Résumé

L'objet de ce cours est de présenter quelques éléments importants de la démarche statistique, afin notamment de permettre un abord raisonné et critique des arguments statistiques rencontrés dans divers contextes.

Bréviaire de réflexion statistique

Jean Bérard

14 décembre 2005

1 Avertissement

Ces notes sont en cours de rédaction!!!

2 Introduction

Pour des raisons diverses, l'intuition et le sens commun nous conduisent souvent à accepter, ou à formuler nous-mêmes, des arguments statistiques incorrects. De fait, analyser et interpréter des données en rapport avec des situations réelles est une tâche difficile, que l'esprit humain semble bien peu à même d'accomplir sans se laisser prendre à des pièges de natures diverses. Il importe pourtant d'apprendre à éviter ces pièges, que ce soit pour être en mesure de tirer des conclusions valables des données dont on dispose, ou pour éviter d'être manipulé ou induit en erreur par d'autres, qu'il s'agisse de tentatives de manipulation délibérée ou d'arguments erronnés défendus de bonne fois, ou de toute situation intermédiaire (par exemple, de la négligence volontaire dans la critique de ses propres arguments).

L'objectif de ce cours est de présenter quelques éléments importants de la démarche statistique, afin notamment de favoriser un abord raisonné et critique des arguments statistiques, quel que soit le contexte dans lequel ils interviennent (scientifique, technique, industriel, civique, personnel,...).

Certains chapitres présentent des idées ou des techniques propres à la statistique, d'autres sont plutôt construits autour de certains types d'erreurs de raisonnement. Nous renvoyons la plupart du temps à des ouvrages spécialisés pour une description précise des techniques abordées, notre objectif étant surtout d'insister sur l'esprit dans lequel ces techniques doivent être employées, et sur les mises en garde qui les accompagnent. Une certaine familiarité avec les notions de base des probabilités et de la statistique est supposée.

La liste des sujets abordés ne doit en aucun cas être considérée comme exhaustive, ni même comme présentant les éléments les plus importants ou les plus fréquemment rencontrés. Ce cours ne présente qu'une sélection personnelle d'un certain nombre de points, parmi ceux que nous considérons comme les plus dignes d'intérêt.

3 Quelques généralités

De manière très générale, l'objet d'étude des probabilités et de la statistique est constitué par les situations et phénomènes faisant intervenir incertitude, va-

riabilité et hasard. Il existe d'importantes nuances sémantiques entre ces termes, qui possèdent eux-mêmes des acceptions variées : l'incertitude peut porter aussi bien sur la mesure d'une grandeur que sur la culpabilité d'un accusé, la variabilité peut se référer à un caractère biologique, – comme la couleur des yeux, ou la réponse à un traitement –, susceptible de varier au sein d'une population, ou encore à des phénomènes variables dans le temps,... et le hasard est, quant à lui, une notion difficile à définir, étant associé simultanément, au moins sous la forme où nous l'envisagerons, à l'idée de régularité statistique et d'imprévisibilité.

Voici un bref schéma censé représenter, de manière très grossière, les domaines respectifs des probabilités et de la statistique.

Probabilités

Formulation et étude des propriétés de modèles mathématiques faisant intervenir la notion de probabilité.

$\Updownarrow$ *Modélisation*

Statistique

$\left\{ \begin{array}{l} \text{Collecter des données (de l'interrogatoire à l'organisation et à la planification d'une enquête)} \\ \text{Décrire des données (graphiques, indicateurs divers et variés, histogrammes, boxplots, acp, ...)} \\ \text{Interpréter les données (pour prédire, expliquer, comprendre)} \end{array} \right.$

4 Des références

4.1 Des références générales

Chance News

How to lie with Statistics

`http://pareonline.net/getvn.asp?v=5&n=5 pitfalls of data analysis`

`http://www.fallacyfiles.org/texsharp.html`

`http://en.wikipedia.org/wiki/List_of_cognitive_biases`

Contradicted and initially stronger effects `http://jama.ama-assn.org/cgi/content/abstract/294/2/218`

why more than 50 pc of published results are wrong... ioannidis.pdf référence !

The great health hoax http ://faculty.samford.edu/ twwoolle/HHoax.htm

Nous renvoyons à un cours de base de probabilités et de statistique pour une introduction aux notions utilisées dans ce qui suit.

Des recouvrements existent entre les chapitres qui suivent, car les sujets ou les exemples qui y sont présentés peuvent souvent, et avec profit, être vus sous plusieurs angles.

5 La probabilité : une notion complexe

Contrairement à ce que l'usage extrêmement courant du mot pourrait laisser supposer, la notion de probabilité est difficile à définir, son acception peut varier considérablement suivant les situations que l'on étudie, et la compréhension de ses diverses interprétations pratiques est fondamentale si l'on veut pouvoir en faire un usage correct dans des situations réelles.

De manière générale, la probabilité peut aussi bien désigner un degré de plausibilité accordé à un fait passé ou futur (Quelle est la probabilité pour que le candidat A soit élu lors de la prochaine élection ? Quelle est la probabilité pour que Jules César ait prévu que l'on allait l'assassiner ?), une fréquence limite au cours d'expériences répétées (Quelle est la probabilité pour qu'une pièce de monnaie retombe sur pile ?), une proportion au sein d'une population (Quelle est la probabilité pour un individu né en France en 1962 de posséder un chien en 2005 ?), ou, la plupart du temps, un mélange plus ou moins clair de ces interprétations – cette possibilité de mélanger les interprétations faisant à la fois la richesse et la difficulté de cette notion. La perception psychologique de la probabilité par les êtres humains constitue encore une autre question, dont l'étude révèle une certaine incohérence avec les définitions scientifiquement acceptées de cette notion.

Du point de vue du formalisme mathématique, cette diversité n'est pas gênante car le formalisme et ses règles sont compatibles avec la plupart des interprétations raisonnables de la notion de probabilité. (On peut donc développer la théorie mathématique des probabilités en toute tranquillité sans se soucier de ce genre de question). En revanche, la portée pratique des raisonnements qui sont menés dans le cadre de ce formalisme est totalement conditionnée par le sens et la validité de l'interprétation concrètes des différentes probabilités qui y sont manipulées ; c'est pourquoi il est extrêmement important, dès qu'il est question de probabilité en pratique, de préciser exactement de quel type de probabilité il s'agit, comment on entend la définir exactement et quels sont les moyens devant permettre de l'estimer (un problème qui se pose très souvent est par exemple de définir une population pertinente pour estimer les probabilités, les proportions n'étant pas nécessairement stables d'une population à une autre). L'absence de clarté à ce stade est une première et abondante source de raisonnements et résultats incorrects, abus et sophismes en tous genres, incorrects non pas car ils sont en désaccord avec le formalisme mathématique, mais parce qu'ils ne sont pas pertinents d'un point de vue pratique.

Nous renvoyons à la discussion effectuée dans le document Introduction aux probabilités et à la Statistique pour (un peu) plus de détails, ou à l'ouvrage de Ian Hacking, The emergence of probability, Cambridge University Press, pour une discussion historique détaillée, ou encore L'ouverture au probable, de Michel Dufour et Ian Hacking, Armand Colin.

6 Utilisation incorrecte des probabilités conditionnelles

L'une des sources d'erreurs de raisonnement que l'on rencontre fréquemment en pratique est l'utilisation incorrecte de la notion de probabilité conditionnelle. En effet, on n'éprouve pas, la plupart du temps, le besoin de poser explicitement en termes de probabilités conditionnelles les raisonnements que l'on effectue. Cependant, la notion est plus subtile qu'il n'y paraît, et l'on est donc souvent amené à des confusions parfois lourdes de conséquence.

Notre conseil : toujours poser explicitement le modèle de raisonnement que l'on utilise, et en particulier toujours préciser exactement (même si leur emploi ne semble qu'implicite) quelles sont les probabilités conditionnelles qui inter-

viennent.

6.1 Origine sociale et études supérieures

Si l'on s'intéresse à la sélection sociale dans les études supérieures, on pourra s'étonner que la probabilité pour un enfant d'ouvrier d'entrer à l'Ecole Polytechnique soit de moins de $0,1\%$ alors que les enfants d'ouvriers représentent plus de 20% d'une classe d'âge. Pourtant, on compare ici ce qui n'est pas comparable : il faudrait, pour se faire une idée du rôle joué par l'origine sociale dans la poursuite d'études prestigieuses, soit comparer la probabilité pour un enfant d'ouvrier d'entrer à l'Ecole Polytechnique à la probabilité pour un individu quelconque de devenir polytechnicien, soit la probabilité pour un polytechnicien d'avoir des parents ouvriers (actuellement, moins de 2%) au poids démographique des enfants d'ouvriers (plus de 20%). Dans un cas, on compare deux probabilités de l'ordre de quelques millièmes, dans l'autre cas, deux probabilités de l'ordre du dixième, et échanger les quantités que l'on doit comparer ne conduit qu'à une remarque vide de signification. Nous avons ici confondu $\mathbb{P}(A|B)$ et $\mathbb{P}(B|A)$, où l'événement A désignerait par exemple le fait pour un individu de devenir polytechnicien, et B celui d'être enfant d'ouvrier. La probabilité se réfère ici sans trop d'ambiguité à des proportions mesurées au sein de la population française.

Source :
`http://users.swing.be/aped/documents/d0055selsocfrance.html`

6.2 Près de soixante pour cent !

Petit exercice : près de soixante pour cent des accidents de voiture graves impliquant de jeunes enfants se produisent dans des véhicules où les enfants ne sont pas correctement attachés (source : la brochure d'information de ma mutuelle). Soixante pour cent, cela fait beaucoup... A quoi faudrait-il comparer ce chiffre ?

Attention, nous n'affirmons nullement que le fait d'attacher correctement les enfants n'est pas utile dans la prévention des conséquences des accidents !

6.3 Comparaisons

Même en comparant des quantités qu'il fait sens de comparer, par exemple $\mathbb{P}(A|B)$ et $\mathbb{P}(A)$, les précautions d'usage lorsque l'on interprète les résulats de comparaison (voir le section **??**) s'impose : le fait que $\mathbb{P}(A|B)$ soit, de manière bien établie, supérieure à $\mathbb{P}(A)$ ne signifie pas qu'elle lui soit très supérieure ; le fait que $\mathbb{P}(A|B)$ soit très supérieure à $\mathbb{P}(A)$ ne signifie pas néanmoins que $\mathbb{P}(A)$ soit élevée (c'est-à-dire proche de 1), etc...

6.4 Le sophisme du procureur

En général, on parle de sophisme du procureur lorsque la faible probabilité des circonstances constatées sous l'hypothèse qu'une personne accusée est innocente, est présentée comme un indice de sa culpabilité. Nous en verrons plusieurs exemples bien réels. Commençons par un exemple fictif. Sur la foi d'un test ADN, M. D*** comparaît devant un tribunal dans le cadre d'une affaire

criminelle, et l'expert invité à la barre explique que, à supposer que M. D*** soit innocent, la probabilité pour que son ADN coïncide avec celui trouvé sur les lieux du crime d'après le test effectué est d'environ 1/10000. Doit-on en déduire qu'il y ait moins d'une chance sur 10000 pour que M. D*** soit innocent ?

Procédons avec ordre et méthode ! Nous envisageons dans ce raisonnement plusieurs possibilités : que M. D*** soit coupable ou innocent d'une part, et, d'autre part, que le test ADN effectué réponde positivement ou non (puisque l'expert fait allusion à la probabilité pour que le test réponde positivement sous l'hypothèse que M. D*** soit innocent). Appelons T l'événement correspondant au fait que le test ADN donne un résultat positif, I l'événement correspondant à l'innocence de M. D***, C l'événement correspondant à sa culpabilité. Dans ce contexte, le témoignage de l'expert revient à proposer d'estimer à 1/10000 la probabilité $\mathbb{P}(T|I)$. Mais ce qui nous intéresse est d'estimer la probabilité pour que M. D*** soit innocent (ou coupable, l'une se déduisant immédiatement de l'autre) sachant que le test ADN a donné un résultat positif, autrement dit la probabilité $\mathbb{P}(I|T)$, et non pas la probabilité $\mathbb{P}(T|I)$, pour laquelle l'expert propose une estimation. Ces deux probabilités sont, bien entendu, liées, d'après la la formule de Bayes :

$$\mathbb{P}(I|T) = \frac{\mathbb{P}(T|I) \times \mathbb{P}(I)}{\mathbb{P}(T|C) \times \mathbb{P}(C) + \mathbb{P}(T|I) \times \mathbb{P}(I)}.$$

Par conséquent, et malgré les apparences, on ne peut rien conclure de l'avis d'expert sans avoir auparavant tenté d'estimer les différentes probabilités qui interviennent dans la formule ci-dessus, savoir $\mathbb{P}(I)$, $\mathbb{P}(C)$ (qui se déduit immédiatement de $\mathbb{P}(I)$) et $\mathbb{P}(T|C)$, la valeur de $\mathbb{P}(T|I)$ étant fournie. Ce qui compte, dans l'évaluation de l'expression ci-dessus, est le rapport

$$\frac{\mathbb{P}(T|C) \times \mathbb{P}(C)}{\mathbb{P}(T|I) \times \mathbb{P}(I}.$$

Si celui-ci est faible, la probabilité pour que M. D*** soit innocent au vu du résultat du test est voisine de 1. Au contraire, si celui-ci est élevé, cette probabilité est voisine de 0. Ce rapport est celui des probabilités des deux explications possibles des circonstances observées : la culpabilité de M. D***, et la coïncidence fortuite alors que celui-ci est innocent. Raisonner sur $\mathbb{P}(T|I)$ seul, même si celle-ci possède une valeur très faible, et conduit donc à admettre que les coïncidences fortuites sont rares, n'a aucun sens : c'est le rapport des probabilités des deux explications concurrentes qui doit être examiné.

Comment évaluer les probabilités $\mathbb{P}(I)$, et $\mathbb{P}(T|C)$ dans ce contexte ? Il s'agit ici clairement de probabilités mesurant des degrés de plausibilité dans un raisonnement en situation d'incertitude. La probabilité $\mathbb{P}(I)$ mesure donc la probabilité *a priori* pour que M. D*** soit coupable, c'est-à-dire, sans tenir compte de l'information selon laquelle le test ADN s'est révélé positif. Si nous ne disposons d'aucune information supplémentaire sur le crime ou sur M. D*** qui nous permette d'affiner notre estimation, il semble raisonnable de poser, dans ce contexte, que $\mathbb{P}(C) = 1/N$, et donc que $\mathbb{P}(I) = 1 - 1/N$, où N désigne le nombre total d'individus dans la population dont est issu M. D*** (et dans laquelle le coupable doit se trouver). Autrement dit, nous attribuons *a priori* à M. D*** la probabilité pour qu'une personne choisie uniformément au hasard dans la population soit coupable. Pour simplifier, nous pouvons supposer que

$\mathbb{P}(T|C)$ est approximativement égal à un, ce qui signifie simplement que le test ADN pratiqué sur le coupable doit presqu'à coup sûr donner un résultat positif.

Nous obtenons donc que

$$\mathbb{P}(I|T) = \frac{\mathbb{P}(T|I) \times (1 - 1/N)}{1/N + \mathbb{P}(T|I) \times (1 - 1/N)}.$$

Ainsi, si $\mathbb{P}(T|I)$ est grand devant $1/N$, la probabilité pour que M. D*** soit coupable est évaluée à une faible valeur, contrairement à ce qu'une impression rapide laisse penser en entendant le chiffre d'une chance sur 10000, qui semble accréditer la culpabilité de M. D***.

Intuitivement, et en admettant donc l'estimation selon laquelle $\mathbb{P}(T|I) = 1/10000$, on s'attend à observer, dans une population de, disons, 100000 individus, de l'ordre d'une dizaine (certainement pas exactement une dizaine, voir la section **??** sur l'oubli de variabilité, pour une discussion de ce point) de personnes répondant positivement au test et n'ayant *a priori* (c'est-à-dire ici en ne tenant compte que du résultat du test ADN) ni plus, ni moins de raisons que M. D*** d'être soupçonnés, d'où une probabilité de l'ordre du dizième pour la culpabilité de M. D*** (attention, il est important dans ce raisonnement que le test ADN constitue la seule raison pour laquelle M. D*** est incriminé parmi les individus de la population en question). Même si une coïncidence fortuite est rare, elle peut l'être suffisament peu pour se produire au sein d'une population importante. Elle ne constitue alors, à elle seule, certainement pas un argument contre une personne caractérisée par cette coïncidence ! Pour raisonner correctement dans ce contexte, et voir à quel point la rareté d'une coïncidence fortuite peut affecter la probabilité de culpabilité, il faut effectuer complètement le raisonnement présenté ci-dessus.

En revanche, si $1/N$ est nettement plus grand que $\mathbb{P}(T|I)$, le raisonnement accrédite plutôt la culpabilité de M. D***, ce qui ne signifie en aucun cas que sa culpabilité soit le moins du monde prouvée.

Exercice 1 *Une question importante est de savoir dans quelle mesure on peut se fier à l'évaluation de 1/10000 proposée par l'expert. Il est clair que de nombreux facteurs peuvent jouer : la fiabilité des méthodes employées pour recueillir et traiter l'ADN sur les lieux du crime, l'état de conservation de celui-ci, la façon dont l'ADN de M. D*** lui-même est recueilli et traité, la diversité génétique au sein de la population considérée... Comment vous y prendriez-vous pour évaluer la probabilité $\mathbb{P}(T|I)$? Votre méthode se réfère-t-elle a la population vivant aux environs du crime, à la population française dans son ensemble, à d'autres populations d'individus ? Quelle pertinence accordez-vous à votre méthode, et quelle fiabilité accorderiez-vous au résultat obtenu en l'utilisant ? Comment des informations spécifiques concernant le génôme de M. D*** ou la structure généalogique de la population peuvent-elles intervenir dans votre évaluation ?*

Exercice 2 *(Variation des estimations)*
1) A quel point les variations dans l'estimation de la probabilité $\mathbb{P}(T|I)$ peuvent-elles affecter l'évaluation valeur de la probabilité $\mathbb{P}(C|T)$? Dans quelle mesure une estimation grossière de $\mathbb{P}(T|I)$ peut-elle s'avérer suffisante ?
2) Deux experts différents proposent deux estimations différentes de $\mathbb{P}(T|I)$, disons p_1 et p_2, obtenues par deux méthodes différentes. Les propositions suivantes vous semblent-elles raisonnables ? Pour quelles raisons ?

- *Evaluer* $\mathbb{P}(C|T)$ *en prenant* $\mathbb{P}(T|I) = \frac{p_1+p_2}{2}$.
- *Evaluer séparément* $\mathbb{P}(C|T)$ *en prenant* $\mathbb{P}(T|I) = p_1$ *puis* $\mathbb{P}(T|I) = p_2$, *et conserver la plus petite des deux valeurs obtenues.*
- *Réexaminer les deux méthodes employées, et ne conserver que la valeur obtenue par la méthode qui semble la plus pertinente.*
- *Multiplier par 10 la plus grande des deux valeurs* p_1 *ou* p_2, *multiplier par 1/10 la plus petite, calculer les estimations de* $\mathbb{P}(C|T)$ *ainsi obtenues, et considérer qu'une valeur raisonnable doit se trouver dans la fourchette ainsi obtenue.*
- *Décider que si les résultats obtenus à partir des deux méthodes pointent dans la même direction (culpabilité ou innocence), on se satisfait de ce résultat.*
- *Décider que si les résultats obtenus à partir des deux méthodes pointent dans deux directions différentes, on ne peut rien dire.*
- *Analyser les deux méthodes employées et tenter de trouver une troisième méthode qui puisse remédier à leurs défauts potentiels avant de faire quoique ce soit.*
- *Essayer d'estimer, pour chaque méthode, les marges d'erreurs susceptibles d'affecter leurs résultats, et raisonner avec des fourchettes de valeur (comment ?) plutôt qu'avec des valeurs fixées.*

*3) Vous faites partie du jury chargé de statuer sur le sort de M. D***. Êtes-vous plus impressionné par une valeur, tous calculs faits, de la probabilité de culpabilité de 0,9998, que par une valeur de 0,9 ou 0,8 ? Décideriez-vous de déclarer M. D*** coupable en fonction de ce seul calcul (sachant que vous n'êtes censé le faire que lorsque sa culpabilité semble établie au-delà de tout doute raisonnable) ? Si oui, jusqu'à quelle valeur de la probabilité de culpabilité vous décidez-vous pour la culpabilité ? 0,99 ? 0,9 ? 0,8 ? 0,55 ? 0,5000001 ? Comment jugez-vous de la fiabilité de l'estimation proposée ? Une estimation à 0,99999 par une méthode qui semble douteuse vous convainc-t-elle davantage qu'une estimation de 0,8 par une méthode qui semble plus fiable ?*

Exercice 3 *M. D***, pour expliquer la similarité observée entre son propre ADN et les traces trouvées sur les lieux du crime, prétend avoir été victime d'une machination, des échantillons de ses propres tissus ayant été récupérés sur lui à son insu, puis déposés sur place, par le meurtrier ou l'un de ses complices, dans le but de le faire accuser à tort. Cette machiavélique possibilité a-t-elle été prise en compte dans les évaluations de probabilité ci-dessus ? Si oui, comment, et sinon, comment pourrait-elle l'être ? Même question avec la possibilité pour que de l'ADN de M. D*** se trouve par hasard sur les lieux du crime (c'est-a-dire, sans que celui-ci soit coupable) ?*

Pour citer la cour supprême de Californie s'exprimant au sujet de l'affaire Collins (voir plus bas), les mathématiques, «while assisting the trier of fact in the search of truth, must not cast a spell over him.»

Ajoutons que la rigueur du formalisme utilisé ne doit pas masquer la possible fragilité de l'interprétation concrète et de l'estimation des quantités qui y sont manipulées. S'appuyer sur la rigueur du formalisme pour prétendre que la même rigueur caractérise les conclusions obtenues est, dans ce contexte, encore un exemple de sophisme.

6.5 L'affaire du testament Howland

Lorsque la riche Mme Sylvia Howland mourut en 1865, il apparut que son testament, daté de 1863, stipulait qu'environ la moitié de sa fortune devait être répartie entre des légataires variés, tandis que l'autre moitié (soit plus d'un million de dollars de l'époque) serait placée, et les intérêts ainsi produits versés à sa nièce, Mme Henrietta Howland Green, à la mort de laquelle le principal serait redistribué entre d'autres légataires.

