ANNALES DE L'I. H. P.

G. DARMOIS

La méthode statistique dans les sciences d'observation

Annales de l'I. H. P., tome 3, n° 2 (1932), p. 191-228

<http://www.numdam.org/item?id=AIHP_1932__3_2_191_0>

La méthode statistique dans les sciences d'observation

PAR

G. DARMOIS

Introduction. — Le but de ces leçons est de mettre en évidence quelques idées essentielles de la statistique et le caractère de cet ordre spécial qu'elle peut mettre dans certains ensembles de résultats expérimentaux.

Un ensemble de recherches se présente généralement, soit comme suite à un désir de vérifier, d'approfondir et de lier un groupe d'idées, soit comme imposé par son importance pratique, par son indiscutable présence.

A mesure que l'observation progresse dans le domaine qu'elle a choisi, les résultats qu'elle accumule sont alors placés, soit dans les cadres parfois assez vagues d'une sorte de théorie préalable, soit dans un ordre de pure commodité, généralement sillonné de quelques filaments logiques.

Cette organisation provisoire, description commode du matériel observé, permet à l'esprit de dominer avec plus d'aisance des ensembles étendus.

On cherche ensuite, ce que d'une manière générale, nous appellerons des permanences (1), certaines relations qui demeurent constantes, et qui constituent des lois empiriques des phénomènes observés.

Par exemple, les premières permanences qui soient apparues étaient relatives à la présence et aux relations de dimensions des corps naturels, et certaines permanences dans les mouvements des astres. Beaucoup plus tard seulement, on apercevra l'ordre dans la Mécanique.

(1) DIVISIA, *Economique rationnelle*, chap. I.

Ces lois empiriques une fois obtenues, les démarches suivantes de l'esprit visent à les comprendre, les expliquer, les relier entre elles.

Parallèlement à l'ordre descriptif se développent l'ordre explicatif, logique, les théories et les lois scientifiques.

Ce développement, qui se construit à partir de certaines notions à la base, doit alors retrouver les résultats de l'expérience.

Il est clair que cette opération de modelage de l'information scientifique dépend de la masse et de la qualité de cette information relative à un sujet donné, qu'elle doit tenir compte des relations qui se manifestent entre des sujets d'abord jugés différents. Ainsi la construction théorique peut être amenée à se modifier, et se modifiera généralement au bout d'un temps plus ou moins long.

Bien entendu, la valeur logique de cette construction n'a pas changé, seulement les notions à la base ont pu se révéler trop grossières, insuffisantes ou inexactes devant des faits nouveaux. La valeur explicative de ces notions disparaît, et le sens même du mot explication peut se trouver profondément modifié.

Par exemple, les permanences de la géométrie expérimentale, la comparaison des grandeurs qu'on y rencontre, et qui trouvent leur place dans le tissu logique de la géométrie, en reçoivent une explication macroscopique, mais qui perd tout son sens dans une conception discontinue de la matière.

A ces moments, on cherchera généralement si les anciens concepts sont capables d'un prolongement, d'une utilisation à échelle différente, ou s'il est devenu nécessaire d'introduire des concepts nouveaux.

C'est ainsi qu'on essaiera des particules de l'ancienne géométrie se mouvant comme l'indique l'ancienne mécanique, jusqu'au moment où cette position sera devenue intenable en présence des faits nouveaux.

Les Permanences fonctionnelles. — Les premières permanences rencontrées par la science moderne avaient un caractère spécial, on peut dire qu'elles étaient rigides, analogues à celles qu'avait étudié la géométrie. Considérons par exemple une masse gazeuse ; une fois acquises les notions de pression, et de température, il apparaît qu'on ne peut fixer ces deux grandeurs sans fixer en même temps l'état de la masse gazeuse. Son volume est fixé. Ces trois grandeurs d'état sont liées rigidement, en ce sens que deux d'entre elles fixent l'état, et déterminent

la troisième. Ces permanences sont fonctionnelles, leur correspondant abstrait étant la notion de fonction ; l'observation impose que le volume est fonction de la pression et de la température, elle donne des valeurs numériques.

L'observation de la chute libre d'un corps impose des conclusions analogues. Dès qu'on sait mesurer le temps convenablement, on voit que l'espace parcouru à partir du point où le corps est lâché sans vitesse est toujours le même pour le même temps de chute. Il y a là une permanence fonctionnelle. L'observation trouve que la fonction a une expression mathématique simple.

Plus généralement, la position au bout d'un certain intervalle de temps ne dépend que de la position et de la vitesse au début de l'intervalle. D'ailleurs on peut mettre tous ces résultats sous la forme simple d'une permanence d'accélération, et généraliser ce résultat aux équations différentielles de la Mécanique rationnelle, qui, partant de ses notions classiques, retrouve à la fin de son développement logique les résultats de l'observation.

Ainsi, disait LAPLACE, l'esprit humain a pu comprendre dans les mêmes expressions analytiques les états passés et futurs du système du monde.

C'est la permanence fonctionnelle de l'espace-temps ou, si l'on veut, la rigidité de l'Univers, conception simple et grandiose qui ne traduit, bien entendu, que le succès momentané d'un système d'explication.

Pour les gaz, si l'on s'en tient aux notions de pression, densité, température, le lien fonctionnel résulte de ce que ces notions sont surabondantes, mais cette explication s'arrête là.

Les permanences statistiques. — Dès la fin du XVII^e siècle, on a vu paraître des régularités d'un ordre tout différent. A Londres, où dès le XVI^e siècle on dressait des statistiques municipales (1), des esprits observateurs tels que JOHN GRAUNT et WILLIAM PETTY remarquèrent un certain ordre dans un grand nombre de faits démographiques et sociaux. Certains rapports de mortalité, de natalité apparaissent comme peu variables.

Cette constance approchée de certains rapports, calculés sur de grands nombres, nous dirons que c'est une permanence statistique

(1) HUBER, *Les Méthodes de la Statistique*, Conférence à l'École Supérieure des P. T. T. *Annales des P. T. T.*, avril 1927.

simple. Les jeux de hasard, tels que pile ou face, en donnent de nombreux exemples.

Une des permanences remarquables de la démographie, signalée et étudiée depuis longtemps, est celle du rapport des naissances masculines au nombre total des naissances dans une population déterminée. Quand les observations sont soigneusement faites, ce rapport se trouve d'une façon remarquablement stable, voisin de 0,51 ; 0,515. On ne connaît pour le moment aucune explication satisfaisante.

Les permanences mendéliennes de l'hybridation. — Les croisements entre races très voisines, végétales ou animales, fournissent des exemples saisissants, qui par leur ensemble et leurs conséquences sont une des plus belles acquisitions de la biologie (1). Dans le cas le plus simple (2), où l'on croise deux races soient A et B, différant par un seul caractère, le premier croisemen donne des individus C que nous supposerons différents de A et de B. Si l'on croise entre eux les individus C, et qu'on obtienne une deuxième génération assez nombreuse, elle comprend des A, des B et des C. Les proportions sont alors très voisines de

$$\frac{1}{4} \text{ pour A}, \qquad \frac{1}{2} \text{ pour C}, \qquad \frac{1}{4} \text{ pour B}.$$

Dans un cas réel, on a trouvé sur 158 produits :

41 A 78 C 39 B

Ces permanences statistiques, étendues à des cas plus complexes, constituent les lois mendéliennes de l'hybridation.

La radioactivité. — La désintégration des noyaux radioactifs présente également des régularités statistiques (3). Pour un élément déterminé la fréquence relative A du nombre d d'atomes désintégrés en un temps donné est un nombre stable. Les constantes de désintégration qu'on peut exprimer par leur inverse, la vie moyenne, varient avec l'élément dans des proportions énormes. La vie moyenne peut aller en effet de milliards d'années à des fractions de secondes.

On trouve là, développées sur une échelle immense, des permanences statistiques fournies par des mécanismes naturels, sur lesquels nos

(1) F. GUYENOT, *L'Hérédité*. G. Doin et Cº.
(2) *Loc. cit.*, p.p 56-57.
(3) SODDY, *Le Radium*, Alcan 1926.

moyens n'ont d'ailleurs aucune action, et dont l'explication pose, nous allons le voir, un problème particulier (1).

L'ordre logique dans les permanences simples. — Le résultat de l'observation est que, pour certains rapports calculés sur de grands nombres, il est très exceptionnel de les voir s'écarter notablement d'un nombre fixe. Pour retrouver logiquement un tel résultat, il faut utiliser un procédé moins rigide que la déduction qui, elle, donne des certitudes. JACQUES BERNOULLI y parvint le premier dans son ouvrage fondamental, *Ars Conjectandi*, en utilisant les probabilités et les quasi-certitudes.

Tout comme la géométrie ou la mécanique rationnelle déroulent leur ordre logique à partir des notions abstraites de point, droite, point matériel, masse, force, la notion abstraite de probabilité est à la base du nouvel ordre logique (2). Dans l'urne abstraite où les boules blanches et noires se tiennent en proportions p, q, la probabilité d'extraction d'une blanche est p. S'il n'y a que des blanches, c'est la certitude. S'il y a une noire sur 100.000, c'est la quasi-certitude à 1/100.000.

BERNOULLI a montré par quel mécanisme on pouvait arriver à des quasi-certitudes. Si, par exemple, les proportions sont égales, et qu'on observe la fréquence relative des blanches apparues sur n tirages, on peut se fixer un intervalle quelconque autour de la fréquence vraie dans l'urne. Supposons l'intervalle 0,49 à 0,51. On peut demander qu'il soit exceptionnel d'en sortir, ou se fixer une quasi-certitude d'y être compris, fixons-la à une valeur quelconque, par exemple 1/1000. BERNOULLI démontre qu'on peut prendre n assez grand pour que ces deux conditions soient remplies. Autrement dit, si la série d'épreuves totalisées est assez longue, on peut être aussi sûr qu'on le désire, de voir la fréquence approcher autant qu'on le veut de la fréquence vraie. On voit que cette notion abstraite de l'urne identique à elle-même et dans laquelle on fait de nombreux tirages, trouve une application dans les permanences simples, et peut être essayée comme explication de ces permanences.

