaire est égale à environ 0,507.
La probabilité pour qu'aucune des 22 autres personnes présentes n'ait la même date anniversaire que vous est égale à $(364/365)^{22}$, et celle pour qu'au moins une des personnes présentes ait la même date anniversaire que vous à $1 - (364/365)^{22}$, soit environ **0,0586**.

HS3403 - Contrairement à ce que l'intuition peut suggérer, **il est possible** (même si c'est en réalité peu probable), **que toutes les personnes interrogées aient répondu correctement**. En effet, supposons que la "bonne conduite" puisse être évaluée de 0 à 200. Supposons ensuite que sur 100 personnes interrogées, 90 personnes aient un "niveau de bonne conduite" égal à 101, et les dix personnes restantes un niveau égal à 91. Le niveau moyen est alors égal à 100 et il est possible que chacun ait répondu correctement. On ne pourrait évidemment pas remplacer le mot "moyenne" par le mot "médiane".

HS3404 - Soit S votre salaire annuel actuel. Examinons les deux cas de figure. Il est évident que **vous serez gagnant en choisissant la seconde proposition**.

	1re proposition	2e proposition
dans 6 mois		S / 2 + 300
dans 1 an	S + 1000	S / 2 + 600
dans 1 an et 6 mois		S / 2 + 900
dans 2 ans	S + 2000	S / 2 + 1 200
dans 2 ans et 6 mois		S / 2 + 1 500
dans 3 ans	S + 3000	S / 2 + 1 800
dans 3 ans et 6 mois		S / 2 + 2100
dans 4 ans	S + 4000	S / 2 + 2400

HS3405 - Désignons par h le nombre de candidats de la catégorie « Haute Compétition », par g celui des candidats de la catégorie « Grand Public », et par L le nombre des candidats de la catégorie « Lycéens ». Les données de l'énoncé conduisent au système d'équations :

$$\left\{\begin{array}{l} 2(0{,}1h + 0{,}5\mathrm{L} + 0{,}4g) = 0{,}4h + 0{,}5\mathrm{L} + 0{,}1g \\ h + \mathrm{L} + g = 1000 \end{array}\right\}$$

Ce système de deux équations à 3 inconnues conduit à l'équation : $7\,\mathrm{L} + 9\,g = 2000$, qu'il faut résoudre en nombres entiers positifs. Le couple (4 ; – 3) est une solution de $7\,\mathrm{L} + 9\,g = 1$, donc (8000 ; – 6000) est une solution de $7\,\mathrm{L} + 9\,b = 2000$. Les solutions seront de la forme $(8000 - 9k ; - 6000 + 7k)$, avec $858 < k < 888$, L et g étant tous deux positifs. Les valeurs de g obtenues sont 6, 13, 20, 27, 34, ...
Mais 0,4 g doit être un entier, ce qui restreint les possibilités à 20 (k = 860), 55 (k = 865), 90 (k = 870), 125 (k = 875), 160 (k = 880), 195 (k = 885). Par ailleurs, 0,5 L étant entier, L doit être pair, ce qui élimine les valeurs 865, 875 et 885 pour k.
Il ne reste donc que les trois possibilités k = 960, 870 ou 880, qui donnent g = 20, 90 ou 160, et L = 260, 170 ou 80.

Il y a donc 3 solutions : le nombre de bulletins de la catégorie « Lycéens » est égal **à 80, 170 ou 260**.

HS3406 - Si les trois restaurants ont vendu le même nombre de couscous et le même nombre de poissons et que pourtant, les trois diagrammes sont différents, c'est que les trois nombres de plats végétariens sont différents. On peut donc en déduire que dans un restaurant, les couscous représentent le tiers des ventes, dans un autre, ce sont les poissons et dans le troisième les plats végétariens. Ces considérations conduisent au tableau suivant.

couscous	poisson	plat v g t.	TOTAL
222	114	330	3 × 222 = 666
222	**114**	6	3 × 114 = 342
222	114	**168**	3 × 168 = 504

Le nombre de plats végétariens vendus dans les trois restaurants est égal à 330 + 6 + 168, soit **504**.

HS3407 - a) Il y a au moins 4/9 de réponses fausses (ou de non-réponses) et au moins 4/9 de réponses correctes, ce qui est parfaitement **possible**. Montrons-le sur un exemple où la classe est constituée de 3 étudiants et ou le test comporte 3 questions.