Mme Howland Green, qui comptait bien hériter de la totalité de la somme, et non pas seulement des intérêts, produisit alors un exemplaire plus ancien du testament, daté de 1862 (donc antérieur à celui effectivement exécuté lors de la succession), qui lui attribuait la quasi-totalité des biens de sa tante Sylvia, accompagné d'une page supplémentaire censée annuler «tout testament rédigé avant ou après celui-ci.» Si l'authenticité du testament de 1862 ne semblait pas devoir être mise en doute (il avait été signé de la défunte Mme Howland et de trois témoins), celle de la page supplémentaire était plus suspecte, celle-ci ne portant la signature que de la défunte et de sa nièce. L'exécuteur testamentaire de Mme Howland refusant d'accorder foi à la seconde partie du document, l'affaire fut portée devant les tribunaux, et plusieurs experts furent convoqués.

Un examen attentif[1]

d'un échantillon de 42 signatures réalisées par la défunte Mme Howland lors de ses dernières années fut mené. Celui-ci révéla d'une part, que chaque signature comportait systématiquement trente traits dirigés vers le bas, et, d'autre part, que entre deux signatures quelconques, en moyenne six traits dirigés vers le bas homologues (c'est-à-dire correspondant à un même élément d'une même lettre de la signature) étaient identiques.

En revanche, en comparant la signature présente sur le testament de 1862 avec celle figurant sur la page supplémentaire de celui-ci, ce fut une coïncidence complète des trente traits qui fut observée, suggérant la possibilité que la signature inscrite sur la page supplémentaire du testament ait été recopiée à partir de l'autre.

Les Peirce affirmèrent qu'au vu de leur étude, on pouvait évaluer la probabilité qu'une telle coïncidence survienne de manière accidentelle à $1/5^{30}$, soit, d'après les Peirce toujours, environ $1/2,666... \times 10^{-21}$, (ce qui est incorrect, la véritable valeur étant $1/9,313... \times 10^{-20}$). La conclusion était qu'une probabilité si faible indiquait que, selon toute raison, la page supplémentaire du testament était un faux.

Exercice 4 *(Questions diverses)*
1) Expliquez en quoi cet argument apparaît comme un (bel) exemple du sophisme du procureur. Quelles probabilités aurait-il également fallu évaluer pour tenter de conclure ? Dans quelles conditions pourrait-on néanmoins considérer que les probabilités mettent sérieusement en cause l'authenticité du document produit par Mme Howland Green ?

[1] Réalisé pour le compte de l'exécuteur testamentaire de Sylvia Howland, Thomas Mandell, par le célèbre mathématicien et astronome américain Benjamin Peirce, assisté de son non moins célèbre fils Charles Peirce. Voir par exemple `http ://www-groups.dcs.st-and.ac.uk/ history/Mathematicians/Peirce_Benjamin.html` et `http ://www-groups.dcs.st-and.ac.uk/ history/Mathematicians/Peirce_Charles.html`.

2) Tentez d'expliquer comment les Peirce ont pu parvenir, à partir de leur étude, à la valeur de $1/5^{30}$. Sur quelles hypothèses ont-ils pu s'appuyer ? Comment jugez-vous la pertinence et la fiabilité de leur argument ?
3) Dans le cadre du procès, un échantillon de 110 signatures tracées par l'ancien président des Etats-Unis John Quincy Adams fut analysé, révélant que les douze signatures de l'échantillon les plus proches entre elles présentaient des similarités supérieures à celles observées entre les deux signatures figurant sur le testament de 1862. L'argument fut employé par les avocats de Mme Howland Green pour affirmer qu'une telle similitude pouvait survenir de manière naturelle. Les avocats de la partie adverse rétorquèrent que le président Adams était connu pour posséder une écriture particulièrement uniforme. D'autres exemples de signatures très voisines produites par une même personne furent donnés (entre autres, à partir de chèques bancaires). Quelle est, selon-vous, la portée de ces arguments ?
4) Il fut également proposé qu'une corrélation importante pouvait exister entre des signatures réalisées par une même personne à peu de temps d'intervalle, à la même place et sur le même bureau, par exemple. Que pensez-vous de cet argument ?
5) En définitive, si vous deviez étudier vous-même la question, de quelles données chercheriez-vous à disposer, et comment procéderiez-vous ?

On notera que les Peirce ont évité l'erreur qui aurait consisté à comparer la signature inscrite sur la page supplémentaire avec simplement une autre signature, ce qui aurait constitué un exemple d'oubli de variabilité. Au contraire, ils se sont attachés à tenter de modéliser et de mesurer cette variabilité afin de l'utiliser pour fournir des arguments. Une autre erreur que l'on pourrait commettre au vu de leurs résultats (mais pas au vu de signatures elles-mêmes) serait de s'attendre à ce qu'il y ait exactement six traits qui coïncident (encore un oubli de variabilité). On peut cependant s'interroger sur le fait que l'utilisation d'une comparaison des traits dirigés vers le bas ait pu être suggérée par les donnée voir ? ? ?s.

Sophisme du procureur ou pas, l'affaire fut tranchée en définitive sur la base d'arguments purement juridiques et complètement indépendants des considérations présentées ci-dessus, qui donnèrent tort à Mme Howland Green. Comme le soulignent Meier et Zabell, cités par CHANCE News (voir ci-dessous), la question de savoir si la cour aurait tranché en sa faveur si la seconde signature avait éét considérée comme authentique, reste ouverte.

Pour plus de détails, voici les sources que nous avons utilisées :

- CHANCE News (Avril-Mai 2001) :
  `http://www.dartmouth.edu/~chance/chance_news/news.html`
- Un article publié dans le Journal of the American Statistical Association : Benjamin Peirce and the Lowland Will JASA Vol. 75, NO. 371, 497-506. Paul Meier and Sandy Zabell
- L'encyclopédie libre et gratuite wikipedia :
  `http://en.wikipedia.org/wiki/Robinson_v._Mandell`

6.6 L'affaire Sally Clark : une chance sur 73 millions !

En 1997, Mme Sally Clark perdit son premier enfant, alors âgé de 11 semaines, et le décès fut attribué à des causes naturelles. L'année suivante, son

deuxième enfant mourut, âgé de huit semaines. Mme Clark fut alors arrêtée et accusée du meurtre de ses deux enfants, puis jugée, reconnue coupable, et condamnée en 1999 à la prison à perpétuité. Pourtant, les éléments de preuve d'ordre médical étaient extrêmement ténus, voire inexistants, et rien ne laissait à penser *a priori* que Mme Clark ait pu être une mère négligente ou violente envers ses enfants. En fait, il semble bien que la conviction du jury ait été emportée par un argument de nature statistique, version moderne du dicton selon lequel la foudre ne frappe jamais deux fois au même endroit, et affirmant en substance qu'il faudrait une coïncidence vraiment extraordinaire pour que l'on observe non pas une, mais deux morts subites du nourrisson successives au sein d'une même famille. Sir Meadow, qui témoigna au procès en tant qu'expert médical, affirma que la probabilité d'une telle coïncidence (que surviennent par hasard deux morts subites du nourrisson dans une famille comparable à celle de Mme Clark) était d'environ une chance sur 73 millions, ce qui fut apparemment interprété comme un argument décisif indiquant la culpabilité de Mme Clark, et présenté comme tel par les médias à l'époque.

Il s'agit ici clairement d'un exemple du sophisme du procureur. Le chiffre proposé ne signifie certainement pas, *a priori* du moins, qu'il y ait seulement une chance sur 73 millions pour que Mme Clark soit innocente. En appelant T l'événement correspondant aux circonstances connues, en l'occurrence le décès des deux enfants, I le fait que Mme Clark soit innocente, et C le fait qu'elle soit coupable, nous cherchons à évaluer

$$\mathbb{P}(C|T) = \frac{\mathbb{P}(T|C) \times \mathbb{P}(C)}{\mathbb{P}(T|C) \times \mathbb{P}(C) + \mathbb{P}(T|I) \times \mathbb{P}(I)}.$$

Ici, on peut raisonnablement poser $\mathbb{P}(T|C) = 1$, l'estimation fournie par Sir Meadow est que $\mathbb{P}(T|I) = 1/73000000$, mais nous ne savons pas grand-chose de $\mathbb{P}(C)$ Pour estimer la probabilité que Mme Clark soit coupable au vu des circonstances (le décès de ses deux enfants), il serait donc nécessaire d'estimer la probabilité pour qu'une mère se rende coupable du meurtre de ses deux enfants. En admettant la validité de l'estimation proposée par Sir Meadow, on voit que la question serait alors de comparer $\mathbb{P}(C)$ à $1/73000000$ (en admettant, ce qui semble raisonnable, que $\mathbb{P}(I) \approx 1$), la valeur de $\mathbb{P}(C|T)$ pouvant être modifiée du tout au tout selon le résultat de cette comparaison. Or ce point avait été complètement ignoré lors du procès. Des études statistiques menées ultérieurement par le Professeur Hill, de l'université de Salford, conduisent à proposer que, raisonnablement, $\mathbb{P}(T|I) \approx 9 \times \mathbb{P}(C)$, d'où une probabilité de culpabilité au vu des circonstances d'environ $1/10$! Une autre erreur grossière réside dans la manière dont a été évaluée la probabilité $\mathbb{P}(T|I)$ lors du procès. Le raisonnement partait du chiffre d'environ $1/8500$ pour la probabilité d'une mort subite du nourrisson au sein d'une famille comparable à celle des Clark, d'où une estimation de $1/8500 \times 1/8500 \approx 1/73000000$ pour la probabilité de deux morts subites au sein de la même famille, ce qui revenait à supposer l'indépendance de ces deux événements. Or il existe, ne serait-ce qu'*a priori*, de sérieuses raisons de douter de cette indépendance : face à une affection mal connue, il semble plausible que des facteurs génétiques ou environnementaux puissent affecter de manière similaire deux enfants nés d'un même couple. Qui plus est, l'étude menée par le Pr Hill semble indiquer que le risque de mort subite est entre 5 et 10 fois supérieur chez un enfant dont un frère ou une sœur

est lui-même décédé de mort subite du nourrisson.

Les Clark firent appel du jugement, s'appuyant en particulier sur des avis de statisticiens dénonçant le sophisme du procureur, ainsi que l'erreur probable d'évaluation liée à l'emploi d'une hypothèse d'indépendance infondée. L'appel fut rejeté, la conclusion du juge étant que le point essentiel était la rareté de l'apparition de deux morts subites au sein d'une même famille, non remise en question par ces remarques. Devant une telle incompréhension, la Société Royale de Statistique écrivit aux autorités judiciaires pour enfoncer le clou. De plus, on découvrit que des éléments médicaux accréditant largement l'hypothèse d'une mort accidentelle du deuxième enfant avaient été dissimulés lors du procès. Un second procès en appel fut alors organisé, et Mme Clark fut finalement acquittée après avoir passé près de deux ans et demi en prison. Sir Meadow a été radié en 2005 par l'ordre des médecins du Royaume-Uni, pour «serious professional misconduct.»

mettre en exo la remarque préc. il y aurait quelques exemples ds la pop mais si la majorité est constituée de meurtriers...

Exercice 5 *(Questions diverses)*
1) D'après vous, que signifie le fait d'évaluer la probabilité de mort subite du nourrisson dans une famille «comparable» à celle des Clark ? Quels critères peut-on ou doit-on retenir pour s'assurer de cette «comparabilité» ?
2) Pour citer la lettre de la Société Royale de Statistique, «The fact that two deaths by SIDS [sudden infant death syndrome] is quite unlikely is, taken alone, of little value. Two deaths by murder may well be even more unlikely. What matters is the relative likelihood of the deaths under each explanation, not just how unlikely they are under one explanation.» Etes-vous (enfin) convaincu ?
3) La «loi de Meadow» citée par les médias lors du procès affirmait : «une mort subite est une tragédie, deux morts subites doivent éveiller les soupçons, trois morts subites : c'est un meurtre.» Etes-vous convaincu ? Comment traduirait-on cette loi en termes de probabilités ?

Pour plus de détails, voici les sources que nous avons utilisées :
- CHANCE news (Juillet-Août 2005) :
  `http://chance.dartmouth.edu/chancewiki/index.php/Main_Page`
- La lettre de la Royal Satistical Society :
  `http://www.rss.org.uk/main.asp?page=1225`
- L'encyclopédie libre et gratuite wikipedia :
  `http://en.wikipedia.org/wiki/Prosecutor's_fallacy`
- Un article de la revue Plus Magazine :
  `http://pass.maths.org.uk/issue21/features/clark/`
- Le site officiel de la campagne pour la libération de Sally Clark :
  `http://www.sallyclark.org.uk/`

6.7 L'affaire Collins

6.8 Le sophisme du sophisme du procureur

Un grave travers serait de déduire de ce qui précède qu'il est la plupart du temps inutile de chercher à tirer des conclusions à partir d'arguments de nature statistique, sous prétexte que l'on peut toujours être en présence du sophisme du procureur (ou de l'un des nombreux types d'erreurs plus ou moins subtiles

que nous rencontrerons dans ce cours). Il est important de garder à l'esprit que les exemples présentés ici sont des exemples parfaitement évitables d'erreurs de méthode dans l'utilisation de la démarche statistique, qui constitue dans des conditions normales un puissant outil au service du raisonnement. Nous n'insistons sur les erreurs possibles que pour mieux vous permettre d'appliquer correctement cette démarche, ou de la critiquer correctement lorsque vous y êtes confrontés, en en comprenant la portée et les limites.

Prendre prétexte des erreurs commises dans l'utilisation d'une démarche pour discréditer la démarche elle-même est déjà en soi un bel exemple de sophisme.

6.9 Encore des exercices

Exercice 6 *(Le sophisme de l'avocat)*

*Revenons sur le cas de M. D***. Admettons que la population considérée comporte 100000 personnes, ainsi que l'évaluation de 1/10000 proposée par l'expert pour la probabilité d'une coïncidence fortuite du test ADN. En plus du test ADN positif, un témoin affirme avoir été présent sur le lieu du crime et reconnaître M. D***. La probabilité que ce témoin reconnaisse correctement une personne est évaluée à* 98%.

*1) Au vu de ce chiffre, le témoin vous paraît-il fiable ? En ne tenant pas compte (pour l'instant) du résultat du test ADN, comment évaluez-vous la probabilité pour que M. D*** soit coupable sur la foi de ce témoignage ?*

2) Comment peut-on parvenir à cette estimation de 98% ? Quelle fiabilité accorderiez-vous à cette estimation ?

*3) En ne tenant compte que du résultat du test ADN, comment évaluez-vous la probabilité pour que M. D*** soit coupable ?*

*4) Les résultats des questions 1 et 3 incitent-ils à tenir M. D*** pour coupable ? En tenant compte des deux éléments à charge : test ADN positif et reconnaissance par un témoin, comment évaluez-vous la probabilité de culpabilité de M. D*** ? En admettant que M. D*** soit innocent, pensez-vous qu'il y ait indépendance entre le résultat du test ADN et le fait que le témoin affirme reconnaître M. D*** ?*

*5) Dans ce qui précède, la possibilité que le témoin produise délibérément un faux témoignage est-elle prise en compte ? La possibilité que celui-ci ait effectivement reconnu M. D***, non pas sur les lieux du crime, mais simplement aux environs ?*

Exercice 7 *Mme G*** se trouve atteinte d'un cancer du sein. En réexaminant des clichés de contrôle pris plusieurs années auparavant, on constate la présence de plusieurs micro-calcifications suspectes n'ayant pas été à l'époque signalées par le radiologue ayant effectué l'examen, le docteur U***. Le docteur U*** est alors poursuivi devant la justice par Mme G*** qui demande des dommages et intérêts, et un expert est chargé de se prononcer sur le fait que le docteur U*** ait commis une erreur. La question se pose de déterminer la probabilité pour que le docteur D***, tout en exerçant correctement sa profession, ait pu passer à côté desdites micro-calcifications (chacun sait ou devrait savoir qu'il n'est pas toujours possible de tout déceler sur un cliché d'imagerie médicale).*

*1) L'un des problèmes rencontré par l'expert est que, étant au courant de la présence d'anomalies sur les clichés de Mme D***, il lui est impossible de retouver*

un regard neutre sur les clichés en ignorant cette information : il ne peut s'empêcher de voir ces anomalies. Quelle méthode proposeriez-vous pour estimer la probabilité recherchée ? Quelle pertinence accorderiez-vous à votre méthode, et quelle fiabilité à ses résultats ?
*2) Admettons que l'on puisse évaluer à 40% la probabilité pour qu'un radiologue exerçant correctement sa profession soit passé à côté des micro-calcifications présentes sur le cliché de Mme G***. D'après vous, cette évaluation conduit-elle à appuyer la demande de dommages et intérêts de Mme G*** ? Si oui, doit-elle selon vous affecter le calcul du montant de ces dommages et intérêts (et si oui, comment) ?*

Exercice 8 *Dans la discussion du cas de M. D***, nous avons expliqué intuitivement la situation par le fait que, si l'estimation de 1/10000 proposée par l'expert est exacte, on peut s'attendre à ce qu'une population de 100000 personnes contienne de l'ordre d'une dizaine de personnes n'ayant ni plus, ni moins de raisons d'être soupçonnées que M. D***, ce qui conduit à s'attendre à une probabilité de culpabilité de l'ordre du dizième. Pour être rares, des coïncidences fortuites peuvent néanmoins se produire dans une population suffisament vaste. Dans le cas de Mme Clark, en admettant qu'une estimation raisonnable pour la probabilité de deux morts subites au sein d'une même famille soit d'une chance sur 7 millions, on peut, de la même façon, s'attendre à ce que plusieurs familles de Grande-Bretagne soient dans ce cas, et il n'y a donc pas de raison d'incriminer particulièrement Mme Clark. Vous noterez que ce raisonnement ne fait pas intervenir la proportion de doubles infanticides. Vous semble-t-il convaincant ?*

Exercice 9 *Un assassinat vient d'être commis, et les suspects se limitent à un ensemble de* 10 *personnes présentes sur les lieux au moment du meurtre. L'enquêteur affirme : «au vu des circonstances, il semble clair que l'assassin doit être gaucher». Après réflexion, il précise sa pensée en affirmant : «la probabilité que l'assassin soit gaucher est de 80%».*
1) Comment traduire les affirmations de l'enquêteur de manière formelle ?
*2) M. H***, gaucher, fait partie des suspects. D'après-vous, a-t-il du souci à se faire ?*

Exercice 10 *M. H*** joue au Loto, et... gagne le gros lot. Quand il tente de faire valoir ses droits, on refuse de lui verser son gain en lui opposant l'argument suivant. «La probabilité de gagner sans tricher est infime, et vous venez de gagner. Le plus probable est donc que vous n'êtes qu'un tricheur ! Estimez-vous heureux que nous ne vous trainions pas devant les tribunaux, et n'y revenez pas !» Que pensez-vous du bien-fondé de cet argument ?*

Exercice 11 *Comment le principe de la présomption d'innocence est-il, selon vous, pris en compte, ou au contraire ignoré, dans les arguments qui précèdent ?*

Exercice 12 *Relevez des exemples de mauvaise utilisation des probabilités conditionnelles dans divers documents (presse, ouvrages) et analysez-les.*

Exercice 13 *Au détour d'une discussion avec M. K***, vous apprenez que celui-ci a deux enfants, dont une fille. A combien évaluez-vous, étant donnée cette information, la probabilité pour que les deux enfants de M. K*** soient des filles ? Même question en admettant que M. K*** ajoute que la fille dont il est en train de vous parler, se prénomme Sophie.*

Exercice 14 *Voici un extrait du journal Le Monde, daté d'août 2005, dans un article consacré à la sécurité aérienne. «(...) Dans le même temps, les vols irréguliers devenaient plus meurtriers : le nombre de tués voyageant sur des charters représentait environ 20% du total des décès dus à des accidents d'avion à la fin des années 1980, contre 50% aujourd'hui.(...)»*

Cette phrase vous semble-t-elle satisfaisante ? Pourquoi ?

Exercice 15 *Une étude réalisée auprès d'adolescents américains appartenant à des gangs a révélé que 40% de ceux qui se déclaraient athées avaient déjà été condamnés pour des délits accompagnés d'actes violents. Cette proportion est plus de cent fois supérieure à celle des personnes condamnées pour des délits similaires au sein de la population totale. Cette étude montre donc clairement que l'athéisme conduit tout droit à la violence. Que pensez-vous de cet argument ?*

7 Coïncidences, événements probables et improbables

7.1 C'est vraiment incroyable !

Commençons par citer trois exemples documentés de coïncidences troublantes. (Source : `http://www.csj.org/infoserv_articles/astop_unlikely_events.htm`)

La romancière britannique Rebecca west était en train d'écrire un récit dans lequel une petite fille trouvait un hérisson dans son jardin. Aussitôt le passage écrit, les domestiques l'interrompirent dans son travail pour lui signaler qu'ils venaient de trouver un hérisson dans son jardin.

L'écrivain américain Norman Mailer n'avait pas initialement prévu, lorsqu'il entama la rédaction de son roman Barbary Shore, d'y inclure un espion russe comme personnage. Il le fit pourtant et, au cours de l'écriture du livre, ce personnage passa progressivement d'un rôle secondaire à celui de personnage principal du roman. Après que la rédaction fut achevée, les services américains de l'immigration arrêtèrent le voisin du dessus de Norman Mailer, que l'on présenta comme l'un des principaux espions russes en activité aux Etats-Unis à l'époque.

Plusieurs noms de code utlra-secrets furent utilisées par les forces Alliées dans la préparation du débarquement du 6 juin 1944 en Normandie, parmi eux : Utah, Omaha (désignant les plages où le débarquement devait avoir lieu), Mulberry (pour désigner le port artificiel qui devait être installé une fois le débarquement entamé), Neptune (pour désigner le plan des opérations navales), et Overlord (désignant la totalité de l'opération). Le 3 mai 1944, le mot Utah apparut comme l'une des réponses dans le problème de mots croisés du London Daily Telegraph ; le 23 mai, ce fut au tour d'Omaha ; le 31 mai, celui de Mulberry ; et enfin, le 2 juin, Neptune et Overlord firent leur apparition dans le même contexte ! Après une enquête poussée des services de renseignement britanniques, l'auteur des problèmes de mots croisés apparut comme totalement innocent, sans aucune idée du projet de débarquement, et ayant apparemment choisi au hasard les mots employés.

Plus loufoque : en 1981, le prince Charles s'est marié, Liverpool a été champion d'Europe, et le Pape est décédé. En 2005, également le prince Charles s'est marié, Liverpool a été champion d'Europe, et le Pape est décédé.

