JOURNAL DE LA SOCIÉTÉ STATISTIQUE DE PARIS

LUCIEN MARCH

# La théorie statistique et la logique formelle à propos de l'« Introduction » de M. Yule

*Journal de la société statistique de Paris*, tome 52 (1911), p. 416-426

<http://www.numdam.org/item?id=JSFS_1911__52__416_0>

## II

## LA THÉORIE STATISTIQUE ET LA LOGIQUE FORMELLE A PROPOS DE L' « INTRODUCTION » DE M. YULE (1)

Dans ses *Principii di statistica methodologica*, publiés en 1906, M. Benini considérait la statistique, successivement comme une branche de la logique et comme une forme d'observation. On pourrait, dans le même esprit, distinguer la *théorie* et la *technique* de la statistique. Or, la technique de la statistique ne diffère de la technique des autres formes d'observation que parce qu'on y applique certains principes empruntés à la théorie. Si l'on dénombre les individus possédant certain caractère, l'opération ne se distingue pas, en soi, de la mesure de la dilatation d'une barre de fer. Elle n'en diffère que par l'instrument. Lorsque l'observation statistique revêt un caractère spécial, comme lorsqu'on associe un grand nombre de mesures du même objet, ou lorsqu'on procède par sondages, ce n'est point là une simple observation ; on interprète l'observation par une généralisation logique.

L'observation, la mise en œuvre et l'interprétation des données statistiques sont donc fondées sur une théorie. L'ouvrage récent de M. Yule constitue un exposé méthodique et clair des points essentiels de cette théorie. On peut dire que cet ouvrage marque une étape : si on le compare aux excellents « Grundzüge » du professeur Westergaard parus il y a vingt ans, on peut mesurer le progrès accompli, durant cette période, dans les idées et dans les procédés.

L'ouvrage est divisé en trois parties :

1° Théorie des attributs, c'est-à-dire des propriétés dont la différenciation est regardée comme discontinue ;

2° Théorie des variables, c'est-à-dire des grandeurs considérées comme variant suivant un mode continu ou pseudo-continu ;

3° Théorie des épreuves (sampling), dont les éléments pourraient être répartis conformément aux deux divisions précédentes.

### I

Si la statistique peut être considérée comme une branche de la logique, ce n'est point, à coup sûr, la logique formelle classique qui suffirait à ordonner et à déterminer les règles dont sa méthode est charpentée.

Il faut faire appel à une logique élargie où les termes des propositions entrent avec leurs quantités, sont *numériquement définis*, suivant l'expression de de Morgan.

La supériorité de cette logique élargie sur la logique péripatéticienne, qui a régi exclusivement la pensée et la didactique pendant plus de deux mille ans, consiste, on le sait, dans la précision que comporte son expression — précision plus grande que ne le permet le langage courant — et dans la possibilité d'analyser des chaînes de propositions plus générales que les formules classiques d'identité.

Stuart Mill a remarqué avec raison que « tous les théorèmes d'Euclide pourraient être mis en séries de syllogismes, réguliers en figures et modes ». Les inférences

(1) *An Introduction to the theory of statistics*, by G. Udny Yule. Londres, Griffin et Cie, 1911.

de la statistique s'appuient sur des chaînes de propositions plus complètement déterminées et d'un caractère plus général que celles du syllogisme classique.

Une proposition (sujet et prédicat ou attribut séparés par une copule) peut être convenablement représentée par une table à double entrée, l'instrument précieux dont Pythagore a fait une application bien connue (1).

Désignons, avec M. Yule et Stanley Jevons, par A un attribut (par exemple : être vivant), par $\alpha$ l'attribut contraire (par exemple : être non vivant); de même par B et $\beta$ un autre attribut et son contraire (par exemple, périssable et immortel); puis, modifiant un peu la notation de ces auteurs, désignons par $Q_{AB}$ la quantité des objets qui possèdent à la fois l'attribut A et l'attribut B, nous formerons la table à double entrée suivante :

| ATTRIBUTS | A | $\alpha$ | TOTAUX |
|---|---|---|---|
| B | $Q_{AB}$ | $Q_{\alpha B}$ | $Q_B$ |
| $\beta$ | $Q_{A\beta}$ | $Q_{\alpha\beta}$ | $Q_\beta$ |
| Totaux. | $Q_A$ | $Q_\alpha$ | Q |

où :

$$Q = Q_A + Q_\alpha = Q_B + Q_\beta$$
$$Q_A = Q_{AB} + Q_{A\beta}$$
etc.

