JOURNAL DE LA SOCIÉTÉ STATISTIQUE DE PARIS

P. THIONET

Note sur quelques fins et moyens de la méthode statistique

Journal de la société statistique de Paris, tome 109 (1968), p. 130-135

<http://www.numdam.org/item?id=JSFS_1968__109__130_0>

NOTE SUR QUELQUES FINS ET MOYENS DE LA MÉTHODE STATISTIQUE

L'auteur de cette note a peu de goût pour les discussions philosophiques ou simplement trop générales. S'il lit, par exemple, dans son journal, un article sur l'agrégation de l'enseignement du second degré, il est agacé par le fait que M. G. pense à celle de philosophie, alors que M. B. pense à celle des lettres; et même si M. G. et M. A. pensent l'un et l'autre à l'agrégation de philosophie, ils peuvent avoir tous deux raison, quand l'un est pour et l'autre contre.

Il ne voudrait donc pas s'engager dans une discussion sur ce qu'est et ce que n'est pas la Statistique, bien qu'il s'agisse d'un mot auquel, dans ce journal même, il est donné bien des sens; sans oublier qu'on lui donne une toute autre signification s'il est au pluriel, ou encore accompagné de l'adjectif « quantique » par exemple.

Il lui paraît au fond très naturel qu'on se fasse de la Statistique une conception fortement marquée (voire déformée) par la formation qu'on a reçue (mathématique notamment) et les professions successives qu'on a exercées. Et les pires ennemis de l'outil mathématique, en Statistique, sont parfois des hommes qui, l'ayant beaucoup manié autour de leur vingtième année, ont totalement cessé de le faire ensuite (ils ont « tourné la page »). Ils sont finalement exaspérés par l'évolution, les transformations survenues là comme ailleurs, parce qu'ils ne sont pas tenus au courant, convaincus qu'ils étaient de l'éternité des mathématiques. Qu'ils se rassurent : les mathématiciens de profession ont les mêmes difficultés dès qu'ils sortent des petits domaines à l'intérieur desquels ils ont travaillé (ce qui ne représente, parfois, que quelques coins perdus dans l'immensité des mathématiques).

Il s'agit ici, simplement, de donner un exemple de problème statistique. Il concerne un sujet biologique, ce qui offre l'intérêt de concerner tout le monde (qui donc oserait, de nos jours, se désintéresser de la médecine?). Bien mieux, il se rapporte à la leucémie, question assez propre à passionner le public. Nous nous pencherons sur une communication de Kimball au Congrès de Stockholm (1957) de l'Institut international de Statistique[1]. En fait, il traite d'un problème qui n'a rien de proprement médical et dont mes amis démographes (médecins ou non) ont fait le tour depuis longtemps : l'appréciation correcte des causes de décès.

Il ne sera pas demandé ici au statisticien de contribuer directement à une décision d'ordre médical, par exemple de juger si tel traitement est plus efficace que tel autre (on ima-

1. M.-L. Dufrénoy, Maupertuis et le progrès scientifique, Studies on Voltaire and the eighteenth century. *Trans. first Int. Congress on the Enlightment*, pp. 519-587, Genève, 1963.
2. J. Dufrénoy, Le Centenaire du Mendélisme, 1865-1965, *Rev. Path. Comp.*, 1965.

gine aisément les heurts qui se produiraient, si c'était le cas, entre statisticiens et médecins). Au fond, c'est un simple problème de *présentation des données*, de la façon la plus parlante; mais cette traduction (des données numériques brutes en données élaborées) peut être une *trahison* et il arrive, bien sûr, qu'on fasse dire aux chiffres bien des choses qu'ils n'ont jamais voulu dire.

Voyons de près l'exemple donné par Kimball : Répartition de 40 décès de souris suivant la cause et le temps. Toutes sont atteintes de leucémie.

