REVUE DE STATISTIQUE APPLIQUÉE

YVES MAINGUY

La statistique comme instrument de connaissance

Revue de statistique appliquée, tome 4, n° 3 (1956), p. 101-114

<http://www.numdam.org/item?id=RSA_1956__4_3_101_0>

LA STATISTIQUE COMME INSTRUMENT DE CONNAISSANCE

par

Yves MAINGUY

Chef du Service des Statistiques et des Études Économiques
à la Direction Commerciale du Gaz de France

Le Centre de Formation des Ingénieurs et Cadres aux Applications Industrielles de la Statistique a organisé cette année, pour la première fois, un cycle d'études sur les Applications de la Statistique à la Gestion Economique des Entreprises. Ce cycle s'adresse au personnel de direction des grandes entreprises. Il a pour objet principal de faire prendre conscience des services que peuvent rendre, pour la gestion des entreprises, les méthodes statistiques. Il fournit également un commencement d'information utilisable. Enfin il évoque certains problèmes qui, liés à l'information statistique, débordent celle-ci.

Les pages qui suivent reproduisent la leçon inaugurale de ce cycle.

1. Il semblera paradoxal à beaucoup que ce cycle d'études, consacré aux applications de la statistique à la Gestion Economique des Entreprises, soit ouvert par un homme qui n'est pas statisticien et qui n'a pas, et n'a jamais eu, de responsabilité de gestion dans une entreprise.

Attaché pendant près de dix ans à la recherche économique, auprès d'un maitre à qui je dois beaucoup, mon ami François Perroux, j'ai connu le besoin de la statistique et rencontré un autre maitre, le Professeur Darmois, dont pourtant je n'ai jamais eu l'honneur de suivre l'enseignement mais dont, par un de ces phénomènes d'osmose plus ou moins conscients que, par paresse d'esprit sans doute, je veux croire efficace dans la communication de la pensée, j'ai reçu et je reçois encore une sorte d'imprégnation intellectuelle que je souhaite féconde.

Il n'en demeure pas moins que je ne suis ni économiste ni statisticien, ayant glané dans l'une et l'autre de ces disciplines, m'étant risqué même trop souvent à en écrire et à en parler, mais n'ayant systématiquement étudié ni l'une ni l'autre, et n'ayant, ni dans l'une ni dans l'autre, produit autre chose que des essais de vulgarisation et, ça et là, de menues contributions à des travaux dont l'ensemble me dépassait.

Je n'ai pas plus de titre industriel que de titre scientifique pour m'adresser à vous aujourd'hui. Je suis ingénieur de formation, mais cette formation est déjà trop lointaine pour qu'il m'en reste quelque chose. Je suis depuis quatre ans Chef de Service dans une grande entreprise publique, ce qui me nourrit de problèmes concrets, mais j'ai la charge d'un service d'études, ce qui me dispense, et même, afin que l'étude demeure objective, m'interdit de prendre aucune décision de gestion.

Engagé dans la recherche économique parce que j'étais réputé ingénieur, j'ai été, près de dix ans plus tard, engagé dans l'industrie parce que j'étais réputé économiste. Soyez assurés, Messieurs, que le paradoxe d'avoir été choisi

pour vous parler de l'apport de la statistique à une gestion industrielle économiquement cohérente se présente à mes yeux aussi lumineusement qu'aux vôtres.

Plus clairement encore que pour nous tous, sans doute, ce paradoxe a été vu par Monsieur Darmois, qui n'est pas seulement un homme de science mondialement écouté, un universitaire respecté de tous ceux qui l'approchent, un éveilleur de vocations scientifiques et un guide sûr, mais aussi un chef d'entreprise. S'il m'a confié la mise en place du nouveau cycle d'études dont l'idée est sienne, ce qui m'oblige à assumer les premiers exposés de ce cycle, c'est en pleine conscience de mon incompétence. Je savais trop bien n'avoir rien à lui apprendre à ce sujet pour hésiter à accepter.

Et ce fut l'occasion d'une nouvelle leçon. Rien n'est aigu, obsédant, et finalement lucide, comme le sens de ce qui manque, dès lors qu'on en a reconnu le besoin. Il arrive que le technicien d'une discipline conduise celle-ci à un degré élevé d'accomplissement sans pour autant percevoir tous les services qu'elle rend. Mais quiconque a reconnu ce qu'une discipline peut lui apporter devient,vis à vis du technicien, d'une grande exigence, lui pose des problèmes, lui "passe des commandes", stimule sa recherche.

Il n'est donc pas absurde de penser que c'est la conscience même de mon incompétence qui me qualifie ici. Elle me qualifie plus spécialement peut-être dans un pays et à une époque où les relations entre université et industrie, entre hommes de science et homme d'action,ne sont pas ce qu'elles devraient être pour que la nation toute entière tire pleinement profit de ses virtualités. J'ai éprouvé, dans un milieu universitaire, le besoin de l'information concrète. J'éprouve aujourd'hui, dans un milieu industriel, le besoin de ce minimum d'abstraction qui écarte les faux problèmes dont est encombrée la vie quotidienne et qui fixe l'attention sur les vrais.

Dans l'un et l'autre cas, cette forme particulière de la connaissance qu'est la statistique est un secours immense : elle fonde l'information sur des données concrètes mais elle l'ordonne pour la rendre intelligible, dégageant le nécessaire du contingent, le permanent de l'accidentel.

Des techniciens de l'analyse statistique diront, dans la suite de notre série d'exposés et de travaux, comment traiter l'évènement pour le comprendre, l'exploiter et , éventuellement, le prévoir. Aujourd'hui, et au cours des deux prochaines séances, un utilisateur de la statistique parlera de ce qu'il en attend et de ce qu'il a appris à n'en pas attendre.

2. Trop de définitions de la statistique ont été données pour que je me sente capable d'en choisir une ou d'en formuler une nouvelle. Mieux vaut sans doute essayer de percevoir ce qui, certainement, est de son ressort, et ce qui certainement n'en est pas, sans se soucier d'enserrer dans le domaine clos d'une définition le contenu d'une discipline qui se prolonge dans beaucoup d'autres en mêlant son apport à ceux de ses voisins.

Dans sa manifestation la plus simple, la statistique décrit brièvement une population d'objets, le mot "brièvement" ayant ici un sens double.

