REVUE DE STATISTIQUE APPLIQUÉE

B. CYFFERS

Analyse discriminatoire II

Revue de statistique appliquée, tome 13, n° 3 (1965), p. 39-65

<http://www.numdam.org/item?id=RSA_1965__13_3_39_0>

ANALYSE DISCRIMINATOIRE II

B. CYFFERS

Ingénieur en chef des Manufactures d'état

Dans un précédent article, nous avons exposé le rôle de l'analyse discriminatoire, et avons donné un exemple d'application.

Nous nous proposons de montrer dans le présent article comment on établit les fonctions discriminantes et comment on peut tester, par analogie avec la théorie de la Régression, les coefficients de ces fonctions.

Nous traitons ensuite deux exemples numériques, le premier se rapporte à deux familles dépendant de quatre paramètres, le second à plusieurs familles dépendant de deux paramètres et à centres alignés.

Notations utilisées.

La population étudiée est divisée en k familles. Une famille est représentée par l'indice i écrit entre parenthèse et en exposant.

Sur chaque individu, on procède à la mesure de p paramètres. Un paramètre est représenté par la lettre s ou r posée en indice à droite.

Une mesure individuelle est représentée par l'indice j placé à gauche. Une mesure individuelle s'écrit donc sous la forme :

$$ {}_jx_s^{(i)} \quad \text{ou} \quad {}_jx_r^{(i)} \quad \text{avec} \begin{cases} (i) = 1, 2 \ldots k \\ s, r = 1, 2 \ldots p \\ j = 1, 2 \ldots n^{(i)} \end{cases} $$

$n^{(i)}$ étant l'effectif des individus prélevés appartenant à la famille (i). L'effectif total de l'échantillon est :

$$ N = \sum_i n^{(i)} $$

A - LES FONCTIONS DISCRIMINANTES

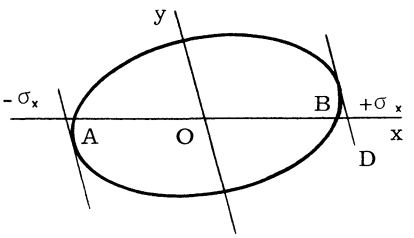
1/ Théorèmes préliminaires.

Démontrons tout d'abord les théorèmes suivants qui nous permettront d'interpréter géométriquement les fonctions discriminantes.

Théorème I - Soit dans un plan un point aléatoire suivant une loi de Laplace-Gauss à 2 dimensions. La projection de ce point sur une droite D parallèlement à une direction Δ suit une loi de Laplace-Gauss à une dimension,

dont le centre est la projection du centre, et dont l'écart-type s'obtient en menant à l'ellipse indicatrice les parallèles à Δ.

Nous pouvons prendre la droite D passant par le centre O de l'ellipse sans rien ôter à la généralité de la démonstration.



Prenons Ox, selon D, Oy parallèle à Δ.

L'équation de l'ellipse indicatrice est :

$$\frac{x^2}{\sigma_x^2} - \frac{2\rho xy}{\sigma_x \sigma_y} + \frac{y^2}{\sigma_y^2} = 1$$

La fonction caractéristique de la loi est :

$$\varphi(u,v) = E\, e^{i(ux+vy)}$$
$$= e^{-\frac{1}{2}(\sigma_x^2 u^2 + 2\rho\sigma_x\sigma_y uv + \sigma_y^2 v^2)}$$
$$= e^{-\frac{1}{2}\Gamma(u,v)}$$

La loi de la projection sur D, parallèlement à Δ de coefficients u, 0, c'est-à-dire la loi de x, a pour fonction caractéristique :

$$e^{-\frac{1}{2}\Gamma(u,0)} = e^{-\frac{1}{2}u^2\sigma_x^2}$$

Ce qui démontre que x suit une loi normale de centre O, d'écart-type σ_x.

$\Gamma(u\ v) = 1$ représente l'équation tangentielle de l'ellipse. Les droites $x = \pm\sigma_x$ satisfont cette équation. Donc les tangentes à l'ellipse parallèles à Δ coupent D aux points d'abscisse $\pm\sigma_x$.

- Faisons pivoter la direction Δ. Les points d'abscisse $\pm\sigma_x$ sont toujours extérieurs au segment AB découpé par l'ellipse sur D. Ces points se confondent avec A et B lorsque Δ est la direction conjuguée de D par rapport à l'ellipse.

La loi projetée sur D a la dispersion minimum lorsque la direction projetante Δ est conjuguée de D par rapport à l'ellipse.

Théorème II - Soit dans l'espace un point aléatoire suivant une loi de Laplace-Gauss à 3 dimensions.

a) La projection de ce point sur une droite D parallèlement à un plan Π suit une loi de Laplace-Gauss à une dimension dont l'écart-type s'obtient en menant les plans tangents parallèles à Π à l'ellipsoïde indicateur.

b) La projection de ce point sur un plan P parallèlement à une direction Δ suit une loi de Laplace-Gauss à 2 dimensions, dont l'ellipse indicatrice est la section de P par le cylindre parallèle à Δ circonscrit à l'ellipsoïde.

a) Supposons encore D passant par le centre de l'ellipsoïde ; prenons Ox_1 selon D, x_2 Ox_3 parallèle à Π.

La fonction caractéristique de la loi de distribution de x_1, x_2, x_3 est :

$$\varphi(u_1\ u_2\ u_3) = E\ e^{i(u_1x_1+u_2x_2+u_3x_3)} = e^{-\frac{1}{2}\Gamma(u_1u_2u_3)}$$

$$\Gamma(u_1\ u_2\ u_3) = \sum_{s,r} \rho_{sr}\ \sigma_s\ \sigma_r\ u_s\ u_r \quad \text{avec} \quad \rho_{sr} = \rho_{rs}$$

$$\rho_{ss} = 1$$

$$s, r = 1, 2, 3$$

La fonction caractéristique de x_1 projection d'un point aléatoire parallèlement à un plan de coefficients u_1, 0, 0, est :

$$e^{-\frac{u_1^2\sigma_1^2}{2}}$$

Donc x_1 suit une loi normale d'écart-type σ_1, et comme les plans $x_1 = \pm\sigma_1$ satisfont $\Gamma(u_1\ u_2\ u_3) = 1$, équation tangentielle de l'ellipsoïde, $\pm\sigma_1$ sont les abscisses des points d'intersection de D avec les plans tangents parallèles à Π.

b) Supposons P passant par le centre, prenons le comme plan $x_1\ Ox_2$ et Ox_3 parallèle à Δ.

