JOURNAL DE LA SOCIÉTÉ STATISTIQUE DE PARIS

LUCIEN MARCH

Essai sur un mode d'exposer les principaux éléments de la théorie statistique

Journal de la société statistique de Paris, tome 51 (1910), p. 447-486

<http://www.numdam.org/item?id=JSFS_1910__51__447_0>

III

ESSAI

SUR UN

MODE D'EXPOSER LES PRINCIPAUX ÉLÉMENTS DE LA THÉORIE STATISTIQUE

1. — A travers la diversité des objets et des phénomènes de la nature, l'observation humaine note des uniformités, des régularités. Le raisonnement relie entre elles les observations, et ses conclusions sont d'autant mieux dégagées des influences personnelles à l'observateur qu'elles portent sur des grandeurs mesurables et qu'elles découlent d'une plus vaste expérience.

Les grandeurs dont on se sert pour comparer un tout à quelque partie ou à quelque objet similaire, peuvent se classer en deux catégories, suivant que la comparaison s'effectue en bloc par l'emploi d'un instrument de mesure approprié, ou suivant qu'elle exige un dénombrement.

Quand l'étude quantitative porte sur un dénombrement de collectivités de faits objectifs ou de représentations mentales, elle est du ressort de la statistique.

Celle-ci comporte une discipline particulière, appropriée à l'observation et à la comparaison des faits par masses ; on la caractériserait convenablement, à cet égard, par le terme *pléthométrie*. Elle comprend en effet les propositions sur quoi doit s'appuyer le statisticien quand il ordonne et mesure des collections de faits.

2. — Cette discipline a pour but principal de réduire les rapports à leur juste valeur. Les raisonnements qu'elle gouverne portent d'ailleurs sur les faits actuels, existants, en quoi elle constitue la partie positive de la statistique. Les généralisations et les hypothèses sur lesquelles se fondent ces généralisations forment le domaine conjectural.

S'il est vrai que toutes les sciences ont des parties positives et des parties conjecturales, on doit reconnaître que, dans les applications de la méthode statistique, le territoire où s'élaborent les investigations conjecturales est plus vaste que le modeste champ où se précise la technique. Mais ce champ fait la valeur de tout le territoire.

3. — Les règles de l'analyse statistique des faits naturels sont généralement empruntées à la théorie des probabilités. J'ai signalé l'inconvénient qu'offre, dans la statistique (1), l'emploi du terme « probabilité ». La partie positive de la statistique

(1) Voir *Journal de la Société de Statistique de Paris,* numéro de septembre 1908, d'après le IV[e] Congrès des sciences mathématiques.

considère des séries d'événements, des rapports, des séquences, des fréquences. Sur ces constatations, ramenées, comme nous l'avons dit, à une juste mesure, nous pouvons fonder des raisons de croire à des éventualités. Mais ces raisons sont presque toujours insuffisantes; d'autres doivent intervenir, fondées sur des données qualitatives ou sur des hypothèses plausibles, et celles-ci ne sont guère susceptibles de mesure.

Quand on extrait d'une urne des boules blanches et des boules noires identiques, sauf la couleur, en proportion donnée, on n'a aucune autre raison de fixer la composition de l'urne que le rapport fourni par l'expérience et précisé par le calcul. Quand on a observé les décès d'une population de même âge, le coefficient de mortalité qui en découle n'a point une valeur comparable à celle que fait apparaître le tirage des boules.

Dans les faits qu'étudie la statistique, toute vue générale implique l'hypothèse que des changements survenus dans les circonstances nombreuses qui gouvernent les phénomènes, la plupart sont assez faibles, ou se compensent suffisamment, pour ne point altérer sensiblement le résultat d'une observation éclairée. Aussi, la confiance dans les conclusions d'observations en apparence analogues à celles du tirage des boules est-elle d'une autre nature : dans un cas on sait que, parmi les influences qui gouvernent l'observation, une circonstance connue intervient d'une façon extrêmement prépondérante, tandis que dans l'autre, ces influences sont souvent mal connues et sont d'ailleurs capables d'interférence.

La statistique aide à discerner l'importance relative de ces influences; afin de conserver une entière liberté d'appréciation elle doit éviter des expressions fondées sur des analogies incertaines et qui peuvent donner une idée inexacte de la valeur des rapports calculés.

4. — La statistique peut d'autant mieux se maintenir sur le terrain qui lui est propre que le principe de compensation, sur lequel est basée toute comparaison de faits collectifs, n'est que la généralisation de la formule d'un marché équitable que les hommes ont sans doute adoptée depuis qu'ils vivent en société.

La règle de la moyenne qui, du point de vue logique, est fondée sur une extension du principe de raison suffisante, traduit cette formule; elle suffit à la rigueur pour justifier les conventions admises dans les comparaisons statistiques, si l'on y ajoute les réserves que comporte l'infirmité de nos moyens d'observation.

Les principes fondamentaux de la logique, le sentiment de l'équité, celui de la faiblesse des capacités d'observations, tels sont les appuis sur lesquels doit se guider le jugement lorsqu'il vise à ordonner les constatations en les simplifiant; nous allons essayer de présenter, très succinctement, le mécanisme des opérations principales sous une forme simple qui laisse apparaître aussi complètement que possible les hypothèses et les conventions, dans la pensée que d'autres rectifieront ou amélioreront cet essai.

I. MOYENNE

5. — La règle de la moyenne à laquelle nous avons fait allusion plus haut peut être illustrée par des exemples tels que le suivant :

J'achète un hectolitre de blé; le marchand mesure et me remet la quantité me-

surée. Après vérification soigneuse, la quantité livrée ne me paraît mesurer en réalité que 99 litres 2 décilitres. Si cette quantité me suffit et que, lors d'un second achat, je trouve 1 hectolitre 8 décilitres, je ne puis reprocher au marchand un gain illicite, puisque l'excédent sur le second achat compense la perte sur le premier. Le marché serait encore équitable si, après m'avoir donné une première fois 99 litres au lieu de cent, le marchand me livrait deux autres fois 100 litres 1/2 au lieu de 100.

La condition du marché équitable est que le *total des quantités livrées soit égal au total des quantités demandées*, quelles que soient les valeurs des livraisons partielles.

6. — En appelant m la quantité demandée à chaque fois, $x_1, x_2 \dots x_i \dots x_n$ les quantités livrées successivement, on doit avoir :

$$x_1 + x_2 + \dots + x_i + \dots + x_n = \underbrace{m + m \dots + m}_{n} = m \times n$$

ce que l'on peut écrire

$$\Sigma x_i = mn$$

d'où l'on déduit

$$\Sigma (x_i - m) = 0 \qquad (1)$$

La moyenne arithmétique, ou plus brièvement la *moyenne* de n quantités est *la grandeur qui, répétée* n *fois, fournit le total des* n *quantités*.

Si certaines quantités se répètent elles-mêmes, par exemple si la quantité x_i est répétée n_i fois, la règle de la moyenne s'exprime alors par la relation

$$\Sigma n_i x_i = m \Sigma n_i = mn \qquad (2)$$

On remarque l'analogie de cette formule avec celle des moments des forces parallèles ou des poids : la règle énoncée par Côtes exprime cette analogie.

7. — La mesure x_1, prise dans la première opération, peut être regardée comme l'une des quantités

$$x_{1,1} \qquad x_{1,2} \dots \qquad x_{n,1}$$

sur lesquelles le marchand, mal servi par ses instruments, eût pu tomber. Ces quantités sont nécessairement de grandeurs limitées, mais elles peuvent être en nombre quelconque. De même, la seconde mesure peut être regardée comme extraite d'une seconde distribution de quantités

$$x_{1,2} \qquad x_{2,2} \dots \qquad x_{n,2}$$

et ainsi de suite.

D'une manière générale, considérons le groupe des n distributions primaires suivantes comprenant $n_1, n_2 \, n_h \, n_n$ termes :

$$\begin{array}{llllll}
x_{1,1} & x_{2,1} & \dots & x_{i,1} & \dots & x_{n_1} \text{ dont la moyenne est } m_1 \\
x_{1,2} & x_{2,2} & \dots & x_{i,2} & \dots & x_{n_{2,2}} \quad — \quad m_2 \\
\dots & \dots & \dots & \dots & \dots & \dots \\
x_{1,h} & x_{2,h} & \dots & x_{i,h} & \dots & x_{n_h,h} \quad — \quad m_h \\
\dots & \dots & \dots & \dots & \dots & \dots \\
x_{1,n} & x_{2,n} & \dots & x_{i,n} & \dots & x_{n_n,n} \quad — \quad m_n
\end{array}$$

Les quantités x sont quelconques, elles peuvent être différentes ou se répéter. Soient de plus

$$n_1 + n_2 + \ldots + n_h + \ldots + n_u = N$$

On peut écrire, pour la distribution de rang h

$$\Sigma x_{i,h} = m_h n_h$$

et en ajoutant toutes les égalités semblables

$$\Sigma x_{i,h} = \Sigma m_h n_h$$

Désignons par M la moyenne des quantités $m_1, m_2 \ldots m_h \ldots m_u$, chacune d'elles étant affectée d'un poids égal au nombre des grandeurs qui entrent dans sa composition, M sera définie par l'égalité

$$\Sigma m_h n_h = M \Sigma n_h = MN$$

On peut donc aussi écrire

$$(3) \qquad \Sigma x_{i,h} = MN \qquad \text{et} \qquad \Sigma (x_{i,h} - M) = 0$$

Si les moyennes partielles sont égales, alors $M = m_1 = m_2 = \ldots = m_u$

Ces relations peuvent être traduites ainsi :

La moyenne générale des termes d'un groupe de distributions, dont les moyennes sont différentes, est égale à la moyenne de ces moyennes partielles, pourvu que chacune d'elles soit affectée d'un poids égal au nombre des grandeurs qui entrent dans sa composition.

La moyenne proprement dite ne saurait être calculée correctement sans tenir compte des poids des éléments qui la composent.

Déterminée conformément à la règle précédente, elle fournit un instrument nécessaire de la comparaison numérique de deux groupes de grandeurs, parce qu'elle permet de comparer les grandeurs indépendamment de leur nombre.

La moyenne jouit de quelques propriétés importantes.

8. — On a vu (**6**) que si l'on détermine les écarts des grandeurs qui composent une distribution à partir de la moyenne de ces grandeurs, le total de ces écarts pris avec leurs signes est nul, ce que nous avons écrit $\Sigma (x_i - m) = 0$. Les règles du calcul algébrique empêchent donc d'utiliser la moyenne de ces écarts pour apprécier leur importance. Afin d'éviter l'influence des changements de signe, on substitue à l'appréciation de la grandeur des écarts l'appréciation de la grandeur de leurs carrés et l'on détermine la moyenne de ces carrés. Pour simplifier le langage, nous appellerons, avec Edgeworth, *fluctuation* cette moyenne de carrés et nous la représenterons par $\mu^2 = \dfrac{\Sigma (x_i - m)^2}{n}$.

Désignons par μ'^2 la moyenne des grandeurs qui composent la distribution considérée au paragraphe **6**. De l'identité

$$(4) \qquad \Sigma (x_i - m)^2 = \Sigma x_i^2 - 2m \Sigma x_i + nm^2 = \Sigma x_i^2 - nm^2$$

on déduit

$$\Sigma x_i^2 = n (\mu^2 + m^2)$$

ou (5)

$$\mu'^2 = \mu^2 + m^2$$

La moyenne des carrés de grandeurs associées est égale à la fluctuation de ces grandeurs augmentée du carré de leur moyenne.

9. — D'après l'égalité précédente, μ' atteint sa plus petite valeur quand $m = 0$, c'est-à-dire quand les grandeurs sont mesurées à partir de leur moyenne. Ainsi :

Les carrés des écarts des grandeurs, à partir d'une certaine valeur prise pour origine des écarts, donnent la moindre somme quand l'origine se confond avec la moyenne.

Cette propriété assure à la moyenne une position unique dans la distribution des grandeurs et engage à la choisir comme origine des mesures quand on étudie leur mode de distribution.

10. — Si, dans l'égalité (4), on suppose un changement d'origine, par exemple si l'on suppose que, dans cette égalité, l'origine des grandeurs a été prise en un point situé à une distance m' d'une autre origine, l'égalité (4) se transforme en celle-ci :

$$(6)\quad \left\{ \text{d'où} \right. \qquad \Sigma (x_i - m)^2 = \Sigma (x_i - m')^2 - n (m - m')^2$$

$$\frac{\Sigma (x_i - m)^2}{n} = \frac{\Sigma (x_i - m')^2}{n} - (m - m')^2$$

La fluctuation est égale au carré moyen des écarts, par rapport à une base quelconque, diminué du carré de l'intervalle entre cette base et la moyenne.

11. — Lorsqu'on groupe plusieurs distributions (7), M étant la moyenne générale des grandeurs qui entrent dans le groupe, on a pour la distribution de rang h :

$$\Sigma (x_{i,h} - m_h)^2 = \Sigma (x_{i,h} - \mathrm{M})^2 - n_h (m_h - \mathrm{M})^2$$

d'où

$$\Sigma (x_{i,h} - \mathrm{M})^2 = \Sigma (x_{i,h} - m_h)^2 + n_h (m_h - \mathrm{M})^2$$

Étendons la sommation à l'ensemble des distributions du groupe et nous pourrons écrire

$$(7)\qquad \Sigma (x_{i,h} - \mathrm{M})^2 = \Sigma (x_{i,h} - m)^2 + \Sigma n_h (m_h - \mathrm{M})^2$$

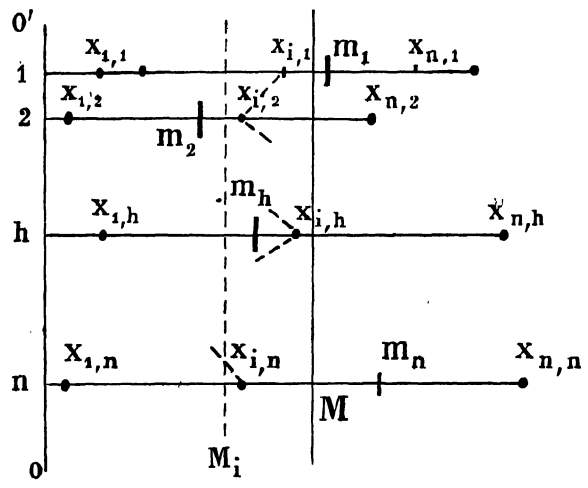
En divisant tous les termes par $\Sigma n_h = \mathrm{N}$, on voit que :

La fluctuation des grandeurs du groupe, par rapport à la moyenne générale, est égale ~~au total~~ à la moyenne des fluctuations des grandeurs autour de leurs moyennes partielles, augmenté de la fluctuation des moyennes partielles, pourvu que dans le calcul de ~~cette dernière~~ le carré de chaque fluctuation et (le carré de l'écart de chaque moyenne partielle, à partir de la moyenne générale, soit affecté d'un poids égal au nombre des grandeurs qui ~~la~~ composent la moyenne particelle

Si les moyennes partielles sont toutes égales, la fluctuation des grandeurs du groupe autour de la moyenne générale est alors égale à la ~~somme~~ moyenne barycentrique des fluctuations autour des moyennes partielles.

