REVUE DE STATISTIQUE APPLIQUÉE

B. CYFFERS

Analyse discriminatoire

Revue de statistique appliquée, tome 13, n° 2 (1965), p. 29-46

<http://www.numdam.org/item?id=RSA_1965__13_2_29_0>

ANALYSE DISCRIMINATOIRE

B. CYFFERS

Ingénieur en Chef des Manufactures de l'État

A - L'ANALYSE DISCRIMINATOIRE

1 - DOMAINE D'APPLICATION

L'analyse discriminatoire trouve son application lorsqu'on est en présence d'une population statistique constituée de plusieurs familles telles que le caractère qui différencie les familles n'est pas directement apparent, et que l'on se propose d'affecter un individu de la population à la famille à laquelle il appartient.

Le domaine d'application de l'analyse discriminatoire peut paraître restreint et, cependant, elle constitue un moyen d'investigation qui, s'il n'a pas reçu jusqu'ici beaucoup d'applications pratiques, n'en semble pas moins susceptible de rendre des services dans de nombreux cas :

- lorsque la reconnaissance du caractère qui définit une famille est délicate ou entièrement subjective.

Exemple : Une population de cigarettes peut être divisée en trois familles : cigarettes de bonne compacité, cigarettes de compacité forte et cigarettes de compacité faible.

- lorsque la mesure du caractère qui définit la famille est très longue et onéreuse ; on cherche alors à identifier la famille à l'aide de caractéristiques plus facilement accessibles.

- lorsque l'identification de la famille à laquelle appartient l'individu entraîne la destruction de l'individu.

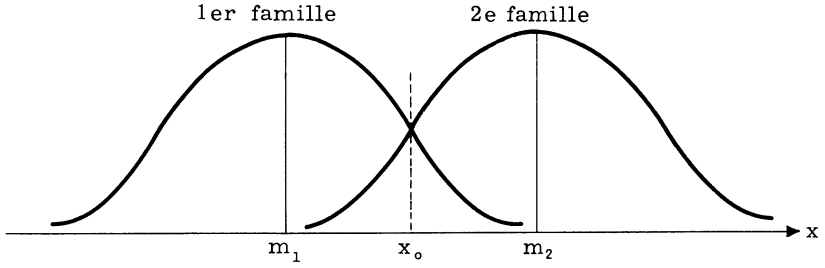
Exemple : savoir si une orange contient des pépins ou non.

- lorsqu'une décision doit être prise alors qu'il est encore impossible de connaître la famille. C'est le cas du sexe des oeufs : on a intérêt à vendre les oeufs mâles et à mettre couver les oeufs femelles. C'est également le cas de certains diagnostics médicaux. Ainsi lorsqu'un malade rentre à l'hôpotal et est reconnu hépatique, doit-on le soigner par voie médicale ou par voie chirurgicale ?

2 - NOTION DE DISCRIMINATION

Considérons le cas le plus simple où la population ne comprend que deux familles.

Soit X la caractéristique mesurable sur chacun des individus, et supposons que la loi de distribution des valeurs x de X à l'intérieur de chaque famille est normale, avec même écart-type σ, mais de moyenne m_1 dans la première famille et m_2 dans la seconde.



Si sur un individu quelconque de la population on ne procède qu'à la mesure de la caractéristique X, la connaissance de la valeur observée x contient toute l'information disponible relative à la famille à laquelle appartient l'individu. Pour identifier cette famille, nous ne pouvons adopter qu'une seule position, à savoir : déterminer une valeur x_o et conclure

si $x < x_o$ l'individu appartient à la première famille.

si $x > x_o$ l'individu appartient à la deuxième famille.

Si, par exemple, on sait que les deux familles sont d'égale importance, on est conduit à prendre $x_o = \frac{m_1 + m_2}{2}$ de manière à ce que le risque de se tromper soit le même quelle que soit la famille à laquelle appartient l'individu. Avec les hypothèses énoncées ce risque est :

$$1 - F\left(\frac{m_2 - x_o}{\sigma}\right) = 1 - F\left(\frac{m_2 - m_1}{2\sigma}\right)$$

où F(u) représente la fonction intégrale de la loi normale.

On constate, ce qui est d'ailleurs assez intuitif, que ce risque est d'autant plus faible que $\frac{m_2 - m_1}{\sigma}$ est grand.

Désignant par d la distance des moyennes ($d = m_2 - m_1$) la discrimination des familles est d'autant meilleure que le rapport $\frac{d}{\sigma}$ est grand.

Supposons maintenant que nous procédions à la mesure de deux caractères mesurables X_1 et X_2 sur chacun des individus et que les lois de distribution de chacune des variables x_1 et x_2 soient normales à l'intérieur de chaque famille

	Famille 1		Famille 2	
x_1	Moyenne	$m_1^{(1)}$	Moyenne	$m_1^{(2)}$
	Ecart-type	σ_1	Ecart-type	σ_1
x_2	Moyenne	$m_2^{(1)}$	Moyenne	$m_2^{(2)}$
	Ecart-type	σ_2	Ecart-type	σ_2

La discrimination réalisée par x_1 est caractérisée par

$$\frac{m_1^{(2)} - m_1^{(1)}}{\sigma_1} = \frac{d_1}{\sigma_1}$$

La discrimination réalisée par x_2 est caractérisée par

$$\frac{m_2^{(2)} - m_2^{(1)}}{\sigma_2} = \frac{d_2}{\sigma_2}$$

Si $\frac{d_2}{\sigma_2} > \frac{d_1}{\sigma_1}$ nous dirons que la variable x_2 assure une meilleure discrimination que la variable x_1.

3 - FONCTION DISCRIMINANTE

Lorsqu'on a mesuré x_1 et x_2 sur le même individu, l'information relative à la famille de cet individu est contenue à la fois dans x_1 et x_2. Or, toute fonction linéaire de x_1 et x_2 est distribuée à l'intérieur des deux familles suivant des lois normales de même écart-type mais de moyennes différentes.