	Question 1	Question 2	Question 3
Etudiant 1	juste	fausse	juste
Etudiant 2	fausse	juste	juste
Etudiant 3	fausse	fausse	juste ou fausse

Les questions 1 et 2 étaient difficiles et les étudiants 1 et 2 ont réussi le contrôle.

b) Il devrait y avoir au moins 9/16 de réponses fausses (ou de non-réponses) et au moins 9/16 de réponses correctes, ce qui est **impossible**.

c) Désignons par E le nombre d'étudiants ayant réussi au moins 7/10 des questions, par D le nombre de questions difficiles, par J le nombre de réponses justes aux questions difficiles obtenues par les étudiants ayant réussi et par F le nombre de réponses fausses aux questions difficiles obtenues par les étudiants ayant réussi. On doit avoir : $J \geq E \times 4 / 7\ D$ (un étudiant ne peut avoir réussi le test s'il a répondu à moins de 4 / 7 des questions difficiles). On doit également avoir $F = D \times 4/7\ E$ (une question difficile doit avoir été ratée par au moins 4 / 7 des étudiants ayant réussi).
Or $J + F = ED \geq 8 / 7\ ED$, **ce qui est impossible**.

Le nombre de plats végétariens vendus dans les trois restaurants est égal à 330 + 6 + 168, soit **504**.

HS3408 - On ne peut pas retrouver tous les numéros des plantes : par exemple, les deux nombres $8\,192 = 2^{13}$ et $16\,384 = 2^{14}$ ne sont pas différenciables à partir des plus grands communs diviseurs. En effet, tous les autres nombres ont comme plus grand commun diviseur avec ces deux là la plus grande puissance de 2 les divisant (car à part 16384, il n'y a pas de multiple de 2^{14} inférieur à 20 000). Par ailleurs, ce n'est pas leur plus grand commun diviseur qui va les départager.

Remarquons que c'est aussi vrai avec *tous* les nombres premiers entre 10 000 et 20 000, mais pour que la réponse soit exacte, encore faut-il démontrer qu'il existe deux nombres premiers entre 10 000 et 20 000.

HS3409 - a) Oui, c'est possible, par exemple si tous les habitants de A ont un nombre Q plus élevé que tous les habitants de B et que des habitants de A ayant les nombres Q les moins élevés émigrent de A vers B.

b) Non, ce n'est pas possible. Si a et b désignent les niveaux intellectuels respectifs des pays A et B, pour que le niveau des deux pays augmente lorsqu'un groupe émigre de A vers B, il faut que la moyenne g des niveaux des émigrés vérifie $a > g > b$. Si a' et b' sont les niveaux des deux pays après émigration et g' le niveau moyen du groupe émigrant de B vers A, on a encore $a' > g' > b'$. Le niveau des deux pays ne peut donc pas augmenter à nouveau.

c) Oui, c'est possible. Il faut pour cela que A et C aient des niveaux plutôt élevés et B un niveau plutôt bas, mais avec deux génies dans sa population.

Exemple

	pays A	pays B	pays C
état initial	50 personnes de Q = 1 50 personnes de Q = 2	98 personnes de Q = 0 2 personnes de Q = 5	50 personnes de Q = 1 50 personnes de Q = 2
après la 1e migration	50 personnes de Q = 2	98 personnes de Q = 0 50 personnes de Q = 1 1 personne de Q = 5	50 personnes de Q = 1 50 personnes de Q = 2 1 personne de Q = 5
après la 2e migration	50 personnes de Q = 2 1 personne de Q = 5	98 personnes de Q = 0 100 personnes de Q = 1	50 personnes de Q = 2 1 personne de Q = 5

HS3410 - Désignons par X le numéro de l'étage d'Aristide, et par n le nombre d'étages de son immeuble. De l'énoncé résultent les deux inégalités suivantes : $2 = X = n$ et $3 = n = 25$, et la double égalité :

distance moyenne pour le rez-de-chaussée = distance moyenne pour le premier étage = distance moyenne pour le X^e étage

$$\frac{1}{2}\times 0+\frac{1+2+3+4+5+\ldots+n}{2n}=\frac{1}{2}\times 1+\frac{0+1+2+3+4+\ldots+n-1}{2n}$$

$$=\frac{1}{2}\left(\frac{1}{2}X+\frac{(X-1)+(X-2)+\ldots+1+0+1+2+\ldots+(n-1)+(n-X)}{2n}\right)$$

soit $$\frac{n(n+1)}{4n}=\frac{1}{2}+\frac{(n-1)n}{4n}=\frac{1}{4}X+\frac{X(X-1)}{8n}+\frac{(n-X)(n-X+1)}{8n}$$

d'où $$n(n+1)=nX+\frac{X(X-1)+(n-X)(n-1+X)}{2}$$

qui aboutit à $n(n+1) = 2X(X-1)$.