Cette représentation suffit pour analyser toutes les propositions de la logique classique, pour sérier les modes et les formes du syllogisme, faire reconnaître les faux syllogismes et les formes illégitimes. A représentant, par exemple, le sujet et B l'attribut dans la majeure, si l'on dit : tout A est B, cela veut dire que dans la colonne A, $Q_{A\beta} = 0$; mais, dans la rangée B, $Q_{\alpha B}$ peut conserver une certaine valeur, et par conséquent de la proposition : de tout A est B on ne peut conclure tout B est A. Tout être vivant est périssable; il n'est pas vrai que tout objet périssable est un être vivant.

Supposons maintenant un troisième attribut C (par exemple : être homme) que Q′ individus (hommes) sont susceptibles de posséder. On peut faire sur les éléments de Q′, classés d'après les attributs A et B, comme dans la table à double entrée ci-dessus, un certain nombre d'hypothèses.

Si, pour Q et pour Q′, aucun des éléments continus dans les 4 cases intérieures de la table n'est nul, cette circonstance peut se traduire ainsi :

Quelques A sont B ; Quelques A sont $\beta$.

Quelques C sont A ; Quelques C sont B, etc.

En principe, deux quelconques de ces propositions réunies n'autorisent pas de conclusion sur la relation entre C et B (2).

Si $Q_{A\beta}$ est nul, cette circonstance s'exprime dans la proposition suivante :

Tout A est B (tout être vivant est périssable).

Si l'on ajoute tout C est A (tout homme est un être vivant) cela s'exprime par $Q'_{A\beta} = 0$, toute la quantité de C correspond à B, ce qui revient à dire : tout C est B (tout homme est périssable).

---

(1) La multiplication est un rapport quantifié entre 4 termes qui représentent des choses dépourvues de toute qualité autre que le nombre.

(2) *Nil sequitur geminis e particularibus unquam.*

Telle est l'une des formes du syllogisme classique. Si l'on annule successivement l'une des grandeurs représentée par la lettre Q (ou Q') affectée d'un indice, on retrouve, par exemple, les 16 figures de syllogisme énumérées par Mill au livre III de son *Système de logique.*

Nous venons de dire avec les anciens logiciens, qu'en principe deux propositions particulières n'entraînent aucune conclusion. En réalité, cela dépend de la *quantité* des termes qui composent les propositions.

Les quatre grandeurs $Q'_{AB}$, $Q'_{A\beta}$, $Q'_{\alpha B}$, $Q'_{\alpha\beta}$ entre lesquelles se distribuent les objets possédant les caractères considérés ne sont point tout à fait indépendantes ; elles sont liées par les relations :

$$Q'_{AB} + Q'_{A\beta} = Q_A \qquad Q'_{\alpha B} + Q'_{\alpha\beta} = Q_\alpha$$
$$Q'_{AB} + Q'_{\alpha B} = Q_B \qquad Q'_{A\beta} + Q'_{\alpha\beta} = Q_\beta$$

Supposons fixées à l'avance les sommes $Q'_A$, $Q'_B$, $Q'_\alpha$, $Q'_\beta$. Elles doivent satisfaire à la condition $Q'_A + Q'_\alpha = Q'_B + Q'_\beta = Q'$, nombre des individus possédant le caractère C. S'il n'en était point ainsi, elles seraient *inconsistantes :* un certain nombre d'objets ne possèdent pas le caractère C, ou bien un certain nombre d'objets ayant ce caractère ne sont pas répartis.

Mais d'autre part, si cette condition est satisfaite, les 4 égalités précédentes n'en forment plus en réalité que 3 distinctes. On peut donc choisir arbitrairement l'une des grandeurs qui figurent dans leurs premiers membres.