- d'une part, elle ne rend pas compte de tous les aspects des objets décrits mais seulement de certains d'entre eux, ceux de leurs "caractères" que l'on juge utiles à l'information cherchée.

- d'autre part, elle ne tient pas compte de l'affectation des caractères sélectionnés aux objets qui les portent, mais seulement de la manière dont se groupent ces caractères autour de valeurs centrales et de la manière dont, éventuellement, ils s'associent.

Contrairement à la monographie qui résume l'état d'un individu, la statistique résume l'état d'un ensemble d'individus considéré d'un point de vue donné. La statistique "dépersonnalise". Les résumés qu'elle donne des groupages de populations par rapport à un caractère donné, ou à quelques caractères donnés, s'appellent des distributions.

Lorsque l'on considère plusieurs caractères des individus qui composent la population analysée, ils peuvent être indépendants les uns des autres, ou au contraire interdépendants. Dans le premier cas, on dit qu'il n'y a pas corrélation entre eux, dans le second cas qu'il y en a une : corrélation simple si deux caractères seulement sont en jeu, corrélation multiple s'il y en a plus de deux.

L'existence de corrélations soumet l'analyste à une grande tentation, la tentation du jugement finaliste que tout homme porte en lui. Il faut s'en défendre avec vigilance. Dire cela n'est d'ailleurs pas nier la finalité ; c'est simplement nier que le mécanisme de l'analyse statistique en rende compte. Une des explications du scepticisme que l'on rencontre trop souvent à l'égard de la statistique se trouve non dans cette discipline mais chez ceux qui s'en servent sans avoir pris conscience avec assez de rigueur de ce qu'elle apporte et de ce qu'elle ne peut pas apporter.

La statistique est neutre devant les phénomènes qu'elle traduit. Si elle révèle une corrélation entre un caractère A et un caractère B, cela ne signifie pas que A soit la cause de B ou B la cause de A, qu'il s'agisse de cause prochaine ou de cause lointaine. Bien avant que naisse la science statistique, Bacon avait déjà mis en garde contre les jugements hâtifs sur la concomitance ou la succession des évènements, mais l'impatience de l'homme est si tenace, en dépit de ce qu'il apprend, génération après génération, qu'il faut sans cesse renouveler de tels avertissements.

La statistique n'est, comme toute science, qu'un instrument, passif et neutre. Si l'on soupçonne une relation de cause à effet entre deux caractères, on peut tenter de vérifier cette relation par l'instrument statistique ; celui-ci dira simplement alors s'il n'est pas incorrect de soutenir l'existence d'une telle relation et, dans l'hypothèse où, en effet, cela n'est pas incorrect, précisera la qualité de cette relation par la valeur d'un indice de liaison entre les caractères considérés. Mais il faut toujours se rappeler que la liaison de cause à effet est une hypothèse introduite par l'analyste et non un argument fourni par l'analyse.

3. Lorsque l'analyste donne ainsi un rôle privilégié à l'un des deux caractères dont il étudie la relation (supposons pour simplifier qu'il n'y en a que deux), il procède à une étude de régression, qui n'est qu'une partie, délibérément choisie, d'une étude de corrélation. Lorsque, dans l'industrie gazière, on s'interroge sur la nature et la qualité de la relation qui existe entre le volume de gaz émis dans une ville donnée au cours d'une journée et la température observée au cours de cette journée à l'observatoire météorologique de cette ville, on est en droit de poser à priori une relation de cause à effet entre la température et l'émission, c'est-à-dire d'étudier les variations de l'émission, variable dépendante, en fonction de la température, variable indépendante. C'est dans ce sens là évidemment qu'il faut lire la relation, mais l'instrument statistique écrit la relation dans les deux sens et permet parfaitement à un aliéné de la lire à l'envers et de construire une théorie des variations de la température à l'observatoire d'une ville en fonction des émissions de gaz dans cette ville.

Tous les cas d'interdépendance ne sont malheureusement pas aussi simples, et il n'est, dans un grand nombre d'entre eux, pas nécessaire d'être aliéné pour se tromper de sens. Un exemple en est fourni par l'étude des relations entre les dépenses et les revenus d'une population de ménages. Il est commode de parler

des variations de la composition des budgets de ménages en fonction des revenus de ces ménages et il n'est pas absurde de penser que, statistiquement, dans une collectivité socialement équilibrée dont l'économie est en expansion, c'est bien dans ce sens là qu'il faut lire la relation. Mais, alors que, dans la relation des températures et des émissions de gaz, un sens de lecture s'imposait, on peut, ici, hésiter, et par conséquent on doit hésiter : le revenu peut, dans certains cas, être une fonction de telle dépense particulière. Je ne parle pas ici de cas personnels, isolés dans un ensemble de cas contraires, car la statistique, je l'ai dit tout-à-l'heure, "dépersonnalise", vise les groupes et non les individus. Je parle de cas de conditionnement de la population toute entière, ou de groupes identifiables au sein de cette population. Une analyse faite dans les deux sens éclairerait les phénomènes d'incompressibilité de certaines dépenses (qui ne sont pas nécessairement les seules dépenses de nourriture,) et rendrait sans doute apparente, selon l'interprétation proposée par mon ami Henri Aujac (1), la genèse de nombreuses inflations monétaires.

L'erreur de sens dans la lecture d'une relation n'est pas seulement une inversion de raisonnement, attribuant à la grandeur x_1 une valeur x_{1i} parce que la grandeur x_2 a la valeur x_{2j}, alors qu'au contraire peut-être x_2 vaudrait x_{2j} parce que $x_1$vaudrait x_{1i} L'erreur de sens est aussi une erreur quantitative car le propre d'une relation aléatoire entre deux grandeurs est d'être résumée par un ensemble de deux fonctions différentes fournissant respectivement deux lois de régression de l'une des variables par rapport aux valeurs de l'autre.

Pour illustrer les conséquences d'une séparation arbitraire des deux lois, on pourrait reprendre l'exemple des dépenses et des revenus. Mais il sera peut être plus instructif de se référer à l'analyse qu'ont présentée MM. GIGUET et MORLAT (2) des causes de la querelle, que tout industriel connait bien, des "ingénieurs" et des "financiers". Le coût d'un programme d'équipement est presque toujours sous-estimé par les ingénieurs qui l'établissent, ce qui vaut souvent à ceux-ci d'être soupçonnés par les financiers de procéder volontairement à des évaluations insuffisantes pour mieux "faire passer" leur programme. MM. Giguet et Morlat ont démontré que la bonne foi des ingénieurs n'était pas en cause ; je voudrais simplement rendre compte sommairement de cette innocence et de ses conséquences, en espérant ne pas déformer la pensée des auteurs par le caractère sommaire, auquel me contraignent la durée et le contenu de cet exposé, de la présentation de ce qu'ils ont démontré avec rigueur.

