LE COEFFICIENT D'APPROXIMATION MOYENNE ET LE COEFFICIENT DE CORRÉLATION

par
R. KAŠANIN

Le but du présent travail est, d'abord, de donner une base pour une approximation moyenne et de former le critérium d'efficacité d'une telle approximation (Chap. I—III); puis d'en faire une application à la corrélation et d'étendre la définition du coefficient de corrélation (Chap IV); et, enfin, d'indiquer brièvement, de ce point de vue, un cas spécial d'approximation moyenne par les polynômes (Chap. V).

I.

Soient a_i $(i=1,2,\ldots,n)$ n nombres réels; il peut y en avoir des égaux parmi ceux-ci, mais ils ne sont pas tous égaux. Prenons un nombre quelconque a (réel, puisque on opère dans le domaine de nombres réels) et formons

$$\alpha_i = a_i - a.$$

Posons

$$\sigma_k = \sqrt[k]{\frac{1}{n}\sum_{i=1}^{n} \alpha_i^k}, \qquad (k=1,2,3,\ldots);$$

si k est un nombre pair, nous prendrons la valeur positive de la racine. Soit α le plus grand parmi les nombres $|\alpha_i|$. On a alors

$$\sum_{i=1}^{n} \alpha_i^{2k} = \alpha^{2k} \sum_{i=1}^{n} \left|\frac{\alpha_i}{\alpha}\right|^{2k}, \qquad \left|\frac{\alpha_i}{\alpha}\right| \leqq 1.$$

Par conséquent la somme figurant en facteur à côté de α^{2k} n'est pas inférieure à 1 ni supérieure à n, de sorte que

$$\lim_{k\to\infty}\sqrt[2k]{\sum_{i=1}^{n}\left|\frac{\alpha_i}{\alpha}\right|^{2k}}=1.$$

Mais comme

$$\lim_{k\to\infty}\sqrt[2k]{\frac{1}{n}}=1,$$

il vient

$$\lim_{k\to\infty}\sigma_{2k}=\alpha,$$

c'est-à-dire σ_{2k} *tend vers le plus grand des nombres* $|\alpha_i|$ *lorsque* k *croît indéfiniment. Par conséquent, si petit que soit* $\varepsilon>0$, *il existe un nombre entier positif* N *tel que*

$$-\sigma_{2k}(1+\varepsilon)<\alpha_i<+\sigma_{2k}(1+\varepsilon)$$

pour tout $k>N$ *et pour tous* i.

Nous désignerons le quotient de α_i par α_{2k} – par K_i, donc

$$\frac{\alpha_i}{\sigma_{2k}}=K_i$$

On a alors

$$\sum_{i=1}^{n}K_i^{2k}=\frac{1}{\sigma_{2k}^{2k}}\sum_{i=1}^{n}\alpha_i^{2k}=\frac{1}{\sigma_{2k}^{2k}}\cdot n\sigma_{2k}^{2k}=n.$$

Considérons un nombre arbitraire $K>0$ et supposons qu'il y ait v des $|K_i|$ supérieurs à K. On aura alors certainement

$$\sum_{i=1}^{n}K_i^{2k}>vK^{2k},\ \text{c'est-à-dire}\ v<\frac{n}{K^{2k}}.$$

Donc, *en dehors de l'intervalle* $(-K\sigma_{2k},+K\sigma_{2k})$ *il y en a des* α_i *moins de* $\frac{n}{K^{2k}}$; c'est-à-dire *dans l'intervalle fermé* $(-K\sigma_{2k},+K\sigma_{2k})$ *il y en a des* α_i *plus de* $n-\frac{n}{K^{2k}}$.

En prenant d'abord $K=2^{\frac{1}{2k}}$, puis $K=n^{\frac{1}{2k}}$, on trouvera que:

Dans l'intervalle fermé $(-2^{\frac{1}{2k}}\sigma_{2k}, +2^{\frac{1}{2k}}\sigma_{2k})$ *il y a plus de la moitié de tous les* α_l, *et en dehors de l'intervalle* $(-n^{\frac{1}{2k}}\sigma_{2k}, +n^{\frac{1}{2k}}\sigma_{2k})$ aucun. En particulier, pour $k=1$, il s'ensuit que; *Dans l'intervalle fermé* $(-\sigma\sqrt{2}, +\sigma\sqrt{2})$ *il y a plus de la moitié des* α_l et en dehors de l'intervalle $(-\sigma_1\sqrt{n}, +\sigma_2\sqrt{n})$ aucun.

