JOURNAL DE LA SOCIÉTÉ STATISTIQUE DE PARIS

LUCIEN AMY

Contribution à l'étude des relations entre variables aléatoires faiblement liées

Journal de la société statistique de Paris, tome 98 (1957), p. 161-178

<http://www.numdam.org/item?id=JSFS_1957__98__161_0>

JOURNAL

DE LA

SOCIÉTÉ DE STATISTIQUE DE PARIS

Nos 7-8-9 — JUILLET-AOUT-SEPTEMBRE 1957

I

CONTRIBUTION A L'ÉTUDE DES RELATIONS ENTRE VARIABLES ALÉATOIRES FAIBLEMENT LIÉES

Notre collègue, M. Delaporte, au cours de l'exposé si intéressant qu'il a présenté en 1955 à la Société de Statistique (1), nous a montré que les particularités des fréquences et des corrélations entre les divers caractères d'objets appartenant à une même série pouvaient s'interpréter comme des combinaisons linéaires de quelques facteurs essentiels. Les exemples venant illustrer son exposé étaient tirés d'études biologiques.

Ayant eu nous-même à rechercher l'existence de telles relations dans un domaine très différent, celui de l'inflammabilité des matériaux, nous avons constaté que ces méthodes y sont rarement utilisables. Bien que M. Delaporte ne l'ait pas spécifié, les méthodes qu'il a exposées supposent, en effet, que les fréquences suivent des lois de Laplace-Gauss à plusieurs variables ou peuvent s'y ramener. Or, les événements que nous avions à étudier s'en écartent profondément. Nous avons ainsi été conduit à utiliser d'autres méthodes moins connues et à en imaginer quelques-unes qui, croyons-nous, sont plus ou moins nouvelles. C'est l'exposé de ces techniques et leur critique en vue des applications pratiques que nous nous proposons d'exposer ici.

Considérons un groupe G d'objets ou d'événements E en nombre illimité. Soit A une première grandeur attachée à ces événements et prenant des valeurs x variables d'un événement à l'autre. On peut caractériser le groupe G par la densité de fréquence $\Delta(x)$ des événements E dont la grandeur A est mesurée par des nombres compris entre x et $x + dx$. Soit ensuite B une deuxième grandeur attachée à ce même groupe, y la mesure de cette grandeur, on peut également caractériser le groupe par la densité de fréquence $\Delta(y)$ définie

(1) DELAPORTE Pierre. *Recherche statistique de facteurs indépendants. Journal de la Société de Statistique de Paris*, 96, p. 162 à 174, 1946.

d'une manière analogue. On obtient une caractéristique beaucoup plus complète en considérant la densité de fréquence $\Delta\ (x, y)$ des événements telle que les grandeurs A et B soient simultanément mesurées par des nombres compris entre x et $x + dx$, y et $y + dy$ respectivement. On généraliserait facilement à un nombre quelconque de variables. La connaissance complète du groupe se ramène à celle de la fonction $\Delta\ (x, y, \ldots)$.

En pratique, le problème se présente d'une manière toute différente. On ne connaît que les valeurs, x_i, y_i, ...d'événements E_i en nombre n limité, et ce nombre étant rarement très grand la fonction Δ n'est pas connue. On peut quelquefois faire une hypothèse à son sujet et tester la validité de cette dernière, mais s'il s'agit d'une fonction purement empirique le problème général est pratiquement insoluble. On se contente alors d'étudier des relations entre les nombres x_i, y_i... En général, ces relations comportent des constantes. Si le nombre n était assez grand, ces constantes pourraient être calculées avec précision. Lorsque n reste faible ces constantes sont comprises entre des valeurs que l'on peut calculer pour des limites de confiance données.

Nous écarterons en général le cas où les valeurs prises par les variables sont complètement indépendantes, c'est-à-dire celui où l'on a :

$$\Delta\ (x, y, \ldots) = [\Delta_1\ (x)]\ [\Delta_2\ (y)]\ [\ldots]; \qquad (1)$$

nous écarterons également celui où il existe une ou plusieurs relations fonctionnelles entre les quantités x, y... Toutefois, il ne faut pas perdre de vue que ces relations constituent des cas limites particulièrement importants. Les solutions envisagées pour le cas général où il existe une interdépendance limitée entre les variables doivent donc garder un sens pour ces cas idéaux.

Enfin bien que ce travail présente un aspect assez général, il a un but essentiellement pratique aussi nous donnerons un certain nombre d'applications concrètes. Celles-ci sont extraites d'une étude sur l'inflammabilité des matériaux, étude qui, nous l'avons dit, est à l'origine de la plupart de ces considérations. Pour faciliter la compréhension de ces exemples nous donnerons quelques indications préliminaires sur la nature de ces recherches.

Leur but essentiel était de tester l'efficacité de produits ignifuges, c'est-à-dire susceptibles de rendre le bois ininflammable. Dans ce but, des éprouvettes de contreplaqué d'okoumé aussi semblables que possible étaient recouvertes de peinture ou d'enduits sur leurs deux faces ou trempées dans des solutions appropriées puis, après séchage, soumises à un essai d'inflammabilité suivant le test du Ministère de l'Intérieur (arrêté du 4 septembre 1951). Cet essai consiste à soumettre le matériau à une source de chaleur rayonnante et à enflammer les gaz qui se dégagent éventuellement sur les deux faces. Nous avons relevé les caractéristiques suivantes :

1° Le temps au bout duquel le matériau s'enflamme sur chaque face;
2° La durée de combustion;
3° La perte de poids;
4° La surface de carbonisation du bois sur chacune des deux faces.

Un peu plus de 400 essais ont été effectués mais les 6 variables précédentes n'ont pas toujours été notées de telle sorte que nous ne disposons que de

321 résultats complets. Dans certaines comparaisons ne portant que sur deux variables nous disposons de 340 à 380 résultats.

Enfin, il y a lieu de noter que ces essais ont été effectués avec des produits industriels et que seuls les fabricants étaient maîtres des différentes variables (composition, concentration des solutions, épaisseur des enduits, nombre de couches, etc...), car les éprouvettes nous étaient remises toutes préparées. Cependant les propriétés de ces éprouvettes n'étaient pas complètement distribuées au hasard. Il y avait, en effet, des industriels qui avaient bien saisi la nature du problème et apportaient des produits efficaces et d'autres qui étaient à côté de la question et dont les techniques n'avaient qu'une efficacité réduite; enfin une minorité des éprouvettes occupaient une position intermédiaire.

