Les probabilités n'existent pas... mais on vous explique quand même comment vous en servir

Gaëlle Chagny, Thierry de La Rue

Les probabilités n'existent pas... mais on vous explique quand même comment vous en servir

Gaëlle Chagny[*], Thierry de la Rue[†]

20 juin 2021

Cet article a été publié initialement dans *The Conversation*, et est accessible ici.

Nous devons chaque jour, dans notre vie personnelle ou professionnelle, prendre des décisions tout en n'ayant qu'une connaissance partielle des informations relatives à la situation : si je choisis cet itinéraire, vais-je me retrouver bloqué dans un embouteillage et arriver en retard ? Dans quelle station-service sur ma route le carburant sera-t-il le moins cher ? Ce chapitre du programme que je n'ai pas encore révisé a-t-il des chances de tomber à l'examen ?

Les exemples sont innombrables où, consciemment ou non, nous parions quotidiennement sur des événements dont nous ne savons pas s'ils vont se réaliser ou non. Dans de telles situations, nous ne pouvons donc pas être certains de faire le bon choix : au final, le résultat relèvera aussi du hasard. Malgré cette part d'incertitude, il nous faut tenter d'optimiser nos chances de succès, et cela passe par le calcul ou l'estimation de la probabilité des événements incertains. Cette probabilité est un nombre entre 0 et 1, d'autant plus proche de 1 que l'événement a des chances de se produire. Mais quels sont les mécanismes mis en jeu dans le calcul de ce nombre ?

Le mathématicien Bruno de Finetti (1906-1985) a passé sa vie à étudier la théorie des probabilités, dont il était un grand spécialiste. Pourtant il clame dans un ouvrage qui leur est consacré que « Les probabilités n'existent pas ! »

Voilà qui commence mal : comment pouvons-nous mettre en pratique la théorie de quelque chose qui n'existe pas ? En fait, par cette provocation, de Finetti voulait souligner que la probabilité d'un événement n'est pas une réalité objective : elle dépend de la personne qui l'estime et évolue en fonction des informations que celle-ci reçoit. Voyons cela concrètement sur un exemple.

Le « problème de Monty Hall »

Le « problème de Monty Hall » se présente sous la forme d'un jeu de hasard inspiré de l'émission télévisée américaine *Let's make a deal* (dont Monty Hall était le présentateur). Le candidat a devant lui trois portes fermées, notées A, B, C, derrière lesquelles sont cachées deux chèvres et une voiture réparties au hasard. Son but est de trouver la porte dissimulant la voiture. Il commence par désigner l'une des trois portes, sans l'ouvrir (disons que c'est la porte A). Le présentateur, qui connaît la répartition, annonce alors qu'il va montrer une chèvre cachée derrière l'une des deux autres portes (c'est toujours possible : puisqu'il n'y a qu'une voiture, au moins une des deux autres portes cache une chèvre). Après avoir ainsi dévoilé une chèvre (disons derrière la porte B), il demande au candidat si celui-ci maintient son choix initial, ou si il préfère aller vers l'autre porte encore fermée (la porte C dans notre exemple). À votre avis, le candidat a-t-il intérêt à changer de porte ?

[*]LMRS, UMR CNRS 6085, Université de Rouen Normandie, gaelle.chagny@univ-rouen.fr
[†]LMRS, UMR CNRS 6085, Université de Rouen Normandie, thierry.de-la-rue@univ-rouen.fr

Au départ du jeu, le candidat ne sait strictement rien sur la position de la voiture et des deux chèvres. Lorsqu'il choisit sa première porte, il a donc une chance sur trois d'avoir choisi celle qui cache la voiture. Mais en montrant la chèvre derrière la porte B, le présentateur apporte une nouvelle information au candidat : ce dernier, qui ignorait tout au début, sait maintenant que la porte B dissimulait une chèvre. En quoi cette nouvelle donnée pourrait-elle l'amener à réviser son choix initial ?

À ce point deux raisonnements s'opposent qui aboutissent à deux conclusions contradictoires. Voici la première façon d'aborder le problème : il reste deux portes fermées, A et C, l'une cache une chèvre et l'autre une voiture. Il y a alors une chance sur deux que la voiture soit derrière la porte A, et donc le candidat aurait autant de chances de gagner en gardant la porte A qu'en choisissant la C. Mais le second raisonnement consiste à remarquer que la voiture n'a pas changé de place depuis le début du jeu. Comme il y avait une chance sur trois qu'elle soit derrière la porte A, elle a maintenant deux chances sur trois d'être cachée derrière la porte C. Selon ce second raisonnement, le candidat doublerait ses chances de gagner en changeant de porte. Quel est parmi ces deux arguments celui qui fournit la bonne stratégie pour le candidat ?

