JOURNAL DE LA SOCIÉTÉ STATISTIQUE DE PARIS

LUCIEN MARCH

# Remarques sur la terminologie en statistique

*Journal de la société statistique de Paris*, tome 49 (1908), p. 290-296

<http://www.numdam.org/item?id=JSFS_1908__49__290_0>

## II

## REMARQUES SUR LA TERMINOLOGIE EN STATISTIQUE [1]

*Fréquences et probabilités.* — La méthode de statistique intervient quand on veut mesurer, dans un ensemble, l'intensité d'un caractère, ou quand on compare des ensembles par rapport au même caractère, et que les parties de l'ensemble ne possèdent pas le caractère au même degré. On appelle alors fréquence d'un degré quelconque le rapport du nombre des parties correspondant à ce degré au nombre total des parties de l'ensemble [2].

En général, les circonstances qui ont imprimé à plusieurs parties de l'ensemble, et au même degré, le caractère étudié, nous sont inconnues. L'analyse statistique cherche à découvrir les plus importantes en comparant entre elles, sous différentes conditions, les fréquences observées. Comme base de comparaison on convient de se reporter au schéma de combinaisons et de répétitions sur lequel est fondé le calcul des probabilités.

Mais une différence essentielle sépare la fréquence observée en statistique et la probabilité mathématique. En statistique, nous ignorons les circonstances initiales des faits observés ; notre connaissance peut tout au plus s'étendre à quelques parties de l'enchaînement intermédiaire entre les conditions originelles et le résultat. Par exemple, nous ne connaissons pas les chances de mort des hommes qui entrent dans leur trentième année, et nous connaissons mal les raisons qui feront que certains seront frappés plutôt que d'autres dans le cours de cette trentième année.

Dans le schéma des probabilités les combinaisons et répétitions qui contiennent en puissance le résultat final sont complètement connues ; seul le jeu du déclenche-

---

[1] **Extrait d'une communication au Congrès de mathématiques de Rome (section des mathématiques appliquées).**

[2] **Le mot « fréquence » impliquant l'idée de répétition, il est logique d'éliminer de l'ensemble les parties qui ne sont pas susceptibles de posséder le caractère étudié.**

ment qui fait apparaître certaines combinaisons demeure imperceptible en raison de l'exiguïté de son action.

L'accord d'une distribution de fréquences, observée en statistique, avec une distribution de probabilités n'implique donc qu'une analogie apparente entre l'enchaînement des faits statistiques et la formation des probabilités. Pour l'ordonnance du travail du statisticien, cette analogie offre une grande utilité puisqu'elle permet de sérier les recherches, de négliger provisoirement les variations imputables en apparence à des cas fortuits. Dans l'exposé des résultats, il convient de ne pas laisser supposer que l'enchaînement des faits étudiés est assimilable au schéma conventionnel que l'on a pris comme terme de comparaison.

Aussi serait-il opportun, en statistique, de renoncer à l'emploi du mot probabilité pour exprimer l'attente que fait naître la constatation d'une fréquence; car, si, dans la théorie des probabilités, la convention sur laquelle repose cette attente inspire une parfaite confiance, en statistique le degré de confiance que mérite cette attente est souvent modifié par l'étude des liaisons des faits, par les enseignements des sciences spéciales.

D'ailleurs, les applications du calcul seraient à peu près aussi commodes si l'on opérait sur des fréquences au lieu d'opérer sur des probabilités; les énoncés des propositions devraient être transformés, mais ils seraient plus rigoureux et éviteraient bien des critiques adressées aux anciens énoncés.

Par exemple, lorsque Laplace contestait le caractère accidentel des écarts du coefficient de natalité masculine observé à Paris, à Londres et à Naples, en assimilant la détermination du sexe à un tirage au sort, il s'exposait aux critiques que Bertrand a adressées à son énoncé. Pourtant le statisticien, obéissant à une convention justifiée par l'expérience, a le droit de décider que les écarts dont il cherchera l'explication par une analyse ultérieure, devront dépasser une certaine limite, en deçà de laquelle les écarts sont provisoirement regardés comme accidentels. Cette limite, l'analogie, d'accord avec l'expérience, engage à la fixer d'après une échelle que détermine la théorie des probabilités.