La fonction caractéristique de la loi de la projection du point aléatoire parallèlement à Δ définie par $x_1 = 0$, $x_2 = 0$, est :

$$\varphi(u_1\ u_2\ 0) = e^{-\frac{1}{2}\Gamma(u_1u_20)}$$

$\Gamma(u_1\ u_2\ 0)$ est une fonction homogène du second degré en $u_1\ u_2$ donc la projection suit une loi de Laplace-Gauss à 2 dimensions dont l'ellipse indicatrice a pour équation tangentielle :

$$\Gamma(u_1\ u_2\ 0) = 1$$

Considérons dans le plan P la famille de droites tangentes à cette ellipse, c'est-à-dire les droites d'équation $u_1\ x_1 + u_2\ x_2 = 1$, telles que $\Gamma(u_1\ u_2\ 0) = 1$. Il lui correspond une famille de plans parallèles à Δ, dont les coefficients satisfont l'équation tangentielle de l'ellipsoïde, et par conséquent tangents à l'ellipsoïde. L'ellipse indicatrice de la loi projetée est donc l'intersection de P par le cylindre circonscrit à l'ellipsoïde parallèle à Δ.

Corrolaire - Dans le cas (a), on démontre comme pour le plan que la loi projetée sur D est la moins dispersée lorsque la direction du plan projetant est conjuguée de D par rapport à l'ellipsoïde.

Dans le cas (b) l'ellipse obtenue est bitangente et extérieure à l'ellipse section de l'ellipsoïde par P. Elle se confond avec cette dernière lorsque P et Δ sont conjugués.

La loi projetée sur P est la moins dispersée lorsque la direction Δ est conjuguée de P par rapport à l'ellipsoïde.

Théorème III- Soit un point aléatoire suivant une loi de Laplace-Gauss à p dimensions. Sa projection sur une multiplicité linéaire F à h dimensions ($h < p$) parallèlement à une multiplicité linéaire Φ à $p-h$ dimensions suit une loi de Laplace-Gauss à h dimensions dont l'élément indicateur (ellipse, ellipsoïde, hyperellipsoïde) est la projection sur F, parallèlement à Φ, de l'hyperellipsoïde indicateur de la loi initiale.

Prenons F passant par le centre de l'hyperellipsoïde indicateur, Ox_1, Ox_2 ... Ox_h dans F, Ox_{h+1} ... Ox_p parallèles à Φ.

La fonction caractéristique de la loi à p dimensions est :

$$\varphi(u_1\ u_2\ \dots\ u_p) = E\ e^{i(u_1x_1+u_2x_2+\ \dots\ +u_px_p)} = e^{-\frac{1}{2}\Gamma(u_1u_2-u_p)}$$

$$\Gamma(u_1\ u_2\ \dots\ u_p) = \sum_{s,r} \rho_{sr}\ \sigma_s\ \sigma_r\ u_s\ u_r \quad \text{avec} \quad \rho_{sr} = \rho_{rs}$$

$$\rho_{ss} = 1$$

$$s, r = 1, 2, \dots, p$$

La loi projetée a pour fonction caractéristique :

$$\varphi(u_1,\ u_2,\ u_h,\ 0,\ 0\ \dots\ 0) = e^{-\frac{1}{2}\Gamma(u_1,u_2\ \dots\ u_h,\ 0,0\dots\ 0)}$$

C'est donc une loi de Laplace-Gauss à h dimensions.

Les multiplicités linéaires d'équation $u_1\ x_1 + u_2\ x_2 \dots + u_h\ x_h = 1$ où $u_1\ u_2 \dots\ u_h$ satisfont à $\Gamma(u_1\ u_2 \dots\ u_h,\ 0,\ 0, \dots\ 0) = 1$ sont tangentes à l'élément indicateur de la loi projetée, et sont les traces sur F de multiplicités linéaires à p – 1 dimensions, parallèles à Φ et tangentes à l'hyperellipsoïde indicateur de la loi initiale. L'élément indicateur de la loi projetée est donc la projection sur F, parallèlement à Φ de cet hyperellipsoïde.

Corollaire - Cette projection sera la plus petite lorsqu'elle sera confondue avec l'intersection de l'hyperellipsoïde par F, donc quand F et Φ seront conjugués.

2/ Fonctions discriminantes.

1er cas : Cas de deux familles.

A - *Familles dépendant de 2 paramètres.*

Soit une population d'individus divisée en 2 familles $F^{(1)}$, $F^{(2)}$. Sur chaque individu on procède à la mesure x_1, x_2, de 2 paramètres X_1, X_2. Supposons que la distribution de ces mesures à l'intérieur de chaque famille suive une loi normale à 2 dimensions, et supposons de plus que les distributions à l'intérieur de chaque famille soient identiques et ne diffèrent que par les valeurs moyennes.

Nous nous proposons de chercher la fonction discriminante :

$$Y = \lambda_1 X_1 + \lambda_2 X_2$$

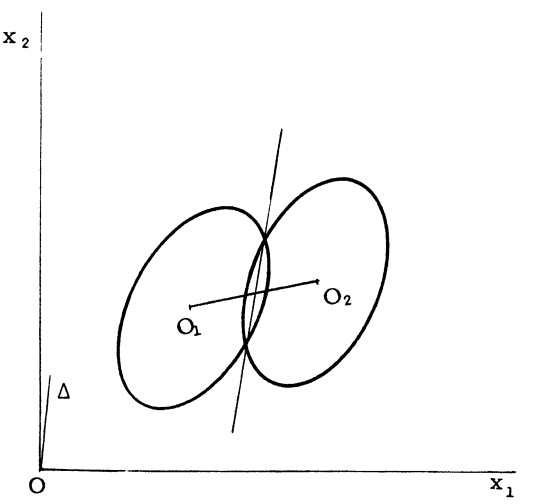
définie par $\dfrac{Y^{(2)} - Y^{(1)}}{\sigma_Y}$ maximum, c'est-à-dire telle que le rapport de la valeur absolue de la différence des valeurs moyennes de Y à l'intérieur de $F^{(1)}$ et $F^{(2)}$ à l'écart-type de la distribution de Y soit maximum.

Interprétation géométrique de la fonction discriminante.

Dans le plan des x_1, x_2 les deux familles sont représentées par deux ellipses de centres O_1, O_2, se déduisant l'une de l'autre par la translation $O_1\ O_2$.

Ces ellipses sont les ellipses indicatrices des lois de distribution de x_1 et x_2 à l'intérieur de chaque famille.

Projetons les 2 lois parallèlement à une direction Δ_1 sur une droite D quelconque. Les centres des lois projetées s'obtiennent par les projections de O_1, O_2, et les écarts-types s'obtiennent à l'aide de tangentes aux ellipses parallèles à Δ_1. Il est clair que le rapport de la distance des centres à l'écart-type commun des lois projetées est constant (pour une direction Δ_1 donnée), quelle que soit la droite D du plan sur laquelle on projette.



Nous pouvons donc supposer, sans rien ôter à la généralité de la démonstration, que l'on projette toujours sur la ligne des centres $O_1 O_2$. La distance des centres est alors constante, et le rapport de cette distance à l'écart-type commun des lois projetées est d'autant plus grand que cet écart-type est petit. Le rapport est donc maximum lorsque la direction projetante Δ_1 est la direction Δ conjuguée de la ligne des centres par rapport aux ellipses.

Δ représente une fonction linéaire $\lambda_1 x_1 + \lambda_2 x_2$ qui n'est autre que la fonction discriminante.

En conclusion, la direction Δ définie par la fonction discriminante est la direction conjuguée de la droite des centres par rapport aux ellipses indicatrices (1).

