STICHTING
MATHEMATISCH CENTRUM
2e BOERHAAVESTRAAT 49
AMSTERDAM

Dantzig D van

Les Problems que pose l'application du calcul
des Probability

(Theory des probabilités I , p 45-65)
Exposes sur ses Fondements et
ses Applications

Gauthier-Villars Paris, 1952)

Les problèmes que pose l'application du Calcul des Probabilités

PAR

D. VAN DANTZIG

1. — L'application d'une théorie mathématique aux sciences empiriques s'effectue toujours de la manière suivante. Le système des formules mathématiques (ou logiques, etc.) est considéré comme un *modèle* (simplifié et régularisé) des phénomènes observés et de ceux qui seront observés, entre lesquels il est intercalé. La transition des *descriptions* des phénomènes observés (les données) au système formel sera dit *l'embrayage* de ce système, celle du système aux *prédictions* des phénomènes à observer sera dit le *débrayage* du système.

A vrai dire l'embrayage et le débrayage s'effectuent en plusieurs étapes, certaines parties du système n'étant qu'imparfaitement formalisées. En effet, déjà la description des phénomènes observés ne les reproduit qu'au prix de l'omission de beaucoup de détails, c'est-à-dire d'une manière simplifiée. En effet la relation entre les phénomènes et leurs descriptions ne peut pas être univoque, car il serait impossible d'appliquer un même mot à diverses situations. L'introduction d'une théorie scientifique est alors une nouvelle simplification de même que sa représentation par un système d'axiomes (lois), et que l'application de la logique déductive à celui-ci. Ici nous ne nous occuperons guère de ces phases partielles. D'ailleurs la description des données repose sur des expériences de nature générale, le plus souvent non-analysées, tandis que la prédiction des phénomènes à observer sert à effectuer des *actions* déterminées, soit de nature utilitaire individuelle ou sociale, soit de nature scientifique (par exemple à faire des expériences nouvelles). On peut même remplacer les prédictions par des décisions d'action, comme l'a fait récemment Abraham Wald, en se basant sur l'idée de Jerzy Neyman, de substituer à la notion classique de raisonnement inductif, celle de conduite inductive. Une telle décision cependant n'a de sens que si elle est considérée comme équivalente avec la prédiction que l'action à laquelle on s'est décidé réussira, ce qui doit être contrôlé par de nouvelles observations. Ce n'est donc que *l'ordre* de la prédiction avec l'observation subséquente et la décision d'action qui est changé. Le processus total a donc globalement la forme suivante (Cf. D. van DANTZIG, 1947) :

Expériences générales

Observations

Descriptions

Embrayage

Théorie scientifique	Modèle mathématique (logique)	———	Axiomes Déductions Conclusions

Débrayage

Décisions d'actions

Il n'est pas nécessaire que cet ordre soit identique à l'ordre chronologique. Lorsque les observations „finales" précèdent dans le temps le système formel, on dira que celui-ci sert à „expliquer" (au lieu de „prédire") ces observations.

Dans les théories classiques l'emploi du système formel est à peu près catégorique: lorsqu'il y a contradiction entre les prédictions et les observations ultérieures, on tâche d'abord d'écarter soit les observations initiales, soit les observations finales comme étant „erronées". Ou bien on s'efforcera de montrer que l'embrayage ou le débrayage ont été effectués d'un manière incorrecte. Mais si ces affirmations s'avèrent insoutenables, on rejette le système formel, en tâchant de le remplacer par un autre. On exige donc *qu'aucune des prédictions résultant du système formel ne soit infirmée* par les phénomènes observés.

2. — Souvent on ne réussit pas à trouver un modèle satisfaisant à cette conditon rigide; ou encore un tel modèle, quoique possible,

deviendrait tellement compliqué qu'il ne serait plus maniable. En effet, le modèle doit satisfaire à deux conditions, généralement à peu près contradictoires: celle de représenter *aussi exactement que possible* les expériences, et celle d'être *suffisamment simple* afin que l'on puisse le manier. Dans plusieurs sciences modernes on ne réussit pas à trouver un modèle qui satisfait à la condition imposée. Dans les sciences classiques, par exemple dans la physique, ceci reste possible aussi longtemps que les obervations ne deviennent pas trop précises. Or, en augmentant la précision des mesures, les modèles classiques deviennent généralement trop grossiers, et l'on ne pourra plus maintenir la condition catégorique.