Fonctions et courbes de régression. — Un physicien ou un ingénieur qui cherche à établir une relation entre deux variables x et y attachées à un même groupe d'expériences, donne à l'une de ces variables (x, par exemple), une série de valeurs $x_1, x_2...$ puis mesure les valeurs correspondantes $y_1, y_2...$ de y (après avoir rendu constantes toutes les autres variables naturellement). Il obtient ainsi une relation empirique ou trace une courbe $y\ (x)$. Mais bien souvent la mesure d'une des variables (y en général) est entachée d'une certaine imprécision. L'opérateur effectue alors un certain nombre d'expériences où x garde une valeur fixe; il obtient plusieurs nombres pour la valeur correspondante de y et c'est la moyenne de ces valeurs qu'il retient. Nous noterons $\bar{y}\ (x)$ la fonction ainsi obtenue.

Les statisticiens ont été amenés à étendre cette notion aux cas où ils n'étaient plus maîtres de fixer à leur gré ni l'une ni l'autre des variables. La fonction $\bar{y}\ (x)$ prend alors le nom de fonction de régression de y par rapport à x. Naturellement on peut aussi considérer la fonction de régression $\bar{x}\ (y)$ de x par rapport à y. Ces deux fonctions sont toujours distinctes sauf dans le cas d'une relation fonctionnelle entre x et y au sens ordinaire du mot.

Pour obtenir directement les courbes de régression, il faudrait disposer d'un nombre considérable de couples de valeurs de x et de y. Il existe certes des méthodes permettant de calculer des fonctions de régression à partir d'un nombre relativement restreint de couples de valeurs. Ces méthodes supposent que l'on se donne *a priori* un type de fonction avec un petit nombre de constantes. En fait, ce que l'on calcule, ce sont les valeurs les plus probables de ces constantes compatibles avec les résultats expérimentaux. En général, on choisit comme fonction un polynôme. Celui-ci est toujours de faible degré car dès que l'on dépasse 4 ou 5 constantes les calculs deviennent extrêmement longs. Si la liaison entre les deux variables est faible et si l'on n'a aucune raison théorique pour guider vers un type particulier de fonction, le résultat peut être complètement dépourvu de sens physique.

Dans le cas général on peut améliorer un peu la méthode en calculant une première courbe approximative au moyen d'une fonction arbitraire (un polynôme, par exemple), puis en recommençant tous les calculs avec une 2e fonction choisie en raison de la forme obtenue au cours de l'essai préliminaire. On augmente malheureusement ainsi très sensiblement la longueur de ces calculs

même si l'on peut réduire le nombre des constantes à adopter pour la fonction définitive.

La solution purement empirique suivante conduit à des résultats pratiquement équivalents avec des calculs beaucoup plus simples.

On classe les événements en fonction d'une des deux variables (x par exemple); puis on divise ces événements en un petit nombre de groupes. Dans chaque groupe la variable x ne varie donc que relativement peu. On prend alors les moyennes des valeurs de x et de y de chaque groupe. Il est facile de voir que la courbe qui passe par les points définis par ces moyennes tend vers une courbe différente de $\bar{y}\ (x)$ mais très peu lorsque le nombre d'expérience de chaque groupe augmente indéfiniment. La nouvelle fonction ainsi définie n'est donc pas une image fidèle de la fonction $\bar{y}\ (x)$, mais l'écart reste négligeable tant que n n'est pas très élevé. En classant les expériences par rapport aux valeurs croissantes de y on obtiendrait de même une courbe approchée de $\bar{x}\ (y)$.

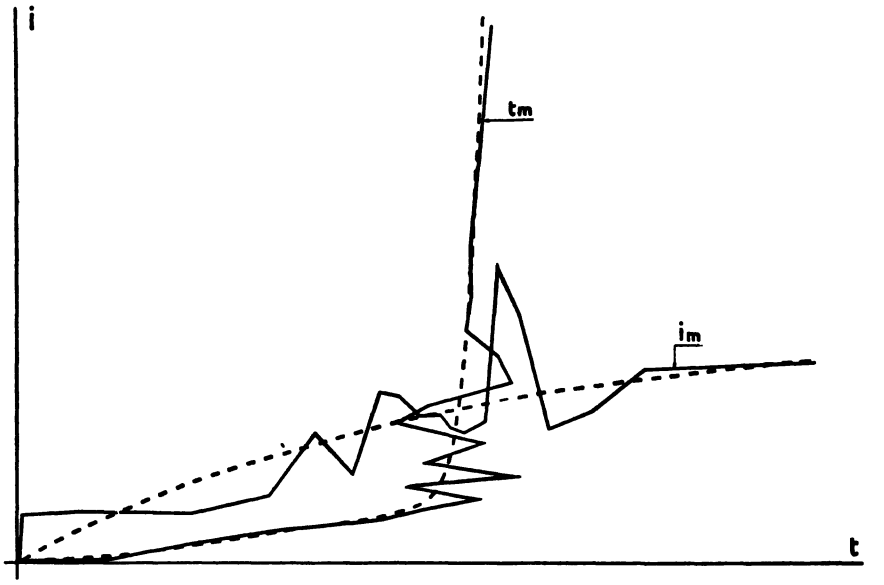


Fig. 1.

La figure 1 donne les résultats obtenus à partir de 380 essais d'inflammabilité. En abscisse nous avons porté l'inverse du temps d'inflammation sur la face exposée au radiateur (ι_e) et en ordonnées le temps de combustion t_c. Les expériences ont été divisées en 20 sous-groupes de 19 et les moyennes de chaque sous-groupe sont reliées par des lignes en trait plein. En pointillé on a dessiné un tracé simplifié approximatif pour les fonctions $\bar{\iota}_e\ (t_c)$ et $\bar{t}_c\ (\iota_e)$. On remarquera que : 1° Les lignes de régression sont très différentes; 2° Pour que les courbes dessinées puissent avec vraisemblance prétendre représenter fidèlement les fonctions de régression il aurait été nécessaire de disposer d'un nombre beaucoup plus élevé de valeurs expérimentales (au moins vingt fois plus).

Nous aurions pu, il est vrai, appliquer aux valeurs des moyennes partielles

le mode de calcul des fonctions de régression auquel nous avons fait allusion plus haut. La simplification aurait été considérable. Les calculs n'auraient en effet porté que sur 20 couples de valeur au lieu de 380 et nous aurions disposé d'une fonction approximative limitant notablement le choix du type définitif à adopter. Il est manifeste en effet que pour $\bar{i}_e\ (t_c)$ on peut se contenter d'une fonction parabolique et pour $\bar{t}_c\ (i_e)$ d'une relation homographique, ce qui, dans les deux cas, réduit à 3 le nombre des constantes arbitraires à calculer.