Vous avez certainement connaissance d'une foule d'autres anecdotes de ce genre, peut-être issues de votre expérience personnelle («Je pensais justement hier soir à mon ami Jojo que je n'avais pas vu depuis deux ans et... chose incroyable, il m'appelle au téléphone ce matin.» «En visitant le château de Blois lors des dernières vacances, c'est incroyable, je tombe sur mon collègue T*** au beau milieu de la cour.», «C'est vraiment surprenant que tu évoques ce sujet, car justement, nous en parlions hier ma femme et moi.»,...), et les exemples les plus frappants sont parfois rapportés dans les journaux. En rédigeant ce cours, j'ai appris qu'un collègue m'avait aperçu la veille (un dimanche) à un péage autoroutier, où nous nous trouvions donc simultanément lui et moi.

On justifie souvent son propre étonnement devant ce genre de coïncidence par des arguments basés sur la probabilité extrêmement faible de l'événement en question.

Exercice 16 *Comment définiriez-vous la probabilité des coïncidences présentées dans les exemples ci-dessus ?*

Il paraît en effet assez raisonnable, dans les exemples évoqués plus haut, de n'attribuer qu'une probabilité assez faible aux coïncidences dont il est question. Mais pourquoi au juste devrait-on s'étonner de les avoir observées ?

7.2 Ce que l'on observe est presque toujours improbable

Prenons l'exemple le plus simple de modèle probabiliste, c'est-à-dire une succession indépendante de lancers de pile ou face. Lançons une pièce dix fois de suite, et notons la suite de résultats obtenus : P pour pile et F pour face. Nous obtenons donc une suite de P et de F de longueur 10, telle que PPFPFFPPFP. Quelle probabilité une suite $(x_1, \ldots, x_{10}) \in \{\text{P,F}\}^{10}$ a-t-elle de sortir dans notre modèle ? Réponse : $1/2^{10}$ quelle que soit la suite, soit moins d'une chance sur 1000. Autrement dit, quel que soit le résultat produit par nos lancers, nous constaterons toujours qu'il n'avait qu'une très faible probabilité de survenir. Cette constatation ne vaut pas seulement pour ce cas particulier, mais pour la plupart des modèles probabilistes et des situations concrètes, dès que l'on cherche à les décrire autrement que par un très petit nombre d'alternatives différentes. De ce point de vue, tout ce que l'on observe, décrit avec suffisament de détail, possède une probabilité extrêmement faible pour la plupart des définitions raisonnables de la probabilité. La probabilité pour que vous vous trouviez exactement là où vous vous trouvez, et non pas quelques centimètres plus loin, que vous ayez exactement la position que vous avez, que vous ayez rencontré aujourd'hui les personnes que vous avez rencontrées, à l'instant exact où vous les avez rencontrées, est vraisemblablement très faible. De ce point de vue, il n'y a pas lieu de s'étonner de la faible probabilité de l'événement que l'on vient d'observer.

7.3 Des coïcidences surprenantes doivent se produire

Une autre manière de raisonner sur les coïncidences frappantes, consiste à les replacer dans un cadre plus général, dans lequel on prend en compte l'ensemble des circonstances susceptibles de nous apparaître comme des coïncidences surprenantes dans un contexte donné (au cours d'une période de temps donnée, parmi un groupe d'individus donnés, etc...). Même si chacune de ces

coïncidences possède individuellement une très faible probabilité de survenir, le grand nombre d'événements que nous sommes susceptibles d'interpréter comme des coïncidences étonnantes peut rendre extrêmement probable le fait que nous observions régulièrement – et donc relevions – un certain nombre d'entre elles.

7.4 Attention à l'interprétation

Le plus souvent cependant, les coïncidences que nous relevons ne nous frappent pas seulement en raison de leur faible probabilité (la plupart du temps bien réelle, comme nous venons de l'expliquer), mais parce qu'elles semblent suggérer une interprétation qui défie le sens commun – un destin mystérieux conduit des amis s'étant perdus de vue depuis longtemps à se retrouver par hasard lors d'un voyage à l'étranger, un étrange don de prémonition vous a fait deviner les trois premiers chiffres du tirage du loto de ce soir, ou penser à un cousin éloigné juste avant que celui-ci ne vous appelle au téléphone, etc...

L'attitude rationnelle face à ces coïncidences consiste bien entendu à tester d'abord de manière systématique les «conclusions» que leur interprétation suggère, avant de gloser plus avant. Par exemple, le fait de penser à une personne accroît-il réellement la probabilité que celle-ci vous appelle peu après ? Pour en juger, il est nécessaire d'enregistrer systématiquement les occasions où il vous arrive d'évoquer une personne de connaissance en pensée, et de mesurer la fréquence avec laquelle ces pensées sont suivies d'un appel de la personne en question dans un délai raisonnablement bref. Ainsi, on évite le **biais de sélection** (ici, d'origine psychologique), consistant à s'étonner, et donc à retenir, les cas où la personne à laquelle vous venez de penser vous appelle, tout en oubliant de remarquer, et donc en négligeant, tous les cas où l'on pense à une personne sans que celle-ci n'appelle dans les minutes qui suivent, et le problème plus évident, mais parfois ignoré, de l'oubli de variabilité qui consisterait à tirer des conclusions à partir de l'observation d'une unique coïncidence.

Il paraît vraisemblable qu'en procédant de cette manière, aucun accroissement significatif de la probabilité d'être appelé ne sera mis en évidence. Toutefois, cela peut parfaitement être le cas sans que cela soit pour autant le signe que vous possédez un don particulier, tout simplement parce qu'il peut être plus probable d'évoquer en pensée des personnes auxquelles on a eu affaire dernièrement, en particulier ses proches, et qui sont par conséquent plus susceptibles de vous appeler que d'autres.

Exercice 17 *Comment définiriez-vous précisément un protocole permettant d'étudier les relations entre le fait de penser à une personne et le fait que celle-ci vous appelle peu après ? Comment comptez-vous procéder pour distinguer le don de la coïncidence ?*

7.5 Quand s'étonner ?

Les observations précédentes sont destinées à vous mettre en garde contre un étonnement infondé ou, pire une interprétation erronnée, face à des coïncidences observées, ou rapportées (à ce propos, se pose toujours le problème de la fiabilité des sources).

Pourtant, si un modèle d'une situation prédit qu'un certain événement ne doit survenir qu'avec une faible probabilité, n'y a-t-il jamais lieu d'être sur-

pris, c'est-à-dire de mettre en doute le modèle, si l'on observe cet événement ? La réponse est positive, mais cela n'est pas incompatible avec les remarques précédentes.

7.5.1 A priori et a posteriori

Dans ce qui précède, nous avons constaté que, la plupart du temps, on pouvait **rétrospectivement** attribuer une très faible probabilité à la manière particulière selon laquelle une situation s'était réalisée. Il est bien évident que, dans ce cas, l'événement dont on examine la probabilité dépend de la manière dont la situation s'est réalisée (c'est complètement évident dans l'exemple des lancers de pile ou face). En revanche, lorsque l'événement de faible probabilité auquel on s'intéresse est fixé indépendamment – par exemple à l'avance – de la réalisation de l'expérience, il y a tout lieu d'être surpris si celui-ci se produit, et cela doit inciter, sinon à rejeter le modèle, du moins à réexaminer les arguments en faveur de celui-ci (de manière systématique, naturellement !).

Quant à savoir à partir de quel niveau de probabilité il convient de s'étonner, tout dépend du contexte, et il n'est pas forcément de bonne politique de fixer une limite *a priori* en-deçà de laquelle les événement sont considérés comme improbables, et au-dessus de laquelle leur apparition doit être considérée comme non-surprenante.

D'autre part, en pratique, il n'est bien entendu pas toujours évident de s'assurer qu'il y a bien indépendance entre l'événement considéré et la réalisation de l'expérience (voir la section «Hypothèses suggérées par les données»).

7.5.2 Familles d'événements

La seconde remarque concernant les événements de faible probabilité consistait à noter qu'un très grand nombre d'événements de faible probabilité susceptibles de survenir seraient remarqués comme des coïncidences, et qu'il était donc plus pertinent de considérer la probabilité de la réunion de la totalité de ces événements, plutôt que la probabilité de l'un d'entre eux (celui qui justement s'est produit) isolément. (On évite ainsi de faire dépendre l'événement que l'on considère de la manière dont l'expérience s'est réalisée.) Dans le cas où l'événement que l'on considère ne dépend pas de la réalisation de l'expérience, il n'y a pas lieu de dresser une telle liste !

7.5.3 Probabilité d'une réunion

De manière générale, on ne peut pas déduire la probabilité d'une réunion $\mathbb{P}(A_1 \cup \cdots \cup A_n)$ des probabilités individuelles, $\mathbb{P}(A_i)$, et l'on dispose seulement d'inégalités, telle que la borne de la réunion :

$$\mathbb{P}(A_1 \cup \cdots \cup A_n) \leq \sum_{i=1}^{n} \mathbb{P}(A_i),$$

que l'on peut utiliser en toute généralité, et qui est une égalité lorsque les événements A_i sont deux-à-deux disjoints.

Dans le cas général toujours, la borne de la réunion peut être raffinée en inégalités plus compliquées (mais en général plus précises), ou en égalité (principe

d'inclusion-exclusion). Précisément, en posant, pour $1 \leq k \leq n$,

$$C_k = (-1)^{k-1} \sum_{i_1 < \cdots < i_k} \mathbb{P}(A_{i_1} \cap \ldots \cap A_{i_k}),$$

on a

$$\mathbb{P}(A_1 \cup \ldots \cup A_n) \leq \sum_{k=1}^{m} (-1)^{k-1} C_k$$

lorsque m est impair,

$$\mathbb{P}(A_1 \cup \ldots \cup A_n) \geq \sum_{k=1}^{m} (-1)^{k-1} C_k$$

lorsque m est pair, et

$$\mathbb{P}(A_1 \cup \ldots \cup A_n) = \sum_{k=1}^{n} (-1)^{k-1} C_k.$$

En revanche, dans le cas d'événements indépendants, par exemple, on peut écrire que

$$\mathbb{P}(A_1 \cup \ldots \cup A_n) = 1 - \prod_{k=1}^{n} (1 - \mathbb{P}(A_i).$$

On voit ainsi que, en toute généralité, si l'on dispose de n événements dont toutes les probabilités sont inférieures à une valeur ϵ, on ne peut en général en déduire que le fait que $\mathbb{P}(A_1 \cup \ldots \cup A_n) \leq n\epsilon$ (ce qui ne veut bien entendu pas dire que cette probabilité est effectivement de l'ordre de $n\epsilon$).

7.6 Un magicien doué

Voici un petit exemple imaginaire destiné à illustrer quelques unes des observations précédentes.

Ce soir, au cours d'une émission de télévision à succès, une expérience de parapsychologie de grande ampleur est en train d'avoir lieu. M. M***, magicien de son état, prétend pouvoir, par la seule force de son esprit, être capable de faire griller les ampoules électriques chez les téléspectateurs. Ceux-ci ont donc été invités à allumer chez eux diverses lampes électriques, et, après plusieurs minutes de concentration intense de la part de M. M***, des téléspectateurs appellent par dizaines le standard de l'émission pour témoigner qu'effectivement, une, et même dans certains cas plusieurs ampoules électriques ont rendu l'âme pendant que M. M*** se concentrait.

En admettant que l'émission en question soit regardée par plusieurs millions de foyers, et que la probabilité pour une ampoule électrique de griller au cours d'une minute d'utilisation soit d'environ $1/60000$ (ce qui correspond à une durée de vie moyenne d'environ mille heures), on s'attend à ce qu'il y ait en moyenne plusieurs milliers de téléspectateurs chez qui des ampoules grillent au cours de l'émission, et, par voie de conséquence (que feriez-vous à leur place, hein ?), contactent le standard de l'émission.

Pris individuellement, le fait qu'une ampoule grille au moment précis ou M. M*** se concentre semble très surprenant, car très improbable sous l'hypothèse

que M. M*** ne détient aucun pouvoir particulier : environ une chance sur 60000. Pourtant, si notre estimation de 1/60000 est correcte, ainsi que celle de plusieurs millions de téléspectateurs, ainsi que l'hypothèse d'une certaine indépendance entre le grillage des ampoules chez les différents téléspectateurs, ce qui serait surprenant serait plutôt que personne n'appelle pour faire part de sa surprise. Replacés parmi l'ensemble des grillages d'ampoules susceptibles de survenir chez les téléspectateurs, la multitude d'appels constatée n'a donc rien de surprenant. Bien entendu, les téléspectateurs chez qui rien de particulier n'est survenu, qui forment pourtant l'écrasante majorité (sans doute plus de 99,9%) ne se précipitent pas forcément sur leur combiné pour composer le numéro (peut-être surtaxé) permettant d'appeler l'émission, car ils ne pensent pas avoir observé quoique ce soit de remarquable. Si l'on ne se fie qu'aux appels passés pour estimer la probabilité de succès de M. M***, on commet tout simplement un (atroce) biais de sélection.

Intéressons-nous maintenant à la manière dont peut raisonner un téléspectateur sceptique chez qui une ampoule électrique vient pourtant de rendre l'âme. N'accordant aucun crédit à M. M***, il cherche pourtant à examiner les faits objectivement, et ne peut que constater le succès de M. M*** en ce qui le concerne. L'événement qui vient d'être observé est très improbable sous l'hypothèse que M. M*** ne possède aucun don, et cet événement a bien été défini indépendamment du résultat de l'expérience, avant que celle-ci n'ait lieu (ou du moins, c'est ainsi que M. D*** voit les choses individuellement, nous savons qu'il n'en est rien puisque nous nous intéressons à M. D*** justement à cause du résultat de l'expérience survenu chez lui). M. D*** devrait donc être amené à remettre en question la validité de son hypothèse selon laquelle M. M*** n'est qu'un charlatan ? Eh bien oui ! Cependant, M.D*** doit tenir compte de l'ensemble des éléments dont il dispose, qui, vraisemblablement, l'incitent très fortement à douter de la réalité des pouvoirs de M. M***, et le résultat de l'émission ne constitue donc pas nécessairement un argument très fort en faveur des pouvoirs parapsychologiques. (Le raisonnement bayésien fournit un cadre pour quantifier ceci de manière précise. Voir la section «Raisonnement bayésien.»)

Si M. D*** cherche à aborder les choses de manière systématique, il tentera à nouveau l'expérience (en admettant que M. M*** réapparaisse plusieurs fois à la télévision, ou que M. D*** va jusqu'à inviter M. M*** chez lui pour en avoir le cœur net), pour constater que M. M*** ne réussit presque jamais. Ou encore, il s'informera des résultats constatés chez un grand nombre de personnes (pas seulement chez celles ayant contacté l'émission, sous peine de biais de sélection, mais au sein d'un échantillon représentatif), pour constater que M. M*** n'a réussi chez quasiment aucune d'entre elles. Cette situation est bien entendu un peu caricaturale, car peu de gens prennent au sérieux les parapsychologues et autres tordeurs de petites cuillères, mais le même genre de phénomène peut apparaître dans bien d'autres contextes.

Imaginons par exemple que 50 équipes scientifiques étudient séparément l'impact d'un nouveau produit, disons la vitamine X, sur la guérison d'une maladie, par exemple le cancer. Chaque équipe conduit son étude dans les règles (essais randomisés en double aveugle, échantillons représentatifs de la population à traiter, constitution de groupes témoins et utilisation de placebos, voir partie... pour une discussion de ces termes). Sur les 50 équipes, 49 observent des résultats non-concluants quant à l'efficacité du médicament. En revanche l'une

des équipes observe un taux de guérison si élevé chez les patients traités à l'aide de la vitamine X, que, sous l'hypothèse que la vitamine X est sans effet sur le cancer, on ne puisse espérer observer un tel taux qu'avec une probabilité d'environ 2%. L'équipe en question, qui travaille seule, estime avoir de bonnes raisons de penser que la vitamine X possède un effet réel sur le cancer. Voir la section «Hypothèses suggérées par les données» pour une discussion plus approfondie de ce type de question.

Dans un autre ordre d'idées, que penser d'une théorie produisant comme résultat le fait que la probabilité d'apparition de la vie sur Terre soit extrêmement faible, et que notre existence doive donc être considérés comme le fruit d'une formidable coincidence ?

Bien entendu, une première difficulté de fond est qu'il semble extrêmement difficile de produire des évaluations de probabilité raisonnables, en rapport avec des mécanismes fort mal connus, ce qui rend très douteuse toute tentative de calcul de ce type. De plus, ce type d'argument se heurte souvent au fait que le calcul porte sur la probabilité d'apparition de la vie sous la forme exacte où nous la connaissons, et non pas d'une forme de vie en général, et nous avons vu le caractère problèmatique de ce raisonnement (qui a pourtant été employé comme). En admettant que ces difficultés soient surmontées, quel sens ce type de remarque pourrait-il avoir, si ce n'est de montrer que la vie doit être rare parmi les planètes présentant des conditions comparables à celles prises en compte dans l'évaluation de la probabilité ? En prenant en compte le nombre de planètes de ce type, l'existence d'au moins une planète abritant la vie peut devenir moins problématique. En définitive, une théorie qui prédirait que la vie ne se produit dans l'univers qu'avec une faible probabilité ne signifierait pas grand-chose ?.....

Simplement de montrer que

Exercice 18 *Au cours d'une émission, on invite une vingtaine de médiums censés deviner des informations sur des membres du public choisis au hasard (par exemple, leur nombre d'enfants, s'ils sont ou non célibataires, etc...). A chaque étape, les médiums ayant deviné juste restent sur scène, tandis que les autres sont éliminés. Après cinq étapes, M. H*** est le seul à rester en lice, et couronné comme possédant un don vraiment exceptionnel. Pensez-vous que cela soit justifié ? En quoi l'élimination progressive peut-elle tendre à accréditer indûment, – auprès des spectateurs non-avertis, bien entendu – M. H*** ?*

7.7 Bibliographie

Nous vous invitons à consulter le chapitre «Les coïncidences exagérées» de l'ouvrage suivant, que nous avons utilisé :

«Devenez sorciers, devenez savants», G. Charpak et H. Broch, Eds Odile Jacob, 2002.

Exercice 19 *Fixer à l'avance, c'est-à-dire avant de réaliser l'expérience en question, un événement de faible probabilité, garantit-il, selon vous, le fait que le choix de cet événement puisse être considéré comme indépendant du résultat de l'expérience ?*

Exercice 20 *Comment, en partant d'une donnée telle que la durée de vie moyenne d'une ampoule, peut-on parvenir à une estimation de la probabilité*

pour qu'une ampoule grille au cours d'une minute d'émission ? Sur quelles hypothèses peut-on s'appuyer, et comment pourrait-on valider celles-ci ? Comment pourrait-on raffiner le modèle en tenant compte de l'existence de différents types d'ampoules, de différents types de comportements d'achat d'ampoules (renouvellement), éventuellement de différents types d'utilisation de celles-ci ? Ceci conduirait-il à modifier beaucoup l'estimation du nombre de personnes pouvant par pur hasard voir griller leur ampoule au cours de l'émission ?

8 Sur les tests statistiques

8.1 Nos objectifs

Les tests statistiques constituent un ensemble de techniques destinées à tester des hypothèses de modélisation de nature diverse, dans une très grande variété de contextes. Ils sont souvent spécialement conçus pour s'adapter à des hypothèses et des contextes particuliers, si bien qu'il existe un nombre considérable de procédures de test différentes, certains tests pouvant toutefois être emplyés dans des conditions plus générales que d'autres. Il n'est pas question d'établir dans ces brèves notes une liste des différents tests statistiques existants. En revanche, comme la plupart des procédures de test sont semblables dans leurs grands principes, ce sont ces principes généraux que nous allons exposer ici.

Il faut savoir que l'utilisation des tests statistiques est courante, et même standard dans un grand nombre de domaines scientifiques et industriels, et que l'on est donc, si ce n'est amené à les utiliser soi-même, du moins très fréquemment confronté aux résultats qu'ils produisent. Bien maîtriser les principes et notions fondamentales en rapport avec les tests statistiques doit vous permettre de :

1. comprendre correctement la signification des sorties produites par un test, le plus souvent mis en œuvre au moyen d'un logiciel spécialisé,
2. parvenir à comprendre les explications décrivant un test statistique dans un document de référence,
3. vous assurer que les conditions de validité d'un test sont bien remplies,
4. apprécier correctement la portée des résultats d'un test, ce qui est primordial lorsque ceux-ci sont utilisés comme arguments dans une prise de décision.

Le point 1) est important car la complexité, qui n'est souvent qu'apparente des sorties produites par une procédure de test, peut décourager le sens critique au point de s'en remettre aveuglément à l'interprétation proposée par d'autres du résultat du test, ce qui n'est pas forcément souhaitable.

Le point 2) est important car il doit vous permettre d'être relativement autonomes, et de vous frayer un chemin au milieu du vaste labyrinthe formé par les différents types de tests, que, comme nous l'avons dit, il n'est pas possible de répertorier dans un cours aussi bref.

Le point 3) est important car, si la validité d'une procédure de test repose la plupart du temps sur un certain nombre d'hypothèses, parfois très spécifiques, parfois plus souples, il n'est en général pas le moins du monde nécessaire que ces hypothèses aient été vérifiées pour que la procédure puisse être mise en œuvre,

et fournisse des résultats qui, pour provenir de l'ordinateur accompagnés de noms savants et/ou ronflants (la plupart du temps anglo-saxons), et de tas de nombres comportant beaucoup de décimales et mis en tableaux, ne signifient absolument rien, et peuvent très facilement mener à des conclusions erronées. Rien ne ressemble plus à une sortie de test dont les hypothèse de validité ont été vérifiées qu'une sortie de test pour laquelle on ne s'est même pas posé la question.

Le point 4), qui repose entre autres sur les points précédents, permet d'adopter une attitude informée, rationnelle et critique face à l'emploi des tests statistiques. La compétence, c'est important !

8.2 Comment teste-t-on une hypothèse (en sciences) en général ?

Un premier point important qu'il convient de noter est que, de manière générale, les hypothèses que l'on cherche à tester sont formulées dans le cadre d'un modèle mathématique de la réalité, et non pas de la réalité elle-même. Une hypothèse de modélisation ne peut donc être testée (à la différence des assertions mathématiques, qui peuvent faire l'objet d'une preuve ou d'une réfutation purement théorique) qu'en la reliant à des éléments mesurables de la réalité, et l'on confrontera donc l'hypothèse à des données mesurées. Par exemple, on teste la validité de la théorie newtonienne de la gravitation en comparant ses prédictions concernant les trajectoires des corps célestes aux trajectoires effectivement mesurées. De plus, une hypothèse de modélisation ne peut généralement être testée spécifiquement que dans un certain cadre interprétatif, qui constitue lui-même un modèle de la réalité que l'on considère comme valable, suffisament validé par l'expérience passée. Par exemple, on testera une hypothèse sur la masse d'un objet céleste en s'appuyant par ailleurs sur les théories physiques communément admises (en particulier la théorie de la relativité générale) afin d'interpréter les résultats expérimentaux et d'en déduire une estimation de la masse de l'objet en question.

