JOURNAL DE LA SOCIÉTÉ STATISTIQUE DE PARIS

PIERRE THIONET

L'école moderne de statisticiens italiens

Journal de la société statistique de Paris, tome 87 (1946), p. 16-34

<http://www.numdam.org/item?id=JSFS_1946__87__16_0>

IV

L'ÉCOLE MODERNE DE STATISTICIENS ITALIENS [1]

(*Suite.*)

ANNEXE II — SÉRIE CYCLIQUE : MOYENNE ET MÉDIANE

Soit $M_1 M_2 \ldots M_i \ldots$ des points fixes (modalités) sur un cercle, affectés des coefficients (fréquences) $f_1 f_2 \ldots . f_i \ldots .$

On peut considérer le vecteur résultant (O étant le centre du cercle) :

$$f_1 \overrightarrow{O M_1} + f_2 \overrightarrow{O M_2} + \ldots . + f_i \overrightarrow{O M_i} + \ldots . = \overrightarrow{O G} . \Sigma f_i .$$

et appeler *point moyen* celui où la demi droite $\overline{OG}$ rencontre le cercle. C'est ce qu'on fera par exemple pour une statistique des directions des vents (provenant d'un office météorologique).

On ne procède pas ainsi pour une série cyclique quelconque. On choisit sur le cercle une unité d'arc, par exemple l'écart entre deux modalités si celles ci sont également espacées; Gini envisage surtout 4 (trimestres), 7 (jours) ou 12 (mois) modalités, également espacées. Soit x_i l'abscisse curviligne de M_i (pour une origine, une unité et un sens de parcours donnés); soit M un point d'abscisse x décrivant le cercle dans le sens positif.

A) Soit la fonction :

$$\mu_1 (x) = \mu_1 (M) = \Sigma f_i \widehat{MM_i}$$

où Σ désigne une somme algébrique et où l'arc algébrique $\widehat{MM_i}$ est *en valeur absolue* le plus petit possible. On voit assez facilement :

— que $\mu_1 (M)$ est une fonction linéaire de x, de pente $-\Sigma f_i$ (fonction décroissante);

— que, chaque fois que M traverse un point M', opposé sur le cercle à une modalité M_i, la fonction $\mu_1 (M)$ subit une discontinuité et s'accroît de $f_i . l$.

Au total, quand M est revenu à son point de départ, $\mu_1 (M)$ s'est accru de $(\Sigma f_i)\, l$ par sauts et de $(-\Sigma f_i)\, l$ par variation continue, c'est à dire a repris sa valeur primitive.

La courbe représentative de cette fonction périodique aura l'allure ci dessous (on a supposé qu'il y avait 7 modalités également espacées).

Par définition, on a $\mu_1 (M) = 0$ *pour les moyennes.*

On voit qu'il y a autant de moyennes que d'intervalles de continuité (c'est à dire de modalités) mais que les moyennes peuvent tomber en dehors de leur intervalle (voir M' et S) : c'est ce qu'on appelle des moyennes *fictives ;* entre la moyenne *ordinaire* et la moyenne *fictive*, il existe la moyenne *limite* (Voir D).

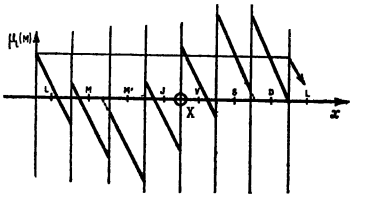


Fig. 1.

En outre, le point X est ce qu'on appelle une *moyenne sui-generis :* on a en effet :

$$\frac{\mu_1 (x - \varepsilon) + \mu_1 (x + \varepsilon)}{2} = \mu_1 (x) = 0.$$

B) Soit de même la fonction

$$\mu_2 (x) = \mu_2 (M) = \Sigma f_i \widehat{MM_i}.$$

On voit assez facilement :

— que contrairement à $\mu_1 (M)$, $\mu_2 (M)$, n'admet pas de discontinuité, mais seulement des changements dans son expression analytique (et des discontinuités dans sa dérivée) quand M traverse les points M'_i;

— que, dans chaque intervalle (M'_i, M'_{i+1}), $\mu_2 (M)$ est un trinôme du second degré en x de la forme $x^2 \Sigma f_i + \ldots$

— que l'on a :

$$\frac{d \mu_2}{d x} = -2 \mu_1 (x).$$

---

(1) Voir *Journal de la Société de Statistique de Paris*, n° de novembre-décembre 1945.

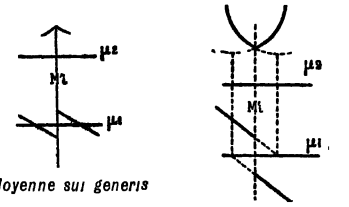
Il en résulte que la courbe représentative de $\mu_2(x)$ se compose d'arcs de paraboles égales tournant leur concavité vers les μ_2 positifs et dont les sommets correspondent à :

$$\mu_1(x) = 0.$$

A chaque *moyenne, ordinaire, limite ou fictive*, correspond un minimum pour μ_2 (M).

Par contre la moyenne *sui generis* correspond à un maximum de μ_2 (M), les deux arcs de parabole étant symétriques.

Aucun minimum de μ_2 ne saurait provenir d'un autre point que d'un sommet de parabole, c'est à-dire que *tout point anguleux* saillant est impossible, car cela supposerait que μ_1 subit un saut négatif en traversant une discontinuité; or cela ne peut être puisque tous les f sont positifs.



Moyenne sui generis

Fig. 2.

En résumé : En dehors des discontinuités X pour lesquelles on a $\mu_1(x-\varepsilon) = -\mu_1(X+\varepsilon)$, les points moyens, définis soit à l'aide de $\mu_1(M) = 0$, soit tels que $\mu_2(M)$ passe par un minimum, coïncident; et il y en a autant que de modalités (de fréquence non nulle).

Médiane d'une série ordinaire.

On sait que la médiane rend minimum la somme :

$$S(x) = \sum_1^s f_i \,|\, x - x_i \,|$$

La fonction S (x) est représentée par une ligne brisée; lorsque

$$x_{k-1} < x < x_k$$

le segment représentatif a pour pente

$$\sum_k^s f_i - \sum_1^{k-1} f_i.$$

Lorsque cette pente peut être nulle $\left(\sum_r^s f_i = \sum_1^{r-1} f_i\right)$ la médiane est indéterminée, car la ligne brisée présente un palier minimum. Jackson introduit :

$$S_p(x) = \sum_1^s f_i \,|\, x - x_i \,|^p$$

c'est-à-dire

$$S_p(x) = \sum_1^{r-1} f_i (x - x_i)^p \quad + \sum_r^s f_i (x_i - x)^p$$

qui est minimum si $\frac{1}{p}\frac{dS_p}{dx}(x) = \sum_1^{r-1} f_i (x - x_i)^{p-1} - \sum_r^s f_i (x_i - x)^{p-1} =$

D'autre part on a : $\sum_1^{r-1} f_i \cdot 1 \qquad - \sum_r^s f_i \cdot 1 =$

En posant $h = p - 1$ et en faisant apparaître les quantités $\frac{(x-x_i)^h - 1}{h}$, qui tendent vers L $(x - x_i)$ quand h tend vers 0, il vient, à la limite :

$$g(x) = \sum_1^{r-1} f_i \,\mathrm{L}(x - x_i) - \sum_r^s f_i \,\mathrm{L}(x_i - x) = 0$$

ou $$\boxed{f(x) = (x-x_1)^{f_1}(x-x_2)^{f_2} \ldots\ldots (x-x_{r-1})^{f_{r-1}} - (x-x_r)^{f_r} \ldots\ldots (x-x_s)^{f_s}}$$

(équation de Jackson.)

Cette équation admet une racine dans l'intervalle (x_{r-1}, x_r) car :

$$f(x_{r-1}) < 0 \qquad f(x_r) > 0$$

Cette racine est unique; en effet :

$$g'(x) = \sum_1^{r-1} \frac{f_i}{x - x_i} + \sum_r^s \frac{f_i}{x_i - x}$$

d'où $g'(x) > 0$ pour $x_{r-1} < x < x_r$

de sorte que $g(x)$ est croissante dans cet intervalle, ainsi que $f(x)$. *cqfd.*

Cisbani a proposé une méthode de résolution de cette équation par approximations successives en partant de $x_0 = \frac{x_{r-1} + x_r}{2}$

et en posant $z = x - x_0$, $z_i = x_i - x_0$, de sorte que l'équation de Jackson devienne :

$$\sum_1^{r-1} f_i \operatorname{L} z_i - \sum_r^s f_i \operatorname{L} z_i + \sum_1^{r-1} f_i \operatorname{L}\left(1 - \frac{z}{z_i}\right) - \sum_r^s f_i \left(1 - \frac{z}{z_i}\right) = 0$$

ou, avec un développement limité au 2e ordre :

$$\sum_1^{r-1} f_i \operatorname{L} z_i - \sum_r^s f_i \operatorname{L} z_i - z\left[\sum_1^{r-1} \frac{f_i}{z_i} - \sum_r^s \frac{f_i}{z_i}\right] - \frac{z^2}{2}\left[\sum_1^{r-1} \frac{f_i}{z_i^2} - \sum_r^s \frac{f_i}{z_i^2}\right] = 0$$

équation du 2e degré en z dont une racine est nulle mais dont l'autre sera $z_0 \neq 0$.