TABLEAU I

Répartition de 40 décès de souris suivant la cause et l'époque

	Périodes				Total
	1	2	3	4	
Causes de décès :					
1.	2	3	5	6	16
2.	7	6	2	1	16
3.	1	1	3	3	8
	10	10	10	10	40

Les causes 2 et 3 sont les leucémies A et B (mais il n'importe guère).

La cause 1 représente les causes diverses de décès par accident.

Le tableau fait ressortir (marge verticale) que la cause 2 tuerait beaucoup plus de malades que la cause 3. Toutefois, si l'on tient compte des changements dans le temps, on constate des évolutions inverses. On peut les expliquer comme suit :

Les accidents (cause 1) surviennent aussi bien à des souris bien portantes qu'à des malades (leucémies A B), et, au fur et à mesure, ils suppriment des malades qui, normalement, auraient décédé en raison des causes 2 ou 3.

Cette superposition des causes de décès (1-2 et 1-3) a pour conséquence une répartition globale (toutes époques) qui ne signifie plus rien.

Les calculs de Kimball ont pour résultat de mettre en évidence que, sur 100 cas de décès de souris par leucémie, on peut imputer le décès à la leucémie A dans 53,5 % des cas contre 46,5 % à la leucémie B. Compte tenu des petits effectifs concernés, on peut conclure que :

— Il n'y a pas de différence réelle entre les pourcentages de décès par leucémie A et leucémie B; on peut, en ce qui les concerne, accepter l'hypothèse 50 %-50 %.

C'est là un résultat des calculs qui contredit les données brutes (16-8) d'après lesquelles la leucémie A semble tuer deux fois plus que la leucémie B.

Méthode de calcul

La méthode de calcul de Kimball nous paraît anormalement compliquée. Il se produit là un fait bien connu pour les chaînes de Markov : le calcul matriciel conduit à de notables simplifications *a posteriori*. C'est ainsi qu'un ouvrage sur les chaînes de Markov écrit avant 1940 (sans faire usage du calcul matriciel), nous paraît illisible aujourd'hui.

Nous imaginerons une séquence 0, 1, 2, 3, 4 de dates successives. Le zéro désigne la date de départ, où 40 souris sont vivantes. Il n'en reste que 30 à la date 1, 20 à la date 2, 10 à la

date 3 et 0 à la date 4. A chaque date, nous distinguons 4 états possibles : O vivant, j mort par cause j (avec $j = 1.2.3.$), d'où une matrice de transition (ou de passage) 4×4.

Ainsi, on postule l'existence de probabilités de passage d'un état à un autre qui soient les mêmes pour chaque souris, mais non à chaque date. On a ainsi des matrices M_{01} M_{12} M_{23} et M_{34} caractérisant les changements d'état d'une souris de la date 0 à la date 1, puis 2, puis 3, puis 4.

Bien entendu, une souris morte ne ressuscite pas; le passage de l'état 0 aux états 1, 2, 3, jouit seul de probabilités non nulles. Ces probabilités sont estimées par la méthode du maximum de vraisemblance, au moyen des seules données du tableau I.

Voici le résultat de ces « estimations »;

$$M_{01} = \frac{1}{40}\begin{bmatrix} 30 & 2 & 7 & 1 \\ 0 & 40 & 0 & 0 \\ 0 & 0 & 40 & 0 \\ 0 & 0 & 0 & 40 \end{bmatrix}; \qquad M_{12} = \frac{1}{30}\begin{bmatrix} 20 & 3 & 6 & 1 \\ 0 & 20 & 0 & 0 \\ 0 & 0 & 20 & 0 \\ 0 & 0 & 0 & 20 \end{bmatrix}$$

$$M_{23} = \frac{1}{20}\begin{bmatrix} 10 & 5 & 2 & 3 \\ 0 & 10 & 0 & 0 \\ 0 & 0 & 10 & 0 \\ 0 & 0 & 0 & 10 \end{bmatrix}; \qquad M_{34} = \frac{1}{10}\begin{bmatrix} 0 & 6 & 1 & 3 \\ 0 & 10 & 0 & 0 \\ 0 & 0 & 10 & 0 \\ 0 & 0 & 0 & 10 \end{bmatrix}$$

Nous renvoyons à Kimball pour le calcul détaillé d'estimation. Nous constatons que les probabilités de passage traduisent tout bêtement le tableau I. Par exemple, examinons la matrice M_{01}.