De manière générale, le schéma employé pour tester une hypothèse est habituellement le suivant :

Modèle admis (M) + Hypothèse à tester (H) $\overset{\text{entraîne}}{\Rightarrow}$ Conséquence testable (C)

Deux cas se présentent alors : soit (C) est effectivement observée, soit elle ne l'est pas. Dans le cas où (C) n'est pas observée, on rejette en général (H), puisque la validité de (H) aurait dû entraîner la réalisation de (C). Bien entendu, toutes sortes de vérifications sont en général nécessaires afin d'éliminer les diverses causes d'erreur (par exemple des effets parasites qui viendraient perturber la réalisation de (C)), et il se peut aussi que la non-observation de (C) aboutisse à la remise en question, voire au rejet de (M), ou tout au moins d'une partie de (M), même si (M) représentait une théorie bien établie jusqu'alors.

Inversement, si (C) est observée, (H) n'est pas prise en défaut et notre confiance en (H) (et en (M) par la même occasion) est donc accrue, même si (C) ne nous fournit pas de raison positive d'accepter la validité de (H). On accepte donc (H) par défaut.

Un problème fondamental, dans ce contexte, est celui du pouvoir de discrimination de (C). En effet, il est parfaitement concevable que (H) ne soit pas

valable mais que (C) puisse néanmoins se produire. Plus ceci est vrai, moins (C) permet de discriminer (H) des hypothèses alternatives. Afin de se poser correctement la question du pouvoir de discrimination de (C), il est nécessaire de préciser quelles sont les hypothèses qui correspondent à (M)+négation de (H), et sont donc en concurrence avec (H).

Au passage, notons qu'il est important de tenir compte de ce qui est admis a priori (ici, (M)) pour juger du pouvoir de discrimination de (C). Si l'on s'autorise comme hypothèse alternative l'existence d'un génie malin qui produit systématiquement les conséquences de (H) lorsque l'on teste celles-ci, mais agit à sa guise et de manière imprévisible le reste du temps, il est impossible de discriminer de quelque manière que ce soit cette «hypothèse» de (H). En général, l'existence de ce genre d'entité est exclue par ce qui est admis a priori (M).

Pour des présentations générales de la méthode scientifique, vous pouvez par exemple consulter

`http://teacher.nsrl.rochester.edu/phy_labs/AppendixE/AppendixE.html`

ou

`http://en.wikipedia.org/wiki/Scientific_method`

ainsi que les références qui s'y trouvent.

8.3 Les tests statistiques

Les tests statistiques s'écartent du schéma général décrit ci-dessus en ce qu'il est en général impossible de déduire d'une hypothèse portant sur un modèle probabiliste une conséquence qui soit à la fois certaine et non-triviale. En effet, un modèle probabiliste n'écarte comme impossibles que les éventualités de probabilité nulle. Toutes les autres éventualités, même affectées d'une très faible probabilité, peuvent survenir. Ainsi, aucune autre conséquence que l'événement certain «l'une des différentes éventualités possibles se réalise» n'est sûre à 100% dans le cadre d'un modèle probabiliste. Le schéma précédent doit donc être remplacé par celui-ci

$$\text{Modèle admis (M) + Hypothèse à tester (H)} \overset{\text{entraîne avec forte probabilité}}{\Rightarrow} \text{Conséquence testable (C)}$$

La plupart du temps, on disposera d'un échantillon $x_1, \ldots, x_N$ issues du phénomène que l'on étudie, et la procédure de test consistera donc à juger à quel point les données sont compatibles avec (H), ou, au contraire, tendent à infirmer celle-ci. La conséquence testable (C) sera donc exprimée en termes de $x_1, \ldots, x_N$.

Dans le contexte des tests statistiques, (M) contient en général l'hypothèse que le phénomène que l'on étudie peut être décrit de manière satisfaisante par un modèle probabiliste interprété de manière fréquentiste (ce qui permet de donner un sens objectif aux éléments du modèle), des hypothèses plus ou moins précises sur la forme de ce modèle, et des hypothèses sur la modélisation de la manière dont les données sont recueillies, tandis que (H) constitue une hypothèse supplémentaire sur le modèle par lequel le phénomène est décrit. Par exemple, (M) pourra supposer que le phénomène que l'on étudie produit des valeurs aléatoires distribuées selon une loi gaussienne de paramètre m inconnu et de variance 1, et que les données données recueillies forment un échantillon indépendant et identiquement distribué de cette loi, tandis que (H) pourra par

exemple spécifier que $m = 2$. On cherchera à tester (H) tout en admettant (M) comme établi.

Comme dans le cas précédent, il est possible de se tromper en acceptant (H) à tort, mais il est à présent également possible de se tromper en rejetant (H) à tort, si (H) est valable et que, par malchance, la conséquence de (H) pourtant valable avec forte probabilité ne se produit cependant pas.

On donne des noms distincts à ces deux types d'erreurs :

- erreur de type I (ou de première espèce) : rejeter (H) à tort,
- erreur de type II (ou de seconde espèce) : accepter (H) à tort.

Dans la même veine, on définit deux types de risques :

- risque de type I (ou de première espèce) : probabilité de rejeter (H) à tort,
- erreur de type II (ou de seconde espèce) : probabilité d'accepter (H) à tort.

Plus le risque de première espèce est faible, plus on dit que la procédure de test est **sensible**. Plus le risque de seconde espèce est faible, plus on dit que la procédure de test est **spécifique**.

Les notations usuelles dans ce contexte sont de désigner par (H0) (appelée «hypothès nulle») l'hypothèse que l'on cherche à tester, et par (H1) (hypothèse alternative) la négation de (H0) dans (M). Dans l'exemple précédent, (H1) sera donc l'hypothèse $m \neq 2$.

L'idéal serait de disposer d'une procédure de test dans laquelle les risques de première et de seconde espèce seraient simultanément faibles. En pratique, on se heurte à la contrainte que les deux risques varient en sens inverse : plus on souhaite limiter la probabilité de rejeter (H) à tort, plus on est conduit à accepter (H) souvent, et donc plus on est conduit à rejeter (H1) à tort, et inversement. Accepter (H0) systématiquement est un bon moyen de limiter le risque de première espèce, mais la puissance du test devient alors négligeable. Il est donc nécessaire de trouver des procédures de test réalisant un compromis entre ces deux contraintes, et c'est ce qui nous guide dans la recherche d'une procédure adéquate. La démarche usuelle consiste à se fixer une limite concernant le risque de première espèce, en imposant un seuil $\alpha \in]0, 1[$ que celui-ci ne doit pas dépasser. Autrement dit, sous l'hypothèse que (H0) est valide, la probabilité de rejeter (H0) doit être inférieure à α. D'autre part, on essaie de rendre le risque de seconde espèce le plus faible possible sous cette contrainte. La valeur $1 - \alpha$ est appelée le niveau de confiance de la procédure de test, tandis que, si β désigne le risque de seconde espèce, la valeur $1 - \beta$ est appelée la puissance du test. Plus le niveau de confiance est élevé, plus la conséquence que l'on teste est probable si (H0) est valable, et donc moins on risque de rejeter (H0) à tort. Plus la puissance est élevée, plus le test est discriminant et donc plus la probabilité d'accepter (H0) si (H1) est valable est faible.

Cette dissymétrie entre les deux risques (l'un étant contrôlé a priori, et l'autre pas) amène en général à choisir pour (H0) l'hypothèse qu'il semble le plus grave de rejeter à tort. Par exemple, on considère généralement qu'il serait plus grave, dans le cadre d'un procès, d'envoyer en prison un innocent, que de laisser libre un coupable. Dans le domaine médical, on pourra considérer comme plus grave de mettre sur le marché un médicament dangereux que de passer à côté d'un traitement plus efficace que ses prédécesseurs.

Si (H1) contient des situations pouvant être arbitrairement proches de (H0) (c'est le cas dans notre exemple où (H0) stipule que $m = 2$ tandis que (H1) si-

gnifie simplement que $m \neq 2$, m pouvant être arbitrairement proche de 2) on ne peut clairement pas espérer donner une borne globale sur le risque de seconde espèce : les données ne permettent pas en général de séparer (H0) d'une hypothèse différente mais néanmoins très voisine, et cela est parfaitement normal. Le mieux que l'on pourra espérer sera donc que, parmi toutes les procédures de tests envisageables, celle que l'on emploie possède un risque de seconde espèce uniformément (c'est-à-dire pour toutes les situations contenues dans (H1)) plus faible que les autres. Il est parfois, mais pas toujours, possible, une limite pour le risque de première espèce étant fixée, de trouver une procédure de test qui soit systématiquement meilleure toutes les autres en ce qui concerne le risque de seconde espèce, et qui dans ce cas s'impose comme la meilleure. Tout dépend de la richesse des situations contenues dans (H1), autrement dit, tout dépend de la précision des hypothèses admises contenues dans (M). Si (H1) est très riche, une procédure de test pourra en général bien séparer (H0) de certaines des hypothèses alternatives contenues dans (H1), mais pas de certaines autres. Si (H1) n'est pas davantage précisée, il n'y a donc pas de choix unique qui s'imposerait, mais bel et bien plusieurs procédures de test différentes, entre lesquelles il n'est pas forcément facile de faire un choix. Des considérations théoriques d'une part, empiriques d'autre part, basées sur l'expérience acquise, sur l'efficacité ou la robustesse plus ou moins grande de telle ou telle procédure peuvent contribuer à guider ce choix. On retient globalement que la définition de (H1) conditionne fortement le choix de la procédure de test.

Même si l'on suppose que la puissance d'un test est bien définie lorsqu'on l'applique à une situation donnée, elle n'est la plupart du temps pas accessible à partir des données. On cherche donc en général à prendre en compte les différentes valeurs que peut prendre la puissance dans toutes les alternatives contenues dans (H1) (ce qui peut faire beaucoup !) Pour une discussion théorique de ces questions d'optimalité, qui dépendent naturellement du contexte spécifique envisagé, vous pouvez consulter ? ? ? Les méthodes bayésiennes, qui introduisent une pondération entre les différentes hypothèses, permettent d'aborder différemment l'évaluation des tests et de leurs risques.

8.3.1 La conséquence testable (C)

La plupart du temps, on dispose d'un échantillon de valeurs mesurées $x_1, \ldots, x_N$, et c'est sur ces données que l'on est censé s'appuyer pour tester l'hypothèse (H0). Dans le cadre du modèle constitué par (M)+(H0), ces valeurs mesurées apparaissent comme une réalisation d'une famille de variables aléatoires $X_1, \ldots, X_N$. Une difficulté conceptuelle est que, une fois que l'on dispose des valeurs mesurées, celles-ci ne sont bien entendu plus aléatoires, mais fixées. Cependant, on pense à ces quantités comme à des réalisations de variables aléatoires, et c'est ainsi que l'on raisonne sur leurs propriétés.

En général, le test repose sur le calcul d'une fonction des données $T(x_1, \ldots, x_N)$ à valeurs réelles, appelée une statistique, et telle que, en supposant (M)+(H0) valables, on soit en mesure de calculer la loi de probabilité de $T(X_1, \ldots, X_N)$. La procédure de test définit un sous ensemble R de $\mathbb{R}$ appelé région critique, vérifiant la propriété que, sous (H0), on ait $\mathbb{P}(T(X_1, \ldots, X_N) \in R) \leq \alpha$.

La mise en œuvre du test consiste simplement, après avoir calculé $T(x_1, \ldots, x_N)$, à regarder si $T(x_1, \ldots, x_N) \in R$ ou si au $T(x_1, \ldots, x_N) \notin R$. Dans le premier

cas, on décide de rejeter (H0), et par conséquent d'accepter (H1). Dans le second cas, on décide de conserver (H0), et par conséquent de rejeter (H1).

Le choix de la statistique et de la région critique déterminent donc la procédure de test. Souvent, (H0) consiste en l'identité de deux objets : une égalité entre paramètres, ou encore entre lois, et T fournit une mesure de l'écart entre ces deux objets, tel qu'estimé à partir des données. Plus T est grand, plus l'écart estimé est important, et donc plus on doit être amené à douter de (H0). Toute la question, à laquelle les procédures de test apportent une solution rigoureuse, est de savoir à partir de quelle valeur de T il convient de rejeter (H0). Il faut pour cela se faire une idée précise des fluctuations «normales» auxquelles on devrait s'attendre si (H0) était vérifiée, compte-tenu de la variabilité inhérente au phénomène considéré et à l'échantillonnage réalisé, et juger si la valeur de mesurée T est de l'ordre de grandeur de ces fluctuations, ou, au contraire, trop importante pour leur être attribuée. La connaissance de la loi de T sous (H0) permet un tel calibrage des fluctuations, et les méthodes de test statistique permettent donc de prendre en compte correctement la variabilité des phénomènes étudiés dans l'évaluation des hypothèses à partir des données. Parfois, T est une quantité signée, qui peut représenter un écart par rapport à (H0) dans un sens ou dans un autre, et la région critique choisie dépendra de (H1), qui parfois exclut les écarts dans un certain sens, tout en rendant les écarts dans l'autre sens probables. De manière générale, la connaissance de (H1) est fondamentale pour déterminer la région critique, de façon à obtenir un test possédant un minimum de puissance : si (H1) n'est pas trop vaste, on pourra espérer discriminer efficacement entre (H0) et les hypothèses de (H1) bien séparées de (H0). Si, au contraire, (H1) est très vaste (autrement dit, si (M) comporte peu d'hypothèses), il n'est pas raisonnable d'espérer pouvoir discriminer efficacement (H0) de toutes les hypothèses alternatives, même celles qui en sont assez éloignées. Au mieux, un test sera efficace pour séparer (H0) d'un certain type d'alternative, mais ne permettra pas de distinguer celle-ci d'une alternative correspondant à une violation de (H0) de nature complètement différente. Le type de violation de (H0) que l'on souhaite étudier détermine le choix du test, et de la zone de rejet, certaines statistiques étant plus ou moins sensibles que d'autres à certains types de violation de (H0), et le type de localisation de T entraîné par une violation de (H0) pouvant varier considérablement.

Exemple du test de comparaison d'une moyenne à une valeur fixée dans le cas asymptotique, TCL. Pourquoi un intervalle symétrique, ou un intervalle semi-infini ? Exemple de région critique «stupide».

Il est clair que, les données étant fixées, la région critique diminue lorsque α diminue : moins on souhaite se tromper en rejetant (H0) à tort, plus on doit accepter (H0) souvent. En général, il est possible de définir une valeur unique α_0 dépendant des données, telle que, pour $\alpha < \alpha_0$, (H0) soit acceptée, et, pour $\alpha > \alpha_0$, (H0) soit rejetée. On appelle α_0 la p-valeur associée aux données (et à la procédure de test), et c'est souvent cette p-valeur qui est donnée en sortie par les logiciels de statistique. La comparaison de α_0 avec le α que l'on s'est fixé fournit la réponse au test.

Exercice 21 *ex :surlestests Un test statistique rejette une hypothèse lorsqu'une conséquence probable de celle-ci n'est pas réalisée, autrement dit, lorsqu'un événement improbable s'est réalisé. En quoi les procédures de test statistique échappent-elles au problème soulevé dans le chapitre sur les coïncidences, montrant qu'il*

peut être complètement infondé de s'étonner de la réalisation d'un événement improbable ?

La discussion qui précède vous paraît peutêtre exagérément compliquée, et peut-être estimez vous qu'il est superflu de se poser tant de questions simplement pour vérifier sur des données si une hypothèse semble vérifiée ou pas. Si des techniques plus élémentaires d'examen des données (notamment par l'intermédiaire de représentations graphiques), correctement employées, peuvent mener à des conclusions correctes, la difficulté de raisonner en prenant en compte la variabilité inhérente aux phénomènes étudiés, ainsi que la propension naturelle de l'esprit humain à interpréter et voir des structures là où il n'y en a en réalité aucune, ne peuvent qu'inciter à rechercher des procédures contrôlées et systématiques pour aider au raisonnement et à la prise de décision. Les tests statistiques sont l'un de ces outils.

8.3.2 Nombre de données disponibles

Clairement, le nombre de données disponibles pour effectuer un test est une quantité importante. Il contribue de manière essentielle à la puissance du test : plus le nombre de données est important, plus il permet de séparer (H0) de ses concurrentes, et, dans une situation donnée, peut orienter le choix du type de test utilisé (paramétrique, ou non-paramétrique, par exemple). Ce nombre est souvent contraint par les moyens (temps, argent, faisabilité pratique,...) disponibles, et nous le considérons dans cette partie comme fixé. Dans le cas où l'on peut calculer, ou tout au moins obtenir des estimations de la puissance du test que l'on compte appliquer, il est également possible de déterminer à l'avance (quitte parfois à estimer des paramètres inconnus au cours d'une première phase de collecte de données) le nombre de données à collecter pour discriminer (H0) de (H1) avec à la fois un niveau de confiance et une puissance fixés.

8.4 Catégories de tests

Plusieurs qualificatifs sont en général attachés aux procédure de tests statistiques. Il est important de les comprendre car elles permettent de se faire une idée du domaine de validité et de la portée des procédures employées. Les intersections entre les catégories présentées ci-après sont non-vides. Quand nous parlons d'hypothèses nécéssaires à la validité d'une procédure de test dans ce qui suit, nous parlons des hypothèses contenues dans (M), qui sont considérées comme un préliminaire à la procédure de test, et ne sont pas explicitement testées par cette procédure, et non pas, bien entendu, (H0) et (H1).

8.4.1 Tests asymptotiques

Il font souvent appel à peu d'hypothèses (hormis, en général le caractère indépendant et identiquement distribué des données collectées) car ils la distribution de la statistique qu'ils utilisent est calculée à partir d'un théorème limite «universels» tels que le théorème de la limite centrale, qui affirme que les fluctuations par rapport à la moyenne d'une somme de variables aléatoires i.i.d. sont à la limite distribuées selon une loi gaussienne, quelle que soit (ou presque) la loi d'origine des variables aléatoires présentes dans la somme. La plupart du temps,

le paramètre par rapport auquel la limite est prise est le nombre d'observations disponibles. Une question qui se pose naturellement avec ce type de test est d'une part la validité de l'hypothèse d'indépendance, et d'autre part la qualité de l'approximation fournie par la limite d'un grand nombre d'observations, pour le nombre nécessairement fini d'observations dont on dispose effectivement. Il existe dans certains cas de bornes explicites sur la qualité de l'approximation obtenue, subordonnées parfois à des hypothèses de modélisation supplémentaires (mais assez générales) sur les données. Des recommandations accompagnent en général la documentation de ce type de test, qui permettent au moins de délimiter les situations où l'approximation est grossièrement prise en défaut. Ces tests sont parfois accompagnés de «corrections» visant à améliorer l'approximation pour des valeurs finies du nombre d'observations disponibles.

Exemples les plus connus : test de comparaisons de moyennes basé sur le TCL, test du χ^2.

8.4.2 Tests exacts

Le contraire des précédents : peu importe le nombre d'observations, ils utilisent la loi de probabilité exacte de la statistique considérée. Leurs caractéristiques doivent parfois être calculées numériquement, car des formules explicites ne sont pas toujours disponibles.

8.4.3 Tests paramétriques

Ils sont appelés ainsi car ils supposent en général que les données sont décrites par une distribution de probabilité appartenant à une famille paramétrique, telle que les gaussiennes, les lois exponentielles, les lois de Poisson, etc... Il est indispensable, avant de les utiliser, de vérifier la validité de cette hypothèse, soit à partir des données, quand on dispose d'un nombre suffisament grand d'entre elles (sans quoi, il est difficile de tester cette hypothèse avec un minimum de puissance), soit en s'appuyant sur une connaissance *a priori* du phénomène considéré. Appliquer un test reposant sur l'hypothèse que les données sont gaussiennes à des données qui ne le sont pas n'a aucune raison de fournir des résultats qui signifient quoique ce soit.

Exemple le plus connu : le test de Student.

8.4.4 Tests non-paramétriques

Ils sont appelés ainsi car ils font en général peu ou pas d'hypothèses quant à la distribution de probabilité susceptible de décrire les données (en revanche l'hypothèse d'indépendance et d'identique distribution est en général présente). Par exemple, on supposera que la distribution de probabilité possède une densité. En tout cas, on ne suppose pas l'appartenance de la distribution à une famille restreinte de lois à paramètres (d'où le nom) telles que les gaussiennes, par exemple. Bien entendu, cet affaiblissement des hypothèses (c'est-à-dire de (M)) s'accompagne généralement d'une diminution de la puissance du test, et les tests paramétriques, lorsqu'ils sont applicables, sont préférables.

En revanche, les tests paramétriques fournissent en général la possibilité de

travailler avec relativement peu de données[2] (trop peu, en tout cas, pour que l'on puisse valider des hypothèses paramétriques avec un minimum de puissance), et se révèlent relativement robustes. De plus, on est souvent conduit à effectuer des tests non-paramétriques afin de valider des hypothèses paramétriques, qui permettent ensuite l'emploi éventuel de méthodes paramétriques (ex. on valide le caractère gaussien des données par un test non-paramétrique, puis on utilise les méthodes gaussiennes).

Exemples classiques : Test de Kolmogorov-Smirnov, des sommes de rangs de Wilcoxon, du signe, tests de permutation de Fisher Pitman (randomization test) pour comparer des groupes.

8.5 Le test du χ^2

Illustration à l'aide d'histogrammes et de diagrammes en bâtons.

Il s'agit de l'un des tests d'adéquation les plus employés et les plus flexibles. Cela ne signifie pas qu'il s'agit du meilleur test dans toutes les situations.

La version la plus simple du test du χ^2 correspond à la situation où l'on dispose d'observations pouvant se répartir dans K catégories ou classes différentes $C_1, \ldots, C_K$. (M) consiste à dire que l'on peut modéliser les observations par une suite de variables aléatoires indépendantes et de même loi, avec donc, pour chaque catégorie $i \in \{1, \ldots, K\}$, une probabilité donnée $\mathbb{P}(X \in C_i)$. L'hypothèse (H0) testée est que $\mathbb{P}(X \in C_i) = p_i$ pour tout i, où $p_1, \ldots, p_K$ est une liste de probabilités fixée. L'hypothèse (H1) est donc simplement le fait qu'il existe au moins un indice i pour lequel $\mathbb{P}(X \in C_i) \neq p_i$.