On se borne ainsi à faire appel à ce que M. Émile Borel a appelé la valeur pratique du calcul (1). L'ampleur théorique des propositions, leur autorité sur les esprits avides d'objectivité s'amoindrissent sans doute, mais la théorie s'établit en meilleure harmonie avec les faits.

En mathématiques on distingue les probabilités *a priori*, et les probabilités *a posteriori*. Il serait parfaitement inutile de modifier cette terminologie, car les mathématiciens ne peuvent se méprendre sur le sens véritable des termes.

Dans les applications à la statistique il ne paraît pas suffisant de substituer à la notion de probabilité *a posteriori* celle de probabilité *statistique* que suggère Blaschke (2). Pour que les personnes insuffisamment averties ne se méprennent pas sur la portée de ces applications, il semble préférable d'écarter complètement de l'analyse des résultats statistiques le terme *probabilité* et le mot *probable* entendu dans le sens mathématique.

*Moyennes.* — Quand on veut comparer rapidement deux ensembles statistiques,

(1) *Revue du Mois*, avril 1906.

(2) *Blaschke, Vorlesungen, etc.*, p. 123.

on est amené à former, pour chacun d'eux, un coefficient unique susceptible de synthétiser l'intensité du caractère étudié, intensité qui n'est d'ailleurs pas uniforme. Puisque l'on a en vue une synthèse, ce coefficient doit être fonction des intensités particulières à toutes les parties ; parmi l'infinité des fonctions possibles on choisit de préférence la plus simple, celle qui exprime l'intensité totale. Enfin, pour que ce coefficient ne dépende pas de la dimension de l'ensemble on rapporte l'intensité totale au nombre des parties : le résultat est appelé moyenne arithmétique.

On fait aussi usage d'autres fonctions, ou, comme on dit, d'autres moyennes : géométrique, harmonique, quadratique, médiane, etc. ; je ne me propose pas de les comparer. Je veux simplement signaler que, dans le langage courant, le mot moyenne a un tout autre sens que dans les sciences mathématiques et physiques où il correspond à la notion de centre de gravité.

A vrai dire, dans le langage courant, le mot moyenne a un sens assez vague, celui de terme intermédiaire entre le plus grand et le plus petit. Quand on cherche à préciser la notion, en examinant, par exemple, l'établissement des cotes sur les marchés publics, on constate que, le plus souvent, le mot moyenne est entendu dans le sens de valeur la plus fréquente.

Ainsi, aux Halles centrales de Paris, les mandataires qui cotent les ventes journalières de chaque denrée ne tiennent compte que des ventes opérées par grosses quantités. Ou bien il s'établit sur le marché un prix courant auquel la quantité offerte trouve aisément acquéreur : c'est ce prix qui est indiqué comme cours moyen ; ou bien la demande n'est pas assez active pour qu'il en soit ainsi ; les grosses quantités sont traitées à des prix divers et le prix moyen s'obtient simplement en formant la demi-somme des prix extrêmes (demi-somme que l'on décore du nom de moyenne mathématique).

On peut admettre que, dans ce dernier cas, les mandataires ont simplement pour but de parvenir le plus simplement et le plus vite possible à la fixation de la valeur courante.

Le prix moyen coté aux Halles est donc autre chose qu'une moyenne arithmétique. Si les ventes par petites quantités étaient très importantes, comme elles s'opèrent généralement à des prix relativement élevés, le prix moyen coté pourrait être très voisin du prix minimum.

A la Bourse des marchandises, le cours moyen du jour est simplement la demi-somme des prix extrêmes enregistrés par les coteurs, sans avoir égard aux quantités qui, pourtant, sont enregistrées.

A la Bourse des valeurs, les conditions d'établissement de la cote sont un peu différentes. Le coteur ignore les quantités vendues ; il inscrit simplement sur un registre, avant l'ouverture du marché, les prix auxquels chaque valeur est demandée et les prix auxquels elle est offerte. Puis, au cours du marché, il note les prix auxquels se sont opérées les diverses transactions. A la fin, le cours moyen est calculé en formant la demi-somme du prix le plus haut et du prix le plus bas.