Dans le plan $x_1 x_2$, Δ est la direction selon laquelle il faut regarder les 2 ellipses pour les voir séparées au maximum l'une de l'autre.

S'il existe une équiprobabilité pour qu'un individu appartienne à chacune des familles, la droite

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

est la droite de direction Δ passant par le milieu de $O_1 O_2$. Si les ellipses se coupent, cette droite est la droite d'intersection.

Une fois Y_o déterminé, on sera amené à classer un individu dans une famille ou l'autre selon que le point représentatif de cet individu est dans l'une ou l'autre des deux régions du plan, séparées sur la droite $Y_o = \lambda_1 x_1 + \lambda_2 x_2$. La probabilité de mal classer est la même pour chaque famille.

Cette probabilité est représentée par la masse de probabilité d'une des lois située dans le demi-plan délimité par la droite $Y_o = \lambda_1 x_1 + \lambda_2 x_2$ ne contenant pas le centre de cette loi.

---

(1) Nous devons à M. G. DARMOIS cette interprétation géométrique de la fonction discriminante.

Si un individu a la probabilité p d'appartenir à la 1ère famille et q d'appartenir à la 2ème (p + q = 1). on déterminera encore Y_o de manière que la droite d'équation $Y_o = \lambda_1 x_1 + \lambda_2 x_2$ partage le plan en deux parties, telles que la masse de probabilité de la loi de la 1ère famille, située du côté de O_2, multipliée par p, soit égale à la masse de la loi de la 2ème famille, située du côté de O_1, multipliée par q.

Calcul de la fonction discriminante.

L'équation tangentielle de l'ellipse indicatrice d'une des lois, ou plutôt de l'ellipse qui se déduit de cette dernière après translation pour amener son centre à l'origine des coordonnées est

$$\sigma_{x_1}^2 u_1^2 + 2\sigma_{x_1} \sigma_{x_2} u_1 u_2 + \sigma_{x_2}^2 u_2^2 = 1$$

ou, sous une forme générale :

$$\Gamma(u_1 u_2) = \sum_{s,r} \rho_{sr} \sigma_s \sigma_r u_s u_r = 1$$

avec :

$$\rho_{sr} = \rho_{rs} \quad \text{et} \quad \rho_{ss} = 1 \quad (s, r = 1, 2)$$

Les premiers membres des équations, correspondant aux ellipses indicatrices des lois de distribution à l'intérieur de chacune des familles, sont identiques.

Pratiquement, on dispose, à partir des mesures effectuées, des sommes des carrés et produits centrés à l'intérieur de chaque famille. Soient :

$$a_{11} \quad a_{12} \quad a_{22}$$

les sommes totales relatives aux deux familles.

Désignons par ${}_j x_1^{(1)}$, la valeur de X_1 sur le jième individu de la première famille, par $\bar{x}_1^{(1)}$ la valeur moyenne de X_1 à l'intérieur de la première famille, etc...

On a, si $n^{(1)}$ et $n^{(2)}$ représentent les effectifs observés dans chaque famille

$$a_{11} = \sum_{i=1}^{2} \sum_{j=1}^{n^{(i)}} ({}_j x_1^{(i)} - \bar{x}_1^{(i)})^2$$

$$a_{22} = \sum_{i=1}^{2} \sum_{j=1}^{n^{(i)}} ({}_j x_2^{(i)} - \bar{x}_2^{(i)})^2$$

$$a_{12} = \sum_{i=1}^{2} \sum_{j=1}^{n^{(i)}} ({}_j x_1^{(i)} - \bar{x}_1^{(i)}) \, (x_2^{(i)} - \bar{x}_2^{(i)})$$

a_{11}, a_{12}, a_{22} sont proportionnels aux estimations des coefficients intervenant dans $\Gamma(u_1 u_2)$.

Comme nous ne nous occupons que de directions conjuguées, on peut considérer l'ellipse d'équation :

$$\Phi(u_1 u_2) = a_{11} u_1^2 + 2 a_{12} u_1 u_2 + a_{22} u_2^2 = 1$$

homothétique de l'ellipse indicatrice estimée.

Les paramètres directeurs de la droite des centres sont :

$$\overline{x}_1^{(2)} - \overline{x}_1^{(1)} = d_1$$

$$\overline{x}_2^{(2)} - \overline{x}_2^{(1)} = d_2$$

La droite d'équation $\lambda_1 x_1 + \lambda_2 x_2 = Y$ sera conjuguée de la ligne des centres si :

$$\frac{\Phi'_{u_1}(\lambda_1 \lambda_2)}{d_1} = \frac{\Phi'_{u_2}(\lambda_1 \lambda_2)}{d_2}$$

ou

$$\frac{a_{11} \lambda_1 + a_{12} \lambda_2}{d_1} = \frac{a_{12} \lambda_1 + a_{22} \lambda_2}{d_2}$$

Les λ n'étant définis qu'à un coefficient de proportionnalité près, on peut prendre pour valeurs de λ_1, λ_2 les solutions des équations :

$$\boxed{\begin{aligned} a_{11} \lambda_1 + a_{12} \lambda_2 &= d_1 \\ a_{12} \lambda_1 + a_{22} \lambda_2 &= d_2 \end{aligned}}$$

La résolution de ces équations est simple. Nous ne l'expliciterons pas, nous la rattacherons à celle des systèmes plus généraux que nous obtiendrons par la suite.

Remarque - Au lieu de l'ellipse indicatrice de la loi de distribution ou une ellipse homothétique, on peut considérer toute ellipse par rapport à laquelle la droite $O_1 O_2$ admet même direction conjuguée, donc toute ellipse appartenant au faisceau linéaire d'enveloppes de seconde classe, défini par l'ellipse indicatrice et une autre enveloppe satisfaisant à cette condition ; par exemple, le point double à l'infini, dans la direction $O_1 O_2$.

On sera donc conduit aux mêmes valeurs de λ_1, λ_2 en partant d'une ellipse d'équation tangentielle :

$$a_{11} u_1^2 + 2a_{12} u_1 u_2 + a_{22} u_2 + \mu(d_1 u_1 + d_2 u_2)^2 = 1$$

μ étant un paramètre quelconque. On aura alors à résoudre des équations de la forme :

$$(a_{11} + \mu d_1^2) \lambda_1 + (a_{12} + \mu d_1 d_2) \lambda_2 = d_1$$

$$(a_{12} + \mu d_1 d_2) \lambda_1 + (a_{22} + \mu d_2^2) \lambda_2 = d_2$$

B - *Familles dépendant de p paramètres.*

Supposons que sur chaque individu des deux familles précédentes, on puisse procéder à d'autres mesures. Soient x_1, x_2 ... x_p les mesures de p paramètres différents X_1 X_2 ... X_p.

Les deux familles sont représentées par deux hyperellipsoïdes dans l'espace à p dimensions. La fonction discriminante

$$Y = \lambda_1 x_1 + \lambda_2 x_2 + \ldots + \lambda_p x_p$$

est l'équation d'une multiplicité linéaire à $p - 1$ dimensions.