La statistique employée par le test sur N données $x_1, \ldots, x_N$ est la suivante :

$$T(x_1, \ldots, x_N) = \sum_{i=1}^{K} \frac{(N_i - Np_i)^2}{Np_i},$$

où N_i désigne le nombre de x_j se trouvant dans C_i, pour $j \in \{1, \ldots, N\}$.

Sous (H0), la loi de $T(X_1, \ldots, X_N)$ converge, lorsque N tend vers l'infini, vers une loi dite du χ^2 à $K-1$ degrés de liberté. La densité de la loi du χ^2 à p degrés de liberté possède la densité sur $\mathbb{R}_+$ suivante :

$$f_{\chi^2,n}(x) = \frac{1}{2^{n/2}\Gamma(n/2)} e^{-x/2} x^{n/2-1},$$

Γ désignant la fonction Gamma d'Euler définie par

$$\Gamma(a) = \int_0^{+\infty} t^{a-1} e^{-t} dt.$$

Il s'agit en fait de la loi d'une somme de p carrés de variables aléatoires de lois gaussiennes standards.

[2]Le fait de disposer de peu de données n'est pas une simple possibilité théorique. Dès lors que l'obtention de données est coûteuse, en temps ou en moyens, – par exemple si plusieurs jours de simulation numérique sont nécessaires avant d'obtenir une valeur (ex. déchets nucléaires), ou si le coût financier associé à l'expérimentation est importante, ou encore si la collecte de données est particulièrement lourde ou fastidieuse (ex. étiqueter et classer les ramifications d'un arbre) – la quantité de données disponibles peut être réellement faible. Le domaine connu sous le nom de planification des expériences (voir par exemple ? ? ?) étudie diverses méthodes permettant de rendre le plus informatif possible pour le but poursuivi un nombre limité d'expériences.

La région critique choisie est en général $[t_\alpha, +\infty)$, où t_α est défini par $\int_{t_\alpha}^{+\infty} f_{\chi^2,n}(u)du = \alpha$.

Etant donnée une classe C_i, $N_i - Np_i$ représente l'écart entre le nombre de valeurs effectivement tombées dans C_i et l'espérance de ce nombre sous l'hypothèse (H0). En effet, sous (H0), N_i suit une loi binomiale de paramètres N et p_i.

On s'attend donc, en vertu du TCL, à ce que l'écart entre N_i et Np_i soit, sous l'hypothèse H_0, de l'ordre de grandeur de $N^{1/2}$, et la normalisation effectuée dans le calcul de T laisse donc à penser que T prend effectivement des valeurs d'ordre fini lorsque N tend vers l'infini (rappelez-vous que K, lui, est fixé). Inversement, sous $(H1)$, il est clair que T doit tendre vers l'infini lorsque N tend vers l'infini.

Plus précisément, lorsque N tend vers l'infini, on voit que la loi de $\frac{(N_i - Np_i)^2}{Np_i}$ doit converger lorsque N tend vers l'infini vers la loi du carré d'une variable aléatoire gaussienne de paramètres 0 et de variance $1 - p_i$. Du fait que les N_i ne sont pas indépendantes (on a en particulier le fait que $N_1 + \cdots + N_K = 1$), on ne peut pas directement déduire de cette observation la loi de T. Pour cela, il est nécessaire de faire appel au théorème de la limite centrale multidimensionnelle, en observant que les vecteurs aléatoires $V_1, \ldots, V_N$ définis par $V_j = (\mathbf{1}_{C_1}(X_j), \ldots, \mathbf{1}_{C_K}(X_j))$ forment une famille indépendante et identiquement distribuée, et que T s'exprime complètement en fonction de $V_1 + \ldots + V_N$. Le TCL multidimensionnel affirme alors que la loi de $V_1 + \ldots + V_N$ correctement normalisée converge vers celle d'une gaussienne multidimensionnelle dont les paramètres peuvent être facilement calculés en fonction des probabilités p_i, et l'on peut donc en déduire la distribution limite de T lorsque N tend vers l'infini.

Une manière naïve de tester (H0) consisterait à appliquer séparément à chaque classe C_i un test de moyenne basé sur le TCL, afin de juger si l'écart entre N_i et Np_i est compatible avec l'hypothèse selon laquelle $\mathbb{P}(X \in C_i) = p_i$. En plus du problème posé par le fait d'effectuer un nombre possiblement grand de tests sur les mêmes données, il arrive que des tests pratiqués individuellement sur chaque classe ne soient pas en mesure de mettre en évidence un écart pourtant significatif détecté par le test du χ^2 entre la répartition observée et ce qu'entraînerait (H0). La statistique employée par le test du χ^2 est une mesure globale de cet écart qui prend simultanément en compte toutes les classes.

La forme de la zone de rejet : la plupart du temps, on rejette (H0) lorsque$T$ dépasse une certaine valeur, mais pas lorsqu'il est proche de zéro, se comprend assez bien : plus la distribution des données s'écarte de celle qui est supposée dans (H0), plus T doit prendre de grandes valeurs. Il paraît donc difficile de faire mieux. Cependant, il est clair que des valeurs anormalement faibles de T peuvent aussi être considérées comme surprenantes. Par exemple, obtenir exactement 500 pile au cours de 1000 lancers d'une pièce de monnaie peut inciter à douter du caractère indépendant des lancers, puisque l'on s'attend à observer des fluctuations typiques autour de la valeur 500 de l'ordre de 15 lancers, en plus ou en moins. Par exemple, si l'on désire tester la qualité d'un procédé de génération informatique de nombres pseudo-aléatoires, on sera amené à utiliser un test du χ^2 possédant une zone critique de la forme $[0, t_1] \cup [t_2, +\infty[$. Dans ce cas, l'hypothèse d'indépendance ne figure plus dans (M), mais dans (H0).

Le test du χ^2 est un test asymptotique basé sur le TCL. Les recommanda-

tions usuelles concernant son utilisation visent à éviter les situations où l'approximation correspondante (du fait que le nombre d'observations est fini) est trop sérieusement prise en défaut (des bornes quantitatives sur l'erreur d'approximation commise peuvent théoriquement être obtenues à partir des bornes sur l'erreur commise dans le TCL, mais elles ne semblent pas être utilisées en pratique). En général, il est au minimum recommandé d'avoir $Np_i \geq 5$ pour tout i, et il est bien entendu nécessaire d'avoir $N_i > 0$ pour tout i. Il est éventuellement possible de regrouper entre elles des classes afin de satisfaire ces hypothèses.

Il est également possible de pratiquer des tests exacts en utilisant la statistique T du χ^2. Les calculs numériques correspondants sont en général beaucoup plus lourds.

Lorsque l'on utilise le test du χ^2 en ayant estimé les p_i sur la base des données, on ne peut plus s'attendre à ce que la statistique T suive une loi du χ^2 à $K-1$ degrés de liberté. Pour s'en convaincre, il suffit de tenir compte du fait que, génériquement, l'erreur d'estimation commise sur p_i est en $N^{-1/2}$, ce qui se traduit par une erreur de l'ordre de l'unité sur T : l'impact de l'utilisation de paramètres estimés ne disparaît pas lorsque N tend vers l'infini.

Lorsque les paramètres sont estimés par la méthode du maximum de vraisemblance, et sous des hypothèses génériques (voir par exemple Dacunha-Castelle tome 2), la distribution de T est encore une distribution du χ^2, mais à $N-1-L$ degrés de liberté, L désignant le nombre de paramètres libres ayant été estimés. Lorsque ces hypothèses ne sont pas complètement vérifiées, la distribution correcte peut se trouver quelque part entre une loi du χ^2 à $N-1$ degrés de liberté et une loi du χ^2 à $N-1-L$ degrés de liberté, et le test perd de sa fiabilité.

On peut utiliser le test du χ^2 pour des variables aléatoires continues, à condition de partitionner en un nombre fini de classes leur domaine de valeurs.

8.6 Les tests reposant sur l'ordonnancement des données

KS, Wilcoxon somme des rangs

8.7 Tests de normalité

Kolmogorov-Smirnov, Shapiro-Wilks, Chi deux, test basé sur le normal probability plot,... Illustration par des histogrammes et d'autres représentations graphiques. Problème du test du chi-deux.

mettre des réf.

Thode, Henry Testing for normality. Statistics : Textbooks and Monographs, 164. Marcel Dekker, Inc., New York, 2002. x+479 pp. ISBN 0-8247-9613-6

8.8 Tests multiples

Il est courant de pratiquer plusieurs tests sur le même échantillon de données, soit que l'on veuille tester des propriétés différentes (par exemple, d'abord valider une hypothèse concernant la forme des lois, puis utiliser un test paramétrique reposant sur cette hypothèse), soit que l'on veuille tester la même propriété par plusieurs moyens différents, pour plus de certitude (différents tests de la même propriété testant en général des conséquences différentes, et pouvant donc être sensibles à des écarts différents vis-à-vis de l'hypothèse testée).

Pour juger de la qualité (sensibilité et spécificité) des résultats ainsi obtenus, il est nécessaire de voir l'ensemble de ces tests comme une procédure de test globale. En effet, si l'on pratique un grand nombre de tests sur les mêmes données, il est raisonnable de s'attendre à ce qu'un certain nombre d'entre eux conduisent à rejeter (H0), même si celle-ci est valide, du simple fait du hasard. Par exemple, si l'on teste, dans un vaste ensemble de données, l'association entre un grand nombre de caractères pris deux-à-deux, il est très probable que l'un des tests finisse par détecter une association significative entre une certaine paire de caractères, du simple fait du hasard. Présenter ce test isolément, sans mentionner les autres tests menés, constitue un exemple simultané de biais de sélection, et un test d'hypothèses suggérées par les données. On nomme en anglais «data dredging» la pratique qui consiste à explorer systématiquement, au sein d'un jeu de données, les relations pouvant apparaître comme significatives, et à ne présenter que celles-ci. Il s'agit au mieux d'une grave erreur de méthodologie, et au pire d'une attitude malhonnête, et pourtant difficile à déceler (pour vérifier l'absence de ce type de pratique, il faudrait que soient définie et déposée à l'avance une liste précisément définie des relations qui vont être recherchées). De manière générale, les conclusions obtenues à partir de tests statistiques de significativité demandent à être appuyées par des études de confirmation menées sur des données indépendantes de celles sur lequel le test a été initialement pratiqué. Lorsque l'on se trouve en présence de plusieurs études effectuées indépendamment, ayant chacune produit des résultats de tests (certaines pouvant avoir donné lieu à des tests rejetant (H0) et à d'autres ne la contredisant pas), il est courant de tenter de regrouper l'ensemble des données recueillies dans ces études, afin de mener un test portant sur cet ensemble de données, ce qui constitue un exemple simple de méta-analyse. Cependant, les conditions dans lesquelles les différentes études sont menées (concernant l'échantillonnage, le protocole expérimental suivi, les quantités ou caractères mesurés) ne sont pas toujours comparables, surtout si elles portent sur un domaine nouveau dans lequel le degré de standardisation est faible, ce qui rend la tâche complexe.

Si l'on cherche à tenir compte correctement du fait que l'on pratique des tests multiples sur un même jeu de données, se pose le problème du fait que les résultats obtenus à l'aide de procédures de tests différentes ne sont pas en général indépendantes, hypothèse qui permettrait pourtant d'évaluer facilement le risque de première espèce associé globalement à l'ensemble des tests pratiqués (c'est-à-dire la probabilité de rejeter à tort au moins l'une des hypothèses (H0) testées). Au pire, on peut avoir recours à la borne de la réunion (voir le chapitre sur les coincidences), qui conduit à estimer que le risque de première espèce associé à un ensemble de tests est inférieur à la somme des risques de première espèce associés à chacun des tests pris séparément, mais on voit facilement que ceci conduit rapidement (dès que le nombre de tests pratiqués est élevé) à des risques de première espèce très élevés, ou à des tests individuels particulièrement peu puissants. Lorsque c'est possible, on préfère donc employer une unique procédure de test qui prend en compte l'ensemble des aspects de l'hypothèse que l'on désire tester, plutôt qu'une série de tests individuels.

8.9 Dépendance de la procédure de test vis-à-vis des données

Nous l'avons vu dans le chapitre sur les coïncidences, il convient d'être méfiant vis-à-vis des tests portant sur des événements construits a posteriori sur les données obtenues. Une erreur en rapport avec cette idée consiste à appliquer une procédure de test dans laquelle (H0) suppose donnés certains paramètres, en estimant ces paramètres sur la base des données et en faisant comme si de rien n'était. Par exemple, le test de Kolmogorov-Smirnov fournit une (pas nécessairement la meilleure) méthode pour tester si un échantillon provient d'une loi normale de paramètres m et v fixés. Si l'on pratique cette procédure en estimant m et v sur la base des données dont on dispose, la conséquence (C) que l'on teste se met elle-même à dépendre des données, ce qui fausse a priori le résultat du test, le niveau de confiance et la puissance étant modifiés de manière parfois incontrôlée (soit que l'on soit plus facilement conduit à rejeter (H0) qu'il ne faudrait, soit que le test conduise à accepter exagérément (H0)). Il est en général nécessaire de tenir compte de la modification apportée par l'estimation des paramètres, en analysant en tant que telle la procédure de test que l'on applique (et non pas en faisant semblant de connaître exactement les valeurs des paramètres alors qu'on les a seulement estimées), et l'on peut obtenir des résultats complètement différents de ce que donnerait l'application (incorrecte) de la procédure initiale avec des paramètres estimés. Le test de Student donne un exemple de situation où l'estimation d'un paramètre (en l'occurrence la variance) modifie la distribution de la statistique employée. Cependant, cette modification disparaît dans la limite où le nombre d'observations est grand. Dans le cas du test du χ^2, par exemple, le problème subsiste même dans la limite d'un grand nombre d'observations. Ce problème se manifeste également lorsque l'on teste la validité d'un modèle comportant un grand nombre de paramètres susceptibles de l'ajuster aux données, en estimant ces paramètres à partir de ce jeu de données (voir le chapitre «Validation et choix de modèles»).

8.9.1 Robustesse

Voir robustesse.pdf et l'auteur

Une question importante en pratique est de déterminer à quel point les procédures de test employées sont robustes vis-à-vis de leurs hypothèses de validité (par exemple, à la forme exacte des lois de probabilité pour des tests paramétriques, ou à l'indépendance pour des tests non-paramétriques), puisque celles-ci ne peuvent pas toujours être établies avec beaucoup de fermeté (notamment l'indépendance, très souvent supposée, mais rarement et difficilement vérifiée). Ainsi, une procédure a priori moins puissante qu'une autre, mais plus robuste, pourra parfois être préférée, ou en tout cas mériter d'être prise en considération à titre de vérification. On distingue en général deux notions de robustesse :

- la robustesse du test en termes de validité, c'est-à-dire le fait que son niveau de confiance soit robuste,
- la robustesse du test en termes d'efficacité, c'est-à-dire le fait que son niveau de puissance soit robuste.

Il est difficile d'aborder la question de la robustesse en toute généralité. Des études théoriques et empiriques étudient la robustesse des procédures de tests vis-à-vis de certains types de perturbation des hypothèses. Tout dépend du test

considéré et du type de modification pris en compte. Il faut se documenter !

8.10 Mises en garde

Voici un ensemble de mises en garde générales et importantes concernant les tests statistiques et leur interprétation.

8.10.1 Les conditions de validité

Plusieurs hypothèses sont en général nécessaires pour assurer la validité d'une procédure de test (entre autres ce que nous avons appelé (M)). La plupart du temps, on suppose ainsi que l'on dispose d'un échantillon de données pouvant être modélisé par une suite de variables aléatoires indépendantes et identiquement distribuées. De nombreuses procédures de test reposent également sur des approximations qui ne sont valables que dans la limite d'un grand nombre de données collectées, la qualité de ces approximations pouvant dépendre d'hypothèses supplémentaires sur le modèle décrivant les données. Enfin, des hypothèseses plus ou moins restrictives sur la distribution de probabilité décrivant les données (par exemple, le fait que les données puissent être décrites par une famille relativement restreinte de lois de probabilité, par exemple des gaussiennes) sont souvent employées.

Il est indispensable de s'assurer que la validité de ces hypothèses a été établie, ou au moins testée, avant d'appliquer une procédure de test. Même si le calcul de la statistique à partir des données et sa comparaison aux valeurs critiques, ou le calcul de la p-valeur peut la plupart du temps être effectué sans que ces hypothèses soient vérifiées, le résultat produit par ces calculs, et a fortiori les conclusions que l'on envisagerait d'en tirer, ne possèdent aucune signification si ces hypothèses de validité ne sont pas satisfaites.

8.10.2 Significativité statistique et significativité pratique

Les procédures de tests statistiques sont basées sur la notion d'écart significatif par rapport à l'hypothèse (H0) : par exemple, lorsque la valeur de la statistique utilisée pour le test dépasse un certain seuil, on considère que ce dépassement est trop improbable si (H0) est valable, pour que l'on puisse attribuer sa survenue à la variabilité «normale» attendue sous l'hypothèse (H0), et l'on parle donc d'un écart significatif par rapport à (H0) (bien sûr, c'est nous qui fixons le seuil de probabilité à partir duquel on décide que l'écart observé est significatif : il s'agit de α). Cependant, un tel écart, pour être significatif au sens statistique du terme, n'a pas forcément de pertinence pratique. Par exemple, si l'on dispose d'un grands nombre de données, de minuscules écarts par rapport à une répartition gaussienne pourront être testés comme significatifs, et même extrêmement significatifs (avec une p-valeur infime) tout en n'ayant guère d'impact sur la validité pratique de la modélisation par une loi gaussienne dans le contexte envisagé. De même, un écart très faible entre les taux de guérison obtenus au moyen de deux traitements (mettons, un augmentation de 0,01 de la probabilité de guérison) pourra être détecté comme significatif si l'on dispose de suffisament de données. Ou encore, une association entre deux facteurs pourra être jugée comme significative alors même que cette association est extrêmement faible (mais pas nulle), et n'est d'aucune valeur prédictive en pratique.

Un écart jugé significatif en statistique ne représente pas forcément un écart devant être pris en considération en pratique compte tenu de son importance.

8.10.3 Des problèmes de traduction

Nous l'avons vu, les hypothèses testées par les procédures de tests statistiques sont en fait des hypothèses de modélisation, formulées dans un cadre probabiliste. Or, les hypothèses que l'on a réellement pour but de tester sont en général formulées en termes beaucoup plus vagues : tel traitement ou procédé est-il plus ou moins efficace que tel autre ? Observe-t-on des différences entre tel et tel comportement ? Tel facteur influe-t-il sur tel autre ? Etc...

Par conséquent, il faut s'assurer que la traduction des questions pratiques que l'on se pose en termes d'hypothèses de modélisation est faite de manière pertinente, sans quoi les résultats des tests pratiqués, pour être éventuellement statistiquement significatifs, ne permettront nullement de répondre à la question initialement posée, et, bien pire, pourront facilement y apporter des réponses totalement erronnées basées sur une illusion d'objectivité statistique.

Or, trop souvent, la vérification de cette étape de traduction est négligée. Plusieurs explications peuvent être invoquées : il s'agit d'une tâche difficile à accomplir, consistant à partir de situation non-formalisées et parfois décrites en termes assez vagues pour parvenir à une formulation mathématique, et qui nécessite donc une bonne compréhension et de la problématique initiale (ou tout au moins une capacité à dialoguer efficacement avec les spécialistes concernés), et de la modélisation probabiliste.

D'autre part, le caractère standardisé et automatique (et l'allure fortement technologique et scientifique) des procédures de tests statistiques et des modèles sur lesquelles elles reposent peut facilement conduire à utiliser sans réflexion préalable ces méthodes (parce que c'est comme ça que l'on fait toujours, parce que c'est standard, parce que c'est fatiguant de réléchir, parce la méthode fournit de toute façon des résultats chiffrés qui ont l'air sérieux, et des réponses tranchées), sans mesurer la fragilité de l'étape de traduction, sur laquelle toute la portée pratique du résultat repose pourtant.

Même une procédure de test effectuée avec beaucoup de soin, dans des conditions de mise en œuvre correctement validées, peut n'apporter aucune information utile pour le problème auquel on s'intéresse si cette question est négligée : la validité de toute la chaîne de la méthode statistique est conditionnée par celle de chacun de ses maillons.

8.10.4 Attention à la puissance

Un autre usage abusif des procédures de tests consiste à accepter comme valables des hypothèses ayant été confirmées par un test dans des conditions où la puissance de celui-ci (c'est-à-dire sa capacité à séparer l'hypothèse testée des hypothèses concurrentes) est très faible, ou inconnue. Par exemple, un test de normalité effectué avec un petit nombre de données rejettera rarement l'hypothèse de distribution gaussienne des données, même si celles-ci ne fournissent aucune raison positive d'accepter cette hypothèse, mais ne peuvent tout simplement pas fournir d'éléments suffisants pour rejeter celle-ci. Utiliser ensuite des techniques gaussiennes pour analyser ces données, en considérant qu'un tel test a validé l'hypothèse de normalité, constitue manifestement une sérieuse erreur

de méthode. Rappelons-nous en tout cas que l'acceptation de (H0) par un test est toujours par défaut : on a tenté de contredire (H0) au moyen d'une procédure de test bien particulière, on n'y est pas parvenu, donc (H0) est acceptée par défaut. (En termes imagés, on n'a pas réussi à prendre (H0) la main dans le sac, donc on présume qu'elle est innocente.) Ce type d'erreur peut être facilité par le fait que l'on a insuffisament examiné les données avant de pratiquer des tests, et par le fait que, si le niveau de confiance du test est connu, la puissance, elle n'est pratiquement jamais connue (ni connaissable), et ne peut donc être indiquée par les sorties automatisées des procédures de test.

De même, lorsque (H1) contient un très vaste ensemble d'hypothèses (par exemple dans les tests d'adéquation), les tests sont en général sensibles à certains aspects seulement de l'hypothèse (H0), et c'est cet aspect qui est réellement testé (par exemple, certains tests de normalité mettront davantage l'accent sur la symétrie, d'autres sur la rapidité de décroissance de la densité,...). Il ne faut pas déduire de l'acceptation de (H0) que les aspects non-testés de (H0) sont en quelque mesure validés.

