JOURNAL DE LA SOCIÉTÉ STATISTIQUE DE PARIS

LUCIEN AMY

Étude statistique de l'expression $\frac{p_1}{q_1} = \frac{p_2}{q_2} = \cdots = \frac{p_i}{q_i} = \cdots = c^{\text{te}}$. Première application à l'étude de la fréquence des formules digitales

Journal de la société statistique de Paris, tome 81 (1940), p. 39-47

<http://www.numdam.org/item?id=JSFS_1940__81__39_0>

III

ÉTUDE STATISTIQUE DE L'EXPRESSION

$$\frac{p_1}{q_1} = \frac{p_2}{q_2} = \cdots = \frac{p_i}{q_i} = \cdots = c^{te}$$

PREMIÈRE APPLICATION A L'ÉTUDE DE LA FRÉQUENCE DES FORMULES DIGITALES

Soient : A_1, A_2, A_3,... A_i,... B_1, B_2..., B_i,... un certain nombre d'événements s'excluant réciproquement; p_1, p_2,... p_i..., q_1, q_2,... q_i..., leurs probabilités respectives au cours d'une épreuve. On fait n épreuves et l'on observe a_1 fois l'événement A_1, a_2 fois l'événement A_2,... b_1 fois l'événement B_1,...

Nous supposerons les probabilités inconnues et nous nous proposons de voir dans quelle mesure l'étude des nombres a_1, a_2,... a_i,... b_1,... b_i... permet de confirmer ou d'infirmer l'existence de la relation

$$(1) \qquad \frac{p_1}{q_1} = \frac{p_2}{q_2} = \cdots = \frac{p_i}{q_i} = \cdots = \rho.$$

Valeurs les plus probables des probabilités. — La probabilité pour que, au cours de n tirages, on ait observé les événements A_1, A_2,... A_i,... B_1; B_2... B_i... un nombre de fois respectif : a_1, a_2,... a_i,... b_1, b_2..., b_i... est :

$$(2) \quad \mathfrak{P} = \frac{n!}{[\Pi(a_i!\, b_i!)]\,[n - \Sigma(a_i + b_i)]!}\, \Pi(p_i^{a_i} q_i^{b_i})\,[1 - \Sigma(p_i + q_i)]^{n - \Sigma(a_i + b_i)}.$$

Nous adopterons comme estimation de p_i et de q_i les valeurs qui, respectant l'équation (1) rendront $\mathfrak{P}$ maximum.

Posons :

$$p_i + q_i = r_i$$

L'équation (1) donne :

$$p_i = \rho\, q_i$$

$$q_i + \rho\, q_i = r_i$$

d'où :

$$(3) \qquad q_i = \frac{r_i}{1 + \rho}$$

$$(4) \qquad p_i = \frac{\rho\, r_i}{1 + \rho}$$

et par conséquent :

$$p_i^{a_i} q_i^{b_i} = r_i^{a_i + b_i} \frac{\rho^{a_i}}{(1 + \rho)^{a_i + b_i}}$$

L'équation (2) devient alors :

$$\mathfrak{L} = \frac{n!}{[\Pi(a_i!\, b_i!)][n - \Sigma(a_i + b_i)]!} \frac{\rho^{\Sigma a_i}}{(1+\rho)^{\Sigma(a_i+b_i)}} [\Pi(r_i^{a_i+b_i})](1 - \Sigma r_i)^{n-\Sigma(a_i+b_i)}.$$

