JOURNAL DE LA SOCIÉTÉ STATISTIQUE DE PARIS

ROBERT FÉRON

Mérites comparés des divers indices de corrélation

Journal de la société statistique de Paris, tome 88 (1947), p. 328-352
<http://www.numdam.org/item?id=JSFS_1947__88__328_0>

© Société de statistique de Paris, 1947, tous droits réservés.

L'accès aux archives de la revue « Journal de la société statistique de Paris » (http://publications-sfds.math.cnrs.fr/index.php/J-SFdS) implique l'accord avec les conditions générales d'utilisation (http://www.numdam.org/conditions). Toute utilisation commerciale ou impression systématique est constitutive d'une infraction pénale. Toute copie ou impression de ce fichier doit contenir la présente mention de copyright.

NUMDAM
Article numérisé dans le cadre du programme
Numérisation de documents anciens mathématiques
http://www.numdam.org/

II

MÉRITES COMPARÉS DES DIVERS INDICES DE CORRÉLATION [1]

AUTEURS UTILISÉS

De Finetti (B.) : Sui metodi proposti per il calcolo della differenza media (*Metron* vol. VIII, n° 1, 1931).

De Finetti (B.) et Paciello (U.) : Calcolo della differenza media (*Metron*, vol. VIII, n° 3, 1930).

Fréchet (M.) : Deux conférences faites au séminaire de calcul des probabilités, le 9 novembre 1945 et le 24 mai 1946 :

Sur le coefficient dit de corrélation (*C. R.*, 1933, p. 107 et 268-269).

Sur le coefficient de corrélation et sur la corrélation en général (*Rev. Inst. Int. Stat.*, 1933, P. 1, p. 1-8).

Gini (C.) : Di una misura, della dissomiglianza tra due gruppi, di quantita e della sue applicazioni allo studio delle relazioni statistiche (Atti del reale istituto veneto di scienze lettere ed arti) (*Année académique*, 1914-1915, t. LXXIV, 2e partie, p. 185-213).

Guidotti : Qualche consideratione sulla differenza media (*Atti*, VII, riunione scientifica, p. 372-385, 1942-1943).

Jordan (C.) : Critique de la corrélation au point de vue des probabilités. Colloque consacré à la théorie des probabilités, VIIe partie, p. 15-33.

Risser (R.) : Les principes de la statistique mathématique, 2e partie, corrélation, covariation (*Journal de la Société de Statistique de Paris*, n° 10, octobre 1936 (1-41).

Risser (R.) et Traynard (C.) : Les principes de la statistique mathématique.

Thionet (P.) : L'école moderne de statisticiens italiens (*Journal Soc. Stat. de Paris*, janvier-février 1946).

A — Mode d'emploi des divers indices de corrélation

Étant donnée une population d'individus, nous mesurons deux caractères X et Y pour chacun des individus. Il peut arriver que ces caractères soient rigidement liés. Ainsi, dans une population de cercles la connaissance du rayon X de l'un d'eux permettra de déterminer sans aucune ambiguïté sa surface Y. Par contre, dans une population humaine, si nous considérons tous les individus ayant le même poids X, nous leur trouverons des tailles différentes Y. Toutefois, de la connaissance du poids, nous pouvons déduire certains renseignements sur la taille; nous exprimerons ce fait en disant que les variables aléatoires X et Y sont liées.

D'une manière plus précise, si la probabilité que $Y = y_j$ correspondant à une valeur donnée $X = x_i$ n'est pas égale à la probabilité de Y_j dans la totalité des expériences nous dirons que x_i et y_j sont liées. Si, au contraire :

$$Pr\,(Y = y_j \mid x_i) = Pr\,(Y = y_j) \qquad (1)$$

X et Y seront dites indépendantes pour les valeurs x_j, y_i. Si les variables aléatoires X et Y sont indépendantes pour toutes les valeurs x_j et y_i, elles seront dites indépendantes.

Nous ne considérerons ici que les indices de corrélation empiriques, c'est-à-dire les indices que nous pouvons déduire du cas pratique où les variables aléatoires X et Y sont susceptibles d'un nombre fini de valeurs $x_1, x_2, \ldots x_n$, $y_1, y_2, \ldots y_m$ et où nous déduisons l'indice du nombre de fois n_{ij} que les caractères x_i, y_j ont été observés. Soit N_i le nombre de fois qu'on observe $X = x_i$, N_j le nombre de fois qu'on observe $Y = y_j$, N le nombre total d'observations.

(1) Communication présentée à la séance du 15 janvier 1947.

On pourra ainsi, par exemple, calculer divers indices de corrélation entre le nombre d'aurores boréales observées une année X déterminée, en un lieu déterminé et le nombre de jours de pluie (Y), entre la variété d'un poirier X et le nombre de poires récoltées Y.

Nous dirons alors que Y est fonction univalente de X si $x = x_i$ entraîne que $Y = y_j$, et nous dirons que nous avons indépendance si, quels que soient i et j:

$$\frac{n_{ij}}{N_i} = \frac{N'_j}{N} \qquad (2)$$

Dès lors, il est logique, comme l'a dit M. Fréchet dès 1934, d'imposer à tout bon indice de corrélation I d'être compris entre 0 et 1 et de vérifier les quatre conditions suivantes :

X	x_1	x_i	x_n	TOTAL
y_1	n_{11}	n_{i1}	n_{n1}	N'_1
y_j	n_{1j}	n_{ij}	n_{nj}	N'_j
y_m	n_{1m}	n_{im}	N_{nm}	N'_m
TOTAL . . .	N_1	N_i	N_n	N

1. Si Y est fonction univalente de X, I = 1.
2. Réciproquement, si I = 1, Y est fonction univalente de X.
3. Si X et Y sont indépendantes, I = 0.
4. Si I = 0, X et Y sont indépendantes.

Examinons quels sont les indices de corrélation vérifiant les quatre conditions.

II — Indices connus vérifiant les quatre conditions

Gini, en 1914, Charles Jordan en 1938, et M Fréchet en 1946 ont successivement donné des formules d'indices vérifiant les quatre conditions, mais si Gini et Jordan ont vraisemblablement eu l'intuition que leur indice vérifiait les quatre conditions, ni l'un ni l'autre ne l'ont démontré complètement, et aucun des deux n'a formule nettement les conditions que doit vérifier un bon indice.

Ce n'est que le 9 novembre 1945 que M. Fréchet a annoncé que l'indice de connexion de Gini vérifiait les quatre conditions et le 24 mai 1946, il a montré que l'on peut construire une infinité d'indices vérifiant les quatre conditions fondamentales. Nous allons montrer que les trois indices mentionnés ci-dessus sont d'un calcul rapide et qu'ils doivent, par conséquent, être employés dans la pratique. Il me semble personnellement que tous trois doivent susbister parce qu'ils répondent à des besoins légèrement différents, mais laissons au temps le soin de décider.

1. *L'indice de connexion de Gini.*

La définition théorique de cet indice est peut-être un peu compliquée, mais son calcul, d'abord très long, a été rendu très simple par de Finetti et Salvi-

mini; il est même plus rapide que celui du coefficient de corrélation linéaire, même dans le cas où l'on dispose de moyens très perfectionnés (machines à calculer électriques pour le coefficient de corrélation linéaire, hélice à calcul pour l'indice de Gini). Nous donnerons donc en annexe la définition théorique de l'indice de Gini et la démonstration du fait qu'il vérifie les quatre conditions et prendrons provisoirement pour définition de l'indice de connexion de Gini G, la formule suivante :

$$G = \frac{N \sum_i \sum_j (y_{i+1} - y_i) \, | s_{ij} - N_i S'_j |}{2 \sum_j (y_{i+1} - y_i) \, NS'_j \, (N - NS'_j)} \qquad (3)$$

avec

$$s_{ij} = \sum_{l=1}^{l=j} n_{il} \quad , \quad S'_j = \sum_{l=1}^{l=j} \frac{N'_l}{N} \qquad N S'_j = \sum_{l=1}^{j} N'_l .$$

On notera la simplification supplémentaire qui intervient dans le cas où les intervalles $y_{j+1} - y_j$ sont égaux entre eux. Dans ce cas, G prend la forme :

$$G = \frac{N \sum_i \sum_j s_{ij} - N_i S'_j}{2 \sum_j N S'_j \, (N - N S'_j)}$$

Application numérique de cette méthode.

Connexion entre la distribution des périodes orbitales et la durée des éclipses :

Y \ X			DURÉE D'UNE ECLIPSE EN HEURES												TOTAL T
			2 1	4 2	6 3	8 4	10 5	12 6	14 7	16 8	18 9	20 10	22 11	24 12	
Période orbitale en heures.	10	0	2	6											8
	40	1		1	3	2	3	2							11
	70	2				3	4	2							9
	100	3					2		1	1					4
	130	4							1		1				2
	160	5												1	1
	190	6										1			1
	220	7								1		1			2
TOTAL			2	7	3	5	9	4	2	2	1	2	0	1	38

Puisque les périodes orbitales forment une progression arithmétique, on obtiendra encore le coefficient de Gini en leur substituant (au numérateur comme au dénominateur) les nombres 0, 1, 2, 3....

Faisons la somme cumulée des fréquences de chaque colonne dans la table. (Nous écrivons ainsi que l'indice de dissemblance est la moyenne des segments compris entre la fonction de répartition *a priori* et la fonction de répartition liée à la variable Y.)

Remarquons qu'une simple hélice à calcul munie de repères multiples permet d'effectuer en bloc la multiplication de la deuxième colonne par n'importe quel nombre. On pourra donc transcrire à vue sans opération manuelle, sans effort pour accoupler deux nombres tous les résultats d'une colonne avec trois ou

quatre chiffres significatifs. Le calcul apparaît ainsi comme très faisable en 35 minutes-temps mentionné par Salvimini.