On trouve très facilement en ligne des simulations du jeu, et nous l'avons expérimenté en situation réelle lors de la Fête de la Science avec un grand nombre de visiteurs. Les résultats sont sans appel : lorsque le candidat conserve la porte qu'il avait choisie initialement, il gagne environ dans 33% des cas, alors que la stratégie de changer de porte aboutit à environ 66% de succès. C'est donc bien le second raisonnement qui semble être correct. Mais alors, qu'est-ce qui cloche dans le premier ?

FIGURE 1 – Expérience illustrant le paradoxe de Monty Hall

Imaginons qu'avant le début du jeu, l'une des trois portes ait été mal fermée et qu'un courant d'air ait permis au candidat d'entrevoir une chèvre derrière cette porte. Compte tenu de cette information, il peut légitimement estimer que la chèvre restante et la voiture ont chacune une chance sur deux de se trouver derrière chacune des deux autres portes. Mais en quoi cette situation est-elle différente de celle du jeu décrit juste avant ?

L'erreur dans le premier raisonnement vient de la mauvaise appréciation de l'information effectivement apportée au candidat : contrairement au courant d'air qui dévoile une chèvre de façon fortuite, le présentateur choisit intentionnellement laquelle des deux autres portes il va ouvrir. Ainsi, non seulement le candidat sait qu'il y a une chèvre derrière la porte B, mais il doit également tenir

compte du fait que le présentateur a délibérément choisi cette porte. Dans le cas où la voiture est derrière la porte A, le choix du présentateur ne dit rien de plus (il peut avoir tiré à pile ou face quelle porte il dévoile), mais dans le cas où la porte A cache une chèvre, ce qui arrive deux fois sur trois, le présentateur choisit la porte B parce qu'il sait que la voiture est derrière la C.

On voit dans cet exemple comment la probabilité de l'événement « La voiture est derrière la porte C » change selon le point de vue. Pour le candidat qui arrive en ne connaissant rien d'autre que les règles du jeu, elle vaut 1/3. Si le candidat a entrevu de manière fortuite une chèvre derrière la porte B, elle passe à 1/2. Pour le candidat qui a d'abord désigné la porte A et qui a vu le présentateur montrer la chèvre derrière la porte B, elle est égale à 2/3. Et pour le présentateur qui sait tout, elle vaut 1 ou 0, suivant que la voiture est ou n'est pas derrière la porte C.

On mesure ici combien la probabilité d'un événement incertain dépend subtilement des informations dont dispose la personne qui l'estime. Si on néglige une partie de notre connaissance de la situation (par exemple, si on ne prend en compte que l'information brute « il y a une chèvre derrière la porte B » en oubliant que cette information résulte d'un choix intentionnel du présentateur dans un contexte précis), on risque d'aboutir à une mauvaise estimation et au final réduire nos chances de succès.

Les probabilités conditionnelles

Pour tenir compte de l'information partielle dont nous disposons dans l'évaluation de la probabilité d'un événement, on fait appel au concept de « probabilité conditionnelle » : la probabilité conditionnelle d'un événement C sachant l'événement B s'interprète comme la probabilité que C se réalise pour une personne qui dispose exactement de l'information que B est réalisé. Elle se calcule comme le quotient de la probabilité que les événements B et C soient réalisés en même temps par la probabilité a priori de l'événement B :

$$\mathbb{P}(C \text{ sachant } B) = \mathbb{P}(B \text{ et } C)/\mathbb{P}(B)$$

Cette notion revêt une importance capitale pour toute la théorie des probabilités. Elle permet de réviser nos estimations, nos chances de succès, en tenant compte d'informations additionnelles.

Or cela peut tout changer : confondre la probabilité d'un événement avec la probabilité d'un événement conditionnellement à certaines informations est une erreur fréquente, qui peut modifier totalement l'appréciation d'une situation. Prenons un exemple d'actualité en démographie. En 2020, l'espérance de vie des femmes à la naissance était de 85,1 ans, selon l'Insee. Pour une femme de 80 ans, cela ne signifie pas qu'il lui reste en moyenne seulement 5 ans à vivre ! Pourtant cet argument, suivi d'une comparaison avec l'âge médian à la date du décès des personnes victimes du Covid-19 (84 ans), est souvent évoqué dans les médias. Or, l'Insee montre aussi que les femmes de 80 ans ont encore une espérance de vie d'environ 11 ans. La différence est que dans l'espérance de vie à la naissance, on ne tient pas compte de l'information additionnelle : « sachant que la personne atteint au moins 80 ans ».

La relative simplicité de la formule de l'espérance conditionnelle cache de nombreuses difficultés pratiques. Nous en avons déjà mis une en évidence : on doit d'abord parfaitement identifier l'information qui nous est connue, représentée ici par l'événement B. Le second problème, sur lequel nous n'insisterons pas davantage, est qu'il nous faut disposer au départ d'une « probabilité a priori » censée représenter l'absence d'information sur la situation, et qui nous sert à mesurer P(B et C) et P(B). La citation de Bruno de Finetti donne une idée de la difficulté de cette question.