Par le même procédé s'expliquent les résultats d'une longue série de pile ou face, et cette explication non seulement réussit pratiquement, mais nous satisfait, et nous donne le sentiment d'un véritable progrès.

(1) BOREL, *Éléments de la théorie des probabilités*. Hermann, 1924, Note I, p. 183.
(2) BOREL, *Loc. cit.*, p. 184.

Pourquoi ? Evidemment, parce que la valeur numérique 1/2 attribuée à la probabilité de face, résulte pour nous d'une analyse exhaustive des conditions du jeu, structure de la pièce, lancement, mouvement, arrivée.

Il est clair que si nous savions que le centre de gravité est très voisin du côté face, et que nous représentions les observations en admettant la probabilité 1/2 pour face, cela pourrait passer momentanément pour une description assez bonne, mais non pour une explication.

Permanences statistiques à une variable. — Il peut arriver que le caractère soumis à l'observation soit capable de diverses valeurs. (Un des exemples les plus célèbres est l'étude des erreurs expérimentales.) Dans une longue série, à chaque valeur est attachée la fréquence relative des épreuves qui ont fourni cette valeur. Il arrive fréquemment que diverses séries, pourvu qu'elles soient longues, fournissent des fréquences stables pour les diverses valeurs. On a donc un ensemble de permanences simples qu'on peut appeler permanence statistique à une variable.

L'étude des erreurs expérimentales conduit parfois à de telles permanences ; on en trouve de très nombreuses, relatives aux caractères les plus divers, dans l'œuvre de QUETELET.

Mais on les rencontre déjà dans l'étude des permanences simples. Dans une série de rapports calculés sur les mêmes populations totales, et manifestant un groupement autour d'un certain nombre, on trouve fréquemment une régularité nouvelle dans la structure même de ce groupement.

Des écarts d'une grandeur fixée reparaissent en conservant à peu près la même fréquence.

Il y a permanence d'une fonction de fréquence, ou d'une courbe de fréquence.

La notion abstraite correspondante. — Considérons une grandeur à qui certaines hypothèses et limitations *a priori* permettent un ensemble d'états déterminés, chacun de ces états possédant une probabilité correspondante.

Une telle grandeur est dite aléatoire, et la correspondance des états à leurs probabilités constitue sa loi de probabilité.

Par exemple une urne renfermant en proportions déterminées des

boules de cinq couleurs différentes, définit une grandeur aléatoire, la couleur de la boule extraite, capable de cinq états, les probabilités étant les fréquences relatives de chaque système de boules dans l'urne.

Le résultat fondamental de Jacques BERNOULLI revient à considérer dans n tirages la variable aléatoire nombre des succès, ou fréquence relative des succès, capable de $n + 1$ valeurs ; la loi de probabilité est fournie par la règle du binôme. Jacques BERNOULLI a commencé l'étude de cette loi de probabilité et montré qu'un intervalle étant choisi qui encadre la fréquence vraie, la probabilité totale de tomber dans cet intervalle est une quasi-certitude arbitrairement fixée, si n est assez grand. Mais sa méthode ne donne que ce résultat global. Abraham DE MOIVRE qui reprit la question après lui, démontra que la courbe du binôme pouvait être remplacée dans la partie intéressante par une courbe dont le seul paramètre est la racine carrée de n. Ainsi paraissait la première permanence d'une courbe de fréquence (c'était la courbe dite maintenant de LAPLACE-GAUSS), la première explication possible des permanences statistiques à une variable.

Description et explication des permanences. — Imaginons maintenant que nous ayons devant nous une série de coups à pile ou face, par exemple 100 parties de 100 coups chacune. Nous aurons 100 valeurs pour la fréquence de face, 100 valeurs groupées au voisinage de 1/2. La moyenne générale est 0.506.

On peut décrire cet ensemble en disant : Tout se passe comme si la probabilité de face était 1/2. Les 100 nombres sont répartis suivant la courbe de GAUSS, avec le paramètre déduit de la racine carrée de 100. On dira que les résultats de l'observation se conforment au schéma de la probabilité 1/2.

On peut aller plus loin ; la pièce qui a servi au jeu est examinée, elle est trouvée très légèrement dissymétrique et les conditions de lancement sont correctes. Imaginons que la position du centre de gravité soit très bien connue, et conduise à une probabilité de face 0,501.

On peut alors expliquer la série en disant :

Il est raisonnable, en présence de la structure de la pièce et des conditions de lancement, d'admettre 0,501 pour la probabilité de face. Les résultats de l'observation se conforment ici à un schéma déterminé par le raisonnement et l'expérience.

Les lois de l'hybridation. — Quelle description pouvons-nous en donner ? On peut toujours tenter un schéma d'urne. Tout se passe-t-il comme si les individus A B C étaient puisés dans une urne dont la composition serait fixée par les proportions $\frac{1}{4}, \frac{1}{4}, \frac{1}{2}$?

Il faut pour cela que les fluctuations observées suivent la loi de la théorie des probabilités. C'est vrai. Nous avons donc une description. Mais il est naturel de chercher à expliquer cette urne, en ramenant sa structure, son mode de remplissage à quelque chose de plus simple.

Il est clair que si, d'une urne renfermant des proportions égales de deux sortes de boules, on extrait des couples, les trois sortes de couples se présenteront avec les probabilités $\frac{1}{4}, \frac{1}{4}, \frac{1}{2}$. Voilà donc une description qui va plus loin.

Si l'on essaie maintenant de découvrir dans la réalité ce qui correspond à ces couples, ce que la théorie des chromosomes croit avoir trouvé dans les noyaux des cellules reproductrices, on construit une explication des permanences statistiques observées, à partir de l'urne à chromosomes, considérée comme un fait.

La radioactivité. — On peut essayer le schéma et dire : tout se passe comme s'il y avait une certaine probabilité de désintégration. On représente ainsi très bien les résultats. On est même amené à dire, tout se passe comme s'il y avait une certaine probabilité de désintégration, la même pour chaque atome, et à chaque instant. Dans chaque atome se trouve l'urne et se font les tirages. A chaque élément radioactif correspond une urne spéciale. Voilà la description.

L'explication paraît bien difficile. Il faudrait, par une démarche logique qui se fonde sur une structure possible du noyau, rendre compte de ces évasions fortuites (1). On ne peut obtenir ce résultat avec une mécanique déterministe du noyau. Il faut expliquer l'urne qui schématise le phénomène par une autre urne, l'urne des états possibles du noyau. Cette substitution peut être faite à l'aide des mécaniques nouvelles (2) et paraît avoir donné de beaux résultats ; sans doute elle explique un résultat statistique par une théorie à base statistique, mais elle met de l'ordre dans les résultats, et unifie les différents

(1) BOREL, *Loc. cit.* p. 188.

(2) Voir par exemple A. HAAS, *La mécanique ondulatoire et les nouvelles théories quantiques.* Traduction française chez Gauthier-Villars, Paris.

schémas. Même si cette mise en ordre devait conserver le caractère statistique à la base de la nouvelle mécanique, elle réaliserait un progrès d'importance et comporterait un véritable élément explicatif, nettement différent d'ailleurs de ce que nous avons rencontré jusqu'ici.

Permanences statistiques à plusieurs variables. — L'étude simultanée de deux ou plusieurs caractères, faites sur les divers éléments d'une population, conduit à une notion très importante, qui généralise largement l'idée de liaison entre grandeurs physiques.

Fixons l'un des caractères g à un état de grandeur déterminée g_1, et déterminons les états de l'autre caractère qui lui sont associés. Nous trouvons une répartition de fréquence pour le deuxième caractère γ. Cette répartition, déterminée en même temps que g_1, peut varier quand on considère les différents états de la grandeur g. La population est généralement étudiée par fragments dans l'espace, ou dans le temps. Si les répartitions changent peu, on a une permanence statistique à deux variables, et si la répartition de γ pour g fixé, change avec la valeur de g, on dit qu'il y a corrélation entre g et γ.

Un exemple important est fourni par l'astronomie stellaire. Les étoiles variables périodiques qu'on appelle Céphéides ont une période bien déterminée, d'ailleurs très variable de l'une à l'autre. Si on met en regard de cette période la luminosité propre ou la grandeur absolue de l'étoile, on constate que l'ensemble des points ne dessine pas la courbe d'une liaison fonctionnelle, mais que la répartition des grandeurs des étoiles de période connue, varie très nettement avec cette période. Ce nuage de points qui s'allonge autour d'une courbe représente la corrélation entre la grandeur absolue et la période des Céphéides.

La notion abstraite correspondante. — On l'obtient par la considération simultanée de deux grandeurs aléatoires définies dans un même champs, soient G et Γ.

A chaque couple d'états $G_i\Gamma_k$ correspond une probabilité p_{ik}.

Si l'on fixe la valeur G_1 et qu'on fasse varier k, on a une loi de probabilité de la variable liée Γ, loi qui est évidemment une fonction de la grandeur G, et qui peut varier ou non quand varie G.

On dit, quand cette loi varie, qu'il y a liaison stochastique entre les deux grandeurs aléatoires.

Quand elle ne varie pas, il y a indépendance stochastique.

Le lien stochastique est le correspondant abstrait de la corrélation. Il est évident que ces notions s'étendent à un nombre quelconque de variables, le lien stochastique résultant du fait que la loi de probabilité d'un groupe de variables, quand les autres variables ont des valeurs fixées, change avec ces valeurs.

Description et explication des permanences à une ou plusieurs variables. — Une loi de probabilité à une variable peut toujours être obtenue par une urne ayant la variété et la composition requises.

Une loi de probabilité à deux variables peut l'être de la même manière en inscrivant sur chaque boule une combinaison $G_i\Gamma_k$ et fixant les proportions convenables.

Tout tirage effectué dans ces urnes fournit des permanences statistiques, avec fluctuation suivant certaines lois autour de la composition réelle de l'urne, les lois des fluctuations variant d'ailleurs avec le mode de tirage. Nous aurons obtenu un schéma d'urnes, une description des résultats expérimentaux, si nous pouvons trouver une urne et un mode de tirage qui soient aptes à redonner les observations.