Lorsque $0<p<1$ et si l'on pose

$$K=(1-p)^{-\frac{1}{2k}},$$

on pourra dire: *Plus que* $p\cdot n$ *des* α_l *se trouvent dans l'intervalle fermé*

$$\left(-\frac{\sigma_{2k}}{(1-p)^{\frac{1}{2k}}}, +\frac{\sigma^{2k}}{(1-p)^{\frac{1}{2k}}}\right).$$

Pour $k=1$, on en conclut: *Plus que* $p\cdot n$ *des* α_l *se trouvent dans l'intervalle fermé*

$$\left(-\frac{\sigma_2}{\sqrt{1-p}}, +\frac{\sigma_2}{\sqrt{1-p}}\right),$$

Pour $p=\frac{1}{2}$ et $p=1-\frac{1}{n}$, on trouverait les résultats spéciaux déjà donnés.

Tout ce que l'on vient de dire sur l'intervalle fermé $(-K\sigma_{2k}, +K\sigma_{2k})$ fut déduit de la supposition qu'il y a ν de quantités $|K_l|>K$. Si l'on admet qu'il y ν de quantités $|K_l|\geqq K$, on aura: *à l'intérieur de l'intervalle* $(-K\sigma_{2k}, +K\sigma_{2k})$ *il y a au moins* $n-\frac{n}{K^{2k}}$ de α_l. Par ex., *à l'intérieur de l'intervalle* $(-\sigma\sqrt{2}, +\sigma\sqrt{2})$ *il y a au moins la moitié des* α_l.

Comme cas extrême, il peut arriver qu'à l'intérieur de l'intervalle $(-\sigma\sqrt{2}, +\sigma\sqrt{2})$ il y ait juste la moitié de tous les α_l, ou tous les α_l. Le premier cas se produira, par ex., si

$$n=4;\ a_1=-1,\ a_2=0,\ a_3=0,\ a_4=+1;\ a=0;$$

où $\sigma_2\sqrt{2}=1$; le second, si

$$n=4;\ a_1=-1,\ a_2=-1,\ a_3=+1,\ a_4=+1;\ a=0;$$

où $\sigma_2\sqrt{2}=\sqrt{2}$.

Pour $K=1$, on trouve que dans l'intervalle fermé $(-\sigma_{2k}, +\sigma_{2k})$ il y des α_l en nombre >0, c'est-à-dire qu'il y en a certainement, et à l'intérieur de cet intervalle qu'il y en a en nombre $\geqq 0$, c'est-à-dire qu'il n'est pas certain qu'il y en ait. En effet, il peut arriver qu'à l'intérieur de cet intervalle il n'y en ait pas. C'est le cas du second des deux exemples cités ci-dessus, car alors $\sigma_{2k}=1$.

De ces considérations il ressort qu'au moyen de σ_2 on peut déjà préciser l'intervalle dans lequel se trouve la majorité des α_l, de même que celui en dehors duquel il n'en existe point. Cette évaluati on est certainement moins précise qu celle au moyen de σ_{2k} pour $k>1$, mais plus simple. En revenant aux nombres a_l on peut dire que: *De quelle manière que soient donnés les n nombres* $a_1, a_2, \ldots a_n$ *et un nombre arbitraire* a, *plus de la moitié des* a_l *se trouvent dans l'intervalle fermé*

$$(a-\sigma_2\sqrt{2},\ a+\sigma_2\sqrt{2}), \qquad \left[\sigma_2=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(a_l-a)^2}\right],$$

au moins la moitié à l'intérieur de cet intervalle et point en dehors de l'intervalle

$$(-\sigma_2\sqrt{n},\ a+\sigma_2\sqrt{n}).$$

Plus généralement: *De quelle manière que soient donnés les n nombres* $a_1, a_2, \ldots a_n$ *et un nombre arbitraire* a, *dans l'intervalle fermé*

$$\left(a-\frac{\sigma_2}{\sqrt{1-p}},\ a+\frac{\sigma_2}{\sqrt{1-p}}\right), \qquad \left[\sigma_2=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(a_l-a)^2}\right]$$

il y aura plus de pn des a_l, *et à l'intérieur de cet intervalle pas moins de pn.*

