MÉMORIAL DES SCIENCES MATHÉMATIQUES

G. DARMOIS

Les mathématiques de la psychologie

Mémorial des sciences mathématiques, fascicule 98 (1940)

<http://www.numdam.org/item?id=MSM_1940__98__1_0>

MÉMORIAL

DES

SCIENCES MATHÉMATIQUES

PUBLIÉ SOUS LE PATRONAGE DE

L'ACADÉMIE DES SCIENCES DE PARIS,

DES ACADÉMIES DE BELGRADE, BRUXELLES, BUCAREST, COÏMBRE, CRACOVIE, KIEW, MADRID, PRAGUE, ROME, STOCKHOLM (FONDATION MITTAG-LEFFLER), DE LA SOCIÉTÉ MATHÉMATIQUE DE FRANCE, AVEC LA COLLABORATION DE NOMBREUX SAVANTS.

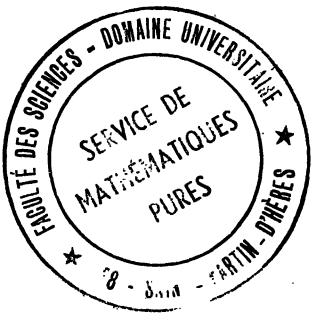
DIRECTEUR :

Henri VILLAT
Membre de l'Institut,
Professeur à la Sorbonne,
Directeur du « Journal de Mathématiques pures et appliquées ».

FASCICULE XCVIII

Les Mathématiques de la Psychologie

Par M. G. DARMOIS



PARIS
GAUTHIER-VILLARS, IMPRIMEUR-ÉDITEUR
LIBRAIRE DU BUREAU DES LONGITUDES, DE L'ÉCOLE POLYTECHNIQUE
Quai des Grands-Augustins, 55

1940

LES

MATHÉMATIQUES DE LA PSYCHOLOGIE

Par M. G. DARMOIS.

CHAPITRE I.

LES CORRÉLATIONS.

Le mouvement qui introduit de plus en plus les estimations numériques dans les études de psychologie s'est développé par utilisation de la statistique mathématique. Nous voudrions rassembler ici l'essentiel des méthodes statistiques utiles et montrer les services qu'elles ont pu rendre, dans le classement, la description et l'interprétation des faits observés.

Nous nous occuperons surtout de ce qui a trait aux différentes aptitudes et aux liaisons qu'elles peuvent présenter entre elles.

Le problème. — Considérons d'abord le cas, tout à fait schématisé, où les aptitudes étudiées seraient, pour un individu donné, des grandeurs bien déterminées, de véritables marques attachées à chacun, mais variables avec l'individu mesuré. Il faut alors rechercher comment ces caractères sont répartis dans la population. Supposons pour fixer les idées, qu'il y ait seulement deux caractères mesurables x et y. Nous considérons que la population soumise à nos mesures est une épreuve faite sur une population plus étendue, où pourraient se présenter toutes les nuances des caractères, toutes les valeurs de x et y. Cette population hypothétique aura sa structure fixée par une fonction de deux variables F(X, Y). Nous pourrons prendre pour

$F(X, Y)$ la proportion, dans la population totale, de la sous-population où l'on a les deux inégalités

$$x < X.$$
$$y < Y.$$

En général, nous aurons plutôt à chercher une densité $f(x, y)$ telle que $f(X, Y)\,dX\,dY$ soit la proportion des individus pour lesquels

$$X < x < X + dX,$$
$$Y < y < Y + dY.$$

Nous aurons donc à préciser la forme de cette fonction de deux variables à l'aide des renseignements fournis par la population réelle que nous étudions. La connaissance de cette fonction suffit évidemment à tout. En effet, si par exemple nous avons $F(X, Y)$, la variable x aura sa répartition connue par $F(X, +\infty)$, proportion où se trouve vérifiée la première inégalité, la variable y par $F(+\infty, Y)$. Mais il est à remarquer que l'intérêt du problème est dans d'autres caractéristiques, que contient $F(X, Y)$, mais qu'il est utile de faire ressortir.

Si, par exemple, nous étudions l'aptitude aux mathématiques et à la musique (en admettant que ces deux termes aient un sens précis), notre but est surtout de voir si ces aptitudes sont liées entre elles.

D'une façon précise, nous voudrions savoir si la connaissance de l'une de ces aptitudes est un apport de valeur positive, ou de nulle valeur, dans la connaissance de l'autre.

Pour en juger, nous considérons la sous-population où le caractère x a une valeur donnée, x_0 et nous étudions la répartition du caractère y dans cette sous-population; cette répartition sera dite répartition liée (à x_0 donné) de y. Deux cas peuvent se présenter :

1° *La répartition liée est toujours la même, quel que soit* x_0. — On dit que y, considérée comme variable aléatoire, est indépendante de la variable aléatoire x. On démontre immédiatement que la répartition liée de y est dans ce cas identique à la répartition de y, dite quelquefois répartition *a priori* ou répartition marginale, qui ne suppose rien sur la valeur de x.

D'autre part, il est facile de voir, par emploi du théorème des probabilités composées, que si l'on fixe la valeur de y, la répartition liée de x est dans ce cas indépendante de y_0. La propriété d'indé-

pendance est donc réciproque. Les deux variables aléatoires x et y sont dites indépendantes.

Quand x et y sont indépendantes, la connaissance de la valeur de l'une d'elles n'a aucune influence sur la loi de répartition de l'autre;

2° *La répartition liée change avec* x_0. — La variable aléatoire, y liée [qu'on peut désigner par $y^{(x_0)}$], a une loi de probabilité qui est fonction de x_0. La connaissance de x_0 a donc une influence sur ce que nous devons penser de y.

Il convient de mettre sous forme concrète cette idée générale. En effet, la connaissance de la loi de répartition de y se traduit par la valeur de ses caractéristiques les plus importantes, en particulier par celles qui correspondent au groupement et à la dispersion de y. Celles qui sont le plus fréquemment employées sont la valeur moyenne, ou espérance mathématique, et l'écart type, ou écart moyen quadratique autour de la valeur moyenne. Nous les désignerons par les notations classiques

$$\text{(1)} \qquad \mathrm{E}^{(x)}(y) = m,$$

$$\text{(2)} \qquad \mathrm{E}^{(x)}(y-m)^2 = \sigma_y^2.$$

Il pourrait d'ailleurs être utile d'introduire des caractéristiques d'ordre supérieur. Ceci posé, quand la variable liée y dépend vraiment de x_0, les deux caractéristiques m et σ_y sont deux fonctions de x_0, dont la connaissance est sans doute insuffisante pour préciser la loi de probabilité de y, mais qui sont déjà de grande valeur.

En particulier, si la fonction de dispersion σ_y reste très petite dans tout le champ de variation de x_0, on voit qu'on pourra assigner, avec une probabilité voisine de l'unité, un champ très petit autour de sa moyenne à la variable y liée.

Comme cas limite, si la fonction σ_y était identiquement nulle, la variable y liée n'est plus une variable aléatoire. La connaissance de x_0 fixe la valeur de y.

Rappelons ici quelques expressions courantes; le point dont les coordonnées sont x_0 et $m(x_0)$ décrit une courbe, qu'on appelle courbe de régression de y en x; il peut être commode d'introduire, de part et d'autre de la courbe de régression, les courbes obtenues en portant, à partir de la moyenne, une grandeur proportionnelle à l'écart type. Pour des raisons tirées des propriétés de la loi de Gauss,

on portera généralement de part et d'autre deux écarts types. On obtient ainsi une bande, qu'on appellera bande de dispersion, qui résume sous forme concrète ce que les deux fonctions m et σ nous apprennent sur la liaison de y à x. Cette bande a une largeur, parallèlement à l'axe des y, de $4\sigma_y$, largeur généralement variable. Si elle est constante, on dit parfois avec Karl Pearson, que la liaison est homoscédastique. Les grandeurs $m(x_0)$ et $\sigma_y(x_0)$ seront appelées respectivement moyenne liée et écart type lié.

Propriétés de l'écart type lié. — Nous prendrons comme origine les moyennes générales de x et de y. Il est clair qu'on aura

$$\mathrm{E}^x[y - m(x)]^2 = \mathrm{E}^x(y^2) - m^2(x),$$

et en prenant les moyennes M dans la loi de probabilité de x;

$$\mathrm{M}[\sigma_y^2(x)] = \sigma_2^2 - \mathrm{M}[m^2(x)], \tag{3}$$

σ_2 désigne l'écart type de la variable y.

On voit bien que s'il peut arriver que l'écart type lié surpasse l'écart type général, cela ne peut se produire pour toute valeur de x; l'écart type lié est en moyenne plus petit, d'après l'égalité (3). Il faut signaler que si $m(x)$ est identiquement nulle, c'est-à-dire si la ligne de régression est une droite parallèle à l'axe des x, la valeur moyenne de σ_y^2 ne diffère pas de σ_2^2. Il peut donc arriver que l'écart type lié soit partout égal à l'écart type général. La liaison des deux variables ne se traduit alors ni dans la moyenne, ni dans l'écart type, qui sont des constantes, et les mêmes que si les variables étaient indépendantes. De telles liaisons ont une importance pratique assez grande. Karl Pearson dit de deux variables pour lesquelles $m(x)$ est identiquement nulle ou constante, qu'elles sont sans corrélation. Lorsque la liaison est homoscédastique, elles ont alors nécessairement la deuxième propriété. On pourrait presque dire qu'elles sont indépendantes au second ordre, mais une telle locution, prise à la lettre, aurait l'inconvénient de laisser supposer que x lié par y a des propriétés analogues. Or, il n'en est rien, on s'en rend compte aisément, si l'on n'impose *a priori* aucune restriction à la loi de probabilité de x liée. Nous verrons tout à l'heure que la réciprocité, dans cette indépendance au second ordre de y à x, peut être assurée par certaines hypothèses.

Étude d'une loi de probabilité par ses moments. — Une loi de probabilité à une variable a comme caractéristiques importantes sa moyenne et son écart type (bien entendu quand ces grandeurs existent, ce qui aura toujours lieu pour les lois que nous étudierons). Ce sont les deux grandeurs :

$$\mathrm{E}(x) \quad \text{et} \quad \sigma \qquad \text{avec} \quad \sigma^2 = \mathrm{E}[x - \mathrm{E}(x)]^2.$$

Bien entendu, ces caractéristiques sont très insuffisantes, il existe une infinité de lois de probabilité ayant leurs deux premiers moments identiques. On peut ajouter à la connaissance d'une loi de probabilité en donnant en outre les valeurs des moments d'ordre supérieur.

On peut se rendre compte aisément, en supposant que la loi de probabilité soit suffisamment approchée par une loi discontinue, de l'effet de la connaissance des moments successifs (1).

En effet, si $x_1, x_2, \ldots, x_k$ sont les valeurs de la variable aléatoire, et $p_1, p_2, \ldots, p_k$ leurs probabilités respectives, le problème est de trouver ces inconnues $p_1, p_2, \ldots, p_k$, par un système d'équations qui expriment que les moments sont connus. S'il y a suffisamment d'équations, la distribution se trouve entièrement déterminée. S'il n'en est pas ainsi, le point de coordonnées positives $p_1, p_2, \ldots, p_k$ se trouve dans une région finie de l'espace, que chaque condition supplémentaire vient en général restreindre.

Pour préciser sur un exemple numérique simple, considérons les quatre points d'abscisses $\pm\frac{1}{2}$, $\pm\frac{3}{2}$; il faut y placer quatre masses positives, de somme égale à l'unité et donnant une moyenne nulle et un écart type unité. On aura les équations :

$$\begin{aligned} p_1 + p_2 + p_3 + p_4 &= 1, \\ 9(p_1 + p_2) + p_3 + p_4 &= 4, \\ 3(p_1 - p_2) + p_3 - p_4 &= 0. \end{aligned}$$

Représentons p_1 et p_2, p_3 et p_4 comme les coordonnées de deux points du plan. On vérifie aisément qu'on peut écrire

$$\begin{aligned} p_1 &= \frac{3}{16} + \varepsilon_1, & p_3 &= \frac{5}{16} - \varepsilon_2 \\ p_2 &= \frac{3}{16} - \varepsilon_1, & p_4 &= \frac{5}{16} + \varepsilon_2 \end{aligned} \qquad (\varepsilon_2 = 3\varepsilon_1).$$

(1) *Voir* l'intéressant article de R. de MISES, *The limits of a distribution function* (*Ann. Math. Statist.*, Vol. X, n° 2, juin 1939, p. 99).

Le premier point P décrit une droite parallèle à la deuxième bissectrice des axes, et qui coupe ces axes en A et B, le deuxième point Q décrit une droite analogue, dans la portion CD comprise entre les axes. On voit immédiatement que Q peut aller de C en D, mais que le point P est réduit à une portion A'B' de AB, trois fois plus petite que CD.

Ainsi, les points les plus éloignés $\left(\pm\frac{3}{16}\right)$ ont des masses qui ne peuvent se permettre que des fluctuations assez petites. Un schéma extrême serait le suivant, où p_1 a la masse la plus grande.

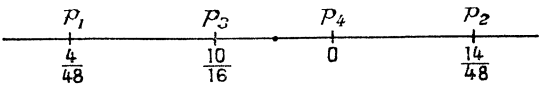


Fig. 1

Le lecteur se rendra compte aisément, par l'étude de masses placées en 6 points déterminés, par exemple ± 1, ± 2, ± 3, de l'indétermination qui subsiste dans une telle loi de probabilité quand on fixe un nombre de moments égal à 2, 3, 4. Il y a détermination complète pour 5 moments.