Calcul du numérateur.

s_{1j}	$2S_j$	DIFF	s_{2j}	$7S'_j$	DIFF	s_{3j}	$3S'_j$	DIFF	s_{4j}	$5S'_j$	DIFF	s_{5j}	$9S'_j$	DIFF	s_{6j}	$4S'_j$	DIFF.
2	0,42	1,58	6	1,48	4,52	0	0,63	0,63	0	1,05	1,05	0	1,90	1,90	0	0,84	0,84
2	1	1	7	3,50	3,50	3	1,50	1,50	2	2,50	0,50	3	4,50	1,50	2	2	0
2	1,47	0,53	7	5,16	1,84	3	2,21	0,79	5	3,69	1,31	7	6,63	0,37	4	2,95	1,05
2	1,68	0,32	7	5,90	1,10	3	2,53	0,47	5	,22	0,78	9	7,58	1,42	4	3,37	0,63
2	1,79	0,21	7	6,27	0,73	3	2,68	0,32	5	4,48	0 52	9	8,05	0,95	4	3,58	0,42
2	1,84	0,16	7	6,45	9,55	3	2,76	0,24	5	4,61	0,39	9	8,29	0,71	4	3,69	0,31
2	1,89	0,11	7	6,63	0,37	3	2,84	0,16	5	4 74	0,26	9	8,53	0,47	4	3,80	0,20
$21_1 = 3,91$			$71 = 12,61$			$31_3 = 4,11$			$51_4 = 4,81$			$91_5 = 7,32$			$41_6 = 3,45$		

s_{7j}	$2S'_j$	DIFF	s_{8j}	$2S'_j$	DIFF	s_{9j}	S'_j	DIFF	s_{10j}	$2S_j$	DIFF				s_{11j}	S_j	DIFF
0	0,42	0,42	0	0,42	0,42	0	0,21	0,21	0	0,42	0,42				9	0,21	0,21
0	1	1	0	1	1	0	0,50	0,50	0	1	1				0	0,50	0,50
0	1,47	1,47	0	1,47	1,47	0	0,74	0,74	0	1,47	1,47				0	0,74	0,74
1	1,68	0,68	1	1,68	0,68	0	0,84	0,84	0	1,68	1,68				0	0,84	0,84
2	1,79	0,21	1	1,79	0,79	1	0,89	0,11	0	1,79	1,79				0	0,89	0,89
2	1,84	0,16	1	1,84	0,84	1	0,92	0 08	0	1,84	1,84				1	0,92	0,08
2	1,89	0,11	1	1,89	0,89	1	0,95	0,05	1	1,89	0,89				1	0,95	0,05
$21_7 = 4,05$			$21_8 = 6,09$			$1_9 = 2,53$			$21_{10} = 9,09$						$1_{11} = 3,31$		

Calcul du denominateur (Méthode de de Finetti Paciello).

On doit ici supposer comme au numérateur que $y_{j+1} - y_j = 1$ et non pas 30. On a alors :

NS'_j	$N - NS'_j$	$NS'_j\ (N - NS_j)$
8	30	240
19	19	361
28	10	280
32	6	192
34	4	136
35	3	105
36	2	72
TOTAL		1 386

Le dénominateur est donc :

$$D_R = 1.386 \times 2 = 2.772$$

$$\Sigma N_i I_i = 3,91 + 12,61 + 4,11 + 4,81 + 7,32 + 3,45 + 4,05 + 6,09 + 2,53 + 9,09 + 3,31 = 61,28.$$

$$G = \frac{61,28 \times 38}{2.772} = 0,840.$$

2. *L'indice de Jordan.*

Cet indice est défini par :

$$J^2 = \frac{1}{m-1}\left[1 + \sum_i \sum_j \frac{(n_{ij})^2}{N_i\ N'_j}\right]. \qquad (4)$$

Son calcul est extrêmement rapide. Dans le tableau suivant, qui donne les $\frac{n^2_{ij}}{N_i\ N'_j}$, il suffit d'effectuer les additions :

TABLEAU

Y \ X	2	4	6	8	10	12	14	16	18	20	22	24	T
10	0,25	0,643											0,893
40		0,013	0,272	0,072	0,090	0,090							0,537
70				0,200	0,197	0,111							0,508
100					0,111		0,125	0,125					0,361
130							0,250		0,5				0,750
160												1	1
190										0,5			0,500
220								0,250		0,25			0,500
	0,25	0,656	0,272	0,272	0,398	0,201	0,375	0,375	0,5	0,75		1	5,049

$$J^2 = \frac{5,049 - 1}{8 - 1} = 0,578 \qquad J = 0,75$$

La valeur de cet indice reste inchangee non seulement quand nous remplaçons x par une fonction de x (Cf. Indice de Gini), mais encore quand on remplace y par une fonction croissante de y. Pour cette raison, quoique cet indice soit valable dans le cas general, il s'impose particulierement quand X et Y sont toutes deux des variables qualitatives (Ex. : corrélation entre la couleur des cheveux et celle des yeux).

3. *L'indice diagonal de M. Fréchet.*

Cet indice est égal à

$$d = \frac{\sum_i N_i \lambda_i}{\sum_i N_i (\lambda_i + \mu_i)}$$

ou λ_i et μ_i sont définis graphiquement de la manière suivante : considérons les courbes en escalier $s_i(y)$, $S'(y)$ obtenues en faisant correspondre à tout nombre y_j les valeurs

$$s_{ij} = \sum_{l=1}^{j} n_{il} \quad \text{et} \quad S'_j = \sum_{l=1}^{j} N'_l.$$

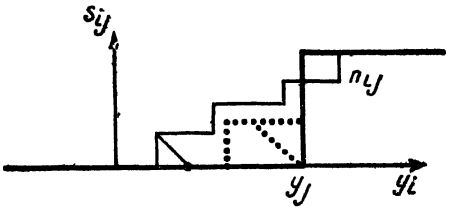

Fig. 1

λ_i est le maximum du segment découpé par ces deux courbes sur des parallèles à la seconde bissectrice. Soit y_{im} la plus petite médiane empirique de Y_{x_i} — c'est-à-dire le nombre y_{im} correspondant au plus petit s_{ij} supérieur à $\frac{1}{2}$. μ_i sera le plus grand des segments découpés sur la deuxième bissectrice par la courbe $s_{ij}(y)$ et la courbe $f_i(y)$ telle que $f_i(y) = 0$ si $y < y_{im}$ $f_i(y) = 1$ pour $y > y_m$.

Mode opératoire. — On tracera les courbes $s_{ij}(y)$ et $f_i(y)$ sur un papier millimétrique et la courbe $S'_j(y)$ sur un calque et on lira la mesure de la

projection des distances indiquées sur l'un des axes de coordonnées l'exemple précédent donne :

$$\Sigma N_i \lambda_i = \sqrt{2}\,[2 \times 0{,}60 + 7 \times 0{,}60 + 3 \times 0{,}50 + 5 \times 0{,}26 + 9 \times 0{,}21 + 4 \times 0{,}30 + 2 \times 0{,}60 + 1 \times 0{,}74 + 2 \times 0{,}89 + 0 \times \ldots + 1 \times 0{,}84] = 17{,}05\sqrt{2}$$

$$\Sigma N_i \mu_i = \sqrt{2}\,[2 \times 0 + 7 \times 0{,}14 + 3 \times 0 + 5 \times 0{,}40 + 9 \times 0{,}33 + 4 \times 0{,}50 + 2 \times 0{,}50 + 2 \times 0{,}50 + 1 \times 0 + 2 \times 0{,}50 + 0 \times \ldots + 1 \times 0] = 10{,}95\sqrt{2}.$$

$$d_{v_x} = \frac{17{,}05}{17{,}05 + 10{,}95} = 0{,}60.$$

Cet indice est à peu près trois fois plus long à calculer que l'indice de Jordan et deux fois plus long que l'indice de Gini, mais il présente sur ce dernier l'avantage d'être beaucoup moins sensible aux perturbations apportées par l'observation aléatoire d'une grande valeur de Y.

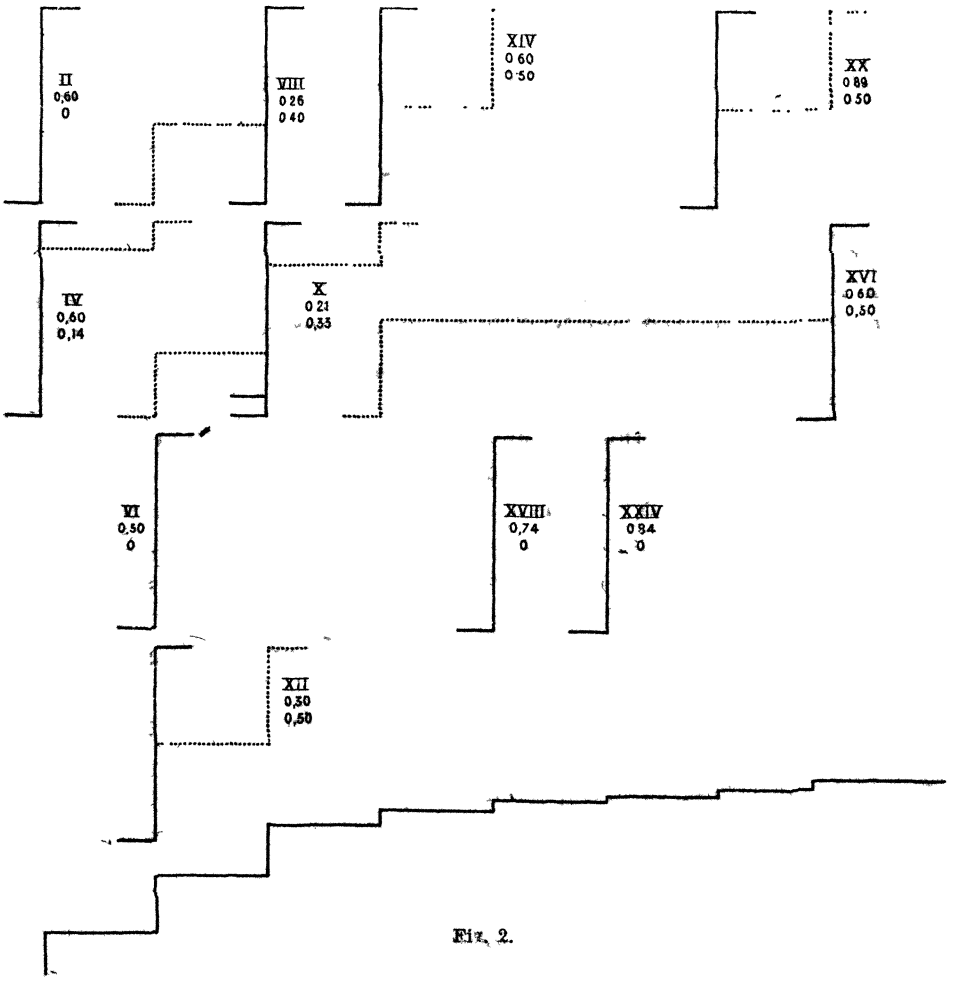

Fig. 2.