La longueur de cet intervalle dans lequel on a trouvé plus de *pn* de nombres donnés est

$$2\sigma_2:\sqrt{1-p},$$

pour $p=\frac{1}{2}$, elle est $2\sigma_2\sqrt{2}$. Cette longueur dépend donc de σ_2; et σ_2, les a_l étant donnés, dépend du choix du nombre a. L'évaluation sera d'autant plus précise que σ_2 est plus petit. Aussi

choisirons-nous le nombre a de manière que σ_2 soit minimum, c'est-à-dire que

$$\sigma_2^2 = \frac{1}{n}\sum_{i=1}^{n}(a_i - a)^2$$

soit minimum. En égalant à zéro la dérivée de σ_2^2 par rapport à a on aura

$$2 \cdot \frac{1}{n}\sum_{i=1}^{n}(a_i - a) = 0,$$

c'est-à-dire

$$a = \frac{1}{n}\sum_{i=1}^{n} a_i.$$

Donc, *l'intervalle fermé*

$$\left(a - \frac{\sigma_2}{\sqrt{1-p}},\ a + \frac{\sigma_2}{\sqrt{1-p}}\right),$$

dans lequel nous avons trouvé plus de pn de nombres donnés a_i sera le plus court si on prend pour a la moyenne arithmétique de ces nombres.

C'est ainsi que nous procéderons dans ce qui suivra, c'est-à-dire nous prendrons pour a la moyenne arithmétique des nombres donnés a_i. Dans ce cas, σ_2 est appelé *la dispersion.* Il serait plus précis de l'appeler: la dispersion autour de la moyenne arithmétique des nombres donnés.

Si l'évaluation se faisait non pas au moyen de σ_2 mais de σ_{2k} pour $k>1$, on aurait à chercher, pour avoir a, le minimum de

$$\frac{1}{n}\sum_{i=1}^{n}(a - a_i)^{2k},$$

c'est-à-dire on aurait à résoudre l'équation, du degré $2k-1>1$,

$$\frac{1}{n}\sum_{i=1}^{n}(a - a_i)^{2k-1} = 0.$$

Le calcul serait donc beaucoup plus long et plus compliqué. Il n'y a que pour $n=2$ que l'on aurait

$$\frac{1}{2}(a-a_1)^{2k-1}+\frac{1}{2}(a-a_2)^{2k-1}=0,$$

c'est-à-dire encore la moyenne arithmétique

$$a=\frac{1}{2}(a_1+a_2),$$

quelque soit k.

II.

Soient $b_1, b_2, \ldots, b_n$ *n nombres donnés;* parmi ceux-ci il peut y en avoir d'égaux, mais ils ne sont pas tous égaux. Soit b leur moyenne arithmétique, donc

$$\frac{1}{n}\sum_{i-1}^{n} b_i = b.$$

Si l'on pose $b_i = b+\beta_i$, on aura

$$\frac{1}{n}\sum_{i-1}^{n} \beta_i = 0.$$

La dispersion relative aux b_i est

$$\tau_2 = +\sqrt{\frac{1}{n}\sum_{i=1}^{n}\beta_i^2}.$$

Substituons à chacun des nombres donnés b_i un autre, c_i, qu'on obtiendrait par un procédé quelconque, par ex., par un calcul; on aura ainsi n nombres calculés c_i. Soit c leur moyenne arithmétique,

$$\frac{1}{n}\sum_{i=1}^{n} c_i = c.$$

En posant $c_i = c+\gamma_i$, on aura

$$\frac{1}{n}\sum_{i=1}^{n} \gamma_i = 0.$$

La dispersion relative aux nombres calculés est

$$\tau_2' = +\sqrt{\frac{1}{n}\sum_{i=1}^{n}\gamma_i^2}.$$

Désignons par ρ le rapport des dispersions relatives aux nombres calculés et donnés

$$(1) \qquad \rho = \frac{\tau_2'}{\tau_2} = +\sqrt{\frac{\frac{1}{n}\sum_{i=1}^{n}\gamma_i^2}{\frac{1}{n}\sum_{i=1}^{n}\beta_i^2}}.$$

En prenant c_i à la place de b_i, on introduit l'écart D_i,

$$D_i = b_i - c_i = (\beta_i - \gamma_i) + (b - c).$$

Cet écart est une somme de deux nombres: de $D = b - c$, le même pour tout i, et $\Delta_i = \beta_i - \gamma_i$, celui-ci étant, en général, différent pour i différant. Nous appellerons le premier *écart systématique* (constant), le second *écart fortuit* (variable). On voit facilement que l'on a

$$\frac{1}{n}\sum_{i=1}^{n} D_i = b - c, \qquad \frac{1}{n}\sum_{i=1}^{n}\Delta_i = 0.$$

Par conséquent, la disperison des écarts est

$$\sigma_2' = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\Delta_i^2} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\beta_i - \gamma_i)^2}$$

$$= \sqrt{\left[\frac{1}{n}\sum_{i=1}^{n}\beta_i^2 - \frac{2}{n}\sum_{i=1}^{n}\beta_i\,\gamma_i + \frac{1}{n}\sum_{i=1}^{n}\gamma_i^2\right]}.$$