Un programme est constitué par un ensemble d'opérations. Chaque opération conduit à l'étude de plusieurs projets. Nous admettrons que la distribution des erreurs relatives commises sur les projets est "normale", c'est-à-dire notamment symétrique (3), surestimations et sous-estimations se compensant sur un ensemble suffisant de projets. L'ingénieur énonce son estimation, qui approche le coût réel, inconnu, tantôt par excès et tantôt par défaut.

Si, en rapportant les dépenses concernant chaque projet à une unité du produit et en supposant les projets très nombreux, on examine alors la régression du coût estimé par rapport au coût réel, pour l'ensemble des projets étudiés, on en obtient une représentation qui n'est autre que la bissectrice des axes sur lesquels

---

(1) Cf. "Economie Appliquée", Avril-Juin 1950 : "l'Influence du comportement des groupes sociaux sur le développement d'une inflation". - 1 - Henri AUJAC : "Une hypothèse de travail ; l'inflation, conséquence monétaire du comportement des groupes sociaux."

(2) "Les causes d'erreur systématique dans la prévision de prix des travaux" dans les annales des Ponts-et-Chaussées de Septembre-Octobre 1952.

(3) Encore que pour certains d'entre eux, GIGUET et MORLAT l'ont montré, cette distribution soit nécessairement dissymétrique, ce qui est une cause supplèmentaire de sous-estimation.

on a porté respectivement les valeurs de l'un et l'autre coûts. Tout parait aller très bien : à chaque valeur du coût réel correspondent des valeurs du coût estimé qui sont en nombre égal en deçà et au-delà du coût réel et symétriques deux a deux par rapport à ce coût réel, de sorte que la somme des erreurs est nulle. En particulier, si tous les projets étudiés étaient réalisés, le coût réel du programme serait égal à son coût estimé.

Mais groupons les points représentatifs des projets parallèlement à l'axe des coûts réels (fig.1). Le lieu des centres de symétrie de chaque ligne ainsi constituée est une **autre** droite, qui passe par le point moyen de l'ensemble mais qui s'écarte de la bissectrice des axes vers les valeurs élevées des coûts réels correspondant aux faibles valeurs des coûts estimés et vers les valeurs faibles des coûts réels correspondant aux valeurs élevées des coûts estimés. La régression des coûts réels par rapport aux coûts estimés est différente de la régression des coûts estimés par rapport aux coûts réels.

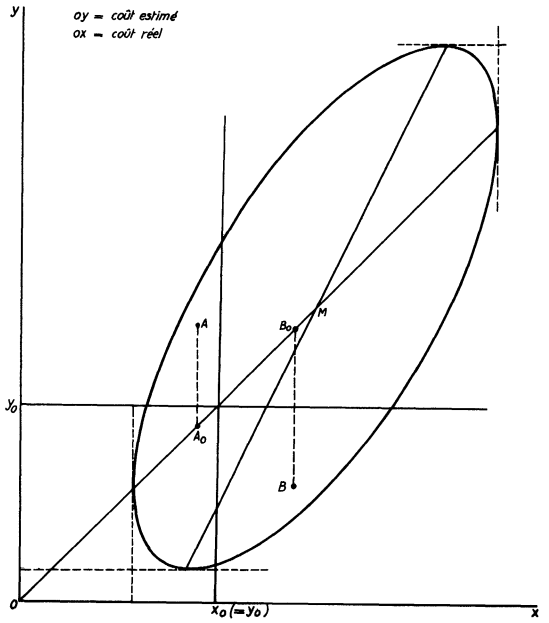

Fig. 1. - Les erreurs d'estimation du coût d'un programme de travaux.

N.B. Il est commode de figurer le nuage des points représentatifs des projets par la plus petite courbe le contenant. En raison de l'hypothèse faite sur la distribution des erreurs, cette courbe est une ellipse. Bien entendu, tous les points contenus à l'intérieur de l'ellipse ne représentent pas des projets, ceux-ci étant en nombre limité ; cette considération évidente est à rapprocher de celle que l'on fera plus loin à l'occasion d'une figure analogue (fig.2).

L'ingénieur devant retenir un certain nombre de projets retiendra ceux dont il a estimé les coûts les plus faibles, par exemple ceux dont les coûts estimés sont inférieurs à y_0. Il éliminera donc un projet tel que A, dont le coût réel est cependant inférieur à celui d'un projet tel que B, qu'il retiendra. Le programme terminé, c'est-à-dire tous les projets retenus ayant été exécutés, le financier aura eu a dépenser une somme supérieure à celle que lui avait annoncée l'ingénieur. Mais si l'ingénieur agissait correctement, présentait le projet A et refusait le projet B, le financier ne protesterait-il pas dès avant la mise en oeuvre du programme ? Accepterait-il un projet estimé sensiblement plus cher qu'un autre qui serait écarté ? Disons tout de suite d'ailleurs qu'une telle éventualité n'est pas réalisable, car si l'ingénieur sait que toutes ses évaluations sont approximatives, il ne peut évidemment pas, à l'avance, prévoir le sens et la valeur de l'erreur commise sur **chaque** projet ; s'il le pouvait, il la corrigerait.

Du fait, par conséquent, du procédé de sélection des programmes, qui est un choix des projets les moins coûteux parmi un ensemble de projets possibles, **le coût réel du programme est bel et bien fonction de son coût estimé** , et ne peut être que supérieur au coût estimé. Si l'on décidait de choisir les projets les plus coûteux, le coût réel serait inférieur au coût estimé ; ni les ingénieurs ni les financiers, cependant, ne recommanderont ce procédé de sélection.