En réalité, comme aucune de ces grandeurs ne peut être négative, comme aucune ne peut excéder la somme obtenue quand on lui ajoute l'une des autres grandeurs, le choix n'est point entièrement arbitraire, il comporte des limites certaines. La réunion de deux propositions particulières comporte donc une conclusion relative à ces limites.

Un raisonnement de Miss Collett, dans le compte rendu de son enquête sur le travail des femmes, fournit un bon exemple d'une conclusion de ce genre.

Supposons que A désigne le fait qu'une femme est ouvrière (α non ouvrière), B le fait qu'une femme est mariée (β non mariée). D'après le Census, sur 100 femmes de vingt à vingt-cinq ans recensées dans un certain district, 82 sont ouvrières et 26 sont mariées.

En langage syllogistique, ces constatations eussent été traduites ainsi :

Quelques femmes sont ouvrières ;

Quelques femmes sont mariées ;

Rien ne résulte de ces constatations.

Formons la table à double entrée où prendront place les chiffres précédents :

| | A | α | |
|---|---|---|---|
| B | $Q_{AB}$ | $Q_{\alpha B}$ | 26 |
| β | $Q_{A\beta}$ | $Q_{\alpha\beta}$ | 74 |
| | 82 | 18 | 100 |

Le nombre des ouvrières mariées $Q_{AB}$ ne peut être *déterminé,* mais on peut déduire de la table des limites de ce nombre. Le maximum de $Q_{AB}$ est évidemment

26 : toutes les femmes mariées seraient ouvrières. Quant au minimum, il correspond évidemment aussi à la plus grande valeur de $Q_{\alpha B}$ ou de $Q_{A\beta}$ et celle-ci correspond à la plus petite valeur de $Q_{\alpha\beta}$, soit zéro. Si $Q_{\alpha\beta} = 0$, les trois autres grandeurs de la table sont $Q_{\alpha B} = 18$, $Q_{A\beta} = 74$, $Q_{AB} = 8$.

Donc au moins 8 ouvrières sont mariées. Contrairement à la conclusion de tout à l'heure un fait certain résulte de la comparaison des données.

Les remarques précédentes fournissent un moyen de contrôler les données des tableaux statistiques, d'autant plus utile que ces données sont souvent rassemblées dans un tableau à simple entrée. Supposons, par exemple, que le Census d'un district ait fait connaître les chiffres suivants : femmes au total 100; femmes ouvrières 82, femmes mariées 6 (par suite d'une erreur qui a fait écrire 6 au lieu de 26). La simple comparaison des chiffres fait ressortir l'erreur : la *consistance* des données est en défaut.

## II

Le classement que présente la table à double entrée permet de faire apparaître sous une forme précise la *relation* qui peut exister entre les attributs d'après lesquels s'opère le classement.

On marque d'abord les états extrêmes de cette relation : absence complète de relation, ou indépendance parfaite; relation la plus étroite possible, telle que la présence de l'un des attributs entraîne nécessairement celle de l'autre, c'est-à-dire dépendance parfaite, déterminisme rigoureux. Puis l'on cherche la forme d'une grandeur qui varie entre des limites correspondant à ces cas extrêmes et dont la valeur représente en quelque sorte l'élasticité plus ou moins grande du lien qui unit les deux attributs.

Pour trouver une forme de grandeur qui mesurera la relation entre les deux attributs, la première chose à faire est de fixer sa valeur dans les cas extrêmes d'indépendance ou de dépendance parfaite.

Dans le cas d'indépendance, la présence ou l'absence de l'un des attributs doit être sans influence sur le nombre des individus qui possèdent l'autre. Si la taille des hommes est indépendante de la couleur des cheveux on doit trouver proportionnellement autant d'hommes grands parmi les bruns que parmi les blonds.

En se reportant à la table à double entrée précédente, la condition d'indépendance peut s'exprimer ainsi :

$$\frac{Q_{AB}}{Q_A} = \frac{Q_{\alpha B}}{Q_\alpha} = \frac{Q_{AB} + Q_{\alpha B}}{Q_A + Q_\alpha} = \frac{Q_B}{Q}$$

ou bien,

$$Q_{AB} = \frac{Q_A \times Q_B}{Q}$$

Telles sont les formes sous lesquelles on peut exprimer le postulat d'indépendance. Si l'on représente par $\delta$ la différence

$$Q_{AB} - \frac{Q_A Q_B}{Q} = \delta$$

on peut encore écrire dans le cas d'indépendance : $\delta = 0$.

