JOURNAL DE LA SOCIÉTÉ STATISTIQUE DE PARIS

JSFS

Vie de la Société

Journal de la société statistique de Paris, tome 69 (1928), p. 104-108

<http://www.numdam.org/item?id=JSFS_1928__69__104_0>

V

BIBLIOGRAPHIE

Statistique mathématique, par G. Darmois, professeur à la Faculté des Sciences de Nancy et à l'Institut de Statistique de l'Université de Paris (*Encyclopédie Scientifique*). Librairie G. Doin. Un volume in-16 de 364 pages, avec 29 figures. 32 francs.

Nous ne pouvons mieux faire pour présenter ce volume que de reproduire la préface de M. Huber, directeur de la S. G. F.

PRÉFACE

La création d'un Institut de Statistique rattaché à l'Université de Paris a marqué en 1922, un sérieux progrès de l'enseignement de la statistique en France. Il n'existait guère alors qu'un seul cours spécial dans notre pays, celui de la Faculté de Droit de Paris, alors que les chaires de statistique sont nombreuses en Allemagne, en Angleterre, aux États-Unis, en Italie, non seulement dans l'enseignement supérieur, mais encore dans les écoles de commerce.

Les résultats que cette innovation permet d'escompter seront encore élargis par la publication en librairie des cours du nouvel Institut. C'est M. Émile Borel qui a professé pendant les deux premières années, le cours sur le calcul des probabilités et ses applications à la statistique. M. Darmois lui a succédé dès 1924-1925 ; le présent volume dans lequel il a réuni ses leçons sur la statistique mathématique, vient à point répondre à un réel besoin. En effet, si l'on trouve généralement dans les traités sur le calcul des probabilités un chapitre, le plus souvent assez court, sur les applications à la statistique, les ouvrages en langue française exclusivement consacrés à la méthode statistique sont plutôt rares.

Cependant, depuis une trentaine d'années, surtout à la suite des travaux de M. Karl Pearson, de nombreuses recherches ont été effectuées dans cette branche des mathématiques appliquées. Disséminées dans les publications spéciales de divers pays, elles ne sont pas toujours aisément accessibles. Les étudiants de langue anglaise, allemande ou italienne trouvent du moins dans de nombreux manuels les éléments essentiels des théories nouvelles. Grâce à M. Darmois, les lecteurs de langue française disposeront, eux aussi, d'un exposé clair, précis et assez complet pour leur permettre de diriger en toute sûreté leurs recherches ultérieures et d'aborder facilement l'étude des mémoires originaux.

Comme il existe à l'Institut de Statistique un autre cours élémentaire sur les méthodes, M. Darmois n'a point été gêné par les restrictions qui s'imposent, quand on ne veut faire appel qu'aux notions les plus simples de l'arithmétique et de l'algèbre. Ses leçons sur les applications des mathématiques supérieures à la statistique, s'adressent à des auditeurs que l'on suppose familiarisés avec les résultats essentiels du calcul différentiel et intégral.

M. Darmois s'est d'ailleurs bien gardé de tomber dans un excès opposé, en donnant à son cours un caractère trop exclusivement théorique. Sans renoncer à aucune des facilités que ménage l'emploi judicieux de l'outil mathématique, il n'a point perdu de vue que le but essentiel n'était pas d'exposer des théories mathématiques, mais de mettre à la disposition du statisticien les méthodes les plus sûres et les plus commodes pour l'analyse des masses de chiffres qu'il a patiemment recueillies et classées.