Il y aura élément explicatif si la loi de probabilité a une ou plusieurs variables, si le mode de tirage, peuvent être dans une certaine mesure, déduits d'hypothèses sur le mécanisme du phénomène observé.

Par exemple, si l'urne peut être considérée comme composée à l'aide d'urnes plus simples, ou seulement par un mécanisme qui a son correspondant dans le phénomène étudié, nous réalisons un progrès sur la description pure, nous aurons une ébauche plus ou moins poussée de théorie.

***

Permanences à une variable. — Schémas d'urnes. — Les premières observations à représenter conduisaient naturellement à la considération de variables aléatoires d'un type spécial, le nombre total des succès relatifs à un évènement attendu. Le succès lui même est une grandeur aléatoire qui prend à chaque épreuve la valeur 1 ou 0.

On était donc amené à additionner un grand nombre de variables aléatoires, et à étudier la loi de probabilité de leur somme. C'est ce pro-

blème qui a fourni les résultats essentiels de la statistique mathématique, et qui est appelé sans doute à en fournir bien d'autres.

Sous la forme où le présentent les permanences statistiques simples, chacune des variables aléatoires, soit x_i correspondant à l'épreuve de rang i a seulement deux valeurs 1 et 0. Les probabilités correspondantes p_i q_i peuvent être prises arbitrairement, mais puisqu'on admet que les épreuves se succèdent, l'épreuve de rang i qui fournit la variable aléatoire x_i fait intervenir la loi de probabilité liée de x_i, toutes les valeurs précédentes étant supposées connues.

La première recherche, celle de BERNOULLI, supposait que toutes ces variables sont indépendantes entre elles, et qu'elles ont la même loi de probabilité.

Les résultats sont très simples. Le nombre des succès est :

$$z = x_1 + x_2 \cdots + x_s$$

la fréquence relative des succès dans n tirages :

$$f = \frac{z}{s}.$$

E étant le symbole de l'espérance mathématique, on a :

$$\mathrm{E}(f) = p, \qquad \mathrm{E}(f - p)^2 = \sigma_f^2 = \frac{pq}{s}.$$

La quantité importante est :

$$\varphi = \frac{f - p}{\sigma_f}$$

écart réduit de la fréquence relative. Le résultat obtenu par DE MOIVRE est que la probabilité p_t de l'inégalité : $|\varphi| \leqslant t$ est très voisine, quand n est grand, de sa valeur limite pour $n = \infty$:

$$\frac{1}{\sqrt{2\pi}} \int_{-t}^{+t} e^{-\frac{t^2}{2}}\, dt.$$

C'est cette dernière valeur qui fixe la quasi-certitude. Il suffit de choisir t assez grand. Quant à la valeur la plus grande de $|f - p|$, soit $t\sqrt{\frac{pq}{s}}$, il suffit de prendre s assez grand pour rendre cette quantité aussi petite qu'on veut.

t fixe donc la quasi-certitude. La précision $t\sqrt{\frac{pq}{s}}$ une fois t fixé, varie comme la racine carrée de s.

Nous avons rappelé ces points classiques pour insister sur une remarque importante de K. PEARSON.

Nous voulons obtenir par la théorie des probabilités une image des permanences statistiques. Il faut pour cela serrer de près l'évaluation du nombre n. S'il ne s'agit que de démontrer la loi des grands nombres, on peut se contenter avec TCHEBICHEF de remarquer que σ_f tend vers zéro avec $\frac{1}{s}$, mais on n'expliquerait pas ainsi la petitesse des fluctuations, ni la permanence à une variable qu'elles présentent.

Cette série de s tirages indépendants dans une urne invariable, répétée un nombre N de fois, constitue le schéma de BERNOULLI. Il est rare qu'on puisse l'appliquer de manière satisfaisante à des séries observées, ces séries ne sont pas, comme on dit, normales.

Pourtant, les phénomènes de désintégration, qui nous donnent le plus bel exemple d'urne invariable, devraient en relever, mais ils correspondent à des valeurs extrêmement petites de la probabilité p, de sorte que les hypothèses nécessaires à la validité de la démonstration ne sont plus vérifiées, la courbe binomiale présentant, même pour de grandes valeurs de s, une dissymétrie marquée.

La variable qui compte ici, c'est le nombre z des désintégrations, dont l'espérance mathématique est sp. On admet que ce produit est fini quand s est grand, p petit. On trouve alors, comme POISSON l'a montré, une distribution discontinue voisine de la suivante, ou loi des petits nombres ;

$$P_r = e^{-\lambda}\frac{\lambda^r}{r!}$$

où $\lambda = sp$. P_r est la probabilité de r succès.

La représentation dans ces conditions des nombres observés de désintégrations radioactives se fait bien.

Séries où l'urne varie, les tirages restant indépendants. — POISSON avait eu l'idée de varier l'urne à chaque fois dans la série des tirages. Dans ces conditions on a :

$$E(f) = \frac{p_1 + p_2 \cdots + p_n}{s} = p_0$$

$$E(f - p_0)^2 = \frac{p_1 q_1 \cdots + p_s q_s}{s^2}.$$

Si cette opération est reprise N fois, on a une série de nombres qui manifestent les deux permanences. Ils sont voisins de p_0 et la série des écarts a une courbe de fréquence voisine de la courbe de GAUSS, si s est grand, mais ici le paramètre de cette courbe est :

$$\frac{1}{\sqrt{s}}\sqrt{\frac{p_1q_1 + \cdots + p_sq_s}{s}}.$$

Il est bien clair que le voisinage du nombre p_0 peut s'obtenir avec n'importe quelle collection d'urnes de moyenne p_0, en particulier avec S urnes p_0 ; le paramètre de dispersion qui fixe la loi des fluctuations varie avec cette collection d'urnes. Il a sa plus haute valeur quand les urnes sont identiques, il serait nul si sp_0 parmi les urnes ne contenaient que des blanches, les s $(1 - p_0)$ qui restent contenant des boules noires. C'est ce qu'on exprime en disant qu'à moyenne égale, une série de POISSON est moins dispersée qu'une série de BERNOULLI.

Ces séries de POISSON ne se trouvent guère dans la réalité.

Ce qu'on est appelé à rencontrer, c'est évidemment une évolution lente de l'urne.

Le schéma de Lexis. — Nous faisons N séries des tirages indépendants, mais cette fois l'urne qui reste la même pendant les s tirages, change d'une série à l'autre parmi les N.

Les phénomènes sont alors tout à fait différents.

La fréquence f_1 a comme valeur probable p, comme écart type $\sqrt{\frac{p_1q_1}{s}}$; es N nombres aléatoires f_i ne sont plus N valeurs d'une même variable aléatoire ; à la dispersion propre à chacun d'eux, s'ajoute la dispersion des p_i, ou dispersion des urnes.

Cette dispersion des f_i peut s'étudier par la variable aléatoire, écart quadratique moyen des f_i autour de leur moyenne, ou mieux par le carré de cet écart, dont l'espérance mathématique peut être désignée par σ_L^2.

Il est clair qu'on aura, si les p_i ne sont pas trop dispersés, un groupement autour de leur moyenne, et qu'on aura la même moyenne pour une infinité de systèmes d'urnes, mais ici la valeur de σ_L^2 dépasse $\sigma_B^2 = \frac{p_0q_0}{s}$. On a :

$$\sigma_L^2 = \sigma_B^2 + \left(1 - \frac{1}{s}\right)\sigma_p^2, \qquad \sigma_p^2 = \frac{\Sigma(p_i - p_0)^2}{N}$$

σ_p indiquant la dispersion des urnes. On voit que le rapport :

$$Q^2 = \frac{\sigma_L^2}{\sigma_B^2} = 1 + \left(1 - \frac{1}{s}\right)\frac{\sigma_p^2}{\sigma_B^2}$$

est plus grand que l'unité. C'est le cas de la très grande majorité des séries statistiques réelles, qui sont plus dispersées que les séries de BERNOULLI donnant même moyenne, ou qui sont, comme on dit, hypernormales.

Ce schéma de LEXIS est très souple, il convient à la représentation d'une série hypernormale quelconque, il suffit de choisir σ_p^2 à peu près égal à :

$$\frac{p_0 q_0}{s}(Q^2 - 1).$$

Mais si l'on veut que le schéma ait une certaine valeur, il faut que les urnes aient quelque réalité.

Par exemple, la mortalité en Allemagne de 1901 à 1910 (1) est $\frac{19,7}{1\,000}$, la population est 60, 737 millions.

En Saxe, la mortalité est la même avec une population de 2.975 millions. Les rapports Q sont respectivement 82,5 et 17,3.

Si le système d'urnes est le même, on devrait avoir :

$$\frac{(82,5)^2 - 1}{60.737} = \frac{Q^2 - 1}{2,975}$$

Q étant le rapport pour la Saxe. Q calculé par cette formule serait 18,3, la valeur réelle est 17,3. Le schéma paraît satisfaisant.

Le Schéma d'urnes de Borel. — On peut obtenir des séries hypernormales par un procédé plus simple, mais de moindre souplesse. Il suffit, dans une série normale de supposer que chaque tirage compte pour K. Le nombre des tirages ne sera plus que $\frac{S}{K}$ et l'écart type sera $\sqrt{K}\sqrt{\frac{pq}{s}}$. Il est multiplié par $\sqrt{K}$.

On voit que les f_i sont ici les N valeurs d'une même variable aléatoire qui suit, autour de la moyenne de BERNOULLI, une loi de GAUSS

(1) L. von BORTKIEVICS, *Homogeneität und Stabilität in der Statistik, Skandinavisk Aktuaritidskrift*, 1918.

surdispersée. On peut obtenir immédiatement un rapport Q de valeur quelconque, mais il faut que les f_i suivent une loi de GAUSS.

Il est clair que c'est, pour les statistiques de mortalité, une mise en œuvre de l'idée de contagion.

Il est donc naturel d'envisager plusieurs urnes affectées chacune d'un cœfficient K. On obtient alors un schéma d'urnes très général, pouvant servir à la représentation des observations.

Cas des variables aléatoires dépendantes. — L'hypothèse simple des variables indépendantes a de grands avantages pour le calcul, elle en a parfois moins pour la vraisemblance.