Les lignes (portant les numéros 0, 1, 2, 3) se réfèrent à l'état occupé à la date 0 par la souris; les colonnes : à l'état occupé à la date 1. Dire que la case ($i = 0, j = 0$) est occupée par la probabilité 30/40 signifie que 30 des 40 souris survivent; la souris prise au hasard a donc une probabilité 30/40 de rester à l'état O (: vivante).

Elle a des probabilités 2/40 de décéder de cause 1, 7/40 de cause 2, et 1/40 de cause 3 (Transformation en probabilités des fréquences indiquées au tableau I).

$$\begin{array}{c} j = \quad 0 \quad 1 \quad 2 \quad 3 \\ \begin{array}{l} i = 0 \\ i = 1 \\ i = 2 \\ i = 3 \end{array} \begin{bmatrix} \frac{30}{40} & \frac{2}{40} & \frac{7}{40} & \frac{1}{40} \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} = M_{01} \end{array}$$

Les lignes $i = 1, 2, 3$, représentent ce qui se passe pour une souris décédée; elle a la probabilité 1 de rester dans son « état », une probabilité 0 de le quitter.

Les matrices M_{12} M_{13} M_{34} s'interprètent de façon analogue.

On effectue alors le produit matriciel :

$$M_{01} \qquad M_{12} \qquad M_{23} \qquad M_{34}$$

On voit qu'il fournit des probabilités (marge verticale du tableau I) 0, $\frac{16}{40}$, $\frac{16}{40}$ et $\frac{8}{40}$ que la souris vivante à la date 0 se trouve à la date 4 : 0/vivante, 1/décédée par cause 1, 2/par cause 2, 3/par cause 3.

$$M_{01} \cdot M_{12} \cdot M_{23} M_{34} = \frac{1}{240\,000} \begin{bmatrix} 0 & 96\,000 & 96\,000 & 48\,000 \\ 0 & 240\,000 & 0 & 0 \\ 0 & 0 & 240\,000 & 0 \\ 0 & 0 & 0 & 240\,000 \end{bmatrix} = \begin{bmatrix} 0 & \frac{16}{40} & \frac{16}{40} & \frac{8}{40} \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

Conséquence :

Ce calcul conduit à l'application suivante : on peut, à présent, éliminer la cause 1 de décès, c'est-à-dire les décès accidentels, et regarder ce qu'il se passe si les Matrices de passage sont privées des lignes et colonnes 1, conservant les lignes et colonnes 0, 2, 3. Désignons-les par :

$$M^*_{01} \qquad M^*_{12} \qquad M^*_{23} \qquad M^*_{34}$$

Faisons à nouveau le produit des 4 matrices, il vient :

$$\begin{bmatrix} \frac{30}{38} & \frac{7}{38} & \frac{1}{38} \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} \frac{20}{27} & \frac{6}{27} & \frac{1}{27} \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} \frac{10}{15} & \frac{2}{15} & \frac{3}{15} \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} 0 & \frac{1}{4} & \frac{3}{4} \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 0 & \frac{32\,940}{61\,560} & \frac{28\,620}{61\,560} \\ 0 & 1 & 0 \\ 0 & 1 & 0 \end{bmatrix}$$

Nota : Bien entendu, on a « normé » les probabilités de passage en remplaçant les dénominateurs respectivement par $40 - 2 = 38$, $30 - 3 = 27$, $20 - 5 = 15$, $10 - 6 = 4$, c'est-à-dire en retranchant de 40, 30, 20, 10 le nombre des décès par cause 1 éliminés du calcul.