Le problème se ramène donc à trouver un produit de deux nombres consécutifs qui soit le double d'un autre produit du même type.

En examinant les produits de ce type pour n compris entre 3 et 24, on trouve $20(20+1) = 2 \times 15(15-1)$

Aristide habite donc au **15^e étage**.

Tangente

Publié par Les Éditions POLE
SAS au capital de 40 000 euros
Siège social : 80 bd Saint-Michel – 75006 Paris
Commission paritaire : 1006 K 80883
Dépôt légal à parution

Directeur de Publication et de la Rédaction
Gilles COHEN

Rédacteur en chef de ce numéro
Philippe BOULANGER

Secrétaire de rédaction
Édouard THOMAS

Comité de rédaction
Stella BARUK, André BELLAÏCHE, Élisabeth BUSSER, Francis CASIRO, Michel CRITON, Nicolas DELERUE, Jean-Jacques DUPAS, Denis GUEDJ, Bertrand HAUCHECORNE, François LAVALLOU, Hervé LEHNING, Alexandre MOATTI, Marie-José PESTEL, Daniel TEMAM, Norbert VERDIER, Alain ZALMANSKI, Chérif ZANANIRI

Autres auteurs d'articles
T. DE LA RUE, Gilles DUPIN, P. DOSSANTOS-DEZARRALDE, N. GAUVRIT, Valérie HENRY, H.-P. JACQUET, E. JANVRESSE, Daniel JUSTENS, J.-P. KRIVINE, Hervé LE BRAS, Jean LEFORT, Jacques LUBCZANSKI, Sylvain MERLUS, Gaël OCTAVIA, Benoît RITTAUD, John WESSON

Publicité au journal
Tél. : 01 47 07 99 10
pub@poleditions.com

Abonnements
Tél. : 01 47 07 51 15 – Fax : 01 47 07 88 13

Maquette
Claude LUCCHINI
Photos : droits réservés
Ce numéro Hors Série de Tangente a été imprimé par Louis Jean, 05000 GAP.

codif : POLE HS34

Bulletin d'abonnement à retourner à :

Espace Tangente - 80, bld Saint-Michel - 75006 PARIS

Nom ..... Prénom .....

Établissement .....

Adresse .....

Code Postal ..... Ville .....

Profession ..... E-mail .....

Oui, je m'abonne à	FRANCE		HORS MÉTROPOLE
	1 AN	2 ANS	
TANGENTE	☐ 32 €	☐ 60 €	☐ + 12 € par an
TANGENTE PLUS	☐ 52 €	☐ 100 €	☐ + 20 € par an
TANGENTE SUPERPLUS	☐ 82 €	☐ 160 €	☐ + 24 € par an
TANGENTE SUP	☐ 24 €	☐ 46 €	☐ + 6 € par an
TANGENTE EDUCATION	☐ 10 €	☐ 18 €	☐ + 2 €

❒ À partir du numéro en cours ❒ À partir du numéro .....

Total à payer

Je joins mon paiement par (établissements scolaires, joindre bon de commande administratif) :

❒ Chèque (uniquement **payable en France**)

❒ Carte (à partir de 30 €) numéro : |_|_|_|_| |_|_|_|_| |_|_|_|_| |_|_|_|_|

Date et Signature : Expiration le : ...../.....

Achevé d'imprimer pour le compte des Éditions POLE
sur les presses de l'imprimerie Louis Jean, 05000 Gap
Imprimé en France - Dépôt légal 24 - Janvier 2009

Tangente

Les Statistiques
et leur décodage

- Croyances et erreurs
- Recueil des données
- Traitement des données
- Interpréter les statistiques
- Le choix collectif

Coordination : Philippe Boulanger

Les statistiques sont omniprésentes dans notre vie quotidienne : résultats sportifs, recensements, quotas, loi des séries, sondages, élections... Elles font usage d'un grand nombre de données, qu'il faut recueillir, classer, traiter, interpréter. Avec elles, les mathématiques réalisent un va-et-vient continu entre la théorie et la confrontation avec la réalité. Sujettes à polémique depuis leurs premières utilisations, les statistiques continuent néanmoins à être invoquées, telles des oracles, pour expliquer ou prédire les phénomènes. En clair, elles fascinent toujours, malgré les erreurs dues aux inévitables incertitudes ou à leur mauvaise utilisation. Elles restent au centre de nombreuses recherches qui devraient permettre de les utiliser à meilleur escient.

9 782848 840963

Diffusion : S391631
Prix : 18 €

EDITIONS POLE