En posant alors $x'_0 = x_0 + z_0$, $z' = x - x'_0$, et en réitérant, il pense qu'on arrivera rapidement à une valeur de x très voisine de la racine de l'équation de Jackson. Il ne démontre d'ailleurs pas la convergence de la suite $x_0, x'_0, x''_0, \ldots$ vers cette racine.

Médiane d'une série cyclique.

On définira le point médian comme un point M d'abscisse x, tel que $\mu = \sum_1^s f_i \mid \widehat{\mathrm{M}\,\mathrm{M}_i} \mid$ soit minimum.

Il est clair que μ change d'expression analytique chaque fois que M traverse une modalité M_i ou un point M'_i opposé à M_i sur le cercle. Dans chacun des intervalles ainsi définis, la courbe représentative de $\mu(x)$ est un segment de droite; donc pour le cercle entier, la courbe est une ligne brisée. Une médiane correspond à un point M_i ou M'_i, où la pente du segment passe du négatif au positif (minimum de μ). Un intervalle médian se rencontre chaque fois que ladite pente passe du négatif à zéro puis au positif.

A) Pour simplifier, supposons les modalités régulièrement espacées et en nombre pair $2k$; la courbe représentative de $\mu(x)$ se compose de $2k$ segments seulement, dont chacun a une équation de la forme :

$$\mu(x) = \sum_k f_h (x - x_h) + \sum_k f_j (x_j - x)$$

en désignant par M_h les k modalités rencontrées avant M et par M_j les k modalités rencontrées après M, quand on parcourt le cercle. La pente du segment est donc :

$$\sum_k f_h - \sum_k f_j.$$

Il existe un intervalle médian lorsqu'on peut avoir : $\sum_k f_h = \sum_k f_j$.

Par exemple, pour la série cyclique :

Printemps	Été	Automne	Hiver
492	313	745	566

on aura :

$\sum f_i = 2.116$

$\sum f_h$	$\sum f_j$	Pente
805	1 311	—
1 058	1 058	0
1.309	805	+
1 058	1 058	0

et l'intervalle entre le point *Automne* et le point *Hiver* est médian.

B) Lorsque les s modalités (régulièrement espacées) sont en nombre impair, le plus simple

est de doubler le nombre de modalités en partageant en deux parties égales chaque intervalle et de donner aux nouvelles modalités M'$_i$ des fréquences nulles La courbe représentative de μ (x) comprend donc (2 s) segments de droite formant une ligne brisée *continue*.

Lorsque le point M, qui décrit le cercle, traverse un point M$_i$, la pente du segment augmente de

$$f_i - 0.$$

Lorsque le même point traverse un point M'$_i$, la pente dû segment augmente de :

$$0 - f_i$$

Si aucune des modalités M i n'a de fréquence nulle, il ne peut y avoir d'intervalle médian; la pente de la courbe doit en effet être négative dans un intervalle, être nulle dans le suivant, être positive dans le troisième. M va traverser successivement les points M$_i$, M'$_{i+\frac{s+1}{2}}$, M$_{i+1}$, ..., et la pente p devient :

$$p + f_i, \qquad p + f_i - f_{i+\frac{s+1}{2}}, \qquad p + f_i - f_{i+\frac{s+1}{2}} + f_{i+1}, \ldots$$

Une condition nécessaire est donc :

$$f_{i+\frac{s+1}{2}} = 0.$$

C) *Définition de la médiane dans l'intervalle médian.*

Cisbani étend le principe de l'équation de Jackson à une série cyclique; le point médian sera X tel que, si l'on pose

$$T_p = \Sigma f_i \mid \overset{\frown}{M M_i} \mid^p$$

X soit la limite du point M qui rend T$_p$ minimum, lorsque p tend vers 1.

Le calcul se fait pour chaque intervalle médian séparément; soit (0, 1) l'intervalle médian d'une série de 2_k modalités, on a :

$$T_p = f_0 x^p + f_{-1}(x+1)^p + f_{-2}(x+2)^p + \ldots + f_{-k+2}(x+k_{-2})^p + f_{-k+1}(x+k_{-1})^p$$
$$+ f_1(1-x)^p + f_2(2-x)^p + \ldots + f_{k-1}(k-1-x)^p + f_k(k-x)^p$$

En écrivant $\frac{1}{p}\frac{d\,T_p}{d\,x} = 0$ et en posant $p = 1 + h$, enfin en passant à la limite, il vient

$$(1-x)^{f_1}(2-x)^{f_2} \ldots (k-1-x)^{f_{k-1}}(k-x)^{f_k}$$
$$= x^{f_0}(x+1)^{f_{-1}}(x-2)^{f_{-2}} \ldots (x+k-2)^{-k+2}(x+k-1)^{-k+1}.$$

L'existence, l'unicité et le calcul numérique de la racine de cette équation dans l'intervalle considéré donnent lieu aux mêmes développements que pour une série ordinaire (1).

Annexe III — Variabilité, Différence moyenne, Concentration

Première partie : Définitions :

Série $a_1, a_2 \ldots . a_n$

Différence moyenne : sans répétition.

Simple : $\Delta = \dfrac{\Sigma_{ij} \mid a_i - a_j \mid}{n(n-1)}$ $\qquad \Delta_R = \dfrac{\Sigma_{ij} \mid a_i - a_j \mid}{n^2}$.

Quadratique : $(^2\Delta)^2 = \dfrac{\Sigma_{ij}(a_i - a_j)^2}{n(n-1)}$ $\qquad (\Delta_R)^2 = \dfrac{\Sigma_{ij}(a_i - a_j)^2}{n^2}$.

Distribution continue: densité de probabilité (f (x)).

Différence moyenne :

Simple : $\Delta = \int\int \mid x - y \mid f(x) f(y)\, dx\, dy = 2 \int\int_{x-y>0} (x-y) f(x) f(y)\, dx\, dy.$

Quadratique : $(^2\Delta)^2 = \int\int (x-y)^2 f(x) f(y)\, dx\, dy = 2 \int x^2 f(x)\, dy \int (y)\, dy$

$$+ 2 \int x f(x)\, dx. \int y f(y)\, dy.$$

Cette dernière formule montre que $^2\Delta$ s'exprime à l'aide des moments des 1er et 2^e ordres de la distribution, ce qui la rend sans intérêt.

(1) Metron, VIII, 1, 2 (1929).

Valeur de Δ pour quelques distributions classiques :

Loi de Gauss-Laplace : $f(x) = \frac{e^{-\frac{x^2}{2}}}{\sqrt{2\pi}}$ $\qquad \Delta = \frac{2}{\sqrt{\pi}}, \quad {}^2\Delta = \sqrt{2}.$

Première loi de Laplace : $f(x) = \frac{e^{-|x|}}{2}$ $\qquad \Delta = \frac{5}{2} - , \quad {}^2\Delta = \sqrt{2}.$

Loi binomiale : $f_k = \frac{s!}{k!\,(s-k)!}$ $\qquad \Delta = \frac{1.3.5\ldots(2s-1)}{2.2.4\ldots(2s-2)}.$

Deuxième partie : Calcul pratique de Δ :

A) *Première méthode de Gini :* Soit la série : $a_1 < a_2 < \ldots < a_i < \ldots < a_n$.

Soit à former : $\Sigma_{i,j} \mid a_i - a_j \mid = 2 \sum_{i>j} (a_i - a_j)$.

On peut n'introduire que les différences des termes symétriques de la série. On voit en effet :

que, si $a < b < c < d$,

$$(b-a) + (c-a) + (d-a) + (c-b) + (d-b) + (c-b) = 3(d-a) + c - b).$$

si $a < b < c < d < e$,

$$(b-a) + (c-a) + (d-a) + (e-a + (c-b) + (d-b) + (e-b) + (d-c) + (e-c) + (e-d) = 4(e-a) + 2(d-b).$$

Plus généralement la somme de tous les termes distincts est :

$$\sum_{i>j} (a_i - a_j) = \sum_{i=1}^{\frac{n+1}{2}} (n + 1 - 2i)(a_{n+1-i} - a_i).$$

On appelle $n + 1 - 2i = d_{i,\,n-i+1}$ la distance graduée entre deux termes symétriques.

On peut faire intervenir la *médiane :* $\mathrm{M} = a_{\frac{n+1}{2}}$ (n impair) ou $a_{\frac{n}{2}} < \mathrm{M} < a_{\frac{n}{2}+1}$ (n pair)

Il vient : $a_{n+1-i} - a_i = (a_{n+1-i} - \mathrm{M}) + (\mathrm{M} - a_i) = 2 \mid a_i - \mathrm{M} \mid$

d'où l'expression $\sum_{i>j} (a_i - a_j) = \sum_{i=1}^{n} d_{i,\,\mathrm{M}} \mid a_i - \mathrm{M} \mid$.

D'après de Finetti, cette méthode de calcul s'adapte mal au cas général où les a_i ne sont pas tous distincts; de sorte qu'on ne l'emploie pas.