Pour avoir les valeurs les plus probables de ρ et de r_i, il suffit d'annuler les dérivées logarithmiques :

$$\frac{d \log \mathfrak{L}}{d\rho} = \frac{\Sigma a_i}{\rho} - \frac{\Sigma(a_i + b_i)}{1+\rho} = 0$$

$$\frac{\Sigma a_i}{\rho} = \frac{\Sigma(a_i + b_i)}{1+\rho} = \frac{\Sigma b_i}{1}$$

et, par conséquent, la *valeur la plus probable du rapport commun* ρ est :

$$\rho = \frac{\Sigma a_i}{\Sigma b_i} \tag{5}$$

De même :

$$\frac{d \log \mathfrak{L}}{dr_i} = \frac{a_i + b_i}{r_i} - \frac{n - \Sigma(a_i + b_i)}{1 - \Sigma r_i} = 0$$

$$\frac{a_1 + b_1}{r_1} = \frac{a_2 + b_2}{r_2} = \ldots = \frac{a_i + b_i}{r_i} = \ldots\ldots = \frac{\Sigma(a_i + b_i)}{\Sigma r_i} = \frac{n - \Sigma(a_i + b_i)}{1 - \Sigma r_i} = \frac{n}{1}$$

et par conséquent :

$$r_i = \frac{a_i + b_i}{n} \tag{6}$$

En portant cette dernière valeur dans les équations (3) et (4) nous avons les estimations les plus probables de p_i et de q_i :

$$p_i = \frac{\rho(a_i + b_i)}{(1+\rho)\, n}$$

$$q_i = \frac{a_i + b_i}{(1+\rho)\, n}$$

Écarts. — Considérons un grand nombre de séries de n épreuves, les événements A_i se produiront en moyenne np_i fois. Or au cours de la série particulière étudiée, ils se sont produits a_i fois; l'écart est donc : $a_i - np_i$.

Si nous remplaçons p_i par sa valeur, la plus probable, nous obtiendrons pour la valeur la plus probable de l'écart :

$$\varepsilon_i = a_i - \frac{(a_i + b_i)\rho}{\rho + 1} = \frac{a_i - b_i \rho}{\rho + 1} \tag{7}$$

On aurait de même l'écart relatif à b_i

$$\varphi_i = b_i - \frac{a_i + b_i}{\rho + 1} = \frac{b_i \rho - a_i}{\rho + 1} \tag{8}$$

On remarquera que les estimations sont égales et de signe contraire. Or, si n est très grand par rapport à a_i et b_i, les événements A_i et B_i sont à peu près indépendants; l'estimation que nous choisissons pour les probabilités ne permet donc pas d'obtenir une valeur précise pour les écarts, correspondant isolément à A_i et B_i. On peut dire par contre que la différence entre la valeur théorique

$\rho = \frac{p_i}{q_i}$ et la valeur expérimentale $\frac{a_i}{b_i}$ dépend de ε_i. Nous adopterons ε_i par conséquent comme étant par définition l'écart commis sur le rapport ρ.

Si l'on a un certain nombre de rapports $\frac{p_i}{q_i}$ chaque valeur de p_i et q_i est nécessairement petite et par conséquent les quantités aléatoires a_i et b_i à peu près indépendantes; ε_i se présente donc comme la somme de deux quantités aléatoires suivant une loi normale d'écart de Gauss; ε_i suit donc la même loi. On sait que dans ces conditions son écart type sera la somme des carrés des écarts types de chaque série, mais on obtient un résultat plus simple en calculant directement, sans aucune approximation, cet écart type Q_i dont le carré est égal à la moyenne du carré de ε_i.

Posons : $a_i + b_i = c_i$, puis donnons à a_i toutes les valeurs comprises entre O et C_i et à C_i toutes les valeurs comprises entre O et n. On a :

$$Q_i^2 = \sum_{c_i=0}^{c_i=n} Pc_i \sum_{a_i=0}^{a_i=c_i} \left(a_i - \frac{c_i \rho}{\rho+1}\right)^2 Pa_i$$

en appelant Pa_i la probabilité pour que A_i se produise a_i fois lorsque l'un quelconque des événements A_i ou B_i s'est produit c_i fois et Pc_i la probabilité pour que cette dernière éventualité ait lieu.