12. — Les propositions énoncées dans les paragraphes **7** et **11** sont susceptibles d'une importante généralisation qui a conduit Bienaymé à formuler en 1853 deux principes de conservation.

Appelons distributions *primaires* les distributions considérées au paragraphe 7 et, pour plus de commodité, représentons-les sur des droites perpendiculaires à un même axe oo' pris pour origine des x, les grandeurs étant représentées par la lettre x affectée d'indices différents suivant la place et suivant la distribution.



Supposons maintenant une distribution *secondaire* de termes formés en prenant une grandeur dans chacune des distributions primaires, et soient par exemple

$$x_{i,1}\ x_{i,2} \ldots\ x_{i,h} \ldots\ x_{i,n}$$

les n grandeurs associées dans le terme secondaire représenté sur la figure par un trajet pointillé à travers les distributions primaires. (Pour plus de commodité dans la suite, la lettre i représente ici des places différentes dans les diverses distributions primaires.)

Ces n grandeurs ont pour moyenne M_i, que nous appellerons une *moyenne secondaire,* et pour somme nM_i. On peut former autant de termes secondaires de ce genre qu'il y a de manières de permuter les $n_1, n_2 \ldots n_h \ldots n_n$ grandeurs des n distributions primaires, soit un nombre exprimé par le produit $n_1\, n_2 \ldots n_h \ldots n_n = \nu$.

Désignons par M la moyenne des moyennes secondaires telles que M_i; toutes ces moyennes secondaires étant composées du même nombre de grandeurs n, on a ;

$$\Sigma n M_i = \nu M n \qquad \text{ou} \qquad \Sigma M_i = \nu M$$

Or

$$nM_i = x_{i,1} + x_{i,2} + \ldots + x_{i,h} + \ldots + x_{i,n}$$

En formant toutes les sommes semblables,

$$x_{i,1} \text{ s'introduit } n_2 n_3 \ldots n_h \ldots n_n = \frac{\nu}{n_1} \text{ fois}$$

$$x_{i,2} \quad — \quad n_1 n_3 \ldots n_h \ldots n_n = \frac{\nu}{n_2} \text{ fois}$$

et ainsi de suite.

Donc on peut écrire :

$$\Sigma n M_i = \nu \Sigma \left[\frac{x_{i,1}}{n_1} + \frac{x_{i,2}}{n_2} + \ldots + \frac{x_{i,h}}{n_h} + \ldots + \frac{x_{i,n}}{n_n} \right]$$

et, par suite, comme $\Sigma n\, M_i = \nu\, Mn$.

$$nM = \Sigma \left[\frac{x_{i,1}}{n_1} + \frac{x_{i,2}}{n_2} + \ldots + \frac{x_{i,h}}{n_h} + \ldots + \frac{x_{i,n}}{n_n} \right]$$

d'où

$$M = \frac{m_1 + m_2 + \ldots + m_n}{n} \tag{8}$$

si $m_1, m_2 \ldots m_n$ sont les moyennes des distributions primaires.

M, moyenne des moyennes secondaires, est donc identique à la moyenne des moyennes primaires, toutes celles-ci entrant avec le même poids, ce qui différencie en général M de la moyenne générale calculée au paragraphe **7**.

La moyenne M se confond avec la moyenne générale des termes des distributions primaires si toutes ces distributions sont composées du même nombre de termes.

On peut alors énoncer l'égalité précédente sous la forme que voici : *Dans le développement de distributions primaires, comprenant toutes le même nombre de termes, en distribution secondaire, la moyenne se conserve.* Cette propriété est une conséquence de la symétrie du mode de formation des distributions secondaires.

Si l'on suppose de plus que les distributions primaires successives du groupe ont même moyenne m, alors $M = m$; dans ce cas, la moyenne des termes de la distribution secondaire est égale à la moyenne commune des distributions primaires.

Exemple : Soit e groupe des distributions primaires des grandeurs suivantes :

1	2	3
4	6	8
5	7	9

La moyenne de ces termes est 5.

Les moyennes des termes secondaires sont les quotients par. 3 des 27 sommes suivantes obtenues en prenant de toutes les manières possibles un terme dans chaque rangée :

10	11	12	13	14	15	16	17	18	19	20
		12	13	14	15	16	17	18		
		12		14	15	16		18		
				14		16				
				14		16				

Le total de ces sommes donne 405 et comme il y en a 27, la somme moyenne est 15, dont le tiers est bien égal à 5.

13. — La moyenne arithmétique n'est point la seule expression qui se conserve dans la transformation ; il en est encore de même d'un certain sous-multiple de la fluctuation.

Reprenons le groupe de distributions primaires figuré dans le paragraphe précédent. Les divers trajets qui unissent les points des distributions 1, 2 ... n ont comme lignes moyennes des droites telles que Mi et la moyenne M de ces moyennes secondaires est aussi la moyenne des moyennes primaires m_1, m_2, ... m_h ...

Considérons la différence entre la somme des termes $x_{i,1}$, $x_{i,2} \ldots x_{i,h} \ldots x_{i,n}$ de la distribution secondaire de rang i, dont la moyenne est Mi, et la somme des moyennes primaires $m_1 + m_2 + \ldots + m_h + \ldots + m_n$, laquelle est égale à nM (**12**).

Le carré de cette différence peut s'écrire

$$[(x_{i,1} - m_1) + (x_{i,2} - m_2) + \ldots (x_{i,h} - m_h) + \ldots + (x_{i,n} - m_n)]^2$$

Ajoutons tous les carrés semblables formés pour les termes secondaires, puis décomposons chaque carré en une somme de carrés tels que $(x_{i,h} - m_h)^2$ et une somme de doubles produits tels que $2\,(x_{i,h} - m_h)\,(x_{i,k} - m_k)$.

La somme de tous ces doubles produits est nulle, puisque $\Sigma(x_{i,k} - m_k) = 0$ pour toute valeur de $x_{i,k} - m_k$.

On peut donc écrire

$$\Sigma [x_{i1} + x_{i,2} + \ldots + x_{i,k} + \ldots + x_{i,n} - n\mathrm{M}]^2 = \Sigma (x_{i,k} - m_k)^2$$

où M représente toujours la moyenne de $m_1, m_2, \ldots m_k, \ldots m_n$.

Il faut maintenant se rendre compte du nombre des termes qui entrent sous chacun des signes Σ. Puisque la somme $x_{i,1} + x_{i.2} + \ldots + x_{i,k} + \ldots + x_{i,n}$ est composée de termes associés de toutes les manières possibles, le nombre des sommes semblables est (**12**).

$$n_1 \times n_2 \times \ldots \times n_k \times \ldots \times n_n = \nu$$

Dans le second membre de l'égalité précédente, si nous étendons d'abord le signe Σ aux seuls termes d'une distribution primaire de rang k, dont le nombre est n_k, chaque somme se trouvera répétée un nombre de fois égal à $\frac{\nu}{n_k}$; on peut donc mettre le second membre sous la forme plus précise

$$\Sigma \frac{\nu}{n_k} \Sigma (x_{i,k} - m_k)^2 \qquad \text{ou} \qquad \nu\Sigma \left(\frac{\Sigma (x_{i,k} - m_k)^2}{n_k} \right)$$

Dès lors en divisant les deux membres de l'égalité par ν, on a :

$$\frac{1}{\nu} \sum_1^{\nu} [(x_{i,1} + x_{i,2} + \ldots + x_{i,k} + \ldots + x_{i,n}) - n\mathrm{M}]^2 = \sum_1^{n} \left(\frac{\sum_1^{n_k} (x_{i,k} - m_k)^2}{n_k} \right)$$

M*i* étant la moyenne secondaire de rang i, c'est-à-dire la moyenne des quantités qui entrent dans le terme secondaire du rang i,

$$x_{i,1} + x_{i,2} + \ldots + x_{i,k} + \ldots + x_i n_k = n\mathrm{M}_i$$

L'égalité précédente peut alors s'écrire

$$(9) \qquad \frac{n^2}{\nu} \sum_1^{\nu} (\mathrm{M}_i - \mathrm{M})^2 = \sum_1^{n} \left[\frac{\sum_1^{n_k} (x_{i,k} - m_k)^2}{n_k} \right]$$

Désignons par μ'^2_k la fluctuation des termes de la distribution primaire de rang k, par μ^2 la fluctuation des moyennes des termes secondaires; l'égalité précédente prend la forme

$$(10) \qquad \mu^2 = \frac{1}{n} \frac{\Sigma \mu'^2_k}{n}$$

Ainsi les moyennes des termes secondaires ont pour fluctuation la n^e partie de la moyenne des fluctuations primaires, ou encore : *Dans la transformation des* n *distributions primaires en distribution secondaire, le rapport entre la moyenne des fluctuations primaires et la fluctuation des moyennes secondaires est égal au nombre des distributions primaires.*

14. — Dans le cas où les distributions primaires comprennent toutes le même nombre de termes et où elles ont toutes même fluctuation μ'^2, on peut remplacer $\frac{\Sigma_h \mu'^2}{n}$ par μ'^2, l'égalité (10) prend alors la forme

$$(10^{bis}) \qquad \mu^2 = \frac{\mu'^2}{n}$$

Il en est ainsi, notamment si toutes les distributions primaires sont identiques. *Le rapport entre la fluctuation commune des distributions primaires et la fluctuation des moyennes secondaires est constamment égal au nombre des distributions primaires associées.* Reprenons l'exemple du paragraphe **12.**

Les moyennes secondaires étant les quotients par 3 des nombres donnés au paragraphe **12,** les carrés des écarts entre ces moyennes et leur moyenne 5 seront les quotients par 9 des carrés des écarts entre les nombres inscrits au paragraphe **12** et leur moyenne 15. Ces écarts sont, avec leurs signes, dans l'ordre déjà adopté :

— 5	— 4	— 3	— 2	— 1	0	+ 1	+ 2	+ 3	+ 4	+ 5
		— 3	— 2	— 1	0	+ 1	+ 2	+ 3		
		— 3		— 1	0	+ 1		+ 3		
				— 1		+ 1				
				— 1		+ 1				

leurs carrés ont pour somme 81×2, dont le quotient par 9 est 18. Le carré moyen des écarts des moyennes secondaires, dont le nombre est 27, est égal à $\frac{18}{27} = \frac{2}{3}$.

D'autre part les distributions primaires ont respectivement pour fluctuations $\frac{1+1}{3}, \frac{4+4}{3}, \frac{4+4}{3}$ dont la moyenne est 2. Le rapport entre 2 et $\frac{2}{3}$ est égal à 3, nombre des distributions primaires.

15. — Considérons maintenant s groupes de distributions primaires semblables à celui des paragraphes précédents.

Désignons par M' la moyenne des moyennes telles que M calculées pour les différents groupes. Chacune de ces moyennes M est, à la fois, la moyenne des moyennes primaires telles que m_h et la moyenne des moyennes secondaires telles que M_i.

D'après le paragraphe précédent, égalité 9, on a, dans un groupe quelconque :

$$\frac{\sum_1^{\nu} (M_i - M)^2}{\nu} = \frac{1}{n^2} \sum_1^n \left(\frac{\sum_1^{n_h} (x_{i,h} - M)^2}{n_h} \right)$$

que l'on peut écrire (**10**)

$$\frac{1}{\nu} \left[\sum_1^{\nu} (M_i - M')^2 - \nu (M - M')^2 \right] = \frac{1}{n^2} \left[\sum_1^n \frac{\sum_1^{n_h} (x_{i,h} - M')^2}{n_h} - n (M - M')^2 \right]$$

D'où l'on conclut

$$\frac{\sum_1^{\nu} (M_i - M')^2}{\nu} = \frac{1}{n} \sum_1^n \frac{\sum_1^{n_h} (x_{i,h} - M')^2}{n n_h} + \frac{n-1}{n} (M - M')^2$$

Lorsque, dans chaque groupe, les distributions primaires comprennent toutes le même nombre de termes, n_h étant constamment égal à n', le premier terme du second membre s'écrit simplement :

$$\frac{1}{n}\frac{\sum_1^{n'}(x_{i,h}-M')^2}{n'}$$

16. — Supposons que tous les groupes de distributions primaires comprennent le même nombre n de distributions. Désignons par μ'^2 la fluctuation moyenne, $\frac{\Sigma(x_i-M')^2}{nn's}$, des distributions primaires de l'ensemble des groupes, par μ^2 la fluctuation des moyennes secondaires $\frac{\Sigma(M_i-M')^2}{\nu}$, par π^2 la fluctuation des moyennes M des groupes autour de la moyenne générale des s groupes. Sommons les s égalités que l'on peut écrire, semblables à celle du paragraphe précédent après substitution des notations nouvelles ; il viendra :

$$\sum_1^s \mu^2 = \frac{1}{n}s\,\mu'^2 + s\frac{n-1}{n}\pi^2$$

Si M est constamment égal à M', tous les groupes ayant même moyenne, alors π^2 est nul et l'égalité précédente se réduit à celle du paragraphe **13**.

Dans le cas général, en divisant par le nombre s des groupes, on a :

$$\frac{\Sigma\mu^2}{s} = \frac{\mu'^2}{n} + \frac{n-1}{n}\pi^2 \qquad (11)$$

La moyenne des fluctuations des moyennes secondaires dépasse la valeur qu'elle prendrait, si les groupes avaient même moyenne, d'une valeur égale à $\frac{1}{n^e}$ *près, à la fluctuation des moyennes des groupes.*

Imaginons par exemple une armée composée de s régiments dont chacun comprend n compagnies. Si, dans chaque régiment, on prend 1 homme par compagnie et que l'on détermine la taille moyenne des n hommes, cette taille moyenne s'écartera généralement de la taille moyenne de l'armée ; elle s'en écartera davantage si les différents régiments donnent des moyennes différentes que s'ils donnent tous la même moyenne.