4. Les considérations et exemples qui précèdent, malgré leur extrême simplicité, ont mis en évidence le rôle de l'opérateur dans l'analyse statistique.. Comme toujours, (c'est une banalité qu'on ne doit pas s'excuser de répéter) l'opérateur est actif et l'instrument passif. Dans une science d'observation, l'opérateur a des phases de passivité nécessaire, qui sont les phases pendant lesquelles il laisse l'instrument enregistrer une action qu'il ne doit pas troubler. Nous venons de voir qu'il trouble parfois le phénomène lui-même et parfois l'observation de celui ci, en adoptant une attitude arbitraire. Il n'en reste pas moins que la statistique ne lui fournira les services qu'il en attend que s'il l'utilise à bon escient et s'il garde son sens critique toujours en éveil devant les résultats qu'elle fait apparaitre.

En d'autres termes, la statistique n'a sa pleine efficacité que si elle est utilisée dans le cadre d'un corps d'hypothèses qu'elle ne peut pas fournir elle-même. Dans la démarche qui permet d'accéder à la connaissance, et qui est faite alternativement d'invention et d'observation, l'observation statistique peut cependant **suggérer** des hypothèses nouvelles et celles-ci auront à être vérifiées par de nouvelles observations statistiques, mais chaque opération statistique repose, explicitement ou implicitement, consciemment ou inconsciemment, sur un corps d'hypothèses formulé ou informulé, que les économistes appellent aujourd'hui un modèle.

Même l'opération la plus simple correspond à un modèle. Etudier les variations des émissions de gaz en fonction de la température, c'est supposer à priori l'existence d'un lien entre ces deux grandeurs et chercher la nature de ce lien ainsi que la valeur des paramètres qui le caractérisent ; c'est aussi , si l'on étudie uniquement cette liaison, supposer à priori que deux variables seulement sont en jeu, ou accepter délibérément de ne donner qu'une représentation extrèmement simplifiée d'une réalité que l'on sait beaucoup plus complexe.

La première hypothèse se révèle valable, puisque l'on constate effectivement une relation entre les deux grandeurs observées, mais un observateur un peu averti ne se contentera pas de cette première hypothèse et, pour complèter son investigation, cherchera à mettre en évidence les autres facteurs qui peuvent avoir un effet sur le volume des émissions de gaz. La première idée qui vienne à l'esprit est que l'émission d'une journée donnée peut dépendre aussi de la température de la veille ou des jours précédents. Une autre idée est évidemment que, tous les jours de la semaine n'étant pas identiques, leurs différences interviennent dans les variations du volume de l'émission. Des essais doivent alors être faits pour éprouver les relations qui correspondent à ces hypothèses.

Mais si l'on est exigeant, ou si l'on est simplement guidé par certaines nécessités, on peut faire d'autres découvertes. C'est ainsi qu'on a cru remarquer dans la région parisienne, au cours de plusieurs hivers successifs, que le gradient d'émission en fonction de la température n'avait pas la même valeur pour des périodes de quelques semaines que pour l'ensemble de l'hiver. Sans qu'un nombre suffisant d'observations ait encore été fait pour qu'on puisse être affirmatif, il semble bien que les gradients correspondant à des périodes partielles soient systématiquement plus faibles que le gradient de l'ensemble. Le nuage de l'ensemble de l'hiver apparait constitué par une superposition de nuages

disposés en escalier, chaque marche de cet escalier étant elle-même inclinée et correspondant à une des périodes analysées. Mais à chaque période correspond aussi approximativement une zone de température ; si le gradient résultant est plus élevé que chacun des gradients composants, c'est parce que la marche la plus élevée de l'escalier correspond à la période la plus froide. La température n'est pas la seule variable saisonnière qui conditionne les émissions de gaz. En limitant l'analyse aux relations entre l'émission et la température, on laisse échapper au moins un facteur dont l'effet, non négligeable, demeure ignoré. La recherche de ce facteur, ou de ces facteurs, est actuellement en cours au Gaz de France.

5. Nous avons vu que la statistique donnait une information résumée. C'est bien cela qu'on lui demande car l'information complète est inutilement encombrante : on ne décrit pas une forêt en en décrivant tous les arbres, ni un arbre en en décrivant toutes les feuilles. Cette information est aussi résumée qu'on le veut, mais une concision excessive ne fournit pas une information efficace. La seule valeur moyenne d'un ensemble de caractères est généralement inexpressive : que la profondeur moyenne d'une pièce d'eau soit 30 cm ne signifie pas qu'un homme ne puisse s'y noyer. En revanche une moyenne assortie d'un indice de dispersion (par exemple l'écart-type dans une distribution normale) suffit généralement à décrire la distribution en cause.

La complexité de certains phénomènes, ou le besoin d'utiliser une expression puissante par sa valeur synthétique, conduit à l'utilisation du langage symbolique. L'usage des équations caractéristiques, familier à l'homme de science, déroute parfois l'homme d'action ou alimente son scepticisme. Cela signifie seulement que la pratique de la statistique n'est pas entrée dans les moeurs. Quel industriel reproche leur concision aux équations chimiques ? Quel industriel met en doute l'aptitude de ces équations à représenter des phénomènes qu'il serait incapable de décrire en langage courant ? Et pourtant, qu'y-a-t-il de plus abstrait qu'une équation chimique ? On ne pense plus à leur abstraction parce qu'on en a pris l'habitude.

Le caractère apparemment abstrait des représentations statistiques ne saurait constituer un obstacle durable à un emploi courant de l'instrument statistique dans les domaines qui le sollicitent. Les obstacles à son emploi efficace ne tiennent sans doute pas tant à des difficultés d'accoutumance qu'à des difficultés d'intelligence. Intelligence de l'objet par le sujet, bien entendu, intelligence au sens originel, l'intelligence du sujet, au sens commun, n'étant pas en cause. Il est en effet difficile, psychologiquement, d'accepter qu'un instrument de connaissance fournisse généralement, tout en étant un des plus pénétrants de la science moderne, ses informations sous forme restrictive et parfois négative. "Généralement", parce qu'il est, en fait, rare, qu'on lui demande simplement de résumer un ensemble exhaustivement connaissable. "Généralement", on lui demande, à partir d'un ensemble restreint, de decrire un ensemble plus vaste, soit que l'on prélève un échantillon sur une population théoriquement connaissable dans son entier, soit, ce qui est plus risqué mais sans doute plus fréquent encore, que l'on utilise des données passées et présentes pour prévoir des évènements futurs. La réponse doit généralement être lue sous la forme "dans telles conditions, il n'est pas incorrect de penser que..."