B) *Méthode de Czuber :* Raisonnons sur la distribution continue :

$$\Delta = 2\int_{x>h}\int (x-y) f(x)\,dx f(y)\,dy = 2\int_{x-y>0}\int xf(x)\,dx.f(y\,(dy - 2\int_{x-y>0}\int yf(y)\,dy.f(x)\,dx$$

$$= 2\int_{x-y>0}\int xf(x)\,dx.f(y)\,dy - 2\int_{y-x>0}\int xf(x)\,dx.f(y)\,dy$$

Posons : $\int xf(x)\,dx = \varphi(x)$, il vient :

$$\Delta = 2\int_{-\infty}^{+\infty} \left[\varphi(x)\right]_y^{+\infty} f(y)\,dy - 2\int_{-\infty}^{+\infty} \left[\varphi(x)\right]_{-\infty}^{y} f(y)\,dy$$

$$= 2\int_{-\infty}^{+\infty} \left\{\left[\varphi(x)\right]_y^{+\infty} - \left[\varphi(x)\right]_{+\infty}^{y}\right\} f(y)\,dy.$$

Cette formule sert ici au calcul numérique, et la méthode convient très bien à une distribution de fréquences :

Soit : $\mathrm{X}_1 < \mathrm{X}_2 < \ldots < \mathrm{X}_s$ les valeurs de la variable.

Soit : $n_1\ n_2\ \ldots\ .$ $\qquad n_s$ les nombres des fois où elles ont été prises.

Le dénominateur est $\frac{(\Sigma n_s)_2}{2}$ ou $\frac{\Sigma n_s\,[\Sigma (n_s) - 1]}{2}$.

Le numérateur de Δ ou Δ_R est :

1er terme : $n_1 (n_2 (X_2 - X_1) + n_3 (X_3 - X_1) + \ldots\ldots + n_s (X_s - X_1)$.

$= n_1 (n_2 X_2 + n_3 X_3 + \ldots\ldots + n_s X_s) - n_1 X_1 (n_2 + n_3 + \ldots\ldots + n_s)$,

$= n_1 (n_1 X_1 + n_2 X_2 + n_3 X_3 + \ldots\ldots + n_s X_s) - n_1 X_1 (n_1 + n_2 + n_3 + \ldots\ldots + n_s)$

$= n_1 \sum_1^s n_i X_i - n_1 X_1 \sum_1^s n_i$

2e terme : $n_2 [n_3 (X_3 - X_2) + n_4 (X_4 - X_2) + \ldots + n_s (X_s - X_2)]$; le même calcul donne :

$= n_2 \sum_2^s n_i X_i - n_2 X_2 \sum_2^s n_i$

.....

et au total : $n_1 \sum_1^s n_i X_i + n_2 \sum_2^s n_i X_i + n_3 \sum_3^s n_i X_i + \ldots + n_s (n_s X_i)$.

$- n_1 X_1 \sum_1^s n_i - n_2 X_2 \sum_2^s n_i - n_3 X_3 \sum_3^s n_i - \ldots - n_s X_s (n_s)$.

Les termes soustractifs peuvent s'écrire autrement en regroupant par colonne :

$$\begin{array}{r} (n_1 + n_2 + n_3 + \ldots + n_s)\, n_1 X_1 \\ (n_2 + n_3 + \ldots + n_s)\, n_2 X_2 \\ \ldots\ldots \\ (n_s)\, n_s X_s \\ \hline n_1 (X_1 n_1) + n_2 \sum_1^2 n_i X_i + \ldots + n_s \sum_1^s n_i X_i. \end{array}$$

Au total, le numérateur est donc :

$$\sum_1^s n_k (A_k - B_k) \qquad \text{avec} \begin{cases} A_k = \sum_k^s n_i X_i \\ B_k = \sum_1^k n_i X_i \end{cases}$$

Cette formule correspond à celle de calcul intégral trouvée plus haut.

En particulier, pour une série $a_1 < a_2 < \ldots < a_n$, on a :

$B_1 = a_1$; $B_2 = a_1 + a_2$; $B_n = a_1 + a_2 + \ldots + a_n$

Posons : $S' = B_1 + B_2 + \ldots B_n$

$A_1 = a_n$; $A_2 = a_n + a_{n-1}$; $A_n = a_n + a_{n-1} + \ldots + a_1$

$S = A_1 + A_2 + \ldots + A_n$.

Le numérateur de Δ et Δ_R est $S' - S$.

tandis que : $S + S' = \sum_{i=1}^{n} (n+1)\, a_i = n (n+1)\, \bar{a}$ ($\bar{a}$ moyenne arithmérique,

d'où des relations

$S' - S = n (n+1)\, \bar{a} - 2 S$,

$= 2 S' - n (n+1)\, \bar{a}$

qui ont été utilisées par divers auteurs pour le calcul numérique.

C) *Seconde méthode de Gini :*

Gini a introduit plus tard le rapport de concentration R et a trouvé que $R = \frac{\Delta_R}{2 \bar{a}}$.

Vérifions cette relation au moyen d'une distribution continue : on a vu que :

$$\Delta_R = \Delta = 2 \int_{-\infty}^{+\infty} \left\{ \left[\varphi (y) \right]_y^{+\infty} - \left[\varphi (y) \right]_{-\infty}^{y} \right\} d F (y)$$

avec $\varphi (y) = \int x f (x)\, dx$.

Introduisons $A = \int_{-\infty}^{+\infty} x f (x)\, dx$ (valeur moyenne)

et posons $\varphi(y) = AG(y)$ avec $\int_{-\infty}^{+\infty} G(y)\,dy = 1$.

Il vient :

$$\Delta_R = \Delta = 2A\int_{-\infty}^{+\infty}\Big[(1 - G(y) - G(y) + 0\Big]\,dF(y)$$

$$= 2A\int_{-\infty}^{+\infty}(1 - 2G)\,dF$$

$$\text{d'où } \frac{\Delta_R}{2A} = \int_{-\infty}^{+\infty} dF - 2\int_{-\infty}^{+\infty} G\,dF = 1 - 2\int_{-\infty}^{+\infty} G\,.\,dF.$$

On retrouvera plus loin cette expression de la concentration R.

D) *Méthode de Paciello et de Finetti.*

La définition de Δ pour une distribution continue étant $2\iint_{x>y}(x-y)\,dF(x)\,.\,dF(y)$

on pose : $x - y = \int_y^x dz$, d'où

$$\Delta = 2\iiint_{x>z>y} dF(x)\,.\,dF(y)\,.\,dz$$

$$= 2\int_{-\infty}^{+\infty}\left\{\int_z^{+\infty} dF(x)\,.\,\int_{-\infty}^{z} dF(y)\right\}dz$$

$$= 2\int_{-\infty}^{+\infty}\Big[1 - F(z)\Big]\,.\,F(z)\,.\,dz.$$

Cette formule sert au calcul numérique comme on va le voir :

1° *Série :* $a_1 < a_2 < \ldots < a_n$. Soit $a_j < a_i$.

$$a_i - a_j = (a_i - a_{i-1}) + (a_{i-1} - a_{i-2}) + \ldots + (a_{j+1} - a_j)$$

d'où : $$\Sigma_{ij}(a_i - a_j) = \sum_{h=1}^{n-1} C_h\,(a_{h+1} - a_h),$$

où C_h est le nombre de termes $(a_i - a_j)$ contenant $(a_{h+1} - a_h)$, c'est à dire pour lesquels on a : $i \geqslant h+1$ et $j \leqslant h$

On a : $n > i \geqslant h+1$ pour $n - h$ termes,

$1 < j \leqslant h$ pour h termes,

d'où : $C_h = (n-h)\,h$

$$\text{et } \sum_{i>j}(a_i - a_j) = \sum_{h=1}^{n-1} h\,(n-h)\,(a_{h+1} - a_h).$$

2° *Distribution de fréquences :* En posant $\sum_1^h n_i = N_h$, on trouve de même :

$$\sum_{i>j}(a_i - a_j) = \sum_{h=1}^{s-1} N_h\,(N_s - N_h)\,(X_{h+1} - X_h).$$

Si toutes les modalités X, sont également espacées, il reste enfin :

$$\Delta = \frac{2\sum_{h=1}^{s-1} N_h\,(N_s - N_h)}{N_s\,(N_s - 1)}$$

Troisième partie : Le Rapport de concentration.

A) De Pareto à Gibrat, le nombre est grand des essais pour représenter la concentration des entreprises, des capitaux, etc., au moyen d'un nombre synthétique. La concentration n'est d'ailleurs, au fond, rien d'autre que la variabilité relative de certaines distributions fortement dissymétriques. On demande à un indice pratique de concentration d'être suffisamment précis, — de n'exiger ni calculs laborieux, ni hypothèses éloignées de la réalité, — de ne pas supposer la connaissance de l'intensité de chaque phénomène *individuel* puisqu'en fait les statistiques connues groupent les entreprises, les capitaux, etc. *par tranches.*

La première idée de Gini (en 1910) ne semble pas avoir été féconde.

Soit une série : $a_1 \leqslant a_2 \leqslant \ldots \leqslant a_i \leqslant \ldots \leqslant a_n$, dissymétrique. La moyenne des m derniers termes est certainement supérieure à la moyenne générale :

$$\frac{1}{m}\sum_{n-m+1}^{n} a_i > \frac{1}{n}\sum_1^n a \qquad \text{ou : } \frac{\sum_{n-m+1}^{n} a_i}{\sum_1^n a_i} > \frac{m}{n} \qquad (< 1)$$

et cette inégalité sera d'autant plus forte que la distribution sera plus concentrée. On considère alors $\delta(m)$ tel que :

$$\left(\frac{\sum_{n-m+1}^{n} a_i}{\sum_{1}^{n} a_i} \right)^{\delta} = \frac{m}{n}.$$

On peut appeler δ un indice descriptif de concentration. D'ailleurs, on ne le calcule pas pour toutes les valeurs de m en raison des tranches déjà signalées.

Peu après, Gini, *changeant complètement de procédé*, allait introduire à la fois une courbe et un indice de concentration.