Projetons les 2 lois de distribution à p dimensions sur la droite des centres $O_1 O_2$ parallèlement à cette multiplicité linéaire : on obtient

deux lois à 1 dimension, de centres O_1 O_2. Y étant la fonction discriminante, le rapport $\frac{|Y^{(2)} - Y^{(1)}|}{\sigma_Y}$ est le plus grand que puisse fournir une fonction linéaire des x_i. Donc pour les lois projetées, la dispersion est minimum, et par conséquent, la direction de la multiplicité linéaire projetante est conjuguée de la droite des centres par rapport aux hyperellipsoïdes.

Ces hyperellipsoïdes sont déterminés à une homothétie près, à partir des données, par la connaissance de la matrice des sommes des carrés et produits centrés à l'intérieur des familles.

$$\bar{\bar{a}} = \begin{Vmatrix} a_{11} & a_{12} & \ldots\ldots\ldots & a_{1p} \\ a_{21} & a_{22} & \ldots\ldots\ldots & a_{2p} \\ \cdot & & & \\ \cdot & & & \\ \cdot & & & \\ \cdot & & & \\ a_{p_1} & a_{p_2} & \ldots\ldots\ldots & a_{pp} \end{Vmatrix}$$

où :

$$a_{ss} = \sum_{i=1}^{2} \sum_{j=1}^{n(i)} ({}_j x_s^{(i)} - \bar{x}_s^{(i)})^2$$

et

$$a_{sr} = \sum_{i=1}^{2} \sum_{j=1}^{n(i)} ({}_j x_s^{(i)} - \bar{x}_s^{(i)}) ({}_j x_r^{(i)} - \bar{x}_r^{(i)})$$

avec

$$a_{sr} = a_{rs} \qquad s, r = 1, 2 \ldots p.$$
$$i = 1, 2 .$$

L'équation tangentielle de ces hyperellipsoïdes est :

$$\Phi(u_1\ u_2 \ldots\ldots\ldots\ u_p) = \sum_{s,r} a_{sr}\ u_s\ u_r = 1$$

La condition de conjugaison entre la multiplicité linéaire :

$$Y = \lambda_1\ x_1 + \lambda_2\ x_2 + \ldots\ldots\ldots + \lambda_p\ x_p$$

et la droite des centres de paramètres directeurs d_1, d_2 ... d_p, coordonnées du vecteur $\overrightarrow{O_1\ O_2}$ s'écrit :

$$\frac{\varphi'_{u_1}(\lambda_1\ \lambda_2 \ldots\ \lambda_p)}{d_1} = \frac{\varphi'_{u_2}(\lambda_1\ \lambda_2 \ldots\ \lambda_p)}{d_2} = \ldots\ \frac{\varphi'_{u_p}(\lambda_1\ \lambda_2 \ldots\ \lambda_p)}{d_p}$$

ou

$$\frac{a_{11}\ \lambda_1 + a_{12}\ \lambda_2 + \ldots + a_{1p}\ \lambda_p}{d_1} = \ldots\ldots\ldots = \frac{a_{1p}\ \lambda_1 + \ldots + a_{pp}\ \lambda_p}{d_p}$$

Les λ n'étant définis qu'à un coefficient de proportionnalité près,

on peut prendre comme système de valeurs λ la solution du système d'équations linéaires :

$$\begin{aligned}
&a_{11}\,\lambda_1 + a_{12}\,\lambda_2 + \ldots\ldots\ldots + a_{1p}\,\lambda_p = d_1\\
&a_{12}\,\lambda_1 + a_{22}\,\lambda_2 + \ldots\ldots\ldots + a_{2p}\,\lambda_p = d_2\\
&\cdot\\
&\cdot\\
&\cdot\\
&\cdot\\
&a_{1p}\,\lambda_1 + a_{2p}\,\lambda_2 + \ldots\ldots\ldots + a_{pp}\,\lambda_p = d_p
\end{aligned}$$

En considérant les λ comme les coordonnées d'un vecteur on peut écrire ces relations d'une manière plus concentrée sous la forme :

$$\bar{\bar{a}}\ \bar{\lambda} = \bar{d}$$

Multipliant à gauche par $\bar{\bar{a}}^{-1}$, il vient :

$$\bar{\lambda} = \bar{\bar{a}}^{-1}\ \bar{d}$$

Cette formule donne $\bar{\lambda}$ en fonction de $\bar{\bar{a}}$ et $\bar{d}$

Elle équivaut à l'ensemble des p équations :

$$\boxed{\lambda_s = \sum_r a^{sr}\ d_r}$$

dans lesquelles les a^{sr} sont les termes de $\bar{\bar{a}}^{-1}$

ou

$$\begin{aligned}
\lambda_1 &= a^{11}\,d_1 + a^{12}\,d_2 + a^{13}\,d_3 + \ldots\ldots\ldots + a^{1p}\,d_p\\
\lambda_2 &= a^{12}\,d_1 + a^{22}\,d_2 + \ldots\ldots\ldots\ldots\ldots + a^{2p}\,d_p\\
&\cdot\\
&\cdot\\
&\cdot\\
\lambda_p &= a^{1p}\,d_1 + a^{2p}\,d_2 + \ldots\ldots\ldots\ldots\ldots + a^{pp}\,d_p
\end{aligned}$$

Remarque - On sera conduit à un système identique de valeurs λ en partant au lieu de la matrice de termes a_{sr}, de la matrice de termes $a_{sr} + \mu d_s\ d_r$, μ étant un paramètre quelconque. Cette matrice définit en effet le faisceau linéaire des hyperellipsoïdes par rapport auxquels la multiplicité linéaire conjuguée de la direction de paramètres directeurs $d_1, d_2 \ldots d_p$ est la même, puisque la matrice des termes d_s, d_r définit l'enveloppe de seconde classe constituée par le point double à l'infini dans la direction $d_1\ d_2 \ldots d_p$.

2e cas : cas de plusieurs familles.

Si l'on dispose de k familles dépendant de p paramètres, on supposera encore que les variances et covariances des mêmes variables sont égales dans les différentes familles.

Ces familles sont représentées dans l'espace à p dimensions par des hyperellipsoïdes à p dimensions.

Si les centres de ces hyperellipsoïdes sont sur une droite, c'est-à-dire une multiplicité linéaire à 1 dimension, il existe une multiplicité linéaire à p – 1 dimensions dont la "direction" est conjuguée de la droite des centres par rapport aux hyperellipsoïdes. Ce cas est analogue à celui de deux familles. Il existe donc une fonction discriminante :

$$Y = \lambda_1 x_1 + \lambda_2 x_2 + \ldots\ldots\ldots + \lambda_p x_p$$

Si les centres des hyperellipsoïdes sont sur une multiplicité linéaire à 2 dimensions, il existe une multiplicité linéaire à p – 2 dimensions qui lui est conjuguée. C'est en projetant parallèlement à cette multiplicité linéaire, sur le plan des centres que l'on obtient les lois projetées les moins dispersées. Ces lois ont pour centres les centres des familles, et leur ellipse indicatrice est l'intersection des hyperellipsoïdes par le plan des centres. Analytiquement, la multiplicité linéaire à p – 2 dimensions est définie comme l'intersection de 2 multiplicités à p – 1 dimensions.

$$\begin{cases} Y = \lambda_1 x_1 + \lambda_2 x_2 + \ldots\ldots\ldots + \lambda_p x_p \\ Y' = \lambda'_1 x_1 + \lambda'_2 x_2 + \ldots\ldots\ldots + \lambda'_p x_p \end{cases}$$

Il y a donc deux fonctions discriminantes.