Désignons par $Y = \lambda_1 x_1 + \lambda_2 x_2$ la fonction linéaire de x_1 et x_2 pour laquelle la différence des moyennes, mesurée en écart-type, c'est-à-dire $\frac{Y^{(2)} - Y^{(1)}}{\sigma_Y}$ est maximum. De toutes les fonctions linéaires, Y est celle qui assure la meilleure discrimination, et cette discrimination ne peut pas être inférieure à celle réalisée par x_1 ou x_2 séparément.

Y est appelé fonction discriminante.

4 - L'ANALYSE DISCRIMINATOIRE

Les considérations qui précèdent nous permettent maintenant de donner la définition suivante de l'analyse discriminatoire :

Soit une population d'individus constituée de k familles distinctes les unes des autres par un certain caractère mesurable ou non. Sur chaque individu on procède à la mesure d'un certain nombre de paramètres x_1, $x_2 \ldots x_p$ L'analyse discriminante établit la ou les fonctions linéaires des x,

$$Y = \lambda_1 x_1 + \lambda_2 x_2 + \ldots + \lambda_p x_p$$
$$Y' = \lambda'_1 x_1 + \lambda'_2 x_2 + \ldots + \lambda'_p x_p$$
$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$

dont la connaissance permet d'identifier avec le minimum de chances d'erreurs la famille à laquelle appartient un individu extrait de la population.

Cette ou ces fonctions sont appelées fonctions discriminantes.

Examinons successivement plusieurs cas :

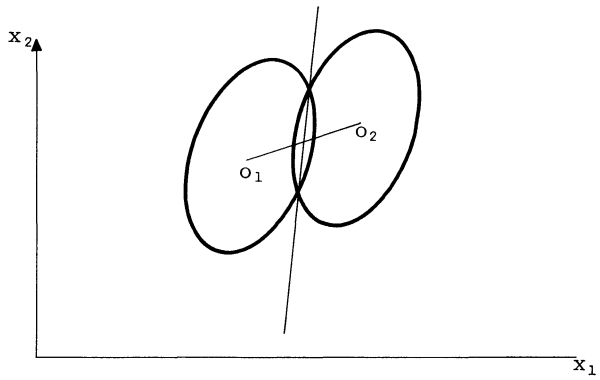
a) Cas de deux familles dépendant de deux paramètres

C'est bien entendu le cas le plus simple.

A l'intérieur de chaque famille les deux variables x_1 et x_2 sont distribuées en loi normale à deux dimensions, dont les ellipses indicatrices dans le plan $x_1 0 x_2$ se déduisent l'une de l'autre par une translation. Nous établirons que la fonction discriminante $Y = \lambda_1 x_1 + \lambda_2 x_2$ définit la direction conjuguée de la ligne des centres par rapport à ces ellipses indicatrices. Ce résultat n'est pas surprenant. Pour employer un langage imagé, supposons que nous déplaçions notre oeil sur la droite de l'infini du plan $x_1 0 x_2$, c'est lorsque nous regarderons les deux ellipses à partir du point correspondant à la direction conjuguée de la ligne des centres que nous les verrons séparées au maximum l'une de l'autre. Le calcul de la fonction discriminante ne présente donc pas de grande difficulté. Il reste ensuite à calculer la valeur Y_o qui permettra d'affecter un individu à la famille 1 ou 2 selon que la valeur de Y observée sur cet individu sera inférieure ou supérieure à Y_o. On peut alors tracer la droite :

$$Y_o = \lambda_1 x_1 + \lambda_2 x_2$$

qui divise le plan en deux régions. Si la probabilité a priori qu'un individu appartienne à l'une ou l'autre famille est identique, cette droite passera par le milieu du segment joignant les centres des ellipses.



b) Cas de deux familles dépendant de p paramètres

Soient $x_1, x_2, \ldots x_p$ les paramètres. La loi de distribution de $x_1, x_2 \ldots x_p$ à l'intérieur de chaque famille est représentée géométriquement dans l'espace à p dimensions par un hyperellipsoïde. La fonction discriminante représente la multiplicité linéaire à p - 1 dimensions conjuguée de la ligne des centres par rapport à ces hyperellipsoïdes. Son calcul ne peut pratiquement être effectué qu'à l'aide de machines électrocomptables dès que $p > 4$. Il reste à choisir la valeur limite Y_o comme dans le cas précédent.

c) Cas de plusieurs familles

Lorsqu'on est en présence de plusieurs familles le problème reste simple lorsque les centres des différentes familles sont alignés. Il n'y a alors qu'une seule fonction discriminante qui définit la direction conjuguée ou la multiplicité linéaire conjuguée de la ligne des centres par rapport aux ellipses ou hyperellipsoïdes indicateurs. Si k est le nombre de familles on doit fixer k - 1 valeurs $Y_o^1 \ldots Y_o^{k-1}$, on peut ainsi "découper l'espace en tranches parallèles" et on affecte l'individu à une famille selon la position du point expérimental.

Lorsque les centres ne sont pas alignés, la théorie devient plus compliquée, la discrimination entre familles nécessite l'emploi de plus d'une fonction discriminante. Par exemple, dans le cas de trois familles, dont les centres sont nécessairement dans un plan, c'est-à-dire une multiplicité linéaire à deux dimensions, il y a deux fonctions discriminantes. Plus généralement, le nombre de fonctions discriminantes est égal au nombre de dimensions de la multiplicité linéaire contenant les centres .

L'utilisation de ces fonctions devient, elle aussi, plus délicate.

