JOURNAL DE LA SOCIÉTÉ STATISTIQUE DE PARIS

LUCIEN MARCH

Différences et corrélation en statistique

Journal de la société statistique de Paris, tome 69 (1928), p. 38-63

<http://www.numdam.org/item?id=JSFS_1928__69__38_0>

II

DIFFÉRENCES ET CORRÉLATION EN STATISTIQUE

Dans des communications anciennes à la Société de Statistique de Paris j'ai signalé des moyens de fonder la méthode statistique sur l'expérience humaine aidée du raisonnement, sans faire intervenir les hypothèses de continuité.

Les détails de la méthode sont alors pleinement compréhensibles et vérifiables par ceux dont les connaissances en mathématique sont rudimentaires, — à la condition toutefois que les notations symboliques ou figurées, indispensables pour la clarté et la précision des raisonnements abstraits, ne les rebutent point. Aujourd'hui la statistique pénètre dans la direction des affaires aussi bien que dans les sciences de la vie : il est donc intéressant que l'esprit de sa méthode soit aisément accessible, et en liaison étroite avec les faits qui,

même groupés en masses, conservent souvent une individualité dont il importe de ne pas méconnaître la valeur.

Comme dans toute recherche scientifique, il s'agit de comparer les choses pour noter leurs analogies et leurs différences, puis de caractériser ces analogies et ces différences par des grandeurs simples qui soient de bons instruments de comparaison. Mais il convient aussi de tenir compte de l'ordre suivant lequel se disposent ces choses.

Souvent en effet les éléments numériques fournis par l'observation des faits se présentent naturellement dans un certain ordre; les caractéristiques à calculer peuvent alors dépendre de cet ordre; d'autres caractéristiques en sont indépendantes. On dispose ainsi de deux catégories de termes de comparaison, soit que l'on se propose de classer plusieurs ensembles d'éléments suivant quelque caractère commun comme la taille, le revenu, etc., soit qu'on cherche à les distinguer d'après l'inégalité de ce caractère d'un élément à l'autre.

C'est ainsi que les ouvriers d'une usine étant classés d'après leurs salaires horaires, le salaire de l'ouvrier qui se trouve au milieu de la série donne la valeur *médiane*, élément de comparaison de plusieurs séries d'ouvriers. On peut de même noter la place de l'ouvrier tel que le total des salaires distribués se répartisse également entre les ouvriers moins payés et les ouvriers plus payés que lui, ce qui détermine la valeur *médiale* des salaires. Ces deux ***médiantes***, on le voit, dépendent de l'ordre des éléments dont elles caractérisent les grandeurs.

Les *moyennes* ne dépendent en aucune façon de cet ordre; on peut en imaginer une infinité; la moyenne arithmétique des valeurs des éléments de la série est la caractéristique la plus simple, indépendante de l'ordre, des éléments de la série.

On dispose, en somme, de divers instruments, dont chacun a ses caractères propres et son intérêt particulier, pour comparer commodément des ensembles divers d'éléments qui sont groupés d'après quelque caractère commun.

Si l'on s'en tient là, l'étude est incomplète et peut conduire à de fausses interprétations. Il faut aussi apprécier les différences. Les grandeurs intermédiaires considérées précédemment ne tiennent aucun compte de l'inégalité des éléments. Deux catégories de problèmes se posent alors :

1° On se propose de caractériser les différences que présentent entre eux les éléments d'un même ensemble, en vue de les comparer à celles que présentent les éléments de quelque autre ensemble. La comparaison porte sur la variabilité des éléments de chaque ensemble sans qu'il y ait lieu de se préoccuper des relations, des affinités, qui peuvent exister entre les éléments d'ensembles différents;

2e On cherche à découvrir ces relations par l'examen des différences élémentaires d'un ensemble à un autre, toutes les fois qu'il existe entre les éléments des deux ensembles une certaine correspondance.

Par exemple, supposons que l'on veuille comparer en France les salaires des ouvriers agricoles en différentes régions et les rendements culturaux. On peut d'abord essayer de mesurer, d'une part l'inégalité des salaires, d'autre part l'inégalité des rendements, c'est le premier problème. Dans cette recherche on ne prête aucune attention à la situation géographique de telle ou telle région, c'est-à-dire à l'individualité propre des éléments de chaque ensemble.

On peut au contraire se demander s'il existe une relation entre les salaires et les rendements. Peu importe d'ailleurs, pour le moment, que l'un des deux éléments en présence détermine l'autre, ou bien que tous deux soient soumis à quelque influence commune. Mais dès que l'on veut découvrir une relation, il faut mettre en présence les éléments qui sont susceptibles de liaison, par exemple le salaire dans une région déterminée avec le rendement dans cette région, la région établissant une correspondance entre les éléments des deux ensembles, constituant le caractère commun qui les unit.

Examinons successivement les deux problèmes.

I. — DIFFÉRENCES

Pour préciser le premier des deux problèmes, prenons comme exemple une répartition de salaires. Dans un tableau inséré dans ce journal, il y a exactement trente ans (1), j'ai comparé plusieurs distributions de salaires, parmi lesquelles celle de 13.000 ouvriers métallurgistes américains dont on a pu apprécier les salaires individuels. Le nombre est naturellement trop considérable pour que l'on puisse raisonner sur les cas individuels et se les représenter. Aussi ai-je décomposé la liste des ouvriers classés par ordre croissant du salaire journalier en vingt-six groupes de 500 ouvriers, admettant que, dans chaque groupe de 500, les salaires sont assez voisins pour que l'on puisse attribuer à chaque groupe le salaire moyen des ouvriers du groupe. D'ailleurs cette distribution de 26 ouvriers théoriques est simplement destinée à illustrer le raisonnement d'après un exemple concret. Et il permet une représentation graphique simple du problème.

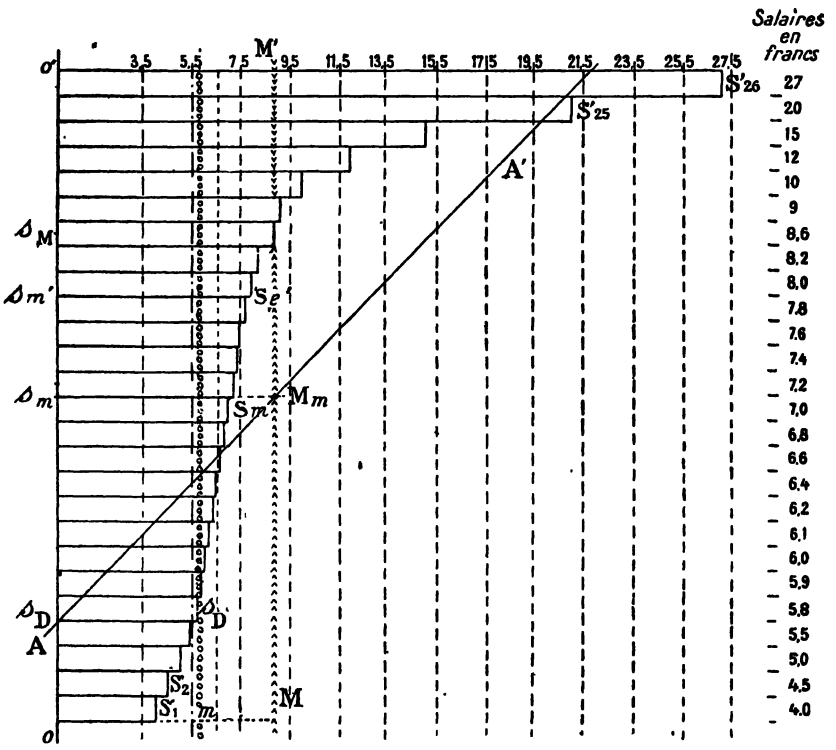

Fig. 1.

(1) Quelques exemples de distribution de salaires, *Journal de la Société de Statistique de Paris*, 1898.

Sur la figure 1, chacun des 26 ouvriers est représenté par une bande d'épaisseur uniforme, pour tous les ouvriers, et assez faible pour que, dans le langage, on puisse la négliger. Les bandes s'appuient à angle droit sur une ligne droite de base; elles sont juxtaposées et se suivent par ordre de largeur croissante, cette largeur étant proportionnée au salaire de l'ouvrier représenté.

Sur cette figure les éléments de la série des salaires sont ordonnés et l'on détermine immédiatement la position et la grandeur dela médiane *sm* S*m* placée au centre de la hauteur de la figure,a insi que de la médiale s_m' S_m', dont la ligne partage en deux parties égales la surface des bandes. Les grandeurs de ces deux caractéristiques, représentées par les largeurs des bandes *sm* S*m*, s_m' S_m', dépendent de l'ordre dans lequel les bandes sont disposées. Si au contraire on trace une parallèle à l'axe de base qui, entre cet axe et les bandes extrêmes, forme un rectangle équivalent à la surface des bandes, on voit que la moyenne correspond à la bande dont l'extrémité serait sur la parallèle. La largeur de cette bande est entièrement indépendante de l'ordre suivant lequel les bandes sont disposées.

Caractéristiques de l'inégalité ou de la variabilité dépendant de l'ordre des éléments.

Pour obtenir une mesure de l'inégalité des bandes, on distingue aussi deux catégories de caractéristiques. Les premières dépendent de l'ordre des éléments. On sait, par exemple, qu'après avoir déterminé la médiane s_m S_m qui partage l'ensemble des ouvriers en deux groupes également nombreux, on détermine de nouveau les médianes des deux groupes que l'on appelle *quartils* (1) L'intervalle de ces quartils ou *interquartil* caractérise assez bien l'inégalité des salaires puisque, dans cet intervalle, est comprise une moitié des ouvriers. Plus il est grand, plus les salaires sont généralement inégaux; plus il est petit plus les salaires sont généralement concentrés autour de la médiane.

