JOURNAL DE LA SOCIÉTÉ STATISTIQUE DE PARIS

VILFREDO PARETO

Quelques exemples d'application des méthodes d'interpolation à la statistique

Journal de la société statistique de Paris, tome 38 (1897), p. 367-379

<http://www.numdam.org/item?id=JSFS_1897__38__367_0>

III.

QUELQUES EXEMPLES D'APPLICATION DES MÉTHODES D'INTERPOLATION A LA STATISTIQUE.

I.

Les chiffres que nous fournit la statistique deviennent d'année en année plus nombreux et permettent ainsi l'application des méthodes d'interpolation, pour les représenter dans leur ensemble.

Tant que ces chiffres sont en petit nombre, on est obligé, soit de les comparer directement, soit de se borner à en tirer certaines moyennes. Mais, quand on dispose d'un grand nombre de chiffres, on peut tracer une courbe qui représente le phénomène et en fait connaître les lois.

Cette courbe est, en général, très compliquée; la forme générale du phénomène se perd dans les détails; il faut l'en dégager, c'est-à-dire substituer une courbe plus simple à celle qui résulte directement de l'observation.

Tel est le but de l'interpolation. En réalité, on a toujours employé, parfois sans s'en rendre compte, une interpolation plus ou moins grossière, pour représenter les résultats de la statistique. Ainsi, quand on dit que le chiffre des naissances va en croissant, on substitue simplement à la courbe réelle, avec toutes ses sinuosités, une autre courbe beaucoup plus simple, dont on ne retient que le caractère d'aller en s'élevant au-dessus de l'axe des temps. Il s'agit de préciser et de compléter ces conceptions, en grande partie intuitives.

Le problème de l'interpolation est un problème essentiellement arbitraire. Objectivement, la courbe des naissances, par exemple, est ce qu'elle est; seule, l'imperfection de nos facultés mentales nous oblige à la simplifier pour en tirer des lois générales. Cette simplification peut se faire d'une infinité de manières, suivant le but que nous nous proposons d'atteindre.

La même chose peut s'exprimer d'une manière différente. Le mouvement, représenté par une courbe du genre de celles dont nous nous occupons, peut en général se décomposer en plusieurs autres. Il y a, par exemple, un mouvement général qui éloigne la courbe de l'axe des temps, d'autres qui produisent des sinuosités assez longues, d'autres, des sinuosités plus courtes, et, enfin, jusqu'à d'insignifiantes irrégularités.

Il y a deux cas limites. Dans l'un, les sinuosités intermédiaires disparaissent; il ne reste plus qu'une partie constante, qui est la partie de beaucoup la plus impor-

tante du phénomène, et de petites irrégularités. Ce cas est celui que considère le calcul dit des « erreurs ». Les petites irrégularités portent ce nom ou bien celui d' « écarts », et on se propose de dégager la partie constante de ces « erreurs » ou de ces « écarts ». On admet, pour cela, une loi de ces écarts, laquelle n'est au fond qu'une formule d'approximation. L'expérience la vérifie en général, de même qu'elle vérifie d'autres formules d'approximation, applicables en d'autres circonstances; telles, par exemple, que la formule de Taylor.

L'autre cas limite est celui où les sinuosités de la courbe passent par degrés insensibles, des plus grandes aux plus petites. On dit alors que le phénomène est tellement irrégulier qu'il est impossible d'en dégager aucune loi générale.

Les cas que nous avons à considérer sont des cas intermédiaires dans lesquels il existe un groupe de grandes sinuosités et d'autres groupes de sinuosités plus petites; et le but de nos recherches est de séparer ce premier groupe des autres.

Entre deux courbes également simples, il est clair qu'il faut préférer celle qui se rapproche le plus de la courbe réelle. Mais que doit-on entendre par ces mots : se rapprocher ? Dans le cas limite du calcul des erreurs, on démontre qu'il convient de rendre un minimum la somme des carrés des écarts $\varepsilon_1, \varepsilon_2, \ldots$, et la *précision* du système est mesurée par une certaine constante k, déterminée par l'équation

$$\frac{1}{2k^2} = \frac{\varepsilon_1^2 + \varepsilon_2^2 + \ldots\ldots}{n},$$

ou mieux, par l'équation

$$\frac{1}{2k^2} = \frac{\varepsilon_1^2 + \varepsilon_2^2 + \ldots\ldots}{n-1}.$$

Mais, comme le dit fort bien M. Bertrand (*Calcul des prob.*, p. 210) : « Si la loi de probabilité n'avait pas une forme toute spéciale, les mots *poids* et *précision*, dont les physiciens font souvent usage, ne pourraient pas avoir de sens exact et précis. »