B - UN EXEMPLE D'APPLICATION LE DIAGNOSTIC DES ICTERES(1)

Parmi les épreuves biologiques dites du "complet hépatique" figure l'analyse électrophorétique des protéines du sérum. Cette analyse, dont il ne paraît pas nécessaire de décrire la technique dans cet article, fournit les renseignements chiffrés suivants :

A	=	masse absolue	d'albumine	par litre	de sérum
α_1	=	"	de globuline α_1	"	"
α_2	=	"	de globuline α_2	"	"
β	=	"	de globuline β	"	"
γ	=	"	de globuline γ	"	"

On obtient également des quantités δ et ε qui sont des anomalies dues à la méthode de mesure. La quantité totale de protéines P' est en réalité la somme

$$P' = A + \alpha_1 + \alpha_2 + \beta + \gamma ,$$

mais on a l'habitude de la mesurer par :

$$P = A + \alpha_1 + \alpha_2 + \beta + \gamma + \delta + \varepsilon$$

La population statistique étudiée est celle des hépatiques, que l'on divise tout d'abord en deux familles "médicaux" c'est-à-dire malades atteints d'hépatite et "chirurgicaux", cette seconde famille se divisant à son tour en deux sous-familles "Calculs" et "Cancers". Il existe d'autres ictères chirurgicaux, mais de fréquence beaucoup plus faible.

---

(1) L'application de l'Analyse discriminatoire au diagnostic différentiel des ictères a déjà fait l'objet des publications suivantes :

- Discrimination entre ictères médicaux et chirurgicaux à partir des résultats de l'analyse électrophorétique des protéines du sérum par A. Charbonnier, B. Cyffers, D. Schwartz et A. Vessereau
  Communication présentée par M. A. Vessereau à la 29ème Session de l'Institut International de Statistique - Rio de Janeiro 1953
- Discrimination entre ictères médicaux et chirurgicaux à partir des résultats de l'analyse électrophorétique des protéines du sérum, des mêmes auteurs.
  Revue International d'Hépatologie - Tome V n° 7 - 1955
- Intérêt de l'Analyse discriminatoire dans l'interprétation des Résultats de l'électrophorèse pour le diagnostic différentiel des Ictères par rétention, des mêmes auteurs, dans La Presse médicale 64ème Année, n° 89, 8 Décembre 1956.

  Le présent article est publié avec l'accord du Docteur Charbonnier, de M. D. Schwartz et de M. A. Vessereau.

Le problème le plus important pour le médecin est de déterminer si le malade doit être opéré ou non, c'est-à-dire s'il est "chirurgical" ou "médical", car il peut être très grave d'opérer un malade qui n'aurait pas dû l'être, ou de ne pas opérer un malade qui aurait dû l'être.

L'examen des résultats de l'analyse électrophorétique des protéines du sérum montre qu'ils apportent une information sur la famille à laquelle appartient le malade. Les médecins avaient énoncé quatre lois, telles que :

"Toute baisse des α_2 globulines au-dessous de leur valeur normale est en faveur d'un ictère par hépatite".

grâce auxquelles ils pouvaient définir environ 2 ictères sur 5 avec un risque d'erreur de 8.5 % du nombre total des ictères, soit 20 % des cas où un diagnostic a été énoncé.

Ils se sont alors demandé s'il n'était pas possible de tirer une meilleure information des données de l'électrophorèse, et c'est pourquoi ils ont fait appel aux statisticiens.

L'analyse discriminatoire s'imposait pour résoudre le problème soumis.

L'étude a porté tout d'abord sur un groupe de 197 malades et a comporté les phases suivantes

a) choix des variables
b) calcul des fonctions discriminantes
c) application au classement de tous les malades
d) application au classement des malades avec possibilité de diagnostic réservé.

Ensuite, les résultats ont été appliqués à un groupe de 81 nouveaux malades.

I - Etude préliminaire sur un 1er groupe de 197 malades

a) Choix des variables. Pour chacun des 197 malades visés, dont l'ictère a pu être rigoureusement identifié (il s'agit de malades traités et suivis à l'hôpital) on possède les résultats d'analyse électrophorétique.

Valeur de $A + \alpha_1$, α_2 , β , γ , P

En effet, la méthode permet mal de séparer A et α_1

Il n'est pas évident a priori que les variables brutes sont les mieux appropriées à discriminer les différentes familles d'hépatiques : certaines fonctions des variables (le logarithme par exemple) peuvent conduire à une meilleure séparation des familles, soit parce que leur dispersion est plus faible que celle de la variable brute, soit parce que cette dispersion dépend moins de la valeur moyenne.

C'est la raison pour laquelle on a envisagé l'emploi des variables suivantes :

- variables brutes $A + \alpha_1$, α_2, β, γ.

- logarithme des variables. On constate en effet que, pour l'une quelconque des variables de l'électrophorèse, le coefficient de variation σ/m reste, d'une famille de malades à l'autre, plus constant que l'écart-type σ.

- variables exprimées en % de leur total $P' = (A + \alpha_1) + (\alpha_2) + (\beta) + (\gamma)$.

On constate que l'écart-type σ ou le coefficient de variation σ/m restent d'une famille de malades à une autre, plus constants, lorsqu'on les calcule à partir du rapport x/P' que lorsqu'on les calcule à partir de la variable brute x. Ce fait est mis en évidence par le tableau ci-dessous, qui donne les valeurs du coefficient de variation σ/m, d'une part pour la famille des hépatites, d'autre part pour celle des cancers:

	Hépatite	Cancer
$A + \alpha_1$	0.19	0.23
α_2	0.36	0.52
β	0.34	0.29
γ	0.28	0.34

	Hépatite	Cancer
$A + \alpha_1/P'$	0.16	0.18
α_2/P'	0.35	0.47
β/P'	0.28	0.25
γ/P'	0.24	0.23

Il est évident d'autre part qu'une erreur éventuelle de dilution, qui fait varier à la fois x et P', influe beaucoup moins que le rapport x/P'. L'adoption de rapports au lieu de données brutes garantit, dans une certaine mesure, contre la variabilité de l'appareillage ou de la technique.