Sensibilité.

M. Fréchet se proposant de construire un grand nombre de variables aléatoires dont le choix ne soit pas subjectif, partit de l'idée qu'une ligne aléatoire convenablement choisie pourrait lui donner une infinité de telles variables

aléatoires. Il choisit comme ligne aléatoire la forme du crâne humain, qui s'exprime en coordonnées polaires par une équation de la forme $\rho = f(\omega)$.

Les coefficients du développement en série de Fourier sont des v. a. qui sont très liées. M. Fréchet a fait effectuer le calcul pour le coefficient de corrélation linéaire, l'indice de Gini et l'indice diagonal. On voit que ces trois indices varient grossièrement dans le même sens et que l'amplitude de variation de l'indice diagonal est plus grande que celle des deux autres. Peut-être serait-il préférable d'employer l'indice diagonal dans le cas où les v. a. X et Y sont très liées et l'indice de Gini dans celui où elles le sont très peu?

I — Indices ne vérifiant pas les quatre conditions

1. *Coefficient de corrélation linéaire.*

$$r = \frac{\sum_i \sum_j n_{ij} (x_i - \overline{X})(y_j - \overline{Y})}{\sqrt{\sum_i N_i (x_i - \overline{X})^2 \times \sum_j N'_j (y_j - \overline{Y})^2}} \tag{3}$$

avec

$$\overline{X} = \frac{\sum_i N_i x_i}{N} \qquad \overline{Y} = \frac{\sum_j N'_j Y_j}{N} \tag{4}$$

1° Si Y est fonction univalente de X, r peut avoir une valeur quelconque. Il peut même être nul si la courbe $y(x)$ est symétrique par rapport à la droite. $y = \overline{Y}$. Ce n'est que dans le cas très particulier où Y est fonction linéaire de X que $r = 1$.

2° Si $r = 1$, Y est fonction univalente de X (et même fonction linéaire.)

3° Si les v. a. X et Y sont indépendantes $r = 0$.

4° Si $r = 0$, les v. a. X et Y ne sont pas en général indépendantes.

D'une manière plus précise, il faut et il suffit pour que $r = 0$ que l'ensemble des points $z = n_{ij}$, $x = x_i$, $y = y_j$ de l'espace à trois dimensions soit symétrique par rapport à l'un ou l'autre des plans $x = \overline{X}$ ou $y = \overline{Y}$.

Le coefficient de corrélation linéaire apparaît donc plutôt comme un indice de dissymétrie que comme un indice de corrélation.

Enfin M. Fréchet a signalé un très grave défaut du coefficient de corrélation linéaire. Son manque de précision. En effet, si pour un seul point x_i et y_j sont très grands, cette observation, qui aurait très bien pu ne pas avoir été faite, fera varier r dans des proportions considérables, car pour un seul $x_i = \infty$, $y_j = \infty$, r prend la forme indéterminée $\frac{\infty}{\infty}$.

2. *Rapport de corrélation de Pearson.*

Son carré est défini par ;

$$\eta^2 = \frac{\sigma_{\overline{Y}_x}^{\ 2}}{\sigma_{\overline{Y}}^2} = \frac{\sum_i N_i \sum_j \left[\frac{n_{ij} y_j}{N_i} - \overline{Y}\right]^2}{\sum_j N'_j (y_j - \overline{Y})^2} \tag{8}$$

On peut d'ailleurs l'écrire également :

$$\eta^2 = 1 - \frac{\mathfrak{M}\,\sigma^2_{\eta_x}}{\sigma_y^2} \qquad 9)$$

De (6) on déduit que :

1. Si Y est fonction univalente de X, $\eta^2 = 1$;
2. Si $\eta^2 = 1$, Y est fonction univalente de X;

De (5), on déduit que :

3. Si X et Y sont indépendantes, $\eta = 0$;
4. Mais si $\eta = 0$, on ne peut affirmer que X et Y sont indépendantes, on peut seulement dire que $\overline{Y}_x$ garde une valeur constante.

On peut écrire :

$$r^2 = \eta^2 \times \frac{\sum_i \sum_j n_{ij} (x_i - \overline{X})(y_j - \overline{Y})}{\sqrt{\sum_i N_i (x_i - \overline{X})^2 \times \sum_i N_i \sum_j \left[\frac{n_{ij} y'_j}{N_i} - Y\right]^2}}$$

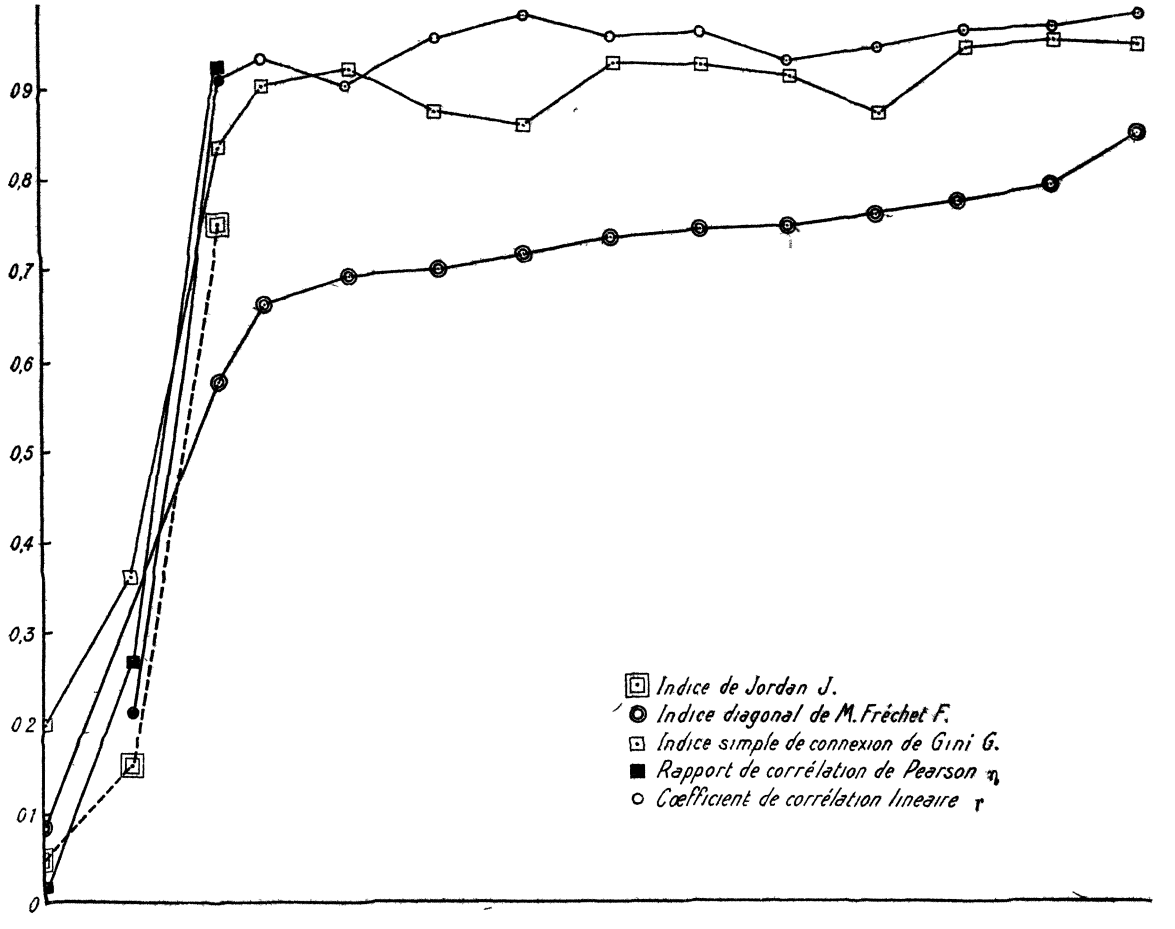

Fig. 3.

Donc r est égal au produit de η^2 par le coefficient de corrélation linéaire entre X et $\overline{Y}_x$. Ce résultat déjà obtenu par M. Fréchet montre que η est toujours préférable à r.

3. *L'indice I' de M. Fréchet.*

M. Fréchet a proposé comme indice, en 1919, et en attendant mieux :

$$I' = \frac{\sum_i N_i \{ | \mu_{x_i} - \mu | + | p_{x_i} - p | \}}{\sum_i N_i \{ | \mu_{x_i} + \mu | + | p_{x_i} - p | \}} \tag{10}$$

p_{x_i} étant la valeur équiprobable (dite aussi valeur probable) et μ_{x_i} étant l'écart probable, quand x est donné égal à x_i. η et μ les quantités correspondantes, quand x n'est pas donné.