Cas d'une loi continue. — Il peut être suffisant, dans certaines recherches pratiques, de substituer à la loi continue une loi discontinue assez approchée. Les considérations précédentes montrent alors l'influence de la connaissance des moments successifs.

Pour une loi continue, envisagée en toute rigueur, il faut connaître, pour la déterminer complètement, des moments en nombre infini, ou mieux connaître la fonction caractéristique (ou génératrice des moments) dont il sera question un peu plus loin.

Retour à l'étude d'une loi par ses moments. — Ce que nous venons de dire peut être étendu aux lois de probabilité à deux ou plusieurs variables. La connaissance des moments jusqu'à un certain ordre limite les lois de probabilité à un certain champ qui se rétrécit quand le nombre des moments augmente.

Pour une loi à deux variables, les moments les plus employés sont ceux du premier et du second ordre. Avec les notations habituelles, nous poserons

$$\mathrm{E}(x)=x_0, \qquad \mathrm{E}(y)=y_0,$$

$$\mathrm{E}(x-x_0)^2=\sigma_1^2, \qquad \mathrm{E}(y-y_0)^2=\sigma_2^2, \qquad \mathrm{E}(x-x_0)(y-y_0)=\mu_{11}.$$

On voit qu'il s'introduit, à côté des quatre premières caractéristiques qui appartiennent aux lois marginales de x et y, un moment mixte du deuxième ordre.

Le moment le plus général a la forme.

$$\mathrm{E}(x - x_0)^f (y - y_0)^g = \mu_{fg}.$$

On voit que ces moments sont des constantes. très différentes des fonctions d'une variable que sont les moments des lois liées.

Relation entre les deux points de vue. — Elle est. dans le cas général, assez complexe. Montrons-le sur un exemple simple.

Soit la loi de probabilité élémentaire

$$0 < x < +\infty, \qquad -\infty < y < +\infty,$$
$$e^{-c}\, dx \frac{1}{\sqrt{2\pi}} e^{-\frac{[y - f(x)]^2}{2}} dy,$$

on reconnaît dans le deuxième facteur, qui donne la loi liée de y, une loi de Gauss, à courbe de régression arbitraire

$$y = f(x)$$

et dont l'écart type est l'unité.

On aura

$$\begin{aligned} \mathrm{E}(x) &= 1, \\ \mathrm{E}(y) &= \mathrm{E}[f(x)], \\ \mathrm{E}(x^2) &= 1, \\ \mathrm{E}(y^2) &= 1 + \mathrm{E}[f^2(x)], \\ \mathrm{E}(xy) &= \mathrm{E}[x f(x)]. \end{aligned}$$

Il est bien clair que le moment lié du premier ordre $f(x)$ ne saurait en général résulter de la connaissance des moments ordinaires du premier ordre [il faudrait que $f(x)$ fût une constante], ou des moments du deuxième ordre [il faudrait que $f(x)$ fût un polynome du premier degré]. En général, il sera nécessaire de connaître une infinité de moments ordinaires. et en tout cas, il faudra des moments ordinaires d'ordre supérieur, pour obtenir une détermination des moments liés.

En résumé. l'étude la plus importante. qui est celle de la variable liée, ne peut en général être résumée par la connaissance d'un nombre

fini de ces moments constants. Il était d'ailleurs bien évident que la détermination des diverses fonctions de la variable x qui sont les premiers résultats de cette étude ne pouvait être remplacée par la détermination de quelques caractéristiques numériques. Toutefois, il nous semble essentiel que ce point soit saisi avec une entière clarté.

Hypothèses simplificatrices. — Il est alors très intéressant de voir que des simplifications importantes sont introduites par les hypothèses suivantes :

1° La ligne de régression de y en x est une droite;
2° L'écart type lié est une constante.

Dans ces conditions, les deux fonctions inconnues $m(x)$ et $\sigma_y(x)$ ne dépendent que de trois constantes; les moments du premier et du second ordre suffisent alors à leur détermination.

Si l'on pose

$$m(x) = \alpha + \beta x,$$

on voit immédiatement que la droite de régression doit passer par le point de coordonnées $E(x)$, $E(y)$. Quant au coefficient angulaire β, il est donné par l'équation

$$\beta = \frac{\mu_{11}}{\sigma_1^2},$$

β est appelé coefficient de régression de y en x.

Le coefficient dit de corrélation. — On posera d'une façon générale

$$\mu_{fg} = \sigma_1^f \sigma_2^g r_{fg}.$$

Les constantes r_{fg} (de dimensions nulles) sont appelées les coefficients de corrélation. Celui d'entre eux qui correspond à μ_{11} est généralement appelé le coefficient de corrélation tout court et désigné par r. On obtient alors la formule

$$\beta = r \frac{\sigma_2}{\sigma_1}. \tag{4}$$

Valeur de l'écart type lié. — On déduit immédiatement de la formule (3) la valeur

$$\sigma_y^2(x) = \sigma_2^2[1 - r^2], \tag{5}$$

la valeur constante de l'écart type lié est donc inférieure à l'écart type général si $r \neq 0$.

La formule (5) montre qu'on a nécessairement $|r| \leqq 1$, ce qu'on peut établir directement.

Remarque. — Si la liaison de x a y vérifie aussi les deux conditions précédentes, on a

$$E^x(y) = \gamma + \delta y, \qquad \delta = r\frac{\sigma_1}{\sigma_2},$$

on voit que, si toutes ces hypothèses sont vérifiées, la seule condition $r = 0$ exprime l'indépendance jusqu'au second ordre, qui est cette fois une propriété symétrique par rapport à x et y.

(Pour toutes ces questions, on trouvera des détails dans [2], [4], [5].)

La fonction caractéristique ou génératrice des moments. — Nous allons rappeler ici la définition d'une fonction qui rend les plus grands services, tant théoriques que pratiques. C'est l'espérance mathématique de l'exponentielle e^{itx}, x est la variable aléatoire, t est une variable réelle (i est le symbole classique de l'imaginaire, qu'on introduit pour obtenir plus de généralité). Nous poserons

$$\varphi(t) = E(e^{itx}).$$

Nous donnerons les deux propriétés principales de $\varphi(t)$.

1° Si l'on ajoute deux ou plusieurs variables aléatoires indépendantes, la variable aléatoire ainsi obtenue a pour fonction caractéristique le produit des fonctions caractéristiques des variables qui composent la somme;

2° La connaissance de la fonction caractéristique détermine entièrement la loi de probabilité.

Cette fonction est appelée génératrice des moments parce que son développement suivant les puissances de t a pour coefficients les moments de la variable aléatoire. Le moment $E(x^k)$ est le coefficient de $\frac{(it)^k}{k!}$.

La deuxième fonction caractéristique. — $\varphi(t)$ étant la fonction caractéristique définie précédemment, le logarithme de $\varphi(t)$ est

appelé deuxième fonction caractéristique, la propriété 2° reste évidemment la même; pour la propriété 1° elle est remplacée par la suivante :

L'addition de variables aléatoires indépendantes se ramene à l'addition des deuxièmes fonctions caractéristiques.

Les semi-invariants d'une loi de probabilité. — Si l'on considère le développement supposé possible de cette deuxième fonction caractéristique $\psi(t)$ suivant les puissances entières de t

$$\psi(t) = k_0 + k_1 it + \ldots + k_m \frac{(it)^m}{m!} + \ldots,$$

le coefficient k_m est appelé, suivant Thiele, le semi-invariant d'ordre m de la distribution.

Et la proposition 1° prend alors la forme très simple :

Dans l'addition des variables aléatoires indépendantes, les semi-invariants s'ajoutent.

Pour cette raison, on les appelle souvent des cumulants.

Extension à une loi de probabilité à plusieurs variables. — On appelle fonction caractéristique d'une loi à deux variables x et y la fonction $\varphi(u v)$ de deux variables

$$\varphi(u v) = \mathrm{E}\, e^{i(ux+vy)}.$$

La propriété 2° garde la même forme. La propriété 1° peut être mise sous la forme suivante :

Nous appellerons point aléatoire le point M de coordonnées x et y, vecteur aléatoire le vecteur OM. On peut dire que la loi de probabilité considérée est celle du point M, ou du vecteur aléatoire OM.

Deux vecteurs aléatoires sont dits indépendants (en probabilité) si la loi de probabilité liée du deuxième vecteur est la même que la loi non liée. Dans ces conditions :

L'addition (géométrique) de deux ou de plusieurs vecteurs indépendants revient à la multiplication de leurs fonctions φ, à l'addition de leurs fonctions ψ.

Rappel de propriétés de la loi de Gauss. — On sait que la loi de

Gauss à deux variables. pratiquement appliquée avec succès à un grand nombre d'observations biométriques, fournit une illustration très simple des différentes propriétés précédentes. La densité de probabilité prend la forme

$$\frac{1}{2\pi\sqrt{1-r^2}}\, e^{-\frac{1}{2}H(xy)} \qquad \left[H(xy)=\frac{1}{1-r^2}\{x^2-2rxy+y^2\}\right].$$

L'origine est prise aux valeurs probables de x et y, les écarts types $\sigma_1\sigma_2$ sont les unités de mesure de x, y ; la constante r, telle que $|r|<1$, est le coefficient de corrélation défini plus haut.

Les deux lignes de régression sont des droites; bien entendu $\beta=\delta=r$. Les écarts types liés sont constants et leur valeur commune est par conséquent $\sqrt{1-r^2}$.

La fonction caractéristique a pour valeur

$$e^{-\frac{1}{2}K(uv)} \qquad [K(uv)=u^2+2ruv+v^2].$$

(Se reporter pour ces différents résultats classiques a l'ouvrage [5].)

Lois de probabilité à n variables. — Bien que les propriétés qui les concernent soient des généralisations toutes naturelles, il est bon d'y avoir réfléchi. Considérons donc une loi de probabilité à n variables. L'étude de cette loi peut prendre des formes assez différentes, suivant le but qu'on se propose.

a. Supposons que nous voulions étudier la liaison de x_n aux $n-1$ premieres variables. L'élément essentiel sera la loi de probabilité liée de x_n, qui dépendra comme paramètres des valeurs fixées pour les autres variables. Nous aurons donc une moyenne liée

$$m[x_1x_2\ldots x_{n-1}],$$

et un écart type lié

$$\sigma_{x_n}[x_1x_2\ldots x_{n-1}].$$

Ce seront les deux caractéristiques les plus importantes.

C'est la généralisation la plus simple; si pour fixer les idées nous supposons trois variables x, y, z, nous aurions à introduire une surface de régression, représentation de la fonction $m(xy)$, et une bande de dispersion comprise entre deux surfaces.

b. Mais on peut avoir en vue d'autres liaisons, par exemple l'autre face de la liaison de x_n à $x_1 x_2 - x_{n-1}$. On peut alors fixer la valeur de x_n et considérer la loi de probabilité à $n-1$ variables, où x_n figure comme paramètre. Dans le cas de trois variables, on aurait une loi de probabilité à deux variables x, y, mais dépendant fonctionnellement de z.

On voit que la variété des problèmes qui peuvent se poser est très grande, puisqu'on peut fixer un groupe de variables et étudier la loi de probabilité liée du groupe complémentaire.

Étude, par les moments, d'une loi à n variables. — Il est clair qu'ici la généralisation est toute naturelle. On considérera les moments successifs

$$\mathrm{E}(x_i) = \xi_i, \qquad \mathrm{E}(x_i - \xi_i)^2 = \sigma_i^2,$$
$$\mathrm{E}[x_i - \xi_i][x_k - \xi_k] = \sigma_i \sigma_k r^{ik}.$$

On a donc, n moments du premier ordre, $\frac{n(n+1)}{2}$ moments du deuxième ordre, etc. (ces moments étant d'ailleurs fournis par les polynomes du premier degré, du second degré, du développement de la fonction caractéristique).

Les difficultés déjà signalées pour deux variables se présenteront à nouveau si l'on veut déduire des ces moments les caractéristiques d'une loi de probabilité liée.

Hypothèses simplificatrices. — On obtient des resultats simples en faisant les mêmes hypothèses :

1° La fonction $m(x_1 x_2 \ldots x_{n-1})$ est linéaire en $x_1 x_2 \ldots x_{n-1}$.

2° L'écart type lié est une constante.

Dans ces conditions, on peut, à l'aide des seuls moments du premier et deuxieme ordre, déterminer m et σ_{1_n}. Les résultats sont tout à fait analogues à ceux relatifs à deux variables, mais un peu plus compliqués.

Quelques formules importantes. — Nous allons faire l'hypothèse, toujours permise, que les espérances mathématiques de toutes les variables sont nulles, et que leurs écarts sont tous égaux a l'unité.