Passons maintenant au cas de dépendance parfaite : il faut le définir.

Dans la proposition : tous les hommes sont mortels, on peut trouver le type de cette dépendance puisqu'il n'y a pas un homme qui ne soit mortel. Si A représente l'attribut homme ($\alpha$ non homme) et B l'attribut mortel ($\beta$ non mortel) la proposition précédente se traduirait ainsi, sous forme de table à double entrée :

| | | |
|---|---|---|
| $Q_{AB}$ | $Q_{\alpha B}$ | $Q_B$ |
| 0 | $Q_{\alpha\beta}$ | $Q_\beta$ |
| $Q_A$ | $Q_\alpha$ | Q |

Dans ce cas :

$$Q_{A\beta} = 0 \qquad Q_{AB} = Q_A$$

M. Yule considère que, dans ce cas, la dépendance des deux attributs est parfaite.

Comment choisir maintenant la quantité qui mesurera la relation possible entre deux attributs dans les cas intermédiaires, par exemple dans des propositions telles que celles-ci : quelques hommes à cheveux bruns ont les yeux bleus. Le critérium d'indépendance complète étant exprimé par $\delta = 0$, il est naturel de choisir comme mesure de la relation, précisément $\delta$ dont la valeur, d'abord nulle dans le cas d'indépendance, s'accroît à mesure qu'on s'écarte davantage de ce cas.

Si la dépendance est parfaite on a, d'après ce qui précède :

$$\delta = Q_{AB} - \frac{Q_A Q_B}{Q} = Q_A - \frac{Q_A Q_B}{Q} = \frac{Q_A (Q - Q_B)}{Q} = \frac{Q_A Q_\beta}{Q}$$

tel est le maximum de $\delta$.

Pour plus de commodité, on admet que l'indice de dépendance doit rester compris entre 0 et 1. Il en sera ainsi si, au lieu de prendre $\delta$ comme indice, on prend de préférence le rapport de $\delta$ à sa valeur maximum

$$\frac{\delta}{\frac{Q_A Q_\beta}{Q}} = \frac{Q\delta}{Q_A Q_\beta}$$

Cet indice s'annule dans le cas d'indépendance et devient égal à 1 dans le cas de dépendance parfaite.

Il offre toutefois un inconvénient. Au lieu de supposer $Q_{A\beta} = 0$, on aurait pu supposer aussi bien $Q_{AB} = 0$, $Q_{\alpha B} = 0$ ou $Q_{\alpha\beta} = 0$. Il est clair qu'on se trouve toujours dans le même cas de dépendance parfaite. Mais la forme de l'indice changerait dans chaque hypothèse.

Pour éviter cet inconvénient, parmi l'infinité des expressions qu'on pourrait choisir comme indice, M. Yule adopte la suivante :

$$\frac{Q\delta}{Q_{AB} Q_{\alpha\beta} + Q_{A\beta} Q_{\alpha B}}$$

Quel que soit le terme de la table à double entrée qui s'annule, cet indice devient égal à 1 et il s'annule lui-même quand $\delta = 0$.

On peut cependant faire à ce choix une objection importante.

Nous avons admis plus haut que la proposition : Tous les hommes sont mortels,

impliquait une relation tout à fait étroite entre l'état d'homme et l'état de mortel. Est-ce bien exact? La relation n'est-elle point encore plus étroite dans la proposition : Tous les êtres vivants sont mortels?

En effet, la première proposition ne peut être renversée, car il est faux que tous les mortels soient des hommes, tandis qu'on peut dire que tous les mortels sont des êtres vivants. La nécessité est unilatérale dans le premier cas, bilatérale dans le second.