Dans un cas simple et général, en particulier, elle est certainement inexacte. C'est quand les boules ne sont pas remises dans l'urne. Ce cas est classique depuis longtemps. On sait qu'une poignée de N boules, puisée dans l'urne à A boules, et de composition p, q fournit pour la variable aléatoire fréquence, la même espérance mathématique, mais l'écart type plus petit :

$$\sqrt{\frac{pq}{N}}\sqrt{\frac{A-N}{A-1}}.$$

Si A et N sont grands, les fluctuations suivent la loi de GAUSS.

On peut faire d'autres conventions. Tchouprov (1) avait examiné le cas où la boule est remise dans l'urne en lui joignant une boule de même couleur. Le problème plus général où l'on remet, avec la boule tirée, un nombre quelconque de boules de même couleur, a été posé et traité par POLYA (2). Pour des raisons évidentes, POLYA dit de ce schéma que c'est celui de la contagion.

Si le nombre de boules remises en sus de la boule tirée est de la forme $A\gamma$, on trouve pour la fréquence f des succès :

$$E(f) = p, \qquad \sigma_f^2 = \frac{pq}{N}\frac{1+N\gamma}{1+\gamma}.$$

En particulier, si N est grand, mais que $N\gamma$ soit fixé, soit C, l'écart type est l'écart de BERNOULLI, multiplié par $\sqrt{1+C}$.

La loi des fluctuations est voisine de celle de GAUSS, mais à dispersion hypernormale, comme dans le schéma de BOREL.

(1) TCHOUPROV, Grundbegriffe und Grundprobleme der Korrelationstheorie. Teubner, 1925.

(2) POLYA et EGGENBERGER, *Zeitschrift für Angewandte Mathematik und Mechanik*, III, 1923, pp. 279-289.

Si l'on envisage maintenant le cas de N infiniment grand, γ infiniment petit, avec $N\gamma = C$, et que d'autre part p soit infiniment petit, avec Np fini :

$$Np = \gamma$$

on trouve une loi limite qui s'appellera la loi des petits nombres avec contagion :

$$p_r = (1 + c)^{-\frac{\lambda}{c} - r} \frac{c(\lambda + c) \cdots (\lambda + (r - 1)c)}{r!}.$$

Etude de la mortalité par diphtérie à Paris, avant et après l'application du sérum. — La moyenne mensuelle de la mortalité à Paris avant le sérum est de 127. L'écart type $\sigma = \sqrt{1500}$ indique une dispersion hypernormale.

On peut obtenir une représentation assez satisfaisante par une loi de GAUSS, avec cœfficient de contagion.

$$C = 10{,}8$$

La courbe intégrale (nombre de mois où les décès sont inférieurs à une valeur donnée) indique les nombres suivants :

Décès	75	100	125	150	175	200	225
Nombre de mois	9	25	48	70	85	94	96

Les observations portent sur 8 ans (1886-1893).

Période après le sérum. — La moyenne mensuelle tombe à 36 et la courbe de fréquence change de caractère. L'ajustement par la loi des petits nombres avec contagion se fait assez convenablement avec le cœfficient :

$$C = 10{,}5.$$

La courbe intégrale donne les nombres suivants :

Décès	10	20	30	40	50	60	70	80	90	100	110
Nombre de mois	4	25	50	74	85	95	102	105	105	106	108

Les observations portent sur 9 ans (1895-1903).

L'addition des variables aléatoires et la théorie des erreurs expérimentales. — Considérons un type déterminé de mesure physique, par exemple la mesure de la grandeur d'une étoile. Cette grandeur, nous

pouvons la supposer bien déterminée, mais nos mesures n'atteignent que l'image du moment, peu déformée si nous prenons une bonne méthode de mesure, mais toujours déformée.

Ce que nous observons dans une série de mesures, c'est donc la population des images, dont la structure dépend évidemment de la méthode de mesure, et de l'observateur.

Les régularités observées dans la répartition des erreurs expérimentales peuvent-elles s'expliquer par des considérations raisonnables sur la structure de cette population.

Si nous pensons à notre exemple, nous voyons que les différences individuelles des images à la grandeur vraie résulte d'un ensemble très complexe d'actions très petites. L'effet de chacune de ces causes peut être considéré comme une variable aléatoire, et c'est la somme de ces variables qui constitue la différence individuelle :

$$x = \varepsilon_1 + \varepsilon_2 \cdots + \varepsilon_n.$$

Il est remarquable qu'une conception aussi générale puisse mener à des conclusions précises, si n est grand. Nous supposerons d'abord que les variables sont indépendantes, et qu'aucune ne soit prépondérante, au sens que les écarts types sont du même ordre de grandeur. Ils sont alors tous petits par rapport à l'écart type de x et l'on démontre que la variable x suit une loi voisine de la loi de GAUSS.

Il est clair que la théorie des erreurs n'est qu'un cas particulier d'application de cette idée et que partout où l'analyse du caractère à mesurer rend plausible l'hypothèse que les différences individuelles soient la somme d'un grand nombre d'effets aléatoires indépendants, on a des chances de pouvoir décrire et expliquer dans une certaine mesure les répartitions observées.

Lois qui se rattachent à la loi de Gauss. — On peut aussi imaginer que les causes aléatoires restant indépendantes, les effets produits par ces causes sur le développement de la différence individuelle x ne sont pas indépendant de la valeur de x.

L'accroissement Δx_i résultant de l'action d'une cause ε_i serait de la forme : (1)

$$\Delta x_i = \varepsilon_i F(x_i)$$

(1) WICKSELL, *On the genetic theory of frequency. Arkiv for Math. Astr. o. Fys.* Bd. 12, nº 20. KAPTEYN, *Skew frequency curves in Biology and Statistics*, Groningen, 1903.

Il est clair qu'en considérant :

$$\Phi(x) = \int \frac{dx}{F(x)}$$

on aura :

$$\Delta\Phi(x_i) = \varepsilon_i$$

$$\Phi(x) - \Phi(x_1) = \varepsilon_1 + \varepsilon_2 \cdots + \varepsilon_n$$

On voit que la propriété de suivre une loi de GAUSS se trouve transféré au premier membre, qui peut être une fonction quelconque de x.

Pratiquement, ces considérations ont un intérêt très grand dans le cas où :

$$F(x) = k(x - s).$$

$\Phi(x)$ est alors un logarithme.

On trouvera dans un travail de R. GIBRAT [1] paru depuis peu, de nombreux exemples de répartitions économiques, répartition des revenus, concentration des entreprises, qui suivent de façon frappante ces lois de GAUSS transformées.

Les courbes de Karl Pearson. — Comme nous l'avons remarqué à plusieurs reprises, le problème de la description des séries statistiques soit par un schéma d'urnes, soit par une courbe de fréquence de forme analytique simple, présente en soi l'intérêt d'une description commode, qu'il suggère ou non une explication.

KARL PEARSON s'est trouvé devant le problème pratique de représenter des observations très nombreuses. Il constitua à cette fin un ensemble de courbes de fréquence qui ont résolu le problème. Il est remarquable que le schéma d'urnes de POLYA comprenne deux courbes de K. PEARSON comme cas limite [2].

Dans les cas où la représentation par un de ces types est possible, on a donc la possibilité d'un schéma explicatif.

(1) R. GIBRAT, Une loi des répartitions économiques. L'effet proportionnel. *Bulletin de la Statistique générale de la France*, t. XIX, fasc. IV, 1930. *Les inégalités économiques*, 1931, Librairie du recueil Sirey.

(2) POLYA, *Sur quelques points de la théorie des probabilités*, p. 34 et 39. *Annales de l'Institut Henri Poincaré.*

***

Corrélations et liaisons stochastiques à plusieurs variables. — L'étude des corrélations existant en fait dans une population entre deux ou plusieurs caractères est un de ces phénomènes qui s'imposent fréquemment par leur importance pratique avant que les idées théoriques correspondantes aient pris une forme bien précise.

Une corrélation se manifeste par ce fait que la répartition liée d'un caractère y, (ou répartition de y quand le caractère x a pris une valeur fixée x_i) ; varie avec x_i, et n'est pas la même que la répartition générale de y. Ainsi pour les Céphéides, la grandeur des diverses Céphéides d'un amas stellaire peut varier de 5 grandeurs ou magnitudes si la période n'est pas fixée, elle ne varie pas de 2 magnitudes pour une période connue. C'est ce qu'on peut exprimer en disant que la dispersion liée est de l'ordre de $1/3$ de magnitude. Ce fait que la dispersion liée est petite donne aux corrélations leur intérêt pratique, et permet d'estimer une grandeur connaissant l'autre, avec une marge d'erreur sans doute, mais de façon tout à fait analogue à ce que donnerait une liaison fonctionnelle entre grandeurs.

Dans le problème des rendements en agriculture [1], l'étude de l'influence des conditions météorologiques, qualitativement évidente et d'une importance énorme, ne peut donner des résultats pratiques, en l'absence de travaux précis de laboratoire, que par l'emploi de la méthode statistique et des corrélations.

C'est pour des raisons analogues que FRANCIS GALTON, désireux d'obtenir des résultats numériques relatifs à l'hérédité, se proposa de voir si, en fait, certaines permanences statistiques pouvaient être mises en évidence sur des générations successives. En vérité au moment où GALTON commençait ses travaux (Hereditary Genius 1869), MENDEL avait publié (1865) des résultats qui, suivant ceux de NAUDIN, pouvaient constituer une base solide pour l'intelligence des phénomènes élémentaires de l'hérédité, mais ces recherches étaient restées tout à fait inaperçues. GALTON partait donc sur un problème difficile, avec des idées très générales, et voulait savoir si d'un caractère connu chez les parents, on pouvait tirer quelque lumière sur le même caractère chez un enfant.

(1) R. A. FISCHER, *Phil. Tr.*, B 213, pp. 89-142, 1925.

Bien évidemment, là comme ailleurs la méthode statistique n'est pas nécessairement la meilleure, il faut lui préférer quand on le peut, l'étude au laboratoire des phénomènes élémentaires, mais elle a l'avantage de fournir une mise en ordre, et des résultats numériquement utilisables, à des moments où les autres méthodes sont totalement désarmées.