D'autre part, la probabilité pour que A_i se produise au cours d'une épreuve sachant que A_i ou B_i a eu lieu, est :

$$p_i' = \frac{p_i}{p_i + q_i} = \frac{\rho}{\rho+1}$$

et par conséquent :

$$\sum_{a_i=0}^{a_i=c} a_i Pa_i = c_i p_i^1 = \frac{c_i \rho}{\rho+1}$$

$$\sum_{a_i=0}^{a_i=c_i} a_i^2 Pa_i = c_i^2 p_i'^2 + c_i p_i' (1 - p_i') = \frac{c_i^2 \rho^2}{(\rho+1)^2} + \frac{c_i \rho}{(\rho+1)^2}.$$

d'où :

$$\sum_{a_i=0}^{a_i=c_i} \left(a_i - \frac{c_i \rho}{\rho+1}\right)^2 = \frac{c_i^2 \rho^2}{(\rho+1)^2} + \frac{c_i \rho}{(\rho+1)^2} - \frac{2 c_i^2 \rho^2}{(\rho+1)^2} + \frac{c_i^2 \rho^2}{(\rho+1)^2} = \frac{c_i \rho}{(\rho+1)^2}$$

$$Q_i^2 = \sum_{c_i=0}^{c_i=n} \frac{c_i \rho}{(\rho+1)^2} Pc_i = \frac{n r_i \rho}{(\rho+1)^2}$$

Nous ne connaissons pas la valeur exacte de r_i, mais nous commettons une erreur aussi faible que possible en remplaçant sa valeur par son estimation (équation 6), d'où :

$$Q_i = \frac{\sqrt{(a_i + b_i)\rho}}{\rho+1}$$

Les équations 7 ou 8 deviennent alors finalement pour l'écart réduit :

$$R_i = \frac{a_i - b_i \rho}{\sqrt{(a_i + b_i)\rho}}$$

Si la relation (1) proposée est exacte, les fréquences des différentes valeurs de R_i devront se répartir suivant la formule de Gauss.

Si l'on ne dispose que d'un nombre restreint de valeurs R_i, on pourra se contenter de constater que leur valeur oscille autour de l'unité, quelques-unes seulement dépassant 2, exceptionnellement 3.

D'autre part, si m est le nombre de ces écarts, la valeur moyenne de $\frac{\Sigma R_i}{m}$ devra être voisine de O et l'écart quadratique moyen $\sqrt{\frac{\Sigma R_i^2}{m}}$ voisin de 1.

Vérification expérimentale. — Étude des classements d'empreintes digitales. — Nous rappellerons qu'il existe à l'extrémité des doigts des dessins papillaires. On peut classer ces différents dessins en un très petit nombre de types, 4 ou 5. Attribuons à chaque doigt un chiffre correspondant au type dans lequel il se classe. L'ensemble des chiffres des dix doigts forme un nombre de 10 chiffres que nous appellerons par définition « une formule digitale ». Le nombre des formules digitales existant est extrêmement grand, mais leur fréquence est très variable.

Considérons deux groupes d'individus choisis de telle manière que la caractéristique qui sépare les deux groupes soit sûrement indépendante de la formule. Par exemple, les dessins digitaux étant immuables de la naissance à la mort, considérons les individus de moins de 22 ans et ceux de plus de 22 ans.

Appelons alors A_i les individus les plus jeunes, ayant une certaine formule digitale, et B_i ceux ayant la même formule et plus âgés, de même A_2 et B_2 correspondront aux individus d'une deuxième formule mais d'âge différent.....

Soit p_1, p_2... les probabilités des événements A_1, A_2..., q_1, q_2..., celles des événements B_1, B_2,..., puisque la formule digitale et l'âge sont indépendants, les rapports $\frac{p_1}{q_1}, \frac{p_2}{q_2}, \cdots$ etc doivent être égaux entre eux et égaux au rapport du nombre total d'individus âgés de moins de 22 ans à celui des individus âgés de plus de 22 ans.