Le résultat est $\frac{32\,940}{61\,560} = 53{,}5\ \%$ $\frac{28\,620}{61\,560} = 46{,}5\ \%$

proportions déjà annoncées et jugées (sans autre calcul) en fait indiscernables de 50 %.

Commentaire :

1. On aurait pu imaginer que les souris décédées du fait de la cause 1 restaient vivantes. Le calcul donne des résultats tout à fait différents :

$$\begin{bmatrix} \frac{32}{40} & \frac{7}{40} & \frac{1}{40} \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} \frac{23}{30} & \frac{6}{30} & \frac{1}{30} \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \frac{15}{20} & \frac{2}{20} & \frac{3}{20} \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \frac{6}{10} & \frac{1}{10} & \frac{3}{10} \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} \frac{66\,240}{249\,000} & \frac{106\,160}{240\,000} & \frac{67\,600}{240\,000} \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

Résultats : il y a plus d'un quart de survivants.
La cause 2 de décès l'emporte assez nettement sur la cause 3.

2. Signalons que le produit matriciel fournit (automatiquement) le résultat du théorème des probabilités composées (comme dans les chaînes de Markov; ici, il s'agit d'une chaîne non homogène, c'est-à-dire que la matrice M se modifie dans le temps). Le calcul de Kimball consiste, lui aussi, en l'application de ce théorème et aboutit aux mêmes résultats que le nôtre.

3. Le calcul ne donne de résultat valable que dans la mesure où la *formalisation* est acceptable, c'est-à-dire où l'on peut parler de probabilité de décès entre 2 dates, suivant la cause 1, la cause 2, ou la cause 3. C'est au biologiste (ou au médecin) de juger si ce concept a un sens. Par exemple, il exclut toute épidémie. Les calculs faits en éliminant la cause 1 doivent être correctement interprétés :

Le premier suppose que toutes les souris sont malades et que la cause 1 n'a pour effet qu'une mort prématurée; mais les causes 2 et 3 s'excluent.

Le second calcul suppose que les causes 1, 2 et 3 s'excluent; il ne paraît guère valable pour la leucémie; il serait fort valable pour des causes de décès 2 et 3, qui ne comporteraient pas une longue période d'incubation et une issue fatale.

Conclusions

1. L'exemple précédent était destiné à montrer que certains moyens modernes de mathématique peuvent avoir pour effet de clarifier des calculs dont l'aspect embrouillé décourage. La discussion reste possible, non sur le calcul lui-même, mais sur la formalisation qui lui a permis de naître. Elle est donc rendue plus facile pour le non-mathématicien (le biologiste).

2. Le résultat du calcul est de mettre en lumière le fait reconnu que la leucémie B a une part aussi grande que la leucémie A dans les décès. C'est parce qu'elle ne tue guère dans les débuts du mal que tous les malades éliminés avant l'issue fatale par quelque cause accidentelle échappent à la statistique des causes de décès.

3. Le risque demeure que le chercheur refuse de se plier à la technique matricielle et à la formalisation de ses concepts, préférant interpréter — ou modifier — les statistiques suivant sa propre subjectivité.

Il peut même tout accepter, sauf les valeurs des probabilités de passage : adepte des probabilités subjectives, il ne se sentira pas tenu de les estimer (par le maximum de vraisemblance ou autrement) à l'aide des seules données du tableau I; s'il est « Bayesien », il peut calculer d'autres probabilités de passage (et je suis entièrement d'accord car, après tout, les données expérimentales sont peu nombreuses pour estimer des probabilités).

Le néo-bayesien aborderait enfin le même problème avec des idées *a priori* sur ce qu'il pourrait en *coûter de se tromper* en acceptant ou en refusant à la cause de décès 3 une importance dissimulée par les statistiques. Pour ma part, je n'aime pas beaucoup ce point de vue, qui mélange trop les rôles du statisticien et du biologiste; mais il a certainement du bon dans bien des situations réalistes.