Si l'on pose alors

$$m = \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_{n-1} x_{n-1},$$

on aura en multipliant cette égalité par $x_1 x_2 \ldots x_{n-1}$ et prenant les espérances mathématiques

$$(I) \qquad \begin{cases} r_{1n} = \beta_1 + \beta_2 r_{12} + \ldots + \beta_{n-1} r_{1,n-1}, \\ r_{2n} = \beta_1 r_{21} + \beta_2 + \ldots + \beta_{n-1} r_{2,n-1}, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots, \\ r_{n-1,n} = \beta_1 r_{n-1,1} + \ldots + \beta_{n-1}. \end{cases}$$

Pour la valeur de $\sigma^2_{x_n}$ (carré de l'écart type lié) on a toujours

$$\sigma^2_{x_n} = 1 - m^2(x_1 x_2 \ldots x_n).$$

Si l'on prend de nouveau la valeur moyenne M (dans la loi de probabilité générale, ou, ce qui revient au même, dans la loi de probabilité de $x_1 x_2 \ldots x_n$) on aura,

$$(II) \qquad M(\sigma^2_{x_n}) = 1 - R^2 \qquad \left[R^2 = M[m^2(x_1 x_2 \ldots x_n)]\right].$$

On a tout de suite la valeur de R^2

$$R^2 = \beta_1^2 + \beta_2^2 + \ldots + \beta_{n-1}^2 + 2\beta_1\beta_2 r_{12} + \ldots,$$

qu'on peut mettre sous la forme

$$R^2 = \beta_1 r_{1n} + \beta_2 r_{2,n} + \ldots + \beta_{n-1} r_{n-1,n}.$$

La quantité R^2 qui est évidemment inférieure à l'unité (généralisation de la propriété du coefficient de corrélation) peut être calculée directement à l'aide des équations (I) par élimination de $\beta_1 \beta_2 \ldots \beta_{n-1}$.

On obtient alors sous forme de déterminants

$$R^2 = -\frac{\Delta_0}{\Delta},$$

$$\Delta_0 = \begin{vmatrix} 1 & r_{12} & \ldots & r_{1n} \\ r_{21} & 1 & \ldots & r_{2n} \\ \ldots & \ldots & \ldots & \ldots \\ r_{n1} & r_{n2} & \ldots & 1 \end{vmatrix}, \qquad \Delta = \begin{vmatrix} 1 & r_{12} & \ldots & r_{1n-1} \\ \ldots & \ldots & \ldots & \ldots \\ r_{n-1,1} & \ldots & \ldots & 1 \end{vmatrix}.$$

Bien entendu le premier membre de (II) donne la valeur de l'écart type lié quand celui-ci est constant (*voir* aussi [2] au chapitre « multiple corrélation »).

Remarque générale. — Il est un peu inquiétant de n'obtenir de conclusions simples qu'à l'aide de ces hypothèses assez restrictives. D'autre part, l'emploi des moments d'ordre supérieur, s'il devient nécessaire, est terriblement lourd. Les calculs précédents, exécutés dans le cas général, ne conduisent plus à m et σ, mais on peut cependant en tirer quelque chose. Plaçons-nous seulement dans le cas de trois dimensions. Les résultats obtenus seront généraux. Considérons le plan

$$z = ax + by$$

et cherchons à le déterminer par la méthode des moindres carrés, en comptant les distances parallèlement à oz, et affectant chaque élément de l'espace de la probabilité qui lui correspond. Nous sommes conduit à former

$$\mathrm{E}(z - ax - by)^2.$$

Les deux conditions de minimum sont justement les conditions (I).

Quant à la valeur du minimum, elle est précisément $1 - \mathrm{R}^2$, comme le montre un calcul très facile.

Ainsi le plan déterminé par les équations ne sera plus, dans le cas général, un plan de régression, il ne contiendra pas les points représentatifs des moyennes liées, mais il leur sera associé par une condition de moindres carrés, et le carré moyen de la distance de la distribution de probabilité à ce plan sera précisémet $1 - \mathrm{R}^2$.

Cette dernière quantité nous indiquera donc, assez grossièrement il est vrai, et en moyenne, si un point de la distribution peut se trouver éloigné du plan des moindres carrés.

Il est à remarquer que la quantité $1 - \mathrm{R}^2$ n'est plus $\mathrm{M}(\sigma_{x_n}^2)$. Soit en effet G le point $z(xy)$ qui représente la moyenne liée, et P le point contenu dans le plan, c'est-à-dire le point d'ordonnée $ax + by$

$$z - ax - by = z - z_{\mathrm{G}} + z_{\mathrm{G}} - ax - by.$$

On en déduit

$$\mathrm{E}(z - ax - by)^2 = \mathrm{M}[\sigma_z^2] + \mathrm{M}(\mathrm{PG})^2. \tag{III}$$

Donc la valeur de $1 - \mathrm{R}^2$ est toujours plus grande que le carré moyen de σ_z, sauf si PG est identiquement nulle, c'est-à-dire si le plan est vraiment de régression.

Relation avec le problème général de la psychologie. — Il faut étudier comment sont liées un certain nombre d'aptitudes, c'est-à-dire étudier une loi de répartition à n variables. Suivant que certaines aptitudes sont plus faciles à atteindre que d'autres, il y aura lieu de les prendre comme variables de base, et d'étudier la loi de probabilité liée des autres variables.

CHAPITRE II.

Nous voici donc en présence d'une population observable, considérée par nous comme une épreuve. comme un extrait de cette population très étendue où pourraient se présenter, nuancés jusqu'à être continus. les caractères dont nous voulons faire l'étude.

Le problème est d'estimer, à l'aide de la population partielle, les caractéristiques de la répartition de la population générale. Nous supposerons d'abord, dans ce chapitre, que les caractères mesurés sont obtenus sans erreur appréciable.

Un cas simple. Estimation de la moyenne. — Considérons le caractère x et proposons-nous d'estimer la valeur moyenne générale, ou l'espérance mathématique de x dans la loi de probabilité cherchée. Nous disposons de n mesures, sur des individus pris au hasard, mesure que nous considérons comme des grandeurs aléatoires suivant la même loi de probabilité : ces grandeurs seront regardées comme indépendantes au sens des probabilités. Toute fonction donnée de ces grandeurs aléatoires est elle-même une grandeur aléatoire, par conséquent l'information que nous pouvons déduire des mesures conserve nécessairement ce caractère aléatoire. Le service rendu dans cette matière par la statistique mathématique est d'évaluer avec précision le risque couru, et de choisir dans certains cas la solution présentant le moindre risque d'erreur.

Nous cherchons ici une caractéristique de la distribution, définie par la propriété d'être la moyenne générale des x. Il nous faut constituer une variable aléatoire, fonction des observations, qui soit une variable assez concentrée, et si possible concentrée exactement autour de la valeur cherchée.

Nous formons pour cela la moyenne arithmétique des mesures,

soit

$$\xi = \frac{x_1 + x_2 + \ldots + x_n}{n}.$$

Bien que nous sachions peu de chose sur la distribution des x, puisque nous l'étudions, nous savons beaucoup sur cette variable ξ, à condition que n soit assez grand. En effet, d'après un théorème classique, ξ a une distribution voisine de celle de Gauss, avec une valeur moyenne qui est justement l'inconnue cherchée, et un écart type qui est $\frac{\sigma_1}{\sqrt{n}}$, σ_1 étant l'écart type de x. Cette variable est donc, si n est grand, concentrée autour de la valeur désirée, et nous savons évaluer avec précision la probabilité pour qu'elle s'écarte de cette valeur (à condition de connaître σ_1).

Remarque. — Si nous connaissions *a priori* quelques propriétés de la distribution de x, il se pourrait que la méthode précédente ne fût pas la seule naturelle, et une autre méthode pourrait être plus avantageuse.

Un cas classique est celui où la distribution de x serait de Gauss; la moyenne de x est alors en même temps valeur médiane, et au lieu de prendre ξ comme estimation, on peut prendre la valeur médiane des mesures. On démontre que cette valeur médiane suit aussi, quand n est grand, une loi voisine de la loi de Gauss, avec l'écart type $\sqrt{\frac{\pi}{2}}\frac{\sigma_1}{\sqrt{n}}$. Elle est donc plus dispersée que la variable ξ, et celle-ci demeure la plus avantageuse, mais pour d'autres lois de probabilité, les circonstances peuvent être autres, et la moyenne des x_i n'est pas nécessairement une estimation privilégiée (1).

Estimation de l'écart type de x. — Pour cela, on forme comme il est classique, la quantité

$$s'^2 = \frac{(x_1 - \xi)^2 + \ldots + (x_n - \xi)^2}{n - 1}.$$

Sa valeur moyenne est σ_1^2, et sa distribution quand n est grand est voisine d'une loi de Gauss, mais la dispersion fait intervenir le moment du quatrième ordre de la variable x, de sorte qu'une idée un peu précise des risques d'erreur exige une appréciation de ce

(1) *Voir* pour ces questions [3], [11], [14], [27].

moment. On se contente souvent d'une valeur de l'écart type de la grandeur s', de la forme $\frac{\sigma_1}{\sqrt{2n}}$; cette expression, valable quand x suit une loi de Gauss, et dans d'autres hypothèses plus générales, résulte tout simplement de la valeur

$$\sigma_{s'} = \sqrt{\frac{\mu_4 - \mu_2^2}{n}}.$$

En résumé, il faut utiliser une hypothèse qui permette le calcul de μ_4.

Tous ces résultats sont simples et clairs dans leurs principes.

Le moment mixte du second ordre et le coefficient de corrélation. — Nous avons vu que le premier élément qui ait à intervenir dans la liaison de deux variables aléatoires est le moment μ_{11} ou, ce qui revient au même, le coefficient de corrélation r. Quand les deux hypothèses simplificatrices que nous avons faites sont vérifiées, ce coefficient de corrélation donne à lui tout seul des renseignements substantiels. Son estimation est donc un problème important, qu'on a traité suivant les mêmes idées générales. En posant

$$\xi = \frac{x_1 + x_2 + \ldots + x_n}{n}, \qquad \eta = \frac{y_1 + y_2 + \ldots + y_n}{n},$$

$$s_1'^2 = \frac{\Sigma(x_i - \xi)^2}{n-1}, \qquad s_2'^2 = \frac{\Sigma(y_i - \eta)^2}{n-1},$$

$$\mu_{11}' = \frac{\Sigma(x_i - \xi)(y_i - \eta)}{n-1}.$$

Le rapport $r' = \frac{\mu_{11}'}{s_1' s_2'}$ est une variable aléatoire, qui lorsque n est très grand, a bien les propriétés requises.

Cette variable est concentrée autour de la valeur inconnue r, elle suit une loi voisine de la loi de Gauss, et son écart type est voisin de

$$\frac{1 - r^2}{\sqrt{n}}.$$

Mais une loi de repartition peut tendre plus ou moins vite vers la loi de Gauss, et notre variable r', qui suit cette loi à la limite, a en vérité une répartition qui dépend très notablement de la valeur r et du nombre n. Ce fait est très gênant parce qu'on ne sait plus bien dans un tel cas ce que signifie l'écart type. Alors qu'avec la loi de Gauss, cet écart type associé à la table classique permet de porter avec

précision des jugements sur les risques qu'on court, il est tout à fait insuffisant pour une loi très dissymétrique. Ces inconvénients, nettement mis en évidence par R. A. Fisher [12], peuvent être corrigés par les méthodes qu'il a données. Nous indiquons ici le résultat; si l'on substitue à la variable r' une variable z donnée par

$$z = \frac{1}{2} \operatorname{Log} \frac{1 + r'}{1 - r'} \qquad (r' = \operatorname{th} z)$$

(*th* étant le signe de la tangente hyperbolique). la nouvelle variable z, qui va de $-\infty$ à $+\infty$, a une loi de répartition qui tend beaucoup plus rapidement vers sa limite, et dont l'écart type a le très grand avantage de ne pas dépendre (en pratique) de r, sa valeur approchée étant

$$\frac{1}{\sqrt{n-3}}.$$

Quelques indications numériques sur les avantages de la méthode de R. A. Fisher. — Il est bien clair qu'il est préférable d'employer une méthode exacte, mais il est bon de voir à quel point la méthode classique devient inexacte pour de faibles valeurs de n.

Supposons que la théorie assigne une loi de Gauss, avec un coefficient de corrélation

$$r = 0.8005.$$

On a fait neuf observations, et obtenu la valeur empirique

$$r' = 0,4053.$$

La méthode classique formera

$$\frac{1 - r^2}{\sqrt{n}} = 0,12.$$

or, l'écart observé est

$$r - r' = 0,3952.$$

Il est supérieur à trois écarts types, et conduirait à rejeter la théorie proposée.

En réalité, avec la variable z, on aura :

$$z = 1,1, \qquad z' = 0,43, \qquad z - z' = 0,67,$$

$$\sigma_{z'} = \frac{1}{\sqrt{6}} = 0,41.$$

Cette fois l'écart est inférieur à deux écarts types, il n'y a pas de raison de suspecter la théorie.

La valeur et la signification des résultats. — Nous avons vu que nos estimations conservent toujours le caractère aléatoire de l'épreuve qui les a fournis. Dans le cas qui nous intéresse, le coefficient de corrélation estimé est affecté d'une erreur, qui peut modifier profondément la conclusion à tirer. Si les hypothèses de régression linéaire et de dispersion constante sont vérifiées, une valeur nulle de r entraîne ce que nous avons appelé l'indépendance au second ordre pour la liaison de y à x. Une valeur non nulle de r entraîne la dépendance. Pour une valeur donnée de l'estimation r', la valeur réelle de r est-elle nulle ou différente de zéro? Aucune conclusion ferme ne peut naturellement être donnée. On peut seulement dire ceci :

Plaçons autour de la valeur expérimentale r' un intervalle qui ait une probabilité donnée (disons 95/100) de contenir la valeur vraie. Adoptons comme règle que si la valeur zéro est en dehors de cet intervalle, la valeur de r' sera dite significative d'une dépendance; si la valeur zéro est à l'intérieur, la valeur r' sera dite non significative. Avec ces conventions les valeurs significatives ne peuvent se produire qu'avec la probabilité 0,05 quand la valeur réelle de r est nulle.