B) *Cas d'une distribution continue :* Soit F (x) et G (x) définies précédemment; on pose

$$\int_{-\infty}^{x} d\,\mathrm{F}(x) = \Big[(\mathrm{F}(x)\Big]_{-\infty}^{x} = p \qquad \int_{-\infty}^{x} d\,\mathrm{G}(x) = \Big[\mathrm{G}(x)\Big]_{-\infty}^{x} = q$$

et on appelle courbe de concentration le lieu du point $(p\,.\,q)$ en coordonnées rectangulaires.

Exemple : Pour une *distribution uniforme* sur un segment $(a.\ b)$ et nulle ailleurs, la courbe de concentration a pour équations $b \geqslant x \geqslant a$:

$$\left.\begin{aligned} p &= \frac{x-a}{b-a} \\ q &= \frac{n_2 - a_2}{b_2 - a_2} \end{aligned}\right|$$

d'où :

$$q = \frac{b-a}{b+a}\,p^2 + \frac{2_a}{b+a}\,p \qquad 0 \leqslant p \leqslant 1$$

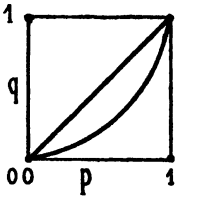


Fig. 3.

C'est un arc de parabole; on a :

$$\mathrm{A} = \int \frac{x\,dx}{b-a} = \frac{b^2 - a^2}{2\,(b-a)} = \frac{b+a}{2}$$

$$\text{puis } \Delta : = 2\,\mathrm{A}\left(1 - 2\int_{-\infty}^{+\infty} q\,dp\right)$$

$$\text{ou } \Delta = (a+b)\left[1 - 2\int_{a}^{b} \frac{(x^2 - a^2)\,dx}{(b^2 - a^2)\,(b-a)}\right] = \frac{b-a}{3}$$

Rapport de concentration. — On appelle ainsi le rapport de l'aire comprise entre la courbe de concentration et la diagonale à l'aire du demi carré, c'est à-dire :

$$\mathrm{R} = \frac{\int p\,d\,q - 1/2}{1/2} = 2\int p\,d\,q - 1$$

Comme, d'autre part, on a : $\int q\,d\,p + \int p\,d\,q = 1$, il vient :

$$\mathrm{R} = \int p\,d\,q - \int q\,d\,p$$

$$\text{ou} \quad \mathrm{R} = 1 - 2\int q\,d\,p.$$

Pour une distribution uniforme par exemple, il vient : $\mathrm{R} = \dfrac{b-a}{3\,(b+a).}$

et en particulier, si $a = 0$ il vient $\mathrm{R} = \frac{1}{3}$ (quelque soit b.) (la courbe a son sommet en 0).

Pour une distribution linéaire la courbe de concentration est déjà un arc de cubique; et le calcul de R est peu élégant.

Quoiqu'il en soit, il faut reconnaître que R présente un grand avantage : on peut le calculer sans faire aucune hypothèse sur la forme de la distribution (contrairement aux indices de Pareto ou Gibrat).

C) *Cas d'une distribution pratique.*

En pratique x sera par exemple un revenu; $p\,(x)$ est alors la proportion de personnes ayant un revenu au plus égal à x, $q(x)$ la proportion du revenu total qui est répartie en revenus au plus égaux à x. On transporte donc ces définitions aux distributions discontinues,

étant entendu qu'on ne connaît, comme valeurs de p et q, que celles qui correspondent aux séparations des tranches. On suppose tous les revenus rangés dans l'ordre croissant.

On connaît donc : $p_i = \frac{1}{n} \sum_{j=1}^{i} n_j$ $\qquad$ $q_i = \frac{1}{N} \sum_{j=1}^{i} n_j x_j$

avec $n = \sum_{j=1}^{s} n_j$ $\qquad$ avec $N = \sum_{j=1}^{s} n_j x_j$.

x_j étant le *revenu moyen* de la tranche considérée comprenant n_j éléments.

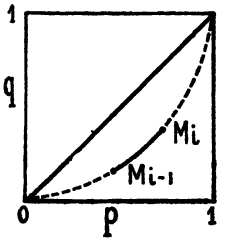


Fig. 4.

On appelle *courbe de concentration* la ligne polygonale brisée joignant les points M_i $(p_i\ q_i)$ dans l'ordre des i croissants, en axes rectangulaires.

Un point M de cette ligne, compris entre les sommets M_{i-1} et M_i, a pour coordonnées :

$$p = p_{i-1} + \rho\,(p_i - p_{i-1}),$$

$$0 < \rho < 1$$

$$q = q_{i-1} + \rho\,(q_i - q_{i-1}).$$

Il représente donc tous les revenus de rang inférieurs à $n\,p_{i-1}$, plus la fraction ρ de ceux compris entre le rang $n\,p_{i-1}$ et le rang $n\,p_i$ en admettant que leur valeur moyenne soit toujours x_i. *Ainsi joindre les points* M_i *par des segments de droite revient à supposer que, dans chaque tranche, tous les revenus sont égaux.* (Si le nombre de tranches augmentait, la ligne brisée aux côtés de plus en plus nombreux tendrait vers une courbe continue.)

Convexité de la courbe de concentration : La pente de la droite $M_{i-1}M_i$ est $\frac{q_i - q_{i-1}}{p_i - p_{i-1}} = x_i \frac{N}{n}$.

Elle croît régulièrement et par sauts; en outre, on admettra ici qu'elle est positive.

Cas de l'égalité des revenus : p et q croissent de 0 à 1 quand on énumère tous les revenus, donc d'après ce qui précède, la courbe de concentration est un segment de droite; c'est donc la *diagonale.*

Cas d'un seul revenu non nul : On a $p_1 = \frac{1}{n}$, $q_1 = 0$; $p_{n-1} = \frac{n-1}{n}$, $q_{n-1} = 0$; $p_n = q_n = 1$.

La courbe de concentration correspondante est tracée sur la figure ci contre.

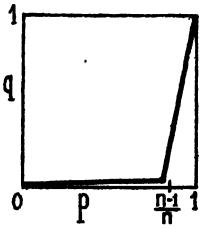


Fig. 5.

Rapport de concentration : Par définition, le rapport de concentration R de Gini s'obtient en rapportant à son maximum l'aire comprise entre la courbe de concentration et la diagonale (courbe de concentration nulle).

Cette définition assure un rapport égal à 0 en cas d'égalité des revenus et à 1 en cas d'accaparement de ceux ci.

Calcul de l'aire comprise entre la courbe de concentration et la diagonale.

Calculons son complément à $\frac{1}{2}$, formé de trapèzes rectangles d'aire :

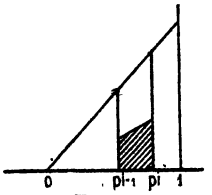


Fig. 6.

$$\frac{1}{2}\,(q_{i-1} + q_i)\,(p_i - p_{i-1}) = \frac{1}{2}\,(q_{i-1} + q_i)\,n_i.$$

Donnons à p_i les valeurs $0, \frac{1}{n}, \frac{2}{n}, \ldots\ldots \frac{n-1}{n}, \frac{n}{n} = 1$,

(en supposant égaux tous les revenus d'une tranche.) L'aire est manifestement :

$$\sum_{1}^{n-1} q_i + \frac{1}{2n}.$$

Donc, si l'on prend un à un tous les revenus, R est donné par la formule :

$$\frac{\frac{1}{2} - \sum_{1}^{n-1} q_i - \frac{1}{2n}}{\frac{n-1}{2n}} = \frac{n-1-\left(\sum_{1}^{n-1} q_i\right) 2n}{n-1} = 1 - \frac{2n}{n-1} \sum_{1}^{n-1} q_i.$$

Si l'on ne peut envisager que les revenus par groupe, il faut calculer :

$$1 - \frac{n}{n-1}\left(\sum_{1}^{n} q_{i-1} n_i + \sum_{1}^{n} q_i n_i\right)$$

Il existe en fait d'ailleurs des formules approchées beaucoup plus simples.

D) *Variabilité relative.*

Il existe deux moyens de comparer les variabilités de deux séries statistiques : le plus connu est de les rapporter aux moyennes arithmétiques respectives (coefficients de variabilité), c'est-à dire de passer à des variables de valeur moyenne unité. Gini emploie un second moyen : il rapporte la variabilité à son *maximum ;* c'est ce qu'on appelle la variabilité relative ou indice de variabilité.

Puisqu'on ne mesure pas la variabilité mais qu'on la *repère* à l'aide de ${}^1S_A$ ${}^1S_M$, ${}^2S_A$ ${}^2S_M$ Δ ou ${}^2\Delta$, il faut rapporter chacun de ses nombres à *son* maximum respectif.

Les écarts moyens et différences moyennes atteignent heureusement leurs maxima en même temps, pour la distribution de concentration maximum (et s'annulent pour une distribution uniforme).

Maximum de ${}^1S_A = 2\frac{n-1}{n}A$	Maximum de ${}^1S_M = A$
» ${}^2S_A = \sqrt{n-1}\,A$	» ${}^2S_M = \sqrt{n}\,A$
» $\Delta = 2A$	» $\Delta_R = 2\frac{n-1}{n}A$

Parmi les autres indices de variabilité possibles, la différence interquartile ou l'intervalle de variation (*range*) n'atteignent pas leur maximum ou leur minimum dans les mêmes conditions.