C'est, sans doute, pour mieux marquer encore cette orientation qu'il a voulu demander à un statisticien professionnel de présenter son travail au public. Cependant, l'esprit qui l'a animé se manifeste en maints passages de son livre et se dégage avec force des conclusions dans lesquelles il précise le rôle des mathématiques en statistique

Il n'est plus nécessaire actuellement de fournir des arguments pour justifier ce rôle, mais le temps n'est pas encore très éloigné où il n'était fait pratiquement, en statistique, qu'un usage modéré des ressources mathématiques. Si l'on a pu dire qu'une science n'est, après tout, qu'une langue bien faite il faut reconnaître que les statisticiens n'ont balbutié pendant longtemps qu'un bien pauvre langage et n'ont connu que des procédés rudimentaires : moyennes arithmétiques et pourcentages. Les progrès réalisés dans les théories statistiques sont surtout dus aux efforts de grands mathématiciens pour adapter les méthodes de l'analyse et du calcul des probabilités au traitement des données numériques fournies par l'observation des faits économiques, sociaux, biologiques, etc.

Toutefois, dans ce domaine, il importe de ne point se méprendre sur les possibilités et les limites d'emploi des mathématiques. Les résultats ne dépendent pas seulement de l'outil, mais de la qualité des matières traitées et aussi de l'habileté de l'opérateur dans le choix et le maniement de l'outil. Trompé par la rigueur des déductions mathématiques, on est trop souvent porté à attribuer une valeur indiscutable et universelle à tel ou tel résultat théorique, alors que pratiquement cette valeur dépend encore des conditions du problème, de la qualité des données. Et M. Darmois rappelle avec beaucoup de raison le mot de Charlier : « La statistique mathématique n'est pas un appareil automatique qui reçoit le matériel d'observation et fournit aussitôt la solution du problème. »

En statistique, comme dans toute recherche scientifique, les méthodes les plus sûres, les mieux éprouvées ne doivent pas être appliquées sans discernement. Il faut les modifier ou même les écarter si elles ne s'adaptent pas à la question traitée; à chaque pas, il faut confronter les résultats théoriques avec les faits pour se rendre compte si quelque hypothèse négligée ou introduite à tort n'a point modifié les conditions qui rendaient valables la méthode employée. Ces précautions, ces réserves sont particulièrement nécessaires en statistique, plus encore que dans beaucoup d'autres branches des mathématiques appliquées.

La nécessité de ces réserves est, à juste titre, une des principales préoccupations de M. Darmois. Son livre présente un exposé clair, ordonné et complet des travaux les plus importants sur la statistique mathématique, même des plus récents. Aucun résultat essentiel n'est omis, rien n'est négligé pour mettre en valeur la portée et la généralité des méthodes, l'ingéniosité et l'élégance de certaines démonstrations, pour faire apparaître la concordance des résultats obtenus par différents chercheurs et que masque parfois la diversité des notations. Mais on sent partout le souci constant de ne point perdre le contact avec la réalité. A chaque pas sont rappelées les précautions indispensables pour appliquer avec fruit les méthodes exposées.

Les méthodes de la statistique mathématique sont généralement présentées comme une application directe du calcul des probabilités. On nous permettra d'indiquer, à cette occasion, qu'il ne serait pas sans intérêt de faire une distinction entre les problèmes statistiques qui ne dépendent pas de la notion de probabilité et ceux qui ne pourraient être abordés sans elle.

Comme l'a fort bien dit M. Darmois au début de son introduction, toute science d'observation comprend deux parties : mise en ordre matériel des faits observés, puis mise en ordre logique pour la recherche des lois.

Or, tous les problèmes relatifs à la mise en ordre matériel des résultats numériques fournis par l'observation statistique, peuvent être traités en partant de la notion de fréquence, qui ne se heurte à aucune des difficultés théoriques, peut-être insurmontables, que l'on rencontre quand on veut rendre entièrement objective la définition de la probabilité.

Sans doute les méthodes utilisées offrent de grandes analogies avec celles du calcul des probabilités; l'examen d'une distribution statistique se fait par les mêmes procédés que l'étude d'une loi de probabilités. Ici et là, on calcule des moyennes, on mesure des écarts, on simplifie la description par l'emploi des moments, de la fonction caractéristique, etc. Mais dans le cas de fréquences statistiques, tous ces développements sont indépendants de la notion de probabilité.