Pour abréger l'écriture, posons

$$\frac{1}{n}\sum_{i=1}^{n}\beta_i\,\gamma_i = [\beta, \gamma].$$

Avec cette notation on a

$$\sigma_2' = \sqrt{(\tau^2 - 2[\beta, \gamma] + \tau_2'^2)},$$

c'est-à-dire

$$\sigma_2' = \tau_2 \sqrt{1 + \rho^2 - \frac{2}{\tau_2^2}[\beta, \gamma]}. \tag{2}$$

Le cas très fréquent où l'on prend c_i, au lieu de b_i, est le suivant: b_i étant les valeurs d'une fonction inconnue ou compliquée, $f(x)$, pour $x = a_i$, les a_i étant des nombres différents entre eux, nous substituons à cette fonction une autre, $F(x)$, connue ou plus simple, en posant $c_i = F(a_i)$. Géométriquement parlant, les points (a_i, b_i) étant situés sur une courbe $y = f(x)$, nous substituons à celle-ci une courbe $y = F(x)$; aux points $M_i(a_i, b_i)$ on substitue, ainsi, les points $N_i(a_i, c_i)$, où $c_i = F(a_i)$.

Pour avoir de cette façon les c_i, il faut calculer les valeurs de la fonction $F(x)$, pour $x = a_1, a_2, \ldots, a_n$. Les opérations élémentaires que l'on effectue facilement, sans aucun appareil auxiliaire de calcul, sont l'addition et la multiplication (la soustraction et la division se ramènent à ces dernières). Par conséquent, au point de vue de calcul, le plus simple serait si l'on choisissait d'avance m fonctions déterminées, linéairement indépendantes $\psi_1(x), \ldots, \psi_m(x)$, dont on peut facilement calculer ou tirer des tables les valeurs pour $x = a_i$, puis posons

$$F(x) = C_0 + C_1\psi_1(x) + \cdots + C_m\psi_m(x), \tag{3}$$

où $C_0, C_1, \ldots, C_m$ sont des constantes. On aura alors

$$c_i = C_0 + C_1\psi_1(a_i) + \cdots = C_m\psi_m(a_i). \tag{4}$$

Nous appellerons l'ensemble de fonctions $\psi_1(x), \psi_2(x), \ldots, \psi_m(x)$ *la base.*

Quelles seront les valeurs de c_i, calculées d'après (4), et des écarts D_i et Δ_i, cela dépendra, une fois la base choisie, des valeurs que l'on attribuera aux paramètres $C_0, C_1, \ldots, C_m$. Si $m+1 \geqq n$, on pourra toujours les choisir de manière que $c_i = b_i$, c'est-à-dire $D_i = 0$, car le système d'équations

$$C_0 + C_1\psi_i(a_i) + \cdots + C_m\psi_m(a_i) = b_i$$
$$(i = 1, 2, \ldots, n)$$

aura alors des solutions en $C_0, C_1, \ldots, C_m$. Mais, si $m < n-1$, ce ne sera pas faisable pour tout i, sauf dans des cas exceptionnels, de sorte qu'il y aura certainement des écarts différents de zéro, quelque soit la manière dont on a choisi les $C_0, C_1, \ldots, C_m$.

De (4), en faisant la somme de $i=1$ à $i=n$ et divisant par n, on a pour moyenne arithmétique der valeurs calculées:

$$(5) \qquad c = C_0 + C_1 \frac{1}{n} \sum_{i=1}^{n} \psi_1(a_i) + \cdots + C_m \frac{1}{n} \sum_{i=1}^{n} \psi_m(a_i).$$

L'écart systématique (constant) est $b-c$.

De même qu'on avait substitué, aux p_i, les nombres c_i, calculés d'après (4), on aura maintenant, au lieu des β_i, les nombres $\gamma_i = c_i - c$:

$$\gamma_i = C_1 \left[\psi_1(a_i) - \frac{1}{n} \sum_{j=1}^{n} \psi_1(a_j) \right] + \cdots + C_m \left[\psi_m(a_i) - \frac{1}{n} \sum_{j=1}^{n} \psi_m(a_j) \right].$$

Si l'on désigne par a la moyenne arithmétique des a_i et l'on pose

$$x = a + \xi, \quad a_i = a + \alpha_i,$$

$$(6) \qquad \psi_k(x) - \frac{1}{n} \sum_{j=1}^{n} \psi_k(a_j) = \varphi_k(\xi),$$

on aura

$$(7) \qquad \gamma_i = C_1 \varphi_1(\alpha_i) + \cdots + C_m \varphi_m(\alpha_i).$$

Par conséquent, la fonction

$$(8) \qquad \Phi(\xi) = C_1 \varphi_1(\xi) + \cdots + C_m \varphi_m(\xi)$$

sert à calculer les valeurs $\gamma_i = \Phi(x_i)$.