En d'autres termes, on demande beaucoup à la statistique de dégager, d'un lot de données variables, des invariants, d'un lot de données transitoires, des permanences. De la statistique, on attend des lois, mais les lois sont rarement universelles et immuables, et, dans le domaine économique et social, ne s'imposent jamais à un évènement donné qu'avec une certaine approximation.

6. Il y a des opérations répétables, telles que, en gros, les opérations de la physique et de la chimie classiques ; on peut, par leur répétition, vérifier leur soumission au déterminisme scientifique, selon lequel aux mêmes causes correspondent toujours les mêmes effets. Mais le rassemblement d'un ensemble donné de causes et l'appréciation de leurs effets ne se font jamais qu'avec approximation : la "théorie des erreurs" est encore, de nos jours, la première leçon de statistique enseignée aux jeunes gens, généralement dans un cours de physique. C'est la répétition, un grand nombre de fois, de la même expérience qui permet d'accéder à une mesure correcte du phénomène observé. Une modification dans les causes apporte une modification dans les effets et conduit, de nouveau, à répéter un nombre suffisant de fois l'expérience ainsi modifiée pour obtenir la mesure du phénomène associé à la nouvelle combinaison de ses causes. Une série d'expériences analogues, chacune répétée assez souvent, fournit une série de mesures, dont, quand on le peut d'une façon simple, on décrit la succession par l'énoncé d'une loi, en admettant qu'on peut traiter par interpolation l'ensemble des valeurs non mesurées.

Si, sous réserve de multiplier suffisamment les expériences (et non simplement les mesures de chacune d'elles), on peut se considérer en droit d'interpoler on comprend aisément qu'on n'ait pas le droit d'extrapoler ou de généraliser autrement que sous forme d'hypothèses, c'est-à-dire dans le cadre de modèles auxquels on n'est pas certain que se plie la réalité. Ainsi, les lois physiques elles-mêmes sont des lois statistiques non seulement en ce sens que les constantes qui les définissent ne sont connues qu'approximativement, mais encore en ce sens que leur validité risque de s'affaiblir à mesure que l'on s'écarte du domaine au sein duquel on les a vérifiées. Un exemple classique de cette approximation des lois physiques est celui de la relation qui existe entre la pression, le volume et la température d'un gaz. Les lois de Mariotte et de Gay-Lussac ne sont que des lois limites caractérisant l'état gazeux parfait. Un gaz parfait est un modèle défini par l'équation :

$$pv = RT,$$

modèle dont s'écartent les gaz réels. Tout ingénieur se rappelle l'équation de Van der Waals qui s'efforce d'exprimer le comportement de chaque gaz réel en corrigeant p et v par l'introduction de deux constantes caractéristiques du gaz considéré et en ajustant ses isothermes non plus à une hyperbole mais à une courbe du 4° degré. L'équation de Van der Waals a enrichi la connaissance du comportement des gaz et des phénomènes de changements de phases, mais elle n'est elle-même qu'une loi approchée que d'autres physiciens ont proposé de corriger à son tour.

7. Dans les collectivités humaines, les opérations ne sont pas répétables parce que les hommes naissent et meurent, parce que chaque homme change à chaque moment de sa vie, parce que les hommes créent incessamment des conditions nouvelles d'existence. A travers tous ces changements, cependant, on peut trouver des permanences : permanences dans les obstacles que la nature des choses oppose à l'invention de l'homme, permanences dans le comportement des groupes sociaux, permanences dans les mouvements de populations et les fluctuations démographiques.

Les phénomènes, et notamment les phénomènes économiques et sociaux, ne sont donc pas inaccessibles à certaines mesures durables, n'échappent pas complètement à l'investigation statistique. Nous en verrons des exemples au cours des prochaines séances. Ce qu'il faut, ici, voir clairment, c'est la manière dont, dans la vie courante et singulièrement dans l'activité industrielle, sont posées les questions aux statisticiens.

On veut savoir ce qu'il faut faire aujourd'hui pour obtenir tel résultat demain ou dans dix ans, ou encore comment se dérouleront, dans l'avenir, les conséquences d'un acte accompli aujourd'hui.

Ce n'est plus alors le langage statistique qu'il faut employer, mais le langage probabiliste. Ce n'est plus une population d'évènements réels que l'on cherche à résumer mais la population des chances de réalisation d'un évènement souhaité ou simplement prévu. Dans le cas, relativement fréquent dans l'industrie, où l'évènement considéré peut être représenté par une grandeur, la question consiste généralement à savoir soit quelle est la valeur la plus probable que cette grandeur atteindra à une date donnée, soit quelle chance a cette grandeur d'atteindre, à une date donnée, une valeur donnée.

L'étude des valeurs les plus probables intervient notamment dans les estimations de rentabilité d'un programme de fabrication projeté. Nous retrouverons ce problème dans la suite de nos réunions.

L'étude des chances d'atteindre une valeur donnée à une date donnée intervient dans la détermination d'une capacité de production, notamment lorsque le produit attendu se prête mal à la mise en stock. Nous retrouverons aussi ce problème, mais il n'est pas inutile d'en préciser dès maintenant le sens, car c'est peut-être lui qui révèle le mieux, parmi les problèmes industriels courants, le changement d'optique qui distingue le raisonnement statistique du raisonnement probabiliste.

L'industriel qui détermine une capacité de production s'intéresse moins à la production moyenne qu'il devra réaliser dans des circonstances données qu'à la production maximum compatible avec les mêmes circonstances ou exigées par celles-ci. Plus précisément, l'industriel est conduit, pour déterminer une capacité de production, à confronter les lois de probabilité des circonstances qu'il envisage et la loi de probabilité du besoin de production associé à chaque jeu de circonstances, puis à retenir la capacité de production nécessaire pour assurer son service avec un risque de défaillance égal à une valeur qu'il se donne.

8. La détermination des émissions de gaz dans une ville donnée constitue un exemple assez typique de ce problème. Utilisons-le donc de nouveau.