On a pu reprocher à Gini que son indice de variabilité relative fut $\frac{\Delta}{2A} = R$, alors qu'on l'avait précisément conçu en vue de se passer de toute valeur centrale; et Vinci a proposé l'emploi d'une autre expression ne présentant pas cet inconvénient :

$$\Delta' = \frac{1}{n(n-1)} \sum_{i=1}^{n} \left[\frac{1}{a_i} \sum_{k=1}^{n-1} |a_i - a_k|\right] = \frac{2}{n(n-1)} \left\{\sum_{i=1}^{n-1} \frac{A_i}{a_i} - \sum_{i=1}^{n-1} \frac{B_i}{a_{n-i+1}}\right\}$$

En fait la théorie n'a pas évolué dans cette direction. On a perfectionné, au contraire, la notion de maximum en tenant compte des limites que les termes de la série peuvent atteindre *effectivement* et qui varient selon les problèmes. Par exemple, telle série de fractions aura ses termes qui ne peuvent jamais dépasser 1 et cela n'aura aucun sens de rapporter sa variabilité à un maximum absolu. On supposera $\frac{nA}{L} = k$ entier.

Si L est le maximum des termes de la série, ceux des indices de variabilité sont :

Pour ${}^1S_A$: $\frac{2A(L-A)}{L}$	Pour ${}^1S_M$: A	Pour Δ $2\frac{n}{n-1}\frac{A(L-A)}{L}$
${}^2S_A$: $\sqrt{A(L-A)}$	${}^2S_M$: $\sqrt{AL}$	Pour Δ_R : $2\frac{A(L-A)}{L}$

Les formules valables lorsque $\frac{nA}{L}$ n'est pas entier sont plus compliquées.

Pour le *rapport de concentration*, en portant $\frac{k}{n} = \frac{A}{L}$ sur l'axe des p, on obtient le point D, tel que la courbe de concentration maximum soit ODB.

$$\text{et } R' = \frac{\text{aire } \mathcal{A}}{\text{aire ODB}}$$

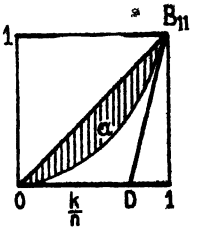


Fig. 7.

On donne également les formules relatives à l'existence d'un minimum $l > 0$ pour les termes de la série.

Enfin le cas de termes négatifs doit être envisagé, lorsque les distributions concernent

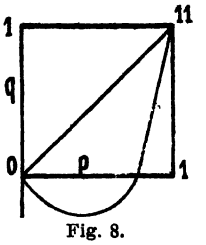


Fig. 8.

d'autres quantités que des revenus, salaires, fortunes,..... La courbe de concentration est donc, au départ, sous l'axe des p ; mais la concavité reste tournée vers les $q > 0$.

Les formules de concentration ne s'appliquent plus alors (1).

Annexe IV — Indice de cograduation

Ce que les Anglo-Saxons appellent *rank correlation* est nommé *cograduation* par Gini.

Étant donné une série de termes : $a_1, a_2 \ldots a_i \ldots a_n$, on peut les ranger par ordre de grandeur croissante; soit alors r_i le *rang* de a_i. Ce rang est susceptible de remplacer la valeur de a_i dans certaines études.

Soit A et B deux caractères prenant les valeurs *jumelées* $a_i\ b_i$; on suppose les b_i rangés dans l'*ordre croissant*. On comparera les deux séries :

$$\begin{matrix} r_1 & r_2 \ldots r_i \ldots r_n \\ 1 & 2 \ldots i \ldots n \end{matrix}$$

en formant $2\,L = |r_1 - 1| + |r_2 - 2| + \ldots + |r_i - i| + \ldots + |r_n - n|$,
qui n'est nul que si les rangs coïncident (cograduation), et qui est maximum si les r_i sont

$$n\ (n-1) \ldots (n-i+1) \ldots 1.$$

Ce maximum est $2\left[(n-1) + (n-3) + (n-5) + \ldots\ldots + \left\{\begin{matrix} 3+1 \\ 4+2 \end{matrix}\right\}\right]$

donc $\dfrac{n^2}{2}$ (si n est pair)

ou $\dfrac{n^2-1}{2}$ (si n est impair).

On comparera de même les deux séries :

$$\begin{matrix} r_1 & r_2 \ldots & r_i \ldots & r_n, \\ n & n-1 \ldots & n-i+1 \ldots & 1 \end{matrix}$$

en formant $2\,L' = |r_1 - n| + |r_2 - n + 1| + \ldots + |r_i - n + i - 1| + \ldots + |r_n - 1.|$

On peut alors repérer :

— la cograduation avec $u = 1 - \dfrac{L}{L\,max}$ $\quad\left\{\begin{matrix} u = 1 \text{ si cogradués} \\ u = 0 \text{ minimum} \end{matrix}\right.$,

— la contregraduation avec $u' = 1 - \dfrac{L'}{L'\,max}$ $\quad\left\{\begin{matrix} u' = 1 \text{ si contregradués} \\ u' = \text{minimum} \end{matrix}\right.$.

Gini forme l'indice de cograduation :

$$\boxed{I = u - u'}.$$

On a bien : $-1 \leqslant I \leqslant +1$.

avec $\left\{\begin{matrix} I = 1 \text{ pour la covariation,} \\ I = -1 \text{ pour la contrevariation.} \end{matrix}\right.$

---

(1) Gini, 1910 : *Indici di concentrazione e di dipendenza* ; 1912 : *Variabilita e Mutabilita* (p. 80). — Vinci, Metron, I, 1 (p. 62); Gini, VIII, 3, 1930 et IX, 3, 4, 1931; Castellano, XIII, 1 (p. 31 à 49); de Finetti Metron, VIII, 3 et IX, 1. — Différence moyenne de la loi de Poisson. (Voir Herman Wold, Metron. XII, 2

Ainsi on a

$$I = \frac{-\sum_{i=1}^{n} \left| r_i - i \right| + \sum_{i=1}^{n} \left| r_i - n - i + 1 \right|}{K} \text{ avec } \begin{cases} K = \frac{n^2}{2} \text{ (si } n \text{ } pair\text{)}, \\ K = \frac{n^2 - 1}{2} \text{ (si } n \text{ impair)}. \end{cases}$$

En appelant p_i et q_i les *rangs* d'un couple, les couples étant dans un ordre quelconque, on aura :

$$I_1 = \frac{-\sum_{i=1}^{n} \left| p_i - q_i \right| + \sum_{i=1}^{n} \left| p_i - q_i' \right|}{K} \quad \text{avec } q'_i = n + 1 - q_i.$$

Par exemple : si A est la production « lourde » rapportée à la superficie agraire et forestière et B la superficie agraire et forestière rapportée à la population active, si ces quantités sont connues par *province*, si l'on classe les 16 provinces suivant leurs rangs pour chacune de ces caractéristiques, on trouve une covariation de — 0,61.

En fait si I n'est pas très voisin de 1 ou de — 1, il perd vite toute signification.

Un indice quadratique de *cograduation* a été en outre imaginé, selon l'habitude de Gini où H est une constante appropriée :

$$H = \frac{n\,(n^2 - 1)}{3} \qquad I_2 = \frac{\Sigma\,(p - q')^2 - \Sigma\,(p - q)}{H}$$

Ces indices sont très utiles si les deux caractères A et B ou l'un seulement ne sont connus que par leur rang. Par exemple, on a composé l'*ordre d'arrivée* des chevaux et les *sommes* misées sur eux, aux courses (on n'a pas en effet l'habitude de mesurer la vitesse des chevaux mais seulement leur rang). On a *renversé l'ordre* pour étudier la cograduation. On n'a trouvé qu'un coefficient de 0,40 en moyenne, c'est à dire peu significatif. (Zingali : *Mesure statistique de l'aptitude à gagner les courses de galop*).

M. Salvemini s'est occupé de définir l'indice de cograduation lorsque les termes ne sont pas tous distincts. Sa façon de faire est connue des Anglo-Saxons sous le nom de « *mid-rank method* ».

Exemple	17	17	16	15	15	14	13	12 .....
	1,5		3	4,5		6e	7e	8e .....

Et une autre méthode est possible : « *Bracket rank method* » :

1e	1e	3e	4e	4e	6e	7e	8e .....

(celle employée aux compositions, examens et concours en France).

La difficulté rencontrée par Salvemini est de donner au dénominateur K la valeur appropriée (: maximum exact). Il a trouvé que celle-ci était

$$\frac{n^2}{2} - 2\,m^2 \qquad \text{si } n \text{ est pair,}$$

$$\frac{n^2 - 1}{2} - 2\,m\,(m + 1) \quad \text{si } n \text{ est impair.}$$

m étant le nombre de termes à droite du terme central qui sont répétés à sa gauche (1).

Annexe V — Dissemblance, Concordance, Connexion

I. — *Dissemblance.*

On considère deux distributions statistiques obtenues *séparément* et on les compare.

Ce qu'on appelle *courbe de probabilités totales* devient ici, après une rotation de $+\frac{\pi}{2}$ et une symétrie, la *courbe de graduation ;* à cet effet, on définit, *comme pour la concentration*, une quantité p_i, qui servira d'abscisse. La valeur x des termes de la série est portée en ordonnée :

Soit x_1 et x_2 les ordonnées sur les deux courbes de graduation, pour une même abscisse F.

(1) Metron, I, 1, p. 131 ; XIII, 4 (1939), p. 27 et 41.