1 et 2. On voit que $0 \leqslant I' \leqslant 1$, que pour $I' = 1$, $\mu_{x_i} = 0$, quel que soit x, donc Y est fonction univalente de X et réciproquement si Y est fonction univalente de X, $I' = 1$.

3. Si on a indépendance $I' = 0$.

4. Si $I' = 0$, on ne peut affirmer que l'on a indépendance, mais seulement que $\mu_{x_i} = \mu$ et $p_{x_i} = p$.

Enfin, dans les distributions théoriques, cet indice n'est pas susceptible de se mettre sous la forme indéterminée $\frac{\infty}{\infty}$. Il est donc bien préférable au rapport de corrélation de Pearson.

4. *Carré moyen de contingence.*

Le premier coefficient trouvé pour caractériser l'indépendance des variables aléatoires X et Y est le carré moyen de contingence

$$\varphi^2 = -1 + \sum_i \sum_j \frac{[n_{ij}]^2}{N_i N'_j} \tag{11}$$

Cet indice est un nombre positif inférieur à $\sqrt{(m-1)(n-1)}$ et tel que $\varphi = 0$ s'il y a indépendance et réciproquement[1]

5. *Coefficient du carré moyen de contingence.*

C'est la quantité

$$C = \sqrt{\frac{\varphi^2}{1+\varphi^2}}.$$

Cet indice est toujours inférieur à 1.

1. Mais si nous avons liaison fonctionnelle univalente, cet indice ne tend pas en général vers 1, même quand m et n tendent vers l'infini.

2. Si $C = 1$, nous avons liaison fonctionnelle univalente.

3. Si nous avons indépendance, $C = 0$.

4. Si $C = 0$, nous avons indépendance.

6. *Le coefficient de Tschuprow.*

C'est :

$$\tau^2 = \frac{\varphi^2}{\sqrt{(m-1)(n-1)}}$$

[1] Il est toujours inférieur à 1

Et en outre :

1. Si nous avons liaison fonctionnelle, univalente, cet indice n'est pas en général égal à 1 ; il n'en est ainsi que si $m = n$.
2. Si $\tau^2 = 1$, nous avons liaison fonctionnelle univalente entre X et Y.
3. Si nous avons indépendance $\tau^2 = 0$.
4. Si $\tau^2 = 0$, nous avons indépendance.

B — Définition théorique des indices de corrélation vérifiant les quatre conditions

Dans la précédente partie, nous avons dit quels sont les indices de corrélation actuellement connus qui vérifient les quatre conditions fondamentales et nous avons donné la formule qui, à notre avis, permet de les calculer le plus rapidement possible. Mais en partant de cette formule, la démonstration pratique du fait que ces indices vérifient les quatre conditions est souvent malaisée. Aussi avons-nous jugé utile de reproduire ici les raisonnements qui nous ont conduits aux formules ci-dessus.

I — L'indice de connexion de Gini

considérons le groupe i' formé des N N_i quantités suivantes :

Y \ X	x_1	x_2		x_i		x_n	T
y_1	n_{11}			n_{i1}		n_{n1}	N'_1
y_2	n_{12}						
y_j	n_{1j}			n_{ij}		n_{nj}	N'_j
y_m	n_{1m}			n_{im}		n_{nm}	N'_m
T	N_1			N_i		N_n	N

$N \times n_{i1}$ quantités égales à y_1
$N \times n_{i2}$ — y_2
$N \times n_{ij}$ — y_j
$N \times n_{im}$ — y_m

et le groupe T′ formé des quantités suivantes :

$N'_1 N_i$ quantités égales à y_1
$N'_j N_i$ — y_j
$N'_m N_i$ — y_m

Rangeons les N N_i quantités du groupe i' par ordre de grandeur croissante (ou au moins non décroissante).

Soit :

$$m_{i1}, m_{i2}, \ldots m_{ij} \ldots m_{i\,NN_i} \quad \text{ces quantités.}$$

De même, rangeons les N N_i quantités du groupe T′ par ordre de grandeur croissante.

Soit

$$m_{T1}, m_{T2}, \ldots m_{Ti} \qquad m_{T\,NN_i} \text{ ces quantités.}$$

(On aura : $m_{T1} = y_1, m_{T2} = y_1 \ldots\ldots m_{T\,NN_i} = y_m$).

Posons :

$$I_i = \frac{1}{NN_i} \sum_{l=1}^{NN_i} | m_{il} - m_{Ti} | \tag{14}$$

$$D_i = \frac{1}{N^2 N_i^2} \sum_{l=1}^{NN_i} \sum_{l'=1}^{NN_i} | m_{il} - m_{Ti'} | \tag{15}$$

L'indice de connexion de Gini sera donné par la formule

$$G = \frac{\Sigma N_i I_i}{\Sigma N_i D_i}. \tag{16}$$

1. *L'indice de dissemblance.*

Pour comprendre les raisons logiques qui ont poussé Gini à définir par cette formule son indice, il est indispensable de se familiariser avec une autre notion également introduite par Gini : celle d'indice de dissemblance entre deux groupes d'individus sur chacun desquels on mesure un certain caractère.

Supposons par exemple que nous voulions comparer la taille des Français à celle des Italiens.

Gini dira que la population française est *semblable* à la population italienne, si la proportion des individus qui ont une taille déterminée est la même en France et en Italie, et cela, quelle que soit cette taille.

S'il n'en est pas ainsi, les deux populations seront dites dissemblables.

Calcul de l'indice diagonal de M. Fréchet.

Ceci posé, il n'est pas difficile de caractériser la dissemblance qui existe entre la taille de ces deux populations, que nous supposons pour simplifier être toutes deux égales à 40 millions d'individus.

Il suffira de supposer les Français rangés par ordre de taille croissante, soit :

$a_1 \quad a_2 \ldots\ldots a_i \qquad a_{40\,000\,000}$ leurs tailles et de même les Italiens.

soit

$b_1 \quad b_2 \ldots\ldots b_i \qquad b_{40.000\,000}$ leurs tailles.

et de calculer la quantité

$$I_{AB} = \frac{| a_1 - b_1 | + | a_2 - b_2 | + \quad + | a_i - b_i | + \quad + | a_{40.000\,000} - b_{40\,000.000} |}{40.000.000}$$

Il est clair que cette quantité appelée indice de dissemblance entre les deux populations est nulle si les deux populations sont semblables.

Si nous voulions comparer la taille des 40 millions de Français et des 180 millions de Russes, on serait amené à considérer la population F′ de 360 millions d'habitants semblable à la population française et la population R′ de 360 millions d'habitants semblable à la population russe et à calculer l'indice de dissemblance, entre ces deux populations, autrement dit nous considérons une population fictive F′, dans laquelle 9 Français ont la taille a_i et une population fictive R′ dans laquelle deux Russes ont la taille b_i, et nous calculons l'indice de dissemblance entre ces deux populations.

Appelons :

$a'_1 \quad a'_2 \quad a'_j \quad a'_{180\,000.000}$ les tailles des individus de F'

et

$b'_1 \quad b'_2 \quad b'_j \quad b'_{180.000\,000}$ les tailles des individus de R'.

On aura :

$$I_{AR} = \frac{|a'_1 - b'_1| + |a'_2 - b'_2| + \quad + |a'_j - b'_j| + \quad + |a'_{180\,10^6} - b'_{180.10^6}|}{180.000.000}$$

Les quantités a'_j, b'_j, qui occupent le même rang dans deux populations F' et R', comprenant un même nombre d'individus, seront dites individus cogradués ou quantités cograduées.

L'indice de dissemblance entre deux populations A et B s'obtient donc en prenant la moyenne arithmétique des valeurs absolues des différences des quantités cograduées des populations A' et B' respectivement semblables à A et B.

On conçoit que sous cette forme le calcul de l'indice de dissemblance peut être très long, mais divers théorèmes énoncés par Gini permettent de simplifier considérablement ce calcul.

Lemme. Soit deux populations A' et B' chacune de n individus contenant t individus ayant la même taille, la somme des valeurs absolues des différences entre les quantités cograduées de ces populations est égale à la somme des valeurs absolues des différences entre les quantités cograduées de deux populations qui s'obtiennent en supprimant les t individus de même taille.

α) La proposition est évidente si les individus qui ont la même taille occupent le même rang dans les deux populations, c'est-à-dire si l'on a $a_K = b_K$. On supprime ainsi des différences nulles, ce qui n'a pas d'inconvénient.

β) Si les individus qui ont la même taille n'occupent pas le même rang dans les deux populations, c'est-à-dire si l'on a : $a_K = b_{K+x}$ dans la somme

$$s = |a_1 - b_1| + \quad + |a_{K-1} - b_{K-1}| + |a_K - b_K| + \quad + |a_{K+x} - b_{K+x}| + \quad + |a_n - b_n| -$$

toutes les différences $a_i - b_i$ pour lesquelles $K \leqslant L \leqslant K + x$ sont positives, on peut donc supprimer les valeurs absolues pour ces différences, et supprimer les deux quantités égales a_K et b_{K+x}. Il vient:

$$s = |a_1 - b_1| + \ldots + |a_{K-1} - b_{K-1}| - b_K + a_{K+1} - b_{K+x-1} + a_{K+x} + |a_{K+x+1} - b_{K+x+1}| + \ldots + |a_n - b_n|$$

Ou encore :

$$s = |a_1 - b_1| + \ldots + |a_{K-1} - b_{K-1}| + |a_{K+1} - b_K| + \\ + |a_{K+x} - b_{K+x-1}| + |a_{K+x+1} - b_{K+x+1}| + \ldots + |a_n - b_n|$$

On peut aussi se servir d'un des trois théorèmes suivants, qui sont à peu près évidents :

Pour obtenir la somme des valeurs absolues des différences entre les quantités cograduées de deux populations A et B, chacune de n individus, il suffit de faire la somme de toutes les tailles a'_i b''_i et de soustraire toutes les tailles a''_i b'_i où a'_i et b'_i désignent les quantités cograduées (ou tailles cograduées) pour lesquelles on a $a_i > b_i$ et où a''_i et b''_i désignent celles pour lesquelles on a $a_i < b_i$.