Ce qui a été dit plus haut s'applique à un moment de la vie de cette ville, pratiquement à un hiver pendant lequel on suppose que la population de la ville ne varie pas et que l'équipement de cette population en appareils d'utilisation du gaz ne varie pas. Il est bien évident que si cet équipement se développe, les émissions augmenteront et notamment que, si le nombre et la puissance des appareils de chauffage au gaz augmentent, le gradient de l'émission en fonction de la température augmentera. Supposons correctement estimé le développement de l'équipement et l'usage qui en sera fait à la date, future, pour laquelle on se pose la question, et correctement calculé le gradient correspondant, sur l'ensemble de l'hiver, compte tenu de tous les facteurs susceptibles d'entrer en jeu. Admettons enfin que la loi, aléatoire, de variation de l'émission en fonction de la température soit effectivement linéaire dans tout l'intervalle dont on a besoin. Comme il serait absurde d'admettre que tous les appareils d'utilisation du gaz fonctionneront simultanément, à plein régime, pendant vingt-quatre heures deux questions importantes demeurent :

1°. Jusqu'à quelle température extrême faut-il garantir les émissions de gaz ?

2°. Quelle peut-être, au maximum, l'émission appelée par les consommateurs ?

A la première question, on ne peut répondre qu'après avoir **décidé** le risque de défaillance que l'on accepte. Il est exclu de refuser tout risque, c'est-à-dire de se mettre en mesure de faire face à toute température dont la probabilité n'est

pas nulle. Il n'y a pas de probabilité nulle ; il faut donc s'arrêter, arbitrairement à une probabilité faible. Constater cela, c'est admettre un risque.

Les températures moyennes de chaque jour sont enregistrées et conservées dans les observatoires. Pour les observatoires de Paris, par exemple, (St Maur et Montsouris), on connait cette température jour par jour depuis une cinquantaine d'années. Relevons la température de la journée la plus froide de chaque année et observons la distribution de la série ainsi constituée ; on constate qu'elle s'ajuste assez bien à la loi dite "normale", caractérisée par une moyenne et un écart-type. On admettra qu'elle est effectivement "normale".

Si l'on retenait, comme température extrême de garantie des émissions de gaz, la valeur moyenne t_o de cette série, on s'exposerait, en admettant que l'émission est une fonction déterminée de la température, à être défaillant, pendant un jour ou quelques jours, une année sur deux. Il est exclu que l'on donne aux consommateurs une aussi faible garantie. Jusqu'où faut-il donc aller ? Faut-il accepter d'être défaillant une fois par millénaire ? Une fois par siècle ? Deux fois par siècle ? Le premier choix n'est pas raisonnable, le deuxième traduit une extrême prudence en même temps qu'une confortable richesse de la collectivité; il ne serait pas raisonnable de condamner le troisième, qu'on fera, en le définissant, par convention de langage, comme l'acceptation d'un risque de 2 %.

Dans une distribution normale, 96 % des valeurs sont comprises entre la moyenne diminuée de deux écarts-types et la moyenne augmentée de deux écarts-types (1) ; 4 % débordent cet intervalle, 2 % en deçà et 2 % au-delà. Seules nous intéressent, en l'occurence, les faibles valeurs de la température ; on assumera donc un risque de défaillance de 2 % en se mettant en mesure de faire face à une température égale à la moyenne des températures moyennes du jour le plus froid de l'année diminuée de deux écarts-types de la distribution de ces températures moyennes. Pour la région parisienne, en observant les températures au Parc de St-Maur, cela donne :

$$- 5°,45 - 2°,93 \times 2 = - 11°,31$$

Reste la dernière question ; quel est le volume maximum de gaz, compatible avec le risque de 2 %, qui peut être appelé, au cours d'une journée de l'hiver considéré, par les consommateurs de la région parisienne ? La réponse parait simple puisque nous avons admis que nous ne nous étions pas trompés dans la détermination de la loi des variations de l'émission en fonction de la température et que, pour toute température inférieure à une température de référence t_r, cette loi était linéaire. Le volume cherché parait donc donné par l'équation :

$$v = a + b\,(t - t_r)$$

dans laquelle on connait a, b et t_r. (2).

Compte tenu de l'approximation des calculs prévisionnels, c'est bien ainsi que l'on procède, en prenant empiriquement la classique "marge de sécurité" que l'on ajoute à la valeur trouvée pour v (3). Il est cependant nécessaire de pousser plus loin l'analyse, non seulement par souci de rigueur, mais encore pour les besoins du contrôle des prévisions.

---

(1) Plus précisément 2,06 écarts-types de part et d'autre de la moyenne.

(2) La rédaction des pages qui suivent a été sensiblement modifiée par rapport au texte de l'exposé oral.

(3) La marge de sécurité couvre non seulement le risque d'erreur dans les estimations prévisionnelles, mais aussi certains risques de défaillance mécanique. Une large part d'empirisme est donc inévitable dans sa détermination. Cette part d'empirisme diminue à mesure que s'améliorent les prévisions de consommation et la connaissance du matériel utilisé.

Soit donc t_i une température quelconque d'une journée d'hiver. L'émission de gaz qui lui correspond a une valeur **moyenne**

$$\overline{v_i} = a + b\,(t_i - t_r),$$

mais l'émission d'une journée **donnée** de température t_i, est

$$v_i = a + b\,(t_i - t_r) + e_i,$$

e étant un résidu aléatoire (ou écart à la moyenne), positif ou négatif, évidemment inconnu à priori.

Ce qui est vrai pour une journée de température t_i quelconque est vrai notamment pour la journée la plus froide retenue, dont nous désignerons la la température par t_m. Ainsi, lorsque l'on fixe la capacité de production à un volume

$$\overline{v_m} = a + b\,(t_m - t_r)$$

on a une chance sur deux, lorsque l'on atteint t_m, d'avoir à émettre un volume inférieur à $\overline{v_m}$, mais on risque une fois sur deux d'avoir à émettre un volume supérieur à $\overline{v_m}$, c'est-à-dire d'être défaillant même pour la température garantie t_m.

Si l'on peut être défaillant pour t_m, on peut l'être aussi pour une température légèrement supérieure à t_m. On peut même l'être, théoriquement, pour une température largement supérieure à t_m puisque les écarts e_i sont aléatoires et qu'il n'a donc pas de probabilité nulle pour que l'un de ces écarts soit très grand.