Auparavant on pensait que dans de pareils cas on ne pouvait plus appliquer les méthodes de la mathématique et de la logique, et que l'on devait nécessairement s'appuyer par exemple sur l'intuition seule. Mais on a constaté bientot que les méthodes mathématiques restent applicables; pourvu que l'on affaiblisse un peu la condition rigide mentionnée ci-dessus.

Au lieu d'exiger qu'*aucun* des énoncés, résultant du modèle, ne soit infirmé, on se contente d'exiger que cela ne se produise que pour un petit nombre d'entre eux seulement; on *admet donc qu'au plus une proportion bien déterminée des prédictions soient contredites par l'expérience*. C'est en renonçant à savoir d'avance *lesquelles* de ces prédictions seront vérifiées, que l'on réussit à réintroduire les méthodes mathématiques.

Pour fixer les idées un peu, soient

$$a, b, c, d, \ldots, u, v$$

des formules appartenant au système formel logique, représentant des énoncés (prédictions) empiriques

$$A, B, C, D, \ldots, U, V.$$

En comparant ces prédictions avec les observations correspondantes, on trouvera parfois que la prédiction est *vérifiée* ce que nous désignerons par un signe $+$, et parfois que la prédiction est *infirmée*, c'est-à-dire *contredite* par l'expérience, ce qui sera noté par le signe $-$. En faisant cela, nous avons supposé que tant pour les prédictions que pour les observations seuls les cas d'accord complet ou de désaccord complèt peuvent se présenter. On obtient donc une suite telle que, par exemple

$$+A, +B, -C, +D, \ldots, -U, -V.$$

Lorsque le nombre d'énoncés dont il s'agit est N, et celui des prédictions infirmées M, nous appellerons M/N le *degré de méfiance* du système de prédictions. La condition affaiblie sera:

Le degré de méfiance ne dépassera pas un nombre α, le seuil de méfiance, fixé d'avance.

Nous verrons que même cette condition affaiblie ne pourra être satisfaite qu'avec deux restrictions sérieuses.

3. — Comment pourra-t-on, — à ces restrictions près précisées ci-dessous —, aboutir à un tel système d'énoncés?

Précisons *d'abord* la forme donnée habituellement au modèle. Il y a de nombreux formalismes, à peu près équivalents, ou plutôt isomorphes, parmi lesquels nous mentionnons seulement ceux de Pierre Simon de Laplace (1812), G. Bohlmann (1901), John Maynard Keynes (1920), Richard von Mises (1921), Hans Reichenbach (1933), Alexander Kolmogoroff (1934), B. O. Koopman (1940), et récemment Rudolf Carnap (1950).

Il s'agit toujours d'une fonction réelle P, ne prenant que des valeurs p avec $0 \leq p \leq 1$. Cette fonction est à une ou, ce qui est plus simple, à deux variables, disons A et B, qui ne sont pas des nombres, mais que parcourent un même ensemble, disons Ω.

Cet ensemble Ω est d'un type spécial, appelé „structure" par Oystein Öre, „treillis" („lattice") par Garrett Birkhoff, „sôme" par Constantin Caratheodory. On peut obtenir de tels ensembles, en prenant pour A et B des sous-ensembles d'un ensemble donné Φ (le „champ de probabilité' de Kolmogoroff) ou aussi des „événements éventuels" (brièvement: „éventualités"), correspondant aux „cas possibles" de Laplace, ou aussi des propositions (Keynes-Koopman, Reichenbach), ou des formules logistiques (Carnap). La négation $\neg A$, (non A), la conjonction $A \wedge B$ (A et B), et la disjonction $A \vee B$ (A ou B), dans les systèmes logiques correspondent respectivement à: l'ensemble complémentaire $\Phi - A$ (les éléments de Φ n'appartenant pas à A), l'intersection $A \cap B$ [1], l'union $A \cup B$ [1]. Il importe peu que le modèle ainsi construit soit un modèle mathématique (fonction additive d'ensemble, Kolmogoroff), ou un mo-

[1] L'intersection $A \cup B$ est l'ensemble des éléments de φ, appartenant à A et à B; l'union $A \cap B$ est l'ensemble des éléments de φ, appartenant à *au moins un* des ensembles A et B.

dèle logique (Keynes - Koopman, Reichenbach) ou un modèle sémantique (Carnap). Dans le dernier cas, P est un prédicat sémantique de deux variables, parcourant les „phrases" („sentences") d'un „language" („object-language") donné.