Liaison stochastique. Notions descriptives fondamentales. — Du point de vue où nous nous plaçons, le correspondant abstrait de la corrélation est la notion de loi de probabilité liée. Ainsi, pour toute valeur fixée de x la loi de probabilité de la variable aléatoire y dépend fonctionnellement de x.

Comme cas particulier, la loi de probabilité de y peut rester la même quel que soit x, c'est le cas de l'indépendance stochastique.

Comme cas limite, l'ensemble des valeurs de y qui correspond à une valeur de x peut se réduire à une valeur unique, y n'est plus aléatoire quand x est connu. C'est le lien fonctionnel entre les variables aléatoires y et x.

Les éléments les plus importants de la loi liée sont l'espérance mathématique et la fluctuation (ou carré de l'écart type) de la variable liée y.

Ces éléments correspondent bien au problème physique de l'estimation de y quand x est connu. L'estimation est l'espérance mathématique et l'écart lié indique l'ordre de grandeur de l'erreur à craindre. Si, pour fixer les idées, nous pensons à deux variables aléatoires continues le point représentatif de la valeur probable décrit une courbe, dite de régression de y en x. Si l'on porte de part et d'autre l'écart type lié, on dessine une bande dont l'épaisseur indique l'erreur à craindre.

Quelques schémas de liaisons stochastiques. — Le plus simple et le plus immédiat est l'urne à trois espèces de boules A, B, C. On fixera l'attention sur les boules A et B. Le résultat d'une épreuve de N tirages fournit deux fréquences aléatoires. Elles seraient rigidement liées s'il n'y avait pas de boules de l'espèce C. Dans le cas général, elles sont en liaison stochastique. Il est évident en effet que si l'on ne sait rien sur le nombre de boules A extraites, le nombre des boules B peut aller de o à N. Mais si l'on connait le nombre N — m des boules A, le nombre de boules B est au plus de m. La loi de probabilité a changé.

L'étude analytique conduit à des résultats analogues à ceux du théorème de BERNOULLI.

Si p_1, p_2, p_3 sont les proportions des boules A, B, C, on a pour les fréquences f_1, f_2 :

$$\mathrm{E}(f_1) = p_1 \qquad \sigma_{f_1} = \sqrt{\frac{p_1 q_1}{\mathrm{N}}},$$

$$\mathrm{E}(f_2) = p_2 \qquad \sigma_{f_2} = \sqrt{\frac{p_2 q_2}{\mathrm{N}}}.$$

Les écarts réduits :

$$x = \frac{f_1 - p_1}{\sigma_{f_1}}, \qquad y = \frac{f_2 - p_2}{\sigma_{f_2}},$$

suivant, quand N est grand, une loi de probabilité à surface en cloche, ou loi de GAUSS à 2 variables, avec la densité :

$$\frac{1}{2\pi\sqrt{1 - r^2}} e^{-\frac{1}{2(1-r^2)}(x^2 - 2rxy + y^2)}.$$

Le paramètre unique r, appelé cœfficient de corrélation, et qui n'est autre que l'espérance mathématique du produit xy, prend ici la valeur :

$$r = -\sqrt{\frac{p_1 p_2}{(p_1 + p_3)(p_2 + p_3)}}.$$

Les courbes de régression sont des droites et l'écart type lié est constant, égal à $\sqrt{1 - r^2}$, inférieur à l'écart type (unité) de la variable non liée.

L'addition des variables aléatoires. — Pensons à un seul tirage dans l'urne ; les deux variables aléatoires représentant le succès pour la boule A, et le succès pour la boule B, sont en liaison stochastique, puisque dès que l'une est égale à 1, l'autre est nécessairement égale à zéro.

Les nombres de succès sont des sommes de la forme :

$$\mathrm{A} = \alpha_1 + \alpha_2 \cdots + \alpha_n$$
$$\mathrm{B} = \beta_1 + \beta_2 \cdots + \beta_n.$$

où chaque couple $\alpha_i \beta_i$ suit la même loi de probabilité, celle du tirage unique. D'ailleurs tous les couples sont indépendants.

Ce problème est l'analogue de celui que nous avons posé en général. On voit qu'il conduit à considérer un nombre quelconque de sommes

de la forme $\Sigma\alpha$, $\Sigma\beta$, $\Sigma\gamma, \cdots$ et à chercher la forme de la liaison stochastique entre ces sommes.

En particulier, l'urne à 4, 5 espèces de boules conduit à généraliser les résultats de BERNOULLI, par l'introduction de lois de GAUSS à un nombre quelconque de variables.

Il est facile de montrer que ces résultats sont plus généraux et que les couples $\alpha_i\beta_i$, supposés indépendants les uns des autres, peuvent avoir des lois de probabilité toutes différentes entre elles ; si cette famille de lois de probabilité satisfait à quelques restrictions d'ordre général, les sommes suivent à peu près la loi de GAUSS à deux variables.

Un cas extrêmement particulier de ce théorème est celui où α et β sont fonctions l'un de l'autre, et plus particulièrement encore, proportionnels. Autrement dit, deux combinaisons linéaires à cœfficients constants de n variables aléatoires sont en liaison stochastique et si n est grand, cette liaison est voisine d'une loi de GAUSS.

Ce dernier schéma fournit un mécanisme assez général de liaisons stochastiques à partir de variables aléatoires (indépendantes). Dans l'hypothèse où différentes grandeurs seraient fonction d'un grand nombre de variables aléatoires, les différences individuelles qui proviennent des fluctuations de ces variables déterminantes satisferaient assez bien à ce schéma spécial.

Ces considérations évidemment applicables à la théorie des erreurs expérimentales, fournissent une explication du rôle dans ce domaine des lois de GAUSS à plusieurs variables.

Dans la conception des phénomènes d'évolution qui découlait des vues de DARWIN, elles devaient aussi s'appliquer, et rendre naturelle la rencontre des lois voisines de celles de GAUSS.

Quelques résultats de Galton. — GALTON et l'école biométrique ayant étudié les corrélations entre générations successives étaient parvenus aux résultats suivants :

« La corrélation de père à fils introduit des régressions linéaires et s'exprime convenablement par une loi de GAUSS à deux variables. La corrélation entre un individu et ses ancêtres introduit des régressions linéaires, et des cœfficients de corrélation diminuant en progression géométrique.

« Elle peut être représentée convenablement par une loi de GAUSS à plusieurs variables.

Cet ensemble de résultats peut être appelé la loi empirique de l'hérédité ancestrale.

On disait quelquefois, sous une forme moins précise, l'influence des ancêtres existe et décroit en progression géométrique.

GALTON avait trouvé pour les cœfficients de corrélation :

$$\frac{1}{3}, \frac{1}{3} \times \frac{1}{2}, \cdots \frac{1}{3}\left(\frac{1}{2}\right)^2, \cdots \frac{1}{3}\left(\frac{1}{2}\right)^n.$$

Les mesures reprises et étendues confirmèrent formellement ce résultat, mais on trouvait plutôt des puissances successives d'un rapport voisin de $\frac{1}{2}$, variable d'ailleurs avec l'espèce et le caractère.

D'autre part, les résultats de MENDEL, redécouverts à partir de 1899, produisaient bientôt une école dont les conclusions parurent s'opposer à celles des biométriciens. Il semblait que la connaissance des seuls parents apportât toute l'information désirable, les ancêtres n'ayant plus d'influence.

En fait, comme l'a montré K. PEARSON dans une série de mémoires, il n'y a là qu'un malentendu et les lois de MENDEL, loin de s'opposer aux lois empiriques de l'hérédité, en fournissent une explication presque complète.

Le but de l'école biométrique est en effet d'obtenir des répartitions liées, autrement dit, connaissant la valeur d'un caractère pour le père et pour le grand-père, d'en déduire la loi de répartition du même caractère pour les fils ayant cette ascendance fixée.

On peut aussi, ne connaissant que le grand-père par exemple, chercher à obtenir le même renseignement ; on est amené aux questions suivantes, très différentes l'une de l'autre,

1° Le père étant connu, la connaissance du grand-père ajoute-t-elle quelque chose ?

2° Le père étant inconnu, la connaissance du grand-père est-elle utile ?

La réponse à 2) résulte des expériences. Il existe une corrélation entre les fils et un ancêtre d'ordre quelconque, le cœfficient de corrélation décroissant d'ailleurs en progression géométrique.

Pour la question 1) K. PEARSON (1) montra qu'en général la con-

(1) K. PEARSON, *Phil. Tr.*, vol. 187, 1896, p. 304.

naissance des ancêtres ajoute quelque chose, sauf si la progression géométrique est de la forme :

$$a) \qquad \rho, \rho^2 \cdots \rho^n$$

mais non de la forme :

$$b) \qquad K\rho, K\rho^2 \cdots K\rho^n.$$

Dans le premier cas, la connaissance du père suffit, la dépendance n'a lieu que par l'intermédiaire du prédécesseur immédiat.

Or, la théorie de MENDEL conduit bien (1) aux régressions linéaires, et aux cœfficients de corrélation décroissant en progression géométrique. De plus (2) (p. 227) si on étudie la corrélation entre les constitutions des gamètes ou cellules sexuelles, les cœfficients de corrélation sont les puissances successives de $1/2$, ce qui est de la forme (a). Il en résulte que la connaissance de la constitution gamétique des ancêtres n'ajoute rien à celle de la connaissance gamétique des parents.

Au contraire, et comme conséquence également des lois de MENDEL, si l'on doit juger par les caractères somatiques, on retrouve bien une progression géométrique mais de la forme b). La connaissance des ancêtres ajouterait quelque chose. Il reste toutefois à établir (3) un accord numérique complet entre la théorie et l'expérience, mais on peut dire avec K. PEARSON qu'il reste peu d'écart entre les conqéquences théoriques des lois de MENDEL et les résultats des biométriciens.

La corrélation entre les âges des époux. — La description d'après l'âge des couples mariés dans l'année ou de l'ensemble des couples d'une population déterminée (par exemple ceux qui résultent d'un recensement) met en évidence le fait naturel que ces âges ne sont pas indépendants.