Corollaire 1. — Si le plus petit individu de la population A est plus grand que le plus grand individu de la population B, l'indice de dissemblance de ces deux populations est égal à la différence des moyennes arithmétiques des tailles des individus de chaque population.

Ce corollaire est vrai même si les deux populations n'ont pas le même nombre d'individus.

Corollaire 2. — Si pour toute taille y supérieure à y_0, les individus de taille y se rencontrent avec une fréquence supérieure dans la populationA et si pour toute taille y inférieure à y_0 les individus se rencontrent avec une fréquence supérieure (ou du moins non inférieure) dans B, l'indice de dissemblance est égal à la différence des moyennes artihmétiques des tailles de chaque population.

Pour démontrer ce théorème, il suffit de supprimer les individus de même taille, pour être ramené au cas précédent, si les deux populations A et B ont le même nombre d'individus. S'il n'en est pas ainsi, on remplace A et B par des populations A′ et B′ qui leur sont semblables et l'on démontre le théorème sur ces populations.

Corollaire 3. — L'indice de dissemblance est égal à la différence des moyennes artihmétiques des tailles des populations A et B, quand les histogrammes de ces tailles se coupent au plus en un point.

2. *Calcul du numérateur de l'indice de Gini.*

Ceci posé, on comprend aisément pourquoi Gini a adopté la formule 16 pour son indice de connexion :

$$G = \frac{\Sigma N_i I_i}{\Sigma N_i D_i}.$$

Dans cette formule, I_i est l'indice de dissemblance entre le $i^{ème}$ groupe partiel et le groupe total, ou, ce qui revient au même, entre un groupe i' semblable à i et un groupe T′ semblable à T, chacun de NN_i éléments.

Les quantités D_i donnent un ordre de grandeur de la dissemblance qu'on devrait attendre.

a) *Méthodes de Gini.*

Les méthodes de Gini sont au nombre de deux.

1° Dans la première, on applique la formule (14) en groupant les quantités qui ont la même valeur.

Calculons ainsi l'indice de Gini sur l'exemple que nous avions choisi (indice de connexion entre la durée d'une éclipse et la période orbitale). En vertu de la remarque précédemment faite, nous savons que nous aurons encore l'indice de Gini en substituant aux diverses périodes observées les nombres 0, 1, 2, 3.

Nous avons groupé ensemble les m_{ie} ayant la même valeur y_l et les m_{TK} ayant la même valeur y_K.

Les différences $|y - y_K|$ entre les modalités du caractère sont indiquées en face de la fréquence par laquelle cette différence doit être multipliée.

On a :

$$\Sigma N_i I_i = 3{,}91 + 12{,}61 + 4{,}11 + 4{,}81 + 7{,}32 + 3{,}45 + 4{,}05 + 6{,}09 + \\ + 9{,}09 + 2{,}53 + 3{,}31 = 61{,}28.$$

0,42	0
0,58	1
0,47	2
0,21	3
0,11	4
0,05	5
0,05	6
0,11	7
2	

$$I_1 = \frac{3,91}{2}$$

1,48	1,48 0	
2,02	2,02 1	
1,66	1,66 2	
0,74	0,74 3	
0,37	0,10 4	0,27 3
0,18		0,18 4
0,18		0,18 5
0,37		0,37 6
7	6	1

$$I_2 = \frac{12,61}{7}$$

0,63	1
0,87	0
0,71	1
0,32	2
0,16	3
0,08	4
0,08	5
0,16	6
3	

$$I_3 = \frac{4,11}{3}$$

1,65	1,05 1	
1,45	0,95 0	0,50 1
1,19		1,19 0
0,59		0,59 1
0,26		0,26 2
0,13		0,13 3
0,13		0,13 4
0,26		0,26 5
5		

$$I_4 = \frac{4,81}{5}$$

1,90	1,90 1		
2,60	1,10 0	1,50 1	
2,13		2,13 0	
0,95		0,37 1	0,58 0
0,47			0,47 1
0,24			0,24 2
0,24			0,24 3
0,47			0,47 4
9	3	4	2

$$I_5 = \frac{7,32}{9}$$

0,84	0,84 1	
1,16	1,16 0	
0,95		0,95 0
0,42		0,42 1
0,21		0,21 2
0,11		0,11 3
0,11		0,11 4
0,21		0,21 5
4	2	2

$$I_6 = \frac{3,45}{4}$$

0,42	0,42 3	
0,58	0,58 2	
0,47		0,47 2
0,21		0,21 1
0,11		0,11 0
0,05		0,05 1
0,05		0,05 2
0,11		0,11 3
2	1	1

$$I_7 = \frac{4,05}{2}$$

0,42	0,42 3	
0,58	0,58 2	
0,47		0,47 5
0,21		0,21 4
0,11		0,11 3
0,05		0,05 2
0,05		0,05 1
0,11		0,11 0
2	1	1

$$I_8 = \frac{6,09}{2}$$

0,21 4
0,29 3
0,24 2
0,10 1
0,05 0
0,03 1
0,03 2
0,05 3
1

$$I_9 = \frac{2,53}{1}$$

0,42	0,42 6	
0,58	0,21 5	
0,47		0,47 5
0,21		0,21 4
0,11		0,11 3
0,05		0,05 2
0,05		0,05 1
0,11		0,11 0
2	1	1

$$I_{10} = \frac{9,09}{2}$$

0,21 5
0,29 4
0,24 3
0,10 2
0,05 1
0,03 0
0,03 1
0,05 2
1

$$I_{12} = \frac{3,31}{1}$$

2° On peut aussi abréger le calcul en se servant d'un quelconque des théorèmes que nous venons d'établir.

Application numérique.

Soit à calculer l'indice de connexion entre le nombre de lobes de la corolle de la *linaria spuria* et la forme de cette corolle.

Corolle / Nb. de lobes	Non velue α	Velue β	T
2	0	1	1
3	4	2	6
4	240	43	283
5	60.250	810	61.060
6	169	52	221
7	7	2	9
8	0	1	1
Total. .	60.670	911	61.581

Nombre de lobes	B′	T′	B′₁	T₁′
2	61.581	911	60.670	0
3	123.162	5.466	117.696	0
4	2.647.983	257.813	2.390.170	0
5	49.880.610	55.625.660	0	5.745.050
6	3.202.212	201.331	3.000.881	0
7	123.162	8.199	114.963	0
8	61.581	911	60.670	0
Total. . .	56.100.291	56.100.291	5.745.050	5.745.050

Nous remplaçons les groupes β et T par les deux groupes semblables β′ et T′ qui ont le même nombre de quantités : 911 × 61.518 = 56.100.291. Puis,

dans β' et T' nous supprimons les quantités communes; nous obtenons ainsi les groupes β'_1 et T'_1.

Les différences entre les quantités cograduées sont alors rapidement obtenues :

$$\begin{array}{rcl} 60.670 \times |2-5| & = & 182.010 \\ 117.696 \times |3-5| & = & 235.392 \\ 2.390.170 \times |4-5| & = & 2.390.170 \\ 3.000.881 \times |6-5| & = & 3.000.881 \\ 114.963 \times |7-5| & = & 229.926 \\ 60.670 \times |8-5| & = & 182.010 \\ \hline \text{dont la somme est} & & 6.220.389 \end{array}$$

L'indice de dissemblance I_β est donc :

$$I_\beta = \frac{6.220.389}{56.100.291} = 0{,}1241.$$

Si nous opérions de même pour I_α, nous trouverions dans la colonne α_1' la même chose que dans la colonne T_1' et pour la colonne T'_2 (correspondante à T_1') la même chose que pour β'_1.

Donc :

$$I_\alpha = \frac{6.220.389}{60.670 \times 61.581} = 0{,}001.665.$$

On trouve pour l'indice de Gini :

$$G = \frac{\Sigma N_i I_i}{\Sigma N_i D_i} = \frac{0{,}003.476}{0{,}017.465} = 0{,}199.$$

c) *Méthode de Salvimini.*

Dans le groupe i' il y a N. n_{ij} quantités m_{il} égales à y_j; convenons d'appeler m_{ij} celle à laquelle nous attribuons le rang le plus petit; celle à laquelle nous attribuons le rang le plus grand sera alors $m_{ij' + N\, nij}$.

De même, dans le groupe T', $N_i N'_j$ quantités sont égales à y_j; nous convenons d'appeler $m_{Tj''}$ celle à laquelle nous attribuons le rang le plus petit $m_{Tj'' + N_i N_j}$, celle dont le rang est le plus grand.

Posons :

$$s_{ij} = \frac{1}{N_i} \sum_{l=1}^{j-1} n_{il} \qquad\qquad S'_j = \frac{1}{N} \sum_{l=1}^{j-1} N'_l$$

Si $S_j > s_{ij}$, on a $j'' > j'$; donc $m_{ij}' > m_{Tj''}$.

Or d'après le corollaire (1) de Gini, l'indice de dissemblance est égal à :

$$I_i = \frac{1}{NN_i} \left| (\Sigma\, m'_{iK} - \Sigma\, m'_{TK}) - (\Sigma\, m''_{iK} - \Sigma m''_{TK}) \right|$$

où m'_{iK}, m'_{TK} désignent les quantités pour lesquelles $m_{iK} > m_{TK}$ et où m''_{iK}, m''_{TK} désignent celles pour lesquelles $m_{iK} < m_{TK}$.