Nous n'avons pas effectué ces calculs. Dès que les variables sont faiblement liées les fonctions de régression présentent en effet de nombreux inconvénients :

1° A moins de disposer d'un nombre très élevé de résultats, les « constantes » ne peuvent être calculées qu'avec une faible précision;

2° Il existe, nous venons de le voir, deux fonctions de régression souvent fort différentes. Doit-on les conserver toutes deux ou une seule et dans ce cas laquelle? Si l'indépendance partielle entre les variables tient à l'imprécision des mesures sur l'une d'elles ce choix est évident, mais ce cas est assez rare. Dans l'exemple cité plus haut, les temps ont été mesurés au 1/5^e de seconde; les moyennes de 19 mesures sont donc connues avec une erreur inférieure au 1/20^e et les écarts sur ces moyennes atteignent plusieurs dizaines de secondes;

3° La forme même des lignes de régression dépend du mode de calcul des moyennes. En général, on utilise la moyenne arithmétique, mais ce choix est arbitraire. Si d'ailleurs on change d'échelle le résultat est différent suivant que les moyennes sont effectuées avant ou après transformation et les diffé-

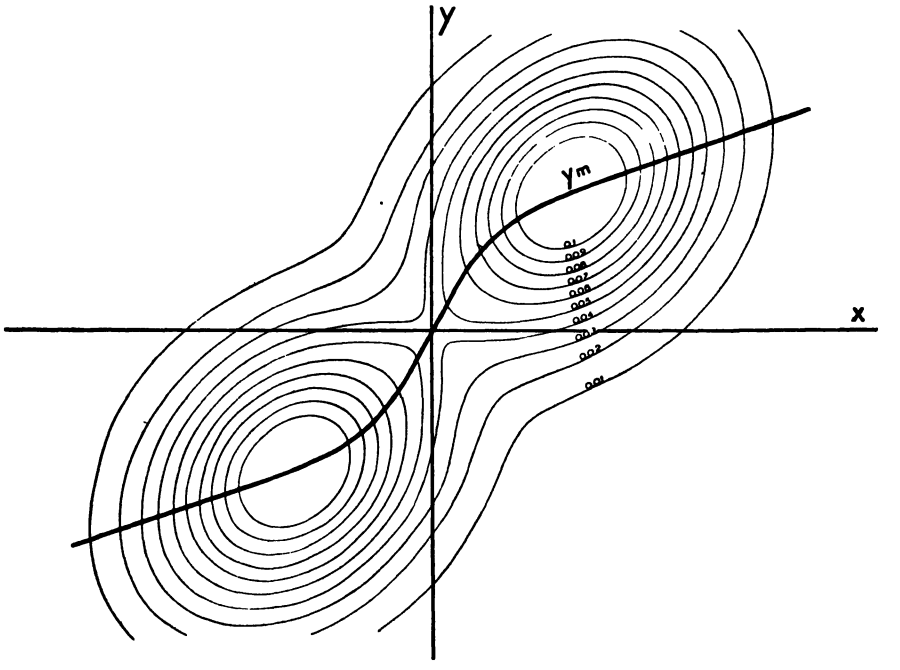


Fig. 2.

rences peuvent être notables si les variables sont faiblement liées. Voici un exemple tiré des valeurs précédentes : 4 essais d'inflammabilité effectués avec

un même produit ont donné des temps d'inflammation presque identiques et des temps de combustion respectifs de 130, 413, 128 et 142 secondes; la moyenne arithmétique est 203 et la moyenne harmonique 159.

4° L'interprétation des courbes de corrélation, même lorsqu'elles sont connues avec précision peut être très délicate. Considérons, par exemple, deux populations suivant une loi de Laplace-Gauss à deux variables ayant mêmes variances marginales et même coefficient de corrélation, seules les valeurs moyennes des variables étant différentes et d'une même quantité. Considérons alors une population mixte formée d'un mélange à parties égales de ces deux populations élémentaires. La densité de fréquence y sera strictement symétrique par rapport aux deux bissectrices des axes de coordonnées des variables (lignes en trait fin de la figure 2). Cependant, la ligne de corrélation $\bar{y}(x)$ n'est pas confondue avec l'un de ces axes et présente une inflexion bien marquée. Cette « anomalie » n'est donc pas due à une dissymétrie entre x et y mais à une répartition particulière. Or, si l'on ne disposait que d'un millier de couples de mesure le point d'inflexion apparaîtrait bien marqué mais l'anomalie de répartition serait fort difficile à percevoir et l'on risquerait d'interpréter ce point d'inflexion d'une manière erronée.

Les considérations suivantes permettent d'établir une correspondance univoque entre deux grandeurs qui remplace avantageusement l'ensemble des deux fonctions et qui échappe aux critiques précédentes. De plus, cette correspondance peut s'obtenir sans aucun calcul, elle est beaucoup plus précise et enfin on peut obtenir très simplement les limites de ses variations pour des limites de confiance données.

Fonction d'ordre (Ω_1). — Classons les événements en fonction des valeurs de x croissantes, notons les valeurs dans l'ordre où elles se succèdent. Reclassons les événements en fonction des valeurs croissantes de y et relevons également leurs valeurs dans ce nouvel ordre. Enfin à toute valeur de x de la première liste faisons correspondre celle de y qui occupe le même numéro d'ordre de la seconde sans se préoccuper de savoir si les valeurs ainsi accouplées appartiennent ou non au même événement. Nous désignerons par fonction d'ordre $\Omega_1(x, y)$, la relation ainsi obtenue et par courbe Ω_1 la ligne représentative de cette fonction.

Propriétés de la fonction Ω_1 :

1° Invariance. Si l'on modifie par anamorphose l'une des variables x ou y ou les deux à la fois sans modifier l'ordre de succession des valeurs, on modifie de la même manière la courbe Ω_1;

2° Lorsque la liaison entre les variables tend vers une relation fonctionnelle stricte la fonction d'ordre tend vers cette même fonction;

3° Lorsque les fréquences sont distribuées suivant une loi de Laplace-Gauss à deux variables la fonction d'ordre est linéaire. En effet, si les fréquences suivent une telle loi les répartitions marginales des variables suivent aussi des lois de Laplace-Gauss et on peut toujours transposer ces fonctions les unes dans les autres au moyen d'une simple transformation linéaire.

Bornes de la fonction d'ordre pour des limites de confiances données :

Soit x_0 une valeur particulière de x. Considérons un très grand nombre de

groupes de n événements et dans chaque groupe celui pour lequel x a pris la valeur la plus voisine de x_o. Son ordre m variera dans chaque groupe. Soit μ la valeur moyenne de m. Si n est assez grand m-μ suit une loi de probabilité assimilable à une loi de Laplace-Gauss de moyenne nulle et d'écart type $\sqrt{\frac{\mu\,(n-\mu)}{n}}$. Considérons d'autre part y_o la valeur de y qui a le même rang moyen μ et m' le rang de y_o dans une série donnée l'écart entre m' et μ suit naturellement la même loi. Pour une série donnée m' est en général différent de m et s'il n'existe aucune corrélation entre x et y la différence $m - m'$ suit une loi de Laplace-Gauss de moyenne nulle et d'écart-type $\sqrt{\frac{2\,\mu\,(n-\mu)}{n}}$.