Pour juger si l'une ou l'autre des variables de l'électrophorèse est influencée de façon significative par la nature de la maladie, on a appliqué le test t de Student-Fisher à la comparaison des moyennes constatées pour cette variable dans les familles "hépatite" et "cancer". On suppose l'égalité des variances dans les deux familles comparées ; cette condition n'est ici réalisée que de façon approximative.

Les valeurs obtenues pour t sont les suivantes :

	Variable brute x	log x	x/P'
$A + \alpha_1$	3.7	3.3	6.1
α_2	7.2	8.7	7.6
β	7.2	5.7	7.6
γ	2.2	1.7	4.6

Ce tableau appelle les commentaires suivants :

1/ A une exception près (log x pour la variable γ) toutes les valeurs de t sont hautement significatives : il y a, de façon quasi-certaine, des différences de valeur moyenne entre hépatite et cancer, pour chacune des variables de l'électrophorèse.

2/ pour une variable électrophorétique déterminée, c'est généralement le rapport x/P' qui donne la valeur la plus élevée de t : ce rapport sépare mieux les deux familles que les autres variables envisagées.

3/ La variable qui donne la moins bonne séparation (valeur de t la plus petite) est la variable γ.

Compte tenu de ces remarques, on adoptera, pour la suite de l'étude statistique :

— d'une part les rapport $A + \alpha_1/P'$, α_2/P', β/P'

— d'autre part la quantité P, masse totale des protéines.

b) Calcul des fonctions discriminantes. Dans le groupe des 197 malades, les valeurs moyennes des variables précédentes sont, par nature de maladie, les suivantes :

	Hépatite (106 malades)	Calcul (41 malades)	Cancer (50 malades)	Calcul + Cancer (91 malades)
$A + \alpha_1/P'$	46.93	46.13	39.82	42.66
α_2/P'	6.25	8.79	11.82	10.45
β/P'	15.86	17.56	21.13	19.52
P	73.51	77.45	77.82	77.65

Les écarts-types ont été calculés sur le groupe des "médicaux" (hépatite) et des "chirurgicaux" (calcul + cancer) ; les écarts-types moyens, pour l'ensemble de ces deux familles de malades ont permis de calculer la discrimination réalisée par chacune des variables retenues.

	Entre Médicaux et Chirurgicaux		
Variable	Différence des moyennes	Ecart-type	Discrimination
$A + \alpha_1/P'$	4.27	7.18	0.595
α_2/P'	- 4.20	3.66	1.148
β/P'	- 3.66	4.97	0.736
P	- 4.14	10.36	0.404

On a calculé deux fonctions discriminantes :

la 1ère Y_1, en vue de séparer Médicaux et Chirurgicaux

la 2ème Y_2, en vue de séparer, à l'intérieur des Chirurgicaux, les Calculs et les Cancers.

Il convient de noter que l'on n'applique pas, dans le cas présent, la méthode générale de discrimination entre 3 familles distinctes, mais que l'on traite, compte tenu de la manière dont se pose le problème au médecin, la discrimination entre deux familles une première fois entre Médicaux et Chirurgicaux, et une deuxième fois pour les malades reconnus chirurgicaux par application de la première discrimination.

Discrimination "médicaux", "chirurgicaux"

$$Y_1 = \frac{A + \alpha_1}{P'} + 12.025\,\frac{\alpha 2}{P'} + 4.513\,\frac{\beta}{P'} + 1.771\ P$$

La valeur moyenne de cette fonction est :

pour le groupe des "médicaux"..... 330.554

pour le groupe des "chirurgicaux"..... 401.476

la différence des moyennes est..... d_1 = 70.922

l'écart-type de Y est..... σ_1 = 51.171

le rapport de discrimination est d_1/σ_1 = 1.386

On vérifie que la discrimination réalisée par Y_1 est meilleure que celle réalisée par la meilleure variable isolée, en l'occurence α_2/P' (1,386 contre 1,148)

Les coefficients de α_2/P', β/P' et P dans la fonction discriminante sont hautement significatifs. Par contre le coefficient d'A + α_1/P' n'est pas significatif ; on pourrait donc, sans nuire sensiblement à la puissance de la discrimination, supprimer cette variable, mais cela obligerait à recalculer les coefficients des autres variables.

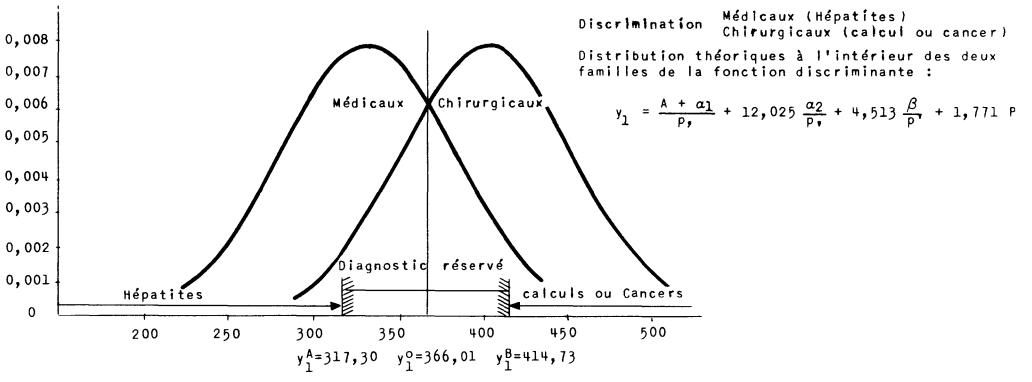
Discrimination "calcul", "cancer" :