Mais cette constatation ne donne point une solution suffisante du problème. On voit bien, par exemple, qu'un petit nombre d'ouvriers seulement reçoit les plus hauts salaires, et qu'entre le plus grand nombre règne une certaine égalité, mais il se peut que ce plus grand nombre reçoive néanmoins une forte partie du salaire total distribué ou qu'au contraire il ne reçoive qu'une faible partie. La même distinction intervient dans tous les problèmes de répartition, dans celui de la répartition des revenus en général aussi bien que dans le cas particulier, examiné ici, du revenu du travail manuel.

La caractéristique que nous avons appeleé médiale permet de compléter l'étude. La droite $s_{m'}$ $S_{m'}$, représentant la médiale des salaires, partage en deux parties égales la surface des bandes qui représente le total des salaires. En partageant de nouveau chacune de ces parties par de nouvelles médiales que l'on peut appeler des *quartals*, l'intervalle compris entre ces quartals, ou l'*interquartal*, caractérise la concentration ou la dispersion des sommes réparties entre les ouvriers. Ainsi, dans l'exemple des 26 ouvriers types américains dont les salaires sont supposés échelonnés entre 4 francs et 27 francs, la médiane a pour valeur 7f 10, la médiale 8 francs, l'interquartil 2f 60, l'interquartal

(1) En anglais quartile, expression proposée par Fr. Galton qui, le premier, a employé ce procédé de comparaison.

5f 65. La comparaison des deux derniers chiffres indique que, si les salaires sont concentrés assez étroitement autour de la valeur médiane, puisque les salaires de la moitié des ouvriers sont compris entre 6 francs et 8f 60, la somme distribuée est moins également répartie : la moitié de cette somme va à des ouvriers qui gagnent de 6f 40 à ceux qui en gagnent 12. D'ailleurs comme la valeur absolue de ces caractéristiques dépend de la grandeur des salaires, ou de l'unité qui sert à les mesurer, il vaut mieux les comparer en valeurs relatives par rapport soit à la médiane, soit à la médiale.

L'interquartil relatif est $\frac{2,6}{7,1} = 0,36$; l'interquartal relatif : $\frac{5,65}{8} = 0,71$.

La comparaison de ces deux fractions nous renseigne sommairement et rapidement sur la double inégalité qu'il s'agissait de caractériser.

On peut aussi employer une autre caractéristique plus simple qui est l'écart entre la médiale et la médiane. Supposons un instant tous les ouvriers payés au même taux de salaire : l'écart est nul. Supposons que, sur les 26 ouvriers, 25 reçoivent le salaire le plus bas, un seul d'entre eux recevant le surplus du total, la médiale serait le salaire de ce dernier ouvrier; elle serait représentée par la plus large bande de la figure, la seule qui se distinguerait des autres et l'inégalité serait extrême. Dans ce cas l'écart entre la médiale et la médiane serait maximum. Sa valeur relative par rapport à la médiale serait égale à l'unité.

Dès lors on peut caractériser par cet écart relatif entre la médiale et la médiane — écart relatif que l'on peut appeler *déviation intermédiante* — la différence constatée entre l'inégalité des salaires payés respectivement aux différents ouvriers et l'inégalité de la distribution du salaire total entre les différents ouvriers. Les deux notions de l'inégalité sont différentes car une fraction considérable de la population peut couvrir la plus grande partie de l'échelle des revenus et cependant ne recevoir que la plus petite partie du total.

Dans l'exemple des 26 ouvriers, la déviation intermédiante a pour valeur $\frac{0,9}{8} = 0,11$, fraction fort éloignée de l'unité, qui correspond à l'extrême inégalité.

Dans le même ordre d'idées, l'écart entre la dominante (1) et la moyenne simple peut aussi être pris comme mesure de la déviation.

Caractéristiques de variabilité indépendantes de l'ordre des éléments.

Cependant on jugera peut-être que les caractéristiques précédentes ne donnent point une idée tout à fait satisfaisante de l'inégale répartition des salaires. En fait, chaque ouvrier compare son gain à celui de ses compagnons pris individuellement. Et d'ailleurs le salaire ne dépend qu'en faible part des conditions économiques générales. Il dépend surtout des caractères individuels, des aptitudes, de l'état physique et mental de chaque ouvrier, de sa productivité.

Or, on résumerait les observations individuelles en notant d'abord, successivement, la différence de chacun des salaires avec tous les autres. Cela fait, la grandeur commune de ces différences peut être caractérisée par l'une des grandeurs intermédiaires considérées plus haut.

En fait, entre deux éléments a et b de la série des salaires, la différence peut

(1) Metron (vol. VI, n° 2, p. 56), *Les mesures de la variabilité.*

être prise soit entre a et b, soit entre b et a. Il suffit donc de considérer les différences formées entre l'élément le plus grand du couple et l'élément le plus petit. Le total de ces différences sera la moitié du total de toutes les différences possibles.

Si l'on fait rentrer dans ce total les différences nulles telles que $a - a$, celui-ci ne change pas : alors le nombre total des différences possibles entre deux quelconques des n éléments d'une série est égal à n^2.

Il s'agit en somme de déterminer le total des différences formées entre deux éléments dont le premier est supérieur au second.

Dans la série des 26 salaires pris comme exemple, le nombre n est égal à 26. Il est pair et nous supposerons qu'en général il est pair; le cas de n impair se traite de la même façon, les formules changeant peu.

Considérons les deux ouvriers dont les salaires Si et S_{-i} sont symétriquement placés par rapport à la médiane.

Formons les différences entre ces éléments de la série et une autre S_k compris entre eux.

On a évidemment $Si - S_{-i} = (S_i - S_k) + (S_k - S_{-i})$.

Les deux différences comprises dans les parenthèses du second membre sont positives.

L'égalité précédente est vraie quelque soit k ($k \leqslant i$), de sorte que la somme des différences entre deux éléments Si, S_{-i} et un élément intermédiaire quelconque est égale à $Si - S_{-i}$ répété autant de fois qu'il y a d'éléments intermédiaires.

De même si m est la grandeur médiane, la différence $Si - S_{-i}$ peut s'écrire $(Si - m) + (m - S_{-i})$. Par suite, le total des différences possibles entre deux éléments S_i, $S - i$, symétriques par rapport à la médiane, et les éléments intermédiaires, est la somme des écarts des mêmes éléments à la médiane, chaque écart étant multiplié par le nombre des éléments compris entre Si et S_{-i}. Ce dernier nombre est lui-même égal au double du nombre des éléments compris, soit entre Si et la médiane, soit entre celle-ci et S_{-i}.

Désignons maintenant par x_i l'écart entre la grandeur d'un élément quelconque et la grandeur de la médiane, par y_i le nombre des éléments compris entre cet élément et la médiane, le total des différents possibles entre deux éléments symétriques quelconques est égal à $2 \sum_{i=\frac{n-1}{2}}^{\frac{n+1}{2}} x_i y_i$ n étant supposé pair et, à cause de la symétrie, tous les produits $x_i y_i$ étant positifs.

Dans ce qui précède, on n'a compté que les différences distinctes et entre éléments différents. Or le total des différences ne change pas si l'on y incorpore les différences nulles d'un élément avec lui-même. En comptant aussi les différences formées entre le plus petit de deux éléments et le plus grand, on double le total qui devient alors égal à $4 \sum_{\frac{n-1}{2}}^{\frac{n+1}{2}} x_i y_i$.

Le nombre des différences devient égal à n^2 et la moyenne de toutes ces différences, ou la *différence moyenne* d[1], est :

(3) $$d = \frac{4 \sum_{i=1}^{\frac{n+1}{2}} x_i y_i}{n^2}$$ puisqu'il y a n^2 différences.

On peut aussi écrire :

$$d = \frac{2 \sum_{i=\frac{1}{2}}^{i=\frac{n+1}{2}} (\overline{S_{n-i} - S_i})\,(n - 2i + 1)}{n^2}$$ d'après l'égalité (1).

Sur la figure 1 la droite qui joindrait l'extrémité de la bande S_{n-i} au point m a pour pente $\frac{x_i}{y_i} = \frac{x_i y_i}{y_i^2}$.

On peut caractériser l'ensemble des pentes de toutes les droites qui joignent au point m les extrémités de toutes les bandes par une pente moyenne obtenue en divisant la somme des numérateurs de cette fraction par la somme des dénominateurs (moyenne arithmétique généralisée). Cette pente moyenne a donc pour expression :

$$p = \frac{\sum_{-\frac{n-1}{2}}^{+\frac{n+1}{2}} x_i y_i}{\sum y^2}$$

Exprimée au moyen de la différence moyenne, sa valeur est :

$$p = \frac{d n^2}{4 \sum y^2}, \text{ or } \sum y^2 = 2\,(1^2 + 2^2 + \ldots + (\tfrac{n-1}{2})^2) = 2\,\frac{\frac{n-1}{2}\,\frac{n+1}{2}\,n}{6} = \frac{n(n-1)(n+1)}{6}$$

D'où :

$$p = \frac{d}{\frac{1}{3}(n - \frac{1}{n})}$$

l Lorsque n est grand, la pente moyenne est celle d'une droite qui, sur une dongueur égale au tiers du nombre des éléments de la série, s'élèverait de la ifférence moyenne de ces éléments.

Une droite quelconque passant par le point médian coupe les axes des bandes en des points tels que la somme des distances de ces points aux extrémités des bandes, ces distances prises en valeurs absolues, est plus petite que, si la droite, conservant la même direction, passait par un autre point de la ligne médiane. Mais nous ne pouvons déterminer d'une manière générale la direction pour laquelle la somme des distances est la plus petite possible. Les grandeurs absolues ne peuvent s'introduire dans les calculs qui sont fondés sur la règle des signes.