Nous avons alors choisi deux classements d'individus arrêtés sur le territoire français et classés justement d'après ces conditions, c'est-à-dire d'une part, par la formule digitale et, d'autre part, suivant l'âge. Nos investigations ont porté uniquement sur les formules digitales ne comportant que les types boucles internes et verticilles. Il y a deux alternatives possibles pour chaque doigt, soit un total de 1.024 combinaisons et nous les avons toutes utilisées pour le calcul du rapport commun des probabilités, soit 8.951 individus de moins de 22 ans et 105.996 de plus de 22 ans, mais pour le calcul des écarts, nous n'avons retenu que les formules comportant un total de dix individus au moins quelque soit l'âge. Il nous est alors resté 499 formules qui nous ont permis de calculer autant d'écarts réduits. Nous avons groupé ceux-ci par échelles de 0,1 en 0,1 et sans tenir compte du signe; nous avons ainsi obtenu le graphique de la figure 1. La courbe qui traverse ce graphique correspond à la courbe théorique.

On voit que le graphique expérimental se superpose d'une manière tout à fait satisfaisante à la courbe théorique. Ce résultat peut être considéré comme une vérification expérimentale de la méthode de calcul proposée.

Application. — Nous avons dit que les fréquences des formules digitales étaient extrêmement variables, nous nous sommes proposé de rechercher si certaines de ces fréquences n'étaient pas reliées suivant une formule du type 1.