Mais, ici encore, si l'on ne peut connaitre des valeurs certaines, on peut calculer les risques de dépassement de la moyenne. La distribution des émissions de gaz, au cours d'un hiver donné, a une dispersion propre caractérisée, si on la suppose normale, par un écart-type σ. La liaison entre émission et température est caractérisée par un coefficient de corrélation ρ dont la signification est de diminuer l'incertitude sur v quand on connait t. Pour une valeur quelconque t_i de t, l'écart-type σ_i de la distribution des écarts à $\overline{v_i}$ est en effet :

$$\sigma_i = \sigma \sqrt{1 - \rho^2}$$

On voit que σ_i est indépendant de t_i. On voit aussi que, si par exemple $\rho^2 = 0,91$, valeur qui a été constatée dans les travaux concrets, on a :

$$\sigma_i = 0,3\sigma$$

Pour chaque valeur t_i de t on a ainsi, en se référant aux tables de la loi normale, en ne retenant que quelques points de repère simples et ne ne tenant compte que des valeurs de v_i supérieurs à $\overline{v_i}$:

- 98 chances sur 100 pour que v_i soit inférieur à $\overline{v_i} + 0,6\,\sigma$
- 82,5 chances sur 100 pour que v_i soit inférieur à $\overline{v_i} + 0,3\,\sigma$
- 75 chances sur 100 pour que v_i soit inférieur à $\overline{v_i} + 0,2\,\sigma$

Revenons à l'émission à attendre pour une journée de température t_m. Si l'on s'est donné un risque de 2 % sur cette température t_m du jour le plus froid de l'année et si, pour l'année qui fait l'objet des prévisions, on estime à un million de mètres cubes (valeur vraisemblable dans la région parisienne) l'écart-type de la distribution des émissions journalières d'hiver, on constate que :

a) par définition, on risque approximativement deux fois par siècle d'atteindre t_m ou une température inférieure à t_m ;

b) **lorsque** l'on atteint t_m, l'émission a :

- 50 chances sur 100 de ne pas dépasser la valeur $\overline{v_m} = a + b\,(t_m - t_o)$
- 75 chances sur 100 de ne pas dépasser la valeur $\overline{v_m} + 200.000$
- 82,5 chances sur 100 de ne pas dépasser la valeur $\overline{v_m} + 300.000$
- 98 chances sur 100 de ne pas dépasser la valeur $\overline{v_m} + 600.000$

Ajouter 200.000, 300.000 ou 600.000 m³/jour à la capacité de production qui correspondrait à t_m dans le cas d'une liaison fonctionnelle entre t et v revient à associer au choix du risque pris sur la température le choix d'un risque pris sur la consommation de gaz qui peut correspondre à cette température, mais cela ne renseigne pas sur le risque pris sur l'émission d'une journée de température t_i supérieure à t_m. Certes, t_i étant supérieur à t_m et σ_i étant invariable, $\overline{v_i} + n\sigma_i$ est toujours inférieur à $\overline{v_m} + n\sigma_i$. Mais y-a-t-il lieu de prendre sur l'émission, la même sécurité pour toutes les valeurs de la température ?

Ce qui intéresse l'exploitant, c'est un risque global, une situation comportant un risque de défaillance donné, que cette situation corresponde à une température ou à une autre. Comme on définit aisément un certain risque de rencontrer telle ou telle température, il est légitime d'utiliser la température comme repère, sans pour autant lier l'émission maximum à cette température.

On peut donc adopter, pour déterminer la capacité de production maximum, une méthode synthétique, consistant à chercher l'émission journalière maximum correspondant à une situation ayant la même densité de probabilité (1) que la situation définie par t_m et $\overline{v_m}$.

La distribution normale d'une variable aléatoire se représente par ce que tout le monde appelle une "courbe en cloche". La distribution normale de deux variables aléatoires liées se représente par une surface, une "cloche" dont les bords sont asymptotes au plan de densité de probabilité nulle (ou, en langage plus concret, de fréquence nulle)(2), et dont les sections par des plans d'égale densité de probabilité (ou d'égale fréquence) sont des ellipses.

Nous sommes ainsi conduits à considérer la loi de probabilité de l'ensemble lié température - émission **pour le jour le plus froid de l'année en examen.** Cette loi se représente par une surface dont le sommet se projette, dans le plan t o v, sur la droite de régression des émissions en fonction des températures, au point d'abscisse t_o (moyenne de la série constituée par le relevé de la journée la plus froide de chacune des années observées. Voir plus haut). Ce point est le centre de toutes les ellipses d'égale densité de probabilité et la droite de regression des émissions en fonction des températures est le diamètre commun de ces ellipses conjugué de la direction o v. L'ellipse qui nous intéresse (fig. 2) est celle dont l'extrémité d'abscisse la plus faible de ce diamètre a pour abscisse t_m.

---

(1) Si l'on a affaire à une variable discrète, on définit sa probabilité de prendre telle valeur dans l'ensemble des valeurs qu'elle peut prendre. Si l'on a affaire à une variable continue x, on définit sa densité de probabilité comme la fonction p(x) telle que, dans le domaine ω de variation de x, on ait $\int_\omega p(x)\,dx = 1$.

Dans le cas qui nous occupe, c'est bien de densité de probabilité qu'il faut parler puisqu'il s'agit de températures futures et d'émissions futures, donc de grandeurs susceptibles de prendre n'importe quelle valeur au sein du domaine exploré.

(2) La notion de fréquence est relativement familière. Elle ne s'applique qu'à la distribution d'un nombre fini de valeurs possibles d'un caractère ou d'un ensemble de caractères. Quand le nombre de ces valeurs possibles croit indéfiniment, la limite de la fréquence est la densité de probabilité.

Parmi toutes les combinaisons de t et de v ayant même densité de probabilité que la combinaison $(t_m, \overline{v_m})$, il y en a une dans laquelle v a une valeur plus élevée que dans n'importe quelle autre ; c'est la combinaison $(t_c, \hat{v_c})$, qui correspond à l'extrémité d'ordonnée la plus élevée du diamètre de l'ellipse conjugué de la direction o t. Ce diamètre n'est autre que la droite de régression des températures par rapport aux émissions. On voit que si cette droite n'a pas de signification concrète (puisqu'il serait absurde de parler de la variation de la température en fonction de l'émission), elle n'en a pas moins une signification instrumentale, et que la liaison entre deux variables aléatoires, même lorsqu'aucune hésitation n'est permise sur la variable indépendante et sur la variable dépendante, n'est valablement exprimée, ou statistiquement résumée, que par un ensemble de deux lignes de régression.