Bien entendu, la délimitation précise de cet intervalle autour de r' ne peut se faire que si l'on a la loi de probabilité de r'. Elle est particulièrement facile à exécuter si l'on a affaire à une loi de Gauss, d'écart type donné. C'est ce qui constitue l'avantage de la variable z de R. A. Fisher. Il suffit, autour de la valeur z', de porter deux écarts types à droite et à gauche pour obtenir (à peu de chose près) l'intervalle à 95/100.

(Pour une étude très générale de ces questions, *voir* [29].)

CHAPITRE III.

LE PROBLÈME RÉEL. RÔLE DES ERREURS.

Nous avons jusqu'ici supposé que les caractères mesurés étaient obtenus sans erreur. Il n'en est rien, et c'est ce qui complique très sérieusement la question. Les aptitudes dont nous parlons, ce sont les réussites dans certaines tâches. Un caractère tel que la largeur des

épaules, les dimensions du crâne, mesuré par comparaison avec une unité déterminée. fournira pour un individu donné, des valeurs très stables. Au contraire, dans une certaine épreuve logique, même maintenue semblable à elle-même, la réussite n'est pas toujours la même. Si on la cote de manière déterminée, les nombres obtenus sont fluctuants. L'épreuve, considérée comme une méthode de mesure, comporte une erreur. Nous admettrons qu'il existe une certaine cote vraie, valable pour la population des épreuves d'une certaine nature, mais que pour une des épreuves de cette population, on obtient la cote vraie x, augmentée d'une erreur ε, soit

$$X = x + \varepsilon.$$

Nous cherchons la loi de répartition des caractères analogues à x, et nos mesures ne nous fournissent que ξ. On voit que le problème, sous cette forme, est très général, et se pose dans d'autres champs que celui de la psychologie.

La loi d'erreur. — Supposons, pour fixer les idées, qu'il s'agit de deux caractères x et y. On aura :

$$X = x + \varepsilon,$$
$$Y = y + \eta.$$

La loi cherchée est une répartition des points xy. Disons que la distribution générale se traduirait par une certaine densité de ces points. Si nous disposions des mesures directes de x et y exécutées sur n individus, nous aurions à estimer cette densité à l'aide du nuage discontinu des n points $x_i y_i$. En réalité, nous n'avons que les n points $X_i Y_i$.

Imaginons que l'on puisse, sur le même individu, exécuter un grand nombre d'épreuves du même type. Nous dirons que l'on vise le même point xy. On obtiendra autant de points XY, répartis autour du point xy; nous considérons ce nuage discontinu comme une épreuve faite sur la loi d'erreur affectée au point xy.

En vérité, cette loi d'erreur, loi de probabilité aux deux variables $\varepsilon\eta$, est la loi de probabilité liée de ces deux variables, pour x et y donnés.

Hypothèse fondamentale. — Pour obtenir des résultats simples, nous ferons l'hypothèse, qui ne paraît pas déraisonnable, que cette

loi d'erreur liée est indépendante du point xy, c'est-à-dire que le couple aléatoire $\xi\eta$ et le couple aléatoire xy sont indépendants.

Dans ces conditions, la loi de probabilité du point observé XY est très simple, ou plutôt se déduit très simplement des deux lois (xy) et $(\zeta\eta)$. Il suffit, nous le savons, de multiplier les fonctions caractéristiques φ, ou d'ajouter les fonctions caractéristiques ψ.

Nous supposerons que l'on a :

$$\mathrm{E}(x) = 0, \qquad \mathrm{E}(y) = 0,$$

et

$$\mathrm{E}(\xi) = 0, \qquad \mathrm{E}(\eta) = 0.$$

Dans ces conditions la fonction caractéristique ψ_{XY} a la forme

$$\psi_{xy}(u v) = -\frac{\sigma_1^2 u^2 + 2\sigma_1\sigma_2 r u v + \sigma_2^2 v^2}{2} + \dots,$$

$$\psi_{\varepsilon\eta}(u v) = -\frac{s_1^2 u^2 + 2 s_1 s_2 \rho u v + s_2^2 v^2}{2} + \dots.$$

On aura donc avec des notations évidentes :

$$(6) \qquad \begin{cases} \Sigma_1^2 = \sigma_1^2 + s_1^2, \\ \Sigma_1\Sigma_2 \mathrm{R} = \sigma_1\sigma_2 r + s_1 s_2 \rho \qquad \text{ou} \qquad \mathrm{E}(XY) = \mathrm{E}(xy) + \mathrm{E}(\xi\eta), \\ \Sigma_2^2 = \sigma_2^2 + s_2^2. \end{cases}$$

Les autres formules, qu'on obtiendrait en considérant les termes de degrés supérieurs, expriment que les semi-invariants de la loi observée sont la somme des semi-invariants de la loi cherchée et de la loi d'erreur. On voit que la méthode naturelle est d'isoler, si l'on peut, la loi d'erreur, de la déterminer d'abord, et d'en déduire la loi du point xy.

Estimation des caractéristiques de la loi d'erreur. — S'il s'agissait en général d'isoler une loi d'erreur, il suffirait de recommencer la mesure, de manière à obtenir au moins deux couples XY pour chaque couple xy. Dans ces conditions, le vecteur

$$\mathrm{X}^2 - \mathrm{X}^1 = \xi^2 - \xi^1,$$
$$\mathrm{Y}^2 - \mathrm{Y}^1 = \eta^2 - \eta^1$$

ne dépend plus que de la loi d'erreur. Nous admettrons que les deux erreurs $\zeta^1\eta^1$, $\zeta^2\eta^2$ sont indépendantes. Nous avons alors à résoudre un problème assez simple :

Estimer les caractéristiques d'une loi de probabilité à deux variables connaissant un certain nombre d'observations faites sur une loi qui lui est liée très étroitement. En effet, les deux variables

$$\alpha = \xi^2 - \xi^1,$$
$$\beta = \eta^2 - \eta^1$$

ont comme fonction caractéristique de leur loi de probabilité

$$\mathrm{E}\, e^{i(u\alpha+v\beta)} = \varphi(uv)\,\varphi(-u-v).$$

Donc les semi-invariants sont liés par les relations suivantes :

Les semi-invariants d'ordre pair sont doubles des semi-invariants cherchés, les semi-invariants d'ordre impair sont nuls.

Ainsi, nous obtiendrons immédiatement tous les semi-invariants d'ordre pair. Ce sont sans doute les plus utiles, et en fait on se contente à peu près toujours des moments du deuxième ordre. Pourtant, on peut indiquer une méthode qui permet d'obtenir les autres moments. Il suffit de faire au moins trois mesures, soit alors

$$\begin{aligned} X^1 &= x + \xi^1, & Y^1 &= y + \eta^1, \\ X^2 &= x + \xi^2. & &\ldots\ldots\ldots, \\ X^3 &= x + \xi^3, & &\ldots\ldots\ldots \end{aligned}$$

Les combinaisons

$$\alpha X^1 + \beta X^2 + \gamma X^3, \quad \alpha Y^1 + \beta Y^2 + \gamma Y^3$$

font disparaître x et y si $\alpha + \beta + \gamma = 0$. On obtient alors pour la deuxième fonction caractéristique

$$(\alpha^2 + \beta^2 + \gamma^2)\,\psi_2(uv) + (\alpha^3 + \beta^3 + \gamma^3)\,\psi_3(uv) + \ldots.$$

Il suffit donc d'employer les diviseurs $\alpha^k + \beta^k + \gamma^k$ pour obtenir tous les groupes homogènes.

Ainsi, le problème ne présente aucune difficulté théorique, dès que chaque mesure est répétée au moins trois fois.

Forme pratique de recherche de la loi d'erreur. — En fait, les épreuves employées pour coter une aptitude sont composées d'une série de questions à résoudre, et la cote est la somme des points obtenus pour toutes ces questions. Chaque question fournit une mesure a_i et l'on utilise les deux sommes $\Sigma a_i\ \Sigma b_i$ pour les deux aptitudes étudiées.

En vérité, et pour être plus net, nous considérerons les moyennes

$$X = \frac{\Sigma a_i}{k} = x + \frac{e_1 + e_2 + \ldots + e_k}{k},$$
$$Y = \frac{\Sigma b_i}{k} = y + \frac{f_1 + f_2 + \ldots + f_k}{k},$$

k étant le nombre des épreuves. Une série de k épreuves est composée de deux séries de $\frac{k}{2} = h$ épreuves.

Ainsi, nous obtenons deux couples de cotes partielles

$$\begin{cases} X = x + \dfrac{e'_1 + \ldots + e'_h}{h}, \\ Y' = y + \dfrac{f'_1 + \ldots + f'_h}{h}; \end{cases}$$
$$\begin{cases} X'' = x + \dfrac{e''_1 + \ldots + e''_h}{h}, \\ Y'' = y + \dfrac{f''_1 + \ldots + f''_h}{h}. \end{cases}$$

On calcule aisément les moments du second ordre en fonction des écarts types a et b de e et f

$$\begin{aligned} E(X'^2) &= E(X''^2) &&= \sigma_1^2 + \frac{a^2}{h}, \\ E(Y'^2) &= E(Y''^2) &&= \sigma_2^2 + \frac{b^2}{h}, \\ E(X'Y') &= E(X''Y'') &&= E(xy) + \frac{1}{2}E(ef), \\ E(XX'') &= \sigma_1^2 &&= E(Y'Y'') = \sigma_2^2. \end{aligned}$$

On en déduit immédiatement

$$E(X' - X'')^2 = \frac{2a^2}{h},$$

En réalité, la grandeur intéressante est $\frac{a^2}{k}$

$$(7) \qquad \begin{cases} \dfrac{a^2}{k} = \dfrac{1}{4}E(X' - X'')^2, \\ \dfrac{b^2}{k} = \dfrac{1}{4}E(Y' - Y'')^2, \\ \text{et, de même,} \\ \dfrac{E(ef)}{k} = \dfrac{1}{4}E(X' - X'')(Y' - Y''). \end{cases}$$

Ces formules sont fondamentales.

Les formules pour l'allongement d'un test. — On peut caractériser la précision d'un test par la quantité $\frac{a^2}{k}$, mais on le fait généralement par le coefficient de corrélation de deux cotes partielles, telles que X′ et X″

$$r_{X'X''} = \frac{\sigma_1^2}{\sigma_1^2 + \frac{a^2}{h}} = \frac{1}{1 + \frac{c^2}{h}} \qquad \left(c^2 = \frac{a^2}{\sigma_1^2}\right).$$

On a donc

$$\frac{c^2}{h} = \frac{1}{r_h} - 1,$$

r_h étant le coefficient de corrélation de deux cotes obtenues par des tests de longueur h.

Si maintenant le même test devient p fois plus long

$$r_{ph} = \frac{1}{1 + \frac{c^2}{ph}} = \frac{pr_h}{(p-1)r_h + 1}. \tag{8}$$

Cette formule (8) est dite généralement formule de Spearman-Brown (*voir* [7] et [1]). On voit qu'elle n'est que la traduction, un peu compliquée par l'écriture, du fait fondamental que la précision croît comme la racine carrée du nombre des épreuves.

S'il n'est pas déraisonnable de supposer p très grand, on voit que r_{ph} serait très voisin de l'unité, ce qui veut dire qu'on éliminerait à peu près l'erreur de mesure.

La formule habituelle. — En général, on coupe seulement un test en deux, par exemple en prenant la série des questions impaires et des questions paires. On obtient alors la formule très employée :

$$r_k = \frac{2r_h}{1 + r_h} \qquad \left(h = \frac{k}{2}\right).$$

Formule de Spearman pour l'atténuation. — Les formules (6) données plus haut mettent en évidence un élargissement du nuage de points (xy) par l'effet des erreurs $(\xi\eta)$. C'est cet élargissement du nuage théorique que nous avons cherché à déterminer.

Spearman, à qui ces théories doivent des résultats essentiels, s'était proposé d'obtenir, par une formule simple, le coefficient de corrélation entre les cotes théoriques xy. Il introduisit une hypothèse

simplificatrice

$$\mathrm{E}(ef) = 0,$$

qui fournit en effet un résultat très élégant.

On a évidemment, dans cette hypothèse,

$$r_{XY} = \frac{\mathrm{E}(xy)}{\sigma_1 \sigma_2} \frac{\sigma_1 \sigma_2}{\Sigma_1 \Sigma_2}.$$

Le deuxième facteur, à cause de l'élargissement du nuage, est toujours inférieur à l'unité. On a donc

$$r_{XY} < r_{xy}.$$

C'est ce que Spearman appela l'effet d'atténuation

Il est bien clair alors qu'en posant

$$r^X_{12} = \frac{\mathrm{E}(X^1 X^2)}{\Sigma_1^2}, \qquad r^Y_{12} = \frac{\mathrm{E}(Y^1 Y^2)}{\Sigma_2^2},$$

on a finalement la formule célèbre

$$r_{X1} = r_{xy} \sqrt{r^X_{12} r^Y_{12}}. \tag{9}$$

Les coefficients r^X_{12}, r^Y_{12}, indices de précision des mesures X et Y, sont généralement appelés « reliability coefficients » ([1]).

Traitement plus général. — Mais nous croyons préférable de ne pas faire l'hypothèse de Spearman et nous ne supposerons rien sur la liaison qui peut exister entre e et f. Nous allons examiner ce qu'on pourrait obtenir avec trois observations.

On suppose donc un test de longueur $3h$, comportant trois séries de longueur h. On aura, bien entendu, les formules précédentes pour $X'X''X'''$, $Y'Y''Y'''$.