De (6) l'on voit immédiatement que les fonctions φ_k possèdent la propriété

$$(9) \qquad \frac{1}{n} \sum_{i=1}^{n} \varphi_k(\alpha_i) = 0.$$

En choisissant arbitrairement non pas les fonctions $\psi_k(x)$ mais les fonctions $\varphi_k(\xi)$ (sous la condition (9), bien entendu),

on aura, au moyen de (7), les nombres γ_i. Les c_i s'obtiennent alors de $c_i = c + \gamma_i$, c étant arbitraire. On aura ainsi

$$c_i = c + C_1 \varphi_1(a_i - a) + \cdots + C_m \varphi_m(a_i - a),$$

c'est-à-dire, pour calculer les c_i on se servira maintenant de la fonction

$$F(x) \equiv c + C_1 \varphi_1(x-a) + \cdots + C_m \varphi_m(x-a), \tag{10}$$

où $c, C_1, C_2, \ldots, C_m$ sont de constantes que l'on peut choisir arbitrairement. En prenant $c = k$, on évitera l'écart systématique.

III.

Admettons que la fonction $F(x)$, qui sert à calculer les c_i, soit donnée sous forme (10). Si ce n'est pas le cas, et qu'elle soit donnée sous forme (3), nous supposerons qu'à l'aide de (6) on l'ait mise sous forme (10).

Comme il vient d'être dit, les fonctions $\varphi_1(\xi), \varphi_2(\xi), \ldots, \varphi_m(\xi)$ étant choisies (à la condition (9)), on peut attribuer aux constantes $c, C_1, C_2, \ldots, C_m$ des valeurs arbitraires. Déterminons les de manières que soient satisfaites les conditions suivantes:

1) $c = b$, car on évite ainsi les écarts systématiques;

2) que la dispersion σ_2', relative aux écarts fortuits $\Delta_i = \beta i - \gamma_i$, soit minimum, car plus la dispersion est petite plus l'intervalle avec la majorité des écarts sera petit (I).

Ces deux conditions sont indépendantes entre elles. Pour la première on aura simplement à porter, dans (10), b au lieu de c. La seconde déterminera les $C_1, C_2, \ldots, C_m$ au moyen de (7).

On a

$$\Delta_i = \beta_i - \gamma_i = \beta_i - \Phi(\alpha_i)$$
$$= \beta_i - [C_1 \varphi_1(\alpha_i) + \cdots + C_m \varphi_m(\alpha_i)],$$

de sorte que

$$\sigma_2'^2 = \frac{1}{n} \sum_{i=1}^{n} \Delta_i^2 = \frac{1}{n} \sum_{i=1}^{n} [\beta_i - \Phi(\alpha_i)]^2.$$

Si l'on veut que $\sigma_2'^2$ soit minimum, on déterminera les paramètres $C_1, \ldots, C_m$ des équations

$$\frac{\partial(\sigma_2'^2)}{\partial C_k} = 0, \qquad (k = 1, 2, \ldots, m).$$

Posons, pour abréger l'écriture,

$$\frac{1}{n}\sum_{i=1}^{n}\varphi_j(\alpha_i)\,\varphi_k(\alpha_i)=[\varphi_j\,\varphi_k],$$

$$\frac{1}{n}\sum_{i=1}^{n}\beta_i\,\varphi_k(\alpha_i)=[\beta\,\varphi_k];$$

on voit immédiatement que l'on a

$$[\varphi_j\,\varphi_k]=[\varphi_k\,\varphi_j],\quad [\beta\,\varphi_k]=[\varphi_k\,\beta].$$

Avec ces notations on aura les équations en $C_1, C_2, \ldots, C_m$:

$$(11)\qquad \begin{cases} C_1[\varphi_1\varphi_1]+C_2[\varphi_1\varphi_2]+\cdots+C_m[\varphi_1\varphi_m]=[\beta\,\varphi_1] \\ C_1[\varphi_2\varphi_1]+C_2[\varphi_2\varphi_2]+\cdots+C_m[\varphi_2\varphi_m]=[\beta\,\varphi_2] \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ C_1[\varphi_m\varphi_1]+C_2[\varphi_m\varphi_2]+\cdots+C_m[\varphi_m\varphi_m]=[\beta\,\varphi_m]. \end{cases}$$

Lorsqu'on calcule les c_i sous la forme (4), c'est-à-dire à l'aide de la fonction $F(x)$, donnée par (3), et, en même temps, on détermine les paramètres $C_0, C_1, \ldots, C_m$ de manière que la moyenne arithmétique des valeurs calculées soit égale à celle des valeurs données et que la dispersion soit minimum, on dit qu'on a procédé à une *approximation moyenne de la fonction* $f(x)$ *au moyen de la base* $\psi_1(x), \ldots, \psi_m(x)$. Se servir de (10), au lieu de (3), n'est qu'une question de technique de calcul.