Considérons dès lors les grandeurs prises avec leurs signes. Pour que des sommes de grandeurs positives ou négatives apparaissent dans le calcul à peu

(1) Expression employée par M. Corrado Gini qui, le premier, a donné une théorie de la différence moyenne dont les explications précédentes sont inspirées.

près telles qu'elles sont représentées graphiquement, on a recours à un artifice qui consiste à les élever au carré. De la sorte, de même qu'avec des grandeurs prises en valeurs absolues, on obtient des quantités toujours positives.

Mesurons les éléments de la série à partir de leur moyenne. Alors la somme des carrés de toutes les différences possibles des n éléments de la série, somme qui peut s'écrire : $\sum_1^n \sum_1^n (x_i - x_k)^2$, est égale au double de la somme des carrés des mêmes éléments puisque les doubles produits donnent une somme nulle.

Mais, dans chaque somme de carrés, un même élément entre n fois pour la même raison que plus haut.

On peut donc écrire :

$$\sum_1^n \sum_1^n (x_i - x_k)^2 = 2n \sum_1^n x_i^2$$

Comme il y a n^2 différences possibles le carré moyen des différences est égal à $2 \frac{\sum_1^n x_i^2}{n} = \frac{\sum_1^n \sum_1^n (x_i - x_k)^2}{n^2}$. D'après cela le carré moyen des différences possibles entre deux éléments quelconques de la série n'est autre chose que le double du carré moyen des écarts des éléments à partir de leur moyenne (1). D'autre part la droite qui, issue du point M, situé sur la ligne médiane à une distance de l'axe de base égale à la moyenne, a pour pente (2) $p = \frac{\Sigma xy}{\Sigma y^2}$. Cette droite coupe les axes des bandes en des points dont les distances aux extrémités, élevées au carré, donnent la somme la plus petite.

II. — CORRÉLATION

Les remarques précédentes s'appliquent à la comparaison des éléments de divers ensembles, par exemple à la comparaison des revenus de populations différentes, quand on ne tient point compte de l'individualité des éléments, tous les éléments du même ensemble intervenant indifféremment dans le calcul des caractéristiques qui sont les instruments de comparaison.

Lorsque les éléments de deux ensembles se correspondent d'après un caractère quelconque, de sorte que, l'un étant fixé, certains de ceux de l'autre ensemble sont également fixés, la question se pose de savoir si cette correspondance implique ou n'implique pas de liaison entre les éléments qu'elle met en présence.

Dans un laboratoire de physique, on opère généralement sur des grandeurs que l'on peut faire varier d'aussi peu qu'on le veut et d'une manière réversible. Par exemple on compare la dilatation d'un métal à la température d'un milieu dans lequel il est plongé. Les observations des deux espèces se correspondent en raison de leur simultanéité. Elles mettent en évidence la formule qui relie ces observations et elles déterminent la grandeur d'un élément d'une espèce d'après la grandeur de l'élément correspondant de l'autre.

Hors du laboratoire d'expériences physiques, on n'obtient pas de relations

(1) Sur les procédés élémentaires qui se rapportent à cette théorie, voir l'article précité dans *Metron*, vol. VI, n° 2.

(2) Cette fois, les x étant mesurés à partir de la moyenne et non à partir de la médiane, les produits xy ne sont pas nécessairement tous positifs.

exprimables aussi simplement par la seule considération des éléments correspondants. On ne peut faire varier les phénomènes d'aussi peu qu'on le veut, d'une manière continue, pour ainsi dire, et réversible. On doit se contenter d'observations spontanées, sporadiques, suivant une évolution imposée. Pour pouvoir juger s'il existe ou s'il n'existe pas de relation entre les manifestations de deux phénomènes, et se faire une idée de l'importance de la relation, il faut tenir compte de toutes les observations. La grandeur qui caractérise l'un des phénomènes ne peut généralement être déterminée par la seule connaissance de la grandeur correspondante de l'autre phénomène.

Dans nos jugements, nous sommes guidés par des coïncidences plus ou moins fréquentes qu'il faut seulement associer et peser avec méthode. Comment distinguer les coïncidences exceptionnelles et sans valeur de celles qui impliquent des relations quasi permanentes? D'abord par le dénombrement des cas qui les laissent apparaître et de ceux qui ne les présentent pas, sans omettre aucun de ceux qui sont accessibles. En second lieu, il convient de soumettre les cas observés à un examen critique attentif, d'après les connaissances acquises sur les faits analogues ou voisins.

La statistique, que Léon Say appelait avec une juste simplicité la science des dénombrements, a la charge de cette méthode. Avant qu'elle fût constituée d'ailleurs, les esprits justes avaient compris que des observations nombreuses pouvaient seules justifier de sérieuses conjectures, qu'il était bon de tracer des courbes, d'en constater le parallélisme ou l'indifférence, ce qui revient à appliquer le principe de logique auquel Stuart Mill a donné le nom de principe des variations concomitantes (1). Encore convient-il d'appliquer ce principe tel que son auteur l'a formulé : « Un phénomène qui varie d'une certaine manière *toutes les fois* qu'un autre phénomène varie de la même manière est une cause ou un effet de ce phénomène ou bien y est lié par quelque fait de causation. »

Les derniers mots s'appliquent au cas de corrélation entre deux phénomènes, sans que l'un soit antécédent de l'autre. Sous réserve des vérifications que Mill lui-même a recommandées, sa formule résume les moyens que l'on a employés de tout temps, avec plus ou moins de sagacité et de scrupules, pour découvrir les relations des choses. Mais il importe de ne point perdre de vue les mots « toutes les fois ». La corrélation de deux phénomènes ne saurait être admise que si l'accord des variations existe pour toutes les variations, tout au moins pour toutes les variations observables distinctement.

Cette remarque est surtout importante quand on se propose, non point d'établir la corrélation parfaite de deux séries de grandeurs — corrélation parfaite qui équivaut à l'identité numérique par l'intermédiaire d'une formule — mais de mesurer un certain degré de corrélation. On essaie alors d'exprimer numériquement l'impression que laisse une représentation graphique des phénomènes comparés.

Prenons un exemple de corrélation à peu près parfaite, par exemple celle qui existe entre la pression d'un poids donné de gaz et son volume à température constante.

On peut traduire le résultat sur un tableau graphique de deux manières, suivant que les observations sont classées dans l'ordre même de leur achève-

(1) *Système de logique*, livre III, ch. VIII, cinquième canon.

ment, ou bien suivant qu'elles sont classées, indépendamment de cet ordre, d'après la grandeur de l'un des éléments mesurés.

Dans le premier système, on tracera la courbe de la pression, puis la courbe du volume, soit aux instants successifs des observations faites au même lieu, soit aux divers endroits d'observations simultanées, soit d'après quelque qualité des observateurs. Les deux courbes, tracées sur le même axe horizontal de base, d'après des échelles judicieusement fixées, se présentent en sens inverse. Mais en faisant tourner l'une d'elles autour de l'axe commun, on constate, après la rotation, un certain parallélisme. En fait les courbes ne sont point parallèles. Or, la relation est presque rigoureuse, à tel point que si l'on transforme les ordonnées de l'une des courbes au moyen d'une formule simple uniforme, les deux courbes deviennent alors presque rigoureusement parallèles.

Ainsi l'examen des courbes nous a bien révélé la liaison qui existe entre les changements des deux phénomènes pression et volume, mais il ne nous a pas permis de déterminer directement le degré de la liaison.

Remarquons d'ailleurs que, si les courbes sont suffisamment régulières, c'est-à-dire si la succession des pressions et des volumes correspondants n'est pas désordonnée, elles permettent de comparer aussi bien les variations d'un point au suivant, que d'un point à un autre plus éloigné, ou que des différents points à partir de l'axe de base. Elles permettent en somme d'apprécier d'un coup d'œil toutes les variations concomitantes. Mais l'impression est souvent confuse; il est nécessaire de la préciser.

Dans le second système de représentation, le plus employé en physique, les éléments numériques de l'une des séries, par exemple de la série des volumes, sont classés par ordre de grandeur et représentés par des longueurs portées sur un axe horizontal à partir d'une certaine origine. Les éléments correspondants de l'autre série (pression) sont représentés sur des verticales dont les extrémités dessinent une courbe et cette courbe représente la loi du phénomène.

Ici les éléments correspondants sont classés de façon à faire apparaître seulement les variations à partir de l'origine des mesures, sans aucun souci de l'ordre dans lequel les variations se sont produites effectivement et par conséquent des variations qui dépendent de cet ordre.

Rappelons maintenant sommairement comment les deux systèmes ont reçu application dans les recherches statistiques (1).

Covariation différentielle.

Pour apprécier la dépendance de deux phénomènes, le psycho-physicien allemand Fechner effectuait le dénombrement des variations concomitantes d'une observation à la suivante (2), c'est-à-dire des variations différentielles. La règle des signes fournit un moyen commode de faire le compte des concordances et des discordances, puisque les variations de même sens (positives ou négatives) donnent un produit positif, les variations de sens contraires un produit négatif. La balance des produits qui s'écrit $I = \Sigma\, v \times v'$ donne ainsi

(1) Pour plus de détails, voir Les Représentations graphiques et la statistique comparative, *Journal de la Société de Statistique,* numeros d'août et septembre 1905, notamment page 270.

(2) *Kollektivmasslehre,* herausgegeben, von C. Lipps, 1897.

une mesure du parallélisme des deux courbes (les unités des quantités v et v' étant convenablement choisies).

Si l'un des phénomènes comparés ne change pas, tandis que l'autre se modifie, on en conclut que le premier est indifférent aux variations du second, alors la balance des produits est nulle.