17. — Le cas particulier relativement simple signalé dans les paragraphes précédents, celui où toutes les distributions primaires d'un groupe ont le même nombre de termes et même fluctuation, possède, comme nous l'avons vu, cette propriété importante que la moyenne générale et le quotient de la fluctuation commune à toutes les distributions primaires par leur nombre, se conservent dans la transformation de ces distributions primaires en distribution secondaire. D'une manière générale, les grandeurs qui subsistent ainsi à travers les transformations d'agrégats ayant quelque origine commune, offrent un grand intérêt parce qu'elles sont des liens qui rattachent les observations à leurs antécédents.

On peut se demander si, en considérant les puissances successives des écarts entre les observations et leur moyenne, les moyennes de puissances d'ordre supérieur au second ne se conserveraient pas aussi. Bienaymé a montré que la moyenne

des cubes jouit encore de la même propriété, mais qu'il n'en est plus ainsi des moyennes de puissances supérieures.

La fluctuation paraît donc la seule moyenne de puissances utilisable pour la comparaison des distributions formées conformément aux paragraphes précédents, car la moyenne des cubes, laissant aux écarts des signes différents, est inutilisable.

Dans le cas particulier où les n distributions primaires ont même fluctuation, la fluctuation des moyennes secondaires est, d'après l'égalité 10, la n^e partie de la fluctuation commune des distributions primaires.

Par conséquent, plus grandit le nombre n des distributions primaires d'où dérivent les moyennes secondaires, plus diminue la fluctuation de ces dernières. Celle-ci décroît indéfiniment, quand n augmente indéfiniment. Or, le carré moyen des écarts autour de la moyenne est composé de termes positifs; il ne peut devenir très petit que si aucun de ces termes n'est grand. Il en résulte qu'au fur et à mesure de sa réduction, les écarts qui, auparavant, avaient une valeur relativement grande (quand n n'était point très grand), doivent nécessairement diminuer de plus en plus, d'autant plus que leur nombre augmente beaucoup plus vite que n.

Un écart de l'ordre $\frac{1}{\sqrt{n}}$ est ici regardé comme très petit, puisque son carré est de l'ordre $\frac{1}{n}$, c'est-à-dire de l'ordre de la fluctuation supposée très petite.

Ces considérations peuvent se résumer comme suit :

Lorsque les distributions primaires, de même nombre de termes et de même fluctuation, dont dérive une distribution secondaire, deviennent de plus en plus nombreuses, en sorte que la distribution secondaire comprend un nombre plus grand de termes, les moyennes des termes de cette distribution secondaire se concentrent autour de leur moyenne, ou encore, leur dispersion diminue de plus en plus.

Les termes de chaque distribution peuvent être tous différents, ou bien certains peuvent se répéter.

18. — Si l'on veut apprécier la grandeur d'un écart entre la moyenne d'un terme secondaire particulier et la moyenne générale M (égale dans ce cas à m) il est naturel de prendre pour module de comparaison la racine carrée de la fluctuation puisque la fluctuation est le carré moyen de tous les écarts.

Désignons par μ'^2 la fluctuation commune des n distributions primaires $\mu'^2 = \frac{\Sigma (x - M)^2}{n'}$ (n' étant le nombre de termes de chaque distribution primaire); puis supposons un écart égal à $\pm t\,\mu'$.

Pour les moyennes secondaires, la moyenne des carrés des écarts est

$$\frac{\mu'^2}{n} = \frac{\Sigma (M_i - M)^2}{\nu}.$$

Les termes de cette dernière somme qui correspondent à des écarts $M' - M$ au delà de $t\mu'$ donnent un total inférieur à la somme de tous les termes. Soit T ce total, on peut donc écrire :

$$T = k \frac{\Sigma (M - M_i)^2}{\nu} = k \frac{\mu'^2}{n}$$

k étant plus petit que 1.

D'autre part, les moyennes secondaires au delà des limites $\pm t\mu'$ étant, par exemple, au nombre de ν' ($\nu' < \nu$), le total des termes de la fluctuation secondaire correspondant aux écarts qui dépassent les limites est nécessairement plus grand que le produit du plus petit de ces termes $t\mu'$ multiplié par leur nombre On peut donc écrire :

$$T = \frac{\nu'}{k'} t^2 \mu'^2 \quad (k' < 1)$$

D'où, en égalant les deux valeurs de T

$$k \frac{\mu'^2}{n} = \frac{\nu'}{k'} t^2 \mu'^2,$$

ce qui donne

$$\nu' = \frac{kk'}{nt^2}$$

Le nombre des termes compris entre les limites $\pm t\mu'$ est dès lors :

$$\nu - \nu' = \nu - \frac{kk'}{nt^2} = \nu \left(1 - \frac{kk'}{nt^2\nu}\right)$$

d'où :

$$\frac{\nu - \nu'}{\nu} = 1 - \frac{kk'}{nt^2\nu}$$

Le nombre des termes compris entre les limites $\pm t\mu'$ est une fraction du total qu'on peut rapprocher de l'unité autant qu'on le veut, en faisant n assez grand (théorème de Bernoulli).

19. — En considérant successivement s groupes de distributions primaires, tous composés d'un même nombre de ces distributions, nous avons vu (**16**) que la moyenne des fluctuations secondaires des différents groupes était le total obtenu en ajoutant à la n^e partie de la fluctuation primaire la fluctuation des moyennes des groupes, celle-ci multipliée par $\frac{n-1}{n}$.

On a :

$$\frac{1}{s} \sum_1^s \mu^2 = \frac{\mu'^2}{n} + \frac{n-1}{n} \pi^2$$

Supposons une observation de la valeur moyenne d'un terme secondaire dans chaque groupe ; la fluctuation des s observations est moyennement $\frac{\Sigma\mu^2}{s}$. Les s' termes, qui, parmi ceux composant cette fluctuation, correspondent aux écarts au delà de $\pm t\mu'$ entrent pour une part nécessairement plus petite que cette fluctuation ; leur somme T peut s'exprimer par

$$T = k \left(\frac{\mu'^2}{n} + \frac{n-1}{n} \pi^2\right)$$

Égalant cette somme à celle qu'on obtient en substituant à chacun de ces s' termes leur valeur minimum $t\mu'$ multipliée par $\frac{1}{k'}$, il vient :

$$k\left(\frac{\mu'^2}{n}+\frac{n-1}{n}\pi^2\right)=\frac{s'}{k'}t^2\mu'^2$$

d'où

$$s'=\frac{kk'}{nt^2}\left(1+(n-1)\frac{\pi^2}{\mu'^2}\right)$$

où k et k' sont des nombres plus petits que 1.

Le nombre des termes situés au delà des limites $\pm t\mu'$ diminue à mesure que n augmente ; le nombre relatif de ces termes diminue à mesure que s et n augmentent.

La concentration autour de la moyenne, même pour des observations extraites de groupes dont les moyennes sont différentes, caractérise ce que Poisson a appelé, improprement comme on va le voir, la *loi des grands nombres.*

20. — Il est intéressant de comparer la fluctuation de ces observations, dans le cas où les groupes successifs de distributions primaires ou secondaires ont des moyennes différentes (§ **15** et **19**), à la fluctuation de ces observations, dans le cas où les groupes successifs ont même moyenne (§ **13** et **18**).

Dans le premier cas, le nombre s' des observations situées au delà des limites $\pm t\mu'$ est égal à

$$\frac{kk'}{nt^2}\left(1+(n-1)\frac{\pi^2}{\mu'^2}\right)$$

dans le second à $\frac{kk'}{nt^2}$.

La quantité $1+(n-1)\frac{\pi^2}{\mu'^2}$ représente donc le rapport des deux nombres et l'on voit que ce rapport est d'autant plus petit que n est plus petit, comme Bienaymé l'a remarqué en 1839.

Une série d'observations empruntées à des groupes pour lesquels le nombre des distributions, la moyenne et la fluctuation ne varient pas d'un groupe à l'autre peut être dite composée de groupes *stables* de distributions. Et quand, d'un groupe à l'autre, la moyenne et la fluctuation varient, la série peut être regardée comme moins stable, en ce sens que la fluctuation secondaire issue de la série moins stable est plus grande que celle de la série stable (**16**). Le rapport des deux fluctuations peut, comme Lexis l'a proposé, être pris pour mesure de la stabilité. Dès lors, *la stabilité relative est d'autant plus grande que* n *est plus petit,* ce qui explique en partie la stabilité de certains phénomènes se manifestant par groupes peu nombreux (masculinité des naissances dans de petits districts, etc.).

21. — Les considérations précédentes justifient l'emploi de la moyenne comme point de départ des comparaisons entre les termes d'une distribution, ou comme premier instrument de la comparaison de diverses distributions.

On a recommandé aussi l'emploi d'une grandeur qui laisse au delà de sa valeur autant d'éléments de la distribution qu'elle en comprend en deçà. Cournot a donné à cette grandeur le nom de valeur *médiane.* La médiane offre l'inconvénient de ne

pouvoir être exactement déterminée quand elle ne se confond pas avec l'un des termes de la distribution (nombre pair de termes). Entre les deux termes qui la comprennent, on peut lui attribuer une valeur quelconque, en sorte qu'elle entrerait difficilement dans un calcul algébrique. Elle est néanmoins d'un grand intérêt pratique.

Si l'on partage une distribution en deux parties contenant le même nombre de termes et si l'on prend les médianes de ces deux parties, l'intervalle de ces médianes (appelé *quartile* par F. Galton) peut servir à comparer la dispersion de deux distributions, d'une manière plus rapide que la fluctuation. De plus, la médiane jouit d'une propriété analogue à la propriété fondamentale de la moyenne. Soit y une grandeur quelconque, x_1 la valeur quelconque des grandeurs plus grandes que y, x_{-1} la valeur quelconque des grandeurs plus petites que y; pour les premières, $x_1 - y$ est un nombre constamment positif; pour les secondes $y - x_{-1}$ est aussi constamment positif.

La moyenne m coïncide avec la valeur de y qui rend

$$\Sigma (x_1 - y) - \Sigma (y - x_{-1}) = 0$$

ou, ce qui revient au même, qui rend minimum $\Sigma (x_1 - y)^2 + \Sigma (y - x_{-1})^2$.

La médiane m *coïncide avec la valeur de* y *qui rend minimum la somme des écarts pris en valeur absolue :* $\Sigma (x_1 - y) + \Sigma (y - x_{-1})$.

Dans la définition de la moyenne, chaque écart entre la moyenne et une grandeur particulière intervient par son carré, ce qui exagère peut-être l'influence des grands écarts; par contre, dans la définition de la médiane, chaque écart n'intervient que par son signe, en sorte que la médiane ne tient pas compte de la valeur de la plupart des grandeurs associées; à cet égard elle constitue un instrument de comparaison plus défectueux que la moyenne.

II — VARIABILITÉ

22. — La comparaison des séries de faits numériques est le but principal de la statistique; mais on ne peut se contenter de comparer des moyennes, il faut encore voir comment les faits se distribuent autour de ces moyennes. D'autre part, la comparaison de séries différentes de faits, permet d'établir entre ces séries de faits, des rapports, des caractères communs.

Nous allons signaler sommairement des méthodes en nous efforçant de mettre en lumière les hypothèses sur lesquelles elles sont fondées. Pour répondre aux exigences de la critique, le statisticien doit faire exactement le départ entre l'observation et l'interprétation.

Dans les distributions primaires prises précédemment comme origine d'une distribution secondaire, l'ordre des termes est indifférent, puisque la distribution secondaire emprunte ces termes de toutes les façons possibles.

Parmi tous les termes secondaires, tels que $n\mathrm{M}_i$, certains se répéteront, car une nouvelle distribution primaire composée de n' termes multiplie par n' le nombre de termes secondaires; les n' valeurs ajoutées se distribuent donc entre un nombre beaucoup plus considérable de termes et donnent forcément des répétitions; les répétitions seront d'autant plus fréquentes que le nombre n des distributions pri-

maires est plus grand. Dès lors, les termes secondaires étant classés par ordre de grandeur, donneront ce que l'on appelle une distribution de fréquence, chaque terme se présentant avec une certaine fréquence.

Cherchons à apprécier l'effet de la variabilité des termes des distributions primaires sur la variabilité de la distribution secondaire. Considérons le plus petit x et le plus grand X des termes des distributions primaires; la variabilité pourrait être appréciée par l'écart $X - x$, à la condition de tenir compte du nombre des termes Σn_h. Le rapport $\frac{X - x}{\Sigma n_h}$ peut être regardé comme une bonne mesure de la variabilité extrême d'un terme isolé quelconque.

Considérons maintenant la distribution secondaire. Le terme le plus grand est nécessairement inférieur ou au plus égal à nX et le terme le plus petit supérieur ou au moins égal à nx. Le nombre des termes est $n_1 \times n_2 \times \ldots \times n_h \times \ldots \times n_n$.

La variabilité des termes de la distribution secondaire, exprimée comme celle des distributions primaires, sera donc représentée tout au plus par $\frac{n(X - x)}{n_1 n_2 \ldots n_n}$, quantité plus petite que $\frac{X - x}{n_1 + n_2 \ldots + n_n}$ quand $n_1 n_2 \ldots n_n > n(n_1 + n_2 + \ldots + n_n)$. Tous les nombres entrant dans cette inégalité étant des nombres entiers, l'inégalité s'accentue avec son sens dès que n prend une valeur supérieure à quelques unités.

Par la manière dont se forment les termes de la distribution secondaire, on voit de plus que la fréquence des termes les plus grands et la fréquence des termes les plus petits sont plus petites que les fréquences des termes intermédiaires, car les termes les plus grands, comme les termes les plus petits, résultent uniquement de l'association des termes les plus grands (ou les plus petits) des distributions primaires, tandis que tout autre terme peut avoir la même valeur, qu'il provienne de l'association de termes primaires compris entre les extrêmes ou de termes primaires extrêmes avec d'autres termes primaires intermédiaires.

C'est d'ailleurs ce qui résulte de la concentration des termes de la distribution secondaire autour de la moyenne (**18**).

23. — D'après l'analyse précédente, la variabilité relative des termes de la distribution secondaire est moindre que celle des distributions primaires. En formant les termes de la distribution secondaire on égalise, dans une certaine mesure, les termes des distributions primaires, on adoucit leur inégalité.