D'une manière générale, si les centres sont sur une multiplicité linéaire P_h à h dimensions, il existe une multiplicité linéaire π_{p-h} à p – h dimensions, conjuguée de la précédente. C'est en projetant parallèlement à π sur P, qu'on obtient les lois projetées les moins dispersées.

π_{p-h} est défini par h équations (π_{p-h} est l'intersection de h multiplicités linéaires à p – 1 dimensions).

On a donc dans ce cas *h fonctions discriminantes.*

Il existe donc autant de fonctions discriminantes que la multiplicité linéaire contenant les centres a de dimensions.

Nous limiterons notre étude au cas des fonctions discriminantes à 1 dimension, c'est-à-dire au cas où les centres sont alignés.

Désignons par T le caractère qui distingue les h familles. L'alignement des centres des familles est alors le reflet d'une réalité physique sous-jacente, T est souvent un caractère mesurable, et on peut considérer qu'il existe une régression linéaire entre les paramètres X et le caractère T. T prend alors la même valeur $t^{(i)}$ pour tous les individus de la famille i.

Désignons par δ_1, $\delta_2 \ldots$ δ_p les paramètres directeurs de la droite de régression de $X_1 \ldots X_p$ par rapport à T.

Les équations de cette droite sont, en désignant par $\bar{x}_1 \ldots \bar{x}_p$ les moyennes générales des valeurs observées :

$$\begin{aligned} x_1 - \bar{x}_1 &= \delta_1 (t - \bar{t}) \\ x_2 - \bar{x}_2 &= \delta_2 (t - \bar{t}) \\ &\cdot \\ &\cdot \\ &\cdot \\ x_p - \bar{x}_p &= \delta_p (t - \bar{t}) \end{aligned}$$

Si $n^{(i)}$ représente le nombre d'individus examinés dans la ième famille on obtient δ_s en rendant minimum la somme des carrés des distances des points expérimentaux aux points d'intersection de la droite de régression avec le "plan" $t = t_1$, pour la 1ère famille, avec le plan $t = t_2$ pour la seconde, etc...

Soit donc :

$$\sum_{i=1}^{k} \sum_{j=1}^{n^{(i)}} \sum_{s=1}^{p} \left[{}_j x_s^{(i)} - \bar{x}_s - \delta_s (t^{(i)} - \bar{t}) \right]^2_{\text{minimum}}$$

ce qui donne :

$$\delta_s = \frac{\sum_{i=1}^{k} \sum_{j=1}^{n^{(i)}} {}_j x_s^{(i)} \ (t^{(i)} - \bar{t})}{\sum_{i=1}^{k} n^{(i)} \ (t^{(i)} - \bar{t})^2}$$

On peut écrire cette expression sous la forme :

$$\delta_s = \frac{\sum x_s \ (t - \bar{t})}{\sum (t - \bar{t})^2}$$

les sommations s'entendant :

- au numérateur ; pour toutes les valeurs de x_s observées ,
- au dénominateur : pour la totalité des observations.

Les paramètres de la droite des centres estimée, dans l'espace des x_1 x_2 ... x_p peuvent donc être pris égaux à

$$\sum x_1 (t - \bar{t}) \quad \sum x_2 \ (t - \bar{t}) \ \ldots\ldots\ldots \sum x_p \ (t - \bar{t})$$

Ce sont ces valeurs qu'on prendra comme coordonnées du vecteur $\bar{d}$.

Les coefficients λ de la fonction discriminante s'obtiendront par la relation :

$$\bar{\lambda} = \bar{\bar{a}}^{-1} \ \bar{d}$$

un terme a_{sr} de la matrice $\bar{\bar{a}}$ ayant la même expression que dans le cas de 2 familles, i prenant alors les valeurs 1, 2 ... k.

(Quand les centres sont dans un plan, c'est en général que les familles se distinguent entre elles par deux caractères U et V, et il existe une régression linéaire entre les X d'une part, U et V d'autre part).

En réalité, pour pouvoir utiliser l'analogie avec la régression pour tester les coefficients λ, on ne calcule pas la fonction discriminante par application de la formule

$$\bar{\lambda} = \bar{\bar{a}}^{-1} \ \bar{d}$$

Compte tenu de la remarque concernant le faisceau d'enveloppes de seconde classe admettant mêmes directions conjuguées, on peut, au lieu de la matrice a, utiliser toute matrice de la forme :

$$\left\| \begin{array}{llll} a_{11} + \mu d_1^2 & a_{12} + \mu d_1 d_2 & \ldots\ldots\ldots & a_{1p} + \mu d_1 d_p \\ a_{12} + \mu d_1 d_2 & a_{22} + \mu d_2^2 & \ldots\ldots\ldots & a_{2p} + \mu d_2 d_p \\ & & & \\ a_{1p} + \mu d_1 d_p & a_{2p} + \mu d_2 d_p & \ldots\ldots\ldots & a_{pp} + \mu d_p^2 \end{array} \right\|$$

où μ est un coefficient quelconque.

En particulier, considérons la matrice :

$$\bar{\bar{r}} = \left\| \begin{array}{ll} \dfrac{\left[\sum x_1 (t - \bar{t})\right]^2}{\sum (t - \bar{t})^2} \ldots\ldots\ldots\ldots\ldots & \dfrac{\left[\sum x_1 (t - \bar{t})\right]\left[\sum x_p (t - \bar{t})\right]}{\sum (t - \bar{t})^2} \\ \dfrac{\left[\sum x_1 (t - \bar{t})\right]\left[\sum x_2 (t - \bar{t})\right]}{\sum (t - \bar{t})^2} \ldots\ldots\ldots\ldots\ldots & \\ \ldots\ldots\ldots\ldots\ldots & \dfrac{\left[\sum x_p (t - t)\right]^2}{\sum (t - \bar{t})^2} \end{array} \right\|$$

C'est la matrice des sommes des carrés et produits centrés dus à la régression linéaire des X par rapport à T. Elle dépend d'un degré de liberté.

On établit la matrice $\bar{\bar{b}} = \overline{\overline{a + r}}$ qui dépend de $N - k + 1$ degrés de liberté $\left(N = \sum_{i=1}^{k} n^{(i)} = \text{nombre total d'individus examinés}\right)$ et on calcule $\bar{\lambda}$ par :

$$\bar{\lambda} = \bar{\bar{b}}^{-1} \bar{d}$$

3 - Analogie avec la régression, test des coefficients.

Nous allons montrer qu'en calculant les coefficients λ par cette dernière formule, on obtient des coefficients qui sont ceux du plan de régression de t par rapport à x_1, x_2 ... x_p, si on suppose les centres de familles rigoureusement alignés comme on le précisera plus loin.