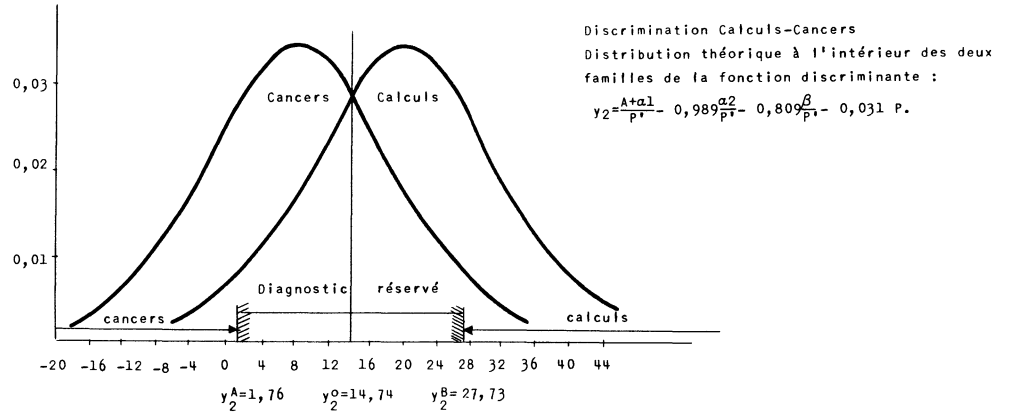
$$Y_2 = \frac{A + \alpha_1}{P'} - 0.989 \frac{\alpha_2}{P'} - 0.809 \frac{\beta}{P'} - 0.031 \; P$$

La valeur moyenne de cette fonction est :

pour le groupe "calcul"	20.847
pour le groupe "cancer"	8.643
La différence des moyennes est d_2 =	12.204
L'écart-type de Y_2 est σ_2 =	11.603
Le rapport de discrimination est d_2/σ_2 =	1.052

Les coefficients des trois premières variables, dans la fonction discriminante, sont peu significatifs ; le coefficient de la quatrième (P)





ne l'est pas du tout. Ces constatations, ainsi que la faible valeur du rapport d_2/α_2, laissent prévoir que la discrimination entre calculs et cancers sera médiocre.

Les distributions théoriques de Y_1 et Y_2 à l'intérieur des familles de malades sont représentées sur un graphique ci-dessus.

c) Application au classement de tous les malades

Supposons tout d'abord que l'on veuille énoncer un diagnostic sur chaque malade. On décidera à l'aide de Y_1 s'il s'agit d'un ictère médical ou chirurgical, et dans cette seconde éventualité on choisira, grâce à Y_2, entre "calcul" et "cancer".

Cette manière de poser le problème conduit à déterminer deux "valeurs de séparation" Y_1° et Y_2° et à décider :

si $Y_1 < Y_1^\circ$ Hépatite

si $Y_1 > Y_1^\circ$ { avec $Y_2 > Y_2^\circ$ Calcul ; avec $Y_2 < Y_2^\circ$ Cancer

Le choix de Y_1° et Y_2° dépend de la proportion de malades dans chacune des trois familles, c'est-à-dire de la probabilité a priori pour un malade d'être atteint de l'un ou l'autre ictère. Il est difficile de déterminer ces probabilités a priori : elles peuvent changer dans le temps (épidemie) et sont probablement variables d'un hôpital à l'autre.

Dans le cas présent, en nous basant sur les nombres d'Hépatites (106), de Calculs (41) et de Cancers (50) observés au cours d'une même période, nous avons admis les probabilités a priori suivantes :

Hépatites 1/2 Calculs 1/4 Cancers 1/4

Ce qui conduit à prendre pour Y_1° et Y_2° la moyenne arithmétique des valeurs moyennes de Y_1° et Y_2° à l'intérieur des familles, soit :

$$Y_1^\circ = 366,015 \qquad Y_2^\circ = 14,745$$

Y_1° et Y_2° sont, sur le graphique, les abscisses des points d'intersection des deux courbes.

1/ Diagnostic "médical" ou "chirurgical" (fonction Y_1) - Risques d'erreur.

Les erreurs possibles dans l'utilisation de Y_1, consistent à déclarer médical un ictère chirurgical, ou à déclarer chirurgical un ictère médical. Avec le choix qui a été fait pour Y_1°, ces deux risques sont égaux, tout au moins théoriquement. Il sont représentés sur le graphique par les aires délimitées par les courbes, l'axe des abscisses, et la droite d'abscisse Y_1°. Par exemple, le risque (c'est-à-dire la probabilité) de déclarer chirurgical un ictère médical est représenté par l'aire comprise à droite de la droite de séparation (366,015), entre la courbe relative aux hépatites et l'axe des abscisses. Cette probabilité peut être évaluée à l'aide de la table de la fonction F(u), intégrale de la loi de Laplace-Gauss.

On trouve ainsi, pour la probabilité de déclarer chirurgical un ictère médical :

$$1 - \left[\frac{Y_1^\circ - \overline{Y}_1 \text{ (Hépatite)}}{\sigma_{Y_1}} \right] = 0,2441$$

soit : 24,41 %

Autrement dit, 24,41 % des ictères médicaux seront classés en ictères chirurgicaux par l'utilisation de Y_1, et de même 24.41 % des ictères chirurgicaux seront classés en ictères médicaux. Comme on a admis qu'il existait autant d'ictères médicaux que d'ictères chirurgicaux, on se trompera, en moyenne, dans 24.41 % des cas.

Il est intéressant de comparer cette proportion à celle que l'on obtient pour chacune des variables de l'électrophorèse prise séparément. Le tableau ci-dessous indique les % théoriques d'erreur selon que l'on utilise l'une ou l'autre de ces variables.

Y_1	$A + \alpha_1/P'$	α_2/P'	β/P'	P
24.41 %	38.67 %	28.27 %	35.68 %	42.07 %

On constate que Y_1 améliore sensiblement la discrimination effectuée par la mesure la plus sélective, à savoir α_2/P'.

Vérification :

Sur les 197 malades étudiés dans ce § (qui ont servi à l'établissement de la fonction discriminante) :

parmi les 106 hépatites,
15 sont classés par Y_1 en ictères chirurgicaux

parmi les 91 ictères chirurgicaux,
31 sont classés en hépatites

soit au total : 46 erreurs sur 197 malades, c'est-à-dire 23.35 %, proportion très voisine de la proportion théorique.