C'est précisément le cas dans lequel nous nous trouvons. Il n'y a aucun motif péremptoire d'employer la méthode des moindres carrés. Souvent même, au lieu de rendre un minimum la somme des carrés des écarts, il conviendrait de rendre le plus grand écart positif égal numériquement au plus grand écart négatif. Malheureusement, ce système donnerait lieu à des calculs fort longs et compliqués. D'autre part, la méthode des moindres carrés présente l'avantage d'être celle qui est la plus convenable pour le cas limite du calcul des erreurs; il peut donc convenir de l'employer pour d'autres cas, qui peuvent, d'ailleurs, se rapprocher insensiblement du cas limite. De même, nous continuerons à prendre comme indice du rapprochement des courbes la quantité k, et cela simplement parce qu'à la limite elle se confond avec la *précision* considérée dans le calcul des erreurs. Du reste, si l'on employait une courbe donnant une même valeur numérique au plus grand écart positif et au plus grand écart négatif, on pourrait, par cette valeur, juger du rapprochement des courbes. Une infinité d'autres systèmes semblables peuvent être employés et sont tout aussi plausibles. Suivant le genre de questions que l'on traite, certains systèmes sont à préférer à certains autres, mais ce sont là des considérations dans lesquelles nous nous abstiendrons d'entrer pour le moment.

Les calculs qu'entraîne la méthode des moindres carrés sont longs et pénibles lorsque les valeurs de la variable sont quelconques, mais ils deviennent relative-

ment faciles quand les valeurs de la variable forment une progression arithmétique; et ce cas se présente fort souvent pour les chiffres de la statistique. En outre, on les facilite beaucoup en faisant usage de tables de multiplication et de plusieurs tables de logarithmes, les unes à sept, d'autres à cinq, et même à quatre décimales. Il ne faut pas poursuivre une précision illusoire, et calculer laborieusement un grand nombre de décimales, qui, au fond, ne signifient rien du tout.

Pour ne pas nous perdre en des généralités trop abstraites, considérons un exemple, celui de la population d'un pays. La statistique nous fournit les chiffres qui représentent la population à des intervalles égaux : 1, 2, 3... n, qui seront les valeurs successives que prend la variable x. Soient : y_1, y_2, ... y_n les chiffres de la population qui correspondent à ces intervalles égaux, et, en général, y correspondra à x. Prenons la formule donnée par M. Tchébychef pour l'application de la méthode des moindres carrés. Nous écrirons cette formule sous la forme

$$(1) \qquad y = A_0 + A_1\Psi_1 + A_2\Psi_2 + A_3\Psi_3 + \ldots\ldots$$

On aura :

$$z = x - \frac{n+1}{2}, \qquad \Psi_1 = z, \qquad \Psi_2 = z^2 - \frac{n^2-1}{12}, \qquad \Psi_3 = z^3 - \frac{3n^2-7}{20}z, \text{ etc.}$$

$$A_0 = \frac{\Sigma y}{n}, \qquad A_1 = \frac{\Sigma y \Psi_1}{\Sigma \Psi_1^2}, \qquad A_2 = \frac{\Sigma y \Psi_2}{\Sigma \Psi_2^2}, \ldots\ldots$$

Dans la formule (1), chaque terme sert de terme de correction aux précédents. Le premier terme nous donne la moyenne des chiffres observés : y_1, y_2, ... y_n. Le second terme corrige le premier, en substituant une droite plus ou moins inclinée sur l'axe des x, à la droite parallèle à cet axe donnée par la moyenne. Le troisième terme donne une parabole, etc.

Cette formule suffisamment prolongée reproduit exactement tous les chiffres y_1, y_2, ... y_n, correspondant aux valeurs 1, 2, ... n de la variable; mais alors, elle ne nous apprend plus rien sur le phénomène que nous étudions. Réduite à un nombre de termes inférieur à celui qui donne exactement y_1, y_2, ... y_n, elle substitue une courbe plus simple à la courbe compliquée que fournit directement l'observation.

Lorsqu'on applique cette formule aux chiffres que donne la statistique, on observe, en général, que les courbes simples qu'on obtient successivement ne vont pas en se rapprochant d'une manière uniforme de la courbe réelle, la *précision* commence d'abord par augmenter rapidement; ensuite, il y a une période où elle augmente lentement, de nouveau elle augmente rapidement, et ainsi de suite. Ces périodes pendant lesquelles la précision augmente lentement séparent les grands groupes des sinuosités dont nous avons parlé; en d'autres termes, elles séparent des groupes d'influences de plus en plus particulières, qui s'exercent sur le phénomène.

Pour expliquer plus clairement la chose, sans entrer dans de trop longs développements, considérons un cas hypothétique. Nous avons, par exemple,

$$y_1 = 1, \qquad y_2 = 3{,}7, \qquad y_3 = 9{,}4, \qquad y_4 = 15{,}5, \qquad y_5 = 25{,}6$$

c'est-à-dire que le phénomène est représenté par une parabole de second degré,

sauf de petites irrégularités. Ici, les groupes se réduisent donc à deux : d'abord une influence générale, qui donne la forme parabolique, ensuite les irrégularités. Dans ce cas fort simple, on s'en aperçoit, à première vue, par l'inspection des chiffres. Nous allons voir que l'application de la formule conduit exactement au même résultat.