Le chiffre ainsi obtenu est naturellement différent de la moyenne arithmétique et il serait également différent de la valeur la plus fréquente sans une circonstance grâce à laquelle il coïncide en fait avec cette valeur : la majeure partie des ordres sont donnés à l'avance au cours moyen, et par conséquent le cours arbitré, comme il vient d'être dit, devient le cours auquel le plus grand nombre des titres ont été négociés.

C'est encore la valeur la plus fréquente que les ouvriers entendent exprimer quand ils calculent la moyenne des salaires; chose remarquable, quand ils revendiquent un minimum de salaire, ce minimum doit encore, dans leur esprit, s'appliquer au salaire gagné par la majeure partie d'entre eux.

Au contraire, les prix moyens calculés pour les marchandises importées dans un pays sont des moyennes arithmétiques.

Diverses expressions ont été proposées pour désigner, dans une série statistique, la valeur la plus fréquente : Lexis l'appelle la valeur normale, Pearson l'appelle le mode. Aucune des deux expressions n'est pleinement satisfaisante. Si le mot *normal* implique une règle générale, il fait naître aussi une idée de loi qui dépasse un peu l'idée de fréquence. Le terme *mode* semble indiquer que la valeur la plus fréquente exprime la manière d'être de la série, ce qui n'est point tout à fait exact, plusieurs éléments étant nécessaires pour caractériser cette matière d'être.

Acceptons le mot normal (1) : nous constaterons que la moyenne, telle qu'on l'entend dans le langage courant est, non pas la moyenne arithmétique, mais la valeur normale.

Conviendrait-il de modifier le langage scientifique pour le conformer au langage vulgaire ?

Ce serait sans doute le meilleur parti si la valeur normale pouvait être déterminée avec précision dans tous les cas. Mais il n'en est point ainsi. D'abord on n'est point tout à fait d'accord sur la question de savoir si, dans son estimation, il faut éliminer ou non les cas jugés exceptionnels. Puis, la ligne qui représenterait une série de fréquences est parfois une ligne polygonale à dents plus ou moins nombreuses. S'il y a plusieurs dents à peu près de même hauteur, sera-ce la plus haute qui devra être regardée comme normale, bien que peut-être elle ne soit la plus haute que par accident?

On convient donc de représenter la série des observations par une courbe continue à un seul sommet, sans écarter aucune observation supposée bien faite, conformément à la règle que s'imposent les physiciens. On admet alors que la position de ce sommet correspond à la valeur normale; ainsi cette valeur dépend de la nature de la courbe choisie et de la méthode d'ajustement.

Elle est différente, par exemple, suivant que l'on emploie comme courbe d'ajustement une courbe binomiale ou l'une des courbes de Pearson (2). Ce dernier auteur, qui a montré par de nombreux exemples combien sont fréquentes dans la nature, et dans la société humaine, les distributions dissymétriques, a indiqué une méthode conventionnelle uniforme propre à faire connaître, par l'application de règles fixes, la valeur normale. Malheureusement, sa méthode d'ajustement à l'aide du calcul des moments de plusieurs ordres, excellente dans une foule de cas, a l'inconvénient de donner une importance excessive aux observations extrêmes. Par exemple, en appliquant son critérium à l'observation détaillée des revenus ou des salaires (3), le calcul conduit à une valeur normale tendant vers zéro, alors que les statistiques démontrent que cette valeur est loin d'être nulle et va en croissant.

---

(1) S'il n'était pas sans grande utilité de créer un néologisme, l'expression valeur « pléistique » serait plus exacte.

(2) *Philos. transactions* Vol. 186 A (1895), p. 361.

(3) Voir *Journal de la Société de statistique de Paris;* juillet 1898, p. 243; avril 1902, p. 154; août 1902, p. 263.

Pour obtenir une meilleure approximation de cette valeur, il faut renoncer au critérium fondamental, choisir une autre courbe et alors on retombe dans l'arbitraire.