La dissemblance simple entre les deux distributions est :

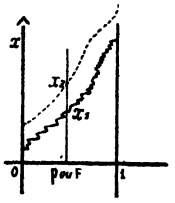


Fig. 9.

$$^1D = \int_0^1 |x_1 - x_2| \, dF.$$

Lorsque les deux courbes ne se coupent pas, elle est égale à $(\overline{x_1} - \overline{x_2})$ (de sorte que l'on n'a ainsi introduit une nouvelle grandeur que si les deux courbes se coupent au moins une fois).

La dissemblance quadratique est donnée par :

$$(^2D)^2 = \int_0^1 (x_1 - x_2)^2 \, dF = \int x_1^2 \, dF + \int x_2^2 \, dF - 2 \int x_1 x_2 \, dF.$$

Cette quantité, s'exprimant à l'aide d'autres bien connues, ne présente pas autant d'intérêt que la première.

Cas de deux séries de chacune n termes. $a_1 \leqslant a_2 \leqslant \ldots \leqslant a_i \leqslant \ldots a_n$; $b_1 \leqslant b_2 \leqslant \ldots \leqslant b_i, \ldots \leqslant b_n$

Les dissemblances sont :

$$^1D = \sum_{i=1}^{n} \left| \frac{a_i - b_i}{n} \right|$$

$$(^2D)^2 = \sum_{i=1}^{n} \frac{(a_i - b_i)^2}{n}.$$

Cas de deux séries ayant des nombres distincts de termes. — Le calcul de l'aire comprise entre les deux courbes de graduation revient à considérer deux séries ayant chacune N termes, N étant un multiple commun des nombres distincts de termes.

II. — *Différence.*

On considère deux distributions ou séries telles que chaque terme de l'une soit accouplé à un terme de l'autre (par exemple : deux séries chronologiques, les termes d'une même année étant associés).

Soit

$$A_1 \quad A_2 \quad \ldots\ldots \quad A_n,$$
$$B_1 \quad B_2 \quad \ldots\ldots \quad B_n.$$

On appelle *différences* les expressions telles que

$$|A_1 - B_1| + |A_2 - B_2| + \ldots\ldots + |A_n - B_n| = AB$$

et

$$(A_1 - B_1)^2 + (A_2 - B_2)^2 + \ldots\ldots + (A_n - B_n)^2 = (\overline{AB}_2)^2.$$

Gini a démontré que la *différence* simple entre deux séries a pour minimum la *dissemblance* simple (obtenue en associant les termes de même rang dans les deux séries) ; tandis que le maximum de différence était obtenu pour deux séries contregraduées.

On peut établir la première partie de la proposition en montrant d'abord que, si l'on a $a < b$ et $x < y$, on a aussi $(a - x) + (b - y) \leqslant (a - y) + (b - x)$; puis en partant des *deux séries cograduées* et en permutant deux termes dans la *dissemblance* autant de fois qu'il le faudra, on pourra reconstituer la *différence*. Or chacune des permutations peut se faire de telle façon que l'expression considérée ne puisse qu'augmenter. *cqfd.*

La seconde partie se démontre de la même façon.

Dissemblance de deux distributions associées par un tableau à double entrée.

Les valeurs des deux séries ne sont pas en général cograduées. Mais on construit facilement un tableau à double entrée ayant les mêmes *marg s* que le tableau donné mais tel que les deux séries qu'il associe soient cograduées.

De la même façon, on peut former le tableau correspondant à la cograduation.

Exemple :

x_2 \ x_1	1	2	3	MARGE
1	1 411	105	0	1 516
2	0	1.281	0	1.281
3	0	21	1 325	1 346
Marge	1 411	1 407	1.325	4 143

Tableau de deux séries cograduées.

$\Sigma |a_i - b_i| = (1411 + 1281 + 1325) \times 0.$
$+ (105 + 21) \times 1,$
$+ 0 \times 2.$
$= 126.$

x_2 \ x_1	1	2	3	MARGE
1	0	191	1.325	1.516
2	65	1.216	0	1 281
3	1 346	0	0	1 346
Marge	1.411	1.407	1 325	4 143

Tableau de deux séries contregraduées.

$\Sigma |a_i - b_i| = (1216 \times 0,$
$+ (196 + 65) \times 1,$
$+ (1325 + 1346) \times 2,$
$= 5.598.$

On considère également la *différence moyenne de deux séries* A_i et B_i : ce sera la moyenne de toutes les différences possibles obtenues en associant au hasard les A_i et les B_j.

Par exemple, pour $\left\{ \begin{matrix} A_1 & A_2 & A_3 \\ B_1 & B_2 & B_3 \end{matrix} \right\}$, ce sera (différence simple)

$$(\overline{A\,B})_0 = \frac{1}{3}\Big\{ |A_1 - B_1| + |A_1 - B_2| + |A_1 - B_3| + |A_2 - B_1| + |A_2 - B_2| + |A_2 - B_3| + |A_3 - B_1| + |A_3 - B_2| + |A_3 - B_3| \Big\}$$

Il est désormais évident que $(A\,B)_0 \geqslant D$. L'égalité a lieu seulement si deux séries n'ont chacune qu'un terme distinct, à moins que la coïncidence terme à terme, n'ait lieu, auquel cas on a $(A\,B)_0 = 0 = D$.

III — *Concordance (Corrélation, Homophilie)*

On a défini au n° II la quantité $\overline{AB}$ et la moyenne $(A\,B)_0$ de toutes les quantités analogues obtenues en associant les A et B au hasard. La différence :

$$d = \overline{A\,B} - (A\,B)_0$$

devra être voisin de 0 si A et B sont *indifférents*, positive si A et B sont *discordants*, négative s'ils sont *concordants*. On considère alors le maximum M de d positif et celui m de $-d$ (d négatif).

On aura un indice de concordance égal à :

$-\frac{d}{M}$ s'il y a discordance (minimum — 1).

0 pour l'indifférence,

$-\frac{d}{m}$ s'il y a concordance (maximum 1),

D'ailleurs $\overline{A\,B}$ peut être remplacé par $\overline{A\,B}_2$, la différence d peut avoir une expression beaucoup plus compliquée, enfin le maximum peut être relatif ou absolu.

Rapporté au maximum relatif, on a un indice d'*homophilie ;* rapporté au maximum absolu, on a un *indice de corrélation* (au sens de Gini).

Les quantités A_i et B_i sont d'ailleurs chacune, soit les grandeurs elles-mêmes a_i, soit leurs *écarts* $l_i = a_i - \bar{a}$, soit leurs variations $\frac{a_i - \bar{a}}{\sigma} = \varphi_i$.

On retrouve ainsi certains indices simples :

Indice quadratique d'homophilie des grandeurs elles-mêmes.

Par maximum relatif on entend celui que l'on obtient en changeant les termes de place (sans changer les termes eux-mêmes).

On trouve :

$$\text{Concordance : } \underset{\text{cogradué}}{\frac{\Sigma\, l_i\, l'_i}{\Sigma\, l_i\, l'_p}} ; \qquad \text{Discordance : } \underset{\text{centregradué}}{-\frac{\Sigma\, l_i\, l'_i}{\Sigma\, l_i\, l'_q}}$$

Les indices analogues, relatifs aux écarts l_i, s'obtiennent en remplaçant les *écarts* l par les *variations* φ ; enfin ceux relatifs aux *variations* φ_i ne changent pas.

Indices quadratiques de corrélation.

1° *Entre variations.* — Il s'agit de repérer la concordance des quantités : $\varphi_i = \frac{a_i - \bar{a}}{\sigma}$ et $\varphi'_i = \frac{a'_i - \bar{a}'}{\sigma'}$, l'indice d'homophilie étant $\frac{\Sigma\, \varphi_i\, \varphi'_i}{\Sigma\, \varphi_i\, \varphi'_p}$.

Pour trouver le maximum du dénominateur, on remarque que :

$$(x^2 + y^2 + z^2)\,(x'^2 + y'^2 + z'^2) - (xy' - yx')^2 - (yz' - zy')^2 - (zx' - xz')^2 \equiv (xx' + yy' + zz')^2$$

et que, si $\frac{x'}{x} = \frac{y'}{y} = \frac{z'}{z}$, le maximum en résulte :

$$(x^2 + y^2 + z^2)\,(x'^2 + y'^2 + z'^2) = (xx' + yy' + zz')^2.$$

Le maximum de $\Sigma\, p_i\, p'_i$ sera donc $\sigma\sigma'$; et l'indice de corrélation est :

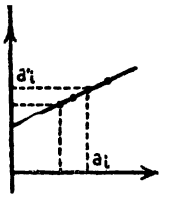


Fig. 10.

$$\frac{\Sigma\,(a_i - \overline{a})\,(a'_i - \overline{a'})}{\sigma^2\,\sigma'^2} \qquad \text{(coefficient } r \text{ de Bravais).}$$

2° *Entre écarts.* — Il s'agit de repérer la concordance entre les quantités $(l_i = a_i - \overline{a})$ et $(l_i = a_i - \overline{a'})$, l'indice d'homophilie étant

$$\frac{\Sigma\; p_i\, p'_i}{\Sigma\; p_i\, p'_p}.$$

On s'était contenté tout à l'heure de supposer p_i et p'_i proportionnels, car

$$\frac{p'_1}{p_1} = \frac{p'_2}{p_2} = \ldots\ldots = \frac{\sqrt{\Sigma\, p'^2_i}}{\sqrt{\Sigma\; p_i^2}} = 1.$$

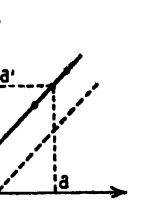


Fig. 11.