Nous désignerons cette fonction par $P[A|B]$ et appellerons sa valeur la probabilité de A sous la condition B. Les notations correspondantes des autres auteurs mentionnés seraient: a/b (Keynes - Koopman), $P_B(A)$ (Kolmogoroff), $\sqsubset$ (a, b) (Carnap), tandis que Reichenbach désigne la relation $P[A|B] = p$ par $B \overset{p}{\supset} A$.

Les règles principales auxquelles la fonction $P[A|B]$ *doit satisfaire* sont, en adoptant les notations logistiques:
Règle d'addition: $P[A_1 \wedge A_2|B] = P[A_1|B] + P[A_2|B]$ pourvu que $\neg (A_1 \wedge A_2 \wedge B)$;
Règle de multiplication: $P[A_1 \wedge A_2|B] = P[A_1|A_2 \wedge B] . P[A_2|B]$.

Nous ne discuterons ici ni les généralisations, ni les conclusions, ni les applications de ces règles. Etant données les valeurs de la fonction P pour un nombre suffisant de paires $[A|B]$, on en peut déduire les autres. Sur ce point il n'y a aucune différence entre les divers auteurs, dès qu'ils s'accordent sur le champ de définition parcouru par les variables — ce qui cependant n'est pas toujours le cas, en particulier pour le principe dit „de Bayes" —. L'extension de la théorie au cas d'ensembles infinis n'est qu'une régularisation subséquente et sert seulement à simplifier les déductions.

Pour nous, il s'agit de deux problèmes tout à fait différents, à savoir:

1. Etant donné un problème empirique déterminé, comment trouver la fonction de probabilité correspondante? (Problème de l'embrayage).
2. Ayant trouvé les probabilités désirées, qu'en tirera-t-on pour le problème empirique donné (Problème du débrayage).

4. — Nous considérons d'abord le problème du débrayage. Supposons qu'on ait obtenu une relation de la forme $P[A|B] = p$, où B correspond aux conditions sous lesquelles une expérience est faite, tandis qu'A désigne un prédicat qui peut être appliqué au résultat des observations. Tandis qu'une division des valeurs de p en intervalles très grossières peut s'exprimer en termes psychologiques („très peu probable", „des chances égales", „très probable", etc.), celà

n'est pas possible pour une distinction plus fine, telle que la différence entre $p = 0,63$ et $p = 0,67$. On ne peut pas davantage les traduire en termes *empiriques*.

Or, au moyen du théorème de Bernoulli, on peut réduire ce dernier problème au cas où p est un nombre qui diffère arbitrairement peu de 1. En effet, δ et ε étant deux nombres positifs arbitrairement petits, on peut déterminer un $n(\delta,\varepsilon)$ tel que pour chaque n $\geq n(\delta, \varepsilon)$, et chaque n-uple de paires indépendantes de prédicats $A_i \cdot B_i$ $(i = 1, \ldots, n)$ tels que $P[A_1|B_1] = \ldots = P[A_n|B_n] = p$ on aura

$$P[| \tfrac{m}{n} - p| \leq \varepsilon|B] \geq 1 - \delta,$$

où B désigne la conjonction des B_i, tandis que m désigne le nombre des A_i prédits (²). En omettant le B (³), supposé connu, nous abrégeons cette inégalité en

$$| \tfrac{m}{n} - p | \leq \varepsilon, \text{ spr } \delta,$$

où l'abréviation spr δ („salva probabilitate δ") veut dire: sauf une probabilité $\leqq \delta$.

Il suffit donc de débrayer des relations de la forme

$$A \text{ spr } \delta$$

avec un δ arbitrairement petit.