Réciproquement si dans une même série d'événements nous considérons deux valeurs x_m et y_m ayant le même ordre m, les rangs moyens correspondants μ et μ' ne sont en général pas identiques et leur différence suit une loi de Laplace-Gauss de moyenne nulle et d'écart-type $\sqrt{\frac{2\,m\,(n-m)}{n}}$. Se donnant des limites de confiance donnée on en déduit la valeur maximum pour cet écart $\mu - \mu'$. Se reportant alors au tableau de la fonction Ω_1 on peut pour toute valeur de x obtenir la valeur la plus probable de y, la valeur de m puis les limites de m' et finalement celles de y pour les limites de confiance choisies.

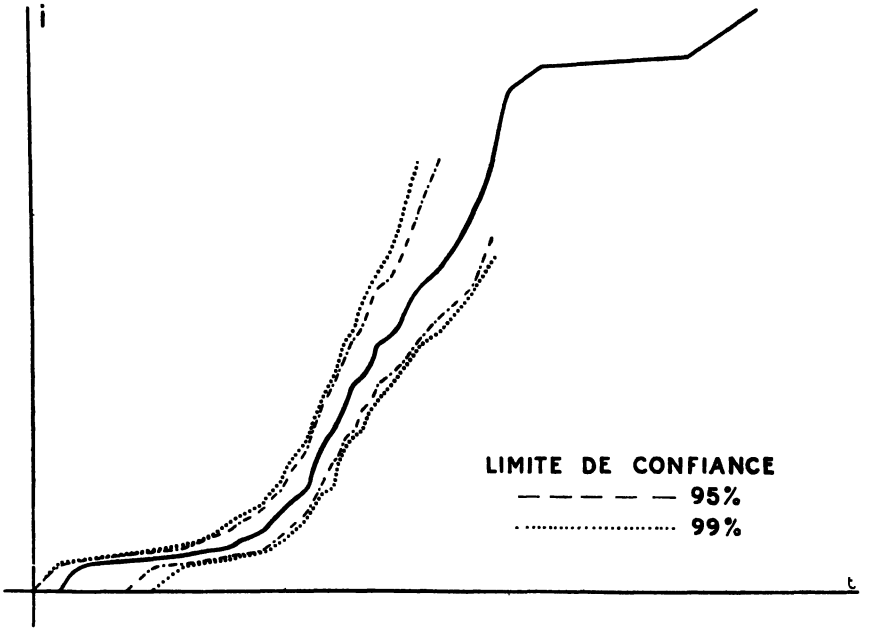


Fig. 3.

Sur la figure 3 nous avons reporté les résultats obtenus avec les 380 couples de valeurs de t_c et i_e déjà utilisés pour tracer les lignes de régression de la figure 1 La courbe Ω_1 est en trait plein, les limites de confiance à 95 % en traits mixtes et celles à 99 % en traits interrompus. Il convient de remarquer que les calculs ont été effectués en négligeant la corrélation entre t_c et i_e; si l'on en avait

tenu compte les zones délimitées auraient été plus étroites. Même en la négligeant on voit que Ω_1 est beaucoup mieux connu que les lignes de régression. Dans ces conditions la recherche d'une fonction théorique pour Ω_1 peut avoir un sens physique même si l'on ne dispose pas d'un nombre élevé de résultats expérimentaux.

Exemple d'utilisation de la fonction Ω_1.

Enfin voici un exemple d'application où la fonction Ω_1 n'aurait pu être remplacée par aucune autre fonction :

On a établi une classification des différents ignifuges au moyen de la durée de combustion du subjectile protégé. Le mode opératoire le meilleur étant celui qui conduit à la durée de combustion la plus courte. Sur cette base on a établi deux coupures ce qui conduit à trois classes d'ignifuges : bons, médiocres et mauvais. On se propose de remplacer le critère durée de combustion par celui du temps d'inflammation. Quelles nouvelles coupures doit-on choisir pour que la nouvelle classification ait la même sévérité que la première, c'est-à-dire comporte le même nombre d'ignifuges dans chaque classe. Il est évident que les coupures doivent se correspondre dans la classification par ordre de valeurs croissantes pour t_c et i_e, c'est-à-dire par la fonction Ω_1. Naturellement les deux modes de classements ne sont pas identiques et certains ignifuges considérés primitivement comme bons deviendront médiocres, voire même mauvais, mais ils seront remplacés par un nombre équivalent d'autres ignifuges qui accéderont à la première classe.

Coefficient de corrélation ordinaire. Tous les statisticiens connaissent le coefficient de corrélation ordinaire entre deux variables aléatoires, calculé à partir des variances et du comoment et en font le plus large usage. Nous n'avons cependant pas pu l'utiliser dans nos études sur l'inflammabilité des matériaux pour les raisons suivantes :

1° La corrélation ainsi calculée n'est invariante que vis-à-vis des transformations linéaires, le choix des variables introduit donc un certain arbitraire dans les calculs. ;

2° Le coefficient de corrélation ne traduit la plus ou moins grande dépendance entre les variables que lorsque la densité de répartition des couples répond à certaines lois particulières par exemple, lorsque celle-ci dépend d'une loi de Laplace-Gauss à deux variables et probablement d'une manière plus générale lorsque la fonction Ω_1 est linéaire. Si l'on applique en effet le calcul de la corrélation à deux variables strictement liées par une relation fonctionnelle on trouve en général un coefficient inférieur à l'unité. Considérons, par exemple, la fonction croissante $y = x^3$ dans l'intervalle 0 à 1 et supposons la densité de répartition de x uniforme dans cet intervalle un calcul élémentaire donne comme coefficient de corrélation 0,93.

Coefficient de corrélation de rang. Pour remédier à ces deux inconvénients on a imaginé de remplacer le coefficient calculé à partir du comoment par un autre obtenu de la manière suivante.

Les objets du groupe étant classés successivement suivant les valeurs croissantes de x puis de y on note leurs rangs m_x et m_y ainsi obtenu dans chacun de ces classements et on forme la quantité $\Sigma\ (m_x - m_y)^2$. Si les deux variables

sont liées par une relation fonctionnelle constamment croissante cette somme est évidemment nulle. On montre facilement qu'elle est égale à $\frac{n(n^2-1)}{6}$ si les deux variables sont indépendantes et à $\frac{n(n^2-1)}{3}$ si elles sont liées par une relation fonctionnelle constamment décroissante. Dans ces conditions on prend pour coefficient de corrélation l'expression :

$$\gamma = 1 - \frac{6\,\Sigma\,(m_x - m_y)^2}{n(n^2-1)} \qquad (2)$$

Il est évident que ce coefficient est un invariant vis-à-vis de toute transformation des variables qui ne modifie pas leur ordre de succession.