Une fois fixé, par exemple, l'âge de la femme, la répartition d'âge des maris est tout à fait différente de la répartition générale, l'âge moyen étant peu éloigné de l'âge fixé pour la femme et la dispersion étant fortement diminuée.

(1) K. PEARSON, *Phil. Tr.*, vol, 203. — A. 1904, pp. 53-86.

(2) R. S. *Proc. B.*, 81, 1909, pp. 219-224 et 225-229.

(3) E. C. SNOW, *Roy. Soc. Proc.*, B. 83, 1911, pp. 37-55.
E. B. WILSON, *Proc. Nat. Ac. of Sc.*, 14, 1928, pp. 137-140.

Cette corrélation peut naturellement présenter des caractéristiques différentes suivant qu'on prend seulement les mariages de l'année, l'ensemble des couples d'une nation, ou l'ensemble des couples ayant des enfants...

Il résulte de cette corrélation qu'on ne saurait, dans une étude quelconque où l'on recherche l'influence des âges des parents, considérer ces âges comme des variables indépendantes.

Le taux de masculinité. — Nous avons déjà parlé de la permanence du rapport, fréquence des garçons dans les naissances (ce rapport étant voisin de 0,515) ou ce qui revient au même, du rapport des naissances masculines aux naissances féminines (voisin de 1,06).

La théorie des chromosomes indique une interprétation possible (1) où les cellules sexuelles mâles, par exemple, seraient de deux types différents, les deux sexes étant obtenus par fécondation de cellules femelles d'un seul type.

Dans une telle interprétation, l'idée la plus simple est celle de l'égalité des deux sexes à la conception, ce qui est tout à fait faux. En attendant des progrès nouveaux dans l'intelligence de ces phénomènes élémentaires, on ne peut guère qu'employer la méthode statistique, pour l'étude des problèmes qui paraissent intéressants. Si l'interprétation des résultats statistiques est souvent difficile, on obtient du moins une mise en ordre qui permet de porter un jugement sur des tentatives plus ou moins complètes d'explication.

Par exemple, on peut se demander si le taux de masculinité conserve sa valeur dans des catégories déterminées, s'il est le même quand les parents sont de races différentes, s'il est influencé par l'âge des parents... C'est à ce dernier point que nous nous attacherons, d'après un mémoire de S. D. WICKSELL (2).

Si l'on considère, dans la population un groupe de couples où les âges de la mère et du père soient voisins de x, y, la densité des naissances masculines provenant des couples d'âges xy est de la forme :

$$n_g \quad \mathrm{G}(xy),$$

ou n_g est le nombre total des naissances masculines. De même la densité des naissances féminines est :

$$n_f \quad \mathrm{F}(xy).$$

(1) GUYÉNOT, *L'hérédité*, p. 358.
(2) *Sex proportion and parental age*. Festkrift C. V. L. Charlier, 1927.

Le rapport P de ces densités, est la grandeur qui nous intéresse. Il est clair que si la répartition des garçons et des filles suivant l'âge des parents était la même ce rapport aurait la valeur constante $\frac{n_g}{n_f}$.

En fait, il varie avec le point xy, en restant bien entendu assez voisin de la valeur globale $\frac{n_g}{n_f} = P_0$. On a donc :

$$P = P_0(1 + f(xy)).$$

Les observations déterminent les valeurs, pour les différents groupes d'âge, de la fonction $f(xy)$ relative à la population étudiée.

Bien entendu, cette fonction $f(xy)$ ne traduit pas l'idée que le rapport P serait rigidement déterminé dans chaque classe, par les âges des parents, mais simplement qu'il n'est pas possible d'attribuer au hasard les différences observées entre P et P_0. Une description complète de ces différences introduit la fonction $f(xy)$.

On pourrait aussi se borner à faire des catégories suivant l'âge de la mère seule, suivant l'âge du père seul, en suivant la différence de leurs âges... Le rapport P change avec la caractéristique de ces classes et l'on peut décrire cette variation par des fonctions de ces caractéristiques.

Donnons quelques résultats sur les naissances à Berlin de 1878 à 1922 :

Pour la période 1878-1905, la fonction $f(xy)$ indique un accroissement de P quand x augmente seul, une diminution de P quand y augmente seul.

De 1906 à 1914, on peut considérer qu'il n'y a pas de variation.

De 1920 à 1922, on constate toujours un accroissement quand l'âge de la mère varie seul, mais plus de variation quand l'âge du père varie seul.

Si l'on divisait le matériel en ne prenant en considération que l'âge du père, on trouverait cependant un accroissement très sensible, mais il tiendrait à l'étroite corrélation qui existe entre l'âge du père et de la mère.

On voit déjà que les résultats changent d'allure pour une même région avec la période considérée, peut-être avec les conditions économiques. On constate également des différences notables, comportant même des changements de sens, quand on examine différents pays.

Il est en tout cas hors de doute qu'on ne peut considérer qu'on puise dans une urne restant la même pour les différents groupes d'âges.

Comme autre résultat net, il est très fréquent de voir P augmenter quand augmente l'âge de la mère, l'âge du père restant le même.

Est-il possible de se représenter ces faits comme conséquences de certaines hypothèses ? Une théorie qui remonte à CHRISTIAN BERNOULLI (1838), prolongée par LEXIS et TCHOUPROV, considère que le sexe est fixé à la conception. A partir de la conception, de taux fixé, la mortalité avant la naissance affecte différemment les deux sexes, le taux à la naissance résultant de ces deux causes.

Si γ_g est le nombre des conceptions masculines, γ_f celui des conceptions féminines, le rapport à la conception est :

$$p_1 = \frac{\gamma_g}{\gamma_f},$$

WICKSELL introduit en outre le rapport du total des avortements au nombre total des naissances :

$$\beta = \frac{A_g + A_f}{ng + nf},$$

dont on possède une estimation (ordre de 0.20 à 0.25). A_g, A_f sont les nombres d'avortements, de rapport p_2. On trouve aisément que la valeur p_3 de $\frac{n_g}{n_f}$:

$$p_3 = \frac{p_1(1 + p_2) - \beta(p_2 - p_1)}{1 + p_2 + \beta(p_2 - p_1)},$$

p_1 et p_2 étant fixés, p_3 dépend de β. On peut essayer d'admettre que p_1, valeur du rapport à la conception ne dépend pas de l'âge, que p_2 n'en dépend pas non plus. Quant au taux β, il paraît raisonnable d'admettre d'abord qu'il peut varier avec l'âge de la mère. Il est clair qu'alors p_3 ne dépendrait que de x mais comme β est fonction croissante de x, p_3 serait décroissant. Ce n'est pas du tout ce qu'on a trouvé, sauf dans les statistiques des Pays-Bas.

Pour la théorie considérée sous cette forme, on peut dire avec WICKSELL que les résultats de l'expérience lui sont opposés, et que le taux doit varier avec les deux âges (d'une façon différente pour les différents pays) ou que l'hypothèse de p_1 constant doit être modifiée.

Avec des notations un peu différentes, on peut considérer les éléments suivants ; les proportions γ_g et γ_f des conceptions masculines

et féminines. Les populations sont de la forme $N\gamma_g$, $N\gamma_f$ sur lesquelles agissent des mortalités de taux $\delta_g\delta_f$ laissant subsister :

$$N\gamma_g(1-\delta_g) \qquad N\gamma_f(1-\delta_f).$$

On aurait donc :

$$p_3 = p_1\left(\frac{1-\delta_g}{1-\delta_f}\right).$$

Dans cette hypothèse générale, tout lien du rapport p_1 avec un élément quelconque doit être cherché soit dans p_3, soit dans la mortalité. L'observation conduit à poser :

$$\delta_g = h\delta_f \qquad h > 1.$$

Nous venons de voir qu'on ne peut se contenter de supposer p_1 constant, δ_f variable avec x seul.

Il ne serait pas déraisonnable de supposer que p_1 est variable avec x et y, δ_f étant variable avec x seul.

On a de cette manière substitué à un schéma simpliste un autre schéma plus souple, peu satisfaisant d'ailleurs et n'allant pas à une explication, puisqu'il devrait avoir des structures différentes pour différentes régions, sans qu'on voie bien nettement les raisons de ces différences.

***

Nous porterons maintenant notre attention sur la structure un peu particulière des problèmes où intervient la variable temps. Il s'agit toujours de rechercher et vérifier des permanences, puis de chercher à les expliquer. C'est ce qu'a fait l'astronomie par exemple, avec cette chance particulière que le mouvement de la terre et des planètes autour du soleil fait apparaître des régularités simples et assez pures, dues à la masse énorme du soleil. Le champ dans lequel se déroulent les phénomènes du système solaire est d'une structure très simple en première approximation, et susceptible d'une étude complète par décomposition du problème et approximation successives.

Mais les problèmes que les sciences sociales et économiques présentent, et que les méthodes statistiques peuvent chercher à préciser et à résoudre, sont très différents.

On voudra suivre la variation des prix du blé, du charbon, du coton, de la fonte... ; la variation de la production des métaux précieux, des

céréales... ; le mouvement de la population, des mariages, des naissances... ; l'évolution de la structure agricole, industrielle, financière... d'un pays ; du mode de vie de certaines classes, du mouvement des salaires, de la consommation de divers produits...

Dans cet immense ensemble, où tout évolue et gravite, il faut mettre de l'ordre, discerner des régularités, voir si elles ont des traits communs. Pour une série statistique unique, on devra la décrire et tenter de l'expliquer. Pour deux ou plusieurs séries, il faut voir de plus si elles ont des liens entre elles.

Prenons d'abord un exemple très schématisé : Un point matériel dont la position d'équilibre serait o, se déplace sur une droite, une force proportionnelle à la distance agissant quand il s'écarte de o. Si le point o est immobile, si la constante de la force élastique n'évolue pas avec le temps, l'observation à intervalles (qu'on peut supposer réguliers) de la position du point fournira une série dessinant une sinusoide :

$$x = x_0 + \frac{v_0}{\omega} \sin(\omega t).$$

Si le point o se déplace sur la droite suivant une loi déterminée, on aura une sinusoide festonnant autour de la courbe représentant $x_0(t)$. On peut concevoir que la force élastique évolue elle-même lentement (ce serait le cas d'un pendule oscillant dans un champ lentement variable, la longueur pouvant elle même varier lentement). L'amplitude et la période évoluent alors.