8.10.5 Le sens de la confiance

Une erreur (grossière, mais néanmoins répandue) dans l'interprétation des résultats de test consiste à interpréter le fait que (H0) a été acceptée (ou rejetée) par un test de niveau de confiance 95% comme le fait que (H0) a donc 95% de chances d'être vraie (ou fausse) d'après le résultat du test. Il n'y a dans le principe des tests aucune probabilité attachée à la validité de (H0) ou de ses concurrentes : soit (H0) est valide, soit elle ne l'est pas, et la procédure de test est telle que, si (H0) est vraie, le test l'acceptera avec une probabilité de 95%. Le test constitue une procédure dont la qualité est jugée dans son ensemble, et mesurée par son niveau de confiance et sa puissance, et la crédibilité du résultat obtenu par cette procédure est bien entendu liée à ces mesures de qualité. Mais en aucun cas on ne juge de la probabilité pour que (H0) soit ou non valable ou fausse. Le raisonnement bayésien fournit un moyen d'aborder correctement cette question, en utilisant les résultats de test pour ajuster des probabilités attribuées a priori aux différentes hypothèses, mais ce n'est pas ce cadre de raisonnement que nous décrivons ici.

8.10.6 Un test ne constitue pas une preuve

Ni lorsqu'il conduit à rejeter l'hypothèse (H0), et encore moins lorsqu'il conduit à ne pas la rejeter, un résultat de test ne constitue une preuve en faveur ou en défaveur de (H0), mais tout au plus un indice pouvant être utilisé dans le raisonnement. Imaginons par exemple que 20 équipes scientifiques étudient séparément l'impact d'un nouveau produit, disons la vitamine X, sur la guérison d'une maladie, par exemple le cancer. Chaque équipe conduit son étude dans les règles (essais randomisés en double aveugle, échantillons représentatifs de la population à traiter, constitution de groupes témoins et utilisation de placebos, voir partie... pour une discussion de ces termes). Sur les 50 équipes, 49 observent des résultats non-concluants quant à l'efficacité du médicament. En revanche l'une des équipes observe un taux de guérison si élevé chez les patients traités à l'aide de la vitamine X, que, sous l'hypothèse que la vitamine X est sans effet sur le cancer, on ne puisse espérer observer un tel taux qu'avec une probabilité

d'environ 5%. Autrement dit, une p-valeur de 5%. L'équipe en question doit-elle conclure qu'elle a prouvé l'efficacité de la vitamine X dans le traitement du cancer ? Certainement pas. En s'informant des résultats obtenus par les autres équipes, ou en tentant de confirmer par des études supplémentaires la validité de ses conclusions, elle devrait rapidement s'apercevoir que son résultat est plus vraisemblablement le résultat d'une fluctuation aléatoire que le signe d'un effet véritable. De fait, sur 20 équipes, il n'est pas surprenant que l'une d'entre elles observe des résultats significatifs avec un niveau de confiance fixé à 5%. Il est même normal qu'une procédure de test possédant un niveau de confiance de $1 - \alpha$ conduise à rejeter (H0) à tort environ α% du temps au cours de tests répétés. Sans parler des erreurs de seconde espèce. (Au passage, le fait que des résultats non-significatifs obtenus soient souvent non-publiés contribue à fausser considérablement le jugement que l'on peut porter sur les études publiées car ayant obtenu des résultats statistiquement significatifs ; il s'agit du phénomène connu sous le nom de biais de publication, voir la section consacrée aux phénomènes de sélection). Le même type de phénomène peut se produire avec le data dredging, intentionnel ou involontaire. Plus généralement, il est difficile d'exclure complètement la dépendance entre hypothèse testée et données (par exemple, c'est souvent un examen exploratoire des données qui suggère la forme des lois que l'on teste et ajuste ensuite). Un sain principe est que, de manière générale, les résultats obtenus par des tests demandent à être confirmés par des études indépendantes.

8.10.7 Attention à la généralisation des conclusions

Il faut toujours garder à l'esprit le fait que la possibilité de généraliser les résultats obtenus à l'aide d'un test statistique effectué sur un échantillon de données à un contexte plus vaste que celui dans lequel il a été pratiqué repose sur la représentativité de l'échantillon choisi vis-à-vis de ce contexte (ceci fait généralement partie de (M), et n'est donc pas remis en question par le test). La validité ou l'efficacité de la procédure de test employée (que les hypothèses nécessaires à sa mise en œuvre soient vérifiées) n'a rien à voir avec ce type de question.

8.10.8 Les pièges de la p-valeur

Rappelons que la p-valeur associée à un test, fournie par la plupart des logiciels de statistique, n'est pas la probabilité pour que l'hypothèse nulle soit vérifiée (lorsque la p-valeur est faible) ou la probabilité pour que celle-ci soit fausse (lorsque la p-valeur est proche de 1). Celle-ci fournit simplement une mesure (parmi une foule d'autres possibles) du caractère surprenant des données sous l'hypothèse (H0). Il est de bonne pratique de fixer à l'avance la valeur de référence à laquelle on va comparer la p-valeur pour décider du résultat du test (c'est-à-dire de α) plutôt que de le fixer a posteriori, ce qui revient à faire dépendre des données la conséquence que l'on teste.

Mentionnons de plus qu'il n'y a rien d'absolu dans le choix de 95% comme niveau de confiance, qui s'est imposé depuis plusieurs années, notamment dans le domaine bio-médical. Tout dépend du contexte, et des conséquences des deux types d'erreur que l'on peut commettre.

8.11 Position des tests statistiques

Les procédures de tests statistiques constituent l'un des éléments de la démarche statistique. Elles permettent de juger quantitativement de la compatibilité entre des hypothèses de modélisation et les données observées, en prenant en compte de manière rigoureuse la variabilité inhérente à la situation considérée. La plupart du temps, les tests statistiques sont employés pour donner une base quantitative à des conclusions suggérées par une analyse exploratoire (graphique) des données recueillies et/ou par une connaissance partielle déjà constituée ou une analyse a priori du phénomène étudié. Il faut garder à l'esprit que la validité de ceux-ci repose en général sur des hypothèses très fortes (telles que la description par des variables aléatoires indépendantes et identiquement distibuées des données recueillies) dont la validité ne peut la plupart du temps être admise qu'en première approximation, et qu'il ne faut donc prendre les résultats produits par les tests statistiques que comme des mesures d'accord (ou de désaccord) entre hypothèses et données obtenues en se plaçant dans une situation assez fortement idéalisée. Plutôt que dans le but d'obtenir une réponse tranchée (hypothèse acceptée ou rejetée de manière stricte), on utilise souvent les tests pour attribuer des scores (par exemples calculés à partir des p-valeurs) à des hypothèses de modélisation dont on admet qu'elles ne peuvent qu'être approximativement vérifiées.

Nombreuses références. Une bonne référence disponible sur le réseau est, par exemple, le «Handbook of Engineering Statistics» du National Institute of Standards and Technology américain, que vous trouverez à l'adresse `http://www.itl.nist.gov/div898/handbook/`

IL MANQUE DANS CE CHAPITRE : - détailler les exemples de tests - donner des exemples d'utilisation concrète pour illustrer la démarche, ses limites, et les erreurs que l'on peut commettre - l'utilisation des simulations pour calibrer des tests (exple. simulation à partir de valeurs estimées sur les données) - commentaires sur la quantité d'informations incorporées (notamment par des tests non param comme celui du signe) - biblio sur plans d'expériences, tests en général, (théorique), tests bayésiens, tests non-param., méthodologie statistique - exemple de Knuth - le cas où les données ne sont pas indépendantes : par exemple des résidus d'ANOVA - commentaire général intelligent sur le fait que les tests que l'on pratique (par exple normalité) sont suggérés par l'examen des données - des exemples de sortie en provenance de R, SAS, etc... - insister sur hypothèse simple (qui détermine la loi des observartions complètement)/ hypothèse composée : qui comprend plusieurs lois possibles, et pour laquelle on peut être amené à estimer des param. exple : suivre une gaissienne (2,3) ou suivre une gaussienne. C'est pas pareil !!! - la robustesse vis-à-vis des hypothèses (à préciser) - la robustesse vis-à-vis des données (données aberrantes, par exple. , ou en tout cas le fait que peu de données concentrent la réponse, ou que la réponse puisse varier si l'on change un peu les données) est-ce la même robustesse que ci-dessus ? ? ? bibli ? validation croisée, bootstrap, jacknife,... (dernier chap. de livre analyse exploratoire des données) Le pb. est)il le même que la robustesse pour l'apprentissage et l'estimation ? (lien avec la puissance des tests ?)

9 Questions d'échantillonnage

L'analyse statistique repose sur la confrontation entre d'une part des données mesurées, et d'autre part des modèles probabilistes censés décrire la situation étudiée et la façon dont les données disponibles sont reliées à celui-ci. Une question fondamentale est donc celle de la représentativité de l'échantillon de données utilisé vis-à-vis de la situation considérée : dans quelle mesure les données obtenues permettent-elles de tirer des conclusions fiables, et de portée générale, sur cette situation ?

Dans cette partie, nous nous référerons préférentiellement aux questions relatives à l'échantillonnage au sein de population d'individus, mais la discussion est de portée beaucoup plus générale et concerne globalement toute analyse statistique fondée sur des échantillons de données mesurées.

De fait, la notion de ce qu'est un échantillon représentatif, par exemple dans le domaine des sondages au sein d'une population d'individus, est difficile à cerner, et le terme de représentativité est utilisé (parfois à tort et à travers) sous des acceptions très diverses. Parmi celles-ci :

- rien de précis, l'utilisation du mot étant purement rhétorique et destinée à décourager les critiques portant sur l'étude menée ou les interprétations qu'elle contient ;
- le fait que l'échantillon de données considéré soit de taille importante ;
- le fait que l'échantillon est une miniature de la population étudiée ;
- le fait que divers quotas sont respectés par la constitution de l'échantillon ;
- le fait que des cas considérés comme typiques se trouvent dans l'échantillon ;
- l'absence supposée de mécanismes conduisant à sélectionner certains types de données plutôt que d'autres ;
- le fait que les données soient issues d'un sondage aléatoire ;
- le fait que tous les cas (lorsqu'il n'existe qu'un nombre limité de modalités envisageables) soient représentés dans l'échantillon.

Pour citer Y. Tillé, «Voir invoquée la «représentativité» dans un rapport d'enquête pour justifier de la qualité d'un sondage peut presque à coup sûr laisse soupçonner que l'étude a été réalisée dans une méconnaissance totale de la théorie de l'échantillonnage. Le concept de représentativité est aujourd'hui à ce point galvaudé qu'il est désormais porteur de nombreuses ambivalences. Cette notion, d'ordre essentiellement intuitif, est non seulement sommaire mais encore fausse et, à bien des égards, invalidée par la théorie.»

Avant même de parler de l'éventuelle représentativité d'un échantillon, il est nécessaire de se demander de quelle population celui-ci est censé être représentatif, – la population ciblée par l'étude menée – et de définir celle-ci de manière non-ambiguë, ou tout au moins avec suffisament de précision pour que les résultats auxquels on s'intéresse ne soient pas affectés par les légères imprécisions qui subsistent, ce qui n'est pas forcément évident. Par exemple, on peut imaginer toute sorte de définitions de «la population de la France en 2005.» Faut-il inclure y les touristes, les personnes installées depuis peu, les personnes nées ou décédées au cours de 2005, etc... ?

Qui plus est, la population au sein de laquelle on collecte des données n'est pas forcément celle sur laquelle on souhaite produire des conclusions, et nous avons donc deux problèmes de définition distincts. Par exemple, on cherchera

à partir d'une enquête sur les personnes ayant vécu en France au cours des 20 dernières années à étudier l'impact de la consommation de tabac sur la santé, dans le but d'en tirer des conclusions ou des recommandations de santé publique concernant les populations des 10 ou 20 années suivantes, qui vivront pourtant dans des conditions (économiques, environnementales, sociales) globalement différentes. Une étude sur les choix de consommation des ménages menée en période de croissance économique ne sera pas forcément pertinente pour décrire le comportement en période de difficultés économiques. De manière plus triviale, une étude sur l'utilisation des véhicules personnels menée au cours des mois de juillet et août ne fournira pas forcément d'informations fiables sur leur utilisation tout au long de l'année. Le fait que de telles extrapolations soient possible repose nécessairement sur des hypothèses, qui doivent être explicitées et autant que possible validées.

Nous nous penchons dans cette partie sur l'extraction à partir d'échantillons d'informations sur la population échantillonnée elle-même. Soulignons néanmoins que la rigueur et la qualité de cette étape d'échantillonnage ne garantit en rien la possibilité d'extrapoler ces informations à une population différente (même si le fait de disposer d'informations fiables sur la population échantillonnée est évidemment important pour pouvoir aborder dans de bonnes conditions une telle extrapolation).

Nous supposerons donc que l'on a défini précisément une population-cible représentée par un espace des possibles Ω, dont les éléments représentent donc les individus, et nous poserons le problème sous la forme suivante : évaluer l'espérance d'une variable aléatoire f définie sur un modèle probabiliste Ω par rapport à la probabilité uniforme $\mathbb{P}$, à partir d'un échantillon d'éléments de Ω, $\omega_1, \ldots, \omega_N$.

La manière la plus évidente de procéder serait de procéder à un recensement complet de Ω, mais, pour des raisons de coût et de temps, cela est la plupart du temps impossible, et c'est pourquoi on ne considère en général que des échantillons de taille relativement faible (souvent quelques milliers d'individus). Qui plus est, il peut être plus facile de réaliser un sondage dans de bonnes conditions qu'un recensement, et un bon sondage peut s'avérer plus précis qu'un recensement de mauvaise qualité.

9.1 Echantillonnage aléatoire

La seconde manière de procéder qui nous est aujourd'hui familière, mais est loin d'avoir toujours été acceptée, est de procéder un sondage aléatoire, en choisissant un échantillon d'éléments de Ω de manière indépendante et uniforme (lorsque la population examinée est suffisament grande, la question de savoir si l'on sonde avec ou sans remise peut être négligée). On peut alors appliquer la théorie de l'estimation, qui fournit des fourchettes d'erreur permettant de prendre correctement en compte la variabilité liée au caractère aléatoire de l'échantillonnage – la précision des estimations n'étant limitée que par la taille de l'échantillon sondé – et l'on peut donc extraire des renseignements fiables des données collectées. En ce sens, un échantillon obtenu de cette manière est représentatif de la population dont il est extrait.

Plus formellement, l'estimateur de $\mathbb{E}(f)$ donné par

$$\frac{1}{N}\sum_{i=1}^{N} f(\omega_i)$$

est sans biais et sa variance peut être estimée (sans biais) par la variance empirique

$$\mathbb{V}(f) = \frac{1}{N-1}\sum_{i=1}^{N}\left[f(\omega_i) - \frac{1}{N}\sum_{j=1}^{N} f(\omega_j)\right]^2 .$$

Pour procéder à ce type d'échantillonnage, le plus simple est, si c'est possible, de partir d'une liste exhaustive des éléments de Ω, comportant des indications permettant de contacter chaque personne y figurant (et éventuellement certaines informations supplémentaires), et de tirer, à l'aide d'une procédure aléatoire a priori non-susceptible d'interagir avec les caractéristiques des éléments de Ω, des éléments de la liste de manière i.i.d. et uniforme.

Même dans ce contexte simple, l'échantillonnage par tirages i.i.d. uniformes ne constitue pas forcément la meilleure solution, au sens où il est possible, si l'on dispose d'informations supplémentaires, d'améliorer la précision des estimations obtenues.

Pour des détails sur les méthodes qui suivent, et notamment l'estimation de variance, nous renvoyons à l'ouvrage d'Y. Tillé cité en bibliographie.

9.1.1 Echantillonnage d'importance

Si l'on effectue des tirages sous une probabilité $\mathbb{Q}$ sur Ω telle que $\mathbb{Q}(\omega) > 0$ pour tout ω, en notant qu'estimer $\mathbb{E}_{\mathbb{P}}(f)$ revient à estimer $\mathbb{E}_{\mathbb{Q}}(f/\mathbb{Q})$, on constate que l'on peut estimer $\mathbb{E}_{\mathbb{P}}(f)$ par

$$\frac{1}{N}\sum_{i=1}^{N} f(\omega_i)/\mathbb{Q}(\omega_i)$$

dans le cadre de la théorie précédente. Si $\mathbb{Q}$ est convenablement reliée à f, on peut parvenir à réduire de manière importante la variance des estimations obtenues. En ce sens, on peut obtenir des échantillons plus «représentatifs» (qui conduisent à des estimations plus précises) que ceux produits par un tirage uniforme, en sélectionnant différemment les points d'échantillonnage.

Bien entendu, pouvoir effectuer un tirage selon la probabilité $\mathbb{Q}$ suppose, dans ces conditions, de disposer à l'avance d'informations sur les individus figurant dans la liste.

9.1.2 Stratification

Si la population étudiée peut se décomposer en strates $C_1, \ldots, C_p$, il peut être avantageux d'échantillonner séparément chacune des strates de manière indépendante, en choisissant le nombre de points d'échantillonnage alloués à chaque strate autrement qu'en proportion de leur taille (ce que fait approximativement un tirage uniforme), de manière à concentrer davantage de points dans les strates où la variabilité est plus importante, ou même simplement en allouant

à chaque strate un nombre de points exactement proportionnel à sa taille, ce qui permet d'éliminer la variabilité liée aux fluctuations aléatoires des nombres de points alloués lors d'un tirage uniforme. Si l'on tire N_i points $\omega_1^i, \ldots, \omega_{N_i}^i$ dans la strate numéro i selon la probabilité $\mathbb{P}(\cdot|C_i)$, l'estimateur obtenu sera

$$\sum_{i=0}^{p} \mathbb{P}(C_i) \frac{1}{N_i} \sum_{j=1}^{N_i} f(\omega_j^i).$$

Ici encore, on constate que l'on peut dans certains obtenir des échantillons plus «représentatifs» que des échantillons résultant d'un tirage uniforme, la composition des échantillons obtenus ne mimant pourtant pas la composition de la population initiale. Bien entendu, tout ceci nécessite que des informations supplémentaires soient disponibles dans la liste utilisée pour l'échantillonnage.

Exercice 22 *Pour une décomposition en strates données, calculez la répartition des N_i permettant de réduire à son minimum la variance des estimations. Quels sont les paramètres qui interviennent, et comment pourrait-on les estimer ?*

Il est également possible de procéder à une stratification a posteriori de la population, notamment lorsque les informations disponibles sur la décomposition en strates sont disponibles globalement, mais pas a priori dans la liste utilisée pour l'échantillonnage. Par exemple, on peut savoir que la proportion d'hommes et de femmes est globalement de 50% dans la population étudiée, mais ne pas savoir a priori pour un individu de la liste s'il s'agit d'un homme ou d'une femme. On pourra alors gagner en précision en repondérant l'échantillon de valeurs mesurées en tenant compte de cette stratification.

On peut bien entendu combiner stratification et échantillonnage d'importance.

9.1.3 Comment réaliser en pratique ce type d'échantillonnage ?

A priori, il suffirait de disposer d'une liste exhaustive des éléments de Ω, comportant des indications permettant de contacter chaque personne y figurant, et de tirer, à l'aide d'une procédure aléatoire des éléments de la liste, de manière i.i.d. et uniforme dans le cas le plus simple, ou selon les procédés plus complexes nécessités par l'échantillonnage d'importance ou la stratification.

Une très sérieuse limitation est que l'on ne dispose la plupart du temps d'aucune liste de ce type, soit que l'existence même d'une telle liste ou son accès soit interdite pour des raisons légales, soit qu'il soit impossible en pratique (ou très coûteux) de la constituer ou d'y accéder. Il n'est en général pas non plus possible de disposer d'une sur-liste suffisamment restreinte de la population visée pour rendre praticable un tirage accompagné de rejets.

Parfois, on dispose effectivement de listes, mais à des échelles plus petites que celle de la totalité de la population étudiée. Par exemple, il peut être inenvisageable de disposer d'une liste centralisant la totalité des élèves de l'enseignement primaire. En revanche, chaque école primaire dispose d'une liste comportant les coordonnées de ses élèves, et l'on dispose d'une liste des différents départements, dans chaque département, d'une liste des différentes communes, et, dans chaque commune, d'une liste des différentes écoles.

On peut ainsi échantillonner la population visée en procédant à des tirages aléatoires successifs à plusieurs niveaux, par exemple en choisissant uniformément dans chacune des listes un département, puis une ville, puis une école, puis un élève. Plus souvent, on échantillonne par grappes en choisissant par exemple un échantillon d'écoles au sein du même département, et un échantillon d'élèves au sein des écoles choisies, ce qui réduit généralement le coût de la collecte (si un enquêteur doit se déplacer sur place, par exemple). Il est possible de tenir compte des dépendances et de la non-uniformité introduites par ces procédures pour fabriquer des estimateurs dont les propriétés essentielles (biais et variance) peuvent être étudiées. Voir par exemple l'ouvrage de Tillé.

9.1.4 Des problèmes incontournables

Même dans le contexte, le plus satisfaisant scientifiquement, d'un sondage aléatoire, les questions de fiabilité des données obtenues, que ce soit au moment de la saisie, de la transmission, de la mesure (lorsque les questions posées, sont vagues ou mal comprises, ou tendancieuses, ou font appel à la mémoire, ou lorsque les personnes interrogées ont des raisons de ne pas répondre correctement, lorsque les sondeurs sont négligents, lorsque l'on fait appel à des appareils de mesure,...), ou encore, et les problèmes de non-réponse, doivent être abordés. Ceux-ci ne sont pas directement liés à l'échantillonnage, mais il est indispensable de les aborder, sous peine d'invalider l'étude menée. Le soin et/ou le raffinement technique apportés à la procédure d'échantillonnage ne pourront rien y changer.

9.2 Autres procédés d'échantillonnage

Les procédés de sondage aléatoire décrits ci-dessus sont les seuls pour lesquels existe une théorie mathématique bien établie permettant de contrôler les erreurs d'estimation commises. Dès que l'on quitte ce cadre, la validité des procédures d'échantillonnage employées devient dans le meilleur des cas soumise à des hypothèses qu'il faut absolument expliciter, et autant que possible valider. Si cela s'avère impossible, la plus grande prudence s'impose dans la manipulation des données ainsi obtenues. Quoiqu'il en soit, il est fondamental, face à des données échantillonnées, de savoir de quel protocole d'échantillonnage elles résultent, afin de ne pas prêter à tort à des données la valeur – liée à la possibilité d'utiliser la théorie mathématique sous-jacente – que seul un véritable échantillonnage aléatoire peut conférer (ce qui ne signifie pas que des données issues d'un échantillonnage non-aléatoire soient sans valeur ; simplement, on ne peut pas prétendre les utiliser de la même façon et avec la même crédibilité).

Les raisons qui font que l'on ne dispose pas d'un échantillon issu d'une procédure aléatoire peuvent être diverses, mais se ramènent souvent à l'impossibilité d'établir une liste (même à plusieurs niveaux) de la population ciblée, par exemple parce que les informations qui permettraient de l'établir ne sont pas accessibles, ou n'existent pas, ou seraient illégales à enregistrer ou encore trop lourdes à collecter... La volonté délibérée de pouvoir manipuler les chiffres produits en cas de besoin a également sa place.