La table à double entrée où A représenterait l'attribut « être vivant » et B l'attribut « mortel » serait donc composée ainsi :

| | | |
|---|---|---|
| $Q_{AB}$ | 0 | $Q_B$ |
| 0 | $Q_{\alpha\beta}$ | $Q_\beta$ |
| $Q_A$ | $Q_\alpha$ | Q |

et l'on aurait

$$Q_{AB} = Q_A = Q_B$$
$$Q_{\alpha\beta} = Q_\alpha = Q_\beta$$

d'où

$$Q_A\, Q_\beta = Q_B\, Q_\alpha$$

Dès lors, dans ce cas, l'indice que nous avions abandonné tout à l'heure $\frac{Q\delta}{Q_A\, Q_\beta}$ pourrait être remplacé par le suivant $\frac{2\,Q\,\delta}{Q_A\, Q_\beta + Q_\alpha\, Q_\beta}$. Dans le cas de dépendance parfaite, que $Q_{AB}$ et $Q_{\alpha\beta}$ soient nuls ou que les deux autres termes soient nuls, cet indice prend la valeur 1 et il s'annule naturellement quand $\delta = 0$ (1).

## III

Les cas d'indépendance parfaite ou de dépendance absolue sont des cas limites qui, en toute rigueur, ne se rencontrent jamais dans la nature. Il y a toujours des influences communes; on ne peut nier qu'entre deux phénomènes il puisse toujours exister une certaine *association*. Le degré de cette association peut être infinitésimal, c'est pratiquement le cas d'indépendance parfaite; il peut être extrêmement voisin du degré maximum conventionnel, dans le cas de dépendance pratiquement absolue. En employant l'un des indices signalés plus haut on mesure sa valeur entre les limites 0 et 1.

Remarquons d'ailleurs que la considération des attributs est elle-même, pour une

(1) L'indice adopté par M. Yule dans son chapitre sur l'Association a été critiqué par M. David Héron (Diométrika, vol. VIII) parce qu'il donne des valeurs très différentes de celles d'une autre expression dont il sera question plus loin. En fait, comme M. Yule l'a indiqué, on peut adopter une infinité d'indices donnant des valeurs différentes. Il vaudrait mieux s'en tenir à celui qui est communément adopté, mais il ne faut pourtant pas oublier que le choix de l'indice dépend avant tout du sens que l'on attache aux mots dépendance parfaite ou corrélation parfaite, ou indépendance parfaite, et de la loi de variation de l'indice que l'on adopte. Dans le précédent numéro de ce journal, M. Niceforo fait usage de l'indice élémentaire de M. Yule. Les valeurs calculées sont comparables entre elles mais elles ne le sont pas aux valeurs habituelles du coefficient de corrélation.

grande part, conventionnelle. Avec la classification dichotomique il semble que le contenu de chaque classe devrait être homogène ; or, souvent on est très loin de l'homogénéité.

L'analyse des relations entre deux attributs doit alors être poussée plus loin à l'aide de classifications plus détaillées, en décomposant chaque attribut en sous-attributs. Par exemple, au lieu que le nombre Q soit décomposé seulement en $Q_A$ et $Q_\alpha$, relativement à l'attribut A, on le décomposera en $Q_{A_1}$, $Q_{A_2}$, $Q_{A_3}$, etc., relativement aux modalités possibles de l'attribut A. De même pour l'attribut B. En sorte que la table à double entrée, au lieu de comporter seulement 2 rangées, 2 colonnes et $2 \times 2$ cases, comportera M colonnes, N rangées et MN cases. Pearson a donné à une table de ce genre le nom de *table de contingence*.

La case située à l'intersection de la colonne $m$ et de la rangée $n$ contient $Q_{A_m, B_n}$ unités, que nous écrirons simplement $Q_{m,n}$.

D'après ce qui a été dit plus haut pour la table à quatre cases, dans le cas d'indépendance parfaite, c'est-à-dire d'association ou de contingence nulle, on doit avoir pour toutes les cases telles que $(m,n)$ :

$$Q_{m,n} = \frac{Q_m \times Q_n}{Q} = q_{m,n}$$

Poursuivant l'analogie, il est dès lors naturel de prendre pour mesure de la contingence l'ensemble des quantités telles que :

$$\frac{Q_{m,n} - q_{m,n}}{q_{m,n}} = \frac{\delta}{q_{m,n}}$$

Ces quantités pouvant être positives ou négatives, leur somme serait nulle sans que chacune d'elles le soit. On est ainsi conduit à les élever préalablement au carré.