Pour l'écart D_i on a

$$D_i=\Delta_i+(b-c),$$

d'où

$$D_i^2=\Delta_i^2+2(b-c)\,\Delta_i+(b-c)^2,$$

et par conséquent

$$\sum_{i=1}^{n}D_i^2=\sum_{i=1}^{n}\Delta_i^2+2(b-c)\sum_{i=1}^{n}\Delta_i+n(b-c)^2=n\,\sigma_2'^2+n(b-c)^2.$$

Dans une approximation moyenne il sera $b=c$ et $\sigma_2'^2$ minimum

et cela indépendamment l'un de l'autre. Par conséquent aussi $\sum_{i=1}^{n} D_i^2$ sera minimum et inversement. Donc, *la caractéristique de l'approximation moyenne est que la somme des carrés des écarts est minimum.*

De (7) on a

$$\gamma_i^2 = C_1^2 \varphi_1^2(\alpha_i) + \cdots + C_m^2 \varphi_m^2(\alpha_i) + \sum_{j=1}^{n} \sum_{k=1}^{n} C_j C_k \varphi_j(\alpha_i) \varphi_k(\alpha_i), \; j \neq k,$$

de sorte que pour la dispersion τ_2' relative aux valeurs calculées on trouve

$$\tau_2'^2 = C_1^2 [\varphi_1 \varphi_1] + \cdots + C_m^2 [\varphi_m \varphi_m] + \sum_{j} \sum_{k} C_j C_k [\varphi_j \varphi_k], \qquad j \neq k. \tag{12}$$

Ils existent, cependant, dans l'approximation moyenne aussi les équations (11). Si, après les avoir multipliées par $C_1, C_2, \ldots, C_m$, on les additionne, on trouvera facilement que le second membre de l'équation (12) est égal à

$$C_0 [\beta \varphi_1] + C_2 [\beta \varphi_2] + \cdots + C_m [\beta \varphi_m],$$

c'est-à-dire à $[\beta\gamma]$. Ainsi donc, *dans l'approximation moyenne la dispersion τ_2' relative aux valeurs calculées est déterminée par*

$$\tau_2'^2 = C_1 [\beta \varphi_1] + C_2 [\beta \varphi_2] + \cdots + C_m [\beta \varphi_m] = [\beta\gamma] \tag{13}$$

En portant, de (13), dans (2)

$$[\beta\gamma] = \tau_2'^2 = \tau^2 \rho^2,$$

on trouve que: *dans l'approximation moyenne la dispersion σ_2' des écarts est donnée par*

$$\sigma_2'^2 = \tau_2 \sqrt{1 - \rho^2}, \tag{14}$$

où $\rho = \tau_2' : \tau_2$, c'est-à-dire le rapport des dispersions relatives aux valeurs calculées et données.

En vertu de ce qui a été dit au Chapitre I, on conclut: *Dans l'approximation moyenne, plus de la moitié des écarts*

ne sont pas, en valeur absolue, supérieurs à $\sqrt{2}\tau_2\sqrt{1-\rho^2}$ *et aucun n'est, en valeur absolue, supérieur à* $\sqrt{n}\tau_2\sqrt{1-\rho^2}$. On peut dire aussi: *Dans une approximation moyenne, les valeurs absolues des écarts sont inférieurs à* $\sqrt{n}\tau_2\sqrt{1-\rho^2}$ *et pour au moins la moitié du nombre total des écarts inférieurs à* $\sqrt{2}\tau_2\sqrt{1-\rho^2}$. Plus généralement: *Dans une approximation moyenne au moins pn des écarts sont inférieurs, en valeur absolue,* à

$$\frac{\tau_2}{\sqrt{1-p}}\sqrt{1-\rho^2}.$$

De (14) on voit que

$$0 \leqq \rho \leqq 1. \tag{15}$$

Si $\rho = 0$, on aura $\tau_2' = 0$, et il sera $\sigma_2' = \tau_2$ et $\gamma_2 = 0$, c'est-à-dire $c_i = b$, de sorte que tous les b_i sont remplacés par un même nombre — par leur moyenne arithmétique. Si $\rho = 1$, il sera $\sigma_2' = 0$, et l'on aura $\Delta_i = 0$ et $\gamma_i = \beta_i$, pour tout i, c'est-à-dire $c_i = b_i$, de sorte que tous les points $M_i(a_i, b_i)$ sont situés sur la courbe $y = F(x)$. Par conséquent, plus ρ est grand, meilleure est l'approximation moyenne. Aussi, dans une approximation moyenne, appellerons-nous ρ *le coefficient d'approximation.* Il se trouve, d'après (15), dans l'intervalle fermé (0,1).