Cependant le parallélisme des courbes, comme on l'a dit tout à l'heure, ne justifie pas nécessairement la corrélation des faits représentés; il ne permet pas toujours de mesurer le degré de cette corrélation. Un autre exemple est celui de la roue qui avance en tournant. L'avance et la rotation sont étroitement liées; cependant les deux lignes qui représentent, soit l'avance, soit la hauteur, d'un point de la roue après une révolution sont indifférentes; le coefficient I est nul, même si les variations des lignes sont prises entre deux points distants d'un intervalle quelconque (1).

En dehors des mouvements périodiques, qui donnent lieu à d'autres singularités, l'étude du parallélisme des courbes fournit cependant le seul moyen dont nous disposions pour caractériser l'accord ou le désaccord des variations de deux phénomènes. Encore convient-il de ne point s'en tenir aux variations différentielles de premier intervalle, c'est-à-dire entre éléments contigus, mais de considérer aussi les variations entre éléments distants d'intervalles plus ou moins étendus. Ainsi l'on se rapproche des conditions d'application de la méthode expérimentale.

Dans ces conditions, le coefficient I ne saurait indiquer que la covariation différentielle.

Covariation tendancielle.

Cherchons maintenant à généraliser ce coefficient quel que soit l'intervalle entre lequel se détermine la variation dans chaque série.

Désignons par x l'ordonnée de l'une des courbes, par y l'ordonnée correspondante de l'autre. Si v et v' désignent cette fois des variations entre éléments séparés par un intervalle quelconque :

$$\Sigma\, v\, v' = \Sigma\, (y_{i+h} - y_i)\,(x_{i+h} - x_i)$$

Or on démontre aisément que le second membre est égal à $\Sigma\, xy - Sx\, Sy$ en désignant par Sx et Sy respectivement les sommes des ordonnées x ou y.

Supposons que les ordonnées de chaque espèce soient mesurées à partir de leur moyenne. Alors $Sx = Sy = o$, de sorte que l'expression caractéristique de l'accord ou du désaccord de toutes les variations que l'on peut observer entre deux points quelconques de l'une des courbes et les deux points correspondants de l'autre est donnée par la somme des produits des ordonnées correspondantes, mesurées à partir de leurs moyennes; c'est-à-dire par l'expression $\Sigma\, xy$.

D'après ce qui précède, cette somme n'exprime pas autre chose que la balance des concordances et des discordances des variations concomitantes, lorsque l'on considère *toutes les variations* différentielles possibles. Elle caractérise, par sa valeur, la *tendance* générale des variations, d'une part entre éléments d'une série, d'autre part entre leurs correspondants dans l'autre, à un certain accord. Pour cette raison, on peut la regarder comme indiquant un

(1) Voir Metron, tome I, fasc. 1, p. 47.

état de *covariation tendancielle* dans la conjugaison des deux séries d'éléments.

Pour obtenir un coefficient variant entre les limites 0 et 1, ce qui est commode, on remarquera que la covariation la plus parfaite existe naturellement quand les éléments x et les éléments y correspondants sont égaux, auquel cas la somme $\Sigma\, xy$ devient égale à $\Sigma\, x^2$ ou à $\Sigma\, y^2$ ou à $\sqrt{\Sigma\, x^2\, \Sigma\, y^2}$. Le coefficient de covariation tendancielle mis sous la forme $r = \dfrac{\Sigma\, xy}{\sqrt{\Sigma\, x^2}\sqrt{\Sigma\, y^2}}$ ne peut être supérieur à 1 et il peut descendre jusqu'à la valeur 0.

De même le coefficient de covariation différentielle prendra la forme

$$j = \frac{\Sigma\, \delta x\, \delta y}{\sqrt{\Sigma\, \delta\, x^2}\sqrt{\Sigma\, \delta\, y^2}}.$$

δx et δy exprimant les variations différentielles désignées plus haut par les lettres v et v'.

Utilité des deux coefficients de covariation.

Voyons maintenant comment on peut obtenir une représentation graphique du coefficient r.

Le coefficient r exprime en somme la balance des concordances et discordances des variations tendancielles, c'est-à-dire comptées à partir d'un axe moyen. On peut lui attribuer une autre signification, d'après l'égalité :

$$\sum_1^n (x-y)^2 = \Sigma\, x^2 + \Sigma\, y^2 - 2\,\Sigma\, xy = \Sigma\, x^2 + \Sigma y^2 - 2r\sqrt{\Sigma\, x^2 + \Sigma\, y^2}$$

d'où

$$r = \frac{1}{2}\left[\frac{\Sigma\, x^2 + \Sigma\, y^2}{\sqrt{\Sigma\, x^2\, \Sigma\, y^2}} - \frac{\Sigma\,(x-y)^2}{\sqrt{\Sigma\, x^2\, \Sigma\, y^2}}\right] = \frac{1}{4}\,\frac{1}{\sqrt{\Sigma\, x^2\, \Sigma\, y^2}}\,[\Sigma\,(x+y)^2 - \Sigma\,(x-y)^2]$$

r est nul quand $\Sigma\,(x-y)^2$ atteint sa plus grande valeur qui est $\Sigma\, x^2 + \Sigma\, y^2$ ou $\Sigma\,(x+y)^2$; il atteint sa plus grande valeur positive quand $\Sigma\,(x-y)^2$ est nul, auquel cas $r = \dfrac{\Sigma\,(x+y)^2}{4\sqrt{\Sigma\, x^2\, \Sigma\, y^2}} = 1$ puisque, dans ce cas, $x = y$ constamment, et sa plus grande valeur négative quand $\Sigma\,(x+y)^2$ est nul, auquel cas y est constamment égal à $-x$.

$\Sigma\,(x-y)^2$ est un indice de l'écartement des deux courbes en x ou en y; r est égal à 1 quand les deux courbes coïncident et à -1 quand les deux courbes sont écartées à tel point que les ordonnées correspondantes soient toujours de signe contraire (1).

Si les deux courbes coïncident, leur parallélisme est évidemment parfait. La corrélation des deux phénomènes qu'elles représentent est parfaite. Mais dès que les deux courbes ne coïncident plus, si peu que ce soit, elles peuvent ne présenter aucun parallélisme, c'est-à-dire qu'elle ne satisfont en aucune façon au principe des variations concomitantes.

D'autre part, les deux courbes peuvent présenter à la fois un parallélisme évident et un écartement considérable. Les figures 2 à 5 représentent des cas assez différents : figure 2, les deux coefficients de covariation sont nuls ou à

(1) La même représentation peut s'appliquer au rapport de corrélation de K. Pearson, qui suppose les éléments répartis par sections dont les poids entrent en compte.

peu près nuls et cependant la corrélation peut être parfaite (cas de la roue); figure 3, le coefficient de covariation tendancielle est voisin de 1 et positif, le coefficient de covariation différentielle est voisin de 1 et négatif; figure 4, les deux courbes sont très voisines, le coefficient de covariation différentielle à peu près nul, le coefficient de covariation tendancielle voisin de l'unité; figure 5, les deux courbes (tracées cette fois en gradins) sont écartées au maximum, le coefficient de covariation tendancielle est égal à — 1, le coefficient de covariation différentielle également. Si l'une des courbes se réduisait à l'axe de base, les deux coefficients seraient nuls.

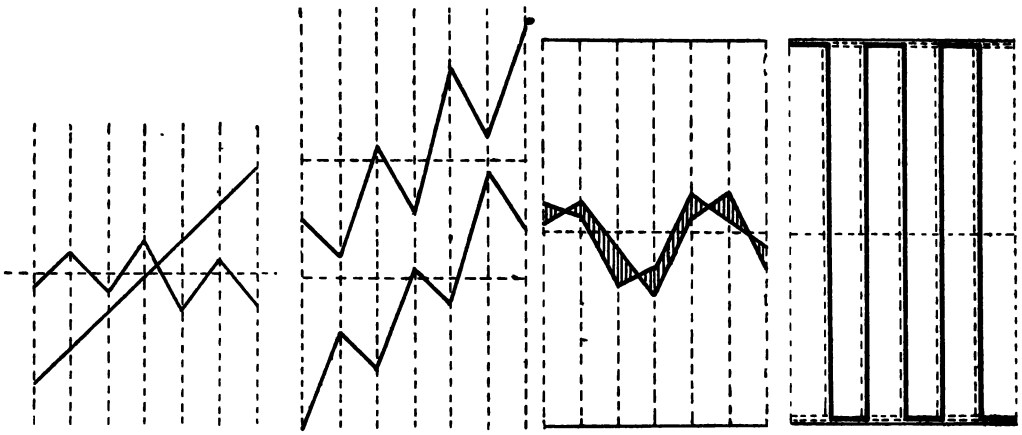

Fig. 2. Fig. 3. Fig. 4. Fig. 5.

Par contre, si les deux courbes sont rapprochées et qu'en même temps leurs mouvements s'orientent dans les mêmes sens, auquel cas les deux coefficients *j* et *r* ont une valeur assez élevée, il semble que les variations différentielles à intervalles supérieurs donneraient également des concordances plus nombreuses que les discordances.

Ainsi un seul des deux coefficients ne suffit pas pour suggérer la corrélation; mais si les deux coefficients ont une valeur assez grande, supérieure à $\frac{1}{2}$ par exemple, la corrélation est généralement établie.

Cas particuliers d'accord entre les coefficients de covariation.

D'ailleurs, en statistique, il peut être également intéressant d'établir entre deux phénomènes une relation qui n'a lieu que pour des variations de faible amplitude, ou bien au contraire une relation qui, n'apparaissant pas entre mouvements peu étendus, se manifeste quand on fait abstraction des changements à court intervalle pour ne considérer que des mouvements d'ensemble ou inversement. Dans le premier cas, il convient de porter l'attention sur les variations différentielles, dans le second il y a lieu de faire disparaître ces variations en opérant sur des moyennes.