L'une des méthodes couramment employées par les instituts de sondage privés est la méthode des quotas : on utilise un échantillon dont la seule contrainte est qu'il doit contenir certains quotas d'individus correspondant à certaines spécifications (tels que l'âge, le sexe, la catégorie socio-professionnelle, etc...), ce

qui est censé assurer sa «représentativité» vis-à-vis de la population ciblée. La conception de l'étude repose sur un «choix raisonné» des individus constituant l'échantillon par les sondeurs, dont l'expérience professionnelle et l'éthique sont mises à contribution. Il s'agit de toute façon d'une méthode empirique dans la mesure ou aucune évaluation théorique du biais et de la variance n'est possible.

Une autre source de données fréquemment disponible résulte de ce que l'on appelle génériquement des enquêtes d'observation, dans lesquelles on s'efforce de collecter l'ensemble, voire seulement certaines, des données disponibles dans un certain contexte sur un certain sujet (par exemple, on réexamine l'ensemble des dossiers de patients traités dans un certain hôpital pour une certaine pathologie). Dans un tel cas, la définition même de ce que pourrait être la population ciblée est déjà problématique, et il est probablement inutile de souligner à quel point les données obtenues dans ce type d'enquête diffèrent de celles produites par un sondage aléatoire.

9.3 Phénomènes de sélection

De manière très générale, on parle de biais de sélection lorsque des données subissent en quelque manière une sélection modifiant la probabilité devant être employée pour les décrire, sans que cela soit pris en compte. Les exemples en sont innombrables, et l'impossibilité de s'assurer de leur absence constitue l'une des tares des échantillonnages non-aléatoires.

Par exemple, des parents peuvent être surpris de découvrir que les résultats scolaires de leur enfant se situent dans la moyenne, alors que toutes les notes dont il leur a fait part sont supérieures à 16/20, car il n'a présenté que les meilleures.

Mettre bcp d'autres exemples, en particulier de biais involontaires !

Par exemple, si la non-réponse à une enquête dépend de la réponse que devraient fournir les individus à cette enquête, il est clairement impossible de se contenter d'ignorer les non-réponses et d'évaluer les quantités d'intérêt à partir des réponses exprimées comme si de rien n'était. Un exemple célèbre est l'élection présidentielle américaine de 1936, dans lequel des prévisions pourtant basées sur des millions de réponses collectées lors d'enquêtes postables s'avérèrent erronnées, les partisans de l'un des candidats (Roosevelt) choisissant beaucoup plus souvent que les partisans de son adversaire (Landon) de ne pas répondre. Plus généralement, on imagine facilement que les personnes particulièrement intéressées par le sujet d'une enquête répondront plus volontiers que les autres, et que l'intérêt pour le sujet peut être lié à l'opinion que l'on s'en fait.

Le nombre de données collectées n'est jamais à lui seul un gage d'estimations fiables ou de représentativité, même s'il constitue un argument rhétorique d'une certaine portée.

Toutes sortes d'autres biais de sélection, plus ou moins subtils, involontaires ou au contraire intentionnels, peuvent se présenter. Citons, à titre d'exemple, la fait de ne choisir de présenter, parmi un vaste ensemble de tests pratiqués, que ceux qui ont produit un résultat significatif, ou, plus généralement, le fait de ne choisir que les études qui appuient une hypothèse en omettant de citer les autres ; le fait de partitionner des données en fonction de leur contenu, tout en leur appliquant des techniques supposant un choix de partition indépendant des données ; le fait de ne rendre compte de certaines observations que lors-

qu'elles semblent surprenantes ou mériter l'attention (parce qu'elles confirment une théorie naissante, ou au contraire, contredisent des hypothèses établies ; c'est ce que l'on appelle le biais de publication), ce qui peut poser de sérieux problèmes lors de méta-analyses ; le fait de n'utiliser un moyen d'interroger qui exclut d'emblée toute une partie de la population ciblée (téléphoner dans une certaine tranche horaire, par exemple).

Plus globalement, le biais de sélection apparaît couramment dans l'argumentation (pas nécessairement fondée sur des statistiques) : ne présenter que les avantages d'une technologie (par exemple la production d'électricité nucléaire) sans parler de ses inconvénients, ou l'inverse ; ne présenter que les bons résultats que l'on a obtenus et passer sous silence les mauvais, ne présenter que les avantages d'une option que l'on préfère, et que les inconvénients de celle que l'on souhaite voire rejetée, etc...

De manière générale, il faut toujours suspecter un biais de sélection dans des données n'ayant pas été obtenues par sondage aléatoire, même en l'absence de doute quant à une sélection délibérée. La sélection peut opérer aussi bien en amont (au moment de la collecte des données), qu'en aval (lors du traitement des données et leur présentation).

Insistons au passage sur la différence existant entre une méthodologie de collecte de données clairement précisée, et une méthodologie correcte. Bien entendu, il est très important que le protocole décrivant, entre autres, la collecte des données utilisées pour une étude, soit décrit de manière détaillée. Cependant, il pourra être précisé clairement dans un épais document de méthodologie que les données de référence sont communiquées sur la base du volontariat (qui plus est par des personnes pouvant avoir intérêt à sélectionner les chiffres les plus avantageux), cela laisse planer un énorme soupçon de biais de sélection, même si la méthodologie retenue est clairement indiquée.

9.4 Redressements

On parle généralement de redressement lorsqu'un traitement a posteriori est appliqué à un échantillon dans le but d'améliorer l'estimation qu'il produit, afin de compenser des effets variés. La post-stratification à partir de divers critères est un exemple de redressement. Souvent, les redressements sont basés sur des hypothèses qui, comme toujours, doivent absolument être rendues explicites et autant que possible testées. Par exemple, on se basera sur une comparaison entre les intentions de vote annoncées en faveur de l'extrême-droite lors de sondages passés et les résultats électoraux observés afin de «redresser» l'effet de censure lié à la réticence de certains électeurs à avouer au sondeur leur intention de vote réelle.

9.5 Bibliographie

Un ouvrage de référence en théorie des sondages.

Yves Tillé. Théorie des sondages. Dunod, 2001.

L'encyclopédie libre et gratuite wikipedia.

`http://en.wikipedia.org/wiki/Selection_bias`

Les documents de méthodologie de l'INED, qui illustrent par des exemples concrets les questions survolées ici.

`http://www.ined.fr/rencontres/seminaires/methodes/`

Un document pédagogique, réalisé par Statistique Canada.
`http://www.statcan.ca/francais/edu/power/toc/contents_f.htm`
Le site d'un institut de sondage privé, à consulter en tant que tel :
`http://www.ipsos.fr/CanalIpsos/cnl_static_content.asp?rubId=35`

Exercice 23

Exercice 24 *Afin de tester le sérieux des répondants dans des enquêtes, on pose parfois des questions redondantes afin de vérifier que les réponses fournies sont cohérentes. Que pensez-vous de cette méthode ?*

10 Corrélation et causalité

L'une des pires erreurs de raisonnement que l'on puisse commettre est de déduire d'une simple corrélation entre deux événements (ou deux variables) l'existence d'une relation de cause à effet entre eux (elles), sophisme connu en particulier sous le doux nom de «cum hoc ergo propter hoc». Par exemple : je me suis mis à chanter ce matin et il s'est mis à pleuvoir peu après, donc ce sont mes multiples fausses notes qui provoquent la colère des éléments et font pleuvoir.

Sauf à avoir un fort biais de sélection, éventuellement psychologique, (voir également le chapitre sur les coïncidences), on ne considère pas en général ce type de coïncidence comme étant représentatives de quelque relation de cause à effet que ce soit, et elles sont, souvent à juste titre, considérées comme fortuites. En revanche, lorsque la corrélation est observée sur un vaste ensemble d'expériences ou une population importante, les réticences à interpréter celle-ci comme le signe d'une relation de cause à effet sont beaucoup moins fortes. Il s'agit de l'une des pires erreurs que l'on puisse commettre, et cela n'a rien à voir avec de possibles problèmes liés à l'échantillonnage. Une corrélation, même établie de manière parfaitement rigoureuse, basée sur des données parfaitement exactes, n'est pas en général le signe d'une relation de cause à effet.

La définition précise de ce que signifie la corrélation peut dépendre du contexte. Par exemple, entre deux événements A et B d'un modèle probabiliste, on dira qu'il y a dépendance entre A et B lorsque $\mathbb{P}(A|B)$ diffère de $\mathbb{P}(A)$, la dépendance étant positive lorsque $\mathbb{P}(A|B) > \mathbb{P}(A)$ et négative lorsque $\mathbb{P}(A|B) < \mathbb{P}(A)$. Lorsque l'on condidère des variables aléatoires, le coefficient de corrélation linéaire entre X et Y est défini comme $\mathbb{E}((X - \mathbb{E}(X))(Y - \mathbb{E}(Y)))/\mathbb{V}(X)^{1/2}\mathbb{V}(Y)^{1/2}$, mais on entend plus généralement par corrélation entre les deux variables le fait que celle-ci ne sont pas indépendantes, par exemple parce que des valeurs élevées de X sont positivement associées à des valeurs élevées de Y, une formalisation précise de la notion devant bien sûr être fournie à chaque fois.

Voici quelques exemples concrets de corrélations.

- en Italie, on a constaté que les régions dans lesquelles les taux d'achat d'ordinateur personnels sont celles où les taux de divorce sont les plus élevés ;
- une étude japonaise portant sur 40000 quadragénaires montre que ceux qui se brossent les dents après chaque repas parviennent mieux que les autres à garder la ligne ;

- il existe une corrélation positive entre utilisation de crème solaire et cancer de la peau ;
- le nombre de noyades est positivement corrélé à la consommation de crèmes glacées ;
- le prix des cigarettes est négativement corrélé qu nombre des agriculteurs en Lozère ;
- en Ecosse, les achats de whisky sont positivement corrélés au montant des dons reçus par les églises ;
- la carte du vote Le Pen lors des élections présidentielles de 2002 se superpose avec celle de l'irradiation due au nuage de Tchernobyl ;
- dans les communes qui abritent des cigognes, la natalité est plus élevée que dans le reste du pays ;
- la confiance des investisseurs est corrélée positivement à la croissance économique ;
- la consommation régulière d'alcool pendant la grossesse est corrélée à des retards de QI et des difficultés d'apprentissage chez les enfants ;
- la hausse des recettes publiques allemandes est positivement corrélée à la hausse des dépenses des ménages espagnols ;
- la proportion de fonctionnaires dans une ville est négativement corrélée au dynamisme économique.
- les enfants P*** acceptent plus volontiers les repas lorsqu'ils sont préparés par leur père que par leur mère ;
- la présence d'un médecin obstétricien lors d'un accouchement accroît la probabilité de complications ;
- le fait d'avoir recours à la péridurale diminue la mortalité lors des accouchements ;
- le nombre d'écoles maternelles dans une ville est positivement corrélé au nombre de crimes et délits ;

Certains exemples ci-dessus paraissent loufoques, d'autres plus recevables. Il est important de comprendre qu'aucune des corrélations mentionnées ci-dessus ne constitue un argument suffisant ou même sérieux pour affirmer l'existence d'une relation de cause à effet. Bien entendu, dans le cas des exemples loufoques, personne ne peut sérieusement penser qu'il existe une telle relation. Cependant, les exemples d'apparence plus sérieux sont de même nature, même s'il est beaucoup plus difficile dans leur cas de se défendre contre le penchant naturel consistant à interpréter une corrélation comme un rapport de cause à effet, par exemple, parce que nous sommes déjà convaincus de l'existence d'un tel rapport, et que nous sommes tentés de voir dans la corrélation une confirmation expérimentale de notre opinion, en omettant d'envisager sérieusement les autres explications possibles.

Dans la plupart des exemples ci-dessus, on peut facilement imaginer qu'une **cause cachée** est réellement responsable de l'«effet» observé, et se trouve simplement liée dans la population avec la «cause» présentée.

Exercice 25 *Suggérez des causes cachées pouvant servir à expliquer les corrélations présentées ci-dessus.*

Par exemple, les régions d'Italie dans lesquelles les achats d'ordinateurs personnels sont les plus élevés sont davantage les régions du nord, à l'économie prospère et au mode de vie moderne, que les régions du sud, moins développées économiquement, et où la tradition catholique est plus présente. Le mode

de vie semble donc une cause susceptible d'expliquer la différence des taux de divorce, naturellement liée aux achats d'équipement informatique. Quoiqu'il en soit, on ne peut pas déduire de cette corrélation que l'utilisation intensive de l'ordinateur a tendance à isoler les époux et détruit les couples.

De même, on peut facilement imaginer qu'une bonne hygiène de vie s'accompagne à la fois d'un brossage de dents réguliers et d'une absence de surcharge pondérale. Le simple fait de se brosser les dents n'est probablement pas à lui seul responsable du maintien de la ligne !

On voit ainsi que des informations parfaitement correctes peuvent conduire à des interprétations complètement erronnées. Rappelons que la qualité des informations de départ ne garantit en rien la qualité de l'interprétation qui en est faite, même si cet argument est souvent employé «Contestez-vous les chiffres ? Ils ont pourtant été bien établis...».

Attention : nous ne sommes pas en train d'expliquer qu'une corrélation observée n'est jamais le signe d'une relation de cause à effet. Par exemple, dans le cas d'une consommation d'alcool au cours d'une grossesse, le risque lié à l'alcool est considéré par les médecins comme parfaitement établi. Simplement, la mention de la corrélation ne suffit pas à entraîner l'existence d'un rapport de cause à effet, et ne constitue pas un argument solide pour l'établir, même si elle peut en constituer un indice. On peut la plupart du temps trouver une cause cachée possible (par exemple, le fait que la consommation régulière d'alcool chez les futures mères soit liée à des difficultés sociales ou relationnelles, qui, à leur tour, peuvent retentir sur les performances scolaires de l'enfant), et le simple constat d'une corrélation ne peut éliminer celle-ci.

Il est donc nécessaire, face à l'affirmation d'une corrélation, de rechercher systématiquement les causes cachées possibles.

Une première manière de tenter d'éliminer la possibilité que la corrélation observée résulte de causes cachées consiste à essayer de tenir compte explicitement de cette cause cachée, et à vérifier si la corrélation persiste une fois celle-ci prise en compte : la corrélation subsiste-t-elle lorsque les conditions qui caractérisent la cause cachée envisagée sont fixées (par exemple, la corrélation entre consommation de crème glacées et noyades subsiste-t-elle une fois que la saison est fixée ?). Ceci est loin d'être toujours facile ou même possible, notamment parce qu'une définition précise de ce qu'est cette cause cachée est nécessaire, et qu'il faut disposer d'informations suffisament précises pour pouvoir décomposer les données en strates correspondant à des valeurs fixées des conditions qui caractérisent la cause envisagée, et, qui plus est, si plusieurs causes possibles sont évoquées, il faut pouvoir tenir compte simultanément de chacune d'entre elles (par exemple, en séparant les données en individus ayant même sexe, même âge, même type de lieu de résidence, même catégorie socio-professionnelle, mêmes antécédents de santé, etc...).

On peut de cette manière éliminer la possibilité que la corrélation observée soit due à un certain nombre de causes cachées identifiées. Cependant, même si aucune cause cachée plausible ne semble pouvoir expliquer la corrélation observée, celle-ci ne prouve pas néanmoins la causalité, des causes qui nous échappent pouvant fort bien être à l'œuvre dans la situation étudiée.

Une solution, qui n'est pas toujours praticable, consiste à effectuer des expériences randomisées. On peut ainsi, au moins en principe, éliminer de l'étude l'ensemble des causes cachées susceptibles de créer une corrélation, grâce à la

randomisation.

Expliquons ceci sur un exemple. On désire tester l'efficacité d'un traitement, disons la vitamine X, sur une certaine maladie, une corrélation ayant déjà été mise en évidence au cours de premières phases de recherche entre guérison et prise de vitamine X. Pour ce faire, on constitue deux échantillons de la population à traiter par tirage aléatoire uniforme. On administre aux membres du premier groupe la vitamine X, le deuxième groupe servant de groupe témoin.

Supposons que la corrélation observée avant l'étude soit due à une cause cachée correspondant à une certaine caractéristique des individus de la population, qui peut être soit présente, soit absente (pour simplifier, on considère qu'il n'y a que deux possibilités), la guérison et la prise de vitamine X devenant indépendantes lorsque l'on conditionne par la présence ou l'absence de la cause en question.

Exercice 26 *Vérifiez que la probabilité de guérison est la même dans le groupe traité par la vitamine X et dans le groupe témoin.*

En pratique, on administre en général au groupe témoin un placebo, et l'on procède à un essai en double aveugle, c'est-à-dire sans que ni les sujets de l'étude, ni le personnel menant celle-ci, ne sache quels sont les patients qui reçoivent le véritable traitement et ceux qui reçoivent le placebo. Ces précautions ne sont en rien superflues (voir par exemple l'ouvrage de Daniel Schwarz), et elles sont destinées à compenser les effets psychologiques que peuvent être la prise d'un traitement (actif ou non) et l'influence des médecins organisant l'étude. Plus généralement, on doit chercher à éliminer les causes cachées autres que la prise de vitamine X qui pourraient expliquer la guérison, et intervenir postérieurement au choix randomisé des individus dans la population (l'effet des causes cachées antérieures à ce choix étant éliminé par la randomisation). Idéalement, il faudrait pouvoir modéliser les individus du premier et du deuxième groupe comme étant choisis aléatoirement dans une population décrite exactement par la même loi de probabilités dans les deux cas, en incluant dans le modèle la totalité des caractéristiques autres que celles de recevoir de la vitamine X (en particulier, le fait de recevoir un traitement, les conditions environnementales et psychologiques au moment où l'étude est effectuée, etc...). Même si toutes sortes de précautions sont prises, il est bien entendu difficile de procéder à une randomisation aussi complète, et l'on se contente souvent d'une randomisation qui porte sur les facteurs plausibles pouvant influencer le résultat (tels que les effets psychologiques décrits précédemment, ou encore une variation systématique des conditions environnementales telles que la température au moment où l'étude est pratiquée). Retenons que, pour se prémunir contre toutes sortes d'effets, la randomisation devrait être la plus systématique possible. Il n'est bien entendu pas toujours possible d'organiser des expériences contrôlées et randomisées sur une échelle suffisante, et dans des conditions complètement satisfaisantes, en particulier lorsque la définition exacte des effets et/ou des causes considérées ne peut pas être donnée, lorsqu'une prise en compte d'effets sur une vaste échelle de temps s'impose, etc...

Pour plus de détails sur l'organisation d'expériences randomisées, et plus généralement sur la planification des expériences, nous renvoyons à renvoyer à un ouvrage sur la planification des expériences ? ? ? ?

En combinant ce qui précède, ainsi que ce qui a été dit sur les questions d'échantillonnage dans la partie précédente, nous pouvons proposer une hiérar-

chie entre les différents arguments (statistiques, sauf pour le premier d'entre eux) visant à établir, ou tout au moins à accréditer, la présence de relations de cause à effet.

1. la mise en évidence expérimentale d'un mécanisme causal, et le confirmation que c'est bien ce mécanisme qui est impliqué, et qu'il suffit à expliquer l'association observée ;
2. les résultats d'expériences contrôlées avec randomisation ;
3. les résultats d'expériences contrôlées sans randomisation ;
4. les résultats obtenus à partir d'études encadrées de cohorte ou de cas-témoins ;
5. les résultats obtenus en collectant simplement les données disponibles ;
6. l'opinion d'experts et de praticiens du domaine considéré, basés sur leur expérience et/ou des études descriptives.

Par décence, nous ne mentionnerons pas les résultats de micro-trottoirs, les enquêtes menés auprès des lecteurs de magazines people, l'opinion du concierge ou du coiffeur.

Une telle classification ne vise certainement pas à remettre en question la validité les opinions des experts ou des praticiens, mais simplement à souligner les différences de force existant entre les différents types d'arguments.

Bien entendu, la négligence ou la fraude, ou encore des effets incorrectement pris en considération peuvent affecter chacun des types d'arguments précédents, et donc leur validité. Retenons que la méthodologie statistique employée pour analyser les résultats doit également être satisfaisante (en prenant correctement en compte les questions liées à la variabilité, aux différents biais possibles, etc...). En général, l'établissement d'un lien de cause à effet (par exemple entre la consommation de tabac et le cancer) repose sur un vaste ensemble d'arguments du type précédent, que l'on collecte au cours de méta-analyses de l'ensemble des données disponibles sur un sujet donné.

Pour en savoir davantage sur la méthodologie d'enquête, dans le contexte des études épidémiologiques.

`http://bmj.bmjjournals.com/epidem/`

En guise de conclusion, mentionnons l'argument suivant : une grande majorité des gens qui meurent ont vu un médecin peu avant leur décès. Conclusion évidente : il est dangereux de voir un médecin ! De plus, une forte proportion des gens meurent dans leur lit. Conseil : si vous ne pouvez éviter de voir un médecin, recevez-le debout, c'est probablement moins risqué !

10.1 Bibliographie

Daniel Schwartz. Le jeu de la science et du hasard. Flammarion, 1994.

Planification des expériences...

11 Estimation et apprentissage de modèles

On regroupe en général sous le nom de techniques d'estimation les méthodes visant à estimer à partir de données les paramètres d'un modèle.

Suivant le contexte, les méthodes d'estimation employées peuvent varier grandement. Deux méthodes couramment employées sont :

– l'estimation à partir de moments, qui repose sur l'évaluation de l'espérance de diverses quantités à l'aide de la loi des grands nombres : on estime $\mathbb{E}(f)$ par $N^{-1}\sum_{i=1}^{N} f(\omega_i)$, où les ω_i représentent les données, en général considérées comme des réalisations, en général indépendantes, du modèle probabiliste $(\Omega, \mathbb{P})$, pour diverses variables aléatoires f. Ces espérances s'expriment en fonction des paramètres du modèle, et l'on inverse donc les équations obtenues en égalant les espérances théoriques à leurs estimations empiriques.
– l'estimation par maximum de vraisemblance : à chaque jeu possible de paramètres, on associe une quantité qui mesure la vraisemblance des données disponibles sous l'hypothèse que ce jeu de paramètres représente les véritables valeurs de ceux-ci. Dans le cas d'un modèle discret, la vraisemblance est simplement la probabilité d'observer les données observées dans le modèle. Dans le cas d'un modèle continu, on remplace la probabilité par la densité correspondante. On estime les valeurs des paramètres en cherchant le jeu qui maximise la vraisemblance des données observées.