Cette méthode particulièrement étudiée par Olds (1 et 2) à un champ d'action beaucoup plus étendu que le coefficient de corrélation ordinaire et nous a donné un certain nombre de résultats utiles. Cependant, la connaissance de ce coefficient de corrélation ne suffit pas à définir la distribution de m_x et m_y. S'il en était ainsi en effet, il serait possible par une transformation convenable des deux variables d'obtenir une densité de répartition suivant une loi de Laplace-Gauss. Il n'en est pas toujours ainsi.

Portons en effet les résultats relatifs à un groupe de n objets dans un tableau carré de n cases de côté de telle manière que l'objet de rangs m_x et m_y soit dans la case commune à la m_x^e colonne et la m_y^e ligne. Il y a n cases sur n^2 du tableau occupée et les marges sont réparties uniformément. Or, M. le Professeur Frechet a montré d'une manière très générale que la distribution des valeurs marginales d'un tableau ne suffisait pas à définir ce dernier et qu'il existait un nombre considérable de tableaux répondant à cette propriété (3). D'une manière plus précise on voit immédiatement qu'il existe dans le cas présent $n!$ tableaux différents. Or, la somme $\Sigma\,(m_x - m_y)^2$ entière et positive ne peut dépasser $\frac{n(n^2-1)}{3}$, le coefficient de corrélation de rang ne peut donc prendre que $\frac{n(n^2-1)}{3}$ valeurs au plus; il y aura donc en moyenne $\frac{3(n-2)!}{n+1} \# 3\,(n-3)!$ tableaux correspondant à une même valeur d'un coefficient de corrélation de rang donné.

D'une manière générale nous dirons qu'une répartition d'événements se fait suivant une loi pseudo-normale lorsqu'une transformation convenable des variables permet d'obtenir des densités de fréquences réparties suivant une loi de Laplace-Gauss. Il résulte du calcul précédent que toutes les répartitions ne peuvent pas être pseudo-normales.

Les différences entre les tableaux correspondant à une même valeur du coefficient de corrélation de rang peuvent-elles du moins s'expliquer par une

---

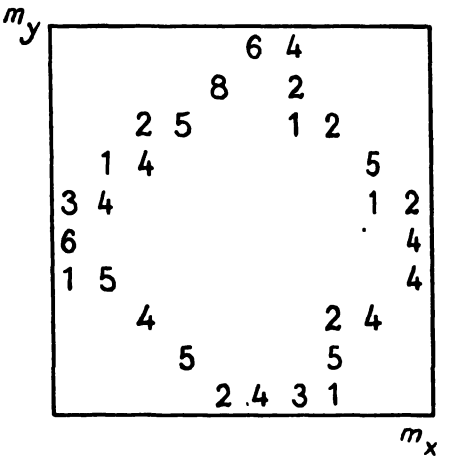
(1) E. G. Olds. *Distribution of sums of squares of rang differences for small numbers of individuals. Annals Math. Stat.* IX, p. 133-148 (1938).

(2) E. G. Olds. *The 5 % significance levels for sums of squares of rank differences and a correction. Annals Math. Stat.*, XX, p. 117-118 (1949).

(3) Maurice Fréchet. *Sur les tableaux de corrélation dont les marges sont données. C. R. Ac. Sc.* 242, p. 2426-2428 (1956). Voir aussi *Annales de l'Université de Lyon*, Section A, p. 53-77 (1951).

dispersion due au hasard des prélèvements à partir d'une population répartie suivant une loi pseudo-normale? L'exemple suivant montre qu'il n'en est rien. Considérons le tableau (1) obtenu à partir de 100 tirages dans une population où les deux variables étaient liées par la relation $x^2 + y^2 = C^{te}$ avec une densité de répartition homogène sur l'arc de cercle. Pour simplifier la présentation le tableau a été réduit à 100 cases avec une répartition marginale uniforme de 10 événements par ligne et par colonne. Le coefficient de corrélation de rang est — 0,044 très voisin de la valeur théorique nulle. Il est évident qu'une telle répartition est très différente de celle obtenue à partir d'une loi de Laplace-Gauss à corrélation nulle. Testons cette répartition. Le coefficient de corrélation étant nul si la loi de répartition était pseudo-normale il y aurait indépendance entre m_x et m_y et par conséquence la probabilité de présence serait de 1 par case. D'autre part, il y a 81 degrés de liberté. Le test du χ^2 donne alors un écart réduit de 10,3.

TABLEAU 1

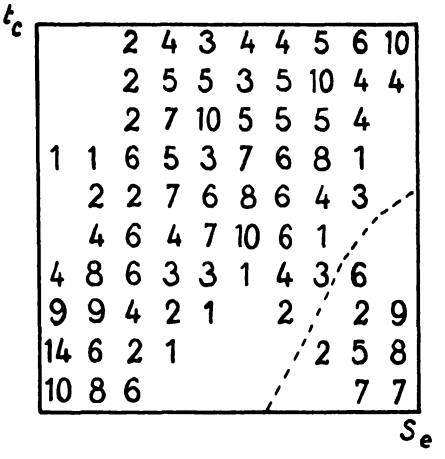


Les densités marginales des deux variables et le coefficient de corrélation de rang ne suffisent donc pas à définir la loi de distribution d'une population. On peut même aller plus loin et dire que le coefficient de corrélation n'a de sens que si la répartition suit une loi pseudo-normale, nous justifierons plus loin cette affirmation.

Il ne s'agit pas là de spéculations purement théoriques et nous allons étudier quelques cas tirés de notre étude sur l'inflammabilité des matériaux.

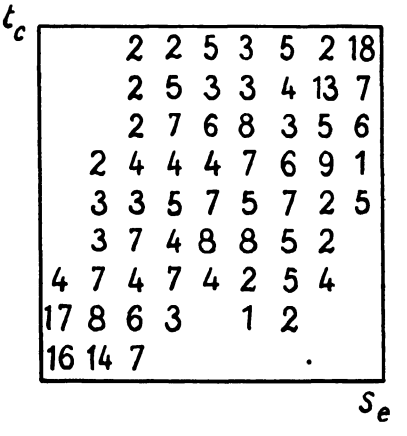
Considérons le tableau 2 relatif à 380 couples de valeurs relatives à la surface de carbonisation sur la face exposée (S_e) et au temps de combustion (t_c) répartis dans 100 cases avec une densité marginale de 38. Le coefficient de corrélation de rangs est 0,32. La dissymétrie de ce tableau est manifeste. Elle ne peut être le simple fait du hasard. En effet, le test du χ^2 appliqué à cette dissymétrie et dont nous indiquons le calcul plus loin conduit à un écart réduit de 7,41. La loi de répartition n'est donc sûrement pas pseudo-normale.