Au contraire, cette notion s'introduit quand on passe à la mise en ordre logique des

résultats statistiques, à la recherche des causes. Le plus souvent, la quasi-régularité révélée par l'observation numérique des faits est apparemment due à un ensemble d'influences variées; on est conduit à rechercher dans quelle mesure elle peut être assimilée aux résultats de tirages de boules dans des systèmes d'urnes plus ou moins complexes. On cherche encore à déterminer avec quelle probabilité s'appliquent à un ensemble, les observations faites sur une partie seulement de cet ensemble, quelle confiance on peut attribuer à des prévisions déduites des constatations passées, etc.

Quelle que soit l'utilité de cette distinction des problèmes statistiques, l'appel au calcul des probabilités reste indispensable pour ceux de la seconde catégorie. C'est pourquoi M. Darmois a fait dans les trois premiers chapitres de son livre, un bref exposé, d'ailleurs excellent, des principes de ce calcul. Il a insisté sur les notions fondamentales, sans se laisser entraîner aux nombreux exemples fournis par les jeux de hasard ou les tirages dans des urnes, que semble imposer la tradition et qui relèvent plutôt de l'analyse combinatoire que du calcul des probabilités comme l'a fait remarquer avec raison M. Paul Lévy.

Les démonstrations des théorèmes de Bernouilli et de Poisson sont traitées avec tous les développements que justifie leur importance capitale dans les applications à la statistique. On saura gré à M. Darmois d'avoir cependant donné au théorème de Tchebicheff la place qu'il mérite par son élégante simplicité et qu'on regrette de ne pas lui voir toujours attribuer. Il faut encore signaler l'heureux emploi de la fonction caractéristique pour l'étude d'une loi de probabilité, notion féconde introduite par Cauchy dès 1853, utilisée par Poincaré et sur laquelle les travaux de Charlier et de Paul Lévy ont ramené l'attention.

Ces fondements posés, la seconde partie du cours, c'est-à-dire les chapitres IV et V, sont consacrés à l'étude des séries statistiques simples ou distributions d'après un seul caractère, représentées graphiquement par des polygones ou courbes de fréquence. Après avoir exposé la méthode classique d'ajustement à l'aide des moments, due à Karl Pearson, et indiqué les 7 courbes-types définies par le même auteur, M. Darmois fait, à juste titre, une large place à la représentation des courbes de fréquence par le développement en séries de fonctions orthogonales. Viennent ensuite les problèmes qui découlent de la comparaison d'une distribution statistique observée, avec celles que peuvent fournir des systèmes plus ou moins simples de tirages dans une ou plusieurs urnes : coefficient de dispersion de Dormoy et de Lexis, schémas de Poisson, de Lexis, de Borel, etc.

L'étude des séries statistiques à deux variables, c'est-à-dire des distributions d'après deux caractères simultanés, remplit les chapitres VI à IX sur un plan analogue à celui des trois précédents. C'est ici que s'introduit la notion de *corrélation* dont sir Francis Galton fut l'initiateur et Karl Pearson le principal artisan. On ne saurait prétendre que les théories développées à ce sujet aient revêtu un aspect définitif; néanmoins de sérieux progrès ont été réalisés, non seulement dans le développement des résultats, mais encore dans la clarification des concepts fondamentaux.

Ce sont les difficiles problèmes de l'hérédité qui conduisirent Galton à l'emploi d'un coefficient de corrélation pour caractériser le degré de liaison entre deux caractères observés. Si, dans les sciences physiques, l'expérimentation permet souvent de découvrir le rapport nécessaire et constant qui s'exprime par une relation fonctionnelle, dans le domaine des sciences biologiques ou sociales, on n'observe guère que des tendances, des liaisons plus ou moins lâches, intermédiaires entre l'indépendance absolue et la relation fonctionnelle. Au lieu d'un déterminisme précis qui permet un calcul exact, on n'a plus qu'une corrélation plus ou moins serrée qui ne donne qu'une approximation. Même si l'on pouvait raisonnablement supposer l'existence d'une relation fonctionnelle, il serait presque toujours impossible de la déceler, parce qu'on ne pourrait l'isoler par l'expérimentation, la séparer des influences perturbatrices dont les effets troublent inéluctablement les observations.