Ces procédures donnent lieu à de vastes généralisation. On retient cependant les deux idées distinctes : estimer les paramètres à partir de combinaison des données, ou optimisation d'un critère de qualité de description de celles-ci.

La manière standard de mesurer la qualité des estimateurs est d'étudier à quel point la distribution de probabilité de l'estimateur sous le modèle est concentrée autour des vraies valeurs des paramètres que celui-ci est supposé estimer. En supposant que les données disponibles sont modélisées par N variables aléatoires $X_1, \ldots, X_N$, un estimateur s'écrit sous la forme d'une fonction des données $T(X_1, \ldots, X_N)$. Lorsque le paramètre recherché est un élément θ de $\mathbb{R}^k$, on mesure en général l'écart entre $T(X_1, \ldots, X_N)$ par la norme euclidienne au carré de la différence $||T(X_1, \ldots, X_N) - \theta||^2$.

L'écart quadratique moyen de l'estimateur est défini par

$$r(\theta) = \mathbb{E}_\theta ||T(X_1, \ldots, X_N) - \theta||^2,$$

et fournit une mesure de la tendance de l'estimateur à se trouver à proximité des valeurs qu'il doit estimer. Notez bien que cette mesure dépend de θ, puisqu'elle suppose donné le modèle, sous lequel la performance de la procédure d'estimation est évaluée.

On décompose en général $r(\theta)$ sous la forme dite biais–variance, soit

$$r(\theta) = b(\theta)^2 + \mathbb{V}_\theta(T),$$

où $b(\theta) = \mathbb{E}_\theta(T - \theta)$ est appelé le biais de l'estimateur. On distingue ainsi deux sources d'erreurs d'estimation : les fluctuations de T autour de son espérance, et l'écart entre son espérance et le paramètre à estimer. Bien entendu, l'idéal est de rendre simultanément petites ces deux quantités, mais on ne peut pas en général descendre au-dessous d'une certaine limite de précision, qui correspond à la quantité d'information apportée par les données sur les paramètres du modèle (tout ceci pouvant être formalisé de manière relativement précise). Comme r est une fonction de θ, et que l'on ne connaît pas a priori la localisation de θ (mieux que son appartenance à un ensemble de valeurs possibles), il faut chercher à rendre $r(\theta)$ petit pour toutes les valeurs possibles de θ simultanément. On peut, dans ce contexte, définir diverses notions de comparaison et d'optimalité de

procédures de tests, et évaluer théoriquement la qualité de diverses procédures de tests par rapport à ces notions.

La théorie se limite souvent à des estimateurs non-biaisés, c'est-à-dire dont le biais est nul pour toute valeur de θ, ce qui n'est pas toujours la meilleure solution, un faible biais pouvant parfois être compensé par une réduction de variance plus importante. Exemple de l'estimateur de Stein en dimension sup à 3.

L'estimation bayésienne, qui suppose donnée une loi a priori sur les différentes valeurs des paramètres, définit quant à elle des critères de qualité moyennés sur cette loi a priori.

En plus d'une estimation ponctuelle, c'est-à-dire constituée par une unique valeur estimées pour les paramètres, on cherche généralement à fournir un intervalle de confiance, ou encore une fourchette de valeurs, parmi lesquels on peut raisonnablement supposer que la véritable valeur doit se trouver. Un intervalle de confiance est, au sens usuel, un intervalle aléatoire – car construit en fonction des données, qui sont vues comme aléatoires – dont la probabilité sous le modèle de contenir la véritable valeur du paramètre est contrôlée, par exemple par un niveau de confiance $1 - \alpha$ similaire à celui défini pour les tests statistiques. Le théorème de la limite centrale fournit les exemples les plus simples de tels intervalles de confiance, généralement approchés et exacts seulement dans la limite d'un nombre d'observations tendant vers l'infini. Les intervalles de confiance bayésiens possèdent une interprétation différente : les paramètres à identifier sont affectés d'une loi de probabilité, et l'on peut fournir un intervalle dans lequel on estime qu'ils ont une certaine probabilité de se trouver.

Dans tous les cas, la construction théorique d'un intervalle de confiance repose sur la capacité à connaître les caractéristiques de la loi de probabilité de l'estimateur sous l'hypothèse que le modèle sous-jacent est correct, quitte à estimer certains paramètres de celui-ci pour obtenir lesdits intervalles.

Comme pour les tests, la validité de la procédure d'estimation, et la fiabilité des indicateurs théoriques de qualité, repose sur la validité du modèle sous-jacent, qui doit impérativement être établie si l'on souhaite utiliser les résultats produits par l'estimation. Il se peut parfaitement que les «paramètres» estimés n'aient tout simplement aucun sens dans le problème considéré, si le modèle n'a aucune validité. Ici encore, l'un des risques est que les méthodes d'estimation peuvent facilement être utilisées sur des données sans que les modèles sous-jacents soient le moins du monde valables, produisant des résultats sans signification mais ayant toutes les apparences de la précision scientifique.

exple des droites de régression proposées dans le handbook du nist.

Se pose également, dans des conditions comparables aux tests statistiques, le problème de la robustesse des procédures d'estimation vis-à-vis d'écarts par rapport au modèle supposé.

12 Validation et choix de modèles

Rappelons, à toutes fins utiles, que la pertinence de l'utilisation d'un modèle dans une situation concrète est entièrement conditionnée par la validité de celui-ci, et qu'il ne faut pas se laisser impressionner par l'aspect technologique ou standardisé d'une procédure de modélisation (accompagnée généralement de procédures standardisées d'estimation et de test) si la validation a été laissée de

côté. Fondamentalement, les hypothèses sous-jacentes au modèle doivent être validées, les procédures de test utilisées à cette fin précisées, afin que l'on puisse juger de la crédibilité de celui-ci. L'un des pièges fréquents à ce stade est lié à la puissance potentiellement très faible des tests de validité employés, qui peuvent conduire à accepter, par défaut, un modèle que les données n'accréditent pas réellement. ERREUR SCANDALEUSE : ne pas chercher à valider les hypothèses ERREUR BIS : se laisser abuser par un test de faible puissance

Pour évidents qu'ils paraissent, ces principes élémentaires ont souvent été ignorés dans de nombreuses études statistiques, voir par exemple l'article (polémique) de Breiman et de Diaconis cités dans la bibliographie, conduisant à des résultats dont la validité est donc plus que douteuse.

Dans de nombreuses situations, le choix d'un modèle stochastique est en partie dicté par l'existence d'une structure sous-jacente et de mécanismes connus dans le phénomène ou le système modélisé. Par exemple, si l'on étudie un système industriel complexe, on pourra chercher à modéliser séparément les différents éléments de ce système, par exemple les temps de traitement de différentes tâches, les taux de panne des différents éléments, etc..., l'interaction entre ces éléments étant supposée connue. Idéalement, on pourra supposer une certaine indépendance, ou une forme simple de dépendance, entre les différentes sources d'aléa, décrire celles-ci par des familles de lois de probabilité classiques (gaussiennes, exponentielles, etc...), et valider les différents modèles employés pour chaque élément, ainsi que la manière dont ceux-ci sont assemblés dans la modélisation du système global, par une analyse exploratoire complétée par divers tests statistiques (éventuellement pratiqués par simulation du système étudié, si l'on fait appel à des statistiques relativement complexes, notamment celles qui font intervenir l'interaction des différents éléments du système), qui permettront de mesurer l'adéquation entre le modèle et les données disponibles avec un minimum de puissance.

Dans ce cas, un nombre de paramètres en général relativement restreint permet de spécifier le modèle, et les paramètres employés, ainsi que les hypothèses de modélisation qui sous-tendent leur utilisation, possèdent en général une interprétation concrète, «physique», relativement naturelle dans le cadre du système étudié.

Le cas le plus simple est probablement celui où l'on dispose simplement d'observations uni-dimensionnelles, provenant en principe de réalisations indépendantes du fonctionnement d'un même système, et que l'on cherche à modéliser.

La situation est au contraire beaucoup moins confortable lorsque l'on dispose d'un phénomène fournissant un grand nombre de données de types différents, ou encore, des données comportant un grand nombre de dimensions (deux ou plus !), ne possédant pas de structure connue préalablement. Par exemple, des mesures d'un grand nombre de caractéristiques d'un individu (régime alimentaire, âge, poids, origine ethnique, niveau socio-économique, antécédents médicaux), effectuées éventuellement à différentes époques, et entre lesquelles on cherche à établir des relations, ce qui suppose un minimum de modélisation. Dans ce contexte, une structure «naturelle» n'est plus forcément imposée par le problème considéré, si bien que toutes sortes de modèles, en général complexes, possédant chacun leur structure propre et reposant sur des hypothèses nombreuses, peuvent être proposés. Parfois, on renonce à employer des modèles dont la structure est censée être explicative (c'est-à-dire, dont les différents élé-

ments sont relativement interprétables en termes de mécanismes intervenant dans le phénomène modélisé), pour se contenter de modèles algorithmiques «boîtes noires» (tels que réseaux de neurones, forêts de décision, SV-machines), dont on attend simplement des performances prédictives correctes en rapport avec le phénomène ou le système considéré.

Tester la validité de ce type de modèles, et, plus globalement, choisir un modèle qui semble bien adapté, ou encore, le mieux adapté parmi un ensemble de modèles possibles, devient alors une tâche beaucoup plus complexe. Les tests pratiqués reposent naturellement sur une mesure de l'adéquation entre le modèle proposé (ce qui inclut les différentes hypothèses que celui-ci pose sur le phénomène considéré) et les données disponibles, souvent (mais pas toujours) en mesurant les p-valeurs obtenues en pratiquant des tests statistiques sur celles-ci.

Mentionnons quelque questions importantes dans ce contexte.

12.1 L'effet Rashômon

Inspiré par un (excellent) film de Kurosawa, dans lequel plusieurs personnages fournissent des explications totalement différentes et incompatibles des mêmes faits constatés, sans qu'il soit possible de départager ces différentes explications. Plusieurs modèles possédant des structures totalement différentes pourront s'adapter de manière apparemment satisfaisante aux données – de nombreux modèles différents obtenant des scores comparables lorsque l'on cherche à évaluer leur qualité, l'ajout ou le retrait d'une seule donnée pouvant modifier l'ordre exact du classement –, tout en fournissant des prédictions ou des interprétations totalement différentes. Ce problème est notamment lié à la faible puissance des procédures de test dont on dispose pour valider les différents modèles employés.

12.2 Parcimonie

De manière générale, on cherche à ce que les modèles que l'on emploie comportent le plus petit nombre possible de paramètres, pour deux raisons au moins [3].

D'une part, la précision avec laquelle il est possible d'estimer les paramètres d'un modèle diminue en général avec le nombre de paramètres à estimer, au point de rendre parfois impossible l'identification des paramètres, même en admettant la validité du modèle, et même lorsque ceux-ci ont un sens concrets vis-à-vis du problème étudié. Une solution possible consiste alors à tenter de simplifier le modèle, en concentrant autant que possible les simplifications sur les aspects du modèle qui ont le moins d'influence sur l'usage que l'on compte faire de celui-ci (les techniques dites d'analyse de sensibilité visent entre autres cet objectif).

exple, chaîne de Markov, chaîne de Markov cachée, d'ordre variable, etc...

ou encore famille de prédictions d'ordre linéaire, non-linéaire, etc... en grande dimension. voir Breiman

[3] La recherche de parcimonie est parfois nommée pompeusement «principe du rasoir d'Ockham», du nom du moine franciscain et philosophe Guillaume d'Ockham, ayant vécu au quinzième siècle, et d'après lequel, «plularitas non est ponenda sine necessitate», que l'on traduit en disant que l'on ne doit pas sans nécessité introduire des entités nouvelles dans une théorie.

D'autre part, plus un modèle possède de paramètres libres, plus il lui est facile de s'ajuster aux données disponibles. On parle de surajustement (overfitting en anglais) lorsque le modèle présente un excellent ajustement au jeu de données particulier utilisé pour l'estimer, mais ne donne pas une description correcte du phénomène lui-même. Ceci est bien entendu lié au fait que l'on juge en général de la validité (ou encore de la qualité, une fois que l'on a admis que le modèle n'est qu'une approximation de la réalité étudiée) d'un modèle par l'ajustement des prédictions de celui-ci aux données disponibles (en utilisant des scores d'ajustement fournis, par exemple, par des p-valeurs de tests). Dit dans les termes précédents, on doit prendre en compte l'impact de l'estimation des paramètres à partir des données sur le comportement des procédures de test pratiquées sur ces données, et l'on peut être conduit à des tests d'adéquation dont la puissance devient négligeable, du fait du trop grand nombre de paramètres par rapport aux données. Ce problème est encore plus gênant lorsque la structure d'un modèle n'est pas naturellement suggérée par une connaissance a priori de la structure du phénomène.

DIRE qqch sur on teste une propriété qui n'est pas vraiment estimée en tant que telle dans l'estimation. Quel type de confirmation ceci représente-t-il ?

Pour limiter ce problème, on introduit en général dans la mesure de la qualité d'un modèle une pénalisation à l'emploi d'un trop grand nombre de paramètres par rapport aux données disponibles. Bien entendu, le choix d'une pénalisation adéquate (autrement dit, du juste compromis entre parcimonie du modèle et qualité d'adéquation aux données disponibles) est un problème difficile et qui ne possède pas de solution générale satisfaisante.

Le recours à la validation croisée (voir le chapitre sur la validation croisée), dans laquelle le jeu de données utilisé pour l'estimation du modèle est distinct de celui sur lequel on teste sa validité, contribue également à éliminer les modèles surajustés.

12.3 Bibliographie

The Elements of Statistical Learning. T. Hastie, R. Tibshirani, J. Friedman. Springer 2001.

Simulation Modeling and Analysis. A. M. Law, W. D. Kelton. McGraw-Hill, 2000.

L. Breiman. Statistical Modeling : The Two Cultures. Statistical Science, 2001, Vol.16, No. 9, 199–231.

P. Diaconis. A place for philosophy ? The rise of modeling in statistical science. Quarterly of Applied Mathematics, Volume 16, No. 4, 1998, 797–805.

13 Bootstrap

L'idée de base des méthodes de bootstrap est d'utiliser une liste de données mesurées, supposées provenir d'une distribution donnée, pour produire des données simulées qui soient approximativement issues de la même distribution, et ce, afin de calculer diverses quantités. Dans les méthodes de bootstrap paramétriques, les données sont utilisées pour estimer les paramètres de la loi paramétrique sous-jacente censée décrire celles-ci, la loi paramétrique ainsi estimée étant employée pour simuler de nouvelles données. Dans les méthodes de

bootstrap non-paramétriques (au moins dans leur forme la plus simple), on se contente de simuler de nouvelles données selon la loi empirique associée aux données mesurées (on peut éventuellement considérer des versions lissées de cette loi empirique). La différence de ces méthodes par rapport aux méthodes de Monte-Carlo les plus classiques est donc que la loi que l'on cherche à simuler n'est pas supposée connue exactement, mais estimée à partir des données.

Les méthodes de bootstrap sont d'une grande utilité dans la construction d'intervalles de confiance, lorsque la loi de probabilité de la statistique employée $T(X_1, \ldots, X_N)$ est difficile à étudier directement, et ne peut simplement être estimée à partir des données disponibles. (Rappelons que $X_1, \ldots, X_N$ représentent des variables aléatoires $X_1, \ldots, X_N$, supposées en général indépendantes et identiquement distribuées, dont la loi commune est inconnue et sur laquelle on cherche à obtenir des informations).

Dans ce contexte, les méthodes de bootstrap sont employées pour produire des familles d'échantillons simulés $X^{(1)}, \ldots, X^{(M)}$, où $X^{(i)} = (X_1^{(i)}, \ldots, X_N^{(i)})$, que l'on utilise pour estimer les propriétés de la distribution de probabilité de T, comme si ceux-ci formaient des réalisations i.i.d. du vecteur aléatoire $(X_1, \ldots, X_N)$. Bien entendu, les échantillons ainsi collectés ne sont pas réellement distribués suivant la loi des observations, mais selon la loi empiriques de celles-ci. L'intérêt de la méthode est qu'elle permet d'obtenir, au prix de cette approximation, et de calculs numériques parfois intensifs (ce qui peut expliquer que ces méthodes ne se soient développées que récemment) autant de réalisations de T qu'on le souhaite, en extrayant de nouveaux échantillons de l'ensemble des données originales (d'où le nom de bootstrap : on désigne ainsi le fait de tirer sur ses lacets pour s'élever dans les airs!!!).

Exemple : estimation du biais, de la variance, construction d'intervalles de confiance.

La facilité d'utilisation de la méthode, et la possibilité de l'appliquer virtuellement à n'importe quelle situation font qu'elle est abondamment utilisée en pratique. Cependant, il est nécessaire de rappeler que celle-ci ne fournit que des résultats approchés, dont le degré d'approximation n'est pas nécessairement facile ou même possible à contrôler. Des études théoriques d'une part, et par simulation d'autre part, confirment que cette méthode produit effectivement des résultats à la fois valides et efficaces dans certaines situations. De manière informelle, le bootstrap fournit une manière naturelle très simple de mesurer ce qu'un échantillon de données indique quand à la variabilité d'une procédure d'estimation, mais l'utilisation de cette méthode ne signifie pas que l'on peut amplifier de manière démesurée un très petit échantillon de données pour obtenir des informations aussi fiables qu'on le souhaiterait sur la distribution de probabilité dont celui-ci provient.

Il existe de très nombreuses extensions et raffinements de l'utilisation du bootstrap et, plus généralement, des méthodes de rééchantillonnage (telles que les tests basés sur la randomisation) en statistique, par exemple, pour traiter des données structurées telles que des séries temporelles, pour effectuer des tests statistiques, pour étudier la sensibilité de diverses procédures statistiques aux données... Nous renvoyons aux références bibliographiques données ci-dessous pour une discussion plus détaillés de ces différentes méthodes, ainsi que des propriétés théoriques du bootstrap et des études empiriques de son efficacité.

13.1 Bibliographie

Deux ouvrages classiques sur le bootstrap.

Efron, B. et Tibshirani, R. J. (1993) An introduction to the bootstrap. New York : Chapman and Hall.

Bootstrap Methods and Their Applications. A.C. Davison, D.V. Hinkley. Cambridge University Press. 1997.

Un livre en ligne sur les méthodes de rééchantillonnage en statistique.

`http://www.resample.com/content/text/index.shtml`

Un exemple de choses plus élaborées sur le bootstrap : l'article d'Efron et al. sur la méthode de Felsenstein sur les arbres de phylogénie.

14 Validation croisée

Idéalement, il est souhaitable de confirmer par une étude indépendante les résultats (positifs ou négatifs) obtenus par une étude statistique reposant sur des tests, ou encore, d'utiliser des jeux de données indépendants pour estimer et pour valider un modèle. Cependant, il n'est pas toujours possible, par exemple pour des raisons de coût ou de comparabilité, de disposer d'un second échantillon (de confirmation, ou de validation).

Le principe de la validation croisée est de partir d'un seul échantillon, en le partitionnant. Sous sa forme la plus simple, elle consiste simplement à partitionner les données disponibles en un ensemble d'apprentissage, et un ensemble d'évaluation. Des formes plus complexes peuvent être employées, dans lesquelles on partitionne aléatoirement les données disponibles en plusieurs sous-ensembles, certains étant utilisés pour l'identification, et d'autres pour la validation.

Le but de cette méthode est notamment de contourner le problème de la dépendance pouvant exister entre l'hypothèse testée et les données utilisées, ou encore le surajustement.

15 Nous avons des principes

Nous espérons avoir illustré, sur divers exemples, le fait que le sens commun est prompt à se laisser prendre aux différents pièges que présente le raisonnement statistique. Faute d'une réflexion suffisante, on est très facilement conduit à tirer soi-même, ou à accepter en provenance d'autres personnes, des conclusions erronnées basées sur des arguments statistiques. Toutes sortes de raisons, qui interviennent aux différents stades de la démarche statistique (tels que par exemple l'échantillonnage, la mesure, le traitement statistique des données, la formulation et la validation des modèles, l'interprétation des résultats,...) peuvent conduire à ces erreurs, et nous avons présenté un certain nombre de celles qui nous semblent les plus fréquentes ou les plus importantes. Se confronter à la réalité d'un phénomène ou d'une situation est une tâche difficile, mais on ne peut pas se permettre de s'abandonner à la facilité si l'on souhaite l'aborder correctement.

15.1 La statistique est un sport de combat

Face à un argument statistique, la seule recommandation générale que nous pouvons donner est de formuler, de la manière la plus précise et la plus explicite qui soit (cela n'est pas très difficile dans la plupart des situations) la totalité des hypothèses, des définitions, des raisonnements, le plus souvent demeurées implicites dans l'argument, concernant la situation étudiée, les données recueillies, et la manière dont on peut les traiter. En aucun cas il ne faut abandonner ce terrain au langage courant ou à une pratique supposée standard ou experte, ou encore à des sources jamais consultées. C'est ainsi que l'on pourra découvrir que les données sont issues d'une enquête d'observation et non pas d'un échantillonnage aléatoire, que les hypothèses sous-jacentes au modèle utilisé n'ont pas été validées, que l'argument confond allègrement corrélation et causalité, compare entre elles des entités non-comparables, utilise des définitions précises qui n'ont qu'un rapport ténu ou fortement biaisé avec les terme employés pour décrire les résultats, ne tient pas compte de diverses variations de composition dans des comparaisons de moyennes, suppose dans ses conclusions que certaines quantités sont fixes d'une période à une autre sans l'avoir vérifié, effectue des quantifications douteuses, utilise des procédures statistiques inadaptées, oublie scandaleusement de prendre en compte la variabilité, qu'une description très précise et jamais lue des protocoles utilisés existe, mais qu'elle invalide totalement la portée des conclusions avancées, etc, etc... Les conclusions elles-mêmes ne seront pas nécessairement invalidées, mais la faiblesse de l'argument, et les questions laissées en suspens qui permettraient de le rendre concluant ou, au contraire l'infirmer, seront ainsi mises en évidence.

Les thèmes abordés dans ce bréviaire fournissent un certain nombre d'angles d'attaque, permettant de formuler correctement un certain nombre de critiques, mais leur liste n'est en rien limitative et il faut à chaque fois tenter de mener une analyse critique la plus complète possible de la démarche utilisée. Une attitude critique aussi systématique n'a pas a être motivée par une posture politique ou morale : elle est simplement nécessaire à une confrontation raisonnée aux arguments de nature statistique.

Bien entendu, il convient d'être critique également vis-à-vis de l'usage de la critique : des critiques valant en général, mais pas dans la situation considérée