Nous indiquerons, en général, par Δ_0 les écarts qu'on obtient en conservant seulement le premier terme de la formule d'interpolation; par Δ_1, ceux qui s'observent quand on conserve deux termes, etc. De même, k_0 sera la *précision* qui correspond à Δ_0; k_1 celle qui correspond à Δ_1, etc. Rappelons, enfin, que lorsque les courbes coïncident, la précision est infinie.

Dans le cas que nous considérons, nous aurons le tableau suivant des écarts :

x	Δ_0	Δ_1	Δ_2	Δ_3	Δ_4
1	— 10,04	+ 2,16	— 0,01143	+ 0,08857	0
2	— 7,34	— 1,24	— 0,15429	— 0,35429	0
3	— 1,64	— 1,64	+ 0,53143	+ 0,53143	0
4	+ 4,46	— 1,64	— 0,55428	— 0,35429	0
5	+ 14,56	+ 2,36	+ 0,18857	+ 0,08857	0

La simple inspection de ce tableau fait voir que les écarts diminuent considérablement jusqu'à Δ_2, ensuite, de Δ_2 à Δ_3, la diminution n'est guère sensible, on ne saurait même dire si elle existe; mais, de nouveau, quand nous passons de Δ_3 à Δ_4, une diminution considérable a lieu. C'est ce qu'indiquera, sous une autre forme, le tableau des *précisions*.

k_0	k_1	k_2	k_3	k_4
0,072	0,341	1,75	1,91	∞

II.

Un grand nombre d'auteurs, lorsqu'ils veulent se rendre compte de l'augmentation de la population, *supposent* soit un accroissement en progression arithmétique, soit un accroissement en progression géométrique, et calculent, en ces cas, les raisons des progressions. Pourquoi cela ? Pourquoi faire des hypothèses, quand nous pouvons interroger les faits et apprendre d'eux quelle est, en réalité, la loi suivant laquelle s'est accrue la population ? Tant que les faits connus ne sont pas assez nombreux pour nous donner ces indications, il est utile de tâcher d'y suppléer au moyen d'hypothèses plus ou moins plausibles; mais à peine les chiffres fournis par la statistique sont assez nombreux, il faut abandonner les hypothèses et étudier la réalité.

Commençons par considérer le mouvement général de la population en Angleterre et Galles depuis 1801, et demandons-nous par quelles courbes simples nous pouvons le représenter, et si l'accroissement de la population se rapproche plutôt d'une progression arithmétique que d'une progression géométrique, ou *vice versâ*. Pour cela, nous interpolerons les chiffres qui donnent la population et les logarithmes de ces chiffres. Nous prenons comme unité 1 000. Les écarts sur les chiffres de la population, directement interpolés, seront indiqués par Δ, les écarts, tou-

jours sur les chiffres de la population, quand l'interpolation de ces chiffres se fait par les logarithmes, seront indiqués par Δ′. Nous aurons le tableau suivant.

Angleterre et Galles.

Années	Population	Écarts		
		Progression arithmétique Δ	Progression géométrique	
			Chiffres bruts Δ′	Chiffres rectifiés Δ″
1801. . .	8 893	— 1 191	— 245	— 449
1811. . .	10 164	— 1 074	— 254	— 445
1821. . .	12 000	+ 452	+ 124	— 47
1831. . .	13 897	+ 1 135	+ 358	+ 217
1841. . .	15 914	+ 1 331	+ 483	+ 380
1851. . .	17 928	+ 816	+ 334	+ 289
1861. . .	20 066	+ 19	+ 8	+ 36
1871. . .	22 712	— 292	— 153	— 33
1881. . .	25 974	— 154	— 92	+ 148
1891. . .	29 003	— 1 054	— 712	— 318

La simple inspection de ces chiffres fait voir que le phénomène réel est bien mieux représenté par la progression géométrique que par la progression arithmétique.

Si l'on pose :

$$y = A\,10^{\alpha z},$$

on a :

$$\log A = 1{,}21692; \qquad \alpha = 0{,}05690.$$

Dans mon *Cours d'économie politique,* I, p. 111, on trouve la valeur :

$$\alpha = 0{,}05637.$$

La différence provient de ce que cette dernière valeur a été obtenue avec la méthode d'interpolation de Cauchy.

Les valeurs qu'on vient de trouver pour A et α ne sont pas les plus favorables. En effet, rendre minima la somme des carrés des écarts des logarithmes n'est pas du tout la même chose que rendre minima la somme des carrés des écarts des nombres de ces logarithmes. On peut vérifier qu'en ce cas particulier la méthode de Cauchy appliquée aux logarithmes donne, pour les nombres, de moindres écarts que la méthode des moindres carrés, appliquée aux logarithmes.