Bien que les calculs qui fournissent $C_1, \ldots, C_m, \tau_2', \rho$ et σ_2' soient élémentaires, ils peuvent être, surtout la résolution de (14), fastidieux et longs. Le plus simple serait si la base avait la propriété

$$[\varphi_j \varphi_k] = \begin{cases} 0 \text{ pour } j \neq k, \\ 1 \text{ pour } j = k. \end{cases} \tag{16}$$

On a alors

$$\begin{cases} C_k = [\beta \varphi_k], \\ \tau_2'^2 = \tau^2 \rho^2 = \sum_{k=1}^{m} C_k^2, \\ \sigma_2'^2 = 1 - \frac{1}{\tau^2} \sum_{k=1}^{n} C_k^2. \end{cases} \tag{17}$$

C'est à cette simplification de calcul que tient son importance la base caractérisée par (16). Mais elle a, en outre, un autre grand avantage. Supposons, en effet, que l'on modifie la base par l'addition d'un terme supplémentaire $\varphi_{m+1}(\xi)$; ceci entraînera l'introduction d'un nouveau paramètre C_{m+1} et, dans le cas général, tous les calculs seront à recommencer, car le système (11) modifié donne, en général, pour les paramètres C_k des valeurs différentes des anciennes. Or, si la base est caractérisée par (16), tous les paramètres C_k, une fois calculés, gardent leurs valeurs; le seul à calculer est le paramètre $C_{m+1} = [\beta\varphi_{m+1}]$ puis, à l'aide de celui-ci, les valeurs corrigées des coefficients ρ et des dispersions σ_2' et τ_2'.

Si les fonctions $\varphi_k(\xi)$ ne satisfont pas à la condition (16), on pourra en former, par des combinaisons linéaires, d'autres, remplissant cette condition. Ce sont les fonctions $\chi_n(\xi)$, définies par la formule recurrente

$$(18)\quad \begin{cases} \chi_1(\xi) = \dfrac{\varphi_1(\xi)}{\sqrt{[\varphi_1\varphi_1]}}, \\ \chi_k(\xi) = \dfrac{1}{\sqrt{[\varphi_k\varphi_k] - \sum_{\nu=1}^{k-1}[\chi_\nu\varphi_k]}} \cdot \left\{ \varphi_k(\xi) - \sum_{\nu=1}^{k-1}[\chi_\nu\varphi_k]\chi_\nu(\xi) \right\}. \end{cases}$$

En effet, on peut se rendre compte par le calcul qu'elles satisfont aux conditions (9) et (16).

Les conditions (9) et (16), étant satisfaites pour les fonctions $\varphi_1(\xi), \ldots, \varphi_m(\xi)$, peuvent être réunies si la base est, de plus, complétée par la fonction $\varphi_0(\xi) = 1$, qui figure à côté de C_0. (9) est alors compris dans (16), à savoir pour $j = 0$.

IV.

Considérons deux suites de nombres a_i et b_i, $i = 1, 2, \ldots, n$

$$(19)\quad \begin{cases} a_1, a_2, \ldots, a_n \\ b_1, b_2, \ldots, b_n \end{cases}.$$

Dans chaque série il peut y avoir de nombres égaux entre eux,

mais nous supposerons qu'ils ne sont pas tous égaux. Nous désignerons les moyennes arithmétiques de ces séries par a et b,

$$a = \frac{1}{n}\sum_{i=1}^{n} a_i, \qquad b = \frac{1}{n}\sum_{i=1}^{n} b_i.$$

Par conséquent, si $a_i = a + \alpha_i$ et $b_i = b + \beta_i$, on aura

$$\frac{1}{n}\sum_{i=1}^{n} \alpha_i = 0, \qquad \frac{1}{n}\sum_{i=1}^{n} \beta_i = 0.$$

Posons en général, pour k positif et entier,

$$\sigma_k = \sqrt[k]{\frac{1}{n}\sum_{i=1}^{n} \alpha_i^k}, \qquad \tau_k = \sqrt[k]{\frac{1}{n}\sum_{i=1}^{n} \beta_i^k};$$

si k est un nombre pair, nous adopterons la valeur positive de la racine. Les grandeurs σ_2 et τ_2 sont donc des dispersions des nombres a_i et b_i. D'après ce qui a été dit au Chapitre I, à l'aide de la dispersion nous pouvons évaluer l'étendue des intervalles autour de a et b où sont situés les nombres a_i et b_i.