La méthode de Salvimini consiste à voir quelle est la contribution dans cette somme des termes m_{il} et n_{Tl} qui sont égaux à y_K.

a) Si $S_j > s_{ij}$, $S_{j+1} > s_{ij+1}$, la contribution apportée par les termes considérés à l'indice de dissemblance est :

$$\alpha_i = [s_{ij+1} - s_{ij} - (S'_j - S_{j-1})]\, y_K$$

ou encore :

$$\alpha_j = [-(S_{j+1} - s_{i,j+1}) + (S'_i - s_{ij})]\, y_K$$

b) Si $S_j < s_{ij}$ $S_{j+1} < s_{i,j+1}$,

on trouve

$$\alpha_j = [-(s_{i,j+1} - S_{j+1}) + (s_{ij} - S_j)]\, y_j$$

c) $S_{j+1} < s_{i,j+1}$ $S_j > s_{ij}$,

on trouve :

$$\alpha_j = [-(s_{i,j+1} - S'_{j+1}) + S_j - s_{ij}]\, y_j$$

d) Si $S_{j+1} > s_{i,j+1}$ $S_j < s_{ij}$, on a :

$$\alpha_j = [-(S_{i+1} - s_{ij+1}) + s_{ij} - S'_j]\, y_j$$

En résumé, on a toujours :

$$\alpha_j = [-|s_{i,j+1} - S_{j+1}| + |s_{ij} - S'_j|]\, y_j$$

D'où :

$$I_i = \frac{1}{N N_i} \sum_{j=1}^{m} \alpha_j$$

$$= \frac{1}{N N_i} \sum_{j=1}^{m} |s_{ij} - S_j|\, (y_{i+1} - y_j)$$

Nous avons donné dans la précédente partie un exemple d'application de la méthode de Salvimini qui est, je crois, la plus rapide.

2 — Calcul du dénominateur

Pour calculer la quantité $\Sigma N_i D_i$, nous sommes amenés à faire les opérations suivantes :

$$D_i = \frac{1}{N^2 N_i^2} \sum_{j=1}^{m} N\, n_{ij} \sum_{k=1}^{N N_i} |y_j - m_{T'k}$$

d'où

$$\Sigma N_i D_i = \sum_i \frac{1}{N^2 N_i} \sum_{j=1}^{m} N\, n_{ij} \sum_{k=1}^{N N_i} |y_j - m_{T'k}|$$

Remarquons que

$$\sum_i n_{ij} = N'_j, \qquad \text{il vient :}$$

$$\Sigma N_i D_i = \sum_{j=1}^{m} N'_j \sum_{k=1}^{N N_i} \frac{|y_j - m_{T'k}|}{N N_i}$$

$$= \frac{1}{N N_i} \sum_{k=1}^{N N_i} \sum_{r=1}^{N} |m_{Tr} - m_{T'k}|$$

Où m_{Tr} désigne la r^e quantité du groupe T.

La quantité

$$\frac{1}{N} \Sigma' N_i D_i = \frac{1}{N^2} \sum_{k=1}^{N} \sum_{r=1}^{N} |m_{Tr} - m_{Tk}| \qquad (20)$$

est appelée différence moyenne (avec répétition) des quantités du groupe total. C'est la moyenne des valeurs absolues des différences obtenues en associant chaque quantité du groupe T avec toutes les quantités de ce même groupe.

Cette différence moyenne a été introduite en 1869 dans la statistique par l'astronome allemand W. Jordan (*Uber die Bestimmung der Genauigkeit mehrfach wiederholter Beobachtungen einer Unbekanten. Astronomische Nach-*

richten 74, 1766-1767). Elle est de nos jours employée d'une manière courante par l'école italienne qui a donné diverses manières de la calculer. Ces résultats ont été rassemblés par de Finetti (3) et récemment Thionnet (8) a reconstitué les diverses démonstrations. Nous signalerons ici les *principales méthodes employées* :

a) *Méthode de Gini.*

On a, si

$$i < a < N + 1 - i$$

$$| m_{Ti} - m_{Ta} | + | m_{Ta} - m_{T\,N+1-i} | = | m_{Ti} - m_{T\,N+1-i} |$$

En opérant de même, quelque soit i, on voit que :

$$\Delta_R = \frac{1}{N^2} \Sigma \mid m_{Ti} - m_{T\,N+1-i} \mid \cdot \mid N + 1 - 2i \mid \qquad (21)$$

Ce qui peut encore s'écrire, $\mathfrak{M}$ désignant une médiane :

$$\Delta_R = \frac{2}{N^2} \sum_{1}^{N} \mid N + 1 - 2i \mid \cdot \mid m_{Ti} - \mathfrak{M} \mid .$$

Application numérique. — Appliquons cette méthode à l'exemple des éclipses.

ÉLÉMENTS *correspondants*		DIFF.	DISTANCE graduelle $N + 1 - 2i$	PRODUIT
7	0	7	37	259
7	0	7	35	245
6	9	6	33	198
5	0	5	31	155
4	0	4	29	116
4	0	4	27	108
3	0	3	25	75
3	0	3	23	69
3	1	2	21	42
3	1	2	19	38
2	1	1	17	17
2	1	1	15	15
2	1	1	13	13
2	1	1	11	11
2	1	1	9	9
2	1	1	7	7
2	1	1	5	5
2	1	1	3	3
2	1	1	1	1
		TOTAL		1.386

$$\Delta_R = \frac{2 \times 1386}{38^2}.$$

b) *Méthode de De Finetti-Paciello.*

On écrit :

$$| m_{Tk} - m_{Ti} | = | m_{Ti+1} - m_{Ti} | + | m_{Ti+2} - m_{Ti+1} | + \ldots + | m_{Tk} - m_{Tk-1} |$$

et on trouve :

$$\Delta_R = \frac{2}{N^2} \sum_{i=1}^{N-1} i\,(N - i) \mid m_{Ti+1} - m_{Ti} \mid \qquad (22)$$

Cette méthode est, croyons-nous, la plus rapide. Nous en avons donné plus haut une application à l'exemple des éclipses, appliquons-le à l'exemple de la *linaria spuria* citée plus haut.

i	N — i	PRODUIT
1	61 580	61 580
7	61 574	431 018
290	61 291	17 774 390
61 350	231	14 271 850
61 571	10	615 710
61 580	1	62 580
		33 216 128

D'où ;

$$\Delta_R = \frac{2 \times 33.216.128}{(61.581)^2} = 0{,}0174.65.$$

c) *Methode de Gini-Czuber.*

Cette méthode s'obtient en séparant les termes positifs et négatifs dans la méthode de Gini (méthode *a*).

Posons :

$$S = \sum_{i=1}^{N} (N + 1 - 2i)\, m_{T_i} \qquad S' = \sum_{i=1}^{N} i\, m_{T_i}$$

Il vient :

$$\Delta_E = \frac{2}{N^2}(S' - S) \tag{23}$$

Ou, en tenant compte que $S' + S'' = N(N + 1)M$ (M désignant la moyenne arithmétique) :

$$\Delta_R = \frac{4\,S' - 2\,N(N+1)\,M}{N^2} = \frac{2\,N(N+1)\,M - 4\,S}{N^2}.$$

Application à l'exemple des eclipses.

	$n\,n'$	$\frac{n'(n'+1)}{2}$	$\frac{n\,(n+1)}{1}$	DIFF	PRODUIT
7	0-2	3	0	3	21
6	2-8			3	18
5	3-4			4	20
4	4-6	21	10	11	44
3	6-10	55	21	34	102
2	10-19	190	55	135	270
1	19-30	465	190	275	275
0	30-38	741	465	276	0
				S' =	750

	$n.n'$	$\frac{n'(n'+1)}{2}$	$\frac{n\,(n+1)}{2}$	DIFF	PROD ×
0	0 8	36	0	36	0
1	9-19	190	36	154	154
2	19-28	406	190	216	432
3	28 32	528	406	122	366
4	32 34	595	528	67	268
5	34 35	630	595	35	175
6	35 36	666	630	36	216
7	36 38	741	666	75	525
				S =	2 136

$$\Delta_R = \frac{2\,(2.136 - 750)}{38^2} = \frac{2.772}{38^2}$$

On peut aussi calculer NM :

$$NM = 11 \times 1 + 18 + 12 + 8 + 5 + 6 + 14 = 74$$

$$\Delta_R = 2\,\frac{(2 \times 2.136 - 74 \times 39)}{38^2} = \frac{2.272}{38^2}$$

L'INDICE DE GINI VÉRIFIE LES QUATRE CONDITIONS FONDAMENTALES

a) La démonstration de ce fait est basée sur le théorème de Gini qui suit :

« La somme des *n* valeurs absolues des differences qui s'obtiennent en accouplant chacune à chacune les *n* quantites du groupe *i'* avec les *n* quantités

du groupe T′ sera minima quand chaque quantité de i' est accouplée avec la quantité cograduée de T′.

En effet :

1° Si nous avons : $m_{ik} < m_{il}$ $\quad m_{Tk'} < m_{Tl'}$,
un calcul élémentaire montre que :

$$|m_{ik} - m_{Tk'}| + |m_{il} - m_{Tl'}| \leqslant |m_{ik} - m_{Tl'}| + |m_{Tk'} - m_{il}| \qquad (24)$$

D'une manière plus précise, on peut écrire que la différence entre le membre de droite et celui de gauche est égale à deux fois la partie commune aux segments m_{ik}, m_{il} et $m_{Tk'}$, $m_{Tl'}$.

2° Ceci posé, nous appelons disposition originelle celle que nous observons en accouplant les quantités du groupe i' avec celles du groupe T′ d'une manière quelconque.

Soit m_{Tl} la quantité qui, dans la disposition originelle, est accouplée avec m_{i1}
Soit m_{ip} la quantité qui, dans la disposition originelle, est accouplée avec m_{T1}

Si nous modifions les couples en accouplant m_{i1} avec sa cograduée m_{T1} et mettons m_{Tl} à la place qu'occupait m_{T1}, nous obtenons une disposition nouvelle que nous appellerons disposition réduite au premier degré.