En général, la fréquence des valeurs différentes des termes de la distribution secondaire varie irrégulièrement. Cette variation comporte pourtant une certaine régularité dans le cas particulier où les distributions primaires sont uniformes et simples.

Nous avons dit (**7**) que les termes de la distribution primaire de base pouvaient être tous différents ou se répéter sans nuire aux propriétés démontrées.

Supposons en premier lieu toutes les distributions primaires composées des mêmes grandeurs et chaque distribution comprenant deux séries de termes égaux; ainsi de $x_{1,1}$ à $x_{p,1}$ tous les termes seraient égaux à x; de $x_{p+1,1}$ à $x_{n,1}$, tous les termes seraient égaux à x', soit q le nombre de ces derniers. Chacune des n distributions primaires comprend donc p termes égaux à x et q termes égaux à x'.

La moyenne des termes d'une distribution primaire est $\frac{px + qx'}{p + q}$; cette moyenne

se conservant dans la distribution secondaire, les éléments d'un terme secondaire ont pour valeur moyenne $\frac{px + qx'}{p + q}$ et la valeur moyenne d'un terme secondaire est $n \times \frac{px + qx'}{p + q}$.

D'autre part, la fluctuation commune des distributions primaires supposées identiques est égale à

$$\frac{1}{p+q}\left[p\left(x - \frac{px+qx'}{p+q}\right)^2 + q\left(x' - \frac{px+qx'}{p+q}\right)^2\right] = \frac{pq\,(x-x')^2}{(p+q)^2}$$

Le quotient de cette fluctuation par n fournissant (**13**) la valeur de la fluctuation des moyennes secondaires, cette dernière est ainsi égale à

$$\frac{1}{n}\,\frac{pq\,(x-x')^2}{(p+q)^2}$$

Or, chaque terme secondaire est le produit par n de la moyenne de ses éléments; la fluctuation des termes secondaires est donc égale au produit par n^2 de la valeur ci-dessus, c'est-à-dire à

$$\frac{npq\,(x-x')^2}{(p+q)^2}$$

Il est facile de trouver la loi de fréquence de ces termes. En effet, on peut les former tous en développant la puissance

$$(px + qx')^n \qquad \text{ou} \qquad \left[\underbrace{x + x + \ldots + x}_{p} + \underbrace{x' + x' + \ldots + x'}_{q}\right]^n$$

Dans ce développement, les coefficients des termes tels que $x^m x'^{n-m}$ auxquels correspond la somme $mx + (n-m)x'$ sont les termes du développement de $(p+q)^n$.

D'après ce qui a été vu aux paragraphes **13** à **18** *les termes de la distribution secondaire convergent autour de leur moyenne à mesure que* n *augmente; il en est par conséquent de même des termes du binôme* $(p+q)^n$.

Dans ce cas, la moyenne tend à se confondre avec la valeur la plus fréquente ou la valeur *dominante* des termes de la distribution secondaire.

En général, cette valeur dominante ne partage pas en deux parties symétriques la suite des termes de la distribution secondaire. Pour qu'il y eût symétrie, il faudrait au moins que la fréquence du terme le plus grand fût égale à la fréquence du terme le plus petit, ce qui exige $p^n = q^n$ ou $p = q$. Cette condition est d'ailleurs suffisante, car il y a alors autant de termes x que de termes x' dans chaque distribution primaire; la formation de la distribution secondaire étant parfaitement symétrique, la symétrie primaire se conserve dans la transformation.

D'une manière générale, si chaque distribution primaire comporte p termes égaux à x_p, q termes égaux à x_q, r termes égaux à x_r..., les coefficients des termes tels que $x_p^\alpha x_q^\beta x_r^\gamma \ldots$ sont les termes du développement de $(p+q+r+\ldots)^n$.

Chacun des termes de ce développement représente donc le nombre de fois que le terme secondaire a la valeur particulière correspondante, telle que $\alpha x_p + \beta x_q + \gamma x_r + \ldots$

La loi de concentration autour de la moyenne, à mesure que n augmente, subsiste dans ce développement, aussi bien que dans celui de la puissance du binôme, mais la convergence est moins rapide.

Exemple. — Supposons :

$$p = 3 \qquad q = 2 \qquad r = 1 \qquad x_p = 0 \qquad x_q = 1 \qquad x_r = 2$$

chaque distribution primaire comprend les quantités 0 0 0 1 1 2.

Soient deux distributions primaires semblables, $n = 2$; en associant deux termes quelconques de ces deux distributions de toutes les façons possibles, la somme des deux termes associés prend les valeurs suivantes :

Les sommes	0	1	2	3	4
se présentent	9	12	10	4	1 fois

les nombres de la seconde ligne sont les termes de $(3 + 2 + 1)^2$.

24. — Les termes secondaires peuvent être formés très simplement lorsque, les distributions primaires étant identiques, les termes de chaque distribution varient en progression arithmétique. Comme le changement de raison de la progression ne modifie pas la fréquence des valeurs différentes des termes secondaires, on peut ramener tous les cas de ce genre à celui où chaque distribution primaire est composée de nombres entiers successifs.

Dans ce cas, la fréquence de chaque valeur, ou le nombre des répétitions des valeurs particulières d'un terme secondaire, s'obtient aisément en construisant ce que l'on a appelé un triangle arithmétique.

Si chaque distribution primaire ne comprend que deux termes, le triangle est celui de Pascal, composé de rangées successives de nombres dont chacun est le total des nombres consécutifs de la ligne précédente (triangle T_2).

0 0 0 0 1 0 0 0 0	0 0 0 0 1 0 0 0 0
0 0 0 0 1 1 0 0 0 0	
0 0 0 1 2 1 0 0 0	0 0 0 1 1 1 0 0 0
0 0 0 1 3 3 1 0 0 0	
0 0 1 4 6 4 1 0 0	0 0 1 2 3 2 1 0 0
0 0 1 5 10 10 5 1 0 0	
0 1 6 15 20 15 6 1 0	0 1 3 6 7 6 3 1 0
. . . . .	. . . . .
T_2	T_3

Le triangle T_2 peut être supposé formé dans un quinconce de zéros ; au milieu de la première ligne on inscrit un 1 au lieu de zéro. On forme ensuite les nombres d'une ligne quelconque en substituant à l'un des zéros la somme des deux nombres de la ligne précédente qui sont à cheval sur lui.

Si chaque distribution primaire comprend trois nombres entiers au lieu de deux, on forme les lignes successives en ajoutant trois nombres consécutifs de la ligne précédente, triangle T_3 (on saute à chaque fois une ligne de zéros pour que chaque nombre soit exactement au-dessous du second des trois nombres de la ligne supérieure) et ainsi de suite.

Si chaque distribution primaire comprend n' nombres entiers en progression

arithmétique, un terme de chaque ligne s'obtient en ajoutant le n' termes de la ligne précédente dont le terme cherché marque le milieu. La n^e ligne comprend les nombres de répétitions des termes secondaires formés à l'aide de n distributions primaires composées chacune des n' nombres en progression arithmétique. Il résulte de ce mode de formation que les nombres de répétitions, ou les fréquences, des termes secondaires se distribuent symétriquement.

Remarquons que la somme des termes de chaque rangée est égale au nombre des termes secondaires, c'est-à-dire à n'^n, tandis que le nombre des termes qui figurent dans chaque rangée est seulement égal à $n(n'-1)+1$; d'autre part, à partir du milieu, les termes de chaque ligne vont en décroissant, le dernier dans chaque sens étant égal à 1. Il en résulte que les termes voisins du milieu deviennent de plus en plus grands à mesure que n augmente, c'est-à-dire que les nombres des répétitions des termes secondaires se concentrent autour de leur moyenne à mesure qu'augmente le nombre des distributions primaires.

Quant aux valeurs des termes secondaires, ce sont les nombres entiers successifs allant de n à nn'.

La loi de formation des termes secondaires est donc pleinement déterminée; le terme de rang $p+1$, égal à $n+p$, est répété un nombre de fois égal au p^e nombre de la n^e ligne du triangle arithmétique dont le module de formation est n'.

D'après ce qui a été démontré précédemment (**18**), l'ensemble de ces termes se concentre autour de la moyenne ; *la moyenne est en même temps la valeur dominante de la distribution secondaire.*

Si les termes primaires en progression arithmétique se répètent, le nombre des répétitions croissant régulièrement, le mode de formation n'est plus symétrique, mais la concentration autour de la moyenne n'en subsiste pas moins.

25. — A mesure qu'augmente le nombre n des distributions primaires, les nombres de répétitions des valeurs identiques des termes secondaires se concentrent autour de leur moyenne. Pour un même écart à partir de cette moyenne, le nombre de répétitions est d'autant plus grand que n est plus grand et la différence de deux nombres successifs devient également de plus en plus grande. Dans un triangle arithmétique d'ordre n' la différence de deux nombres consécutifs d'une même rangée est égale à la différence des deux nombres de la ligne précédente qui sont séparés par $n'-1$ nombres intermédiaires. Il y a donc lieu d'espérer que le rapport calculé entre la différence de deux nombres successifs et le premier des deux nombres sera plus fixe que cette différence elle-même. D'autre part le dénominateur de ce rapport diminue aussi à mesure que l'on s'écarte de la moyenne. On obtiendra donc finalement un nombre beaucoup moins variable que ceux de la table triangulaire en déterminant le rapport qui existe entre la différence de deux nombres successifs et le produit du premier par sa distance à la moyenne.

Désignons par x la grandeur de cette distance, par z_x et z_{x+1} les nombres successifs d'une même rangée; le rapport $\frac{z_{x+1}-z_x}{z_x \times x}$ varie beaucoup moins que les nombres de la table. On pourrait donc représenter approximativement la loi de variation des nombres d'une même ligne par l'équation

$$\frac{z_{x+1}-z_x}{z_x \times x} = c$$

et, comme les nombres z représentés par cette équation se distribuent symétriquement de chaque côté de la moyenne, c est un nombre constamment négatif, on le remplacera par $— h^2$ pour bien marquer ce caractère et l'équation s'écrira :

$$\frac{z_{x+1} — z_x}{z_x \times x} = — h^2$$

Le cas particulier considéré dans le paragraphe précédent peut servir de schéma pour représenter simplement les résultats des observations humaines.

Quand on observe une grandeur, comme dans l'exemple considéré au paragraphe **5** (mesure d'un hectolitre de blé), on commet des erreurs élémentaires de différente nature dont l'association empêche l'exactitude du résultat. Ainsi dans l'exemple, on apprécie plus ou moins exactement la capacité de la jauge ; erreur aussi dans l'appréciation du moment où la jauge est remplie ; erreur dans l'appréciation du tassement des grains, etc. Nous ne connaissons pas la loi de ces erreurs. Nous pouvons pourtant distinguer deux catégories : les erreurs qui, en raison de leur petitesse ou faute d'attention, échappent entièrement à nos moyens de perception et celles dont nous avons plus ou moins conscience.

Naturellement, en ce qui touche la première catégorie, notre ignorance est absolue. Nous admettrons que chaque erreur élémentaire de cette catégorie étant répétée sur plusieurs opérations, l'observation qui en découle, si elle pouvait être appréciée, donnerait des grandeurs régulièrement échelonnées et également fréquentes à chaque échelon. Cette hypothèse est évidemment la plus simple. Dès l'instant où les erreurs commises échappent à nos moyens de perception et demeurent inconscientes, il n'y a pas de raison pour supposer que les unes sont plus ou moins fréquentes que les autres ; il n'y a, dans ces limites, aucune raison pour que les petites erreurs soient plus fréquentes que les grandes.

Eu égard à chaque cause d'erreur particulière, la grandeur observée peut prendre l'une quelconque des valeurs voisines $x_1, x_2 \dots x_n$ se succédant en progression arithmétique. Le résultat de l'observation est une association de valeurs empruntées à un groupe de distributions telles que $x_1, x_2 \dots x_n$: c'est-à-dire que ce résultat peut être assimilé à un terme secondaire issu d'un groupe de distributions primaires identiques dont chacune comporte des termes en progression régulière.

On a vu, dans le paragraphe précédent, que le nombre z des répétitions d'un même terme secondaire était très variable, mais qu'on obtenait un rapport beaucoup moins variable en calculant la fraction $\frac{z_{n+1} — z_n}{z_n \times x}$ *où* x *représente l'écart entre l'une des grandeurs observables* $x_1, x_2 \dots x_n$ *et la moyenne de ces grandeurs.*

On pourrait donc représenter approximativement la distribution des termes secondaires, c'est-à-dire des observations possibles, par la formule du paragraphe précédent.

$$\frac{z_{n+1} — z_n}{z_n x} = — h^2$$

26. — Indépendamment des erreurs imperceptibles et inconscientes, nous commettons des erreurs dont nous avons plus ou moins conscience. Celles-ci obéissent peut-être à une loi dont il convient de chercher une expression plausible.

Puisque, dans une certaine mesure, nous avons conscience de l'erreur, nous pouvons admettre que nous avons tendance à l'atténuer le plus possible ; par conséquent les erreurs de ce genre ne sauraient être, comme les premières, également nombreuses, quelles que soient leurs grandeurs.

Cette hypothèse étant écartée, on pourrait supposer simplement que la fréquence z d'une erreur décroît en raison inverse de sa grandeur, mesurée à partir de la moyenne, ce qu'exprimerait la relation $z = \frac{c}{x}$. Dans cette hypothèse le nombre des erreurs croîtrait indéfiniment quand celles-ci deviendraient très petites ; les erreurs appréciables ne seraient pas possibles, ce qui est contraire à ce que nous savons de nos facultés.

On éviterait cet inconvénient en substituant à x une expression telle que a^x, ce qui donnerait la relation $z = \frac{c}{a^x}$.

Mais, cette formule, comme la précédente d'ailleurs, ne peut convenir pour les erreurs imperceptibles supposées également nombreuses quelle que soit leur grandeur (**25**). D'après cette hypothèse, le coefficient différentiel $\frac{z_{x+1} - z_x}{z_n}$ doit être presque nul aux environs de la moyenne.

En fait, ce coefficient représente le déchet relatif du nombre des erreurs. Si, à mesure qu'on s'écarte de la valeur dominante, c'est-à-dire à mesure que x augmente, ce déchet augmente, le nombre des erreurs ira toujours en diminuant. La manière la plus simple de concevoir cet accroissement simultané, les autres hypothèses étant écartées, est de supposer que les deux quantités qui varient dans le même sens sont proportionnelles, c'est-à-dire de poser

$$\frac{z_{x+1} - z_x}{z_x} = -h^2 x \qquad (12)$$

On est ainsi amené, pour cette catégorie d'erreurs, à une forme identique à celle du paragraphe précédent, cette forme convenant à la fois pour les deux catégories.