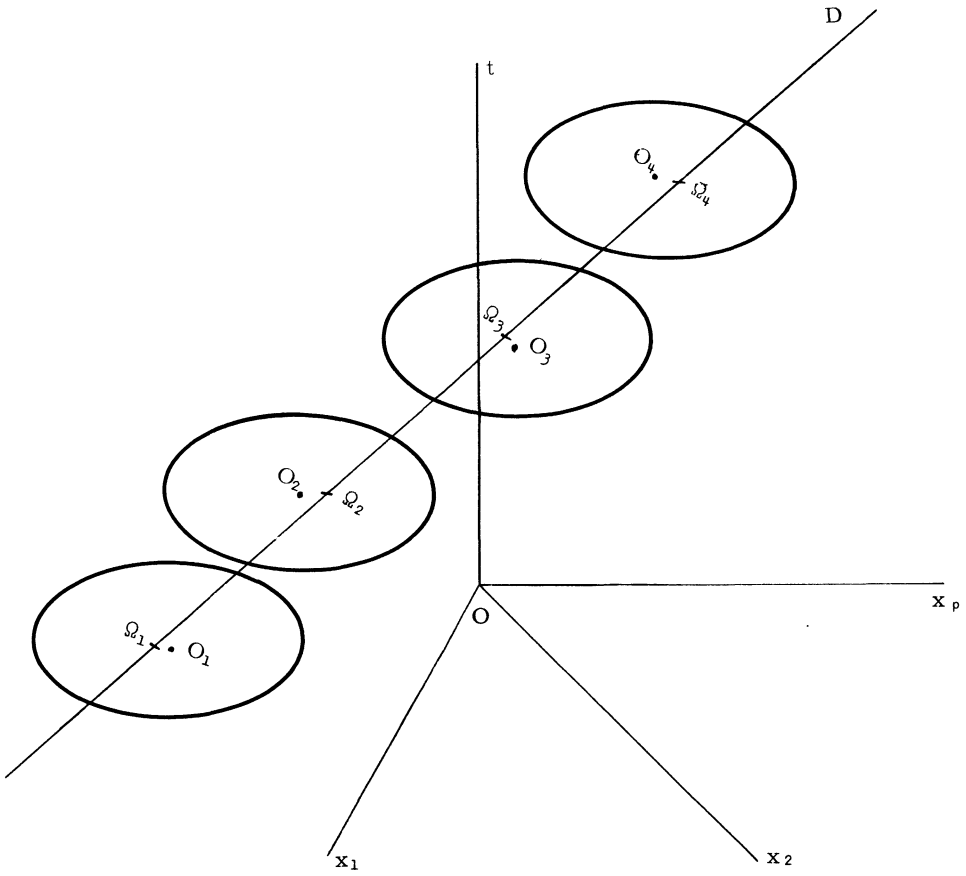
Considérons dans l'espace à p + 1 dimensions p + 1 axes rectangulaires Ox_1, Ox_2, Ox_p, Ot. Les familles y sont représentées par des "ellipses" homothétiques les unes des autres et situées dans des "plans" t = constante.

La droite des centres estimée, D, passe par le point moyen général de coordonnées $\bar{x}_1$, $\bar{x}_2$ ... $\bar{x}_p$, $\bar{t}$ et a pour paramètres directeurs :

$$\sum x_1 (t - \bar{t}), \quad \sum x_2 (t - \bar{t}), \quad \sum x_p (t - \bar{t}), \quad \sum (t - \bar{t})^2$$

Désignons par O_1, O_2 ... O_k les centres réels des ellipses, Ω_1, Ω_2 ... , Ω_k les points d'intersection de la droite D avec les "plans" $t = t^{(1)}$ $t = t^{(2)}$ ... $t = t^{(k)}$.

On a :



- coordonnées de O_i $\quad \bar{x}_1^{(i)}, \bar{x}_2^{(i)} \ldots\ldots\ldots \bar{x}_p^{(i)}, t^{(i)}$
- coordonnées de Ω_i $\quad {}_0x_1^{(i)}, {}_0x_2^{(i)} \ldots\ldots\ldots {}_0x_p^{(i)}, t^{(i)}$

et on peut écrire :

$$ {}_jx_s^{(i)} - \bar{x}_s = [{}_jx_s^{(i)} - \bar{x}_s^{(i)}] + [\bar{x}_s^{(i)} - {}_0x_s^{(i)}] + [{}_0x_s^{(i)} - \bar{x}_s] $$

et en élevant au carré et sommant :

$$ \sum_{ij} [{}_jx_s^{(i)} - \bar{x}_s]^2 = \sum_{ij} [x_s^{(i)} - \bar{x}_s^{(i)}]^2 + \sum_i n^{(i)} [\bar{x}_s^{(i)} - {}_0x_s^{(i)}]^2 + \sum_i n^{(i)} [{}_0x_s^{(i)} - \bar{x}_s]^2 $$

Degré de liberté $N - 1 = (N - k) + (k - 2) + 1$.

Le premier terme de la décomposition de la somme des carrés, qui dépend de N - k degrés de liberté, n'est autre que a_{ss}, terme de la diagonale de la matrice $\bar{\bar{a}}$.

Considérons le dernier terme, qui ne dépend que d'un degré de liberté. C'est le terme dû à la régression de X_s par rapport à T. Comme dans le plan des x_s, t, les valeurs ${}_0x_s^{(i)}$ sont situées sur la droite de paramètres directeurs $\sum x_s (t - \bar{t})$, $\sum (t - \bar{t})^2$ qui passe par le point $\bar{x}_s$, $\bar{t}$,

on a :

$$ {}_{o}x_s^{(i)} - \bar{x}_s = (t^{(i)} - \bar{t}) \frac{\sum x_s (t - \bar{t})}{\sum (t - \bar{t})^2} $$

et

$$ \sum_i n^{(i)} \left[{}_{o}x_s^{(i)} - \bar{x}_s \right]^2 = \frac{\sum x_s (t - \bar{t})^2}{\sum (t - \bar{t})^2} = r_{ss} $$

On établit des relations analogues pour les termes rectangles, et la matrice $\bar{\bar{b}}$ apparaît ainsi comme la matrice des sommes des carrés et produits centrés autour de la moyenne générale si l'on suppose les centres parfaitement alignés

$$ \left({}_{o}x_s^{(i)} = \bar{x}_s^{(i)} \right) $$

Si $t = \lambda'_1 x_1 + \lambda'_2 x_2 + \lambda'_p x_p + C^{te}$ est l'équation du plan de régression de t par rapport à x_1, x_2 ... x_p, les λ' peuvent être considérés comme les coordonnées d'un vecteur $\bar{\lambda}'$, et on l'obtient par :

$$ \bar{\lambda}' = \bar{\bar{b}}^{-1} \bar{d} $$

Il y a donc identité entre les coefficients λ et λ'.