En substituant à des points successifs de l'une et de l'autre courbe un point moyen et en substituant aux courbes primitives deux courbes tracées par ces points moyens, les courbes nouvelles ne présentent plus de petites sinuosités. Leur forme est adoucie, leur allure uniforme : alors les variations diffé-

rentielles ont très généralement le même sens que les variations tendancielles. Les deux coefficients j et r donnent des valeurs qui concordent suffisamment; comme exemple je rappellerai une communication antérieure sur la relation qui existe entre les mouvements des prix à la longue période observés depuis le début du XIX^{e} siècle et le taux d'accroissement du stock des métaux précieux. Cette relation n'apparaît pas quand on compare les variations à courte période; elle ressort de la comparaison de moyennes calculées sur un certain nombre d'années (1).

Ainsi, dans le cas de courbes à allure régulière, les deux coefficients de covariation ont à peu près la même signification; on peut employer l'un ou l'autre pour établir la corrélation.

Dans un autre cas particulier, qui présente un grand intérêt pour les observations physiques, et en même temps pour l'histoire de la théorie de la corrélation, le calcul du coefficient de covariation tendancielle suffit aussi pour établir la corrélation. Il convient d'en chercher la raison.

Lorsque Fechner reconnaissait la dépendance de deux phénomènes en comparant leurs variations d'un moment de l'observation au suivant, — ce qui revient à comparer les courbes du point de vue du parallélisme — il employait un procédé empirique : il constatait des concordances ou des discordances sans chercher la forme de la liaison qui expliquait ces rapports,

Un autre savant, l'anthropologiste anglais Fr. Galton, a abordé le même problème d'un autre point de vue. Étudiant l'hérédité des caractères, il cherchait sous quelle forme se manifeste cette hérédité (2). Considérant la taille, par exemple, il représentait par un point d'un plan le couple formé par la taille d'un individu et la taille moyenne de ses parents. L'observation de 928 individus lui permit de dresser une table à double entrée dont les rangées correspondaient aux divisions de taille des parents, les colonnes aux degrés de taille des enfants. Dans chaque case, à la rencontre d'une rangée et d'une colonne, s'inscrit le nombre des individus dont la taille correspond au degré indiqué par l'entête de la colonne et dont les parents ont la taille moyenne indiquée par l'entête de la rangée. Dans chaque rangée Galton détermina la taille moyenne des enfants issus de parents dont la taille correspond à la rangée et il représenta les données de sa table par un graphique disposé comme suit.

Dans le plan, supposé divisé en quatre quadrants par deux axes rectangulaires, sur l'axe horizontal **OX** on porte les tailles des enfants, sur l'axe vertical OY les tailles de parents. Ces tailles sont supposées mesurées à partir de leurs moyennes respectives, de sorte que l'origine des coordonnées est au point qui représente un couple formé l'individu moyen et le parent moyen.

Une rangée de la table est représentée par une bande parallèle à l'axe horizontal, une colonne par une bande parallèle à l'axe vertical.

Si la taille de l'enfant à l'âge adulte était indépendante de celle de ses parents, les points représentatifs des couples : parent et enfant, se répartiraient indifféremment d'un côté ou de l'autre de l'axe vertical, quelle que soit la taille parentale. Tous les points moyens déterminés pour les bandes horizontales successives se trouveraient au moins à peu près, sur la verticale $o\,y$.

---

(1) *Journal de la Société de Statistique,* mars 1912. p. 111
(2) *Natural inheritance.* London, 1889.

Si la taille de l'enfant parvenu à l'âge adulte était égale à celle de ses parents, tous les points représentatifs des couples seraient situés sur la bissectrice de l'angle des axes. S'il n'en est pas constamment ainsi, mais si néanmoins cette égalité existe *en moyenne*, les points qui représentent la taille moyenne des enfants dans chaque rangée seront sur la bissectrice.

Galton constata que, d'après sa table, les points moyens de bandes horizontales étaient bien à peu près sur une ligne droite passant par le centre des coordonnées, mais cette droite était moins inclinée sur la verticale que la bissectrice. Elle *régressait* vers l'axe sur lequel toutes les moyennes se fussent rassemblées si les âges des fils n'avaient généralement eu aucun rapport avec les âges des parents. C'est la loi de retour au type moyen.

L'observation de son tableau a révélé à Galton un autre fait intéressant, c'est que les cases qui contiennent à peu près le même nombre de couples semblent se distribuer sur des ellipses concentriques dont les axes principaux sont inclinés sur les axes des coordonnées, ce qui lui faisait pressentir une certaine loi de distribution des points représentatifs dans leur plan.

Galton était un grand naturaliste; il n'était pas mathématicien. Il ignorait qu'une quarantaine d'années auparavant le physicien français Bravais avait obtenu des résultats analogues en étudiant la répartition des erreurs de situation d'un point soit dans un plan soit dans l'espace (1). Ainsi la loi d'évolution des tailles présente quelque analogie avec la loi de distribution des points de chute d'un projectile, par exemple. C'est ainsi que Bravais a déterminé l'orientation des points au moyen du rapport $\frac{\Sigma xy}{\Sigma y^2}$ ou $\frac{\Sigma xy}{\sqrt{\Sigma x^2}\sqrt{\Sigma y^2}} \times \frac{\sqrt{\Sigma x^2}}{\sqrt{\Sigma y^2}}$ et qu'il a déterminé la surface dont les sections horizontales sont des ellipses analogues à celles qu'a trouvées Galton.

Voyons maintenant comment dans le cas particulier qu'ont considéré, indépendamment l'un de l'autre, Bravais et Galton, les deux coefficients de covariation donnent des indications concordantes.

L'un des caractères communs des deux espèces de grandeurs dont ils se sont occupés c'est que, dans chaque rangée horizontale, les points se distribuent à peu près symétriquement autour de leur moyenne. Il en est de même dans chaque colonne verticale.

Supposons d'abord que la distribution soit la même dans les deux directions; il est facile de voir qu'elle est encore la même dans une direction quelconque. De sorte que, quelle que soit la loi commune de distribution des points, les îlots de même surface qui contiennent un même nombre de points sont situés sur une circonférence de cercle tracée du centre O des coordonnées (fig. 6). Cela suppose que les coordonnées des points soient mesurées avec la même unité.

Si les ordonnées sont mesurées avec une autre unité que les abscisses, par exemple avec une unité plus grande égale à K fois l'unité d'abscisse la distribution des points se modifie, les ordonnées sont réduites dans le rapport de 1 à K. Les îlots de points également nombreux, qui étaient également nombreux sur la circonférence, se trouveront après la réduction également nom-

(1) *Mémoires de l'Institut*, tome IX. Analyse mathématique sur les erreurs de situation d'un point.

breux sur une ellipse dont les axes principaux coïncident avec les axes de coordonnées. Supposons que O A soit l'unité de mesure des abscisses, les diamètres principaux de l'ellipse sont, l'un égal à O A, l'autre $O B = \frac{O A}{K}$.

Considérons maintenant une parallèle quelconque à l'axe Ox. Sur cette parallèle les îlots de points également nombreux se répartissent symétriquement par rapport à l'axe oy ; les nombres changent quand on passe de l'une à l'autre des ellipses concentriques. S'il en est ainsi pour toutes les parallèles à $o\,x$, il n'existe aucune concordance entre les ordonnées et les abscisses des points ni entre leurs variations puisque, pour chaque valeur de y, les points sont aussi nombreux d'un côté que de l'autre : les concordances entre coordonnées de même signe sont aussi nombreuses que les discordances entre ordonnées de signe contraire.

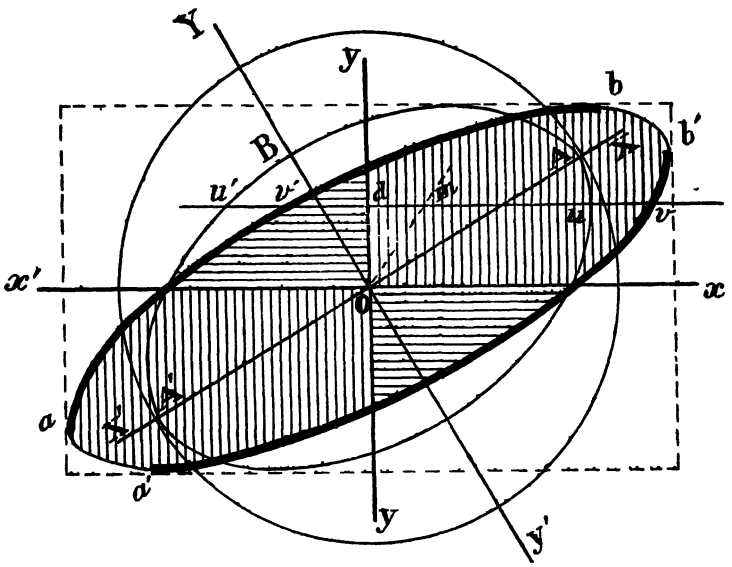

Fig 6

Tel sera à peu près le cas si, par exemple, on compare les tailles de nombreux individus d'âge donné, ou bien les sexes d'enfants nouveaux-nés, suivant les rangs qu'occupent dans l'alphabet les initiales de leurs noms de famille. Cependant il arrive souvent que les variations des deux éléments d'un couple ne sont point indépendantes : par exemple, la taille du père et celle du fils, l'âge du mari et celui de la femme, la production et le prix d'un objet, etc.