Pour trouver les valeurs plus convenables de A et de α, la méthode classique consiste, en général, à partir de valeurs approchées et à en déduire les valeurs exactes par des applications répétées des formules suivantes :

$$\Sigma\left(y - A\,10^{\alpha z}\right) 10^{\alpha z} = \Delta A \Sigma\, 10^{2\alpha z} + \Delta\alpha \Sigma \frac{A z}{M} 10^{2\alpha z},$$

$$\Sigma\left(y - A\,10^{\alpha z}\right) \frac{A z}{M} 10^{\alpha z} = \Delta A \Sigma \frac{A z}{M} 10^{2\alpha z} + \Delta\alpha \Sigma \left(\frac{A z}{M}\right)^2 10^{2\alpha z};$$

où $M = 0{,}43429\ldots$ est le module des logarithmes naturels.

Dans le calcul des erreurs, les quantités ΔA et $\Delta \alpha$, etc., sont souvent fort petites, et alors les formules précédentes sont assez utiles; mais dans les calculs qui nous occupent, les quantités ΔA, etc., peuvent avoir des valeurs considérables, et alors le calcul de ces formules devient assez long et pénible. Si l'on voulait continuer à suivre cette voie, il faudrait conserver d'autres termes des développements en série, dont ces formules ne donnent que les premiers termes. Mais dans le cas particulier dont nous traitons maintenant, on peut trouver une méthode bien plus satisfaisante pour la pratique. Il faut simplement donner à chaque équation du type

$$\log y_1 = \log A + \alpha z_1$$

un *poids* égal au nombre du $\log y_1$; c'est-à-dire multiplier ces équations par $y_1, \ldots$ ce qui donne des équations du type :

$$y_1 \log y_1 = y_1 \log A + \alpha z_1 y_1.$$

C'est à ces équations qu'on applique directement la méthode des moindres carrés.

On trouve ainsi :

$$\log A = 1{,}21887, \qquad \alpha = 0{,}05520,$$

et les écarts prennent les valeurs indiquées dans la colonne Δ'' du tableau précédent.

Nous avons ainsi une représentation générale du mouvement de la population de l'Angleterre, mais l'examen d'un tracé graphique fait voir que nous négligeons des détails fort intéressants. Ainsi, la courbe qui représente les chiffres de la population, de 1855 à 1894, présente vers 1880, presque brusquement, un changement très marqué de forme. Pour mieux nous en rendre compte, nous allons calculer séparément les courbes de ces deux périodes.

Les logarithmes présentent ici des différences bien moindres que dans le cas précédent, et il suffira d'employer simplement la méthode de Cauchy pour les interpoler.

Angleterre et Galles.

Années	Population	Δ'	Années	Population	Δ'
1855. . .	18 829	+ 101	**1880**. . .	25 714	+ 8
1860. . .	19 903	— 34	**1883**. . .	26 627	+ 2
1865. . .	21 145	— 78	**1886**. . .	27 523	— 11
1870. . .	22 501	— 91	**1889**. . .	28 448	— 27
1875. . .	24 045	— 6	**1892**. . .	29 402	+ 16
1880. . .	25 714	+ 111	**1895**. . .	30 383	+ 13

De 1855 à 1880, l'origine des z correspond à 1867, 5, et l'on a :

$$\log A = 1{,}34039, \qquad \alpha = 0{,}027160.$$

De 1880 à 1895, l'origine des z correspond à 1887, 5, et l'on a :

$$\log A = 1{,}44671, \qquad \alpha = 0{,}028860.$$

Pour comparer les valeurs que nous venons d'obtenir pour α, il faut tenir compte que de 1801 à 1891 l'unité est de 10; qu'elle est de 5, de 1855 à 1880; et de 3, de 1880 à 1895. En réduisant uniformément l'unité à une année, nous obtenons pour α les valeurs suivantes :

1801 a 1891	1855 a 1880	1880 à 1895
0,005520	0,005432	0,004810

On voit maintenant, d'une manière très claire, que l'accroissement de la population n'est plus aussi rapide que par le passé.

La courbe que nous venons d'obtenir pour les chiffres de la population en Angleterre ne représente le phénomène que d'une manière très générale. Pour en connaître les détails, il faut pousser plus loin l'approximation. Nous allons faire cela pour la population en Angleterre et Galles, de 1855 à 1895. Nous prendrons les chiffres de cinq en cinq années. Il vaudrait mieux les prendre pour toutes les années, bien que, d'autre part, les chiffres qui ne correspondent pas aux recensements soient assez incertains; mais le temps nous manque pour exécuter de trop longs calculs.

En poussant l'approximation jusqu'aux termes en Ψ_6, nous avons :

$$(2)\quad y = 24278{,}33 + 1469{,}4\,z + 33{,}6818\,\Psi_2 - 6{,}88215\,\Psi_3 + 0{,}48951\,\Psi_4 + 0{,}630128\,\Psi_5 + 0{,}081111\,\Psi_6.$$

Et nous formons le tableau suivant :

Angleterre et Galles.