TABLEAU 2



Une étude critique de l'origine des données expérimentales nous a d'ailleurs montré que ces valeurs devaient être partagées en deux groupes approximativement séparés sur le tableau 1 par le pointillé. Si l'on élimine les 47 valeurs étrangères (coin inférieur droit du tableau) et que l'on reclasse les 343 valeurs résiduelles dans un tableau carré à 9 cases de côté on obtient la distribution du tableau 3 dont l'aspect symétrique est beaucoup plus satisfaisant. Le test symétrie conduit en effet à un écart-type du χ^2 qui reste un peu élevé 1,38 mais cependant encore acceptable. Le coefficient de corrélation de rang est également plus grand : 0,68.

TABLEAU 3



Une distribution dissymétrique ne provient pas nécessairement de la coexistence de plusieurs types de distribution. Considérons le tableau 4 relatif aux relations entre le temps d'inflammation sur la face exposée (t_e) et le temps de combustion (t_c). La dissymétrie se traduit cette fois par une accumulation de valeurs dans le coin supérieur gauche. Or, une telle répartition est propre

au phénomène étudié. Lorsqu'il n'y a pas d'inflammation le temps de combustion est naturellement nul. Mais quel que soit le type d'ignifuge étudié on peut toujours obtenir la protection totale car il suffit d'utiliser une couche protectrice assez épaisse. On conçoit donc que lorsque l'inflammation est tardive il ne se dégage que très peu de gaz inflammables et que la combustion reste

TABLEAU 4

t_e									t_c
32	3	1	1						1
	19	6	4	4	2	2		1	
1	4	10	3	6	6	1	3	1	3
2	4	3	3	3	4	6	8	1	4
1	1	1	3	5	5	7	6	6	3
1	1	7	3	5	3	7	3	2	6
	5	3	7	3	5	3		9	3
1	1	2	4	1	2	4	6	10	7
		4	5	4	5	6	7	5	2
		1	5	7	6	2	5	3	9

brève. Au voisinage du coin supérieur gauche il doit donc y avoir une relation étroite entre t_e et t_c et la corrélation ira en diminuant au fur et à mesure que l'inflammation sera plus brève.

Avant de calculer le coefficient de corrélation de rang il faut donc s'assurer que la répartition correspond à une loi pseudo-normale. Pour cela on commencera par dresser le tableau des répartitions puis l'on testera ce tableau. Malheureusement ce test comporte des calculs compliqués. La densité de répartition théorique varie en effet avec chaque case dès que le coefficient de corrélation diffère de zéro ou de l'unité.

On peut heureusement effectuer un test très simplifié qui permet une première discrimination. Remarquons en effet que la répartition doit être symétrique par rapport aux deux diagonales du tableau. Le long de ces diagonales les fréquences sont égales deux à deux; elles le sont quatre à quatre dans le reste du tableau et l'on voit facilement qu'un tableau de n^2 cases possède $\frac{n\,(n+2)}{4}$ degrés de liberté vis-à-vis de cette propriété si n est pair et $\frac{(n-1)\,(n+3)}{4}$ si n est impair.

Naturellement, il peut exister des répartitions très différentes d'une loi pseudo-normale et qui satisfont cependant à ce test préliminaire. La répartition du tableau 1 est par nature symétrique. En effet, l'écart réduit du χ^2 du test symétrique est de — 0,60 pour 70 degrés de liberté. Nous avons vu cependant que la répartition s'écartait profondément de celle d'une loi pseudo-normale.

Lorsque la répartition est symétrique il convient donc de compléter ce test. Nous proposons dans ce but la méthode suivante :

Soient r un nombre entier positif au plus égal à n le nombre total d'événements et t le nombre d'événements tels que l'on ait simultanément :

$$m_x \leqslant r \text{ et } m_y \leqslant r \tag{3}$$

Posons

$$\Theta = \frac{t}{n} \text{ et } \rho = \frac{r}{n} \tag{4}$$

et considérons la fonction :

$$\Theta = \Omega_2 (\rho) \tag{5}$$

Il est évident que si la loi de distribution de x et y est pseudo-normale cette fonction ne dépend que de la corrélation γ. Appelons N_γ la fonction correspondant à une loi pseudo-normale de corrélation γ. La méthode que nous proposons consiste à vérifier que Ω_2 s'identifie avec une fonction N ou plus exactement à tester les écarts $\Omega_2 - N_\gamma$. Pour cela pour chaque valeur de r nous calculerons t (c'est un simple décompte puis ρ et Θ et nous vérifierons que les écarts sont compatibles avec des limites de confiance choisies à l'avance.

Avant d'indiquer comment s'effectue ce calcul, il est nécessaire de donner quelques indications sur les propriétés des fonctions Ω_2.

On a par définition :

$$0 \leqslant t \leqslant r \tag{6}$$

et par conséquent :

$$0 \leqslant \Theta \leqslant \rho \tag{7}$$

D'autre part, considérons le tableau de distribution de m_x et m_y; dans chaque colonne et chaque ligne il y a une unité. Donc lorsque r augmente de une unité, t ne change pas, augmente de 1 unité ou de deux au maximum. Si l'on assimile Θ à une fonction continue on peut donc écrire :

$$0 \leqslant \frac{d\Theta}{d\rho} \leqslant 2 \tag{8}$$

Considérons alors une relation fonctionnelle croissante entre x et y, on a constamment $m_x = m_y = \rho$ donc $t = \rho$ ce qui est une limite en raison de l'équation (7). Si au contraire, la relation est fonctionnelle et décroissante $m_x + m_y = n$ et par conséquent pour $\rho \leqslant \frac{1}{2}$ $t = 0$. La relation (8) montre alors que si $\rho \geqslant \frac{1}{2}$, $\Theta = 2\rho - 1$ et la fonction formée de ces deux parties constitue une deuxième limite. Toute fonction Ω_2 est donc nécessairement comprise entre ces deux limites.

Pour calculer les écarts possibles de Θ, nous admettrons que toutes les valeurs sont également probables entre ces deux limites. Si n est assez grand l'écart ε entre la valeur expérimentale Θ et la valeur que prendrait cette quantité

si le nombre d'événements était infini est assimilable à une loi de Laplace-Gauss de moyenne nulle et un calcul classique donne pour écart-type

$$\sqrt{\frac{\Theta(\rho - \Theta)}{n\rho}} \text{ si } n\rho \leqslant \frac{1}{2}$$

$$\sqrt{\frac{(1-\Theta)(\rho - \Theta)}{n(1-\rho)}} \text{ si } n\rho \geqslant \frac{1}{2} \qquad (9)$$

Au lieu de tracer directement les valeurs limites calculées à partir des formules (9) pour des limites de confiance données il est plus commode de calculer pour chaque valeur de ρ les valeurs des corrélations γ des fonctions N qui ont même valeur que Θ et ses limites. On trace alors les variations de γ et de ses limites en fonction de ρ. Pour qu'une loi de distribution soit pseudo-normale il faut qu'il existe une droite parallèle à ρ entièrement comprise entre les dites limites. L'opération donne alors immédiatement la corrélation et ses limites pour la loi pseudo-normale.