Ici il faut imposer l'égalité aux l_i et l'_i. On utilise la formule :

$$(x^2 + y^2 + z^2) + (x'^2 + y'^2 + z'^2) - (x - x')^2 - (y - y')^2 - (z - z')^2 = 2\,(x\,x' + y\,y' + z\,z').$$

Le maximum atteint par $\sigma\,\sigma'$ lorsque $x = x'$, $y = y'$, $z = z'$, est $\dfrac{\sigma^2 + \sigma'^2}{2}$, comme le montre également la formule :

$$\frac{\sigma^2 + \sigma'^2}{2} - \sigma\,\sigma' = \frac{1}{2}\,(\sigma - \sigma')^2.$$

L'indice de corrélation est donc : $\dfrac{\Sigma\,(a_i - \overline{a})\,(a'_i - \overline{a'})}{\left(\dfrac{\sigma^2 + \sigma'^2}{2}\right)^2}$.

3° *Entre grandeurs elles-mêmes.* — Il s'agit de repérer la concordance des a_i et a'_i. L'indice d'homophilie est : $\dfrac{\Sigma\; l_i\, l'_i}{\Sigma\; l_i\, l'_p}$.

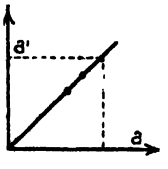


Fig. 12.

On a : $\Sigma\; l_i\, l'_p = \Sigma\; a_i\, a'_p - \overline{a}\,.\,\overline{a'}$,

et le maximum de $\Sigma\; a_i\, a'_p$ est d'après ce qui précède $\dfrac{\Sigma\; a_i^2 + \Sigma\; a'^2_i}{2}$.

L'indice de corrélation est donc :

$$\frac{\Sigma\,(a_i - \overline{a})\,(a'_i - \overline{a'})}{\dfrac{\Sigma\; a_i^2 + \Sigma\; a'^2_i}{2} - \overline{a}\;\overline{a'}} = \frac{\Sigma\,(a_i - \overline{a})\,(a'_i - \overline{a'})}{\dfrac{\sigma^2 + \sigma'^2 + (a - \overline{a'})^2}{2}}$$

On comprend mieux toute la portée de ces distinctions, quand on sait que Pietra ayant comparé les répartitions des revenus et des fortunes en Australie a trouvé les indices de corrélation suivants : 0,084 entre les valeurs, 0,086 entre les écarts, 0,64 entre les variations

IV — *Connexion.*

La *dépendance* en probabilités étant supposée connue, on sait comment on la transpose en statistique; c'est ce que Gini appelle connexion.

A et B sont indépendants (en probabilités) lorsque :

Probabilité de A = Probabilité de A lorsque B est donné,

ce qui entraîne aussi :

Probabilité de B = Probabilité de B lorsque A est donné.

A l'opposé, la dépendance complète consiste dans la relation de fonction entre A et B. Dans l'intervalle, on parle de dépendance stochastique.

En statistique, on considère deux caractères X et Y dont les valeurs sont accouplées (par exemple sont prises simultanément). On suppose que X ne peut prendre que les valeurs $x_1\, x_2 \ldots x_i \ldots$ et Y les valeurs $y_1\, y_2 \ldots y_j \ldots$ Le nombre de fois où le couple $x_i\, y_j$ est ren-

contré est désigné par n_{ij} dans un tableau à double entrée; on pose $\Sigma_i n_{ij} = n_{oj}$; $\Sigma_j n_{ij} = n_{io}$; $\Sigma n_{io} = \Sigma n_{oj} = n_{oo}$.

Il y aura *dépendance de* Y *en* X, dans la mesure où la distribution des cas suivant les diverses valeurs de y dépendra ou non de x_i, c'est à-dire où les distributions constituées par une colonne ne seront pas semblables à la distribution marginale n_{oj}.

	x_1	x_2	..	x_i	...	
y_1	n_{11}	n_{12}	...	n_{i1}	...	n_{o1}
y_2	n_{12}	n_{22}	...	n_{i2}	...	n_{o2}
....	...	...	...	...	...	
y_j	n_{1j}	n_{2j}	...	n_{ij}	. .	n_{oj}
....	...	...	...	...	...	
	n_{10}	n_{20}	...	n_{i0}	...	n_{oo}

La dépendance de x en y se définira de même en comparant les lignes à la distribution marginale n_{io}.

Comparaison de la colonne n_{oi} et de la colonne marginale.

Soit Dx_i la dissemblance et δx_i la différence moyenne; on a :

$$0 \leqslant D_{xi} < \delta_{xi}$$

$$\text{avec } \delta_{xi} = \sum_{hj} |y_h - y_j| \frac{n_{oh}}{n_{oo}} \frac{n_{ij}}{n_{io}}.$$

L'égalité à zéro a lieu si les deux distributions sont proportionnelles.
L'égalité à δ_{xi} se produit si tous les n_{ij} sauf un sont nuls.

Indice de connexion de Y *en* X.

En définissant ses mêmes grandeurs pour chaque colonne, on forme alors par sommation :

$$0 < \Sigma \frac{n_{io} D_{xi}}{n_{oo}} \leqslant \Sigma \frac{n_{io} \delta_{xi}}{n_{oo}}$$

d'où

$$0 < \frac{\Sigma n_{io} D_{xi}}{\Sigma n_{io} \delta_{xi}} \leqslant 1.$$

L'égalité à 0 suppose $\frac{n_{i1}}{n_{o1}} = \frac{n_{i2}}{n_{o2}} = \ldots\ldots = \frac{n_{ij}}{n_{oj}} \ldots$ quelque soit i, d'où $n_{ij} = \frac{n_{io} \cdot n_{oj}}{n_{oo}}$.

L'indépendance ou non connexion de x et y est symétrique.

L'égalité à 1 suppose que, dans chaque colonne, tous les n_{ij} sauf un sont nuls, soit n_{ih}. On a alors $y(x_i) = y_h$ (relation fonctionnelle).

L'indice de connexion de Y en X, $\frac{\Sigma n_{io} D_{xi}}{\Sigma n_{io} \delta_{xi}}$, peut comporter des différences simples ou quadratiques; on peut former de la même façon l'indice de connexion de X en Y.

Indice de connexion des valeurs moyennes de Y *aux valeurs de* X.

Pour $X = x_i$, la valeur moyenne de Y est : $\frac{\Sigma_i n_{ij} y_j}{n_{io}} = \overline{y_i}$.

Quelque soit X, la valeur moyenne de Y est : $\frac{\Sigma_i n_{io} \overline{y_i}}{n_{oo}} = \overline{y}$.

On aurait $y_i = \overline{y}$ s'il y avait non-connexion. On va repérer la connexion à l'aide de l'écart moyen, simple ou quadratique :

$$\sigma^2_{\overline{y_i}} = \frac{\Sigma (\overline{y_i} - \overline{y})^2 n_{io}}{n_{oo}}; \qquad e_{\overline{y_i}} = \frac{\Sigma |\overline{y_i} - \overline{y}| n_{io}}{n_{oo}}.$$

Pour avoir un indice de connexion, on rapporte chacune de ces expressions à son maximum, calculé en supposant qu'il y a relation fonctionnelle, de sorte que tous les n_{ij} de la colonne sont nuls sauf $n_{ih} = n_{io}$. Dans le cas particulier où $n_{io} = n_{ii} = n_{oi}$ quel que soit i, ces expressions sont égales aux quantités analogues calculées à l'aide de la *colonne marginale* :

$$\sigma^2_{\overline{y_j}} = \Sigma \frac{(y_j - \overline{y})^2 n_{oj}}{n_{oo}}; \qquad e_{\overline{y_j}} = \Sigma \frac{|y_j - \overline{y}| |n_{oj}}{n_{oo}}.$$

On montre facilement que, dans le cas général, on a :

$$0 \leqslant \sigma^2_{\overline{y}_i} \leqslant \sigma^2_{\overline{y}_j}; \qquad 0 \leqslant e_{\overline{y}_i} \leqslant e_{\overline{y}_j}$$

d'où

$$0 \leqslant \frac{\sigma^2_{\overline{y}_i}}{\sigma^2_{\overline{y}_j}} \leqslant 1 \qquad ; \quad 0 \leqslant \frac{e_{\overline{y}_i}}{e_{\overline{y}_j}} \leqslant 1.$$

(Coefficient η de Pearson.)

La distinction faite par Gini entre la *connexion* et la *concordance* surprend généralement. « On peut dire... que les indices de connexion (dont le rapport de corrélation r est un cas particulier) caractérisent l'étroitesse de la dépendance, tandis que les indices de concordance (dont le coefficient de corrélation η représente un cas particulier) caractérisent aussi la direction de la dépendance (1) », c'est à dire sont négatifs s'il y a discordance et nuls dans le cas de l'indifférence, alors que les premiers restent compris entre 0 et 1.

Ajoutons qu'on ne s'occupe guère de l'*étroitesse de la dépendance*, tant que le nombre de cas observés est trop faible pour qu'on les disperse dans un tableau à double entrée (1).

Annexe VI — Extensions aux séries cycliques ou non connexes

I — *Indices de mutabilité.*

Gini parle de *variabilité* pour les caractères quantitatifs et de *mutabilité* pour les qualitatifs.

Les indices employés pour les séries linéaires ne présentent aucune originalité nouvelle.