Lorsqu'il en est ainsi, les points représentatifs des couples d'éléments numériques ne sont plus symétriquement placés sur les parallèles à Ox. Ils sont déviés, le point moyen étant à une distance de l'axe $o\,y$ d'autant plus grande que la parallèle est plus éloignée de $o\,x$. Dans la table dressée par Galton, plus la taille des parents augmente, plus la taille moyenne des enfants s'éloigne de la moyenne générale. Dans ce cas les deux éléments du couple, *taille parentale*,

taille filiale varient à peu près proportionnellement. Supposons parfaite la proportionnalité, autrement dit, admettons que la régression soit linéaire (1).

Pour représenter cette situation nouvelle, nous emploierons un système d'axes rectangulaires *ox*, *oy* qui d'abord coïncide avec le précédent OX, OY puis tourne autour du centre O jusqu'à la position *ox*, *oy*.

Une parallèle à O*x* coupe l'ellipse A B A′ en deux points où les ilôts sont également nombreux mais qui ne sont pas symétriquement placés par rapport à O*y*. Pour que ces ilôts soient représentés symétriquement il faut les dévier parallèlement à O*x* d'une quantité égale à la distance du point moyen à l'axe *oy*. Par exemple, on a $uv = u'v' = \delta m$.

La figure obtenue au moyen des déviations analogues pour toutes les parallèles à O*x* est encore une ellipse. Celle-ci et l'ellipse primitive ABA′ ont mêmes tangentes parallèles à O*x*.

La série des ellipses concentriques telles que ABA′ est ainsi remplacée par une série d'ellipses concentriques déviées parallèlement à O*x* (2).

On obtient, d'ailleurs, une autre série d'ellipses analogues à la précédente en déviant l'ellipse primitive parallèlement à O*y*. Les diamètres conjugués des tangentes qui sont parallèles à O*x* dans la première série, parallèles à O*y* dans la seconde, sont les droites de régression de Galton.

L'une des ellipses de la première série, par exemple, coïncidait avec l'ellipse ABA′ avant la rotation et elle avait pour axes principaux OX OY. Alors, dans les quadrants opposés compris entre les deux directions, soit positives, soit négatives, des axes, chaque ordonnée d'un point pris dans un îlot à la circonférence de l'ellipse a le même signe que l'abscisse : le produit des deux coordonnées est donc toujours positif. Dans les deux autres quadrants le produit est toujours négatif. Et comme l'ellipse est symétrique par rapport aux deux axes, comme les îlots sont également nombreux sur tout le parcours de l'ellipse, la balance des produits négatifs et des produits positifs est nulle.

Or, dans les quadrants où les coordonnées des points sur la circonférence de l'ellipse ont le même signe, les variations de ces coordonnées entre deux points ont des signes contraires, quelle que soit l'amplitude de la variation. L'inverse a lieu dans les autres quadrants.

Il en résulte, pour le même raison de symétrie que tout à l'heure, que la balance des produits positifs et des produits négatifs est nulle aussi bien pour les variations des coordonnées que pour les coordonnées elles-mêmes. Entre points situés dans des quadrants différents, le même équilibre existe attendu qu'à tout segment de droite unissant deux points, correspond un segment identique et symétriquement placé par rapport à O*x*.

Entre des points *a* et *b* situés sur des ellipses différentes on peut passer d'abord d'une ellipse à l'autre sur un même rayon. Dans ce trajet la balance des concordances des coordonnées reste de même sens que la balance des variations. Puis le cheminement sur la seconde ellipse laisse encore aux deux balances le

---

(1) Dans ce cas le rapport de corrélation de Pearson équivant au coefficient *r*. On peut d'ailleurs imaginer d'autres caractéristiques par exemple, celles qui résulteraient de la substitution aux sommes de carrés de sommes d'éléments, ou de différences, comptés en valeurs absolues.

(2) Sur la détermination de ces ellipses. voir : Essai sur un mode d'exposer, etc. *Journal de la Société de Statistique*, 1910, p. 481.

même sens. Dans le cas particulier de l'ellipse ABA′, ces deux balances restent nulles.

Supposons maintenant qu'il existe une relation entre les faits que représentent les deux coordonnées d'un même point représentatif. C'est ainsi qu'en déterminant la position d'un point par des mesures indirectes les écarts de la position vraie, soit dans le sens Ox, soit dans le sens Oy, varient dans le même sens; il en est de même des écarts des points de chute d'un projectile, d'autant plus prononcés dans le sens perpendiculaire à la trajectoire que le point de chute est plus éloigné.

Or, la représentation des couples d'éléments ainsi liés, au moyen de points qui se répartissent également sur des ellipses concentriques, nous permet d'apercevoir la raison pour laquelle les variations des coordonnées donnent des concordances généralement de même sens que les coordonnées elles-mêmes.

En effet, la rotation des axes ox, oy transforme avons-nous vu, l'ellipse ABA′ en une ellipse dont les axes principaux sont inclinés sur les axes de référence. Cette ellipse est inscrite dans un rectangle formé par des tangentes parallèles à ces axes de référence, de même que l'ellipse ABA′. Mais, tandis que pour cette dernière les quatre points de contact partageaient la circonférence de l'ellipse en quatre parties égales, à mesure que l'ellipse s'aplatit et tourne avec les axes Ox, Oy les arcs compris entre les points de contact deviennent de plus en plus inégaux. Les arcs ab, $a'b'$ sont de plus en plus longs, les arcs aa', bb' de plus en plus courts.

Supposons la relation entre les coordonnées telle que les points moyens sur les parallèles à Ox s'orientent dans les quadrants à coordonnées de même sens xoy, $x'oy'$, suivant l'apparence que donne la figure. Dans la partie hachurée parallèlement à Oy, les coordonnées sont de même sens; elles sont de sens contraire dans la partie hachurée parallèlement à Ox. Mais cette fois il est visible que les portions d'ellipse à coordonnées de même sens sont plus grandes que les autres. Le nombre des concordances dépasse certainement celui des discordances.

D'autre part, entre deux points de contact a, b situés sur l'arc de la circonférence de l'ellipse l'ordonnée varie dans le même sens que l'abscisse, de même pour les points sur $a'b'$, tandis que les variations sont de sens contraire pour les points des arcs aa', bb'. Les premiers arcs étant plus longs que les seconds, la balance des variations est donc aussi en faveur des concordances.

Pour ce qui est de points situés sur des arcs différents ou sur des ellipses différentes, on peut répéter ce qui a été dit plus haut à propos de l'ellipse ABA′ et des points dont les coordonnées sont indépendantes.

En résumé, lorsque les éléments correspondants de deux séries sont tels qu'à un petit groupe d'éléments de l'une, ayant à peu près la même valeur, correspondent des éléments de l'autre qui soient répartis symétriquement autour de leur moyenne et que cette moyenne varie proportionnellement à d'élément correspondant de la première série, les variations tendancielles des éléments autour de leurs moyennes et les variations différentielles d'intervalle quelconque donnent des balances, entre concordances et discordances, qui sont le même signe. Et ainsi l'une des deux balances suffit pour affirmer la corrélation des deux séries, soit que les éléments de l'une dépendent directement des

éléments de l'autre, soit que tous deux dépendent de quelque autre série ou de plusieurs autres séries. La distinction des deux cas résulte de la comparaison des deux séries corrélatives après décalage de l'une d'elles par rapport à l'autre.

Lorsque la distribution des éléments de chaque série n'est pas symétrique, la répartition des moyennes des éléments de l'une, pour chaque valeur de l'élément correspondant dans l'autre, n'a plus lieu suivant une ligne droite. La ligne est plus ou moins compliquée et les propositions précédentes ne sont plus vraies. La corrélation ne peut guère être établie qu'en considérant successivement les variations tendancielles et les variations différentielles, au moins celles de première amplitude (1), ce double examen semblant devoir suffire dans la généralité des cas.

Cas particulier de variations ordonnées.

Il reste cependant un point à éclaircir, c'est celui des relations qui naissent d'un mouvement périodique ou oscillatoire, tel que dans l'exemple donné plus haut, le déplacement d'une roue. Considérons un point de la roue et notons ses positions au-dessous ou au-dessus du plan de l'essieu, ainsi que l'avancement de la roue, à chaque quart de tour.

Les points représentatifs des couples formés par la hauteur du point et le chemin que parcourt l'essieu en même temps se disposent dans un plan de la façon suivante. Portons sur oy les hauteurs du point de la roue, nous obtenons trois points p', o p tels que $op = op'$: ce sont les seules représentations de la hauteur. A mesure que la roue avance, la projection verticale du point représentatif des couples de grandeurs va de o en p puis revient en o, ensuite en p', retourneen o et ainsi de suite. Quant aux abscisses des points, à chaque fois que l'ordonnée repasse par la même valeur elle augmente d'une quantité égale à la longueur d'un quart de circonférence de la roue.

Les points représentatifs sont donc symétriques par rapport aux axes de coordonnées, les points moyens sur des parallèles à Oy sont situés sur l'axe Oy. Cette représentation semble confirmer la remarque faite plus haut à l'aide de la représentation par deux courbes. Les deux grandeurs : hauteur d'un point de la roue et déplacement de cette roue ne semblent avoir entre elles aucune relation puisqu'un changement de la hauteur donne des déplacements indifféremment situés au-dessus ou au-dessous de la moyenne, et de même, à un déplacement donné de la roue, correspondent des hauteurs indifféremment au delà ou en deçà de la moyenne.

Et cependant il n'est pas douteux que les mouvements sont corrélatifs. D'où vient cette contradiction?

Remarquons d'abord que l'analyse de Bravais suppose autre chose que la simple symétrie des points. Elle suppose une distribution conforme à la loi des erreurs.

Cependant, dans ce qui précède, nous n'avons point eu besoin de faire intervenir cette loi qui exprime simplement la distribution, dans le plan ou dans l'espace, des erreurs de situation.

(1) Dans son ouvrage *Forecasting the prices of Cotton*, le professeur Moore a considéré les variations tendancielles et les variations différentielles relatives.