Fonction de corrélation. Que peut-on dire d'une distribution lorsqu'elle n'est pas pseudo-normale. La notion de corrélation n'a alors plus de sens et il faut la remplacer par une notion plus complexe. Désignant alors par fonction de corrélation les variations de γ en fonction de ρ

$$\gamma = \Gamma(\rho)$$

nous caractériserons les dites distributions par les deux fonctions Ω_1 et Γ. Le calcul précédent donne alors immédiatement le moyen de déterminer deux courbes entre lesquelles est comprise la fonction Γ pour des limites de confiance données.

Les figures 4 et 5 donnent deux exemples des dites fonctions.

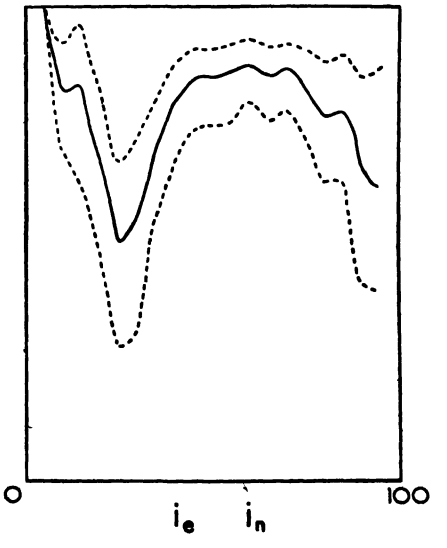


Fig. 4.

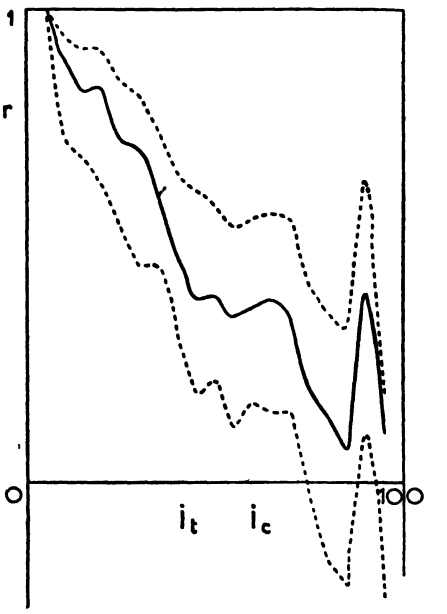


Fig. 5.

Nous ne possédons pas encore une expérience suffisante des fonctions de corrélation pour indiquer les avantages exacts et les inconvénients de cette notion nouvelle dans l'étude des distributions mais on conçoit que la présence d'un minimum bien marqué comme celui de la figure 4 correspond nécessairement à un phénomène physique difficile à détecter autrement et qu'il importe d'interpréter.

Nous nous proposons d'étendre ces notions au cas de plus de deux variables et d'introduire, en particulier, la notion de corrélation multiple.

Appendice

Calcul des fonctions N

Une fonction N ne dépendant que de la valeur de la corrélation nous pouvons utiliser pour la calculer une fonction de Laplace-Gauss à deux variables dont les densités marginales ont une moyenne nulle et une variance unité. La densité de probabilité est alors :

$$\frac{1}{2\pi\sqrt{1-\gamma^2}}\, e^{-\frac{x^2-2\gamma xy+y^2}{2(1-\gamma^2)}} \tag{10}$$

Posons alors u tel que :

$$\varphi = \frac{1}{\sqrt{2\pi}} \int_{x=-\infty}^{x=u} e^{-\frac{x^2}{2}}\, dx = \Phi(u) \tag{11}$$

on aura par définition :

$$\Theta = \frac{1}{2\pi\sqrt{1-\gamma^2}} \int_{x=-\infty}^{x=u} dx \int_{y=-\infty}^{y=u} e^{-\frac{x^2-2\gamma xy+y^2}{2(1-\gamma^2)}}\, dy \tag{12}$$

Donnons à u un accroissement du, Θ s'accroît de la valeur de l'intégrale double dans deux zones étendues respectivement de $x = u$ à $x = u + du$, $y = -\infty$ à $y = u$ et $x = -\infty$ à $x = u$, $y = u$ à $y = u + du$. En raison de la symétrie par rapport à la première bissectrice, la valeur de l'intégrale est la même dans ces deux zones. On a donc :

$$\frac{d\Theta}{du} = \frac{1}{\pi\sqrt{1-\gamma^2}} \int_{y=-\infty}^{y=u} e^{-\frac{u^2-2\gamma uy+y^2}{2(1-\gamma^2)}}\, dy$$

En posant $Y = \dfrac{y-\gamma u}{\sqrt{1-\gamma^2}}$, on a facilement :

$$\frac{d\Theta}{du} = \sqrt{\frac{2}{\pi}}\, e^{-\frac{u^2}{2}}\, \Psi\left(\sqrt{\frac{1-\gamma}{1+\gamma}}\, u\right) \tag{13}$$

$$\Theta = \sqrt{\frac{2}{\pi}} \int_{-\infty}^{u} e^{-\frac{u^2}{2}}\, \Psi\left(\sqrt{\frac{1-\gamma}{1+\gamma}}\, u\right) du \tag{14}$$

Au moyen d'une table de la fonction Ψ les équations (11) et (13) permettent de calculer Θ en fonction de ρ par intégration numérique pour chaque valeur de γ.

Pour certaines applications il peut être utile de calculer Θ en fonction de γ pour une valeur donnée de ρ. On a successivement :

$$\frac{\partial}{\partial \gamma}\sqrt{\frac{1-\gamma}{1+\gamma}} = \frac{-1}{(1+\gamma)\sqrt{(1-\gamma^2)}}$$

$$\frac{\partial \Psi}{\partial \gamma}\left(\sqrt{\frac{1-\gamma}{1+\gamma}}\,u\right) = \frac{\partial}{\partial \gamma}\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{u\sqrt{\frac{1-\gamma}{1+\gamma}}} e^{-\frac{x^2}{2}}\,dx = \frac{u}{\sqrt{2\pi}}\,e^{-\frac{u^2}{2}\left(\frac{1-\gamma}{1+\gamma}\right)}\frac{\partial}{\partial \gamma}\sqrt{\frac{1-\gamma}{1+\gamma}}$$

$$= \frac{-u}{(1+\gamma)\sqrt{2\pi}\sqrt{1-\gamma^2}}\,e^{-\frac{(1-\gamma)u^2}{2(1+\gamma)}}$$

$$\frac{d\Theta}{d\gamma} = \frac{-1}{2\pi\sqrt{1-\gamma^2}}\int_{-\infty}^{u}\frac{2\,u\,e^{-\frac{u^2}{1+\gamma}}}{1+\gamma}\,du = \frac{-e^{-\frac{u^2}{1+\gamma}}}{2\pi\sqrt{1-\gamma^2}} \qquad (15)$$

On a donc cette fois Θ par une intégrale numérique plus simple que l'équation (14).