On obtient aussi ce dernier résultat en supposant qu'à la force élastique s'ajoutent des forces aléatoires, de petits chocs (1) v_1 et ω se modifient alors, mais avec un caractère aléatoire et non fonctionnel.

Il est bien clair que, quel que soit le mécanisme envisagé, à chaque valeur de t correspond une valeur de x, et qu'on aura une courbe festonnée, plus ou moins régulière, autour d'une courbe pouvant avoir une forme quelconque.

Si d'une façon générale, on envisage les petits mouvements d'un système au voisinage d'une position d'équilibre stable, on aura une ample généralisation de ce schéma, avec un nombre quelconque de paramètres.

(1) G. UDNY VULE, *Phil. Trans.*, A. 226, pp. 267-298.

Si l'on en prend deux, x, y, on aura en particulier :

$$x = x_0 + a \sin \omega_1(t - \alpha) + b \sin \omega_2(t - \beta)$$
$$y = y_0 + a' \sin \omega_1(t - \alpha) + b' \sin \omega_2(t - \beta).$$

Chaque paramètre pris à part dessinera une courbe festonnée autour d'une autre courbe dont la forme peut être quelconque, l'ensemble $x_0(t)$, $y_0(t)$ figurant l'évolution de la position d'équilibre.

Il est bien clair qu'amplitudes et périodes dépendent de l'énergie du système, donc de sa structure, et peuvent évoluer avec lui.

Variables aléatoires. — L'intervention la plus simple des variables aléatoires est obtenue quand on exécute dans une urne de composition variable avec le temps, des tirages dont le nombre N peut lui-même varier avec t.

On aura pour la fréquence relative f et pour la fréquence absolue n :

$$f = p(t) + \varepsilon$$
$$n = Np + \eta,$$

ε et η sont deux variables aléatoires, dont la loi, voisine d'une loi de GAUSS, évolue avec le temps.

Ce deuxième type n'est pas très éloigné de celui où l'on totalise la production d'une période déterminée, l'importance des moyens mis en œuvre dépendant du temps, mais où le résultat peut-être favorisé ou contrarié par des circonstances aléatoires.

Supposons en effet que cette production soit fixée rigidement par la connaissance de deux grandeurs dont l'une fixerait le début du développement et l'autre la structure d'un milieu nutritif :

$$x = f(a, b).$$

On est maître de la grandeur a, qu'on fixera chaque fois, et qui dépendra donc du temps par la politique adoptée, mais on n'est maître que dans une certaine mesure de la grandeur b qui peut subir des fluctuations aléatoires. Il est clair qu'en première approximation, on aura :

$$x = f(a(t), b_0(t)) + \varepsilon,$$

$b_0(t)$ étant la valeur moyenne de b, ε une variable aléatoire, de structure généralement complexe, fonction du temps.

Il est clair que ce dernier schéma peut s'étendre à une grandeur dépendant d'un nombre quelconque de paramètres déterminants, pour

un groupe desquels les valeurs sont fixées par une certaine politique, les valeurs des autres n'étant fixées qu'en moyenne. On aura toujours :

$$x = \Phi(t) + \varepsilon,$$

où $\Phi(t)$ représente ce qu'on aurait dû avoir si les circonstances aléatoires avaient eu exactement leurs valeurs moyennes.

Le développement continu (1). — Donnons encore un exemple simple qui conduit à une idée un peu différente. On observe une grandeur dont le développement est contrôlé par une équation différentielle du type :

$$\frac{dx}{dt} = f(x, a).$$

On connaîtrait son développement si a était connu à chaque instant. Mais supposons que a subisse des fluctuations aléatoires et prenons le cas très simple :

$$\frac{dx}{dt} = a.$$

a au lieu d'avoir une valeur constante a_0, possède au début la valeur $a_0 + \varepsilon_0$, puis prend au bout du temps Δt la valeur $a_0 + \varepsilon_0 + \varepsilon_1$ et ainsi de suite jusqu'au temps t. Il est clair que :

$$x = a_0 t + \Delta t[n\varepsilon_0 + (n - 1)\varepsilon_1 \cdots + \varepsilon_n].$$

Nous aurons encore une série fluctuante autour de la droite $x = a_0 t$, mais la variable aléatoire ajoutée est d'un type spécial. Sa valeur probable est bien nulle, mais son écart type dépend du temps t. Si les ε_i sont indépendants, on voit que cet écart est de l'ordre de $t\sqrt{t}$.

Il est clair d'ailleurs qu'on obtiendrait des résultats analogues en totalisant les résultats de tirages faits à la cadence de un par intervalle Δt, mais l'on aurait ici :

$$x = pt + \varepsilon_1 + \varepsilon_2 \cdots + \varepsilon_n,$$

et cette fois la variable aléatoire aurait seulement des écarts de l'ordre de $\sqrt{t}$.

Les séries statistiques réelles. — On trouve très fréquemment des séries présentant le caractère dont nous venons de parler, de festons autour d'une autre courbe. On voit qu'on ne manque pas de schémas

(1) Voir H. HOTELLING, *Differential equations subject to error. J. of the Amer. Stat. Assoc.*, septembre 1927, pp. 283-314.

pour essayer de les représenter, et l'on pourrait en fabriquer d'autres (1) mais il faut encore qu'ils aient quelque rapport avec la question étudiée.

La variation dans le temps des taux de masculinité. — Nous empruntons à BOREL l'exemple de la variation des taux de masculinité sur une période de 32 ans (1866 à 1897) en Autriche.

Donnons ici seulement les écarts par rapport à la moyenne 0,51486.

1886	— 151	79	17	— 17	89	— 13	— 128	
	776	108	2	27	— 163	86	— 113	(1897)
	15	68	— 88	— 36	— 25	76		
	89	125	— 41	— 108	— 81	— 31		
	28	89	— 7	— 47	9	— 77		

Le graphique qu'on construira montre immédiatement que la fluctuation a lieu autour d'une courbe descendante, qu'il est raisonnable d'essayer de représenter par une droite qu'on ajustera par la méthode des moindres carrés. On aura donc pour le taux observé p' :

$$p' = p_0 - ht + \varepsilon.$$

Si cette vue est raisonnable, on doit trouver que les valeurs de ε sont les valeurs indépendantes d'une variable aléatoire suivant une loi de GAUSS relative à l'urne moyenne p_0, le nombre total des tirages étant le nombre total des naissances d'une année.

On constate en effet que l'écart quadratique moyen des résidus est bien l'écart type σ_B de BERNOULLI.

Un tel traitement de la question laisse intacte la question de savoir pourquoi le taux de masculinité évolue ainsi.

Les séries de la production agricole. — On trouve encore des graphiques de même allure. Comme nous l'avons vu, il n'est pas déraisonnable de supposer que la courbe autour de laquelle se produisent les fluctuations serait celle qui aurait été le résultat de l'expérience si la politique de la production restant ce qu'elle a été en réalité, les aléas des conditions météorologiques avaient été supprimés, laissant subsister certaines conditions moyennes.

(1) On peut citer les suggestions fournies par les oscillations de relaxation.
L. HAMBURGER, *Analogie des fluctuations économiques et des oscillations de relaxation*, Public. de l'Institut de Statistique, janvier 1931.

Etant donné notre ignorance encore bien grande des relations qui existent entre la production agricole et ses conditions déterminantes, il ne saurait être question de deviner une forme nécessaire de la courbe de production normale. Nous pouvons seulement, à partir des renseignements connus sur la politique de production et la physionomie météorologique de l'année, juger si une courbe ajustée aux données répond convenablement à nos exigences. Il est clair que si rien de particulier n'ayant été tenté, nous enregistrons deux récoltes exceptionnelles en deux années successives très favorables, nous devons laisser au-dessous la courbe de production normale.

Il arrive que cette courbe normale soit appelée courbe de tendance. Une telle locution est bien imprécise, et malgré le vague où l'on est souvent forcé de demeurer en ces matières, il serait bon de spécifier à quoi se rapporte cette tendance, et quel effort précis elle traduit.

Explication d'une série unique. — Imaginons que nous ayons des raisons de séparer ainsi en composante normale et composante aléatoire un phénomène déterminé. Il faut alors approfondir la signification de ces deux composantes. Pour la première, il conviendrait de dégager les paramètres déterminants, c'est-à-dire les grandeurs dont la connaissance suffit pratiquement à la prévision de la valeur normale. Ce problème est très difficile. Par exemple, pour les récoltes, en se bornant à l'influence de la pluie, qui est un élément très important, il faut faire intervenir la physionomie de l'année, c'est-à-dire au fond la courbe de répartition en fonction de la date de chute, de la quantité d'eau tombée (1). On a donc en fait, comme élément déterminant, une fonction du temps. Le nombre qu'on cherche est une fonctionnelle de la physionomie météorologique.

Si l'on n'étudie que l'influence de la pluie, on peut obtenir expérimentalement les variations de la récolte qui correspondent à des déformations en un point quelconque de l'année, de la fonction de répartition normale.

Il resterait à étudier les autres éléments déterminants, pour deviner l'allure de la courbe de tendance.

En fait, on est très loin de posséder les résultats de laboratoires, de

(1) R. A. FISCHER, *On the influence of rainfall on the yield of wheat at Rothamsted. Ph. Trans. R. Soc. of London*, B, 213, 1925.

champs d'expérience, ou même d'exploitations normales qui éclaireraient suffisamment cette première question.

On se contente généralement de tracer une courbe d'une forme analytique simple, et qui fournit une description en fonction du temps des résultats normaux du passé. Même si cette courbe n'est pas très éloignée de la signification qu'on lui attribue, elle ne peut servir qu'à exprimer des vues à très courte échéance sur l'avenir.