27. — Cette formule ne saurait fournir une expression tout à fait générale de la distribution des erreurs d'observation. D'abord, les erreurs systématiques ou tendancieuses y échappent entièrement. Celles-ci peuvent être déterminées et corrigées.

En second lieu, il y a des cas où les erreurs s'écartent notablement de la loi représentée par la formule (12). Exemple : les erreurs commises en se servant d'une équerre pour mesurer la taille.

Mais, en dehors de ces cas particuliers, qui peuvent être isolés par une analyse et une critique attentive des conditions de l'observation, *l'équation (12) convient pour représenter la loi de variation des erreurs d'observation* et se trouve généralement assez exactement vérifiée.

28. — Quételet a remarqué qu'un grand nombre de faits naturels se distribuent conformément à la loi des erreurs d'observation. On peut s'en rendre compte par un raisonnement tel que le suivant.

Un phénomène naturel, examiné sous le point de vue quantitatif, atteint une grandeur déterminée au milieu de circonstances extrêmement nombreuses qui peuvent, à chaque instant, influer ou ne point influer sur son développement.

Considérons en particulier une circonstance A agissant sur un objet O. Supposons qu'intervenant seule elle produise une variation élémentaire x de O, indépendante de la grandeur de O. Si, au cours de l'opération, d'autres circonstances exercent aussi une influence, le progrès élémentaire dû à A ne sera point simplement déterminé par A et fixe; il comportera des possibilités variables. L'hypothèse la plus simple que l'on puisse faire pour représenter le champ de telles possibilités, est de supposer que les effets de A varient régulièrement; dès lors, les effets possibles de A peuvent être représentés par les termes d'une progression arithmétique $x_1, x_2 \ldots, x_n$ (**24**).

Toutefois, comme l'a remarqué Kapteyn, les effets possibles de A dépendent souvent de la grandeur de l'objet O; plus O est grand, plus peut grandir l'effet de A. D'ailleurs O grandit précisément sous l'effet des influences autres que A, ce qui établit un lien entre toutes ces influences.

Le schéma des valeurs en progression arithmétique ne représentera donc le cas le plus habituel que s'il est modifié de façon que la quantité x_p, par exemple, puisse s'accroître avec la dimension de O. Le mode d'accroissement le plus simple est de supposer que x_p est susceptible de se répéter d'autant plus de fois que O est plus grand.

Le schéma le plus général représentant, d'une manière simple, les possibilités de l'effet de A sera donc une suite de termes progressifs pouvant se répéter, et l'on supposera que le nombre des répétitions varie régulièrement.

Remarquons maintenant que la circonstance A peut être décomposée, au moins par la pensée, en parties élémentaires. Plus la division est poussée loin et plus se restreint le champ des possibilités de l'effet, en sorte que, poussant la division à l'extrême, il ne subsiste finalement que l'alternative : ou l'absence d'effet ou un petit effet, les deux termes de cette alternative pouvant comporter des répétitions (**23**).

29. — Nous rappellerons brièvement, maintenant, les principales propriétés des lignes représentées par la formule (12) que, pour plus de commodité, nous écrirons en notation infinitésimale

$$\frac{1}{z}\frac{dz}{dx} = -h^2 x \tag{13}$$

Cette équation ne satisfait pas tout à fait à l'une des conditions posées au paragraphe **24**. Elle comporte, en effet, une variation de z pour toute valeur de x et, par suite, elle représente une distribution illimitée. Toutefois, comme en valeur absolue la réduction de z est proportionnelle à x, on voit déjà qu'elle doit ramener z à une très petite valeur pour de grandes valeurs de x.

Si l'on représente par une courbe la relation (**13**) entre z et x, cette courbe a un sommet à l'aplomb de l'élément $x = o$ et deux branches qui s'abaissent vers l'axe des x de chaque côté du point o. Elle est symétrique; par conséquent l'élément dominant, l'élément moyen et l'élément médian (**21**) sont confondus.

En différentiant l'équation (13) on constate aisément que la courbe change de

courbure, d'abord convexe vers l'axe des x, puis concave; elle a deux points d'inflexion, symétriques par rapport à l'ordonnée dominante, dont l'écartement — ou *dispersion* de la distribution — est égal à $2\sigma = \frac{2}{h}$, d'où $h^2 = \frac{1}{\sigma^2}$.

L'équation (13) peut dès lors se mettre sous la forme

$$\frac{1}{z}\frac{dz}{dx} = -\frac{x}{\sigma^2} \tag{14}$$

Multipliant les deux membres par zx on a :

$$x dz = \frac{1}{\sigma^2} z x^2 dx$$

Intégrant de $-\infty$ à $+\infty$; le premier membre donne la surface de la courbe, soit le nombre N des observations, et l'on peut écrire :

$$N\sigma^2 = \int_{-\infty}^{\infty} z x^2 dx$$

ce qui fait connaître la valeur de $\sigma^2 = \frac{\int_{-\infty}^{\infty} z x^2 dx}{N}$: σ^2 est égal à la fluctuation des observations; Pearson appelle σ l'*écart type* de la distribution.

Intégrant l'équation (14) il vient :

$$z = z_0 e^{-\frac{x^2}{2\sigma^2}}$$

D'autre part, une formule connue donne $z_0 = \frac{N}{\sqrt{2\pi}}$. L'équation de la courbe qui représente la distribution prend finalement la forme

$$z = \frac{N}{\sigma\sqrt{2\pi}} e^{-\frac{x^2}{2\sigma^2}} \tag{15}$$

soit la formule de Gauss, laquelle exprime la loi de distribution des erreurs d'observation et représente la *courbe normale* de distribution. Le coefficient $\sigma\sqrt{2}$ a été désigné par Cournot sous le nom de *module de convergence*.

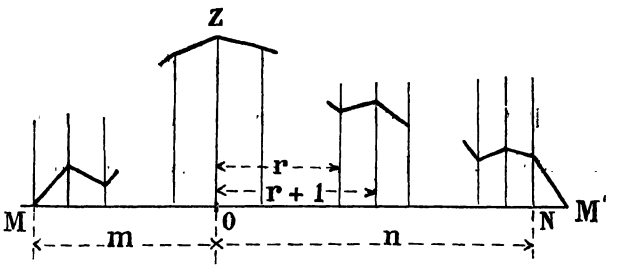
Pour comparer deux courbes il est bon de les rapporter à la même échelle, et de fixer une relation entre l'unité des x et celle des z ou entre les mesures de σ et de N.

30. — La courbe normale représente convenablement la distribution des observations humaines telles que celles considérées au paragraphe **5** dans l'exemple d'un marché équitable. Sa symétrie implique la compensation des erreurs.

En toute rigueur, pourtant, la plupart des faits naturels ne se distribuent pas symétriquement; pour représenter convenablement leur distribution, il convient de

chercher d'autres formes de courbes. Pour les motifs invoqués au paragraphe 28 et par raison de continuité, nous admettrons toujours qu'aux environs de leur valeur dominante, les grandeurs dont on étudie la distribution tendent à devenir également fréquentes. L'étendue de la distribution est, d'ailleurs, nécessairement limitée, puisqu'il s'agit de grandeurs observables.

De part et d'autre de l'élément dominant 0, la distribution des grandeurs observées peut être représentée par une ligne polygonale, en général fort irrégulière. Au centre de chacun des éléments régulièrement espacés de l'axe de représentation, élevons une perpendiculaire de longueur proportionnelle au nombre des grandeurs qui aboutissent à des points de cet élément. Les extrémités de toutes les perpendiculaires (ou ordonnées) semblables sont les sommets de la ligne polygonale. De part et d'autre de l'ordonnée dominante Oz, les ordonnées ont tendance à diminuer; elles s'annulent en certains points M et M'.



Cherchons une loi aussi simple que possible qui exprime la croissance ou la décroissance successive des ordonnées. Soit $MO = m$ et désignons par n la distance du point O au pied de la dernière ordonnée ayant une valeur appréciable.

Numérotons les ordonnées à partir du point O dans les deux sens; le point M étant situé sur la m^e ordonnée à gauche, nous dirons que sa distance au point O est égale à m; le point N étant sur la n^e ordonnée à droite, nous dirons que sa distance au point O est égale à n.

Considérons maintenant deux ordonnées consécutives de rangs r et $r + 1$, dont nous désignerons les longueurs par z_r et z_{r+1} et cherchons une expression générale du rapport $\frac{z_{r+1}}{z_r}$.

Lorsqu'on fait $r = -m$, on doit avoir $\frac{z_{-(m+1)}}{z_{-m}}$ infiniment grand, puisqu'au point M, Z est nul; d'autre part, si l'on fait $r = n$, le rapport $\frac{z_{n+1}}{z_n}$ doit être nul, puisque le point de la ligne polygonale qui vient après le point N est situé sur l'axe des abcisses. D'ailleurs la position de ce point peut rester indéterminée

Le rapport $\frac{z_{r+1}}{z}$ doit donc être une fonction de r de la forme

$$\frac{z_{r+1}}{z_r} = \frac{n - r}{r + m} \varphi(r)$$

De plus, la forme de la fonction $\varphi\left(\frac{x}{\sigma}\right)$ doit être telle qu'aux environs de l'ordonnée oz les ordonnées de la ligne polygonale soient presque égales.

On doit donc avoir, approximativement

$$\frac{z_1}{z_0} = 1$$

Supposons que l'on fasse exactement

$$\frac{z_1}{z_0} = 1$$

La forme générale du rapport $\frac{z_{r+1}}{z_r}$ (1) doit être telle que, pour $r = 0$,

$$\frac{z_1}{z_0} = \frac{n}{m}\varphi(0) = 1$$

$\varphi(o)$ doit donc être de la forme $\frac{m}{n}\psi(r)$, $\psi(r)$ étant égal à 1 pour $r = o$.

Le plus simple est de supposer $\psi(r) = 1$ dans tous les cas. La forme la plus simple du rapport $\frac{z_{r+1}}{z_r}$ satisfaisant aux conditions énoncées sera donc

$$\frac{z_{r+1}}{z_r} = \frac{n-r}{m+r}\frac{m}{n} \tag{16'}$$

A l'aide de cette formule on calculera successivement les ordonnées de la ligne polygonale théorique à laquelle peuvent être comparées toutes les lignes polygonales qui représentent des distributions d'observations, lorsque l'une de ces ordonnées est donnée ou lorsque leur somme totale, c'est-à-dire le nombre total des observations, est donnée. Il est facile de déduire de la formule (16) les valeurs successives de z quand r varie d'unité en unité et de se rendre compte que ces valeurs sont les termes du développement de

$$N \times \left(\frac{m}{m+n} + \frac{n}{m+n}\right)^{m+n-1}$$

L'équation (16) peut être mise sous la forme

$$\frac{z_{r+1} - z_r}{z_r} = -\frac{(m+n)\,r}{(m+r)\,n} = -\frac{m+n}{n}\left(1 - \frac{m}{m+r}\right) \tag{17}$$

ou, en développant la fraction $\frac{m}{m+r}$.

$$\frac{z_{r+1} - z_r}{z_r} = -\frac{m+n}{n}\left[\frac{r}{m} - \frac{r^2}{m^2} + \frac{r^3}{m^3} - \dots\right] \tag{18}$$

31. — Pour simplifier cette expression dans les cas où r est petit et où l'on peut négliger $-\frac{r^2}{m^2} + \frac{r^3}{m^3} \dots$ au regard de $\frac{r}{m}$, on peut la remplacer approximativement par

$$\frac{z_{r+1} - z_r}{z_r} = -\frac{m+n}{n}\frac{r}{m} \tag{19}$$

(1) La conclusion finale n'est pas modifiée si l'on part de l'égalité $\frac{z_0}{z_1} = 1$ ou $\frac{z_{\frac{1}{2}}}{z_{-\frac{1}{2}}} = 1$.

Supposons maintenant que le nombre $m + n$ des divisions de l'axe ox aille en grandissant, les deux lignes polygonales représentées par les expressions (18) et (19) tendront vers des courbes dont les équations, en notation infinitésimale, seront

$$\frac{1}{z'}\frac{dz'}{dx} = -\frac{(m+n)\,x}{n\,(m+x)} = -\frac{m+n}{n}\left(1 - \frac{m}{m+x}\right) \tag{20}$$

et

$$\frac{1}{z}\frac{dz}{dx} = -\frac{(m+n)\,x}{nm} \tag{21}$$

celle-ci ayant la même forme que l'équation (14) du paragraphe **29**.

Si l'on dérive par rapport à x les équations (20) et (21), on constate que les deux courbes représentées par ces deux équations ont même forme, à cela près que l'une (21) est symétrique et l'autre non.

Pour toutes deux, l'intervalle des points d'inflexion a la même grandeur, ce que l'on peut exprimer en disant que la *dispersion* des deux distributions est la même : $2\sqrt{\frac{mn}{m+n}}$ ou 2σ (**29**), soit le double de l'écart type de la courbe normale (**29**). Mais σ n'est égal au carré moyen des écarts que pour cette courbe normale. D'après l'équation (18), pour obtenir l'écart type de la courbe déviée, équation (20), il faut compléter le carré moyen des écarts par des termes de degrés supérieurs.

Si l'on pose $\frac{m}{m+n} = p$, $\frac{n}{m+n} = q$, p et q étant les grandeurs relatives des deux fractions de la distribution que sépare la valeur dominante, on peut écrire

$$\sigma = \sqrt{(m+n)\,pq}$$

L'écart type relatif qui sert à comparer deux distributions d'étendue différente (29) s'exprime alors par le rapport $\sqrt{\frac{(m+n)pq}{m+n}}$; il diminue proportionnellement à la racine carrée de $m + n$, grandeur qui est elle-même prise pour mesure du nombre des observations afin que les ordonnées et les abscisses des courbes comparées soient ramenées à la même unité.

Au delà de M et de N, dans les parties du plan qui ne sont point utiles pour la représentation, la seconde courbe, équation (21), reste très voisine de l'axe $0x$ et lui est asymptote des deux côtés. La première courbe, équation (20), est limitée au point M, à gauche de l'origine. Vers la droite, au delà de N, elle est asymptote à $0x$.