Nous formerons alors. en supposant

$$\alpha + \beta + \gamma = 0,$$

$$\mathrm{E}(\alpha X' + \beta X'' + \gamma X''')^2 = (\alpha^2 + \beta^2 + \gamma^2)\frac{a^2}{h},$$

$$\mathrm{E}(\alpha Y' + \beta Y'' + \gamma Y''')^2 = (\alpha^2 + \beta^2 + \gamma^2)\frac{b^2}{h},$$

$$\mathrm{E}(\alpha X' + \beta X'' + \gamma X''')(\alpha Y' + \beta Y'' + \gamma Y''') = (\alpha^2 + \beta^2 + \gamma^2)\frac{\mathrm{E}(ef)}{h}.$$

([1]) Coefficients de fidélité (terminologie fixée à la Conférence Psychotechnique de Moscou en 1931).

On aurait donc, pour le test de longueur $3h$,

$$\frac{a^2}{3h} = \frac{1}{3(\alpha^2+\beta^2+\gamma^2)} E(\alpha X' + \ldots)^2,$$

$$\frac{E(ef)}{3h} = \frac{1}{3(\alpha^2+\beta^2+\gamma^2)} E(\alpha X' + \ldots)(\alpha Y' + \ldots),$$

$$\frac{b^2}{3h} = \frac{1}{3(\alpha^2+\beta^2+\gamma^2)} E(\alpha Y' + \ldots)^2.$$

Les moments d'ordre supérieur s'obtiennent par des formules analogues qu'il est inutile d'écrire.

Quelques formules classiques. — L'introduction du coefficient de corrélation r_k qui est équivalent à la dispersion de l'erreur ε donne quelques formules simples, que nous signalons

$$\Sigma_1 \sqrt{r_k} = \sigma_1, \qquad s_1 = \Sigma_1 \sqrt{1 - r_k}.$$

Considérons d'autre part la cote observée X, il est naturel de la prendre comme estimation de x, et le risque d'erreur est caractérisé par s_1. Cependant, X *étant connu*, quelle est la valeur moyenne de x. On l'obtient immédiatement *dans l'hypothèse d'une régression linéaire de x en* X. En effet, la relation

$$X = x + \varepsilon$$

établit la régression linéaire de X en x, donc le coefficient de corrélation est $\rho = \frac{\sigma_1}{\Sigma_1}$ et l'équation de régression de x en X est

$$\frac{M(x)}{\sigma_1} = \frac{\sigma_1}{\Sigma_1}\frac{X}{\Sigma_1}, \qquad M(x) = X\left(\frac{\sigma_1}{\Sigma_1}\right)^2 = X r_k.$$

Ainsi, la valeur moyenne de x est $r_k X$, et non pas X.

L'écart type lié, d'après la formule générale (si la liaison est à dispersion constante) a pour valeur

$$\sigma_1 \sqrt{1-\rho^2} = \Sigma_1 \sqrt{r_k(1-r_k)}$$

(*voir* pour ces questions [1], [2] et [4]).

Estimations de ces différentes grandeurs. — En vérité, la situation fournie par les observations est la suivante : on dispose des mesures faites sur n individus, dans une épreuve de longueur k.

Cette épreuve est fractionnée en deux parties de longueur $h = \frac{k}{2}$.

On peut ainsi construire une estimation r'_h de la grandeur r_h, et nous savons comment apprécier l'erreur de cette estimation.

La valeur de l'estimation r'_k de r_k est donnée par la formule

$$r'_k = \frac{2 r_h}{1 + r_h}.$$

Des risques d'erreur sur r'_h, nous pouvons déduire le risque sur r'_k.

Il serait, à notre avis, beaucoup plus simple et plus net de considérer, comme nous l'avons fait, les valeurs de $X' - X''$ et d'utiliser les formules classiques qui donnent la précision de l'écart type ainsi obtenu.

Bien entendu il faudrait estimer, comme il a été indiqué, la valeur $E(ef)$ et par les mêmes méthodes, indiquer la précision avec laquelle cette grandeur est connue.

CHAPITRE IV.

RÉDUCTION AU NOMBRE MINIMUM D'APTITUDES DÉTERMINANTES.

Lois réductibles. — Considérons n aptitudes dont nous étudions la loi de répartition. Imaginons encore que la mesure puisse être faite exactement et fournisse pour chaque individu un groupe de n valeurs, $x_i^1, x_i^2, \ldots, x_i^n$. Nous associons à cet individu un point de l'espace à n dimensions. A la population générale correspond un nuage continu dont nous étudions la structure. Nous pouvons *a priori* obtenir différents types. Si tous les individus étaient identiques, nous aurions un seul point (nuage à zéro dimension). Les points peuvent être répartis suivant une courbe, ou suivant une multiplicité à 2, 3, ... dimensions, jusqu'a n.

Si le nuage est à moins de n dimensions, nous dirons que la loi est réductible.

Il est clair, par exemple, que si l'on considère une population d'êtres semblables entre eux, une dimension linéaire, la surface et le volume d'un individu constituent trois variables aléatoires telles que le nuage soit réduit à une courbe. Si l'on suppose que ces êtres n'ont

pas une densité constante, et que la troisième variable soit la masse au lieu du volume, on aura un nuage aplati sur une surface, la répartition sur cette surface achevant de déterminer la loi étudiée.

Pour une multiplicité à $n - p$ dimensions, il doit exister p relations entre les variables $x_1, x_2, \ldots, x_n$. Nous nous proposons de rechercher ces relations, en les supposant du premier degré.

Cas d'une seule relation. — Supposons, pour rendre les choses plus intuitives, que nous étudions seulement trois caractères. Nous nous trouvons dans l'espace à trois dimensions. S'il existe une seule relation, les points x_1, x_2, x_3 doivent se trouver dans un plan. Si nos trois caractères étaient mesurables avec beaucoup de précision, si par exemple il s'agissait des trois angles d'un triangle, nous aurions une excellente détermination de ce plan par une méthode à peu près quelconque, et les considérations mathématiques ne seraient guère que raffinements sans grande utilité.

Mais nous nous trouvons dans un cas tout différent, où les erreurs sont assez importantes, et le choix de la méthode peut avoir une très grande influence sur la précision des résultats. Nous avons à faire passer au mieux, un plan au travers d'un nuage de points. Pour cela il est essentiel de connaître quelque chose sur la loi d'erreur.

Forme adoptée pour la loi d'erreur. — Soit M_i un point aux trois coordonnées x_1, x_2, x_3, soit P_i le point obtenu par les mesures, ses coordonnées sont X_1, X_2, X_3. Nous admettrons que le vecteur aléatoire M_iP_i suit une loi de Gauss de centre M_i. Cette hypothèse est peut-être un peu forcée, mais elle n'est pas déraisonnable. Elle permet d'introduire dans la loi d'erreur tous les moments du second ordre, et ces moments sont ceux auxquels on se borne généralement. Ainsi, posant

$$\begin{aligned} X_1 &= x_1 + \varepsilon_1, \\ X_2 &= x_2 + \varepsilon_2, \\ X_3 &= x_3 + \varepsilon_3, \end{aligned}$$

le vecteur $\varepsilon_1 \varepsilon_2 \varepsilon_3$ a pour loi de probabilité élémentaire

$$A e^{-\frac{1}{2} H(\varepsilon_1 \varepsilon_2 \varepsilon_3)} d\varepsilon_1 d\varepsilon_2 d\varepsilon_3,$$

$H(\varepsilon_1 \varepsilon_2 \varepsilon_3)$ étant une forme quadratique homogène qui égalée à une

constante positive, représente un ellipsoïde. Ces ellipsoïdes de la loi d'erreur peuvent avoir une forme quelconque, nous ne faisons sur eux aucune hypothèse restrictive. Rappelons que la fonction caractéristique de cette loi de Gauss est exprimée par

$$E[e^{i(u_1\varepsilon_1+u_2\varepsilon_2+u_3\varepsilon_3)}] = e^{-\frac{1}{2}K(u_1u_2u_3)},$$

K étant une forme quadratique qui est la réciproque de H, c'est-à-dire que l'ellipsoïde

$$H(\varepsilon_1\varepsilon_2\varepsilon_3) = 1$$

est tangent au plan

$$u_1\varepsilon_1 + u_2\varepsilon_2 + u_3\varepsilon_3 = 1$$

quand on a la condition

$$K(u_1u_2u_3) = 1.$$

On peut mettre cette condition sous une forme un peu différente. L'ellipsoïde

$$H(\varepsilon_1\varepsilon_2\varepsilon_3) = \lambda^2$$

est tangent au plan

$$u_1\varepsilon_1 + u_2\varepsilon_2 + u_3\varepsilon_3 = u_4$$

quand on a la condition

$$K(u_1u_2u_3) = \frac{u_4^2}{\lambda^2}.$$

La forme quadratique $K(u_1u_2u_3)$ est, nous le savons, liée directement aux moments du second ordre de la loi d'erreur. On a évidemment

$$E[u_1\varepsilon_1 + u_2\varepsilon_2 + u_3\varepsilon_3]^2 = K(u_1u_2u_3).$$

Méthode de recherche du plan. — Supposons maintenant qu'on dispose de n points P_i. Il s'agit de rechercher le plan passant par les points M_i correspondants. Or, *si les points* M_i *sont considérés comme fixes et donnés*, la probabilité d'obtenir l'ensemble observé des points P_i a la valeur

$$\Pi = A^n e^{-\frac{1}{2}(H_1+H_2+\ldots+H_n)}\, dP_1\, dP_2 \ldots dP_n.$$

Les inconnues sont les $3n$ coordonnées des points M_i, qui sont liées par n relations, mais où figurent trois nouvelles inconnues. Nous prendrons le plan sous la forme

$$a_1x_1 + a_2x_2 + a_3x_3 = a_4.$$

Nous allons chercher à déterminer les $2n+3$ inconnues de manière à rendre minimum la somme $H_1+H_2+\ldots+H_n$, ce qui rendra maximum la grandeur considérée plus haut. Cette méthode suppose que la loi d'erreur est connue. On voit qu'elle n'est autre que la méthode générale de R. A. Fisher, dont les avantages de précision sont bien connus.

La solution du problème. — Fixons le plan et déterminons d'abord les points M_i pour qu'il y ait minimum. Il faut évidemment que H_i soit minimum. Or, la signification géométrique de cette condition est très claire. Considérons l'ellipsoïde de probabilité ayant son centre en P_i, son équation est de la forme

$$H[x_1-X_1,\ y_1-Y_1,\ z_1-Z_1]=\lambda^2.$$

Il faut que cet ellipsoïde soit tangent au plan considéré. Son équation peut s'écrire

$$a_1(x_1-X_1)+a_2(x_2-X_2)+a_3(x_3-X_3)+a_1X_1+a_2X_2+a_3X_3-a_4=0.$$

On a donc la condition

$$\lambda^2=\frac{(a_1X_1+a_2X_2+a_3X_3-a_4)^2}{K(a_1a_2a_3)}.$$

Le point M_i est évidemment le point de contact, sa position ne nous intéresse pas, mais seulement la valeur de λ^2. Finalement, en ajoutant les valeurs obtenues, on a

$$\Sigma H_i=\frac{\Sigma(a_1X_1'+a_2X_2'+a_3X_3'-a_4)^2}{K(a_1a_2a_3)}.$$

La position du plan s'obtiendra en cherchant les valeurs de a_1, a_2, a_3, a_4 qui rendent minimum cette somme.

La valeur de a_4 s'obtient immédiatement. On a

$$na_4=a_1\Sigma X_1+a_2\Sigma X_2+a_3\Sigma X_3.$$

Par conséquent, le plan passe par le centre de gravité des points P_i.

Après transport de l'origine en ce point, on est ramené à résoudre le même problème, a_4 étant nul. On a donc à chercher le minimum du rapport de deux formes quadratiques en $a_1a_2a_3$. Ce problème est classique : on sait qu'il est équivalent à la recherche des directions

conjuguées communes aux deux formes quadratiques. Si l'on pose

$$\theta = \frac{L(a_1 a_2 a_3)}{K(a_1 a_2 a_3)} = \frac{m^{ij} a_i a_j}{n^{ij} a_i a_j},$$

le minimum est donné par la plus petite racine de l'équation

$$| m^{ij} - \theta n^{ij} | = 0,$$

en représentant par cette notation abrégée le déterminant qui est le discriminant de la forme quadratique $L - \theta K$. La direction qui correspond à cette racine donne les coefficients $a_1 a_2 a_3$ du plan cherché.

Pour voir bien clairement les choses, on peut supposer faite la réduction aux mêmes carrés. le rapport prend la forme

$$\frac{L}{K} = \frac{m^{11} A_1^2 + m^{22} A_2^2 + m^{33} A_3^2}{n^{11} A_1^2 + n^{22} A_2^2 + n^{33} A_3^2} \qquad \left(\frac{m^{11}}{n^{11}} \geqq \frac{m^{22}}{n^{22}} \geqq \frac{m^{33}}{n^{33}} \right).$$

Les trois rapports sont justement les trois racines de l'équation en θ. On voit bien que

$$\frac{L}{K} - \frac{m^{33}}{n^{33}} = \frac{\left(m^{11} - n^{11} \frac{m^{33}}{n^{33}} \right) A_1^2 + \left(m^{22} - n^{22} \frac{m^{33}}{n^{33}} \right) A_2^2}{K} \geqq 0.$$

Le minimum est $\frac{m^{33}}{n^{33}}$. Il est atteint pour $A_1 = 0$, $A_2 = 0$. Ces deux équations déterminent la direction $a_1 a_2 a_3$.