Déjà d'Alembert (1761) a dit qu'il était erroné d'admettre qu'une éventualité de très petite probabilité ne se produit que très rarement, mais qu'on doit dire qu'elle ne se produit point du tout. Plusieurs de ses contemporains et de leurs successeurs ont nié ce principe, l'ont même ridiculisé quelque peu jusqu'à nos jours, où M. Emile Borel a réhabilité ce principe tout en précisant les degrés de petitesse requise. Citons un exemple que j'ai mentionné une autre fois. Il y a une probabilité extrêmement petite — une estimation tout à fait grossière conduit à 1 sur $10^{10^{31}}$ — que presque toute la radiation du soleil (que nous savons avoir une distribution aléatoire) se concentre soudainement en un certain lieu, tandis que la radiation sur le voisinage reste inaltérée. Si ce phénomène extraordinaire se produisait, on pourrait voir s'évaporer soudainement, disons, une arche du pont de

(²) Précisément: considérons la conjonction $A_{i_1, \ldots, i_m}$ d'une m-aine des A_i avec les négations de tous les autres. Alors $\left| \frac{m}{n} - p \right| \leqq \varepsilon$ remplace la disjonction de *tous* les $A_{i_1, \ldots, i_m}$ pour lesquels m satisfait à cette inégalité.

(³) On peut aussi conserver le B, en écrivant: $(B. \supset .| m/n - 1| \leqq \varepsilon)$ spr δ.

chemin de fer du Moerdijk. Il y a même une probabilité encore beaucoup plus petite, mais toutefois positive, que, peu après, la vapeur de fer ainsi produite se recondense exactement dans la forme originale. On peut s'imaginer la consternation des voyageurs d'un train passant peu après sur ce pont sans rien apercevoir, lorsque le soir, ils liront dans les journaux à quel danger ils ont échappé si miraculeusement!

Un autre exemple se rattache au jeu de bridge. La probabilité que le donneur des cartes recoive les treize cartes d'une couleur déterminée est environ 1 sur 635 milliards. On peut donc gager un franc contre le budget annuel d'un état quelconque, que cette éventualité ne surviendra pas. Autrement dit: si 10000 teams jouaient jour et nuit avec une vitesse d'un jeu par minute, un tel événement se présenterait en moyenne une fois en 120 ans. Néanmoins il n'y a guère de joueur de bridge ou d'amateur de ce jeu qui ne dira, et peut-être même ne *croira,* avoir assisté à un tel miracle.

Ces exemples ne peuvent qu'affirmer l'avis de d'Alembert. On ne peut pas soutenir que des probabilités si minimes, puissent être distinguées de zéro. Car leur calcul, quelque soigneusement qu'il soit fait, repose sur tant de suppositions de constance et d'indépendance des probabilités, que les valeurs numériques obtenues sont complètement illusoires.

On peut donc tâcher d'user le principe de d'Alembert pour le débrayage du modèle. Et c'est celà qui est souvent fait aujourd'hui. Il y a, cependant, une petite subtilité de nature logique dont on doit tenir compte. Supposons que l'on dise: une éventualité ayant une probabilité $\leqq 10^{-10}$ ne se réalise pas. Or, la probabilité que le joueur de bridge obtienne treize cartes déterminées mais tout-à-fait arbitraires a la même valeur d'un sur 635 milliards; donc *aucune* répartition des cartes ne se présenterait. Cette conclusion repose sur le fait que le prédicat „se réalise", attribué à une éventualité, est distributif par rapport à la disjonction. En dénotant ce prédicat par R, et les éventualités par $A_1, \ldots, A_n$, on aura:

$$R(A_1 \vee \ldots \vee A_n) = R(A_1) \vee \ldots \vee R(A_n)$$

c'est-à-dire: dire que l'éventualité (A_1 ou ... ou A_n) se réalise est équivalent à dire que l'une au moins des éventualités $A_1, \ldots, A_n$ se présente. En dénotant la négation de R par $\overline{R}$ on a donc:

$$\overline{R}(A_1 \vee \ldots \vee A_n) = \overline{R}(A_1) \wedge \ldots \wedge \overline{R}(A_n)$$

(si $A_1 \vee \ldots \vee A_n$ ne se réalise pas, alors aucune des éventualitées $A_1, \ldots, A_n$ ne se réalise). A cause de cette difficulté, on ne peut *pas* utiliser le prédicat R pour débrayer le modèle. Il nous faut pour cela un prédicat qui n'est *pas* distributif. D'autre part le débrayage consiste en le passage d'un prédicat quantitatif, par exemple $P[A] \leqq \delta$ à un prédicat qualitatif catégorique, comme $R(\bar{A})$.