32. — En intégrant les équations (20) et (21), on obtient les relations qui déterminent z et z' en fonction de x

$$z' = \mathrm{H}e^{-\frac{m+n}{n}x}(m+x)^{\frac{(m+n)m}{n}} = z'_0 e^{-\frac{m+n}{n}x}\left(1 + \frac{x}{m}\right)^{\frac{(m+n)m}{n}} \tag{23}$$

$$z = z_0 e^{-\frac{x^2(m+n)}{2mn}} \tag{24}$$

La courbe (24), dite normale, est symétrique par rapport à l'ordonnée dominante; la courbe (23) est dissymétrique.

Si dans l'équation (23) on fait $x + m = x'$, l'expression de z' prend la forme

$$z' = z'_0 m^{-\frac{(m+n)m}{n}} e^{\frac{(m+n)m}{n}} e^{-\frac{m+n}{n}x'} x'^{\frac{(m+n)m}{n}}$$

Posons $\frac{m+n}{n} = \frac{1}{q}$, substituons à la variable x' la variable y définie par $x' = qy$ et intégrons de 0 à ∞ pour obtenir la surface de la courbe

$$z' = z'_0 m^{-\frac{m}{q}} e^{+\frac{m}{q}} e^{-\frac{x}{q}} x^{\frac{m}{q}} \tag{25}$$

l'intégration donne

$$\frac{N}{q} = z_0 \left(\frac{m}{q}\right)^{-\frac{m}{q}} e^{+\frac{m}{q}} \times \Gamma\left(\frac{m}{q} + 1\right)$$

d'où l'on déduit la valeur de l'ordonnée dominante

$$\frac{z'_0}{N} = \left(\frac{m}{q}\right)^{\frac{m}{q}} e^{-\frac{m}{q}} \times \frac{1}{q \Gamma\left(\frac{m}{q} + 1\right)}$$

Si l'on remplace $\Gamma\left(\frac{m}{q} + 1\right)$ par la valeur approchée par défaut que fournit la formule de Stirling, il vient,

$$\frac{z'_0}{N} = \left(\frac{m}{q}\right)^{\frac{m}{q}} e^{-\frac{m}{q}} \times \frac{1}{q \left(\frac{m}{q}\right)^{\frac{m}{q}} e^{-\frac{m}{q}} \sqrt{2\pi \frac{m}{q}}}$$

d'où

$$\frac{z'_0}{N} = \frac{1}{\sqrt{2\pi mq}} = \frac{1}{\sqrt{2\pi\sigma}}$$

valeur approchée par excès; l'ordonnée dominante de la courbe déviée est plus petite que l'ordonnée dominante de la courbe normale (1).

33. — Lorsqu'on dispose d'une série statistique faisant connaître la répartition d'un ensemble d'individus classés suivant la grandeur d'un caractère commun, on peut souvent substituer à la ligne polygonale qui représente cette répartition une courbe de la forme (25). On y parvient par les procédés connus d'ajustement; en particulier, on détermine aisément les valeurs de q et de m comme l'a fait K. Pearson, en égalant les mouvements des premiers ordres, calculés pour la courbe et pour la ligne polygonale. Lorsque q et m sont déterminés, on obtient la valeur de $n = \frac{qm}{1-q}$.

(1) Des applications de courbes déviées ont été données dans le *Journal de la Société de Statistique de Paris*, numéros de juillet 1898 (*Quelques exemples de distributions de salaires*) et de juillet 1902 (*Salaire des ouvriers mineurs belges*).

On connaît dès lors l'intervalle des points d'inflexion

$$2\sqrt{\frac{qm^2}{(q-1)\left(\frac{m}{q-1}\right)}} = 2\sqrt{mq}$$

Les courbes (24) et (23) se rapportant au même nombre d'observations et ayant par conséquent même surface, alors que $\frac{x}{m}$ demeure assez petit, la substitution de la courbe (24) à la courbe (23) se fait sans écarts importants entre les ordonnées des deux courbes. Lorsque $\frac{x}{m}$ est grand, les ordonnées des deux courbes deviennent très différentes, mais en même temps très petites. Dans tous les cas, les deux courbes ont même dispersion.

34. — Les formules (23) et (24) ont été obtenues en établissant une relation simple entre les ordonnées successives de la courbe représentative des observations, de façon que l'étendue de la courbe soit pratiquement limitée de part et d'autre d'un point culminant où la tangente soit horizontale.

La seconde de ces formules comporte deux paramètres; elle représente une courbe symétrique par rapport à l'ordonnée dominante; la première contient trois paramètres et représente une courbe déviée.

Au lieu de chercher la loi de décroissance des ordonnées consécutives de la courbe représentative des observations, on peut aborder la question d'une façon plus générale. Proposons-nous de trouver l'équation d'une courbe assujettie uniquement à avoir un seul sommet arrondi d'où partent deux branches qui se terminent en deux points de l'axe des x situés de part et d'autre de l'abscisse du sommet.

La courbe devant passer par les points de ox situés aux distances $-m$ et n de o, et ne devant avoir aucun point réel au delà de ces points, son équation peut utilement être mise sous la forme

$$z'' = (x+m)^p (x-n)^q \varphi(x)$$

p et q étant des nombres positifs quelconques non entiers. On peut aisément déterminer la forme de la fonction φ de façon que pour $x = o$, $\frac{dz}{dx} = 0$.

Mais, puisque nous cherchons une expression de z aussi simple que possible et que les deux premiers facteurs du second membre renferment déjà quatre paramètres m, p, n, q, il vaut mieux supposer tout de suite que $\varphi(x)$ se réduit à une constante A. On peut alors écrire

$$z'' = \mathrm{A}\,(x+m)^p (x-n)^q$$

en prenant la dérivée des deux membres

$$\frac{dz''}{dx} = \mathrm{A}\,(x+m)^{p-1} (x-n)^{q-1} [p(x-n) + q(x+m)]$$

La dérivée devant s'annuler pour $x = 0$, on doit avoir

$$A + m^{p-1}(-n)^{q-1}[-p\alpha + q\alpha] = 0$$

condition qui exige

$$\frac{p}{m} = \frac{q}{n} = \alpha$$

Divisant membre à membre les deux équations précédentes, puis remplaçant p par $m\alpha$, q par $n\alpha$, il vient :

$$(26)\qquad \frac{1}{z''}\frac{dz''}{dx} = \frac{\alpha(m+n)x}{(x+m)(x-n)} = -\frac{\alpha(m+n)x}{(x+m)(n-x)} = \alpha\left(\frac{m}{m+x} - \frac{n}{n-x}\right)$$

d'où, en intégrant

$$(27)\qquad z'' = z''_0\left(1+\frac{x}{m}\right)^{\alpha m}\left(1-\frac{x}{n}\right)^{\alpha n}$$

équation d'une courbe définie par les quatre paramètres m, n, α, z''_0.

L'équation différentielle (26) est de la forme

$$\frac{1}{z}\frac{dz}{dx} = \frac{-x}{ax^2+bx+c}$$

K. Pearson a classé les courbes représentées par cette équation en sept types, répartis en deux divisions suivant que les racines du dénominateur sont réelles ou imaginaires. Les expressions (26) ou (27) conviennent au cas des racines réelles.

Il n'y a aucun avantage à chercher des formes de distribution plus complexes dans lesquelles interviendraient plus de quatre paramètres. En effet, pour déterminer les paramètres, il faut ajuster la courbe théorique aux séries d'observations. Or, dans le cas général, pour cette opération, il n'y a pas d'autre méthode algébrique que celle des moments. Celle-ci est déjà défectueuse dans certaines applications des courbes à quatre paramètres qui exigent l'emploi des moments d'ordre 4 (Σax^4), parce que ces moments donnent une importance excessive aux observations extrêmes.

35. — L'équation différentielle (26) devient, lorsqu'on fait croître n indéfiniment

$$\frac{1}{z''}\frac{dz''}{dx} = \alpha\left(\frac{m}{m+x} - 1\right) = \frac{-\alpha x}{m+x}$$

Cette équation est identique à l'équation différentielle de la courbe déviée à trois paramètres

$$\frac{1}{z'}\frac{dr'}{dx} = \frac{m+n}{n}\frac{x}{m+x}$$

à la condition que $\alpha = \dfrac{m+n}{n}$.

Sous cette condition, l'équation (27) s'écrit

$$z'' = z''_0\left(1+\frac{x}{m}\right)^{\frac{m+n}{n}m}\left(1-\frac{x}{n}\right)^{n\frac{(m+n)}{n}}$$

Lorsque n augmente indéfiniment $\left(1 - \frac{x}{n}\right)^n$ tend vers e^{-x}. Donc z'' tend vers

$$z'' = z''_0 \left(1 + \frac{x}{m}\right)^{\frac{(m+n)m}{n}} e^{-\frac{m+n}{n}x}$$

expression identique à l'équation (23)

Déterminons les points d'inflexion de la courbe générale définie par l'équation (27).

En prenant la dérivée seconde des deux membres on obtient la valeur de l'abscisse des points d'inflexion :

$$\frac{d^2 z''}{dx^2} = \frac{\alpha^2 z''(m+n)}{(x+m)^2(x-n)^2}\left[\alpha(m+n)x^2 - x^2 - mn\right]$$

expression qui s'annule pour $x^2 = \frac{mn}{\alpha(m+n) - 1}$.

Dans l'équation de la courbe déviée à trois paramètres, n représentait le dernier degré de graduation des abscisses avant le contact de la courbe et de l'axe des abscisses. Dans l'équation de la courbe générale on a supposé que n était le degré correspondant au point de contact. Donc l'intervalle des points d'inflexion a la même valeur pour les deux courbes, et l'on a vu que cette même valeur déterminait la courbe normale. *Les trois courbes représentées par les équations* (24), (23), (27) *ont même dispersion,* tant que m et n ne varient pas. Or, les parties convexes des trois courbes sont celles qui comprennent le plus grand nombre des observations dans un même intervalle; la dispersion 2σ caractérise ainsi, dans les trois cas, la concentration des observations autour de la valeur dominante.

L'équation générale de la courbe à quatre paramètres, si l'on fait $\alpha = \frac{m+n}{n}$ peut être mise sous la forme

$$z'' = z''_0 \left(1 + \frac{x}{m}\right)^{\frac{m^2}{\sigma^2}} \left(1 - \frac{x}{n}\right)^{\frac{m}{\sigma^2}}$$

Cette courbe se transforme en courbe déviée à trois paramètres quand n augmente indéfiniment. Elle devient normale quand on suppose $m = n$ et que l'on fait croître indéfiniment m et α, leur rapport demeurant constant.

Son équation prend alors, en effet, la forme :

$$z'' = z''_0 \left(1 + \frac{x}{m}\right)^{\alpha m} \left(1 - \frac{x}{m}\right)^{\alpha m} = \left(1 - \frac{x^2}{m^2}\right)^{\frac{m^2\alpha}{m}}$$

Si m croît indéfiniment ainsi que α, le rapport $\frac{\alpha}{m}$ restant fini, a la limite $z'' = z''_0 e^{-\frac{\alpha x^2}{m}}$, courbe normale pour laquelle $2\sigma^2 = \frac{m}{\alpha}$.

36. — Les formules à deux, trois ou quatre paramètres permettent de représenter plus ou moins fidèlement les distributions d'observations naturelles classées

simplement par grandeur. Les distributions théoriques qu'elles représentent peuvent être considérées comme des étalons pour l'étude comparative des distributions naturelles.

On a recours à deux modes de comparaison. L'un, assez sommaire, consiste à comparer l'écart type de la distribution observée à l'écart type σ de la distribution théorique symétrique, en déterminant leur rapport, appelé par Dormoy *coefficient de divergence*.

Dans une autre méthode, on utilise la table qui fait connaître la surface comprise entre deux ordonnées symétriques, par rapport à la moyenne, et la courbe normale de même écart type, ou même dispersion, que la distribution étudiée, et l'on compare le nombre théorique ainsi obtenu au nombre observé.

Ces deux procédés ne sont, ni l'un ni l'autre, bien rigoureux, puisque la dispersion, qui conserve la même valeur dans les trois courbes étalons, se mesure autour de la dominante et non autour de la moyenne. On devrait tenir compte de la déviation.

L'équation (23) de la courbe déviée permet de calculer l'écart entre la valeur moyenne des observations et leur valeur dominante, soit d cet écart.

En désignant par μ'_2 le carré moyen des écarts des observations par rapport à la valeur dominante, μ_2 étant le carré moyen des écarts autour de la valeur moyenne, on sait que (**8**)

$$\mu'_2 = \mu_2 + d^2$$

En substituant à μ'_2 la valeur μ_2, comme on le fait d'ordinaire, on adopte une valeur trop petite.

Si donc la fluctuation, calculée d'après les observations, est plus grande que sa valeur théorique, donnée par la substitution de la courbe normale à la courbe déviée (coefficient de divergence supérieur à 1), le carré moyen des écarts à partir de la valeur dominante sera *a fortiori* plus grand que la fluctuation théorique.

Mais, si la fluctuation réelle est inférieure ou égale à la fluctuation théorique, on ne peut en conclure immédiatement qu'il en est de même du carré moyen des écarts mesurés à partir de la valeur dominante. Or, c'est cette valeur qui correspond à l'axe de symétrie de la courbe normale substituée à la courbe déviée dans le calcul de la fluctuation théorique.

37. — On sera peut-être surpris que, dans la recherche de courbes représentatives des observations naturelles, il suffise de poser la question sous la forme très générale : « Trouver une courbe aussi simple que possible, limitée en deux points de l'axe des x et offrant un point culminant intermédiaire où la tangente soit parallèle à l'axe des x », pour obtenir finalement des formes de courbes auxquelles s'adaptent assez bien les observations.

C'est que, quand nous avons cherché à rattacher le développement des observations naturelles à leurs causes, nous avons procédé avec un souci aussi grand de la simplicité.

On peut concevoir un fait naturel ou une observation, sous le point de vue quantitatif, comme rattaché aux circonstances adjuvantes ainsi que les termes d'une distribution secondaire se rattachent aux termes de distributions primaires. Si ces

distributions primaires sont quelconques, les termes secondaires ne suivent aucune loi. Nous ne sommes parvenus à leur assigner une loi que moyennant une hypothèse simplificatrice, en supposant toutes les distributions primaires identiques et réduites à des termes en progression régulière.

Nous avons adopté les hypothèses les plus simples quant à l'origine des observations; il n'est donc point étonnant qu'en s'imposant, sous une autre forme, la même condition d'extrême simplicité, on obtienne des courbes qui représentent convenablement les termes de la distribution secondaire issue de systèmes de distributions primaires.