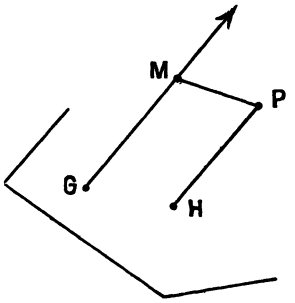
Le problème est donc entièrement résolu. Nous verrons comment on peut juger s'il l'est de manière acceptable.

Cas où les points sont situés sur une droite. — Il est clair que, posé géométriquement, le problème est le même. Il faut d'abord se donner une droite, déterminer sur cette droite les points M_i, en cherchant les ellipsoïdes tangents, puis chercher le minimum de l'expression ainsi obtenue. On peut résoudre le problème sans calcul, en supposant que par une transformation linéaire de variables. les ellipsoïdes soient devenus des sphères. Il est évident alors que les points M_i sont les projections, orthogonales au sens habituel, des points P_i sur la droite. D'autre part, la grandeur désignée par H_i devient le carré de la distance du point P_i à cette droite (à un facteur près). On est donc ramené à trouver une droite telle que la somme des carrés des distances de n points P_i soit un minimum. Pour une

direction donnée de cette droite, le minimum est obtenu quand la droite passe au centre de gravité des points P_i. On peut donc l'y faire passer et si l'on considère le plan perpendiculaire à la droite mené par G, il est clair que la somme des carrés des distances des points P à ce plan doit être un maximum.

Or, ce problème est celui que nous venons de traiter, mais à condition de prendre la plus grande racine de l'équation en θ.

Fig. 2.



On peut encore dire, puisque les deux autres plans principaux contiennent cette droite, qu'il suffit de considérer les deux plus petites racines de l'équation en θ. L'intersection des plans qui leur correspondent est la droite cherchée, et la valeur du minimum est la somme des deux plus petites racines. Sous cette forme, le résultat obtenu peut se transporter au cas tout à fait général de l'espace à n dimensions. (Nous donnons ici les résultats sans démonstrations.)

Résultat général. — Les racines de l'équation en θ seront supposées rangées par ordre de grandeur croissante,

$$k_1^2, \quad k_2^2, \quad \ldots, \quad k_n^2.$$

Dans ces conditions, si l'on prend un plan quelconque passant au centre de gravité, et qu'on cherche dans ce plan les N points qui fournissent le minimum de la somme $H_1 + H_2 + \ldots + H_N$. la plus petite valeur de ce minimum est k_1^2, et les coefficients de l'équation du plan sont les solutions d'un système de n équations homogènes du premier degré, compatibles en vertu de l'équation en θ.

Si l'on cherche une multiplicité à $n - 1$ dimensions, le minimum

est alors $k_1^2 + k_2^2$, et la multiplicité est l'intersection des deux plans déterminés par la méthode précédente et correspondant aux deux racines k_1^2 et k_2^2. On peut continuer ainsi; en particulier pour la multiplicité à une dimension, il faut prendre $k_1^2 + k_2^2 + \ldots + k_{n-1}^2$, et la droite commune aux $n - 1$ plans correspondant à $k_1^2 \ldots k_{n-1}^2$.

Enfin, si l'on cherchait un point, ce ne pourrait être que le centre de gravité et le minimum correspondant serait $k_1^2 + k_2^2 + \ldots + k_n^2$.

On voit que si la plus petite racine est très petite, la loi sera réductible. Mais on peut essayer de préciser les ordres de grandeur admissibles (*voir* pour ces questions [17], [19], [23]).

Comment juger si l'on a une solution qu'on puisse admettre. — En somme, nous avons supposé connue la loi d'erreur, et nous nous demandons si l'épaississement du nuage théorique, épaississement constaté par l'observation, est en accord raisonnable avec les erreurs que nous pouvons prévoir. C'est le problème qui se pose après chaque ajustement ou estimation.

Si l'on considère les mesures comme visant un groupe déterminé, mais inconnu, de points $M_1, M_2, \ldots, M_N$, le groupe des points P_1, $P_2, \ldots, P_N$ est un groupe aléatoire, dans la loi de probabilité duquel figurent comme paramètres les coordonnées des points visés, liées par des relations linéaires contenant les coordonnées de la multiplicité. Dans le cas de l'espace à 3 dimensions, il reste, tout compte fait, $2N + 3$ paramètres.

D'après les résultats généraux, utilisables quand n est très grand, la valeur probable du minimum, c'est-à-dire de la valeur R_1^2, est égale à

$$3N - (2N + 3) = N - 3.$$

L'écart type du minimum (aléatoire) a d'autre part la valeur

$$\sqrt{2(N - 3)}.$$

S'il s'agissait de n caracteres au lieu de 3, il faudrait prendre

$$N - n \quad \text{et} \quad \sqrt{2(N - n)}.$$

C'est ainsi qu'on jugera si la valeur du minimum n'est pas trop grande. Par exemple, supposons qu'on ait

$$N = 80, \qquad n = 8, \qquad N - n = 72, \qquad \sqrt{2(N - n)} = 12.$$

Si la valeur trouvée pour le minimum est 80, elle ne diffère que de moins d'un écart type, elle est par conséquent fort acceptable. Des règles analogues sont valables pour une multiplicité d'ordre inférieur à $n-1$. Supposons qu'au lieu d'une relation linéaire, il en existe un nombre k, on aura toujours un groupe aléatoire à $n\mathrm{N}$ coordonnées, mais le nombre total des paramètres est ici de

$$\mathrm{N}(n-k)+k(n-k+1).$$

D'où la valeur probable

$$h=k(\mathrm{N}-n+k-1).$$

L'écart type serait $\sqrt{2h}$.

Si le nombre des observations n'était pas très grand, il y aurait avantage à utiliser les tables classiques de Karl Pearson ou de R. A. Fisher pour la valeur de χ^2 (test of goodness of fit). Le minimum que nous avons rencontré dans les méthodes précédentes suit en effet cette loi avec un nombre de degrés de liberté (terminologie de R. A. Fisher) égal au nombre h (*voir* pour ces questions [6], [13]).

La méthode générale de l'ellipsoïde. — On remarquera que la quantité $\mathrm{L}(a_1, a_2, a_3)$ (qui figure au numérateur du rapport étudié précédemment) a pour coefficients les quantités

$$\sum_i (x_k^i - x_k^0)^2 = \mathrm{S}_{kk},$$

$$\sum_i (x_k^i - x_k^0)(x_h^i - x_h^0) = \mathrm{S}_{hk}.$$

Les coordonnées x_h^0 sont celles du centre de gravité des points P_i. Si l'on considère ces points P_i comme affectés de la même masse unité, on obtient les moments et produits d'inertie du nuage expérimental. C'est ce nuage qui, par l'effet des erreurs. est un véritable nuage alors que les points M_i sont dans un plan, ou sur une droite. Ce problème a été rencontré en économétrie, et longuement étudié par Ragnar Frisch, qui a donné à cet effet d'épaississement le nom de « Cushion effect » (effet coussin). Le problème est de dégonfler convenablement ce coussin expérimental.

Une idée toute naturelle est de chercher les axes de l'ellipsoïde

d'inertie, mais il manque quelque chose pour que cette idée ait un appui solide, car on n'aperçoit aucune raison pour appeler rectangulaires le système de coordonnées $x_1, x_2, \ldots, x_n$. Comme l'a fort bien remarqué H. Hotelling, il n'y a pas *a priori* de métrique dans une telle question, on peut, dit-il, en introduire une en admettant que les erreurs commises sur les différentes mesures sont indépendantes et d'égale importance. On voit immédiatement que cette hypothèse revient à introduire (si la loi d'erreur est gaussienne)

$$K(a_1 a_2 a_3) = \sigma^2(a_1^2 + a_2^2 + a_1^2),$$

c'est-à-dire à prendre comme ellipsoïdes des sphères.

Il semble que malheureusement cette hypothèse ne soit pas toujours justifiée. Si on l'adopte, on est conduit à chercher les axes de l'ellipsoïde d'inertie.

Nous pensons qu'il est préférable d'introduire franchement les ellipsoïdes d'erreur et de prendre les axes conjugués communs à cet ellipsoïde et à l'ellipsoïde d'inertie. La méthode, sous cette forme, n'est qu'une généralisation de la méthode des moindres carrés.

Si cette méthode est recommandable, c'est qu'en psychologie, malgré toutes les difficultés de cette tâche, on peut vraiment espérer isoler la loi d'erreur par mesures répétées.

Au contraire, en économie politique, si l'on veut résoudre des questions analogues, la méthode de l'ellipsoïde est assez arbitraire. Elle réussit d'ailleurs plutôt mal, comme l'a montré R. Frisch, qui a mis au point pour l'étude de ces questions une autre méthode d'analyse [18].

Grandeurs déterminantes. — Nous voyons d'après ce qui précède, qu'on peut arriver à estimer la position du nuage vrai, en cherchant combien de relations (linéaires) existent entre les différentes grandeurs étudiées. Supposons cette réduction faite, on n'a plus en réalité que $n - k$ caractères à étudier. Les autres en sont des fonctions linéaires. Nous dirons qu'il existe $n - k$ grandeurs déterminantes.

On est ramené à l'étude d'une loi de répartition dans un espace à $n - k$ dimensions, aucune réduction ultérieure n'étant plus possible. Il subsiste un certain arbitraire dans le choix de ces grandeurs déterminantes, qui pourraient *a priori* être des combinaisons linéaires assez générales de $n - k$ des variables primitives.

Bien entendu, de telles combinaisons linéaires, qui peuvent convenir à une expression mathématique, n'auraient souvent aucune signification directe. De sorte que les grandeurs qui doivent être retenues doivent l'être par une collaboration du psychologue et du mathématicien (*voir* [23]).

La liaison entre grandeurs déterminantes. — Les grandeurs qui restent sont réduites au nombre minimum, mais elles ne sont pas indépendantes. Est-il possible d'aller plus loin et de leur substituer des grandeurs indépendantes au sens des probabilités? En général, c'est impossible.

Une loi de probabilité doit avoir une structure spéciale pour qu'une telle transformation puisse être faite. Pourtant, à l'approximation généralement adoptée, qui ne dépasse pas les moments du deuxième ordre, on peut, et d'une infinité de manières, choisir des combinaisons linéaires des anciennes variables pour lesquelles les moments mixtes soient nuls.

Géométriquement, cela revient à choisir comme nouveaux plans de coordonnées des plans conjugués par rapport à l'ellipsoïde d'inertie, ce qui est possible d'une infinité de manières, dont aucune n'est privilégiée.

Mais il faut insister sur le fait que des combinaisons linéaires de grandeurs concrètes ne sont pas nécessairement des grandeurs concrètes, de sorte que cette transformation n'a pas beaucoup d'intérêt.

En somme, on en restera à la considération de grandeurs déterminantes, choisies pour leur importance, et il reste à étudier, par les méthodes précédentes, leur loi de répartition.

CHAPITRE V.

L'INTERPRÉTATION DES CORRÉLATIONS. LA THÉORIE DE SPEARMAN.

Nous abordons maintenant le dernier problème, celui d'une explication possible pour les lois de répartition obtenues par l'expérience. Dans le cas de la psychologie, le fait le plus important est le suivant : Les aptitudes étudiées sont toutes en corrélation, avec des coeffi-

cients de corrélation positifs, dont la valeur peut être assez élevée.

Ce résultat domine toute la question. C'est lui qu'il s'agit d'interpréter, d'expliquer dans une certaine mesure.

Les liaisons par variables communes. — Considérons pour plus de netteté deux variables seulement, x et y et supposons que a, b étant deux variables aléatoires indépendantes, on ait les relations

$$x = a \qquad \text{avec} \quad E(a) = E(b) = 0,$$
$$y = a + b \qquad \text{avec} \quad E(x) = E(y) = 0.$$

Les variables x et y ne sont pas indépendantes, puisque $E(xy) = E(a^2)$ n'est pas nul. On voit bien sur cet exemple que si l'on fixe la valeur de x, y n'est plus aléatoire que par la variable b, la valeur moyenne de y liée est donc égale à x, et la dispersion est constante. La régression de y en x est linéaire, l'écart type lié est constant, égal à σ_b.

On peut obtenir des types plus généraux sous la forme

$$x = \alpha' a + b,$$
$$y = \alpha'' a + c,$$

α', α'' sont des constantes, a, b, c trois variables indépendantes. Les variables x et y sont liées.

Bien entendu, de telles lois de probabilité sont assez particulières. On voit tout de suite qu'en appelant φ_1, φ_2, φ_3 les fonctions caractéristiques de a, b, c, la fonction caractéristique de la loi (xy) est

$$E\{e^{i(ux+vy)}\} = \varphi_1(\alpha' u + \alpha'' v)\,\varphi_2(\beta u)\,\varphi_3(\gamma v).$$

A cette condition, la loi xy sera analysable par trois variables indépendantes. dont une est commune, et les lois de probabilité des variables composantes sont entièrement données par les fonctions caractéristiques.