On ne peut donc pas sauver la situation en disant qu'il est „extrêmement probable" ou „pratiquement certain" qu'A ne se réalise pas, puisque ces prédicats restent une fois précisés — quantitatifs. Or, un tel prédicat non distributif, satisfaisant à toutes les conditions, ne m'est connu ni dans le modèle lui-même, ni dans le métasystème sémantique qui décrit ce qui arrive en réalité. Le seul prédicat de ce type que je connaisse, se rapporte à un méta-métasystème dans lequel on inclut la conduite humaine par rapport aux événements. Par exemple le prédicat C qui exprime que quelqu'un (par exemple un individu I ou un groupe d'individus déterminés) *tient compte* de la réalisation d'une éventualité. Car, dire que je ne tiens compte ni de A ni de B n'exclut pas que je tienne compte de (A ou B). Avec ce changement, à savoir le remplacement de „A n'arrive pas" par „I ne tient pas compte de la possibilité qu'A arrive", on peut donc employer la remarque de d'Alembert pour débrayer le modèle, en l'appliquant aux éventualités de petite probabilité $\leqq \delta$.

Remarquons encore que le seuil de méfiance α („level of significance") accepté habituellement dans les sciences biologiques et médicales, à savoir $\alpha = 0{,}05$ n'est pas la borne δ que nous utilisons pour le débrayage. Lorsqu'on a une suite de N prédictions $A_1, \ldots, A_N$, faites toutes spr 0,05, on tiendra certainement compte que quelques-unes parmi elles soient erronées, et même que leur proportion soit à peut près d'une sur 20. Mais ce dont on ne tient pas compte, c'est que le nombre de prédictions erronées soit sensiblement supérieur à $0{,}05\,N$, disons, pour $N = 1000$, plus grand que $0{,}07\,N = 70$ au lieu de $0{,}05\,N = 50$. C'est précisément en réduisant le débrayage *à une seule* application du principe de d'Alembert que nous pouvons choisir un δ assez grossier (par exemple 1,5 ‰ ou même encore plus grand). Il va sans dire que, la valeur 0,05 étant une borne *supérieure* admise par convention, la plupart des degrés d'incertitude réellement obtenus, donc aussi le nombre attendu de prédictions erronées, seront sensiblement plus petits que $^1/_{20}$, pourvu, bien entendu, que *l'embrayage* ait été effectué d'une manière correcte, autrement dit, que le modèle utilisé soit bien adapté au problème empirique.

5. — Passons donc au problème de l'embrayage. Le temps disponible ne me permet pas de m'y attarder. Je dois donc me contenter de quelques remarques sans pouvoir donner une théorie complète.

Dans la plupart des applications on aura à faire à un problème empirique se rattachant à une série d'expériences éventuelles, dont les dépendances mutuelles sont négligeables. On les supposera donc mutuellement indépendantes. Une hypothèse plus restrictive consiste à les supposer avoir toutes la même fonction de probabilité. En particulier, les conditions expérimentales sous lesquelles les données ont été obtenues doivent être sensiblement les mêmes que celles auxquelles les prédictions se rapportent. Si, par exemple, un médecin a obtenu de bons résultats avec une cure en l'appliquant aux malades de sa clinique et s'il en veut déduire une prédiction sur l'application générale de cette thérapeutique,, il devra se demander si la guérison de la maladie ne dépend pas par exemple des conditions sociales des malades, et si celles-ci sont les mêmes pour les patients de son hôpital et pour les malades de tout le pays. Strictement dit, les événements observés doivent constituer un „random choice", disons sans nous arrêter à la signification précise de ce terme: *un choix „asélect"*, parmi l'ensemble de tous les malades auquel la prédiction doit s'étendre. Un autre exemple se rapporte à des essais de diètes différentes sur plusieurs groupes de poussins. Nous avons rencontré un cas où aucune différence de croissance des poulets ne pouvait être trouvée d'une diète à l'autre, tandis qu'une même diète, donnée à deux groupes différents donnait une différence de croissance bien significative. Il se faisait que les poussins d'un groupe étaient déjà un petit peu plus lourds au commencement de l'expérience. Quoique la différence des poids initiaux fut négligeable en soi-même, cette petite différence indiquait peut-être que les poussins de ce groupe croissaient constitutionnellement mieux que les autres, ce qui probablement se manifestait avec force plus tard. Si le choix des animaux eût été fait d'une manière asélecte, cette difficulté ne se serait pas présentée.