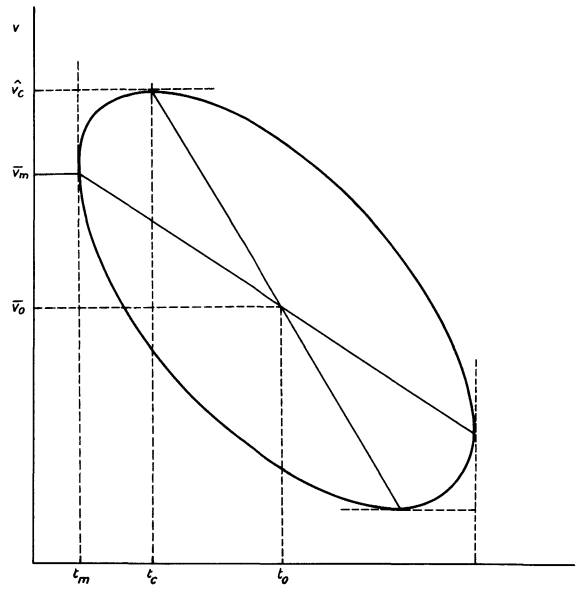

Fig.2. - Liaison entre la température et l'émission le jour le plus froid de l'année.

N.B. 1) Tout point contenu dans l'ellipse représente une possibilité compatible avec le "risque de 2 % " défini plus haut sur le jour le plus froid de l'année.
2) On a déformé la figure pour la rendre plus lisible. Voir figure 2bis.

On détermine aisément par le graphique une valeur approchée de $\hat{v_c}$, en admettant (ce qui n'est pas absurde) que le coefficient de corrélation de la liaison se conserve quel que soit l'intervalle de température observé ou supposé (1), et en utilisant la relation fondamentale :

$$\frac{dv(t)}{dt} \times \frac{dt(v)}{dv} = \rho^2$$

signifiant que le produit des pentes des deux droites sur leurs axes respectifs est égal au carré du coefficient de corrélation.

Pour la région parisienne, avec un gradient d'émission journalier voisin de 150.000 m³ par degré et une valeur de ρ^2 voisine à 0,90, on trouve un gradient de température journalier voisin de 0°,6 par centaines de milliers de mètres cubes. Plaçant l'origine des coordonnées en $(t_o, \overline{v_o})$, on trouve approximativement 900.000 mètres cubes comme valeur de $\overline{v_m} - \overline{v_o}$ et 950.000 mètres cubes comme valeur de $\hat{v_c} - \overline{v_o}$, t_c étant supérieur de près d'un demi degré à t_m (fig. 2bis). On

(1) C'est-à-dire que la droite v(t) est fixe et que la droite t(v) se déplace parallèlement à elle-même.

a donc, concrètement, autant de chances d'atteindre une émission égale à $\overline{v_o}$ + 950.000 mètres cubes par une température de - 10°,9 qu'une émission de $\overline{v_o}$ + 900.000 mètres cubes par une température de - 11°,3.

Dans la pratique, ainsi qu'il a été dit plus haut, on se contente de calculer $\overline{v_m}$ et de prendre une marge de sécurité qui couvre à la fois les erreurs d'estimation et certains risques de défaillance du matériel (gel de certains organes ou incidents mécaniques). Les comparaisons entre $\hat{v_c}$ et $\overline{v_m} + n\sigma_i$ n'interviennent donc pas dans les décisions d'équipement, mais il est clair qu'elles interviennent dans le contrôle des prévisions. Si l'on n'est pas averti de l'ampleur et de la forme des fluctuations de l'émission autour des valeurs données par la loi qui lie l'émission et la température, ainsi que de la répartition des risques dans tout le champ opératoire, on s'expose à prendre des mesures incohérentes. La pratique de la statistique est une école de sang-froid.

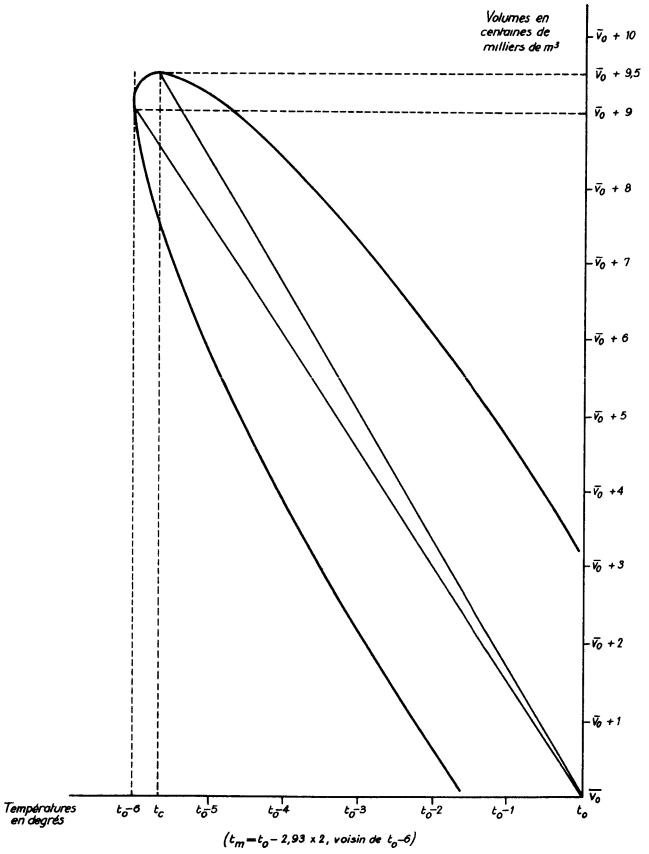

Fig.2bis. - Le coefficient de corrélation détermine l'angle des deux droites une fois fixé le rapport des échelles des coordonnées.

9. Cet exposé n'a pas été un cours de statistique. Il a tenté de faire entrevoir la forme de connaissance qu'apportait la statistique et d'introduire le sens du jugement probabiliste. Parmi les nombreux domaines dans lesquels la statistique et la probabilité trouvent un champ fécond d'applications, l'économie tient une place de choix. L'économie est une science du comportement des hommes et des groupes humains devant certains types de problèmes. Contrairement à ce que disent les sceptiques ou les ignorants, la statistique, discipline mathématique, rapproche de la vie et de ses aléas ; cest à ce titre peut-être qu'elle apparait comme un élément irremplaçable de la culture moderne.