Application. La loi de Gauss. — Soit une loi de Gauss à deux variables dont la fonction caractéristique peut être écrite

$$\psi(uv) = -\frac{1}{2}[u^2 + 2ruv + v^2].$$

On a immédiatement

$$\frac{\partial^2 \psi}{\partial u\,\partial v} = \alpha'\alpha''\frac{\partial^2 \varphi_1}{\partial \lambda^2} = -r \qquad (\lambda = \alpha' u + \alpha'' v).$$

φ_1 est donc un polynome du deuxième degré et par conséquent, la variable a doit suivre une loi de Gauss. Puisque x et y suivent des lois de Gauss, les fonctions caractéristiques de b et c sont celles de lois de Gauss. On a donc tout simplement à résoudre une équation

$$u^2 + v^2 + 2ruv = A(\alpha u + \alpha'' v)^2 + B u^2 + C v^2.$$

Il existe une infinité de solutions, qu'on peut écrire

$$\begin{aligned} \sqrt{A}\,\alpha' &= \cos\varphi, & \sqrt{A}\,\alpha'' &= \cos\psi, \\ \sqrt{B} &= \sin\varphi, & \sqrt{C} &= \sin\psi, \\ & & \cos\varphi\cos\psi &= r. \end{aligned}$$

Ainsi, la loi de Gauss ou loi de corrélation normale à deux variables est représentable d'une infinité de manières, par des combinaisons linéaires de variables de Gauss, indépendantes.

Au contraire, dans le cas général à deux variables, le problème est impossible.

Loi à trois variables. — On peut généraliser de bien des manières les considérations précédentes. Contentons-nous de considérer le cas

$$\begin{aligned} x &= \alpha' a + b, \\ y &= \alpha'' a + c, \\ z &= \alpha''' a + d, \end{aligned}$$

a, b, c, d étant quatre variables indépendantes. Le problème est encore impossible dans le cas général. Que devient-il pour une loi de Gauss. On doit avoir

$$\psi(uvw) = \psi_1(\alpha' u + \beta'' v + \alpha''' v) + \psi_2(u) + \psi_3(v) + \psi_4(w).$$

On voit encore aisément que les variables composantes doivent suivre des lois de Gauss, et l'on est ramené à une équation

$$\begin{aligned} & u^2 + v^2 + w^2 + 2r_{12}uv + 2r_{23}vw + 2r_{31}wu \\ & \quad \equiv A(\alpha' u + \alpha'' v + \alpha''' v)^2 + B u^2 + C v^2 + D w^2. \end{aligned}$$

On a aisément la représentation

$$\begin{aligned} \sqrt{A}\,\alpha' &= \cos\varphi, & \sqrt{A}\,\alpha'' &= \cos\psi, & \sqrt{A}\,\alpha''' &= \cos\chi, \\ \sqrt{B} &= \sin\varphi, & \sqrt{C} &= \sin\psi, & \sqrt{D} &= \sin\chi, \\ & & \cos\varphi\cos\psi &= r_{12}, & & \\ & & \cos\psi\cos\chi &= r_{23}, & & \\ & & \cos\chi\cos\varphi &= r_{31}. & & \end{aligned}$$

Mais cette fois, le problème est déterminé

$$\cos^2\varphi\, r_{23} = r_{12} r_{31}.$$

Il existe une solution unique, à condition qu'elle soit réelle.

$$\cos\varphi = \varepsilon\sqrt{\frac{r_{12}r_{31}}{r_{23}}},$$

$$\cos\psi = \varepsilon\sqrt{\frac{r_{23}r_{12}}{r_{13}}},$$

$$\cos\chi = \varepsilon\sqrt{\frac{r_{31}r_{23}}{r_{12}}}.$$

Les deux signes ne donnent qu'une solution.

Notion générale du facteur commun. — On dira que n variables aléatoires sont représentables avec un facteur commun, ou facteur général, si l'on a

$$x_i = m_i g + s_i,$$

les $n+1$ variables aléatoires g, s_1, s_2, ... s_n étant indépendantes, les m_i étant des constantes. Il est clair que n variables x_i ainsi déterminées sont en liaison, mais cette liaison est très particulière. Pour qu'une loi de probabilité soit représentable de cette manière, il faut et il suffit que l'on ait l'identité, pour la fonction caractéristique

$$\psi(u_1 u_2 \ldots u_n) \equiv \psi_g(m_1 u_1 + \ldots + m_n u_n) + \psi_{s_1}(u_1) + \ldots + \psi_{s_n}(u_n).$$

La variable g est dite facteur général, les variables s_i sont dites les facteurs spécifiques, relatifs aux aptitudes mesurées par les x_i.

Détermination des facteurs. — Une question se pose. Quand l'identité précédente est vérifiée, les lois de probabilités des différents facteurs sont déterminées par le second membre. Nous dirons que ces variables aléatoires sont connues. Cela veut dire uniquement que leur loi de probabilité est entièrement déterminée. Cette détermination est-elle unique ? Nous allons voir qu'en général, il n'y a qu'une solution. En effet, supposons que les variables g et s_i sont rapportées à leur valeur probable. On aura, en prenant les termes du second degré de l'identité

$$(m_1 u_1 + \ldots + m_n u_n)^2 + \lambda_1^2 u_1^2 + \ldots + \lambda_n^2 u_n^2 = \text{forme connue}.$$

Il résulte des calculs précédents, faits pour les formes de Gauss, qu'à partir de $n = 3$, ces équations ont au plus une solution : $m_1 \ldots m_n$ sont donc connus. On connaît alors la dérivée seconde de Ψ_g et par conséquent ψ_g puisque les termes du premier ordre sont nuls. Il en résulte qu'on connaît aussi les dérivées secondes des Ψ_{si}, donc toutes les fonctions caractéristiques. Ainsi :

Quand une loi de probabilité à n variables $[n > 3]$ admet un facteur commun et n facteurs spécifiques, les $n + 1$ facteurs suivant des lois de probabilité bien déterminées.

Il n'existe donc qu'une manière (s'il en existe une) de reconstruire par le schéma précédent la loi de probabilité.

Autre sens de la détermination des facteurs. — Une légère confusion, très naturelle, pourrait s'introduire ici. Si la représentation d'un groupe d'aptitudes par facteur commun est possible, le psychologue qui étudie un individu voudrait connaître quels sont, pour cet individu, les valeurs de g, s_1, s_2, ..., s_n qui lui sont particulières. Or, ce problème est tout différent du premier. En effet, il s'agit, ayant ainsi mesuré sur cet individu les caractères x_1, x_2, ... x_n, d'en conclure quelque chose sur g, s_1, s_2, ... s_n. Il est évident que toute conclusion rigide est impossible, puisque les individus où l'on a fixé x_1, x_2, ..., x_n ne constituent qu'une sous-population où peuvent varier g, s_1, s_2, ..., s_n suivant une certaine loi de probabilité liée.

L'ensemble des $2n + 1$ variables aléatoires, x_1, ..., x_n, g, s_1, s_2, ..., s_n possède une certaine loi de répartition, loi réductible au sens du chapitre précédent, mais alors que la fixation des $n + 1$ grandeurs g et s_i fixe les n autres, la fixation des n premières laisse subsister une loi de répartition liée.

En particulier, il existe, dans cette loi liée, une espérance mathématique liée, un écart type lié pour chacune des $n + 1$ grandeurs. Plus particulièrement, la grandeur g spécialement intéressante, possède des moments liés dont les deux premiers sont les plus importants.

La théorie de Spearman. — La théorie proposée par Spearman est précisément celle du facteur commun. On l'appelle aussi quelquefois la théorie des deux facteurs, parce qu'elle propose de décomposer

chaque aptitude en deux facteurs, l'un général et l'autre spécifique. La première question posée par cette théorie est celle-ci : Comment la vérifier ?

Nous avons vu que la connaissance de la fonction caractéristique le permettrait. Mais on ne connaît pas cette fonction, on ne connaît que les premiers moments. L'attention s'est donc portée spécialement sur l'identité, restreinte aux moments du deuxième ordre

$$\mathrm{E}[u_1 x_1 + \ldots + u_n x_n]^2 = (m_1 u_1 + \ldots + m_n u_n)^2 + \lambda_1^2 u_1^2 + \ldots + \lambda_n^2 u_n^2.$$

Il est à peine besoin de dire que cette identité est beaucoup moins étroite que la première.

Elle suppose seulement les conditions

$$\mathrm{E}(g^2) = 1, \qquad \mathrm{E}(s_i^2) = \lambda_i^2, \qquad \mathrm{E}(g s_i) = 0, \qquad \mathrm{E}(s_i s_k) = 0,$$

c'est-à-dire des conditions imposées aux moments du deuxième ordre des $n + 1$ facteurs. Ainsi, les $n + 1$ variables g, s_i peuvent avoir une loi de probabilité quelconque, mais à moments nuls pour le premier ordre, à coefficients de corrélation nuls pour le second. Si alors l'identité restreinte est vérifiée, les variables

$$m_i g + s_i$$

ont les mêmes moments du premier et du second ordre que les variables $x_1, x_2, \ldots, x_n$.

On voit qu'au sens indiqué précédemment, les facteurs sont tout d'abord largement indéterminés, et qu'ensuite ils ne reproduisent qu'au second ordre la loi $x_1, x_2, \ldots, x_n$.

En particulier, la loi de probabilité $(g, x_1, x_2, \ldots, x_n)$ ne reproduit qu'au second ordre la loi $(x_1, x_2, \ldots, x_n)$. Peut-elle la reproduire complètement? Certainement oui, et d'une infinité de manières. En particulier, choisissons une loi liée de g, qui soit de Gauss, du type

$$\frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{1}{2\sigma^2}[g-h]^2}\,dg$$

h peut être pris égal à la combinaison linéaire qui fournit le plan de régression, σ étant l'écart type lié donné par la formule classique. La loi de $x_1, x_2, \ldots, x_n$ est alors prise identique à la loi véritable. On peut évidemment choisir une loi de régression différente, et un écart

type qui dépende de $x_1, x_2, \ldots, x_n$. Les conditions imposées à la loi $g x_2, \ldots, x_n$ se traduiraient par le fait que l'hyperplan des moindres carrés est fixé et que la valeur moyenne du carré de la distance des points de la distribution à cet hyperplan a une valeur donnée.

En somme, si incomplète que puisse paraître une solution basée sur les moments du deuxième ordre, et bien qu'elle laisse subsister une large incertitude sur la loi même de probabilité du facteur g, elle est pourtant très intéressante parce qu'elle ne permet pas à la moyenne liée de s'écarter beaucoup d'un certain plan, et que la variable aléatoire g, pour $x_1, x_2, \ldots, x_n$ donnés ne peut fluctuer qu'entre des limites fixées.

Des considérations analogues sont valables pour le groupe des variables liées $g, s_1, s_2, \ldots, s_n$. Elles comportent quelques longueurs, mais aucune difficulté.

Les conditions d'application. — Bornons-nous donc à la vérification de l'identité au sens restreint

$$(10)\quad \mathrm{E}(u_1 x_1 + \ldots + u_n x_n)^2 = (m_1 u_1 + \ldots + m_n u_n)^2 + \lambda_1^2 u_1^2 + \ldots + \lambda_n^2 u_n^2.$$

Nous supposons que les variables $x_1, x_2, \ldots, x_n$ sont ramenées à avoir l'écart type unité. On aura donc

$$1 = m_i^2 + \lambda_i^2, \qquad r_{ik} = m_i m_k.$$

On voit qu'à partir de $n = 4$, et en outre des conditions de réalité des m_i et λ_i, nous aurons des conditions de compatibilité; on peut les écrire

$$r_{ik} r_{hl} = r_{ih} r_{kl} = m_i m_k m_h m_l.$$

Les différences qui doivent être nulles

$$r_{ik} r_{hl} - r_{ih} r_{kl}$$

s'appellent des tétrades. Si elles sont toutes nulles, on aura

$$\frac{r_{ik}}{r_{ih}} = \frac{r_{kl}}{r_{hl}} = \frac{\rho_k}{\rho_h} = (\text{rapport qui ne dépend que de } h \text{ et } k)$$

$$(r_{ik} = \rho_k \lambda_i)$$

avec

$$\lambda_i = p \rho_i, \qquad r_{ik} = p \rho_i \rho_k,$$

ce qui donne, en modifiant les ρ, $\rho_i \rho_k$ ou $-\rho_i \rho_k$, suivant le signe de p.

Si c'est la premiere forme, on n'a plus que la condition de réalité, qui est que tous les ρ_i soient inférieurs ou égaux à l'unité.

On aura évidemment

$$E(x_i g) = m_i.$$

Si c'est la deuxième forme $-\rho_i\rho_k$, la réduction au type précédent est impossible.

Sur l'estimation du facteur général. — Si l'on a mesuré sur un individu les quantités $x_1, x_2, \ldots, x_n$, la valeur de g est, comme nous l'avons dit, une variable aléatoire, g liée. Elle a même une loi de probabilité insuffisamment connue si l'on se borne à utiliser les moments du deuxième ordre. Mais, même en admettant qu'on la connaisse entierement, g ne serait pas en liaison fonctionnelle avec $x_1, x_2, \ldots, x_n$, mais seulement en liaison de probabilité. Autrement dit, il n'y a aucune raison pour que l'écart type lié de g soit nul. Mais si la théorie reste valable pour un grand nombre d'aptitudes, on peut espérer voir l'estimation de g devenir meilleure quand n grandit. Pour voir ceci bien nettement, exécutons les calculs qui donnent l'équation de régression et la quantité R^2. On a, appelant $\beta_1, \beta_2, \ldots, \beta_n$ les coefficients de régression

$$\begin{aligned} M_1 &= \beta_1 + \beta_1 m_1 m_2 + \ldots + \beta_1 m_1 m_n, \\ &\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots, \\ R^2 &= \beta_1 m_1 + \ldots\ldots\ldots + \beta_n m_n. \end{aligned}$$

La quantité R^2 inconnue auxiliaire, introduite dans les équations, donne

$$\frac{m_i}{1 - m_i^2} = \beta_i + \frac{m_i}{1 - m_i^2} R^2.$$

D'où sa valeur

$$R = \frac{S}{1 + S},$$

en posant

$$S = \sum_{i=1}^{n} \frac{m_i^2}{1 - m_i^2}.$$

Par conséquent, la quantité $1 - R^2$ qui nous intéresse a pour valeur

$$1 - R^2 = \frac{1}{1 + S}.$$

On voit bien qu'elle diminue à chaque adjonction d'un nouveau terme à la somme S.