La composante aléatoire. — Elle représente, nous l'avons vu, les effets des causes aléatoires qui diffèrent de leur valeur normale. Dans notre ignorance du mécanisme de ces causes, il n'est pas commode de raisonner sur leurs effets. On peut bien essayer dans certains cas de croissance d'envisager comme élément fondamental l'accroissement relatif, ce qui conduit à multiplier la cause aléatoire par la grandeur de l'organisme, ou à considérer le logarithme de la grandeur sous la forme :

$$\log (X) = \log (X_0) + \eta.$$

Supposons en tout cas, qu'on ait isolé une telle composante aléatoire. On peut se demander quelle est sa loi de fréquence (à la supposer unique et indépendante du temps). Il faut alors déterminer cette loi par les observations.

Il reste encore à se demander si les valeurs successives de la variable aléatoire peuvent être considérées comme indépendantes ; dans le cas dont nous avons parlé on se demandera si les aléas sont liés d'une année à l'autre, ou s'ils sont indépendants.

Dans le schéma des tirages ajoutés, ce ne serait pas le cas, ni pour l'opération :

$$\frac{dx}{dt} = a.$$

avec le mécanisme que nous avons adopté. On trouverait pour le premier schéma :

$$\eta_{it} = \varepsilon_1 + \varepsilon_2 \cdots + \varepsilon_t$$
$$\eta_{it+1} = \varepsilon_1 + \varepsilon_2 \cdots + \varepsilon_t + \varepsilon_{t+1}.$$

Il y a liaison presque rigide lorsque t est grand.

Dans le cas général, on est conduit à chercher quelle loi de probabilité unit deux termes consécutifs, trois termes consécutifs... Dans les schémas précédents, tous les termes sont liés mais de façon décroissante avec leur distance.

Si ce travail peut être fait de façon satisfaisante, on peut dire qu'on a décrit :

La composante normale en fonction du temps ;

La loi de fréquence et les liaisons internes de la série aléatoire.

La comparaison de deux phénomènes. — Supposons que deux phénomènes soient capables d'une telle représentation, chacun possédant sa composante normale, à laquelle se superpose une composante aléatoire.

Etudier les relations des deux phénomènes, en logique, c'est comprendre leurs causes, dégager la manière dont ils en dépendent, et voir si ces groupes de causes ont une partie commune. Si cette partie commune est de grande importance, les deux phénomènes seront étroitement liés, mais pourront l'être de deux manières bien différentes.

Si le groupe commun est celui des causes a fixées par une politique, il y aura liaison des composantes normales par l'intermédiaire des a ; mais liaison assez difficile à mettre en évidence sans une esquisse de théorie.

Les composantes aléatoires seront alors sans lien.

En revanche, si le groupe commun est celui des b, il y a une liaison des composantes normales par l'intermédiaire des b_0, et une liaison des composantes aléatoires par les fluctuations autour des b_0.

Cette liaison des composantes aléatoires, c'est une loi de probabilité à deux variables à mettre en évidence et à décrire. C'est à cette tâche, la plus facile, qu'on s'est généralement borné, on a étudié la corrélation des composantes aléatoires.

Une telle méthode bien entendu, ne peut donner de résultats sérieux dans la comparaison des composantes normales, elle n'a même aucun sens précis. On ne possède en général que des descriptions empiriques, en chaque instant du passé, de ces deux composantes normales. Ces descriptions n'ont aucune valeur explicative, car ce n'est pas en tant que fonctions du temps que les composantes normales doivent être rapprochées, c'est comme fonctions d'autres éléments, qui ont été des fonctions du temps.

Il semble qu'il y ait danger à employer le mot de corrélation pour l'étude de ces relations entre composantes normales. Peut être pourrait-on comme je l'ai suggéré, lui réserver le mot de covariation (1) qui laisse subsister le véritable problème, comprendre pourquoi et de quelle manière les phénomènes varient ensemble.

(1) G. DARMOIS, *Metron*, vol. VIII, 1929, pp. 2-42.

La liaison des composantes aléatoires par la méthode des différences (1). — Il est naturel de penser que la courbe de l'évolution normale, résultat d'une politique continue, lentement variable, sera généralement une courbe assez tendue, convenablement représentée dans un intervalle par un polynôme de degré peu élevé. Dans ces conditions, on voit que les différences premières secondes... seront représentées par une composante normale qui sera constante, si l'on pousse assez loin la différenciation, et qui deviendra nulle ensuite.

On démontre alors que, si cette opération est faite pour deux séries, les séries aléatoires qui subsistent après élimination de la composante normale permettent de trouver les liaisons internes de chaque série et les liens qu'elles présentent entre elles. Cette opération n'exige donc pas qu'on recherche la composante normale.

Elle s'applique immédiatement à la variation des taux de masculinité examinés plus haut. On ne trouve alors pas de liaison entre les termes aléatoires.

Nous emprunterons un exemple de liaison interne aux résultats d'un travail d'EGON S. PEARSON (2). (On the variations of personal equation and the correlation of successive judgments).

La corrélation des jugements successifs. — Si on considère une série d'épreuves identiques (bisection, trisection, etc.), on constate que les erreurs commises manifestent le caractère fluctuant autour d'une courbe, généralement une droite inclinée.

Pour un observateur déterminé, il y a une erreur systématique qui évolue, et une composante aléatoire qu'on voudrait étudier, surtout au point de vue des liaisons internes de cette série aléatoire.

Si l'on admet un lien entre les erreurs successives mais que, pour simplifier, on suppose que l'erreur de rang i est liée seulement à l'erreur commise précédemment, on trouve que deux termes quelconques d'une série sont en corrélation, les cœfficients de corrélation décroissant en progression géométrique, du type :

$$\rho \quad \rho^2 \quad \rho^3 \cdots \rho^n,$$

déjà rencontré à propos de l'hérédité.

(1) Voir O. ANDERSON, *Die Korrelationsrechnung in der Konjonkturforschung*, Schroeder, Bonn, 1929.

(2) BIOMETRIKA, vol. XIX, pp. 23 à 102, 1922-1923.

Or, ce n'est pas ce que donne l'observation. Dans le cas le plus simple, on trouve des cœfficients de corrélation de la forme

$$k\rho \quad k\rho^2 \cdots k\rho^n \cdots,$$

dont une explication peut être fournie de la manière suivante. A l'erreur d'estimation suivant la loi précédente, peut s'ajouter une erreur accidentelle affectant cette estimation :

$$e = \alpha_t + \beta_t.$$

Les erreurs accidentelles β_t sont supposées indépendantes entre elles et indépendantes des erreurs α_t.

Dans ces conditions, on trouve que la décroissance des cœfficients de corrélation prend bien la forme $K\rho^n$ donnée par l'expérience.

Les phénomènes économiques. — Les séries statistiques du mouvement des prix, de la production présentent aussi ce caractère d'une courbe fluctuante. L'interprétation, si primitive qu'elle soit, donnée dans les cas précédents, ne peut convenir ici. Il s'agit de mouvements relatifs à cet ensemble complexe qu'on peut appeler le monde économique, son développement et son évolution ? Si l'on admettait que, à certains points de vue simplifiés, cet ensemble pût se représenter par un nombre fini n de paramètres, les courbes fluctuantes représentant leurs variations en fonction du temps seraient des graphiques résultant du mouvement du système autour d'un mouvement général d'évolution.

Ce mouvement général d'évolution serait analogue à la croissance d'un organisme. Quant aux fluctuations, elles seraient des sortes d'oscillations, de pulsations en relation avec la structure actuelle du système. Si vague et incomplète que soit une telle analyse, elle peut suggérer des points de vue analogues aux mouvements d'un système autour d'un équilibre stable et mobile, ou aux oscillations de relaxation mais elle ne permet pas de conserver la notion de tendance au sens ou nous avons tenté de la préciser. Il serait plus raisonnable de parler de mouvement moyen d'une grandeur, et d'oscillations autour du mouvement moyen.

Etant donnée une grandeur quelconque en relation avec le système qui évolue, son mouvement moyen est une composante du mouvement moyen du système, et ses oscillations sont un indice de la structure du système.

Deux ou plusieurs grandeurs sont donc, de ce point de vue deux

traits, deux aspects d'une même chose ; en particulier les oscillations peuvent dépendre des mêmes éléments de structure ou avoir de nombreux éléments communs, elles peuvent aussi être relatives à des éléments distincts et n'avoir rien de commun.

Mouvement des mariages et situation économique. — Il s'agit d'un problème déjà ancien qui a été étudié pour l'Angleterre par HOOKER en 1901.

La grandeur en relation directe avec la situation économique donnée par HOOKER était le commerce extérieur (valeur par tête d'habitant de la somme exportation plus importation).

Dans toutes ces questions, il est assez difficile de dégager le mouvement moyen.

HOOKER le définit en chaque point comme la moyenne d'un nombre impair d'années dont l'année considérée est l'année centrale. Le choix de ce nombre, en fait suggéré par l'importance des fluctuations, est un peu arbitraire (HOOKER prenait 9 ans) toutefois cette méthode est parfaitement raisonnable et plus appropriée à ces recherches que la méthode des différences.

Ayant ainsi obtenu le mouvement moyen des deux phénomènes, on a par différence le mouvement oscillatoire. On peut alors examiner si ces deux mouvements paraissent avoir une importante composante commune. En fait, on est réduit à regarder si les courbes se ressemblent. Dans le cas étudié par HOOKER, elles se ressemblaient beaucoup.

Il reste évidemment, après une telle constatation, à voir si la psychologie du peuple étudié rend compte des résultats obtenus (1).

Si les courbes se ressemblent peu, cela ne prouve nullement l'absence de relation, car si par exemple l'un des phénomènes avait en commun avec l'autre un certain nombre de termes périodiques, ce caractère analytique ne paraîtrait pas nécessairement et ne pourrait être aperçu que par une analyse plus soignée (2).

(1) Des études analogues ont été faites pour la France par Henri BUNLE, *Journal de la Société de Statistique de Paris*, 1911 ; la ressemblance des courbes est très faible.

(2) Il semble que les travaux actuels de Ragnar Frisch puissent fournir la base nécessaire à une étude plus approfondie.
Voir *A Method of decomposing an empirical, series, Amer. Stat. Journal.* March Supplement, 1931.

(Conférences faites à l'Institut HENRI-POINCARÉ en 1929)

Manuscrit reçu le 25 juillet 1931.

Imprimerie des PRESSES UNIVERSITAIRES DE FRANCE, Paris-Saint-Amand. 22-10-1932.