Pour tester les coefficients λ, on va donc décomposer en 2 termes la somme des carrés totale $\sum (t - \bar{t})^2$, qui ne dépend plus que de $N - 1 - (k - 2)$ degrés de liberté puisqu'on a rigoureusement aligné les centres des familles. Cet alignement a été sans effet sur la valeur de $\sum (t - \bar{t})^2$ puisque les déplacements des nuages de points ont été réalisés dans les plans $t = t^{(1)}$, $t = t^{(2)}$ ..., mais on a utilisé $K - 2$ relations entre les x et t.

On a :

$$ t - \bar{t} = \lambda_1 (x_1 - \bar{x}) + \lambda_2 (x_2 - \bar{x}) + \ldots\ldots\ldots + \lambda_p (x_p - \bar{x}) + \varepsilon $$

$$ = y - \bar{y} + \varepsilon $$

$$ \sum (t - \bar{t})^2 = \sum (y - \bar{y})^2 + \text{résidu} $$

On établit aisément que :

$$ \sum (y - \bar{y})^2 = \lambda_1 \sum x_1 (t - \bar{t}) + \lambda_2 \sum x_2 (t - \bar{t}) + \ldots\ldots + \lambda_p \sum x_p (t - \bar{t}) $$

Ce terme dépend de p degrés de liberté.

On peut alors calculer le terme résiduel, par différence, et calculer le carré moyen résiduel. Le terme résiduel dépend de $(N - 1) - (k - 2) - p = N - k + 1 - p$ degrés de liberté.

On obtient les variances estimées des coefficients λ en multipliant ce carré moyen résiduel par les termes diagonaux de la matrice $\bar{\bar{b}}^{-1}$.

1er exemple - 2 familles dépendant de 4 paramètres.

Discrimination entre Ictères médicaux et chirurgicaux à partir des résultats de l'analyse Electrophorétique des protéines du sérum.

Nous sommes à l'origine en présence de 3 familles (Hépatites, Calculs, Cancers) dépendant chacune de 4 paramètres. La première

question qui se pose est celle de l'alignement des centres. Le tableau ci-dessous indique les valeurs moyennes observées pour chacun des paramètres à l'intérieur des 3 familles.

	Hépatites $n_1 = 106$	Calculs $n_2 = 41$	Cancers $n_3 = 50$
$\frac{A + \alpha_1}{P'}$	46. 934905	46. 126829	39. 822000
$\frac{\alpha_2}{P'}$	6. 248113	8. 790243	11. 818000
$\frac{\beta}{P'}$	15. 864150	17. 556097	21. 132000
P	73. 514150	77. 446341	77. 820000

Les centres sont sensiblement alignés en ce qui concerne les paramètres $\frac{\alpha_2}{P'}$ et $\frac{\beta}{P'}$. Ils ne le sont pas en ce qui concerne $\frac{A + \alpha_1}{P'}$ et P. Par conséquent, on ne peut considérer, dans l'espace à 4 dimensions, les 3 familles comme ayant leurs centres alignés, et l'on doit effectuer la discrimination 2 à 2.

Toutefois, comme le problème consiste d'abord à distinguer les ictères médicaux des ictères chirurgicaux et ensuite à séparer les ictères chirurgicaux en "calculs" et "cancers", nous sommes conduits à calculer successivement la fonction discriminante entre Hépatites d'une part, et calculs et cancers d'autre part, et ensuite la fonction discriminante entre calculs et cancers.

Nous indiquons comment a été établie la 1ère de ces fonctions appelée Y_1.

Nous avons rassemblé les 91 groupes de 4 mesures relatifs aux "Chirurgicaux" (Calculs + Cancers).

Nous cherchons la fonction discriminante.

$$\lambda_1 \frac{A + \alpha_1}{P'} + \lambda_2 \frac{\alpha_2}{P'} + \lambda_3 \frac{\beta}{P'} + \lambda_4 P$$

Tableau I - Valeurs moyennes

	Hépatites $n_1 = 106$	Calculs-Cancers $n_2 = 91$	Différence des moyennes
$\frac{A + \alpha_1}{P'}$	46. 934 905	42. 662 637	+ 4. 272 268
$\frac{\alpha_2}{P'}$	6. 248 113	10. 453 846	- 4. 205 733
$\frac{\beta}{P'}$	15. 864 150	19. 520 879	- 3. 656 729
P	73. 514 150	77. 651 648	- 4. 137 498

Le tableau suivant représente la matrice des sommes des carrés et produits centrés à l'intérieur des familles.

Cette matrice dépend de 195 degrés de liberté.

Tableau II

Matrice des sommes des carrés et produits centrés 195 degrés de liberté

	$\frac{A + \alpha_1}{P'}$	$\frac{\alpha_2}{P'}$	$\frac{\beta}{P'}$	P
$\frac{A + \alpha_1}{P'}$	10 044. 633	- 1 902. 065	- 3 092. 476	- 1 919. 616
$\frac{\alpha_2}{P'}$		2 609. 191	482. 600	- 788. 355
$\frac{\beta}{P'}$			4 823. 554	614. 986
P				20 943. 156

Le caractère qui distingue les deux familles, dans le cas présent, n'est pas mesurable. Mais, comme dans le cas des 2 familles les centres sont rigoureusement alignés, on peut, pour utiliser l'analogie avec la régression et tester les coefficients λ, introduire un paramètre t arbitraire : choisissons t de manière que sa moyenne générale soit nulle : par exemple :

- pour les Hépatites, posons :

$$t_1 = \frac{n_2}{n_1 + n_2} = 0.461\,928\,93$$

- pour les Calculs-Cancers posons :

$$t_2 = \frac{-n_1}{n_1 + n_2} = -0.538\,071\,07$$

On a alors :

$$\bar{t} = 0$$

$$\sum (t - \bar{t})^2 = \frac{n_1 n_2}{n_1 + n_2} = 48.964\,467\,00$$

On considère alors la régression des paramètres $\frac{A + \alpha_1}{P'}$ ... par rapport à t.

La droite de régression a pour paramètres directeurs :

$$\sum \frac{A + \alpha_1}{P'} (t - \bar{t}) = 209.189\,325$$

$$\sum \frac{\alpha_2}{P'} (t - \bar{t}) = -205.931\,475$$

$$\sum \frac{\beta}{P'}(t - \bar{t}) = -179.049786$$

$$\sum P\ (t - \bar{t}) = -202.590384$$

(On remarquera qu'on obtient facilement les nombres ci-dessus en multipliant les différences des valeurs moyennes par $\sum(t - \bar{t})^2$ ainsi

$$\sum \frac{A + \alpha_1}{P'}(t - \bar{t}) = 48.964467\,00 \times 4.272268).$$

De même, les termes dus à la régression des paramètres par rapport à t forment la matrice du tableau III. Le calcul des termes de cette matrice est aisé. Par exemple : le terme à l'intersection de la ligne $\frac{A + \alpha_1}{P'}$ et de la colonne $\frac{\alpha_2}{P'}$ a pour expression :

$$\frac{\left[\sum \frac{A + \alpha_1}{P'}(t - \bar{t})\right]\left[\sum \frac{\alpha_2}{P'}(t - \bar{t})\right]}{\sum(t - \bar{t})^2}$$

soit $\frac{1}{48.964467\,00} \times 209.189325 \times (-205.931475)$,

mais on connaît déjà le quotient de chacun de ces deux derniers nombres par $\sum(t - \bar{t})^2$.