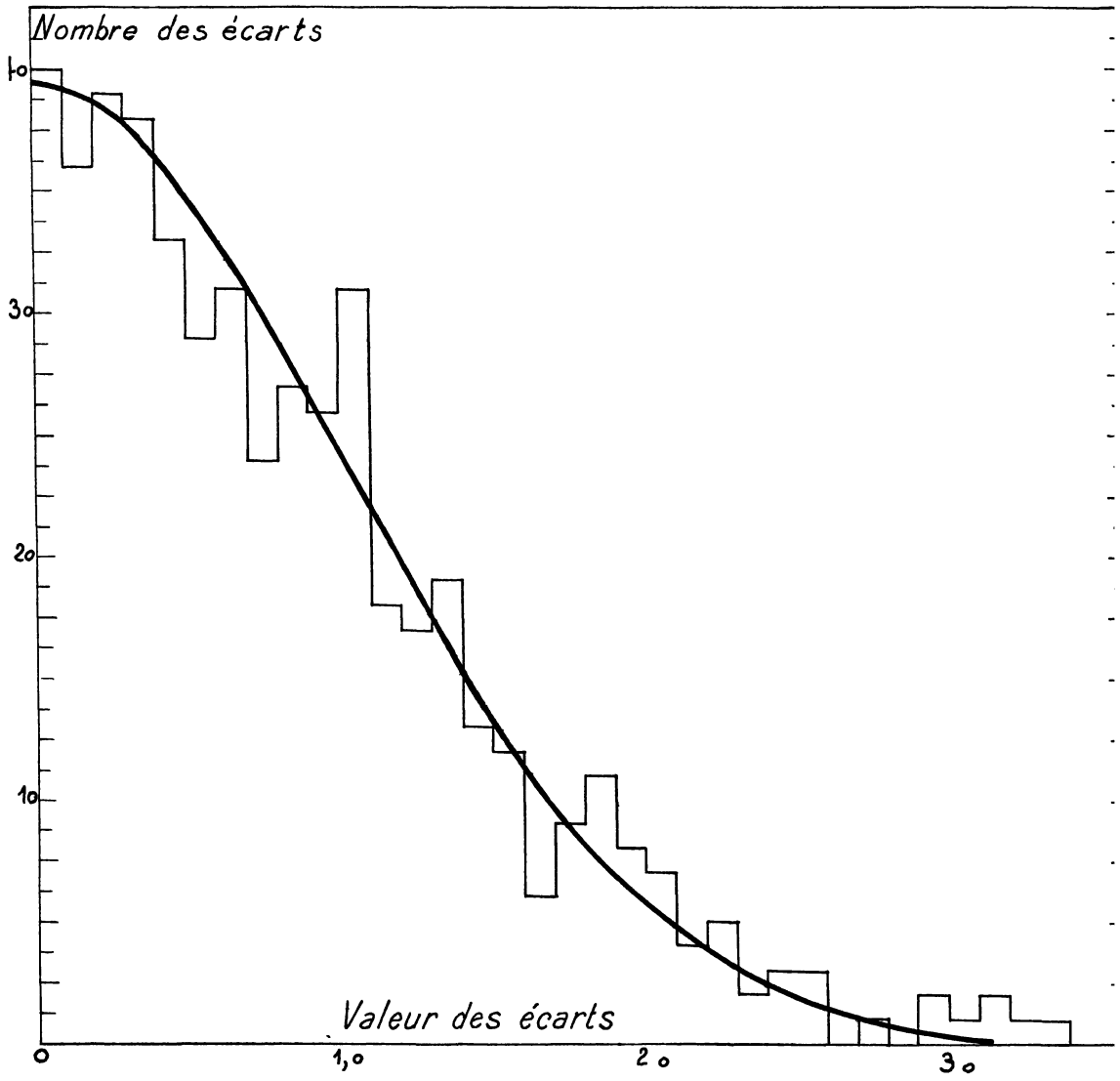

Fig. 1.

Nous dirons qu'une formule digitale est symétrique si l'on observe un dessin digital de même type à chaque paire de doigts correspondants de la main droite et de la main gauche, par exemple deux verticilles ou deux boucles internes, etc...

Nous dirons de même qu'une formule est dissymétrique pour un seul doigt si tous les autres doigts sont symétriques.

Considérons alors le tableau des 1.024 formules digitales précédentes ne comportant que des boucles internes et des verticilles, et extrayons de ce tableau les formules ne comportant de dissymétrie que sur le pouce. Dans chacune des

formules considérées, l'un des pouces est un verticille, l'autre une boucle interne. Pour chaque combinaison des quatre autres doigts, il existe deux formules correspondantes suivant que le verticille est au pouce droit ou au pouce gauche.

TABLEAU II

Formule digitale	a_i	b_i	$a_i'b_i$	Écart type	Écart réduit
IIII	2.114	879	2,45	17,4	—1,32
VIII	21	10	2,10	1,95	—1,33
IVII	613	209	2,83	10,10	+1,24
VVII.	186	79	2,35	9,3	—1,34
IIVI	5	2	2,50	0,93	—0,26
IVVI.	56	13	4,31	2,92	+2,29
VVVI.	35	9	3,89	2,33	+1,41
IIIV	762	298	2,55	11,43	—2,62
VIIV.	18	7	2,57	1,76	—0,37
IVIV.	643	228	2,82	5,31	—0,38
VVIV.	266	89	2,99	6,62	+0,68
IIVV.	85	22	3,86	3,63	+2,17
IVVV.	1,116	342	3,27	13,40	+3,77
VVVV	955	249	3,83	12,19	+6,08

Moyenne : 2,84.

Les fréquences de ces formules figurent dans les colonnes 2 et 3 du tableau II. Le groupe de quatre lettres situé dans la colonne I correspond aux formes d'empreintes des autres doigts dans l'ordre : Auriculaire, Annulaire, Médius et Index. On voit qu'il existe manifestement une relation entre les deux nombres d'une même ligne; leur rapport étant approximativement constant.

Nous nous sommes proposé de voir si les écarts entre les rapports étaient ou non compatibles avec la loi des grands nombres. Pour cela, nous avons calculé l'écart type et l'écart réduit. Ces valeurs figurent colonnes 4 et 5. Il est manifeste que les écarts réduits sont beaucoup trop élevés. Sur 14 valeurs, 5 dépassent 2, l'une d'elles atteint la valeur 6,08; enfin l'écart quadratique moyen est de 2,35.

Le rapport considéré n'est donc qu'approximatif.

Il n'y a pas lieu de s'étonner de ce résultat. Dans le classement étudié, les figures digitales n'ont été divisées qu'en 4 types. Le type verticille se trouve de ce fait englober des formes qui sont en réalité très différentes les unes des autres. Il est naturel qu'il en résulte une certaine altération des rapports de fréquences.