Années	Population	Écarts Δ_1	Δ_2	Δ_3	Δ_4	Δ_5	Δ_6
1855	18 829	+ 428,27	+ 113,90	— 1,72	— 13,17	+ 3,34	+ 1,57
1860	19 903	+ 32,87	— 45,72	+ 12,08	+ 29,71	— 16,50	— 8,98
1865	21 145	— 194,53	— 104,71	+ 2,65	+ 11,88	+ 28,68	+ 18,95
1870	22 501	— 307,93	— 117,07	— 42,74	— 50,29	— 12,49	— 12,93
1875	24 045	— 233,33	— 8,79	— 8,79	— 23,89	— 23,89	— 15,01
1880	25 714	— 33,73	+ 157,13	+ 82,80	+ 75,25	+ 37,41	+ 37,00
1885	27 221	+ 3,87	+ 93,68	— 13,68	— 4,45	— 21,25	— 30,98
1890	28 764	+ 77,47	— 1,12	— 58,93	— 41,31	+ 4,90	+ 12,42
1895	30 383	+ 227,07	— 87,30	+ 28,32	+ 16,57	— 0,23	— 2,00
Indice de précision k		0,0030	0,0070	0,0173	0,0183	0,0328	0,0347
Somme des carrés des écarts. .		4 303	809	134	120	37	33

L'unité pour la somme des carrés des écarts est 100.

On voit que les indices de précision croissent rapidement jusqu'à celui qui correspond à Δ_3; ensuite, ils croissent beaucoup plus lentement. Dans le cas que nous examinons, on trouve donc que sur la population agit un premier groupe de forces qui donnent au phénomène la forme indiquée par les quatre premiers termes de la formule (2); les autres termes représentent des « perturbations », des « irrégularités ». Dans l'état actuel de nos connaissances nous ignorons même si ces perturbations sont celles qui se présentent, en réalité, ou si elles n'ont pas été introduites, au moins en partie, par les méthodes employées pour évaluer le chiffre de la population.

La formule (2), réduite à ses quatre premiers termes, c'est-à-dire :

$$(3) \qquad y = 24278{,}33 + 1469{,}4\,z + 33{,}6818\,\Psi_2 - 6{,}88215\,\Psi_3,$$

représente donc tout ce que nous pouvons actuellement connaître sur la forme générale du phénomène dans le cas considéré.

Le dernier terme de cette formule est négatif. Il représente l'influence qui s'est fait récemment sentir pour diminuer l'accroissement de la population en Angleterre.

On prétend souvent pouvoir calculer ce que sera le chiffre de la population à l'avenir, en se réglant sur l'accroissement qu'elle a eu par le passé. Cela revient à étendre l'usage de formules analogues à la formule (3), hors des limites pour lesquelles ces formules ont été calculées. S'il ne s'agit que d'un laps de temps fort court, la chose peut se faire sans danger de trop grandes erreurs, mais s'il s'agit d'un laps de temps assez long, par exemple un siècle, on n'arrive qu'à des résultats absurdes. Ainsi, par exemple, si l'on voulait connaître la population de l'Angleterre en l'an 2000, en se réglant sur l'accroissement qu'elle a eu de 1855 à 1895, il faudrait adopter la formule (3), qui représente ce que nous savons de plus certain sur la forme générale de la courbe. Or, cette formule donne un chiffre négatif pour la population en l'an 2000! Si l'on n'est pas satisfait de ce résultat, l'on n'a qu'à supprimer le terme en Ψ_3; alors on trouvera que la population de l'Angleterre en 2000 sera, à peu près, de 82 millions. Si l'on n'est pas encore content, et si, comme il arrive généralement quand on se livre à ces beaux calculs, on n'éprouve aucune difficulté à nourrir cette multitude, on peut prendre la formule qui interpole les logarithmes, et l'on aura un chiffre bien plus considérable pour la population.

On peut ainsi démontrer tout ce que l'on veut; mais une seule chose est certaine, c'est que la réalité sera différente de ces chiffres: ils peuvent bien nous faire connaître ce qui ne sera pas, ils ne peuvent pas nous faire connaître ce qui sera. On peut voir la chose encore plus clairement en examinant le mouvement de la population en Prusse depuis 1816; mais nous ne développerons pas ici ces calculs.

On croit parfois détourner ces difficultés en disant : *supposons* que la population s'accroisse en progression géométrique ou en progression arithmétique. On paraît ainsi croire qu'il n'y a que ces deux formes de courbes pour représenter les chiffres de la population, et que la réalité doit se trouver entre les limites que donnent ces deux hypothèses. Rien n'est plus faux; ces hypothèses sont absolument arbitraires, et les chiffres qu'on prétend en déduire pour la population que pourra avoir un pays dans un ou deux siècles ont autant de rapport avec la réalité que des chiffres qui auraient été écrits au hasard.

III.