III — COVARIATION

38. — Au cours des paragraphes précédents on a signalé les moyens de comparer une distribution d'observations statistiques à des distributions types, et par conséquent d'apprécier l'importance relative des écarts des observations particulières, par rapport à leur moyenne ou par rapport à la grandeur dominante.

L'étude comparative des faits, surtout quand ceux-ci s'ordonnent par rapport à des caractères autres que leur grandeur, exige d'autres procédés.

Pour comparer deux séries de n faits ordonnés d'après un certain attribut commun, le plus commode est de les représenter par des courbes [1] grâce auxquelles on embrasse d'un coup d'œil les régions des deux séries qui donnent des mouvements soit concordants, soit discordants.

Afin d'accroître la précision de ce rapprochement graphique, on considère chaque terme d'une des séries comme exprimant une variation v ou v' à partir de la moyenne m ou m' de la série. Quand les deux séries donnent des variations de même sens, le produit vv' est positif; il est négatif quand les variations sont de sens contraire. En sorte que la somme algébrique $\Sigma vv'$ exprime le résultat du dénombrement des concordances et des discordances; la moyenne $\frac{\Sigma vv'}{n}$ caractérise l'accord moyen des variations des deux séries. Si l'on veut que la grandeur moyenne des variations d'une série, par rapport à la grandeur moyenne des variations de l'autre, n'ait point d'influence sur l'expression, il faut rapporter les variations de chaque série à leur moyenne. Mais il importe d'éviter que celle-ci subisse l'effet des variations négatives; on substitue donc à la somme des variations la racine carrée de leur carré moyen. Ainsi, l'expression $\frac{\Sigma vv'}{n}$ est remplacée par

$$\frac{1}{n}\Sigma \frac{v}{\sqrt{\frac{\Sigma v^2}{n}}} \times \frac{v'}{\sqrt{\frac{\Sigma v'^2}{n}}}$$

et l'on adopte cette expression comme instrument de comparaison des variations des termes des deux séries à partir de leurs moyennes.

La quantité ci-dessus varie entre 0 et 1. Pearson l'a appelée *coefficient de corrélation*.

Par analogie avec les notations des paragraphes précédents, si l'on désigne par

(1) Comparer *Journal de la Société de Statistique de Paris*, numéro de janvier 1905 (*Les représentations graphiques et la statistique comparative*).

σ^2 la moyenne $\frac{\Sigma v^2}{n}$ et par σ'^2 la moyenne $\frac{\Sigma v'^2}{n}$ où n représente le nombre commun des variations comparées, la valeur du coefficient de corrélation peut s'écrire ainsi

$$r = \frac{\Sigma vv'}{n\sigma\sigma'} \tag{28}$$

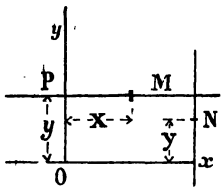
La grandeur de ce coefficient r mesure la ressemblance des deux courbes ou des deux séries de faits comparés, si l'on admet que les degrés de leur caractère commun sont assez petits pour qu'à un degré assigné ne corresponde qu'un seul fait dans chaque série.

39. — La comparaison revêt une forme un peu différente, et plus commode à certains égards, quand on représente les termes des deux séries, qui correspondent à la même valeur du caractère commun, par les deux coordonnées d'un point dans un plan.

Ce mode de représentation est analogue à celui dont nous avons fait usage (**29**) pour l'analyse de la distribution d'une série de faits statistiques. A chaque grandeur correspondait un point d'une droite et la distribution était représentée par la densité variable des points aux divers éléments de la droite.

Maintenant, nous représenterons la variabilité des couples de grandeurs possédant un certain caractère au même degré par la densité variable de points compris dans les divers éléments d'un plan.

Par exemple, nous pouvons représenter la distribution des couples mariés suivant les âges des deux époux au moyen de points dont chacun, tel que M, représente un couple. Et, de même que nous avons cherché des formes de distribution simple auxquelles pussent être comparées les distributions de points en ligne droite, de même nous chercherons une forme simple à laquelle comparer les distributions de points dans le plan. Soient deux axes rectangulaires tracés dans le plan à partir d'un point O. Considérons la tranche horizontale sur laquelle se trouvent tous les points situés à la distance y de ox. Nous devons admettre que la distribution simple cherchée, pour le plan, est telle que, sur la tranche horizontale PM, les points se distribuent conformément à l'une des lois obtenues précédemment pour les points en ligne droite. De même pour les points disséminés sur une parallèle quelconque à oy. Jusqu'à présent on n'a étudié les distributions dans le plan que dans le cas où les points en ligne droite se distribuent conformément à la loi normale.



Prenons pour origine commune des axes ox et oy l'élément où la densité des points par élément de plan est la plus grande.

Sur l'axe ox, la densité des points est maximum en o; si nous voulons que la distribution de ces points soit conforme à la loi normale, o est en même temps le centre des moyennes distances des points. De même o est le centre des moyennes distances des points situés sur OY.

Pour une parallèle quelconque PM à OX, le point de densité maximum étant en M, si la distribution a la forme normale et si l'on pose OM = X, on aura

$$\frac{1}{z}\frac{dz}{dx} = -\frac{x - X}{\sigma_1^2} \tag{29}$$

σ_1 est l'écart type des points sur PM.

De même sur une parallèle à OY, en appelant Y la distance du point de densité maximum à ox, on aura

$$\frac{1}{z}\frac{dz}{dy} = -\frac{y - Y}{\sigma_1'^2} \tag{30}$$

On admet encore, pour simplifier, que la distribution type à adopter pour les points du plan est celle qui donne sur toutes les droites parallèles à OX et OY des distributions normales.

Naturellement, dans les expressions précédentes, σ_1' et σ_1 sont des quantités variables. En général, les distributions des points sur les différentes parallèles à Ox n'auront point même dispersion, même en les supposant soumises à la même loi différentielle.

Cependant, comme nous cherchons un modèle de distribution aussi simple que possible, nous admettrons que, dans cette distribution idéale, l'écart type σ_1 des distributions parallèles à ox demeure constant et qu'il en est de même de l'écart type σ_1' des distributions parallèles à OY.

Même sous cette hypothèse, la valeur de z, c'est-à-dire le nombre des observations représentées dans un élément du plan, ne sera déterminable en valeurs des coordonnées x et y de cet élément que s'il existe une fonction z satisfaisant à la fois aux équations différentielles (29) et (30), ce qui implique la condition

$$\frac{1}{\sigma_1^2}\frac{dX}{dy} = \frac{1}{\sigma_1'^2}\frac{dY}{dx}. \tag{31}$$

Cette égalité devant avoir lieu quel que soit X pour toute valeur donnée de Y, et σ_1, σ_1' demeurant constants, la condition qu'elle exprime revient à supposer $\frac{dx}{dy}$ constant, c'est-à-dire que tous les points tels que M doivent être en ligne droite. De même les points moyens des distributions parallèles à OY doivent aussi se trouver en ligne droite. Comme le point O est nécessairement un point de ces deux droites, celles-ci convergent en O.

F. Galton, qui s'est servi de ces droites pour comparer des statistiques anthropométriques, les a appelées les *lignes de régression*. U. Yule a proposé une expression plus générale et meilleure : *axes d'estimation*, parce que chaque point de l'une fournit l'estimation convenable d'une des variables qui correspond à une grandeur donnée de l'autre variable.

D'après ce qui précède, les équations de ces droites, dont on verra plus loin la signification géométrique, sont

$$X = KY$$

et

$$Y = K'X$$

Pour déterminer K, considérons les distributions parallèles à OX et situées aux distances $y_1, y_2, \ldots y_m$ de OX, leurs points moyens ayant $X_1, X_2, \ldots X_m$ comme abscisses.

On a les égalités

$$\frac{X_1}{y_1} = \frac{X_2}{y_2} \ldots = \frac{X_m}{y_m} = K$$

d'où également

$$\frac{n_1 X_1 y_1}{n_1 y_1^2} = \frac{n_2 X_2 y_2}{n_2 y_2^2} = \ldots = \frac{n_m X_m y}{n_m y_m^2} = K = \frac{n_1 X_1 y_1 + n_2 X_2 y_2 + \ldots + n_m X_m y_m}{n_1 y_1^2 + n_2 y_2^2 + \ldots + n_m y_m^2}$$

Si n_1 est le nombre des points de la distribution situés sur la parallèle à OX distante de y_1

$$n_1 X_1 = x_{1,1} + x_{2,1} + \ldots + x_{n_1,1} = \Sigma x_1$$

$x_{1,1}$, $x_{2,1}$, $x_{n_1,1}$ étant les abscisses de ces points, et de même pour les autres distributions, en sorte que l'on peut écrire

$$K = \frac{y_1 \Sigma x_1 + y_2 \Sigma x_2 + \ldots + y_m \Sigma x_m}{n_1 y_1^2 + n_2 y_2^2 + \ldots + n_m y_m^2}$$

ce qui s'écrit plus simplement encore

$$K = \frac{\Sigma xy}{\Sigma y^2}$$

le signe Σ s'étendant à tous les points du plan et n étant le nombre total des points du plan. Or $\frac{\Sigma y^2}{n} = \sigma'^2$, σ' étant l'écart type calculé sur les directions parallèles à OY, mais pour tous les points du plan et non plus seulement pour une distribution linéaire, en sorte que K est défini par l'égalité

$$K = \frac{n \Sigma xy}{\sigma'^2}$$

En utilisant la formule (28) du paragraphe **38** qui exprime la valeur du coefficient de corrélation r, et remplaçant v et v' par x et y, on a

$$r = \frac{\Sigma xy}{n \sigma \sigma'}$$

d'où

$$K = r \frac{\sigma}{\sigma'}$$

L'équation de la première droite de régression peut donc s'écrire

$$X = r \frac{\sigma}{\sigma'} Y$$

on trouverait de même

$$Y = r \frac{\sigma'}{\sigma} X$$

De ces relations, on déduit aisément les valeurs de σ_1^2 et $\sigma_1'^2$.

L'égalité (31) donne, en effet, après substitution des valeurs de X et Y

$$\frac{1}{\sigma^2} r \frac{\sigma}{\sigma'} = \frac{1}{\sigma'^2} r \frac{\sigma'}{\sigma}$$

d'où

$$\frac{\sigma_1^2}{\sigma^2} = \frac{\sigma_1'^2}{\sigma'^2} = \alpha^2$$

Or (**8**)

$$n_1 \sigma_1^2 = \Sigma (x - X)^2 = \Sigma x^2 - n X^2$$

ou, en remplaçant X par sa valeur en y

$$n_1 \sigma_1^2 = \Sigma x^2 - n_1 \frac{\sigma^2}{\sigma'^2} y_1^2$$

Ajoutant toutes les équations semblables en remarquant que σ_1 conserve la même valeur dans toutes les tranches parallèles, il vient

$$n\sigma_1^2 = n\sigma^2 - \frac{\sigma^2}{\sigma'^2}\Sigma n_1 y_1^2 = n\sigma^2 - \frac{\sigma^2}{\sigma'^2} r^2\, n\,\sigma'^2$$

d'où

$$\sigma_1^2 = \sigma^2(1 - r^2) \text{ et de même } \sigma_1'^2 = \sigma'^2(1 - r^2)$$

En résumé, la fonction z se trouve définie par les deux équations différentielles :

$$\frac{1}{z}\frac{dz}{dx} = -\frac{1}{\sigma_1^2}\left(x - r\frac{\sigma}{\sigma'}y\right) = -\frac{1}{\sigma^2(1-r^2)}\left(x - r\frac{\sigma}{\sigma'}y\right)$$

$$\frac{1}{z}\frac{dz}{dy} = -\frac{1}{\sigma'^2}\left(y - r\frac{\sigma'}{\sigma}x\right) = \frac{1}{\sigma'^2(1-r^2)}\left(y - r\frac{\sigma'}{\sigma}x\right)$$

que l'on peut écrire aussi :

$$\frac{1}{z}\frac{dr}{dx} = -\frac{x}{\sigma^2(1-r^2)} + \frac{ry}{\sigma\sigma'(1-r^2)}$$

$$\frac{1}{z}\frac{dr}{dy} = \frac{rx}{\sigma\sigma'(1-r^2)} - \frac{y}{\sigma'^2(1-r^2)}$$

Pour la surface définie par les équations précédentes on a

$$\log z = -\frac{x^2}{2\sigma^2(1-r^2)} + \frac{rxy}{\sigma\sigma'(1-r^2)} - \frac{y^2}{2\sigma'^2(1-r^2)} + \log z_0$$

d'où

$$z = z_0 e^{-\frac{x^2}{2\sigma^2(1-r^2)} + \frac{rxy}{\sigma\sigma'(1-r^2)} - \frac{y^2}{2\sigma'^2(1-r^2)}} \tag{32}$$

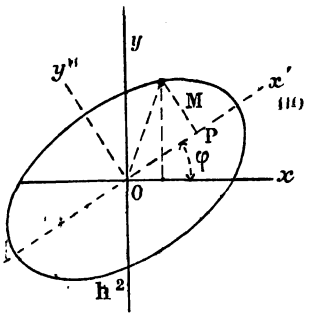
40. — Dans cette expression, l'exposant de e changé de signe prend, pour les valeurs particulières x et y, la valeur

$$h^2 = \frac{x^2}{2\sigma^2(1-r^2)} - \frac{rxy}{\sigma\sigma'(1-r^2)} + \frac{y^2}{2\sigma'^2(1-r^2)}$$

Pour une valeur donnée de h, cette équation représente une ellipse et, comme à cette valeur de h correspond une valeur fixe de z, cette ellipse est le lieu des éléments du plan où la densité des points est constante.

En faisant varier h, on obtient une série d'ellipses concentriques ayant mêmes axes principaux. Les diamètres conjugués des directions parallèles à ox et oy ont pour équation :

$$x = r\frac{\sigma}{\sigma'}y \qquad y = r\frac{\sigma'}{\sigma}x$$



ce sont les axes d'estimation ou de régression.