Disposition originelle	.....m_{i1}..... .....m_{ip}..... .....m_{Tl}..... .....m_{T1}.....
Disposition réduite au premier degré	.....m_{i1}..... .....m_{ip}..... .....m_{T1}..... .....m_{Tl}.....

En accouplant de même m_{i2}, m_{T2}, on obtient la disposition réduite au second degré.

Il est clair que la disposition réduite au $NN_i^{ème}$ degré est celle dans laquelle toutes les quantités cograduées sont accouplées.

Appelons S_0 la somme des valeurs absolues des différences qui correspondent à la disposition originaire.

S_1 la somme de ces mêmes quantités dans la disposition réduite au premier degré.
S_n la somme de ces mêmes quantités dans la disposition réduite au n^{me} degré.

On a, d'après (24) :

$$(24) \qquad S_0 \geqslant S_1 \geqslant \qquad \geqslant S_p \geqslant \qquad \geqslant S_{NN_i} = I_i \qquad (25)$$

Donc le théorème est démontré.

b) Or D_i est la moyenne arithmétique de quantités S_l (obtenues en associant une fois et une seule, tout m_{ik} à tout $m_{Tk'}$), qui sont toutes supérieures ou au moins égales à I_i; donc $D_i \geqslant I_i$ et

$$G_a = \frac{\Sigma N_i I_i}{\Sigma N_i D_i} \leqslant 1.$$

1° *Si* Y *est fonction univalente de* X, G = 1.

En effet, dire que Y est fonction univalente de X, c'est dire que dans chaque groupe partiel i, Y n'est susceptible que d'une valeur y_j. Dès lors, toutes les quantités S_l sont égales à I_i et, par conséquent, aussi leur moyenne I_i.

$$D_i = I_i \qquad (25)$$

Et comme 25 a lieu quel que soit le groupe i considéré,

$$\Sigma\ N_i D_i = \Sigma\ N_i I_i \quad \text{donc} \quad G = 1.$$

2° *Si* G = 1, Y *est fonction univalente de* X.

En effet G = 1 entraîne $\Sigma\ N_i I_i = \Sigma\ N_i D_i$; ce qui ne peut se produire en vertu de $I_i \leqslant D_i$ que si quel que soit i $\quad D_i = I_i$ (26).

Mais on a :

$$D_i = \mathfrak{M}\ S'_i$$

Ou, d'après (25) :

$$S_l \geqslant I_i$$

On ne peut avoir $D_i = I_i$ que si, quelle que soit la disposition envisagée, on a :

$$I_i = S_l \tag{27}$$

Ceci posé, supposons que dans le $i^{\text{ème}}$ groupe partiel se trouvent au moins deux quantités différentes y_j et y_k; ces deux quantités se retrouvent certainement dans le groupe total; dès lors les deux permutations, où l'on a :

Groupe i.....	y_j..... y_k.....	et	..... y_j..... y_k.....
Groupe T.....	y_k..... y_i.....		..... y_j..... y_k.....

ont des sommes S_l et $S_{l'}$ différentes, ce qui est inconciliable avec (27).

Donc, à chaque valeur $X = x_i$ ne correspond qu'une seule valeur de $Y = y_j$, ce qui caractérise une liaison fonctionnelle.

3° *Si les variables aléatoires* X *et* Y *sont indépendantes* G = 0.

En effet, dans ce cas, tous les groupes partiels sont semblables au groupe total; donc tous les I_i sont nuls et, par suite également leur moyenne arithmétique I.

D'autre part, le dénominateur de l'indice de Gini ne sera nul, lui aussi, que si toutes les quantités du groupe total sont égales entre elles; on a en effet dans ce cas, à la fois indépendance et liaison fonctionnelle. Y = Cte quel que soit X et l'indice de Gini prend la forme indéterminée $\frac{0}{0}$.

4° *Si* C = 0, X *et* Y *sont indépendantes.*

Le dénominateur étant une somme d'un nombre fini de quantités finies, ne peut être infini. On doit donc avoir :

$$\Sigma\ N_i I_i = 0 \quad \text{ou} \quad I_i = 0,$$

quel que soit i, les I_i étant des quantités positives. Donc chaque groupe i est semblable au groupe T; donc X et Y sont indépendantes.

II — L'indice de Charles Jordan

Nous pouvons indifféremment écrire l'indice de Jordan (de Y en X) :

$$J^2 = \sum_i \sum_j \frac{[N\, n_{ij} - N_i\, N'_j]^2}{(m-1)\, N_i N_j' \times N^2} \tag{28}$$

Soit :

$$J^2 = \frac{1}{(m-1)} \left[-1 + \sum_i \sum_j \frac{n^2_{ij}}{N_i N'_j} \right]$$

La deuxième forme étant manifestement d'un calcul plus rapide, c'est cette dernière que nous avions préalablement indiquée.

I — L'INDICE DE JORDAN VÉRIFIE LES QUATRE CONDITIONS

Du fait que $\frac{n_{ij}}{N_j} \leqslant 1$ il résulte que :

$$\sum_i \sum_j \frac{n_{ij}}{N_j} \geqslant \sum_i \sum_j \frac{n^2_{ij}}{N_i N'_j} \tag{29}$$

Le premier membre est égal à m et le second à $(m-1)J^2 + 1$.

Donc :

$$m \geqslant (m-1)\,J^2 + 1 \quad \text{et} \quad J^2 \leqslant 1.$$

1° *Si* Y *est fonction univalente de* X, $J^2 = 1$.

En effet, nous avons alors $\frac{n_{ij}}{N_i} = 1$ pour une valeur j et une seule, quel que soit i fixé.

Donc :

$$\begin{aligned} \sum_i \sum_j \frac{n_{ij}}{N_j} &= \sum_i \sum_j \frac{n_{ij}}{N'_j} \frac{n_{ij}}{N_j} \\ m &= (m-1)\,J^2 + 1. \end{aligned} \tag{30}$$

Donc $J^2 = 1$.

2° *Si* $J^2 = 1$, Y *est fonction univalente de* X.

Cette démonstration manque dans Jordan, mais il est aisé de la rétablir :

Si $J^2 = 1$, chacun des deux membres de l'inégalité (29) est égal à m; donc on a l'égalité (30).

Or de $\frac{n_{ij}}{N_i} \leqslant 1$, résulte que (30) ne peut avoir lieu que si à tout $n_{ij} \neq 0$ correspond un $\frac{n_{ij}}{N_j} = 1$.

Mais $\sum_{j=1}^{m} n_{ij} = N_i$ nous ne pourrons donc avoir $\frac{n_{ij}}{N_i} = 1$ que si dans la $i^{\text{ème}}$ colonne il n'y a qu'un seul élément, autrement dit que si Y est fonction univalente de X.

3° *Si* X *et* Y *sont indépendantes* $J^2 = 0$.

En effet, on a alors $\frac{n_{ij}}{N_i} = \frac{N'_j}{N}$ quel que soient i et j.

Donc :

$$n_{ij}\,N - N_i\,N'_j = 0. \tag{31}$$

Et l'on voit sur la forme (28) que $J^2 = 0$.

4° *Si* $J^2 = 0$, X *et* Y *sont independantes.*

En effet, on voit sous la forme (28) que l'égalité (31) est vérifiée quelle que soient i et j; donc X et Y sont indépendantes.

Jordan a montré que dans le cas des tables à deux lignes et deux colonnes son indice est égal à r et à η.

Il a également fait remarquer que son indice ne peut être égal à l'unité que dans le cas où $m < n$. C'est en effet dans ce cas seulement que cet indice peut

être égal à l'unité. Aussi M. Fréchet a-t-il suggéré de prendre pour X la variable aléatoire qui possède le plus grand nombre de classes (c'est-à-dire de lire le tableau dans un sens tel que l'on ait toujours $m < n$).

DISTRIBUTION DE J^2

Dans la pratique, le statisticien ne trouvera jamais un indice de corrélation égal à zéro ou à un.

En effet :

a) Il est aisé de montrer que si dans une population pour laquelle les caractères X et Y sont indépendants, on prélève un grand échantillon, il y a une probabilité nulle pour que l'indice de corrélation soit nul.

b) D'autre part, si Y est rigoureusement fonction univalente de X, il y a une probabilité nulle pour que les classes aient été choisies de telle sorte que l'indice de corrélation soit égal à l'unité.

Il paraît donc indispensable d'établir des tests permettant d'affirmer que nos indices de corrélation sont compris entre certaines limites avec une certaine quasi-certitude.

Remarquons toutefois que la question que nous venons de soulever pose un problème de probabilité inverse sur la solution duquel les divers probabilistes ne sont pas d'accord.

Il sort du cadre de cet exposé de donner leurs positions respectives. Nous nous bornerons simplement ici à considérer un échantillon de N individus prélevé dans une population infinie (d'indice de Jordan J). Nous trouverons sur l'échantillon un indice de Jordan J'.

Il est aisé de calculer EJ'^2 et $\sigma J'^2$. Ceci résulte immédiatement d'un calcul fait dans Risser et Traynard (*Les principes de la statistique mathématique*, p. 193) pour la quantité :

$$\varphi'^2 = (m-1)\, J'^2.$$

Appelons :

p_{ij} la probabilité d'observer le couple $x_i\, y_j$
P_i la probabilité d'observer x_i
P'_j la probabilité d'observer y_j

M. Risser montre que :

$$E(\varphi'^2) = \varphi^2 + \frac{1}{N} \sum_i \sum_j \frac{p_{ij}(P_i - p_{ij})(P'_j - p_{ij})}{P_i^2 P'^2_j}$$
$$+ \frac{1}{N^2} \sum_i \sum_j \frac{p_{ij}(P_i - p_{ij})(P'_j - p_{ij})(2\, p_{ij} - P_i P'_j)}{P_i^3 P'^3_j} + \ldots\ldots$$
$$= \varphi^2 + \frac{a}{N} + \frac{b}{N^2} + \ldots\ldots$$

où a et b sont uniquement fonction des $P_i\, P'_j\, p_{ij}$.