Si parmi les nombres a_i il n'y en a pas d'égaux entre eux, la table (19) établit la liaison fonctionnelle $y = f(x)$, d'une part, entre la variable indépendante x dont le domaine est l'ensemble des a_i, et, d'autre part, la fonction y, dont le domaine est constitué par les nombres b_i, le terme fonction (uniforme) étant pris dans son sens le plus général. Géométriquement parlant, on aura, dans le plan Oxy, des points $M_i(a_i, b_i)$. On peut, de même, envisager le problème de l'approximation, en particulier de l'approximation moyenne. On aura alors la fonction $F(x)$ et une nouvelle série de n nombres $c_i = F(a_i)$. On en a, d'ailleurs, parlé au Chapitre III.

Cependant, si parmi les a_i il en existe d'égaux entre eux la correspondance (19) entre les deux séries ne peut plus être qualifiée de fonctionelle. $x_1, x_2, \ldots, x_v$ étant des valeurs différentes de a_i, au nombre x_j correspond tout un groupe de b_i. On dit alors que les séries (19) sont *en corrélation.* Les α_i et β_i aussi sont alors en corrélation.

En cas de corrélation nous choisirons pour base $\psi_1(x)$, $\psi_2(x), \ldots, \psi_m(x)$ et nous formerons la fonction

$$F(x) = C_0 + C_1\psi_1(x) + \cdots + C_m\psi_m(x).$$

$E(a_i)$, pour $i = 1, 2, \ldots, n$ donnera alors une troisième série de nombres

$$C_1, C_2, \ldots, C_n.$$

Si $a_j = a_k$, on aura $c_j = c_k$. C'est en cela que consiste la différence essentielle entre les séries b_i et c_i. De cette façon, au lieu d'une corrélation entre les a_i et b_i, nous avons établi une liaison fonctionelle entre les a_i et c_i. Entre les b_i et c_i on trouvera les écarts $D_i = b_i - c_i$, composés des écarts systématiques (constants) et des fortuits (variables): $D_i = (b - c) + \Delta_i$. La grandeur de ces écarts et leur dispersion σ_2', pour une base choisie, dépend de la manière dont on a déterminé les coefficients C_0, $C_1, \ldots, C_m$. Nous les prendrons comme au Chapitre III, c'est-à-dire de manière que l'on ait $c = b$ et que σ_2' soit minimum, ou que $\sum_{i=1}^{n} D_i^2$ soit minimum. On retombera ainsi encore sur les équations (11) et les formules (13), (14) et (15). Dans ce cas nous appellerons ρ – *le coefficient de corrélation.*

Lorsque les séries (19) sont données, le coefficient de corrélation dépend du choix de la base $\psi_1(x), \ldots, \psi_m(x)$. Si $\rho = 1$, on aura $b_i = c_i$, en d'autres termes, la corrélation se transforme en liaison fonctionnelle (3). Si la base a la propriété (16), on arrivera aussi pour la corrélation aux formules (17).

V.

Lorsqu'il est question de l'approximation moyenne, que ce soit d'une liaison fonctionnelle ou d'une corrélation, on peut prendre pour base $\psi_k = x^k$ (l'approximation moyenne par les polynômes algébriques). A l'aide de la formule récurrente (18) on pourra former les polynômes $\varphi_k(\xi)$ ayant les propriétés (9) et (16). Ces polynômes des deux premiers degrés sont

$$L_1 = \frac{\xi}{\sigma_2}, \qquad L_2 = \frac{\left(\frac{\xi}{\sigma_2}\right)^2 - \left(\frac{\sigma_3}{\sigma_2}\right)^3 \cdot \frac{\xi}{\sigma_2} - 1}{\sqrt{\left(\frac{\sigma_4}{\sigma_2}\right)^4 - \left(\frac{\sigma_3}{\sigma_2}\right)^6 - 1}}.$$

Ce genre de polynômes, ainsi que l'approximation par ces derniers ont été l'objet d'une étude plus détaillée, dans ma note „Sur les divers procédés d'interpolation", parue dans les „Publications mathématiques de l'Université de Belgrade", T. VI et VII, 1937-38, p. 240-266. Pour les approximations linéaire et quadratique les calculs y ont été effectués jusqu'au bout; avec tous les détails y a surtout été traité le coefficient d'approximation (resp. de corrélation) dans le cas d'une approximation linéaire.