Si l'on pouvait penser que n devienne très grand, et que la somme S devienne en même temps très grande (ce qui n'est pas une conséquence). la quantité $1 - R^2$ tendrait vers zéro, et l'on aurait une estimation qui convergerait en probabilité vers la vraie valeur, c'est-à-dire qui aurait une probabilité aussi grande qu'on veut de l'approcher autant qu'on veut.

Mais il nous semble que la théorie de Spearman ne saurait, en admettant son exactitude, être forcée jusque-là. Elle peut bien décomposer le mécanisme mental en $n+1$ grandeurs indépendantes capables de reconstituer l'essentiel des différentes aptitudes, mais supposer que n est très grand revient à donner à ce mécanisme mental une complication infinie (*voir* [20, 21, 22, 24, 25]).

Effet des substitutions linéaires. — E. B. Wilson, dans de remarquables contributions apportées à la théorie de Spearman, a signalé un point curieux. Si l'on suppose le mécanisme du facteur général applicable à n aptitudes, il ne le sera généralement pas a des combinaisons linéaires des nombres $x_1, x_2, \ldots, x_n$. Il y a lieu de distinguer ici ce qu'on entend par la conservation du facteur commun dans une substitution linéaire. Il peut arriver :

1° qu'apres la transformation les nouvelles variables résultent de $n+1$ facteurs indépendants. dont l'un, commun, est le même que pour les variables primitives. C'est évidemment le sens que le psychologue serait porté à donner, car un individu donné, dans la théorie proposée, a une valeur donnée de g, et c'est elle qui devrait intervenir dans les autres aptitudes;

2° la propriété de facteur commun subsiste, mais il s'agit d'un autre facteur. Ce point de vue serait plutôt de mathématicien.

Il est clair qu'au sens 1° il n'y a pas conservation en général, car il faudrait que dans les nouvelles variables

$$y_k = a_{ik} x_i = a_{ik} m_i g + a_{ik} s_i,$$

les nouveaux facteurs spécifiques qui sont $a_{ik} s_i$. soient indépendants. Or, il est bien clair qu'il n'en est rien en général. Ces nouveaux facteurs spécifiques ne sont même pas en non-corrélation, il faudrait

pour cela des conditions

$$\mathrm{E}(a_{ik}s_i)(a_{jh}s_h) = 0,$$

c'est-à-dire

$$a_{i1}a_{j1}\lambda_1^2 + \ldots + a_{in}a_{jn}\lambda_n^2 = 0.$$

Ce sont des conditions analogues à des conditions d'orthogonalité. Ainsi : même au sens restreint des conditions du deuxième ordre, une transformation linéaire quelconque ne conserve pas le facteur commun.

Y aurait-il là une difficulté grave pour la théorie de Spearman? Il nous semble que non. S'il y a un mécanisme des aptitudes, il n'a aucune raison de s'appliquer à des combinaisons linéaires qui n'ont pas de sens concret.

La théorie de Spearman peut parfaitement être acceptable pour une dizaine, une vingtaine d'aptitudes, connues ou à découvrir. Pourquoi le serait-elle à une somme de deux aptitudes, pourquoi le serait-elle à 100 aptitudes? (*voir* [10. 16, 24, 28]).

Les théories à plusieurs facteurs communs. — Il est clair qu'on peut généraliser ces théories, et introduire un nombre quelconque de facteurs communs. Par exemple

$$x_i = m_i g^1 + p_i g^2 + s_i,$$

g^1, g^2 étant deux facteurs généraux. La base de la théorie sera encore une identité entre fonctions caractéristiques, et au sens restreint

$$\mathrm{E}(u_1x_1 + \ldots + u_nx_n)^2 = (m_1u_1 + \ldots + m_nu_n)^2 + (p_1u_1 + \ldots + p_nu_n)^2 + \lambda_1^2u_1^2 + \ldots + \lambda_n^2u_n^2.$$

Les conditions, un peu plus compliquées, sont analogues. Il suffit de former les dérivées partielles du premier membre

$$u_1 + u_2r_{12} + \ldots + u_nr_{1n} = m_1[m_1u_1 + \ldots + m_nu_n] + p_1(p_1u_1 + \ldots + p_nu_n) + \lambda_1u_1,$$
$$u_1r_{12} + u_2 \ldots\ldots\ldots = m_2[\ldots\ldots\ldots\ldots] + \ldots\ldots\ldots\ldots + \lambda_2u_2,$$
$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$

Il est clair que si l'on prend les termes en u_4, u_5, ... u_n, ils sont des combinaisons linéaires des deux expressions

$$m_nu_n + \ldots + m_nu_n, \qquad p_nu_n + \ldots + p_nu_n.$$

Par conséquent, les déterminants du troisième ordre

$$\begin{vmatrix} r_{14} & r_{15} & r_{16} \\ r_{24} & r_{25} & r_{26} \\ r_{34} & r_{35} & r_{36} \end{vmatrix}$$

sont nuls. Ce sont eux qui remplacent les tétrades.

On voit que les calculs à exécuter dans ces hypothèses se compliquent assez fortement (*voir* [4], Crossroads in the mind of man).

Les fluctuations aléatoires. — Pour juger si la théorie de Spearman permet de rendre compte des observations, il faut estimer l'ensemble des paramètres $m_1, m_2, \ldots, m_n$, et voir si la loi de probabilité ainsi obtenue est en accord suffisant avec les mesures. En réalité, comme nous l'avons vu, la donnée de $m_1 m_2 \ldots m_n$ et la connaissance, qui en résulte, de $\lambda_1 \lambda_2 \ldots \lambda_n$, ne suffisent pas a faire connaître cette loi.

On se borne alors à examiner si les observations permettent l'estimation du groupe des m_i. La condition théorique est que toutes les tétrades soient nulles, il ne reste ensuite qu'une condition de realité.

Or, les tétrades qui résultent de l'expérience sont affectées d'erreurs aléatoires. On se contentera donc de voir si les tétrades expérimentales peuvent raisonnablement être considérées comme ayant une valeur théorique nulle. Cet examen pose un problème assez lourd, du point de vue des calculs. Il faut en effet connaître la loi de probabilité d'une tétrade autour de la valeur zéro. Si l'on pouvait admettre que la tétrade expérimentale suit une loi de Gauss, il suffirait de connaître l'écart type, et c'est en effet à quoi l'on se borne en pratique.

Il faut bien remarquer cependant que le coefficient de corrélation suit une loi assez différente de celle de Gauss, quand les observations ne sont pas très nombreuses. La tétrade des quatre coefficients suivra bien, à la limite, une loi de Gauss, mais pour elle aussi, il est un peu risqué de lui appliquer la loi de Gauss avec un nombre d'observations qui ne soit pas très grand.

Spearman et Holzinger ont donné la partie principale de l'écart type d'une tétrade sous la forme

$$\sigma_t^2 = \frac{A}{N},$$

N est le nombre des observations, A est une fonction des coefficients de corrélation vrais.

On pourrait opérer autrement. En effet, nous avons vu que

$$E(x_i x_k) = m_i m_k.$$

Ces quantités, qu'on appelle les covariances C_{ik}, permettent donc de former des tétrades nulles, mais qui, cette fois, sont des fonctions entières des observations. La loi de probabilité de ces fonctions entières est beaucoup plus facile à obtenir rigoureusement, et a été étudiée par Wishart; il est d'ailleurs raisonnable de penser que ces grandeurs suivent une loi qui se rapproche plus rapidement de la loi de Gauss que les tétrades des coefficients de corrélation. Il semble donc qu'il y avait quelque avantage à utiliser plutôt les tétrades C.

Des difficultés, tenant à la longueur des calculs numériques, se présentent d'ailleurs dès que le nombre des tests est un peu élevé.

Dans les expériences de Brown et Stephenson portant sur 20 tests, il s'introduit 14535 tétrades ($3 \times C_{20}^4$).

En résumé, et bien que la méthode actuelle soit pratiquement suffisante pour porter un jugement sur la théorie, on peut désirer des perfectionnements d'ordre mathématique et technique, qui permettraient d'enlever quelque lourdeur aux calculs numériques qui sont actuellement nécessaires (1).

(1) D'intéressantes recherches sont faites en ce moment au Laboratoire de la S. N. C. F., à Viroflay, par M. Pierre Delaporte. Elles permettent de traiter beaucoup plus aisément ces problèmes.

INDEX BIBLIOGRAPHIQUE.

Ouvrages généraux.

1. Spearman (C.). — *The abilities of man* (Londres, 2e édit., 1932. Traduction française de F. Brachet aux Éditions du Travail humain, Paris).
2. G. Udny Yule. — *An introduction to the theory of statistics* (Londres, nombreuses éditions).
3. R. A. Fisher. — *Statistical Methods for Research Workers* (Londres et Edinburgh, nombreuses éditions).
4. Truman L. Kelley. — *Statistical Method* (New-York).
   Truman L. Kelley. — *Crossroads in the mind of man* (Stanford University Press, 1928).
5. G. Darmois. — *Statistique mathématique* (Paris, 1928).

Mémoires et Communications.

6. K. Pearson. — On the criterion that a given system..... (*Phil. Mag.*, série V, 1900, p. 157-175).
7. Spearman (C.). — *Am. Journ. Psych.*, vol. XV, 1904, p. 202.
8. Garnett (J. C. M.). — *Proc. Roy. Soc.*, A, 1919, p. 96.
9. Garnett (J. C. M.). — *Brit. Journ. Psych.*, vol. X, 1920, p. 242-58.
10. Garnett (J. C. M.). — *Nature*, t. 132, n° 3339, octobre 1933, p. 676.
11. Fisher (R. A.). — On the mathematical foundations of theoretical Statistics (*Ph. Trans.*, A, t. 212, 1921, p. 309-368).
12. Fisher (R. A.). — On the probable error of a coefficient of correlation deduced from a small sample (*Metron*, t. 1, Part. IV, 1921, p. 1-32).
13. Fisher (R. A.). — The conditions under which χ^2 measures the discrepancy between observation and hypothesis (*Journ. Roy. Stat. Soc.*, t. 87, 1924, p. 442-449).
14. Fisher (R. A.). — Theory of statistical estimation (*Proc. of the Cambridge Phil. Soc.*, t. 22, 1925, p. 700-725).
15. Pearson (K.) et Marg. Moul. — *Biometrika*, vol. XIX, 1927, p. 246-291.
16. Wilson (E. B.). — *Proc. Nat. Ac. Sci.*, vol. XIV, 1928, p. 283-296.
17. Frisch (Ragnar). — Correlation and scatter in statistical variables (*Nordik Statistical Journal*, t. 1, 1928, p. 36).
18. Frisch (Ragnar). — *Statistical Confluence analysis* (Oslo, 1934).
19. St Georgesco (Nicolas). — Le problème de la recherche des composantes cycliques d'un phénomène (*Journ. Soc. Stat. Paris*, octobre 1930).

20. Heywood (H. B.). — On finite sequences of real numbers (*Proc. Roy. Soc.*, A, vol. 134, 1931, p. 486-501).

21. Piaggio (H. T. H.). — *Mathematical Gazette*, vol. XVII, n° 232, 1933, p. 40-42.

22. Piaggio (H. T. H.). — *Brit. Journ. of Psych.*, vol. XXIV, 1933, p. 88-105.

23. Hotelling (H.). — *Analysis of a complex of statistical Variates into principal components* (Columbia University, Baltimore, 1933).

24. Irwin (J. O.). — *Statistical Methods in Psychology. The present position of the theory of two factors* (XXII[e] Session de l'Institut International de Statistique, La Haye, 1934).

25. Irwin (J. O.). — On the indeterminacy in the estimate of *g* (*Brit. Journ. of Psych.*, vol. XXV, Part. III, janvier 1935).

26. Darmois (G.). — Sur la théorie des deux facteurs de Spearman (*C. R. Acad. Sc. Paris*, t. 199, 1934, p. 1176 et 1358).

27. Darmois (G.). — *L'emploi des observations statistiques. Méthodes d'estimation* (Actualités scientifiques et industrielles, Paris, 1936).

28. Darmois (G.). — Sur l'indétermination de *g* dans la théorie de Spearman (*Mathematica*, 1936.

29. Neyman (J.) et E. S. Pearson. — *Statistical Research Memoirs*, vol. I (Londres, juin 1936).

30. Holzinger (Karl J.). — *Preliminary Reports on Spearman. Holzinger unitary trait Study*, n° 5, *Introduction to bifactor theory* (The University of Chicago Press, 1935). *Student Manual of factor analysis* (The Univerty of Chicago Press, 1937).

31. Thurstone (L. L.) — *The Victors of mind.* University of Chicago Press, 1936. *Primary mental abilities.* University of Chicago Press, 1938.

32. Thomson (Godfrey H. T.). — *The Factorial analysis of human ability.* University of London Press, 1939.

TABLE DES MATIÈRES.