Les méthodes d'interpolation peuvent être employées non seulement pour représenter un phénomène, mais encore pour rechercher les rapports qu'ont des phénomènes entre eux. Il est des auteurs qui, actuellement, ont une tendance à n'admettre de corrélation entre les chiffres de la statistique, que si cette corrélation est exprimée par une formule du genre de :

$$(4) \qquad y = e^{-\alpha x^2}$$

C'est ainsi qu'on a fait à notre formule pour représenter la répartition des revenus une objection fort singulière. On a dit qu'il *devait* y avoir une formule du genre de la formule (4) qui représenterait mieux les faits. Cela est fort possible, mais il y a une seule manière de le prouver, c'est d'indiquer quelle est cette formule, et c'est précisément la seule chose qu'on s'est bien gardé de faire. Si ces auteurs avaient vécu au temps de Kepler, ils auraient dit qu'il y avait des combinaisons d'épicycles qui représentaient les orbites planétaires mieux que l'ellipse. La chose est vraie en partie, car avec des épicycles on peut reproduire une courbe à peu près quelconque.

La formule (4) ne représente même pas toute la courbe des probabilités. Poisson (*Rech. sur la prob. des jug.*, p. 180) n'a pas manqué d'insister sur le fait, qu'à une certaine distance du maximum, une autre formule d'approximation pourrait donner une valeur qui ne coïnciderait pas avec celle qui se déduit de la formule (4), « de telle sorte que le rapport de l'une de ces valeurs approchées à l'autre pourrait différer beaucoup de l'unité ».

Considérons la formule binormale :

$$y = \frac{m(m-1)\ldots\ldots(m-n+1)}{1.2\ldots\ldots n} p^m q^n ;$$

$$p + q = 1, \qquad m + n = \mu ;$$

et posons :

$$m' = p\mu, \qquad m = m' + t,$$

$$n' = q\mu, \qquad n = n' - t,$$

$$P_0 = \frac{1.2\ldots\ldots\mu}{1.2\ldots\ldots m'.1.2\ldots\ldots n'} \left(\frac{m'}{\mu}\right)^{m'} \left(\frac{n'}{\mu}\right) ;$$

nous aurons, en substituant, comme d'habitude, aux factorielles leurs valeurs approchées :

$$y = P_0 \left(\frac{m'}{m'+t}\right)^{m'+t+\frac{1}{2}} \left(\frac{n'}{n'-t}\right)^{n'-t+\frac{1}{2}}. \tag{5}$$

Pourvu que l'on ait en même temps :

$$\frac{t}{m'} < 1, \qquad \frac{t}{n'} < 1,$$

on peut, après avoir pris les logarithmes de la formule (5), les développer en série, et l'on obtient alors une formule du genre de la formule (4). Si, au contraire, on avait :

$$\frac{t}{m'} > 1, \qquad \frac{t}{n'} < 1,$$

cette formule ne serait plus valable, et devrait être remplacée par une autre. On ne fait pas cette substitution, dans le calcul des erreurs, simplement parce qu'elle ne conduirait qu'à remplacer, dans une intégrale, certains éléments très petits par d'autres également fort petits, ce qui est indifférent. Mais il est d'autres cas où cette

substitution s'impose : par exemple s'il s'agit d'étudier les rapports de ces éléments (1).

Nous croyons, pour notre part, que la statistique doit, autant que possible, s'en tenir aux faits. Une formule d'interpolation est, à notre avis, d'autant meilleure qu'elle subit moins l'influence de conceptions théoriques et d'idées *à priori*. Nous n'entendons certes pas dire par là que le statisticien doit renoncer à user d'un judicieux esprit de critique, et chercher, au hasard, des rapports entre des faits qui manifestement n'en peuvent avoir. Une préparation théorique est toujours nécessaire. Nous voulons seulement dire que quand nous interrogeons les faits, nous devons, autant que possible, nous abstenir de leur dicter la réponse que nous en attendons. Il est bien entendu aussi qu'il est toujours licite de faire telle hypothèse qu'on voudra. Mais il faut ensuite la soumettre, sans aucun parti pris, à l'épreuve des faits.

Considérons, par exemple, le rapport entre le nombre des mariages, en Angleterre, et la prospérité économique du pays. Des considérations, *à priori*, font voir que l'existence d'un rapport entre ces deux phénomènes n'est nullement absurde ni illogique. Mais quel est précisément ce rapport ? Ici, il faut laisser la parole aux faits, et ne pas nous substituer à eux.

Si nous traçons la courbe qui donne le nombre des mariages en Angleterre, depuis 1855, et différentes autres courbes qui indiquent les exportations, les quantités de charbon extrait des mines, etc., nous observons, à première vue, une tendance très marquée des ondulations de ces courbes à devenir parallèles. Voilà un fait brut, qu'il faut analyser.

Les courbes qui représentent les exportations de marchandises, la production des mines de charbon, etc., ne sont évidemment ici que des indices, qui se complètent l'un l'autre, de l'état économique du pays. Considérons-les un après l'autre.

Le parallélisme des courbes indique plutôt un rapport entre les tangentes des courbes qu'entre les ordonnées, *directement*. Soit v le nombre des mariages à l'époque x, et u le chiffre des exportations, si l'on avait l'équation rigoureuse :

$$\frac{dv}{dx} = A\frac{du}{dx}, \tag{6}$$

(1) La courbe que nous avons trouvée pour la repartition des revenus correspond précisément à la portion d'une courbe loin du maximum; il est donc fort naturel qu'elle soit représentée par une autre formule que celle qui est valable pour le maximum des ordonnées de la courbe, qui, ici, est proche du minimum des revenus.