Il est vrai que, dans la représentation du mouvement d'un point de la roue, les déplacements de celle-ci s'échelonnent à intervalles fixes, bien loin de se concentrer autour de leur moyenne comme des observations de visée. Mais on peut imaginer qu'au lieu de rester fixe par rapport à la roue, le point considéré soit mis en mouvement, indépendamment du mouvement de la roue, de façon que les déplacements de celle-ci correspondent à des hauteurs différentes déterminées mécaniquement, les déplacements observés étant plus fréquents pour de petites hauteurs que pour des grandes.

Les points représentatifs sembleraient alors distribués comme ceux qui représentent des erreurs dans le plan. Seulement, tandis que si l'on numérote les erreurs successives, les numéros ne manifestent aucun ordre, le numérotage des points représentatifs des positions de la roue ferait apparaître au contraire la succession régulière due à l'appareil mécanique.

Le calcul des coefficients de covariation des deux espèces et leur confrontation ne dispensent donc pas d'examiner les conditions des phénomènes comparés, de suivre leur évolution. Les observations précédentes engagent surtout à multiplier et à fractionner les observations de façon à observer les concordances à intervalles plus ou ou moins rapprochés.

L'exemple de la roue suggère une remarque générale. Nous savons que les écarts accidentels, tels que ceux que l'on constate sur une cible, suivent la loi de concentration bien connue, à laquelle toute expérience se conforme. Que l'on nous présente une cible sur laquelle les points d'impact sont distribués conformément à cette loi, en conclurons-nous que les écarts sont accidentels, imputables au hasard? Sans doute, mais à la condition toutefois que, les points étant numérotés d'après l'ordre de tir, les numéros ne se suivent pas régulièrement. S'ils se suivent dans un ordre évident, nous serons convaincus que le tir a été effectué par un appareil mécanique, bien combiné au moyen de leviers et de cames.

Les résultats de l'expérience doivent être interprétés avant que l'on apprécie, d'après eux et d'après les éléments numériques qui en découlent, les relations de cause à effet ou même de simples probabilités (1). Dans les faits complexes qu'étudie la statistique il est toujours prudent de prendre garde aux liaisons cachées.

Les difficultés, les applications douteuses, les contradictions même, que fait apparaître une théorie incomplète de la corrélation, ont frappé de nombreux esprits.

Après avoir étudié des relations de faits de toute nature avec une précision que la matière ne comporte pas (2), on a parfois reconnu qu'il y avait de fausses corrélations (3). D'autres fois on a jugé sans signification des corrélations indirectes qui, sans avoir la signification causale des corrélations directes, peuvent parfois suggérer l'étude de causes jusque-là ignorées ou mal connues. On a dit aussi qu'il y avait une différence fondamentale entre l'étude des séries

---

(1) Sur une connaissance imparfaite des causes on peut fonder des paris, mais on n'a pas le droit de calculer des probabilités (Cournot, *Théorie des chances*, p. 161).

(2) Voir une communication de Guldberg dans *Skandin. Actuarietids krift* sur l'écart probable du coefficient de corrélation.

(3) Différentes études de Karl Pearson sur les « spurious corrélations ».

dont les éléments se correspondent à un moment donné et l'étude des séries chronologiques (1).

Il semble qu'une théorie complète de la corrélation comprend, sous un même cadre logique, tous les cas suggérés par l'expérience, si l'on ne néglige pas d'analyser les faits complexes dans leurs qualités et propriétés en même temps qu'on les saisit numériquement par des indices assez nombreux, suffisamment variés, et qui sont plus ou moins instructifs suivant l'interprétation qu'autorise l'analyse des particularités dont on connaît les analogies avec des éléments déjà étudiés.

Dans les premiers des tableaux ci-après, on a indiqué les calculs qui permettent de déterminer un certain nombre de caractéristiques de la grandeur ou de la variabilité des éléments de la série des 26 salaires considérée plus haut.

Dans le dernier tableau, on a attribué arbitrairement des âges aux différents ouvriers, en vue d'illustrer le calcul des covariations, les ouvriers étant supposés rangés par ordre alphabétique sur la feuille de paie.

Lucien MARCH.

(1) Études de Tschuprow (*Grundbegriffe der Korrelationstheorie*), de Yule (*Journal of the Royal, Statistical Society*, janvier 1926), de Crum et Persons dans *The Review of economic statistics*, n° d'avril 1927, de Zinn dans la même revue en octobre 1927, de Slutsky dans le *Bulletin économique de l'Institut russe de conjoncture*, vol. III, n° 1, 1927.

APPENDICES

APPENDICE I. — Le tableau ci-après présente le calcul de di verses caractéristiques,

TABLEAU I. — *Salaires de 13.000 ouvriers métallurgiques des États-Unis en 1890, exprimé en francs et réduits*

N° d'ordre (1)	SALAIRES journaliers en francs Ordre croissant (2)	SALAIRES journaliers en francs Ordre décroissant (3)	ÉCARTS des salaires (4)	ÉCARTS des rangs (6)	PRODUIT des deux écarts (6)	ÉCARTS à la médiane en valeur absolue dessalaires (7)	ÉCARTS à la médiane en valeur absolue des rangs (8)	PRODUIT des deux écarts (9)	CARRÉS des écarts des rangs partie entière (10)	SALAIRES cumulés (11)	DIFFÉRENCES successives des salaires (12)	CARRÉS de ces différences (13)
1	4,	27,0	23,0	25	575,0	— 3,0	— 12,5	+ 38,75	144	4,0	—	—
2	4,5	20,0	15,5	23	356,5	— 2,6	— 11,5	29 90	121	8,5	0,5	0,25
3	5,0	15,0	10,0	21	210,0	— 2,1	— 10,5	22,05	100	13,5	0,5	0,25
4	5,5	12,0	6,5	19	123,5	— 1,6	— 9,5	15,20	81	19,0	0,5	0,25
5	5,8	10,0	4,2	17	71,4	— 1,3	— 8,5	11,05	54	24,8	0,3	0,09
6	5,9	9,0	3,1	15	46,5	— 1,2	— 7,5	9,00	49	30,7	0,1	0,01
7	6,0	8,6	2,6	13	33,8	— 1,1	— 6,5	7,15	36	36,7	0,1	0,01
8	6,1	8,2	2,1	11	23,1	— 1,0	— 5,5	5,50	25	42,8	0,1	0,01
9	6,2	8,0	1,8	9	16,2	— 0,9	— 4,5	4,05	16	49,0	0,1	0,01
10	6,4	7,8	1,4	7	9,8	— 0,7	— 3,5	2,45	9	55,4	0,2	0,04
11	6,6	7,6	1,0	5	5,0	— 0,5	— 2,5	1,25	4	62,0	0,2	0,04
12	6,8	7,4	0,6	3	1,8	— 0,3	— 1,5	0,45	1	68,8	0,2	0,04
13	7,0	7,2	0,2	1	0,2	— 0,1	— 0,5	0,05	0	75,8	0,2	0,04
14	7,2	7,0				+ 0,3	+ 0,5	0,05		83,0	0,2	0,04
15	7,4	6,8				+ 0,5	+ 1,5	0,45		90,4	0,2	0,04
16	7,6	6,6				+ 0,7	+ 2,5	1,25		98,0	0,2	0,04
17	7,8	6,4				+ 0,9	+ 3,5	2,45		105,8	0,2	0,04
18	8,0	6,2				+ 0,9	+ 4,5	4,05		112,8	0,2	0,04
19	8,2	6,1				+ 1,1	+ 5,5	6,05		122,0	0,2	0,04
20	8,6	6,0				+ 1,5	+ 6,5	9,75		130,6	0,4	0,16
21	9,0	5,9				+ 1,9	+ 7,5	14,25		139,6	0,4	0,16
22	10,0	5,8				+ 2,9	+ 8,5	24,65		149,6	1,0	1,00
23	12,0	5,5				+ 4,9	+ 9,5	46,55		161,6	2,0	4,00
24	15,0	5,0				+ 7,9	+ 10,5	82,95		176,6	3,0	9,00
25	20,0	4,5				+ 12,9	+ 11,5	148,35		196,6	5,0	25,00
26	27,0	4,0				+ 19,9	+ 12,5	248,75		223,6	7,0	49,00
Totaux	223,6	223,6	72,0	169	1 472,8	72,0	169,0	736,40	650		22,8	89,60

(1)

somme des rangs (parties entières) 78
carré de 0.5 0.25
somme des carrés 725.25

(1) $\frac{169}{2} = 84.5$ $\quad 84.5 - \frac{13}{2} = 78$

TABLEAU II — RÉSULTATS DU CALCUL DES CARACTÉRISTIQUES.

1° *Caractéristiques dépendant de l'ordre*

Médiane (col. 2) : $m = \frac{7.0. + 7.2}{2} = 7.1$

Médiale (col. 14 et 15) : $m' = 8{,}0 - \frac{113{,}8 - \frac{223{,}6}{2}}{113{,}8} = 8{,}0 - 0{,}002 = 7{,}998$

Déviation intermédiante : $\Delta' = m' - m = 0{,}9$; intermédiante relative : $\frac{0{,}9}{7{,}228} = 0{,}12$

1er quartil : Q1 = 6,0 (col. 2)

2e quartil : Q2 = 8,6 (col. 2)

Interquartil : I = Q2 − Q1 = 2,6 ; interquartil relatif $= \frac{2{,}6}{7{,}1} = 0{,}36$

1er quartal : $Q'1 = 6{,}4 + \left(\frac{6{,}6 - 6{,}4}{6{,}6}\right) \times \frac{223{,}6}{4} - 55{,}4 = 6{,}4 + 0{,}016 = 6{,}416$ (col. 2 et 11)

soit des grandeurs d'une série, soit des différences de ces grandeurs.

aux salairs de 26 ouvriers dont chacun recevrait le salaire moyen de 500 ouvriers de la série primitive.