2/ Diagnostic "calcul" ou "cancer" (fonction Y_2) - Risques d'erreur.

Nous donnons ci-dessous les % d'erreurs qu'entraînent l'utilisation de Y_2 ou de chacune des variables électrophorétiques prise isolément, lorsqu'on veut séparer en Calculs et en Cancers les malades dont on est déjà certain qu'ils sont atteints de l'une ou l'autre maladie.

Y_2	$A + \alpha_1/P'$	α_2/P'	β/P'	P
30.00 %	31.56 %	37.26 %	37.03 %	49.32 %

Y_2 n'améliore que de très peu la discrimination effectuée par $A + \alpha_1/P'$ considéré isolément.

Vérification :

Sur les 91 malades "chirurgicaux" de l'étude actuelle :

parmi les 41 calculs, 12 sont classés par Y_2 en Cancers,
parmi les 50 cancers, 16 sont classés en Calculs

soit au total 28 erreurs sur 91 malades, c'est-à-dire 30.77 %, proportion très voisine de la proportion théorique.

3/ Diagnostic par utilisation successive de Y_1 et Y_2 - Risques d'erreur

Les risques d'erreur dans l'utilisation successive de Y_1 et Y_2 sont les suivants (en admettant les probabilités a priori 1/2, 1/4 et 1/4 pour un malade d'être atteint d'hépatite, de calcul ou de cancer) :

pour les hépatites : $1/2 \times 0.2441$
pour les calculs : $(1/4 \times 0.2441) + (1/4 \times 0.7559 \times 0.3000)$
pour les cancers : $(1/4 \times 0.2441) + (1/4 \times 0.7559 \times 0.3000)$

soit au total : 0.35750

La proportion théorique d'erreurs est donc de 35.75 %

Il est à noter que le risque d'erreur dans l'utilisation successive de Y_1 et Y_2 est supérieur à ceux rencontrés dans l'utilisation séparée de Y_1 ou de Y_2. Ceci provient de ce que, dans la pratique, lorsqu'on utilise Y_2, on ignore si les malades sont réellement atteints d'ictères chirurgicaux.

Vérification :

Pour les mêmes 197 malades, le tableau ci-dessous donne dans chaque case le nombre d'individus atteints de la maladie correspondant à la ligne, classés par les fonctions discriminantes dans la maladie correspondant à la colonne.

Maladie réelle \ Classement par Y_1 et Y_2	Hépatite	Calcul	Cancer	Total
Hépatite.....	91	8	7	106
Calcul.....	19	13	9	41
Cancer.....	12	7	31	50
Total.....	122	28	47	197

Le nombre total d'erreurs s'obtient en ajoutant les chiffres des cases n'appartenant pas à la diagonale principale, soit 62 sur 197 malades. La proportion observée est de 62/197 = 31.47 %, légèrement inférieure à la proportion théorique.

d) Application avec possibilité d'un diagnostic réservé.

L'utilisation des fonctions discriminantes exposée ci-dessus permet de connaître la proportion d'erreurs commises lorsqu'on veut prononcer un diagnostic sur tous les malades à partir des résultats de l'électrophorèse, et uniquement à partir de ces résultats. La bonne concordance entre la proportion d'erreur théorique, et la proportion observée sur les 197 malades ayant servi à l'étude apporte une confirmation de la validité de la méthode et des hypothèses faites.

Mais, on vient de le constater, les risques d'erreur sont loin d'être négligeables. Il paraît donc plus raisonnable de ne pas chercher à énoncer un diagnostic sur tous les malades. On se donnera à priori la proportion d'erreurs à ne pas dépasser (proportion suffisamment faible, par exemple 5 %), et, compte tenu de cette condition, on n'énoncera un diagnostic que sur une partie des malades. Pour les autres, le diagnostic sera "réservé", et remis à la connaissance du résultat d'autres investigations médicales.

Il convient donc de déterminer, pour une probabilité d'erreur donnée à l'avance (nous prendrons 5 %), la proportion des cas où un diagnostic pourra effectivement être prononcé.

1/ Diagnostic "médical" ou "chirurgical" (fonction discriminante Y_1)

Reportons-nous au graphique représentant les distributions de Y_1, au sein des familles "médical" et "chirurgical".

Il s'agit de délimiter sur l'axe des Y un segment $(Y_1^A - Y_1^B)$ tel qu'un diagnostic ne sera prononcé que lorsque la valeur de Y_1 déduite des mesures effectuées sur un individu sera extérieure à ce segment.

En acceptant le risque d'erreur 5 %, Y_1^B est défini par l'équation :

$$F\left[\frac{Y_1^B - \overline{Y}_1 \text{ (Hépatite)}}{\sigma_{Y_1}}\right] = 0.95$$

d'où

$$Y_1^B - \overline{Y}_1 \text{ (Hépatite)} = 1.645\ \sigma_{y_1}$$

De même Y_1^A est donnée par :

$$\overline{Y}_1 \text{ (chirurgical)} - Y_1^A = 1\ 645\ \sigma_{y_1}$$

On trouve

$$Y_1^A = 317,30 \qquad Y_1^B = 414,73$$

La proportion de malades atteints d'hépatite, dont les Y_1 sont extérieurs au segment $(Y_1^A - Y_1^B)$ est de 5 % $(Y_1 > Y_1^B)$ plus 39,78 % $(Y_1 < Y_1^A)$, soit au total 44.78 %.

Cette proportion est la même pour les malades de la famille "chirurgical".

La fonction discriminante Y_1 permet donc de dire si un malade est atteint d'un ictère médical ou d'un ictère chirurgical dans 45 % des cas environ, avec une proportion d'erreur de 5 %. Ce dernier chiffre est rapporté au nombre total des malades. Rapportée au nombre de malades pour lesquels un diagnostic est énoncé, la proportion d'erreurs est de 5/45, soit 1/9.