Bien que nous ayons répété à satiété : 1° que cette formule n'est pas valable pour les revenus proches du revenu minimum; 2° qu'en aucun cas ce revenu minimum ne saurait être zéro; car nous considérons comme revenu la somme des biens dont jouit un homme, et cette somme ne peut descendre au-dessous de ce qui est indispensable pour la conservation de la vie, M. le professeur Edgeworth a voulu, avec insistance, juger notre formule par les résultats qu'elle donne pour un revenu zéro!

En raisonnant de la sorte, on jugerait d'un développement en série par les résultats qu'il donne hors de son cercle de convergence. Ainsi, par exemple, on a, pour $x < 1$,

$$\frac{1}{1-x} = 1 + x + x^2 + x^3 + \ldots.$$

et il faudrait rejeter ce développement en série, parce que, si l'on y fait $x = 2$, il donne une somme infinie.

on en déduirait :

$$v = B + Au, \tag{7}$$

B étant une constante. Mais quand l'équation (6) n'est qu'approchée, l'équation (7) peut ne représenter le phénomène que d'une manière fort imparfaite, ou même ne pas le représenter du tout.

Ainsi, par exemple, supposons qu'en réalité A, au lieu d'être une constante, contienne un petit terme périodique. Ce petit terme pourrait être négligeable (6) et donner des termes qui ne sont pas négligeables dans (7). Voyons mieux la chose avec des chiffres. Soit, en réalité,

$$A = 1 + 0{,}06 \cos x, \qquad \frac{du}{dx} = \cos x.$$

L'interpolation nous donne l'équation fort approchée :

$$\frac{dv}{dx} = \cos x = \frac{du}{dx}$$

et pour une valeur de x aussi grande qu'on le désire, l'erreur ne peut dépasser 0,06. Mais si l'on intègre cette équation pour avoir l'équation (7), on aura, en prenant zéro pour la constante :

$$u = \sin x; \tag{8}$$

tandis que l'équation rigoureuse aurait donné :

$$u = 0{,}03\,x + 0{,}015 \sin 2x + \sin x; \tag{9}$$

et l'erreur, pour $x = 100$, atteindra trois unités. La formule (8) n'est pas même une expression approchée de (9).

Il y a donc lieu d'étudier séparément les expressions (6) et (7). Il n'y aurait rien d'impossible, par exemple, à ce que, d'une part, les ondulations de la courbe des mariages fussent *à peu près* parallèles aux ondulations des courbes qui servent d'indices des conditions économiques du pays, et que, d'autre part, le nombre des mariages diminuât *en moyenne* quand la prospérité économique augmente.

Cette communication est déjà trop longue pour que je développe l'étude des deux relations (6) et (7); je me bornerai à considérer la première, et encore ne pourrai-je donner une solution complète, mais seulement indiquer la voie qui pourrait y faire parvenir. Nous emploierons la méthode d'interpolation de Cauchy, fort suffisante pour une étude qui n'est nécessairement que préliminaire.

Dans le tableau suivant (p. 378) :

u indique la valeur des exportations. L'unité est 1 000 livres sterling.

t indique la quantité de houille extraite des usines. L'unité est 100 000 *tons*.

v indique le nombre des mariages. L'unité est 1 000.

Δu, Δv, Δt, indiquent les différences d'une année à une autre.

E_1 indique les écarts, quand on tient compte seulement des exportations.

E_2 indique les écarts, quand on tient compte des exportations et des quantités de charbon.

La formule donnée par l'interpolation est :

$$\Delta v = 1{,}9 + 0{,}3535\,(\Delta u - 3{,}25) + 0{,}05523\,\big(\Delta t - 32{,}05 - 1{,}875\,(\Delta u - 3{,}25)\big).$$

E_1 indique donc les écarts quand cette formule est arrêtée au second terme, et E_2 indique les écarts quand elle est complète.