On convient de plus d'affecter chaque carré d'un poids proportionnel au contenu théorique de la case dans le cas d'indépendance parfaite, c'est-à-dire d'un poids égal à $q_{mn}$, sous réserve de rapporter le total à la somme de tous ces poids, c'est-à-dire à Q.

En définitive, on mesure le degré de contingence par l'expression

$$\Phi^2 = \Sigma \frac{\delta^2_{m,n}}{Q \times q_{m,n}}$$

que Pearson appelle le *moyen carré de contingence* et que l'on peut écrire

$$\frac{1}{Q} \Sigma \frac{\delta^2_{m,n}}{q_{m,n}} \qquad \text{ou encore} \qquad \frac{1}{Q} \Sigma \frac{Q^2_{m,n}}{q_{m,n}} - 1$$

Le cas limite de dépendance complète, de contingence parfaite ou de nécessité absolue, est réalisé quand l'un des attributs est, comme l'on dit, fonction de l'autre. Tout individu qui possède l'attribut $A_m$, possède nécessairement l'attribut $B_n$, aucun ne peut posséder l'un des attributs $B_1$ à $B_{n-1}$ ou $B_{n+1}$ et au delà. Il en résulte que la table à double entrée ne comporte qu'un seul nombre par colonne ou par rangée et, comme l'ordre des colonnes et des rangées est arbitraire, on peut supposer que

les nombres sont disposés sur une diagonale de la table, celle qui va du coin supérieur de gauche au coin inférieur de droite. D'ailleurs, comme à toute valeur de A correspond une valeur de B et inversement, le nombre des colonnes est alors égal au nombre des rangées. Que devient, dans ce cas, la quantité

$$\Phi^2 = \frac{1}{Q} \Sigma \frac{Q_{m,n}^2}{q_{m,n}} - 1 = \Sigma \frac{Q_{m,n}^2}{Q_m Q_n} - 1$$

Puisque chaque colonne ou chaque rangée ne contient qu'un nombre, par exemple, $Q_{m,n}$ à l'intersection de la colonne de rang $m$ et de la rangée de rang $n$, il est clair que $Q_{m,n} = Q_m = Q_n$. Donc, dans le cas de dépendance parfaite, s'il y a N colonnes et autant de rangées, $\Phi^2$ devient égal à N — 1.

D'après cela, la valeur limite de $\Phi^2$ augmente à mesure que la classification est poussée plus loin. Pour la commodité des comparaisons, on substitue à $\Phi^2$ l'expression $\frac{\Phi^2}{1 + \Phi^2} = C^2$, C étant appelé *coefficient de contingence*. Cette quantité C varie entre 0 et 1, quand on passe du cas d'indépendance ou de liberté parfaite des attributs comparés, aux cas de contingence et finalement au cas de dépendance ou de nécessité absolue.

## IV

On peut imaginer que les divisions de la table à double entrée soient rendues de plus en plus nombreuses. C'est ce qui arrivera par exemple, si l'on différencie les degrés possibles d'un même attribut à l'aide d'instruments de plus en plus perfectionnés et précis. A la limite on conçoit des divisions tellement nombreuses que beaucoup de cases restent vides et qu'aucune ne contienne plus d'une unité.

Dans ce cas, le postulat d'indépendance ne peut plus conserver la même forme. D'abord l'ordre des rangées et des colonnes ne peut plus être arbitraire. S'il n'y a plus d'unités accumulées dans certaines cases plutôt que dans telle autre, la disposition de cases contenant chacune une seule unité n'est qu'un chaos, tant que ces cases ne comportent pas un certain ordre. Si elles comportent un certain ordre, c'est alors cet ordre qui fixe la dépendance ou l'indépendance des deux attributs. On caractérise cet ordre par les distances rectangulaires du centre de chaque case à deux axes $ox$, $oy$ menés parallèlement aux côtés de la table, à partir du centre de gravité des cases occupées; celles-ci sont d'ailleurs toutes d'égal poids puisque chacune ne contient qu'une unité.