Galton étudiant la relation entre les tailles des parents et celles de leurs descendants introduisit son coefficient de corrélation r de forme identique à celui que Bravais avait obtenu dans l'étude des erreurs de situation d'un point sur un plan. Comme toute distribution statistique d'après deux caractères peut être représentée par un

tableau à double entrée ou par des masses ponctuelles réparties sur un plan, l'introduction dans l'étude statistique des écarts types (écarts moyens quadratiques) et du coefficient de corrélation est aussi naturelle que celle des moments d'inertie dans le problème mécanique corespondant.

Le coefficient de corrélation définit la direction du plan qui correspond à la plus grande concentration des masses représentatives. Il est nul si le groupement est le même dans toutes les directions autour du point central (centre de gravité). Il tend vers 1 en valeur absolue si tous les points s'alignent sur une droite.

Mais on s'aperçoit bien vite que ce coefficient ne fournit pas toujours une bonne appréciation du degré de liaison entre les deux caractères considérés. Des relations fonctionnelles de forme convenable peuvent, en effet, donner au coefficient r toute valeur inférieure à 1, y compris zéro.

Il fallait donc trouver autre chose. On peut imaginer bien des manières de déterminer le degré de liberté de l'association entre deux caractères qui résulte d'une table statistique à double entrée. Pour étudier la corrélation entre les âges des nouveaux époux, par exemple on peut suivre la variation de l'âge moyen des femmes qui épousent des hommes appartenant à des groupes d'âge croissant, ou inversement; on peut encore étudier la différence d'âge des époux, etc.

Pearson a défini un *rapport de corrélation* qui varie en valeur absolue de 0 à 1 comme le coefficient r et résulte de la comparaison entre l'écart type de la distribution tout entière et les écarts types des diverses lignes ou des diverses colonnes du tableau. Ce rapport s'annule quand toutes les lignes (ou toutes les colonnes) présentent le même écart type égal naturellement à l'écart type de l'ensemble. Il est égal à 1 quand tous les écarts types s'annulent, il y a relation fonctionnelle, les masses représentatives sont toutes sur une même courbe.

Le rapport de corrélation résout donc bien le problème posé, quand les deux caractères associés sont mesurables, c'est-à-dire qu'on peut caractériser chacun d'eux par un nombre variable. C'est le cas par exemple pour les tailles, les âges, etc. Mais on doit opérer parfois sur des caractères qualificatifs (couleur des cheveux ou des yeux, etc.) ou sur des répartitions dont l'ordre n'est pas imposé par la nature des choses (divisions territoriales d'un pays, etc.). L'indice qui sert à apprécier le degré de dépendance, de corrélation, doit donc être indépendant de l'ordre des colonnes ou des lignes du tableau à double entrée; il doit pouvoir servir à l'étude de distributions ne comprenant que quelques lignes ou quelques colonnes.

Mieux adapté à ces cas que le rapport de corrélation est le *coefficient de contingence* défini aussi par Pearson à partir du moyen carré de contingence.

Ces quelques observations peuvent permettre de se représenter simplement les données fondamentales de la corrélation et la nature des coefficients les plus fréquemment employés. La question est traitée par M. Darmois avec tous les développements qu'elle comporte, notamment en ce qui concerne la loi de corrélation normale (qui joue dans le cas de deux variables un rôle analogue à celui de la loi normale de Laplace-Gauss pour une variable), la fonction caractéristique à deux variables, les coefficients de corrélation partielle, les généralisations possibles dans le cas de distributions d'après trois caractères ou plus, etc.