Nous avons alors considéré un deuxième classement dérivé du premier, mais dans lequel le type verticille était subdivisé en verticilles vrais et en formes plus complexes. Les résultats paraissent en première approximation nettement meilleurs; nous les avons étendus à toutes les formules dissymétriques pour les pouces (boucles internes et verticilles) et symétriques pour les quatre autres doigts. Nous avons ainsi obtenu les valeurs du tableau III.

On voit que, cette fois, ces derniers constituent une série satisfaisante. La moyenne des écarts réduits est en effet de 0,038 et l'écart quadratique moyen de 1,19. On peut donc dire que la formule (1) est vérifiée d'une manière aussi satisfaisante que possible. Cependant, il convient de remarquer que les calculs ont porté sur des nombres sensiblement plus petits que ceux du tableau II. La comparaison précise des écarts types montre qu'il n'y a pas à craindre une

divergence aussi grande que dans le premier cas, mais il faudrait disposer d'un nombre plus important de formules pour émettre une opinion plus précise au sujet de la relation supposée.

TABLEAU III

Formule digitale	a_i	b_i	a_i/b_i	Écart type	Écart réduit
AAAA	6	0	»	1,03	+1,36
IAAA	15	5	3,0	1,88	—0,16
IIAA	51	11	4,6	3,32	+1,12
IIIA	108	21	5,14	4,80	+1,92
IIAI	6	2	3,0	1,19	—0,08
IIII	333	111	3,0	8,90	—0,83
IVII	84	23	3,6	4,38	+0,45
VVII	24	10	2,4	2,46	—0,85
IVVI	5	2	2,5	1,11	—0,35
IIAE	14	4	3,5	1,80	+0,13
IIIE	164	41	4,00	6,03	+1,13
IVIE	37	4	9,3	2,69	+2,08
IVVE	11	4	2,7	1,63	—0,31
IIIV	103	43	2,40	5,11	—1,74
VIIV	4	2	2,0	1,03	—0,58
IVIV	59	28	2,1	3,93	—1,92
VVIV	33	7	4,7	2,66	+0,88
IIVV	7	5	1,4	1,46	—1,50
IVVV	139	47	2,96	5,76	—0,40
VVVV	114	28	4,07	5,03	+1,02
IIID	4	5	0,8	1,26	—2,31

Moyenne : 3,29.

Nous nous proposons de revenir sur cette question lorsque nous disposerons des éléments nécessaires pour cela.

En résumé, on voit que la méthode que nous proposons permet de vérifier ou d'infirmer certaines relations entre les probabilités de certains événements et éventuellement de contrôler la rectification introduite dans certaines méthodes de classement.

Lucien AMY.

DISCUSSION

Monsieur le Président, après avoir remercié M. Amy de sa très intéressante causerie, donne la parole à ceux de nos collègues qui auraient des observations à présenter ou des questions à poser.

M. BARRIOL demande si la population un peu spéciale de délinquants qui a été examinée ne présente pas des particularités telles qu'on ne pourrait pas généraliser les résultats obtenus et les appliquer à la population du pays.

M. AMY répond que la généralisation des résultats obtenus n'est pas impossible; il s'agit bien, en effet, d'une catégorie spéciale d'individus; mais, en fait, on y trouve aussi bien de simples délinquants, en grand nombre, à côté de criminels, en petit nombre.

Mlle GRANDJEAN demande à M. AMY si, pour les sélections automatiques de ses fiches pour lesquelles il y avait eu autrefois des pourparlers, il avait utilisé ses formules mathématiques comme base de codification pour les recherches des empreintes digitales ou s'il pensait les rechercher différemment. M. AMY

répond que comme il s'agissait de 1.800.000 fiches, les recherches et le classement de ces fiches étaient très compliqués. En Amérique du Sud, le fichier d'empreintes digitales est encore plus important puisque chaque individu a sa fiche à la Police Judiciaire, qu'il soit ou non délinquant.

Enfin, est-il encore possible de déterminer les dessins papillaires, lorsque les doigts présentent, au cours de l'existence d'un individu, des cicatrices laissées par des blessures, des brûlures, etc...