Il en résulte que, jusqu'à présent, la valeur normale n'est pas fixée par une règle uniforme comme la moyenne arithmétique ; elle sera d'ailleurs toujours d'un calcul beaucoup moins simple. La moyenne arithmétique doit donc être préférée comme caractéristique uniforme du caractère de l'ensemble statistique.

Il y aurait quelque présomption à proposer une modification du langage ordinaire qui, souvent à juste titre, ne se pique pas de précision. Mais on peut demander que cette précision s'impose dans les travaux et publications statistiques. Dans ces travaux le mot moyenne devrait toujours avoir le sens de moyenne arithmétique ; on désignerait sous le nom de valeur normale, la valeur la plus fréquente, et, quand il s'agirait de prix ou de cours, on substituerait aux expressions inexactes de prix moyen, cours moyen, des expressions plus correctes telles que prix courant, cours arbitré. On éviterait de la sorte beaucoup de méprises auxquelles donnent lieu les discussions relatives aux moyennes.

Quételet, et beaucoup d'auteurs l'ont suivi, attachait une grande signification au cas particulier dans lequel la moyenne et la normale sont confondues. Pour lui il n'y a de moyenne véritable que dans ce cas. Il admettait que les erreurs des mesures physiques se distribuent symétriquement et, à son avis, la nature dans ses créations typiques opère suivant cette loi d'erreurs, vise un but dont elle ne s'écarte qu'accidentellement et indifféremment dans un sens ou dans l'autre.

Les mesures physiques ne suivent pas toujours la loi symétrique des erreurs ; Bravais en a signalé des exemples à Quételet lui-même (1). Mais surtout rien n'autorise à attribuer à la nature, par une sorte d'anthropomorphisme peu justifié, la tendance à ne s'écarter du type que suivant une loi uniforme. Les observations météorologiques, biologiques, sociales, aujourd'hui fort nombreuses, démontrent que dans la nature, les types, au sens où l'entendait Quételet, sont rares : ce sont presque des accidents.

Cependant, les observations typiques, ou à peu près typiques, sont assurément beaucoup mieux synthétisées que les autres par le calcul de la moyenne ; elles satisfont mieux l'esprit parce qu'elles éveillent l'idée d'une tendance commune. D'autre part, l'hypothèse que ces observations se conforment approximativement à la loi d'erreurs dispense à peu près de s'inquiéter de la répartition des faits autour de la moyenne. Mais il n'y a là aucune raison de contester la valeur comparative de la moyenne arithmétique, on doit seulement reconnaître que cette valeur est mieux représentative des faits quand la moyenne et la normale sont confondues ou très voisines, quand la moyenne arithmétique est une *moyenne normale.*

Arrivons à un dernier abus du mot moyenne. Nous avons vu que pour caractériser la valeur moyenne d'une série de prix on se contentait souvent de former la demi-somme des termes extrêmes.

En réalité on n'obtient de la sorte qu'un indice commode de la véritable valeur de la moyenne.

Il en est encore ainsi quand, au lieu de prendre la demi-somme des prix extrêmes, on divise la somme des prix par le nombre de ces prix.

---

(1) On sait que les épreuves dites au hasard ne suivent pas toujours la loi des probabilités. D'après Pearson, la roulette de Monaco donne des séries cahotiques (Pearson, *The chances of death*, t. I, p. 56).

De même qu'en physique le terme vitesse moyenne a un sens différent de celui de moyenne des vitesses, il ne faut pas confondre la moyenne des prix et le prix moyen. La première n'est qu'un indice, ce qui ne lui enlève pas d'ailleurs sa valeur comparative : outre que souvent l'indice est très voisin de la moyenne et peut la remplacer pratiquement, l'indice a parfois une valeur comparative spéciale d'un réel intérêt.