Supposons donc qu'on ait choisi un type de modèle bien adapté au problème. En général celui-ci ne déterminera pas complètement la fonction de probabilité $P[A, B]$. Au contraire, il contiendra souvent non seulement des paramètres indéterminés, mais même une fonction de distribution indéterminée. Or, l'embrayage d'un modèle unique demanderait la détermination complète de cette fonction. Souvent on fait celà en posant *l'hypothèse* que cette fonction ait une forme déterminée et en appliquant la théorie de l'essai des hypothèses de J. Neyman et Egon S. Pearson. Le principe de cette théorie, — dont nous ne pouvons pas donner aujourd'hui une exposition plus détaillée — consiste en ce qu'on détermine un ensemble (dit l'ensemble critique) de résultats possibles de l'expérience, tel que la probabilité de cet ensemble reste sous le seuil de méfiance admis, et que l'on rejette l'hypothèse *chaque fois* que les données expérimentales se trouvent dans cet ensemble. Or, lorsqu'on n'a pas réussi à rejeter une hypothèse H sous un seuil de méfiance α, qu'on adopte ensuite cette hypothèse comme si elle fût vraie et qu'on en dérive une prédiction avec un même seuil de méfiance α, la conclusion ne vaut pas spr α,, mais seulement spr 1. En effet si H n'est pas rejetée spr α, on peut seulement conclure spr α que H appartient à l'ensemble, disons Ω_1, de *toutes* les hypothèses qui ne seraient pas rejetées par les observations, si elles avaient été essayées. Or, une prédiction, valable spr α si H est vraie, peut avoir un seuil de méfiance arbitrairement près de un (ou aussi $= 1$) si une autre hypothèse appartenant à Ω_1 est vraie.

Remarquons encore que des deux relations

$$A \text{ spr } \alpha \qquad B \text{ spr } \alpha$$

on ne peut pas déduire la relation $(A \wedge B)$ spr α, mais seulement $(A \wedge B)$ spr 2α. Une conjonction d'affirmation valable spr α ne peut donc être faite qu'en tenant compte de cette restriction.

Il y a un autre principe sur lequel on peut parfois s'appuyer et qui permet *d'éviter* un embrayage complet. Soit Ω l'ensemble de *toutes* les fonctions de propabilité (hypothèses") admises. Soit Q un prédicat vrai pour les données expérimentales, et soit A le prédicat qu'on veut prédire. Ceci pourra être fait spr α lorsqu'on réussit à partager Ω en deux sous-ensembles Ω_1 et Ω_2 tels que

$$P[Q|\Omega_1] \leqq \alpha,$$
$$P[A|\Omega_2] \geqq 1 - \alpha$$

Il s'ensuit que pour *chaque* hypothèse appartenant à Ω on a

$$P[Q \wedge \neg A] \leqq \alpha,$$

c'est-à-dire

$$\neg (Q \wedge \neg A) \qquad \text{spr}\, \alpha,$$

ou

$$Q \supset A \qquad \text{spr}\, \alpha.$$

Puisque Q est supposé vrai, on obtient

$$A \qquad \text{spr}\, \alpha$$

Supposons, par exemple qu'il y ait seulement un paramètre p désignant une probabilité, qui reste indéterminé. On déterminera alors un ensemble Ω_1, sur l'axe des p, tel que pour chaque $p \varepsilon \Omega$, l'énoncé Q sur les données expérimentales soit peu probable ($P \leqq \alpha$). On rejettera donc ces valeurs de p. Si pour *toutes* les *autres* valeurs de p, constituant le soi-disant *intervalle de confiance*, l'énoncé A est très probable (valable spr α), la prédiction A vaut spr α pour *toutes* les valeurs de p.

En résumant, nous pouvons dire qu'une prédiction spr α fondée sur une hypothèse non-rejetée spr α ne vaut que spr 1, tandis qu'une telle prédiction fondée sur tout un intervalle de confiance à seuil de méfiance α vaut elle-même spr α.