M. Amy répond affirmativement : grâce aux recoupements de ces dessins papillaires, on retrouve les empreintes digitales sans erreur sur l'individu qui les a fournies; des lésions très graves, des brûlures par exemple, peuvent compliquer les recherches, mais non les fausser.

M. R. Humery pose deux questions :

1° Le classement des empreintes n'a-t-il pas été fait selon des critères purement pratiques (facilités de reconnaissance et de classement) et non selon des critères biologiques. Il se peut alors que la statistique ne soit pas assise sur des bases naturelles et, par-là, que ses résultats soient faussés. On pourrait déceler cette erreur méthodique si la loi des écarts ne suit pas la loi de Gausse.

Le Conférencier répond qu'il en est bien ainsi et que, précisément, les études des écarts ont été effectuées pour mettre les chercheurs sur la piste d'un classement naturel plus satisfaisant au point de vue théorique.

2° Une science très ancienne et toujours vivace, la chiromancie, prétend établir une correspondance entre les lignes de la main et le caractère, ainsi qu'avec la destinée qui, et dans une certaine mesure, en est la conséquence. Les études dactyloscopiques ont-elles permis de déceler une corrélation entre les empreintes digitales et les caractères psycologiques, la criminalité et la déficience mentale, par exemple?

M. Amy indique que le fameux pouce des criminels n'a pas été constaté; en ce qui concerne la répartition des groupes, il a trouvé des différences sensibles pour le groupe des enfants arriérés, mais les statistiques sont encore imparfaites et on ne peut conclure; des études entreprises à l'étranger ont montré qu'elles ne donnaient aucune différence par groupe, mais malheureusement, elles proviennent de l'étude de populations hétérogènes et il est certain qu'il existe des caractéristiques de races.

M. Hibbert fait observer que le classement paraît fondé sur des formes de courbes et leurs points singuliers qui font l'objet de travaux bien connus en mathématiques. Sur chaque doigt, en effet, il y a un nombre fini de ces points singuliers, et le nombre de ces types distincts de ces points ne dépasse pas cinq. Comme il y a dix doigts et que sur chaque doigt il y a au plus cinq points singuliers dont les types distincts sont au plus au nombre de cinq, il y aurait peut-être là un moyen de trouver tous les types de mains possibles et d'attacher à chaque type dans une population de mains une probabilité.

La question posée par M. Hibbert est en effet fort intéressante, dit M. Amy, mais on se heurte à de grosses difficultés techniques en ce qui concerne la prise même des empreintes : il suffit d'une petite erreur de prise pour changer le caractère des courbes, des plages et des points singuliers; en tout cas, la question pourrait être soumise à l'analyse mathématique.

Pour terminer, le Président souligne la portée générale de la communication de M. Amy et de la discussion qui l'a suivie.

Cette communication est un très bel exemple d'application de la statistique à des phénomènes de l'ordre qualitatif. Les lignes de la main pourraient, théoriquement, être représentées par des équations de la géométrie analytique, mais cette représentation paraît assez inextricable. Au contraire, la répartition des éléments étudiés en catégories numérotées et traitées statistiquement se présente, dès à présent, comme très féconde.

Certes, une telle répartition, purement formelle, peut paraître bien arbitraire. Mais précisément, on peut trouver un critérium d'appréciation de ces catégories dans le fait qu'elles conduisent ou non à des régularités statistiques. L'étude des corrélations peut permettre, de même, d'apprécier le bien-fondé de bien d'autres classifications, ainsi que l'a montré M. Amy à propos d'un exemple de recherches des rapports pouvant exister entre les empreintes digitales et certaines malformations des individus.

C'est là, on le voit, une méthode très générale, susceptible d'intéresser tous ordres de recherches. La statistique manifeste ainsi, une fois de plus, sa fécondité, due à son caractère universel.

C'est là une raison particulière de maintenir l'activité de notre Société dans les circonstances actuelles : les applications dans un domaine pouvant être transposées à d'autres domaines très différents, nous devons maintenir, à cet égard, un étroit contact entre tous les efforts que notre pays a à fournir à l'heure actuelle.