Lorsque u n'est pas trop grand, en limitant l'exponentielle de l'équation (11) aux deux premiers termes de son développement on obtient :

$$\rho - \frac{1}{2} = \frac{1}{\sqrt{2\pi}}\left(u - \frac{u^3}{6}\right) \qquad (16)$$

d'où l'on tire :

$$u = \left(\rho - \frac{1}{2}\right)\sqrt{2\pi} + \frac{1}{6}\left(\rho - \frac{1}{2}\right)^3 2\pi)^{3/2}$$

$$u^2 = 2\pi\left(\rho - \frac{1}{2}\right)^2 + \frac{4\pi^2}{3}\left(\rho - \frac{1}{2}\right)^4 \qquad (17)$$

d'autre part, en limitant l'exponentielle de l'équation (15) à ses trois premiers termes on a :

$$\frac{d\Theta}{d\gamma} = \frac{1}{2\pi\sqrt{1-\gamma^2}} - \frac{u^2}{2\pi(1+\gamma)\sqrt{1-\gamma^2}} + \frac{u^4}{4\pi(1+\gamma)^2\sqrt{1-\gamma^2}} \qquad (18)$$

une intégration facile donne alors :

$$\Theta = \frac{\arcsin\gamma}{2\pi} + \frac{u^2}{2\pi}\sqrt{\frac{1-\gamma}{1+\gamma}} - \frac{u^4}{12\pi}\frac{2+\gamma}{1+\gamma}\sqrt{\frac{1-\gamma}{1+\gamma}} + C^{te} \qquad (19)$$

en remplaçant u^2 par sa valeur tirée de l'équation (17) :

$$\Theta = \frac{\arcsin\gamma}{2\pi} + \left(\rho - \frac{1}{2}\right)^2\sqrt{\frac{1-\gamma}{1+\gamma}} + \frac{\pi}{3}\left(\rho - \frac{1}{2}\right)^4\frac{\gamma}{1+\gamma}\sqrt{\frac{1-\gamma}{1+\gamma}} + C^{te}$$

Variations de Θ pour les fonctions pseudo-normales

γ	ρ → 0,05	0,10	0,15	0,20	0,25	0,30	0,35	0,40	0,45	0,50
0,00	0,002	0,010	0,022	0,040	0,062	0,090	0,122	0,160	0,202	0,250
02	03	11	24	42	64	92	125	163	206	253
04	3	11	25	43	66	95	128	166	209	256
06	3	12	26	45	68	97	131	169	212	260
08	3	13	27	46	70	100	133	172	215	263
10	4	13	28	48	72	102	136	175	218	266
12	4	14	29	50	74	105	139	178	221	269
14	4	15	31	51	77	107	142	181	225	272
16	5	16	32	53	79	110	145	184	228	276
18	5	16	33	55	81	112	148	187	231	279
20	5	17	35	57	83	115	150	190	234	282
22	6	18	36	59	85	117	153	193	237	285
24	6	19	37	60	87	120	156	196	241	289
26	6	20	38	62	90	123	159	200	244	292
28	7	21	40	64	92	125	162	203	247	295
30	7	22	41	66	94	128	165	206	250	298
32	8	23	43	68	97	130	168	209	254	302
34	8	24	44	70	99	133	171	212	257	305
36	8	25	46	72	101	136	174	215	260	309
38	9	26	48	74	104	139	177	219	264	312
40	9	27	49	76	106	142	180	222	267	315
42	9	28	51	78	109	145	183	225	271	319
44	10	29	52	80	111	147	186	229	274	322
46	11	30	54	83	114	150	190	232	278	326
48	12	31	56	85	117	153	193	236	281	330
50	12	32	58	87	119	156	196	239	285	333
52	13	34	59	89	122	159	200	243	288	337
54	13	35	61	92	125	163	203	246	292	341
56	14	36	63	94	128	166	206	250	296	345
58	15	38	65	97	131	169	210	254	300	348
60	15	39	67	99	134	172	214	257	304	352
62	16	40	69	102	137	176	217	261	308	356
64	17	42	71	104	140	179	221	265	312	360
66	18	43	74	107	143	183	225	269	316	365
68	19	45	76	110	146	186	229	273	320	369
70	20	47	78	113	150	190	233	278	324	373
72	20	48	81	116	153	194	237	282	329	378
74	21	50	83	119	157	198	241	286	334	383
76	22	52	86	122	161	202	246	291	338	387
78	24	54	88	125	164	206	250	296	343	392
80	25	56	91	129	168	211	255	301	348	398
82	26	58	94	133	173	216	260	306	354	403
84	27	61	97	137	177	220	265	312	359	409
86	29	63	101	141	182	226	271	317	365	415
88	30	66	104	145	187	231	277	324	372	421
90	32	69	108	150	192	237	283	330	379	428
92	34	72	113	155	199	244	290	338	386	436
94	36	76	118	161	206	252	299	346	395	445
96	38	80	124	168	214	260	308	356	405	455
98	42	86	132	178	224	272	320	369	418	468
1,00	50	100	150	200	250	300	350	400	450	500

Or, si $\gamma = 1$, $\Theta = \rho$ et si $\gamma = o$ $\Theta = \rho^2$ la constante est donc égale à $\rho - \frac{1}{4}$ et l'on peut écrire finalement :

$$\Theta = \frac{\arcsin \gamma}{2\pi} + \left(\rho - \frac{1}{4}\right) + \left(\rho - \frac{1}{2}\right)^2 \sqrt{\frac{1-\gamma}{1+\gamma}} + \frac{\pi}{3}\left(\rho - \frac{1}{2}\right)^4 \frac{\gamma}{1-\gamma}\sqrt{\frac{1-\gamma}{1+\gamma}} \tag{20}$$

Cette formule donne Θ avec 3 décimales exactes pour ρ compris entre 0,15 et 0,85 et l'on peut supprimer le dernier terme lorsque ρ est compris entre 0,25 et 0,75.

par Lucien Amy,
Ingénieur en chef au Laboratoire Municipal,
Docteur-ès-Sciences.