Considérons maintenant les ordonnées telles que $y_n$ des cases occupées et leurs abcisses telles que $x_m$ ; $x_m$ et $y_n$ sont les deux coordonnées de l'une de ces cases par rapport aux axes moyens.

On peut convenir de ne pas tenir compte de la grandeur relative, soit des $x$ entre eux, soit des $y$ entre eux, de façon à n'envisager que la relation entre les $x$ et les $y$ indépendamment de leurs grandeurs moyennes.

Dans ce but, au lieu de comparer chaque $x$ à chaque $y$, on comparera leurs rapports $\frac{x}{\sigma}$ et $\frac{y}{\sigma}$ à certaines moyennes $\sigma,\sigma'$. Désignons par X et Y ces rapports, c'est-à-dire les valeurs de $x$ et de $y$ quand on les mesure à l'aide des unités $\sigma$ et $\sigma'$.

Avec ces unités, $X_m$ est la quantité dont la position de la case $(m, n)$ dépasse la

moyenne dans le sens horizontal et $Y_n$ la quantité dont la position de la même case $(m, n)$ excède la moyenne dans le sens vertical. Comme je l'ai indiqué dans un précédent article (1), l'expression $r = \frac{1}{Q} \Sigma X Y$ fait connaître le nombre moyen des cases pour lequel les deux excédents sont de même sens, et concordent par conséquent, déduction faite des cases pour lesquelles les deux excédents sont de sens contraire. De plus $r$ tient compte du poids de ces excédents; Pearson lui a donné le nom de *coefficient de corrélation*. M. Yule obtient la valeur de $r$ en appliquant la méthode des moindres carrés.

On voit que $r$ s'annule quand les excédents de même sens donnent une somme de produits égale à celle des excédents de sens contraire. La distribution des cases occupées comporte alors une certaine symétrie par rapport aux axes moyens.

A première vue, il ne paraît pas (2) que le postulat de l'indépendance parfaite soit le même dans l'application du coefficient de corrélation et dans l'application du coefficient de contingence. Le postulat de dépendance parfaite est au contraire exactement le même.

Dans le cas particulier de corrélation normale, les deux coefficients de corrélation et de contingence sont identiques. Dans le cas de décomposition dichotomique où le nombre des cases se réduit à quatre, M. Yule montre que le coefficient de corrélation $r$ est égal au rapport $\frac{Q\delta}{\sqrt{Q_A Q_\alpha Q_B Q_\beta}}$ (avec la notation utilisée au début de cet article). Cette formule ne diffère de l'expression signalée au même endroit $\frac{2Q\delta}{Q_A Q_\beta + Q_\alpha Q_B}$ que par la substitution de la moyenne géométrique à la moyenne arithmétique des quantités $Q_A Q_\beta$ et $Q_\alpha Q_B$. On peut avoir des raisons de préférer l'une des formes à l'autre.

Quoi qu'il en soit, l'emploi de ces coefficients permet de déterminer avec quelque précision l'étroitesse du lien qui unit deux phénomènes, lorsque les manifestations de ces phénomènes sont classées, soit par degrés qualitatifs, soit par degrés quantitatifs.

Sous des réserves dont se pénétreront ceux qui liront avec soin l'ouvrage de M. Yule, ces coefficients contrôlent les relations causales qui peuvent exister entre des séries de faits et ils mettent sur la voie de la découverte de liens insoupçonnés.

## V

Les mêmes méthodes peuvent s'appliquer à l'étude des relations de plusieurs attributs. On pourrait encore représenter une proposition comprenant un sujet et deux attributs, au moyen d'un stéréogramme à triple entrée. Au delà de trois termes il faut se contenter de la représentation par des lettres. Celle-ci aide déjà beaucoup à s'orienter dans le dédale des classifications. Or, une classifica-

(1) *Journal de la Société de Statistique de Paris*, numéro de juillet, 1905, page 269.

(2) Pourtant il conviendrait d'étudier la question plus à fond dans les travaux de Pearson.

tion détaillée et complète est indispensable pour autoriser des conclusions précises de propositions particulières où les termes sont donnés avec leurs quantités.