Remarquons que les valeurs de Y_1^A et Y_1^B sont indépendantes des probabilités à priori pour qu'un malade appartienne à l'une ou l'autre famille et ne dépendent que du risque d'erreur consenti.

Vérification :

Le tableau ci-dessous donne, pour les 197 malades de l'etude, le classement effectué par Y dans les trois catégories : diagnostic non prononcé, hépatite, calcul ou cancer.

Classement par Y_1 / Maladie réelle	Diagnostic non prononcé	Hépatite	Calcul ou Cancer	Total
Hépatite.....	61	43	2	106
Calcul.....	28	6	7	41
Cancer.....	27	1	22	50
Total.....	116	50	31	197

Nombre de diagnostics énoncés : 81 sur 197, soit 41. 12 %
Nombre d'erreurs..... : 9 sur 197, soit 4. 56 %

ou par rapport au nombre de diagnostics prononcés : 9/81 = 1/9

2/ Diagnostic "calcul" ou "cancer" (fonction discriminante Y_2)

Partant cette fois de la fonction Y_2, on calcule, de la même manière que précédemment, les valeurs Y_2^A et Y_2^B correspondant à une probabilité d'erreur de 5 %.

On trouve : $Y_2^A = 1.76$ $Y_2^B = 27.73$

Le pourcentage de malades pour lesquels le diagnostic est prononcé est de 32. 66 %. La proportion d'erreurs, par rapport au nombre de diagnostics prononcés, est de 5/32. 66 soit près de 1/6.

Vérification :

Sur les 91 malades de l'étude, atteints réellement de calcul ou de cancer, le classement obtenu par l'emploi de la fonction Y_2 est le suivant :

Classement par Y_2 / Maladie réelle	Diagnostic non prononcé	Calcul	Cancer	Total
Calcul.....	30	10	1	41
Cancer.....	32	2	16	50
Total.....	62	12	17	91

Nombre de diagnostics énoncés : 29 sur 91, soit 31, 87 %
Nombre d'erreurs..... : 3 sur 91, soit 3. 3 %

ou par rapport au nombre de diagnostics prononcés 3/29 # 1/10

3/ Diagnoctic par utilisation successive de Y_1 et Y_2

L'utilisation successive des fonctions Y_1 et Y_2 conduit à donner pour chaque malade une des 5 réponses suivantes : diagnostic non prononcé, hépatite, calcul, cancer, calcul ou cancer. C'est, semble-t-il, dans l'état actuel de nos connaissances, la meilleure façon d'utiliser les fonctions discriminantes aux fins d'un diagnostic.

Le calcul théorique de la proportion d'erreurs et de la proportion de diagnostics énoncés, calcul qui fait intervenir la corrélation des quantités Y_1 et Y_2 n'a pas été effectué. Il est certain que la proportion d'erreurs (par rapport à l'ensemble des malades) est supérieure à ce qu'elle est dans l'utilisation de Y_1 seul ou de Y_2 seul. En effet, l'utilisation de Y_2 sur les malades déjà classés par Y_1 dans la famille "calcul" ou "cancer" ne peut qu'augmenter le nombre des erreurs déjà commises lors de l'utilisation de Y_1.

L'application de la règle shcématisée ci-dessous

$Y_1 < 317.30$		Hépatite
$Y_1 > 414.73$	avec $Y_2 > 27.73$	Calcul
	avec $1.76 < Y_2 < 27.73$	Calcul ou Cancer
	avec $Y_2 < 1.76$	Cancer

aux 197 malades de l'étude donne les résultats suivants :

Maladie réelle \ Classement par Y_1 et Y_2	Diagnostic non prononcé	Hépatite	Calcul	Cancer	Calcul ou Cancer	Total
Hépatite.....	61	43	-	1	1	106
Calcul.....	28	6	-	1	6	41
Cancer.....	27	1	-	21	1	50
Total.....	116	50	0	23	8	197

Le nombre de malades classés est évidemment de 81 comme dans l'utilisation de Y_1 seul. Le nombre total d'erreurs est de 10 : une de plus que dans l'utilisation de Y_1 seul.

Aucun des malades n'a été classé en "Calcul". Ce fait provient de ce que les valeurs moyennes de $A + \alpha_1/P'$, α_2/P', β/P' et P observées sur les malades de la famille "Calcul" sont comprises entre les valeurs moyennes correspondantes dans les familles "Hépatite" et "Cancer". La valeur moyenne de Y_1 à l'intérieur de la famille "Calcul" est comprise entre $\overline{Y}_1$ (Hépatite) et $\overline{Y}_1$ (chirurgical). Il en résulte que les valeurs observées de Y_1 supérieures à Y_1^B ont une probabilité plus grande de provenir d'un malade atteint d'un cancer que d'un malade ayant un calcul.

Ceci est confirmé par le fait que sur les 22 Cancéreux classés par Y_1 dans la famille Calcul ou Cancer, 21 sont reconnus effectivement Cancéreux par Y_2, tandis que sur les 7 "Calculs" classés Calcul ou Cancer par Y_1, Y_2 ne permet de préciser qu'une fois s'il s'agit de l'un ou de l'autre ictère, en donnant d'ailleurs une réponse erronée.

II - Application à un nouveau groupe de 81 malades ictériques

Les fonctions discriminantes Y_1 et Y_2 sont appliquées, dans ce paragraphe, à 81 nouveaux malades, dont les analyses électrophorétiques ont été effectuées après l'établissement de ces fonctions, mais pour lesquels un diagnostic a été prononcé avant que les médecins n'aient encore eu connaissance de l'existence de ces fonctions et de la manière de les utiliser.

L'application à ces nouveaux malades de la méthode préconisée au paragraphe précédent pour l'exploitation des résultats de l'électrophorèse constitue donc une épreuve cruciale de la validité de cette méthode.