Pour les séries cycliques, on mesure l'écart entre chaque modalité M_i et la moyenne $\overline{M}_i$ de l'intervalle où se trouve M_i (Voir à l'Annexe II la détermination de $\overline{M}_i$) et l'on forme :

$$^1S = \frac{\Sigma f_i \left| \overline{X}_i - X_i \right|}{\Sigma f_i} = \qquad (^2S)^2 = \frac{\Sigma f_i (\overline{X}_i - X_i)^2}{\Sigma f_i}$$

On peut également calculer l'écart par rapport à la médiane M : $\frac{\Sigma f_i (MX_i)}{\Sigma f_i}$ et $\frac{\Sigma f_i (M - X_i)^2}{\Sigma f_i}$.

On définit en outre les différences moyennes par les formules :

$$\Delta = \frac{\Sigma f_i \mu (x_i)}{(\Sigma f_i)^2} \qquad ; \quad ^2(\Delta)^2 = \frac{\Sigma f_i \mu_2 (X_i)}{(\Sigma f_i)^2}.$$

où μ et μ_2 ont les expressions données à l'Annexe II.

II. — *Dissemblance et connexion, pour deux séries cycliques.*

A) *Dissemblance.*

Deux séries étant données sans que leurs éléments se correspondent deux à deux, on les considère comme les marges d'un tableau à double entrée; et l'on détermine les éléments du tableau de façon qu'en outre la somme des écarts soit minimum. On a déjà procédé ainsi pour une série ordinaire; la seule différence réside dans le fait que les *écarts* entre modalités sont comptés sur un cercle.

Exemple :

MOIS DE NAISSANCE	DE LA MÈRE	DE L'ENFANT
1° Janvier à avril. .	1 516	1 411
2° Mai à août . . .	1 281	1 407
3° Sept à décembre	1 346	1.325
Total . . .	4 143	4.143

1°	2°	3°	TOTAL
1.411	105	0	1 516
0	1 281	0	1.281
0	21	1.325	1 346
1.411	1.407	1.325	4.143

Une fois choisis les éléments du tableau, la dissemblance se calcule à l'aide de :

$$0 \times (1411 + 1281 + 1325) + 1\,(105 + 0 + 0) + 1\,(21 + 0 + 0) = 126$$

rapportée au nombre total 4.143 d'éléments :

$$126 : 4143 = 0{,}0304.$$

(1) Gini, *Revue de l'I. I. S.*. 1936, p. 358.

En fait, dès que le nombre de modalités dépasse 3, la détermination des éléments du tableau se complique. De toutes façons, le numérateur de la dissemblance est de la forme :

Différence simple : $S = 0 \times d_0 + 1\ (d_1 + d_{-1}) + 2\ (d_2 + d_{-2}) + \ldots$

Différence quadratique : ${}_2S = 0 \times d_0 + 1^2\ (d_1 + d_{-1}) + 2^2\ (d_2 + d_{-2}) + \ldots$

où d_0 est la somme des termes de la diagonale principale, d_1 celle des termes bordant celle-ci à droite et à gauche, etc... On démontre alors que, pour rendre S minimum, il faut d'abord rendre d_0 maximum, puis d_1 et d_{-1} compte tenu de d_0, puis d_2 et d_{-2}, etc... Cette règle permet à elle seule de reconstituer le tableau.

Par exemple, voici la répartition donnant la somme des écarts minimum :

Janvier à mars	1.082	26	101	0	1 209
Avril à juin	0	931	0	0	931
Juillet à septembre.	0	0	996	0	996
Octobre à décembre	0	0	39	968	1.007
	1.082	957	1.136	968	4.143

On forme : $S = (1082 + 931 + 996 + 968) \times 0 + (26 + 39) \times 1 + 101 \times 2 = 267.$

Le maximum de dissemblance se détermine en choisissant les éléments du tableau (mêmes marges imposées) à l'aide des règles inverses.

B) *Connexion.*

Deux séries sont données par leurs éléments *accouplés*, qui ont permis de constituer un tableau à double entrée. On définit, comme pour les séries ordinaires, la dissemblance entre une colonne et la colonne marginale; on la pondère avec $\frac{n_{i0}}{n_{00}}$ et on totalise pour toutes les colonnes.

III. — *Étude des séries non connexes.*

On a renoncé à donner ici une idée de la complication des calculs et de certains résultats.

1° On considère un espace à s dimensions et les points P_1 (100 . . . 0), P_2 (010 . . . 0), P_s (00 . . . 01) affectés des masses $f_1 f_2 \ldots f_s$. Ces points représentent une série non connexe.

On supposera que $f_1 f_2 \ldots f_s$ sont les fréquences *relatives*, de sorte que $\Sigma f_i = 1$.

2° Le barycentre G des points P_i a pour coordonnées $\left(\frac{f_1}{\Sigma f_s} = f_1, f_2 \ldots f_s\right)$. Ce point sera le *point moyen*.

Effectivement la somme $\Sigma f_i \overline{GP_i}$ est nulle.

De même $\Sigma f_i \overline{MP_i^2}$ est minimum pour M = G.

3° Pour trouver la médiane, on cherche la modalité P_i telle que $\Sigma f_i \mid P_i P_j \mid$ soit minimum.

Or par hypothèse tous les $\mid P_i P_j \mid$ sont égaux, sauf $\mid P_i P_i \mid = 0$. La somme sera donc minimum si f_i *est maximum*. La médiane serait donc la modalité de fréquence maximum; d'ailleurs $\Sigma f_i \overline{P_i P_j}^2$ également étant minimum, cette modalité pourrait aussi bien être tenue pour moyenne, ce qui montre tout ce que ces considérations ont d'artificiel.

4° On calcule l'écart moyen simple ou quadratique :

$$ {}^1S = 1 - \sum_1^s f_i^2 \qquad ({}^2S^2) = \frac{{}^1S}{2}. $$

(Pour 2 modalités, on a : $f_1 = p$, $f_2 = q$; $({}^2S^2) = pq$, formule bien connue).

On retrouve ${}^1S$ quand on calcule la différence moyenne avec répétition Δ_R.

5° Des indices ont été calculés par Cisbani (1938 en utilisant le barycentre comme une moyenne (indice d'inégalité, de variabilité, écart moyen, écart quadratique moyen).

6° Gini a calculé des indices de dissemblance.

7° Pietra a calculé des indices de connexion et de concordance (Metron, IV, 3 4). Par exemple, il trouve une connexion de 0,0056 entre l'âge du père et le sexe de l'enfant, de 0,0042 entre l'âge de la mère et celui de l'enfant. On aime à penser que ceci pouvait s'observer à l'œil nu.

P. THIONET.

DISCUSSION

M. BARRIOL félicite M. Thionet de son très bel exposé qui a fait revivre pour les membres de l'Institut international de Statistique les discussions avec M. Gini pendant les dernières sessions.

M. BARRIOL cherche vainement l'intérêt et surtout la philosophie de toutes ces moyennes qui paraissent être des jeux d'esprit. Étant donnés trois nombres par exemple, on com-

prend bien la moyenne arithmétique que les enfants eux-mêmes se représentent dans leurs jeux de construction. La moyenne géométrique est déjà une spéculation difficile à justifier mais que dire de la moyenne $\sqrt[2]{\frac{ab + bc + ca}{3}}$. Quel est le raisonnement qui y conduit?

M. Gini et ses disciples ont l'air de définir des expressions algébriques de moyennes pour leurs besoins dans tel ou tel cas. Cela ne semble pas sérieux et on arrive aux résultats étranges indiqués par le conférencier, savoir des moyennes dont la valeur est en dehors des données servant à l'établir!...

M. Bertrand. — A la suite d'une intervention de M. Barriol sur l'inutilité de créer des moyennes fantaisistes ne correspondant pas à une nécessité, j'ai fait remarquer que j'utilisais régulièrement depuis plusieurs années dans les séries chronologiques, une moyenne progressive géométrique de raison $(1 + r)$, r étant déterminé empiriquement et variable suivant les cas. Le but d'une telle moyenne est de donner aux divers termes de la série un poids d'autant plus grand que le terme est plus récent.

La formule d'une telle moyenne est :

$$\frac{a + b\,(l + r) + c\,(l + r)^2 \ldots + m\,(l + r)^{n=1}}{l + (l + r) + (l + r)^2 \ldots + (l + r)^{n-1}}$$

elle présente l'avantage de pouvoir être prolongée.

J'ai également signalé avoir utilisé il y a plusieurs années une formule de calcul de la dispersion moyenne exposée par M. Thionet sous le symbole Δ, formule qui consiste à calculer les écarts de tous les termes pris deux à deux : $a\,b$, $a\,c$, $a\,d$, $b\,c$, $b\,d$, etc... Je l'ai abandonnée à cause de la longueur de l'opération lorsque le nombre des termes (ou le nombre de groupes dans lesquels on fait rentrer les termes) prend une certaine importance. Je n'y ai pas trouvé d'avantage compensant la longueur du calcul.

Le conférencier renvoie M. Bertrand à l'Annexe de son exposé (Annexe III) où seront exposées les méthodes de calcul de la différence moyenne lorsque le nombre des termes est grand.

Il pense que beaucoup de moyennes, soupçonnées d'être artificielles par M. Barriol, finissent par trouver leur intérêt (exemple de la corrélation $\frac{ab + bc + ca}{a^2 + b^2 + c^2}$ rencontrée dans les travaux de R. A. Fisher, qui ignore bien probablement Gini).

Pierre Thionet.