* * *

Il est à craindre que le lecteur n'ait été un peu dérouté par cet exposé; l'économiste ne se sent pas concerné; le démographe est à peine rassuré parce qu'il est question de décès de souris et non d'hommes; le médecin est rebuté par le calcul matriciel; le mathématicien lui-même (comme l'éléphant) a peur de ces souris. (Si notre ami Depoid ne nous avait si tragiquement quittés, je pense qu'il rirait de bon cœur ici.)

Nous avons surtout voulu faire sentir qu'il y a loin entre les mythes et la réalité. Le mythe des données statistiques compilées et recopiées avec commentaires prudents, opposé au mythe des statistiques auxquelles on peut faire dire ce qu'on veut, doivent l'un et l'autre s'effacer devant des analyses variées, s'appuyant sur des formalisations. Le mythe de la « mathématique rigide » fournissant un indice statistique correct à 5 décimales et dénué de sens réel doit s'effacer de l'esprit de nos économistes (confondant mathématicien et machine à calculer), comme cet autre mythe de la mathématique n'abordant jamais aucun problème concret.

Il est vrai que nous n'avons envisagé qu'un seul aspect de la statistique : *l'analyse des données*. Il y en a bien d'autres. C'est d'ailleurs pourquoi, en France, tant de gens nient l'existence même de la statistique, tant elle peut changer de visage.

Considérez une science à ses débuts, en train de passer du qualitatif au quantitatif, par exemple la sociométrie, la psychométrie. On ne peut tout de même pas prétendre que tous les gens qui font passer des tests font de la statistique. Pourtant, il serait profondément inexact de ne pas voir les liens qu'ils ont avec la statistique, et, précisément, la mathématique statistique.

C'est que cette mathématique peut leur apporter des moyens d'étalonner leurs observations, d'en faire quelque chose d'un peu comparable aux mesures des sciences physiques. Là aussi, il ne faut pas se payer de mots, et un exemple précis éclairera notre lanterne. Les sociologues étudient (disons) les petits groupes sociaux et leurs structures; dans certains, les individus sont bien intégrés, dans d'autres, il existe beaucoup d'isolés. Certaines expériences permettent de dénombrer les isolés, et, par suite, de caractériser le degré d'intégration du groupe.

Exemple : Soit une classe de lycée de jeunes gens. On demande à chacun de donner 3 noms de camarades avec qui ils désirent être « binomés » pour les manipulations. On évite tout conciliabule, toute tractation : on constate dans la classe l'existence *d'isolés*, c'est-à-dire de jeunes dont le nom n'a jamais été indiqué par personne.

Le problème se pose alors de savoir s'il y a beaucoup, ou peu, d'isolés dans cette classe. Un moyen banal consiste à comparer entre eux les résultats d'expériences analogues. Un moyen mathématique consiste à calculer la distribution du nombre d'isolés, dans une classe de N personnes, sachant que chacune désigne des noms pris au hasard dans la liste des N-1 autres. Cette distribution de référence n'a, bien entendu, aucun rapport avec les distributions statistiques obtenues en rassemblant les résultats de nombreuses expériences. Elle fournit cependant un nombre moyen d'isolés, sorte de zéro pour une échelle de l'isolement, et toute une graduation en probabilités, constituant une véritable échelle de mesure.

Ce problème de calcul des probabilités a été résolu vers 1952 par Léo Katz [2]; le nombre moyen avait été calculé par Lazarsfeld dès 1938. Katz utilise, notamment, des résultats de notre maître et ancien président, M. le professeur Fréchet. Il reste que, pour l'instant, ce sont des algébristes (Guilbaud, Barbut) et non des probabilistes qui, en France, se passionnent pour la sociométrie.

P. THIONET

Références

[1] KIMBALL (A. W.). — Disease incidence estimation in populations subject to multiple causes of deaths. *Bull. Inst. Intern. Statis.*, 1957, (papier n° 18).

[2] KATZ (Leo). — The distribution of the number of isolates in a social group. *Annals of Math. Statistics*, 23, 1952, 271-276.