Le dernier chapitre est consacré à l'examen critique des solutions proposées pour la comparaison de deux ou plusieurs séries statistiques, le plus souvent de séries qui représentent les mouvements de deux phénomènes dans le temps. Il s'agit de rechercher s'il y a une relation plus ou moins nette entre ces mouvements, une ressemblance plus ou moins parfaite entre les courbes représentatives. L'affirmation ne permet naturellement pas de conclure qu'il y a une liaison directe entre les deux phénomènes : on pourrait citer de nombreux cas de ressemblance parfaite entre les mouvements de phénomènes de toute évidence indépendants. Elle suggère seulement la possibilité d'une liaison, dont l'existence et la nature ne peuvent pas être précisées par la statistique, mais exigent des recherches directes.

Le coefficient de Galton a été utilisé, peut-être avec quelque précipitation, comme indice de ressemblance de deux courbes statistiques. Peut-être conviendrait-il, comme M. Darmois le suggère, d'éviter toute ambiguïté, en ne reprenant pas ici le

mot de corrélation réservé au degré d'association de deux ou plusieurs caractères d'un même fait et de le remplacer par un autre, celui de *covariation* par exemple, proposé par M. March.

Par une analyse serrée qui constitue une des parties les plus originales de son ouvrage, M. Darmois a montré toute la complexité du problème et l'insuffisance du coefficient de Galton qui, dans ce cas, permet seulement d'apprécier jusqu'à quel point les deux phénomènes peuvent être considérés comme grossièrement liés par une loi de proportionnalité. Le coefficient n'a donc de sens que s'il est très voisin de l'unité; sinon le résultat signifie simplement que l'hypothèse implicitement admise est inacceptable.

En fait, l'étude de la ressemblance de deux courbes statistiques exige une analyse plus complète des mouvements qu'elle représente. On peut souvent distinguer : le mouvement d'ensemble sur une longue période, que l'on appelle parfois tendance séculaire; les mouvements oscillatoires de part et d'autre de la droite ou de la couche moyenne qui caractérise cette tendance séculaire. Dans ces mouvements, il est parfois possible d'apercevoir des périodes assez nettes (variations saisonnières ou mensuelles, hebdomadaires, journalières, etc.), et aussi des régularités quasi-périodiques à périodes plus ou moins longues comme les cycles économiques.

Par des méthodes appropriées, on peut éliminer les oscillations dont la période est sensiblement fixe (variations saisonnières); mais les autres sont difficilement séparées des mouvements purement accidentels. En fait, ce problème ardu est encore loin d'une solution complète, s'il doit jamais en recevoir une, pleinement satisfaisante.

On voit que si M. Darmois a su exposer avec une remarquable clarté les solutions acquises, il n'a pas manqué de soumettre à une pénétrante critique, celles qui semblent insuffisamment adaptées aux besoins de la pratique. Il aura rendu ainsi un double service.

Placer à la portée du public français un exposé bien coordonné de travaux un peu épars et insuffisamment connus dans notre pays, était son but essentiel; il l'a pleinement atteint. Non seulement il a su mettre en relief les résultats les plus importants des Pearson, des Charlier et des Tschouprow, pour ne citer que quelques noms; mais il a su les grouper et montrer leurs liens, pas toujours très apparents. Le lecteur qui aura assimilé la substance de son livre, abordera sans difficulté les mémoires de plus en plus nombreux dans cette branche des mathématiques appliquées; il pourra les étudier avec fruit, en les situant à leur place exacte dans l'ensemble de nos connaissances sur ce sujet.

Mais le livre de M. Darmois ne sera pas moins utile s'il contribue efficacement à répendre chez nous le goût de ces études, à raviver l'intérêt qu'on a toujours porté, dans le pays de Laplace, de Poisson et de Cournot, aux applications du calcul des probabilités à la statistique. Les nouvelles recherches dont il montre l'importance et la nécessité, profiteraient à toutes les branches de l'activité scientifique où s'impose de plus en plus l'étude numérique des fait collectifs, qu'il s'agisse de la physique moléculaire ou atomique, de la biologie, de la psychologie appliquée ou des sciences économiques et sociales.

Michel HUBER.

Le Gérant : A. VALANTIN.

IMPRIMERIE BERGER-LEVRAULT, NANCY-PARIS-STRASBOURG. — 1928