Années	v	Δv	u	Δu	t	Δt	E_1	E_2
1855. . .	152	+ 7	96	+ 20	615	+ 51	— 0,8	— 0,1
1856. . .	159	0	116	+ 6	666	— 12	— 2,9	— 0,2
1857. . .	159	— 3	122	— 5	654	— 4	— 2,0	— 0,8
1858. . .	156	+ 12	117	+ 13	650	+ 70	+ 6,7	+ 5,6
1859. . .	168	+ 2	130	+ 6	720	+ 80	— 0,9	— 3,2
1860. . .	170	— 6	136	— 11	800	+ 36	— 2,9	— 4,6
1861. . .	164	0	125	— 1	836	— 20	— 0,4	+ 2,0
1862 . .	164	+ 10	124	+ 23	816	+ 47	+ 1,1	+ 2,3
1863 . .	174	+ 6	147	+ 13	863	+ 65	+ 0,7	— 0,2
1864. . .	180	+ 5	160	+ 6	928	+ 54	+ 2,1	+ 1,2
1865. . .	185	+ 3	166	+ 23	982	+ 34	— 5,9	— 3,9
1866 . .	188	— 9	189	— 8	1 016	+ 29	— 6,9	— 7,9
1867. . .	179	— 2	181	— 2	1 045	— 14	— 2,0	0
1868. . .	177	0	179	+ 11	1 031	+ 43	— 4,6	— 4,4
1869. . .	177	+ 5	190	+ 10	1 074	+ 30	+ 0,7	+ 1,5
1870. . .	182	+ 8	200	+ 23	1 104	+ 69	— 0,9	— 0,9
1871. . .	190	+ 11	223	+ 33	1 173	+ 62	— 1,4	0
1872. . .	201	+ 5	256	— 1	1 235	+ 35	+ 4,6	+ 4,0
1873. . .	206	— 4	255	— 15	1 270	— 19	+ 0,6	+ 1,5
1874. . .	202	— 1	240	— 17	1 251	+ 68	+ 4,3	+ 0,2
1875. . .	201	+ 1	223	— 22	1 319	+ 14	+ 8,0	+ 6,4
1876. . .	202	— 8	201	— 2	1 333	+ 13	— 8,4	— 7,5
1877. . .	194	— 4	199	— 6	1 346	— 20	— 2,6	— 0,7
1878. . .	190	— 8	193	— 1	1 326	+ 11	— 8,4	— 7,6
1879. . .	182	+ 10	192	+ 31	1 337	+ 133	— 1,7	— 4,4
1880. . .	192	+ 5	223	+ 11	1 470	+ 72	+ 0,4	— 1,0
1881. . .	197	+ 7	234	+ 8	1 542	+ 23	+ 3,4	+ 4,5
1882. . .	204	+ 2	242	— 2	1 565	+ 72	— 2,0	— 0,8
1883. . .	206	— 2	240	— 7	1 637	— 29	— 0,3	+ 2,0
1884. . .	204	— 6	233	— 20	1 608	— 14	+ 0,3	+ 0,4
1885. . .	198	— 2	213	0	1 594	— 19	— 2,6	+ 0,1
1886. . .	196	+ 5	213	+ 9	1 575	+ 46	+ 1,1	+ 0,9
1887. . .	201	+ 3	222	+ 12	1 621	+ 78	— 2,0	— 3,6
1888. . .	204	+ 10	234	+ 15	1 699	+ 70	+ 3,9	+ 3,1
1889. . .	214	+ 9	249	+ 14	1 769	+ 47	+ 3,3	+ 3,6
1890. . .	223	+ 3	263	— 16	1 816	+ 39	+ 7,9	+ 5,5
1891. . .	226	+ 1	247	— 20	1 855	— 37	+ 7,3	+ 8,7
1892. . .	227	— 9	227	— 9	1 818	— 175	— 6,6	+ 3,6
1893. . .	218	+ 8	218	— 2	1 643	+ 240	+ 8,0	— 4,1
1894. . .	226	+ 2	216	+ 10	1 883	+ 14	— 2,3	— 0,6
1895. . .	228	»	226	»	1 897	»	»	»

Sur 40 valeurs de Δv, il y en a 31 dont le signe est le même que celui de Δu, et seulement 9 dont le signe ne coïncide pas avec celui de Δu. Nous comptons parmi ces cas de non-coincidence les valeurs zéro. Il est donc clair qu'il y a des causes qui font croître et décroître ensemble les mariages et les exportations. C'est là le résultat le plus sûr auquel nous arrivons. Quant à préciser les rapports de ces variations, la formule (10), réduite à ses deux premiers termes, ne donne pas une approximation suffisante; les écarts sont trop considérables. Mais ces écarts diminuent si, outre les chiffres de l'exportation, on considère aussi ceux de la produc-

tion de la houille. La somme des carrés des écarts, qui était de 722, devient 566. Ces valeurs ne sont que grossièrement approchées, pour en avoir de plus exactes, il faudrait calculer les écarts avec plus de décimales que nous ne l'avons fait. Mais ces valeurs, telles qu'elles sont, remplissent parfaitement le but dans lequel elles ont été calculées, et qui est de faire voir la diminution de la somme des carrés des écarts. Il est permis de croire que cette somme continuerait à diminuer si nous tenions compte d'autres indices de l'état économique du pays. Il faut observer que les deux indices que nous avons considérés se rapportent à l'industrie, il faudrait en avoir aussi pour l'agriculture. Si l'on avait des chiffres tant soit peu plausibles pour la valeur des récoltes, depuis 1855, ce serait là un fort bon indice à considérer. Nous n'avons pas pu nous procurer ces chiffres.

Les calculs nécessaires pour élucider complètement cette question sont fort longs. Ils dépassent probablement les forces d'un calculateur isolé. Il serait fort désirable de les voir entreprendre par quelque bureau de statistique.

Vilfredo PARETO.