Par exemple, supposons que l'on étudie dans un groupe Q d'enfants anormaux l'influence de l'état nerveux et de la débilité physique sur l'intelligence. Représentons par A l'attribut « nerveux », par B l'attribut « chétif », par C l'attribut « esprit borné ». Si l'observation fait connaître $Q_A$, $Q_B$, $Q_C$, $Q_{AB}$, $Q_{\beta C}$, on peut fixer des limites à $Q_{\delta C}$, $Q_{B\gamma}$, etc. ; on peut déterminer s'il existe une certaine *association*, et à quel degré, entre deux des attributs et le troisième.

Toutefois, la multiplicité des attributs crée des embûches auxquelles il faut prendre garde. Par exemple, M. Yule montre que deux attributs peuvent être indépendants l'un de l'autre dans une catégorie d'individus qui tous possèdent un troisième caractère, être indépendants aussi dans la catégorie des individus qui ne possèdent pas le troisième caractère, et cependant, dans le groupe total des individus des deux catégories, il peut se révéler une liaison des deux premiers attributs. Il suffit que le caractère distinctif des deux catégories dépende lui-même de l'un au moins des deux premiers attributs considérés.

A ce propos, on peut signaler une classe de problèmes dont l'auteur ne s'est point occupé, et qui ne manque cependant pas d'intérêt. Bien avant que la théorie de la corrélation et celle de la contingence fussent exposées sous une forme générale par le professeur Pearson, notre collègue, M. Jacques Bertillon, avait appliqué une méthode analogue à celle de la contingence pour étudier l'influence de la confession religieuse sur la nuptialité (1). Il se demandait comment il convenait d'apprécier la fréquence relative des mariages mixtes.

Soient

| | | | |
|---|---|---|---|
| A | représentant le caractère | | homme |
| $\alpha$ | — | — | femme |
| B | — | — | catholique |
| $\beta$ | — | — | non catholique |

L'état des mariages conclus et des mariables étant donné par les deux tables suivantes :

| Mariages : | $\alpha$B | $\alpha\beta$ | |
|---|---|---|---|
| AB | $x_1$ | $y_1$ | $M_1$ |
| A$\beta$ | $x_2$ | $y_2$ | $M_2$ |
| | X | Y | M |

| Mariables : | B | $\beta$ | |
|---|---|---|---|
| A | $u_1$ | $v_1$ | $N_1$ |
| $\alpha$ | $u_2$ | $v_2$ | $N_2$ |
| | U | V | N |

M. Bertillon a pris avec raison pour mesure de la puissance des antipathies religieuses le rapport $x_2 : \frac{N_2 \times V}{N}$ au lieu du rapport $x_2 : \frac{XM_2}{M}$ et il a eu pleinement raison.

(1) *Annales de démographie internationale*, année 1882.

## VI

Je me suis arrêté trop longuement sur la partie la plus nouvelle des recherches dont rend compte l'ouvrage de M. Yule pour pouvoir donner une idée de son exposé des procédés mathématiques appliqués au traitement des observations.

Je signalerai seulement la détermination, à l'aide des coefficients de corrélation, de l'écart type d'une fonction linéaire de variables, quand les variables ne sont point indépendantes (p. 208). Je signalerais bien aussi les modes élémentaires suivant lesquels sont établies les propriétés essentielles d'où l'on peut déduire les théorèmes de Bernouilli et de Poisson (p. 252 et 278), mais les étudiants ne trouveront pas là, à mon sens, de véritables démonstrations.

J'insisterai plutôt sur la quantité de procédés ingénieux par lesquels des choses assez ardues sont rendues aisément accessibles, sur le grand nombre d'exemples bien choisis, sur la variété des exercices qui permettront aux étudiants de bien pénétrer le sens et la portée des théories.

Celles-ci se recommandent surtout par leur généralité qui fait de la logique statistique la partie la plus étendue de la logique formelle, mais la logique statistique est surtout indispensable au statisticien pour l'obliger à raisonner avec méthode et précision.

Lucien MARCH.