INVERSES des salaires	CARRÉS des salaires	CUBES des salaires	LOGARITHMES des salaires	ÉCARTS à la moyenne		PRODUITS de ces écarts par ceux de la col. 8		CARRÉS des écarts	CUBES DES ÉCARTS	
				négatifs	positifs	négatifs	positifs		négatifs	positifs
(14)	(15)	(16)	(17)	(18)	(19)	(20)	(21)	(22)	(23)	(24)
0,250000	16,00	64.000	0,602060	4,6			57,50	21,16	97,336	
0,222222	20,25	91.125	0,653213	4,1			47,15	16,81	68,921	
0,200000	25,00	123.000	0,698970	3,6			37,80	12,96	46,656	
0,181818	30,25	166.375	0,740363	3,1			29,45	9,61	29,791	
0,172414	33,64	195.112	0,763428	2,8			23,80	7,84	21,952	
0,169491	34,81	205.379	0,770852	2,7			20,25	7,29	19,683	
0,166666	36,00	216.000	0,778551	2,6			16,90	6,76	17,576	
0,163934	37,21	226.981	0,785530	2,5			13,75	6,25	15,625	
0,161290	38,44	238.328	0,792392	2,4			10,80	5,76	13,824	
0,156250	40,96	262.144	0,806180	2,2			7,70	4,84	10,648	
0,151515	43,56	287.496	0,819544	2,0			5,00	4,00	8,000	
0,147059	46,24	314.432	0,832509	1,8			2,70	3,24	5,832	
0,142857	49,00	343.000	0,845098	1,6			0,80	2,56	4,096	
0,138889	51,84	373 248	0,857332	1,4		0,70		1,96	2,744	
0,135125	54,76	405.224	0,869232	1,2		1,80		1,44	1,728	
0,131579	57,76	448,975	0,880814	1,0		2,50		1,00	1,000	
0,128205	60,84	474.552	0,892095	0,8		2,80		0,64	0,512	
0,125000	64,00	512.000	0,903090	0,6		2,70		0,36	0,216	
0,121951	67,24	551.268	0,913814	0,4		2,20		0,16	0,064	
0,116279	73,96	636.056	0,934498	0,0		0,00	0,00	0,00	0,000	
0,111111	81,00	729.000	0,954243		0,4		3,00	0,16		0 064
0,100000	100,00	1.000.000	1,000000		1,4		11,90	1,96		2.744
0,883333	144,00	1.728.000	1,079181		3,4		32,30	11,56		39 304
0,076923	225,00	3.375 000	1,176691		6,4		67,20	40,96		262.144
0,066667	400,00	8.000 000	1,301030		11,4		131,10	129,96		1.481 544
0,037037	729,00	19 683.000	1,431364		18,4		230,00	338,56		6.229 504
3,667615	2.860,76	40.641.795	23,080474	41,4	41,4	12,70	749,10	637,80	366.204	8.015.304
				82.8		736,40			7.649,100	

2e quartal : $Q'2 = 12{,}0 + (15 - 12)\dfrac{3/4\ 223{,}6 - 161{,}6}{15} = 12{,}0 + 0{,}62 = 12{,}62$
(col. 2 et 11)

Interquartal : $1' = Q'2 - Q'1 = 6{,}2$; interquartal relatif $\dfrac{6{,}2}{7{,}998} = 0{,}77$

2° *Caractéristiques dépendant exclusivement des grandeurs*

Moyenne : $M = \dfrac{223{,}6}{26} = 8{,}6$
(col. 13)

Moyenne quadratique : $M2 = \dfrac{2560{,}75}{26} = 9{,}85$
(col. 15)

Moyenne cubique : $M3 = \dfrac{40641{,}79}{26} = 5{,}39$
(col. 16)

Moyenne harmonique : $H = \dfrac{26}{3.667} = 7{,}09$
(col. 14)

Moyenne logarithmique : $L = \frac{23,0805}{26} = 0,8877$
(col. 17)

Moyenne géométrique : G = No (log. 0.8877 = 7,773.

Écart moyen : $e = \frac{82.8}{26} = 3.18$
(col. 18 et 19)

Différence moyenne : $d = \frac{1472,8 \times 2}{26^2} \times \frac{736,4}{\overline{13}^2} = 4,36$
(col. 6)

Pente moyenne ou Orientation moyenne $p = \frac{1472,8}{4 \times 728,25} = 0,506 = d_x \frac{\overline{13}^2}{728,25}$
(col. 6 à 10 ou 20, 21)

Fluctuation : $\mu_2 = \frac{637,8}{26}$ 24.5
(col. 22)

Écart type : $\sigma = \sqrt{\mu_2}$ = 4.95
(col 22)

Dispension : $\delta = \sigma \sqrt{2} = 5.83$

Limites de la déviation intermoyenne : $\Delta = \frac{24.5}{8,6} = 2.85$ et $\frac{7649,1}{637,8} = 9,1$
(col. 23, 24 et 22)

9,1 dépassant 8,36, on adopte pour Δ la valeur 2,85.

Dominante : D = M — Δ = 8,6 — 2,85 = 5,75

Contrôle des calculs

$2560,76 = 637,80 - 26 \times \overline{8,6}^2 = 637,8 - 26 \times 73,96$
(col. 15) (col. *b* 22)

$40.6417,96 = 7649,1 + 3 \times 8,6 \times 2560,76 - 3 \times \overline{8,6}^2 \times 223,6 + 26 \times 8,6^3$
(col. 16) (col. 23-24) (col. 15) (col. 13)

$= 7649,1 + 66067,608 - 49612.368 + 16537,456$
(col. 23-24)

APPENDICE II. — Dans le tableau ci-après on a réuni la série des salaires portés dans le tableau précédent à une série d'âges supposés des ouvriers qui gagnent ces salaires. Les ouvriers ont été supposés sériés dans l'ordre alphabétique, d'après lequel ils seraient rangés sur la feuille de paie.

Initiale	Salaires journaliers	Ages en années	VARIATIONS DIFFÉRENTIELLES des salaires +	des salaires —	des âges +	des âges —	PRODUITS DES variations différentielles +	—	CARRÉS DES variations différentielles des salaires	des âges	VARIATIONS TENDANCIELLES (écarts à la moyenne) des salaires +	des salaires —	des âges +	des âges —	PRODUITS des variations tendantielles +	—	CARRÉS des variatio s des âges
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)
A	5,8	19	—	—	—	—	—	—	—	—		2,8		12	33,6		144
B	10,0	55	4,2		36		155,2		17,64	1.266	1,4		24		33,6		576
C	4,0	15		6,0		40	240,0		36,00	1.600		4,6		16	73,6		256
D	5,9	18	1,9		3		5,7		3,61	9		2,7		13	35,1		169
E	6,6	21	0,7		3		2,1		0,49	9		2,0		10	20,0		100
F	8,0	43	1,4		22		30,8		1,96	484		0,6	12			7,2	144
G	8,6	34	0,6			9			0,36	81		0,0	3			0,0	9
H	8,2	30		0,4		4	1,6	5,4	0,16	16		0,4		1	0,4		1
I	7,2	22		1,0		8	8,0		1,00	64		1,4		9	12,6		81
J	9,0	36	1,8		14		25,2		3,24	196	0,4		5		2,0		25
K	27,0	33	8,0			3		5,4	324,00	9	18,4		2		36,8		4
L	7,4	26		19,6		7	137,2		384,16	49		1,2		5	6,0		25
M	4,5	53		12,9	27			78,3	8,41	729		4,1	22		90,2		484
N	6,8	24	2,3			29		66,7	5,29	841		1,8		7	12,6		49
O	12,0	45	5,2		21		109,2		27,04	441	3,4		14		47,6		196
P	5,5	16		6,5		29	188,5		42,25	841		3,1		15	46,5		225
Q	7,6	31	2,1		15		31,5		4,41	225		1,0		0	0,0		0
R	7,0	23		0,6		8	4,8		0,36	64		1,6		8	12,8		64
S	6,0	40		1,0	17			17,0	1,00	289		2,6	9			23,4	81
T	20,0	35	14,0			5		7,0	1,96	25	11,4		4		45,6		16
U	6,1	44		13,9	9			125,1	193,21	81		2,5	13			32,5	169
V	15,0	42	8,9			2		17,8	79,21	4	6,4		11		70,4		121
W	6,4	26		8,6		16	137,6		73,96	256		2,2		5	11,0		25
X	6,2	20		0,2		6	1,2		0,04	36		2,4		11	26,4		121
Y	7,8	38	1,6		18		28,8		2,56	324		0,8	7			5,6	49
Z	5,0	17		2,8		21	58,8		7,84	441		3,6		14	50,6		196
Totaux . .	223,6	806	62,7	63,5	185	187	1.166,2	371,3	1.220,16	8 410	141,4	141,4	120	126	667,8	68,7	3.230
Moyennes .	8,6	31	(17 +)	(11 —)													

A l'aide des résultats calculés dans le tableau ci-dessus et dans le précédent, on calcule les caractéristiques suivantes de la relation qui paraît exister entre les salaires des ouvriers et leurs âges.

Indice simple de covariation différentielle : $i = \frac{17 - 8}{25} = 0{,}36$

Indice pondéré de covariation différentielle : $I = \frac{1165{,}2 - 371{,}3}{1536{,}5} = 0{,}52$

Coefficient de covariation différentielle : $j = \frac{793{,}9}{\sqrt{1220{,}16 \times 8410}} = 0{,}25$

Coefficient de covariation tendancielle : $r = \frac{598{,}7}{\sqrt{637.8} \times \sqrt{3130}} = 0.7$