Par exemple si, à l'aide de tables de mortalité successives, on veut comparer l'état de la mortalité des adultes de vingt à quarante ans au moyen d'un coefficient synthétique, on peut adopter comme élément de comparaison le rapport du nombre des décès survenus entre vingt et quarante ans au nombre des individus de la génération de vingt ans. Mais la comparaison se fera sous un autre aspect, intéressant à d'autres égards, si l'on adopte comme élément de comparaison la moyenne des taux annuels entre vingt et quarante ans. Il en est de même des indices de comparaison des prix de diverses marchandises : suivant le but de ces comparaisons l'indice de la moyenne peut être mieux approprié au but que la moyenne proprement dite qui tient compte des quantités vendues. Des indices de ce genre sont donc fort utiles ; il importe seulement d'observer que rien n'autorise à les confondre avec les moyennes proprement dites.

*Variabilité comparative.* — Pour comparer des faits variables on fait souvent usage de représentations graphiques. Il ne suffit pas, en effet, de rapprocher les moyennes et les écarts quadratiques.

On évite toute fausse apparence de l'allure des mouvements comparés en se servant d'échelles logarithmiques. On obtient d'ailleurs le même avantage plus simplement et moyennant une représentation plus exacte de ces mouvements, si, pour construire chaque courbe, on se borne à substituer aux nombres observés leurs rapports à la moyenne d'une certaine série de ces nombres.

La juxtaposition des courbes permet de se rendre compte de l'accord ou du désaccord des mouvements ; la comparaison reste néanmoins un peu vague et imprécise. Pour lui donner la valeur d'une mesure on calcule des coefficients moyens qui synthétisent l'accord des variations observées, en tenant compte ou non de l'importance de ces variations.

Un type de coefficient de ce genre a reçu de Fechner le nom de coefficient de dépendance. Un autre a été appelé par Pearson coefficient de corrélation.

Ces expressions ont l'inconvénient de laisser supposer que la grandeur du coefficient mesure effectivement la dépendance ou la relation plus ou moins étroite qui existe entre les faits comparés. Or, il n'en est point ainsi. Le coefficient révèle simplement la concomitance des variations, c'est pourquoi il semble plus correct de ne l'employer en statistique que comme coefficient *de covariation.*

En résumé, les statistiques sociales s'enrichissent de documents de plus en plus nombreux, applicables à des catégories de population plus étendues et à de longues séries d'années.

Des progrès analogues dans le domaine des observations biologiques font que, sans parler des applications possibles dans la physique proprement dite, l'intervention de la méthode statistique s'étend à des documents plus nombreux et plus soigneusement recueillis.

Les applications des mathématiques à l'analyse de ces documents peuvent apporter

l'ordre et la précision nécessaires et aider à orienter les investigations des statisticiens. Peut-être les théories auraient-elles besoin de quelques développements pour mieux s'adapter aux réalités, notamment en ce qui concerne les cas de probabilité variable, l'étude des ensembles concrets et limités, l'interpolation, etc.

Mais il importerait que, dans ces applications, dont les conclusions sont destinées au public, les modes de comparaison et la terminologie fussent uniformes et débarrassés aussi bien que possible des risques d'ambiguïté.

D'après ce qui précède, les précautions suivantes semblent devoir être recommandées :

1° Éviter le terme probabilité pour caractériser l'attente à laquelle donne lieu l'observation d'une fréquence. Indiquer, quand on le peut, les limites conventionnelles entre lesquelles il est légitime d'admettre que cette fréquence peut varier fortuitement ;

2° Réserver le terme « valeur moyenne » à la moyenne arithmétique ; on désignerait sous les noms de valeur normale, prix courant, cours arbitré, indice de la moyenne, les rapports exprimés dans le langage courant sous les noms de : valeur moyenne, prix moyen, cours moyen ;

3° Dans les comparaisons de variations, si l'on a recours à la méthode graphique, choisir les unités de façon à assurer l'uniformité de la représentation ; si l'on calcule un coefficient synthétique de l'accord ou du désaccord des variations, éviter que son expression paraisse préjuger l'existence de liaisons entre les faits ; le terme coefficient de covariation, par exemple, n'implique que l'accord des variations ; il semble devoir être préféré aux expressions qui supposent implicitement la dépendance ou la corrélation des faits.

Lucien MARCH.