Soit φ l'angle que fait le grand axe ox de ces ellipses avec ox, cherchons l'équa-

tion générale de ces ellipses rapportées aux axes principaux en usant des formules de transformation

$$x = x' \cos\varphi - y' \sin\varphi$$
$$y = y' \cos\varphi + x' \sin\varphi$$

L'équation de l'ellipse h^2 devient

$$\frac{(x'\cos\varphi - y'\sin\varphi)^2}{2\sigma^2(1-r^2)} - \frac{r(x'\cos\varphi - y'\sin\varphi)(y'\cos\varphi + x'\sin\varphi)}{\sigma\sigma'(1-r^2)} + \frac{(y'\cos\varphi + x'\sin\varphi)^2}{2\sigma'^2(1-r^2)} = h^2$$

Cette équation représente une ellipse rapportée à ses axes principaux si le terme en xy disparaît. La condition pour que le coefficient de xy s'annule s'écrit

$$-\frac{\cos\varphi\sin\varphi}{\sigma^2(1-r^2)} + \frac{r}{\sigma\sigma'(1-r^2)}(\cos^2\varphi - \sin^2\varphi) + \frac{\sin\varphi\cos\varphi}{(\sigma'^2(1-r^2)} = 0$$

ou

$$\cos\varphi\sin\varphi\left(\frac{1}{\sigma^2} - \frac{1}{\sigma'^2}\right) + \frac{r}{\sigma\sigma'}(\cos^2\varphi - \sin^2\varphi) = 0$$

d'où

$$\operatorname{tg} 2\varphi = \frac{2r\sigma\sigma'}{\sigma^2 - \sigma'^2}$$

Le grand axe principal de l'ellipse caractérise la direction suivant laquelle les points du plan sont concentrés (axe principal d'inertie).

En effet, le carré de la distance MP d'un point quelconque M à cet axe est égal à

$$x^2 + y^2 - (x\cos\varphi \pm y\sin\varphi)^2 = x^2\sin^2\varphi + y^2\cos^2\varphi - 2xy\sin\varphi\cos\varphi$$

La somme des expressions semblables pour tout le plan est

$$\sin^2\varphi\Sigma x^2 + \cos^2\varphi\,\Sigma y^2 - 2\sin\varphi\cos\varphi\Sigma xy$$

Cette somme sera la plus petite possible, et, par conséquent, on obtiendra la direction de concentration des points, pour la valeur de φ qui annule la dérivée, c'est-à-dire pour

$$2\sin\varphi\cos\varphi\Sigma x^2 - 2\cos\varphi\sin\varphi\Sigma y^2 - 2\cos^2\varphi\Sigma xy + 2\sin^2\varphi\Sigma xy = 0$$

c'est-à-dire pour

$$2\sin\varphi\cos\varphi(\Sigma x^2 - \Sigma y^2) = 2\Sigma xy(\cos^2\varphi - \sin^2\varphi)$$

ou

$$\operatorname{tg} 2\varphi = \frac{2\Sigma xy}{\Sigma x^2 - \Sigma y^2}$$

valeur identique à la précédente.

Examinons maintenant la répartition géométrique des points.

Suivant les axes ox', oy', l'équation de la surface, le coefficient de xy étant nul, prend la forme

$$z = z_0\, e^{-\frac{x^2}{2\sigma^2} - \frac{y^2}{2\sigma'^2}}$$

les grandeurs de s et s' peuvent être trouvées en substituant à φ sa valeur. Pour une ellipse particulière

$$\frac{x^2}{2s^2} + \frac{y^2}{2s'^2} = h^2.$$

s^2 et s'^2 sont proportionnels aux carrés des diamètres principaux. Sur les directions parallèles à ox' les points sont donc symétriquement placés par rapport à y' et leur distribution est représentée par une courbe normale. De même les distributions parallèles à oy' sont également normales et ont leurs centres sur ox'. Il en résulte que les valeurs de x' correspondant à une valeur quelconque de y' n'ont aucun lien avec cette valeur. Quand y' augmente, le nombre des points de la distribution parallèle à ox' qui tendent vers la direction des x' positifs est aussi important que le nombre des points qui tendent vers les x' négatifs, de sorte que la moyenne se transporte perpendiculairement à ox'. Les grandeurs des x' et des y n'ont entre elles aucune corrélation.

Considérons maintenant l'axe ox, supposons-le d'abord coïncidant avec ox', puis faisons-le tourner autour de ox. Dans l'équation de la surface rapportée aux axes ox et oy, le terme en xy va grandissant, à mesure que grandit l'angle φ.

Suivant les directions parallèles à ox, la distribution des points est normale et le lieu des centres est le diamètre conjugué de la direction ox. Lorsque l'angle φ est très petit, le diamètre conjugué des directions parallèles à ox est très voisin de oy; le diamètre conjugué des directions parallèles à oy est très voisin de ox : l'angle de ces deux diamètres qui sont en même temps les axes d'estimation est voisin de 90°. A mesure que, pendant la rotation, φ augmente, c'est-à-dire à mesure que les axes des coordonnées s'écartent des axes principaux de l'ellipse, les diamètres conjugués des directions parallèles aux axes de coordonnées s'écartent de moins en moins l'un de l'autre, jusqu'à un angle minimum. En même temps, le terme en xy, $\sin 2\varphi \left(\frac{1}{\sigma^2} - \frac{1}{\sigma'^2}\right) + \frac{\sigma\sigma'}{r} \cos 2\varphi$, s'accroît constamment jusqu'à un maximum qui est atteint pour $\sin 2\varphi = 1$, $\cos 2\varphi = 0$, d'où $\varphi = \frac{\pi}{4}$.

La rotation continuant, l'angle des diamètres conjugés augmente de nouveau jusqu'à 90°, valeur atteinte quand ox coïncide avec oy'.

Observons maintenant ce que devient, pendant la rotation, la relation entre les x et les y.

Lorsque φ est nul, comme nous l'avons vu, les points situés sur des parallèles à ox sont symétriquement groupés autour de oy, quelle que soit leur distance à ox.

Dès que φ prend une certaine valeur, la distribution des points d'une même tranche parallèle à ox n'est plus symétrique par rapport à oy; le centre de la distribution est d'autant plus éloigné de oy que y est plus grand et que φ est plus grand. Ainsi quand φ augmente, une variation donnée de y déplace d'une quantité croissante le centre des points de la distribution des x, ce qui indique qu'un plus grand nombre des valeurs de x varient sous l'influence de la variation de y.

La corrélation entre les valeurs de x et les valeurs de y augmente donc à mesure qu'augmente, dans la valeur de z, le coefficient du terme en xy, jusqu'à un certain maximum atteint pour $\sin 2\varphi = 1$ ou $\varphi = \frac{\pi}{4}$.

Elle diminue ensuite, le coefficient de xy se rapprochant de o à mesure que φ tend vers 90°.

Tel est le mécanisme suivant lequel la surface représentée par l'équation (32) donne l'image de la covariation de deux grandeurs, dans l'hypothèse de corrélation normale, c'est-à-dire dans le cas où la distribution des points du plan est telle que, suivant deux directions rectangulaires, les points se distribuent normalement et, dans chaque direction, avec un écart type constant.

41. — Pour compléter la formule (32) il resterait à calculer z_0, mais il nous suffira de renvoyer à la méthode appliquée par Bravais en 1837 et qui se trouve dans les cours relatifs à la théorie des erreurs.

Cette valeur est

$$z_0 = \frac{1}{2\pi\sigma\sigma'\sqrt{1-r^2}}$$

en sorte que l'équation de la surface de corrélation (32) devient

$$z = \frac{1}{2\pi\sigma\sigma'\sqrt{1-r^2}}\, e^{-\frac{x^2}{2\sigma^2(1-r^2)}+\frac{rxy}{\sigma\sigma'(1-r^2)}-\frac{y^2}{2\sigma'^2(1-r^2)}}, \tag{33}$$

Cette équation est analogue à celle qui exprime (**29**) la loi de distribution normale de points en ligne droite. On peut comparer les distributions de couples de faits classés suivant un caractère commun à la distribution normale représentée par cette équation, soit en utilisant le nombre des points compris à l'intérieur d'une ellipse type analogue à l'écart type de la distribution linéaire, soit en utilisant la table qui donne le volume de la surface normale en fonction du paramètre h de l'ellipse type.

42. — Les procédés qui viennent d'être exposés permettent d'apprécier l'accord ou le désaccord des variations de deux séries de faits. Ils donnent par suite une base quantitative précise au jugement que l'on est amené à porter sur la relation des faits des deux séries.

La méthode a été étendue au cas où l'on veut apprécier l'influence de deux séries de faits sur une troisième. Udny Yule a démontré que, dans ce cas, le coefficient de corrélation qui conduit à une formule analogue à celle du paragraphe **38**, est celui qui combine le coefficient de corrélation entre la troisième série et la première avec le coefficient de corrélation entre la troisième et la seconde, de manière à donner un coefficient résultant maximum.

Nous renverrons aux mémoires de Pearson et de Yule. La théorie de la corrélation à trois variables a d'ailleurs été donnée, en même temps que la théorie à deux variables, par Bravais qui s'appuyait sur la petitesse et sur l'indépendance des changements relatifs des variables (erreurs d'observation).

Dans ce cas comme dans le précédent, il importe de bien distinguer entre le calcul du coefficient de corrélation et la représentation complète des observations par une surface ou un volume. Le coefficient de corrélation exprime la ressemblance numérique des variations des grandeurs comparées, sans aucune hypothèse sur le mode de distribution de ces grandeurs, tandis que la représentation par une surface ou un volume ne s'accorde avec la distribution des faits observés que si celle-ci est normale.

De plus, comme nous l'avons déjà remarqué, le coefficient dit de corrélation n'implique pas précisément une mesure de la relation de deux ou plusieurs faits; il précise seulement l'accord des variations des faits et serait plus exactement désigné sous le nom de coefficient de covariation, expression employée par J.-P. Norton.

Pearson et Yule ont pensé que l'on pouvait étendre l'étude de la corrélation à un nombre quelconque de variables. De même que par les équations d'estimation (**39**) on exprime la variation y d'une grandeur à l'aide de la variation x d'une autre par une relation linéaire

$$y = ax,$$

de même que la théorie de la corrélation à trois variables fournit une équation semblable

$$z = ax + by,$$

de même on pourrait exprimer la variation d'une grandeur u en fonction des variations de plusieurs autres

$$u = ax + by + cz + \dots$$

les coefficients a, b, c étant déterminés par l'application aux observations de la méthode des moindres carrés.

Mais cette méthode n'est valable que pour de très petites variations. Jusqu'à présent, l'étude de la corrélation à deux variables est à peu près la seule qui ait été véritablement féconde.

43. — Des réserves de même ordre semblent s'imposer, à propos du calcul des erreurs commises sur les quantités entrant dans les formules de distribution ou de corrélation.

Les calculs supposent que ces erreurs se distribuent normalement et, d'autre part, on ne les mène le plus souvent à bonne fin qu'en négligeant des quantités dont il est difficile d'apprécier l'importance réelle.

Il est sans doute nécessaire, pour l'objectivité et la loyauté des éléments d'appréciation numérique, que des limites soient assignées à ces éléments dont la détermination ne peut être exacte. Mais la fixation de ces limites ne saurait guère être que conventionnelle, leur valeur probante paraît difficilement mesurable.

Par exemple, comparant deux séries de faits dont les variations concomitantes à partir des moyennes sont v et v', nous avons déterminé le coefficient de corrélation

$$r = \frac{1}{n} \Sigma \frac{vv'}{\sigma\sigma'}$$

L'observation qui a fourni ce coefficient peut être regardée comme l'une de celles qui eussent été possibles dans un vaste ensemble de faits de même nature dont les séries considérées ne seraient qu'un fragment ou un groupe. Classons les groupes semblables d'après le nombre des concordances c ou des discordances d qu'on observe entre les deux catégories de faits qui le composent; ces groupes se distribueront suivant un mode que nous pouvons convenir de rapporter aux schémas étudiés dans les paragraphes **23** à **27**, l'amplitude étant $c + d$, valeur égale au nombre total n des variations comparées dans chaque groupe.

La distribution de ces groupes comportera pour les grandeurs c ou d une valeur dominante; cette valeur dominante, ou la plus fréquente, est celle que nous devons supposer se présenter à l'observateur de préférence aux autres.

On peut donc écrire, si le plus grand nombre de groupes donne c concordances et d discordances ($c + d = n$).

$$\frac{c-d}{c+d} = r \qquad \text{d'où} \qquad \frac{c}{c+d} = \frac{1+r}{2} \quad \frac{d}{c+d} = \frac{1-r}{2}$$

Mais, dans les schémas de distribution auxquels nous assimilons la distribution de nos groupes, c et d sont représentés par les lettres m et n et nous avons vu que le module de la distribution était, en valeur absolue, égal à $\sqrt{2\,\frac{mn}{m+n}}$ et, relativement à l'étendue de la distribution, $\sqrt{2\,\frac{mn}{(m+n)^3}}$.

Dans le cas actuel, cette dispersion sera donc

$$\sqrt{\frac{2\,cd}{(c+d)^3}} = \frac{1}{2}\sqrt{2\,\frac{1-r^2}{n}} = \sqrt{\frac{1-r^2}{2\,n}}$$

Tel serait l'indice qui caractériserait convenablement le degré de confiance à attribuer au coefficient r. D'après l'expression de cet indice, la valeur comparative de r doit inspirer d'autant plus de confiance que n est plus grand et r plus grand : c'est alors, en effet, que l'écart type des groupes, dont chacun fournit une valeur particulière de r, est le plus petit (1).

44. — En procédant comme nous l'avons fait au cours de ce trop rapide exposé, il nous semble que la théorie statistique se lie étroitement au processus logique d'où elle découle. La loi des grands nombres se ramène directement au principe de compensation équitable sur lequel est fondée la règle de la moyenne.

Le schéma de distributions conjuguées qui sert d'instrument à l'exposé n'est, d'ailleurs, qu'une forme particulière de l'hypothèse sur laquelle est fondée la logique de beaucoup de recherches scientifiques. On admet qu'un fait d'observation résulte de la superposition des petits effets de causes variables que l'on peut regarder comme invariables pendant l'instant où on les considère, et dont on associe les possibilités successives.

La théorie statistique aide à comprendre comment la régularité apparaît dans les résultats moyens, et elle explique dans une certaine mesure ces résultats. La variation des causes se manifestant par la variabilité des effets, il importe de mesurer cette dernière si l'on veut porter un jugement éclairé sur la première.

Cela fait, l'esprit dispose d'un point d'appui objectif pour l'interprétation des faits à la lumière des autres connaissances; les conjectures sont limitées et les chances d'erreur amoindries.

Lucien MARCH.

---

(1) K. Pearson, par un procédé plus complexe et en négligeant certains termes, emploie une valeur différente $(1 - r^2)\sqrt{\frac{2}{n}}$. Les deux valeurs sont égales quand $r = 0{,}87$.