Les 81 nouveaux ictères se répartissent ainsi :

39 hépatites, 21 calculs, 17 cancers et 4 autres ictères chirurgicaux.

Bien que dans la pratique on ne cherche pas à prononcer un diagnostic pour tous les malades, on considérera tout d'abord, à titre d'information et de vérification, les résultats de l'application des fonctions discriminantes au classement de tous les malades.

1/ Application au classement de tous les malades

a) Utilisation de Y_1 seule :

Parmi les 39 ictères médicaux, 7 sont classés comme chirurgicaux

Parmi les 42 ictères chirurgicaux, 12 sont classés médicaux

soit au total 19 erreurs sur 81 diagnostics, c'est-à-dire 23.45 % des cas (proportion théorique 24.41 %).

b) Utilisation de Y_2 seule :

Parmi les 21 calculs, 6 sont classés en cancers

Parmi les 17 cancers, 4 sont classés en calculs

soit au total 10 erreurs sur 38 diagnostics, c'est-à-dire 26.32 % des cas (proportion théorique 30.00 %).

c) Utilisation successive de Y_1 et Y_2 :

Le tableau ci-dessous donne pour les 77 malades atteints d'hépatites, calculs ou cancers, le classement effectué par l'utilisation successive de Y_1 et Y_2 :

Maladie réelle \ Classement par Y_1 et Y_2	Hépatite	Calcul	Cancer	Total
Hépatite	32	5	2	39
Calcul	9	8	4	21
Cancer	2	3	12	17
Total	43	16	18	77

Au total, 25 erreurs sur 77 diagnostics, soit 32,47 % (proportion théorique 35.75 %).

2/ Application avec diagnostic réservé

a) Diagnostic "médical" ou "chirurgical" à partir de Y_1

Le classement effectué par Y_1 est donné ci-après sous forme de tableau :

Maladie réelle \ Classement par Y_1	Diagnostic non prononcé	Médicaux	Chirurgicaux	Total
Hépatite	22	17	–	39
Calcul	14	4	3	21
Cancer	6	–	11	17
Autres chirurgicaux	3	–	1	4
Total	45	21	15	81

Un diagnostic ("médical" ou "chirurgical") est prononcé 36 fois sur 81, soit 44,4 % des cas (proportion théorique 45 %). Les erreurs, au nombre de 4, correspondent toutes à des "calculs" classés en "hépatite". Elles sont en proportion de 4/81, ou 4 sur 36 diagnostics prononcés, soit exactement 1/9.

La concordance des résultats observés avec les résultats attendus est extrèmement satisfaisante.

b) Diagnostic "calcul" ou "cancer" à partir de Y_2

L'utilisation de Y_2, avec possibilité de diagnostic réservé, sur les 38 malades atteints réellement de calcul ou cancer donne les résultats suivants :

Maladie réelle \ Classement par Y_2	Diagnostic non prononcé	Calcul	Cancer	Total
Calcul	17	4	–	21
Cancer	12	1	4	17
Total	29	5	4	38

Le nombre de diagnostics prononcés est de 9 sur 38, alors que le nombre attendu (38 × 32.66 %) est de 12.4. Le nombre d'erreurs attendu est de 38 × 0,05 = 1,9 ; on en constate une seule.

c) Diagnostic par utilisation successive de Y_1 et Y_2

Les résultats de l'application successive de Y_1 et Y_2 sont les plus intéressants à considérer, puisqu'ils correspondent à l'utilisation envisagée des fonctions discriminantes.

Ils sont consignés ci-dessous :

Maladie réelle \ Classement par Y_1 et Y_2	Diagnostic non prononcé	Hépatite	Calcul	Cancer	Chirurgicaux	Total
Hépatite	22	17	–	–	–	39
Calcul	14	4	–	–	3	21
Cancer	6	–	–	4	7	17
Autres chirurgicaux	3	–	–	–	1	4
Total	45	21	–	4	11	81

Le nombre de diagnostics énoncés est de 36 sur 81 (comme pour Y_1) soit 44.4 % des cas. Il n'y a que 4 erreurs, correspondant à des malades atteints de calculs et classés en hépatites. On a déjà signalé que l'utilisation de Y_1 et Y_2 était peu apte à déceler les "calculs" : on constate ici encore qu'un tel diagnostic n'est jamais prononcé. Enfin, il est intéressant de noter que pour les 4 malades chirurgicaux, autres que "calculs" ou "cancers", les fonctions discriminantes conduisent à ne pas énoncer de diagnostic dans trois cas, et à conclure "chirurgical" pour le quatrième.

CONCLUSIONS

L'application des fonctions discriminantes au diagnostic différentiel à partir des résultats de l'électrophorèse, a conduit aux résultats suivants :

En consentant à ne prononcer un diagnostic que dans 45 % des cas, la séparation des ictères entre "médicaux" (hépatites) et "chirurgicaux" (calculs ou cancers) peut être effectuée avec un pourcentage de 5 erreurs sur 100 malades examinés (soit 11 erreurs sur 100 diagnostics prononcés). Ces résultats sont meilleurs que ceux que l'on obtenait par les méthodes antérieures, qui donnaient au mieux 40 % de diagnostics, avec 20 erreurs sur 100 diagnostics prononcés.

La séparation des ictères chirurgicaux entre "calculs" et "cancers" ne donne pas des résultats aussi satisfaisants.

Il faut insister sur le fait que les résultats qui viennent d'être donnés ne sont valables que pour la technique électrophorétique utilisée, et pour des groupes de malades comparables à celui qui a servi de base à l'étude.

Après avoir présenté dans cette étude l'analyse discriminatoire et montré, par cet exemple, les services qu'elle est susceptible de rendre, nous en ferons dans un prochain article un exposé théorique suivi d'un nouvel exemple indiquant la manière d'exécuter les calculs pratiques.