D'où :

$$E(J'^2) = J^2 + \frac{a}{N(m-1)} + \frac{b}{(m-1)N^2} + \ldots\ldots$$

Cette formule est d'autant plus utile que N est plus grand.

On a toujours :

$$E(J'^2) \geqslant J^2; \text{ donc si } J^2 = 1, \ E\ J'^2 = 1.$$

résultat évident *a priori* puisque alors $J'^2 = 1$ presque certainement.

En cas d'indépendance des variables aléatoires, X et Y on trouve :

$$E(J'^2) = (n-1)\left[\frac{1}{N} + \frac{1}{N^2} + \ldots\ldots\right].$$

M. Risser a, de même, calculé l'écart type de φ'^2. Il trouve :

$$\sigma[\varphi'^2] = \frac{\alpha}{N} + \frac{\beta}{N^2} + \frac{\gamma}{N^3} + \ldots\ldots$$

Il donne l'expression de α et constate que $\alpha = 0$ si X et Y sont indépendantes. On a évidemment :

$$\sigma(J'^2) = \frac{\sigma(\varphi'^2)}{m-1}$$

Il est à remarquer que dans le cas du coefficient de corrélation linéaire, l'écart type est beaucoup plus grand. Quand les variables aléatoires X et Y sont indépendantes, il est en effet de l'ordre de $\frac{1}{\sqrt{N}}$ (il est même égal à $\frac{1}{\sqrt{N}}$ en première approximation dans le cas où X et Y obéissent à une loi de Bravais).

Robert Féron.

DISCUSSION

M. Fréchet. — Je signalerai d'abord que M. Féron a trop modestement omis de rappeler que c'est lui qui a reconnu, dans l'indice de Jordan, un véritable indice de corrélation et m'en a informé.

En ce qui concerne les quatre conditions que j'ai énoncées et auxquelles doit satisfaire un véritable indice de corrélation, il doit être entendu que ces conditions étaient déjà dans l'esprit de tous ceux qui avaient réfléchi sur ce sujet mais qu'elles n'étaient généralement ni énoncées explicitement ni démontrées complètement.

Il n'est pas très étonnant que l'indice J de Jordan publié peu de temps avant la guerre se soit peu fait connaître pendant cette période tragique. On pourrait au contraire s'étonner d'avoir vu le coefficient de connexion, G, de Gini tomber dans l'oubli ou l'indifférence pendant un si long intervalle de temps.

Ce fait s'explique sans doute, d'une part parce que sa définition primitive qui a l'avantage de mettre bien en lumière pourquoi il repère la corrélation — est assez longue à exposer; d'autre part, et plus encore parce que cette définition conduisait à des calculs très longs exigeant six ou sept fois plus de temps que celui de *r*. Mais la formule simplifiée qu'a rappelée M. Féron conduit au contraire à un calcul très rapide du même ordre que celui de *r*, peut être plus bref quand les intervalles de valeurs de 1 sont égales.

En ce qui concerne l'indice diagonal, *d*, il semble que pour un calculateur pourvu de machines le calcul de *r*, G ou J soit un peu plus court. L'indice diagonal, *d*, possède peut-être cependant une sensibilité plus grande comme je

l'ai signalé ailleurs (1). Il n'est d'ailleurs pas impossible qu'on puisse substituer plus tard au calcul graphique que j'ai proposé un calcul numérique de d, dirigé de façon à le rendre plus automatique et plus rapide.

En terminant, je félicite M. Féron pour la clarté qu'il a su maintenir dans le résumé où il a condensé l'essentiel de plusieurs travaux.

M. Roy. — 1° En écoutant l'exposé de M. Féron, complété par les explications de M. Fréchet, je n'ai pu m'empêcher d'établir une certaine analogie entre l'évolution des recherches concernant les indices de corrélation, et celle que l'on a observé pour la définition des indices de prix ou de quantités : recours à des procédés empiriques, nécessité de procéder à des recoupements, difficultés de faire appel à des principes d'ordre rationnel, etc... »

2° J'ai demandé ensuite la manière dont se comportaient les divers indices de corrélation pour un même matériel statistique.

M. Fréchet m'a répondu en invoquant l'exemple des retards dans l'influence des variations d'une grandeur sur celle d'une autre grandeur qui en dépend, et en insistant sur le fait que les écarts présentés par ces divers indices étaient peu significatifs, tandis que les courbes traduisant les variations de ces indices au voisinage du maximum permettaient, au contraire, de faire un choix entre ces indices.

3° J'ai indiqué, à titre d'exemple susceptible de permettre la comparaison des indices de corrélation, le retard entre les variations de l'indice des prix de gros et celles de l'indice des prix de détail. J'ai signalé à ce propos, les études faites par M. Bowley il y a près de trente ans.

M. Baticle a été vivement intéressé par la communication de M. Féron et par les explications complémentaires de M. Fréchet. Il souligne l'utilité d'un indice de corrélation répondant aux quatre conditions énoncées par M. Fréchet ainsi que celle du choix possible entre les indices de corrélation corrects pour l'application qu'on a en vue.

En particulier, l'étude de l'*auto-corrélation* d'un ensemble d'observations d'un même phénomène, mais faites à des époques différentes, peut présenter un grand intérêt, et lorsqu'on dispose d'un indice de corrélation qui est « sensible » au voisinage de zéro, on peut déterminer avec une certaine précision, l'intervalle de temps au bout duquel les séries statistiques correspondant au début et à la fin de l'intervalle deviennent pratiquement indépendantes.

M. Risser. — Dans sa très intéressante communication, M. Féron a bien voulu tout d'abord nous indiquer les sources bibliographiques auxquelles il a eu recours, et nous signaler tout particulièrement les travaux de M. Gini en 1914, de MM. de Finetti et Paciello en 1930, de Finetti en 1931, et enfin ceux de MM. Jordan et Fréchet sur la délicate question de la correlation. Rappelons à ce propos qu'en 1934, l'Institut international de Statistique ayant constaté que certains statisticiens employaient sans précautions probables le coefficient de corrélation, avait jugé utile de mettre à l'étude l'emploi de ce coefficient, et chargé M. Fréchet d'établir un rapport sur cette question de la corrélation.

Nous savons que de nombreux statisticiens utilisaient à cette époque, pour

(1) *Anciens et nouveaux indices de corrélation.* Leur application du calcul des retards économiques. *Econometrica*, vol. 15, janvier 1947.

représenter le *coefficient* de dépendance entre deux variables statistiques (x, y), le coefficient

$$r = \frac{\sum_i \sum_j n_{ij} (x_i - a)(y_j - b)}{\sqrt{\sum n_{i|} (x_i - a)^2} \sqrt{\sum n_{|j} (y_j - b)^2}},$$

où $n_{i/j}$ n'est autre que le nombre de fois que le couple $(x, y.)$ prend le couple de valeurs (x_i, y_j) avec $n_{i/} = \sum_j n_{i/j}$, $n_{/j} = \sum_i n_{i/j}$, et où a et b sont les valeurs moyennes de x_i et de y_j.

Les fausses interprétations observées jusqu'alors de ce coefficient, auraient été probablement évitées en modifiant l'appellation de r, et en le désignant avec M. Fréchet, coefficient de linéarité.

Or b_i étant le centre de gravité de la file i, on constate, comme l'a montré M. Fréchet, que l'on peut écrire :

$$r = \rho\, \eta, \text{ avec } \rho = \frac{\sum (x_i - a) \,.\, n_i \,.\, (b_i - b)}{\sqrt{\sum_i n_{i|} (x_i - a)^2} \sqrt{\sum_i n_{i|} (b_i - b)^2}}$$

et η n'étant autre que le rapport de corrélation de Pearson.

$$\frac{\sqrt{\sum_i n_{i|} (b_i - b)^2}}{\sqrt{\sum_j n_{|j} (y_j - b)^2}}.$$

On voit ainsi apparaître dans r un nombre ρ, qui est un facteur étranger à la dépendance de (x, y), et vient, en quelque sorte, fausser sa mesure.

M. Fréchet a le premier montré que tout bon indice de corrélation doit être compris entre 0 et 1, et vérifier les conditions suivantes :

1° Si y est fonction univalente de x, l'indice I est égal à 1 (en module);

2° Et réciproquement, si cet indice est égal à 1, y est fonction univalente de x;

3° Si x et y sont indépendants, $I = 0$;

4° Réciproquement si $I = 0$, x et y sont indépendants.

Si le *rapport* η *de corrélation* de Pearson, jouit bien des propriétés 1°, 2°, 3° indiquées ci dessus, on ne peut pas affirmer que x et y sont indépendantes, mais dire seulement que $\bar{y}_{x_i}$ garde une valeur constante. Or MM. Gini et Jordan ont les premiers donné des formules d'indices qui ne pouvaient être critiquées, sans montrer toutefois que ces indices vérifiaient les quatre conditions; c'est M. Fréchet qui a démontré d'une part que l'indice de M. Gini satisfaisait aux conditions exigées, et d'autre part que l'on pouvait construire toute une série d'indice répondant à la question.

M. Féron nous a fait remarquer que l'indice de M. Gini mérite une mention spéciale ainsi que ceux de MM. Jordan et Fréchet doivent tous subsister, du fait qu'ils répondent à des besoins un peu différents; il lui revient le mérite d'avoir apporté des modifications heureuses dans la présentation des calculs de M. Gini, et de faciliter grandement la tâche des calculateurs.

Pour ma part, je serais fort reconnaissant à M. Féron, s'il voulait bien à la suite de sa communication, montrer — grâce à un exemple — comment l'on procède au calcul des indices de corrélation de MM. Gini et Fréchet, en ayant soin de donner au lecteur et cela sans craindre d'entrer dans le détail des opérations — tous les renseignements utiles.