Nous voudrions surtout ici mettre l'accent sur un troisième piège : si la notion de probabilité conditionnelle constitue un outil universel, sorte de « couteau suisse » pour survivre dans un monde aléatoire, il s'agit de l'utiliser dans le bon sens et de ne pas confondre le manche avec la lame ! Il

existe en effet de nombreuses situations, dans lesquelles une mauvaise évaluation d'une probabilité a des conséquences autrement plus graves que pour le candidat du jeu de Monty Hall, et où il est si tentant d'utiliser les probabilités conditionnelles « à l'envers ». Nous illustrons ce phénomène à travers deux exemples concrets.

Le premier exemple, qui résonne malheureusement avec l'actualité de la pandémie, considère un test de dépistage d'une maladie sur lequel nous formulons les hypothèses suivantes : on suppose que si on teste une personne infectée, le test sera positif dans 99% des cas (soit un taux de faux négatifs égal à 1%), et qu'inversement si on teste une personne non infectée, le résultat sera positif dans 1% des cas (le taux de faux positifs est lui aussi supposé égal à 1%).

On fait subir le test à une personne prise au hasard dans la population, et le test est positif. Quelle est la probabilité que cette personne soit réellement porteuse de la maladie ? Nous sommes très tentés ici de répondre directement que cette probabilité est de 99%. Cependant, ce nombre est donné comme la probabilité conditionnelle que le test soit positif sachant que la personne est infectée. Et ce qui nous intéresse ici est la probabilité conditionnelle « inverse » : celle que la personne soit infectée sachant que le test est positif. Or, en général les deux probabilités conditionnelles ne sont pas identiques, elles peuvent même être très différentes !

C'est la célèbre formule de Bayes, l'un des résultats les plus importants de toute l'histoire des probabilités, qui permet de relier les deux. Elle s'écrit sous la forme suivante :

$$\mathbb{P}(C \text{ sachant } B) = [\mathbb{P}(B \text{ sachant } C) \times \mathbb{P}(C)]/\mathbb{P}(B)$$

Dans l'exemple du test de dépistage, B représente l'événement « le test est positif » et C l'événement « la personne est infectée ». Pour calculer la probabilité cherchée ici, celle de l'événement C sachant que l'événement B est réalisé, il nous manque une donnée essentielle qui s'interprète comme la probabilité a priori : le taux d'incidence de la maladie dans la population, c'est-à-dire la probabilité de l'événement C.

Supposons que cette maladie touche une personne sur mille. Sur un million de personnes, on compterait environ 1000 malades, dont 990 seraient détectés positifs par le test. Sur les 999 000 personnes non infectées, le test détecterait environ 9 990 faux positifs. Au total, la proportion de personnes malades parmi celles détectées positives au test serait donc $990/(9\,990 + 990)$, soit environ 9% ! Ainsi il y aurait moins d'une chance sur 11 que la personne testée positive soit réellement atteinte par la maladie. Notons toutefois que ce calcul suppose que l'on ne dispose d'aucune autre information sur la personne testée. Si par exemple la personne qui se fait tester présente des symptômes de la maladie, il faut intégrer cette information supplémentaire et cela fera certainement augmenter sa probabilité d'être réellement atteinte. Inversement, si on sait que la personne ne présente aucun symptôme, cela aboutira à une probabilité d'infection plus faible encore.

Dans le domaine judiciaire, l'emploi inversé de mauvaises probabilités conditionnelles constitue un piège classique appelé le « sophisme du procureur », qui peut aboutir à des conclusions dramatiques comme dans le second exemple que nous présentons. Dans les années 1990, un couple d'Anglais, Steve et Sally Clark, perdent successivement leurs deux bébés de mort subite du nourrisson (MSN, dans la suite). La mère est condamnée pour meurtre sur la base des conclusions d'un expert pédiatre. Celui-ci a convaincu les jurés du procès en tenant l'argumentation suivante : il estime (par des méthodes déjà très discutables) que la probabilité d'observer 2 MSN consécutives dans une même famille est de l'ordre de 1 sur 72 millions, et en conclut que ceci représente la probabilité que Sally Clark soit innocente.

Autrement dit, la mère est coupable avec une probabilité extrêmement proche de 1 ! L'erreur principale dans cet argument repose ici encore un mauvais usage des probabilités conditionnelles : la probabilité que la mère soit innocente sachant que ses deux enfants sont décédés est confondue à tort avec la probabilité que les enfants décèdent sachant que la mère est innocente. Le procès fut

révisé quelque temps plus tard, avec notamment une intervention de la Royal Statistical Society, et Sally Clark fut libérée. Malheureusement elle ne se remit jamais de ces épreuves et décéda peu après.

Si les raisonnements probabilistes sont un outil incontournable pour structurer nos raisonnements quotidiens et nous guider dans nos choix, il convient toutefois d'être prudent dans leurs usages et leurs interprétations, pour se garder de toute erreur aux conséquences plus ou moins importantes.

Connaître les principes de base sur lesquels s'est développée la théorie des probabilités peut nous aider à déjouer les pièges posés par la tentation de la facilité.