Tableau III

Matrice des termes de régression 1 degré de liberté

	$\frac{A + \alpha_1}{P'}$	$\frac{\alpha_2}{P'}$	$\frac{\beta}{P'}$	P
$\frac{A + \alpha_1}{P'}$	893.712859	−879.794450	−764.948671	−865.520415
$\frac{\alpha_2}{P'}$		+866.092800	+753.035594	+852.041063
$\frac{\beta}{P'}$			+654.736545	+740.818132
P				+838.217309

En ajoutant les termes respectifs des matrices des tableaux II et III on obtient la matrice des sommes des carrés et produits centrés autour de la moyenne générale, dépendent de 196 degrés de liberté, faisant l'objet du tableau IV.

On pourrait objecter qu'on peut obtenir directement cette matrice à partir des données. Nous croyons cependant préférable d'adopter la méthode utilisée ci-dessus, à cause de sa généralité. Elle est indispensable lorsqu'on a affaire à plus de 2 familles, et dans tous les cas on a besoin de la matrice des sommes des carrés et produits centrés à l'intérieur des familles (tableau II) pour calculer la variance de la fonction discriminante à l'intérieur des familles.

Tableau IV

Matrice des sommes des carrés et produits centrés autour de la moyenne générale 196 degrés de liberté

	$\frac{A + \alpha_1}{P'}$	$\frac{\alpha_2}{P'}$	$\frac{\beta}{P'}$	P
$\frac{A + \alpha_1}{P'}$	10 938. 346	-2 781. 859	-3 857. 425	-2 785. 136
$\frac{\alpha_2}{P'}$		+3 475. 284	+1 235. 636	+ 63. 686
$\frac{\beta}{P'}$			+5 478. 290	+1 355. 804
P				21 781. 373

La matrice inverse de cette matrice figure dans le tableau V.

Tableau V

Matrice inverse 10^{-4}

	$\frac{A + \alpha_1}{p'}$	$\frac{\alpha_2}{P'}$	$\frac{\beta}{P'}$	P
$\frac{A + \alpha_1}{P'}$	1. 449 965 730	0. 877 318 701	0.790 002 908	0.133 664 197
$\frac{\alpha_2}{P'}$		3. 661 848 245	-0.236 952 875	0.116 223 393
$\frac{\beta}{P'}$			2.447 630 787	-0.050 646 554
P				0.479 011 950

Les coefficients s'obtiennent en multipliant chaque ligne de la matrice inverse par :

$$\sum \frac{A + \alpha_1}{P'} (t - \bar{t}), \quad \sum \frac{\alpha_2}{P'} (t - \bar{t}), \quad \sum \frac{\beta}{P'} (t - \bar{t}), \quad \sum P\,(t - \bar{t}).$$

Par exemple :

$$\lambda_1 = 10^{-4} (1.449\,965\,730 \times 209.189\,325 + 0.877\,318\,701 \times (-205.931\,475) + \ldots)$$

On obtient :

$$\lambda_1 = -0.004\,587\,911\,44$$
$$\lambda_2 = -0.055\,168\,348\,37$$
$$\lambda_3 = -0.021\,393\,103\,37$$
$$\lambda_4 = -0.008\,394\,789\,18$$

La fonction discriminante est donc :

$$y_1 = -0.004\,587\,911\,\frac{A+\alpha_1}{P'} - 0.055\,168\,348\,\frac{\alpha_2}{P'} - 0.021\,393\,103\,\frac{\beta}{P'} - 0.008\,394\,789\,P$$

On peut remplacer les coefficients λ par un groupe de valeurs proportionnelles (on a intérêt à conserver un grand nombre de chiffres significatifs pour chaque coefficient λ tant que l'on n'a pas effectué le test de signification de ces coefficients).

Test des coefficients λ : Les coefficients λ obtenus par la méthode que nous avons suivie sont ceux de l'équation de régression :

$$t - \bar{t} = \lambda_1\left[\frac{A+\alpha_1}{P'} - \frac{\overline{A+\alpha_1}}{P'}\right] + \lambda_2\left[\frac{\alpha_2}{P'} - \frac{\overline{\alpha_2}}{P'}\right] + \lambda_3\left[\frac{\beta}{P'} - \frac{\overline{\beta}}{P'}\right] + \lambda_4(P - \overline{P})$$

Pour un ensemble de valeurs $\frac{A+\alpha_1}{P'}$, $\frac{\alpha_2}{P'}$, $\frac{\beta}{P'}$, P observées sur un individu, on a :

$$t - \bar{t} = \lambda_1\left[\frac{A+\alpha_1}{P'} - \frac{\overline{A+\alpha_1}}{P'}\right] + \lambda_2\left[\frac{\alpha_2}{P'} - \frac{\overline{\alpha_2}}{P'}\right] + \lambda_3\left[\frac{\beta}{P'} - \frac{\overline{\beta}}{P'}\right] + \lambda_4(P - \overline{P}) + \varepsilon.$$

En élevant au carré, et en sommant, on obtient :

$$\sum(t - \bar{t})^2 = \sum(y_1 - \overline{y_1})^2 + \sum\varepsilon^2$$

Dans cette équation, $\sum(y_1 - \overline{y_1})^2$ représente la somme des carrés de y_1 autour de la moyenne générale de y_1.

On établit aisément que :

$$\sum(y_1 - y_1)^2 = \lambda_1 \sum \frac{A+\alpha_1}{P'}(t - \bar{t}) + \lambda_2 \sum \frac{\alpha_2}{P'}(t - \bar{t}) + \ldots$$

$$= 15.\,932\,291\,400$$

D'où le tableau VI d'analyse de la variance :

Tableau VI

Analyse de la variance

Somme des Carrés		Degrés de Liberté	Quotient
Régression	15. 932 291 40	4	
Résidu	33. 032 175 60	192	0. 172 042 58
Total	48. 964 467 00	196	

D'après la théorie de la régression multiple, on obtient les variances estimées des coefficients λ en multipliant le carré moyen résiduel 0. 172 042 58 par chacun des termes de la diagonale principale de la matrice inverse.

Ainsi var $\lambda_1 = 10^{-4}\ 1.449\ 965\ 730 \times 0.172\ 042\ 58$

d'où :

- estimation de var $\lambda_1 = 0.000\ 024\ 945\ 584$ et $s_\lambda = 0.004\ 994$
- " var $\lambda_2 = 0.000\ 062\ 999\ 382$ $s_\lambda = 0.007\ 937$
- " var $\lambda_3 = 0.000\ 042\ 109\ 671$ $s_\lambda = 0.006\ 489$
- " var $\lambda_4 = 0.000\ 008\ 241\ 045$ $s_\lambda = 0.002\ 870$

Il vient donc :

$$\frac{|\lambda_1|}{s_{\lambda_1}} = 0.919 \qquad \frac{|\lambda_3|}{s_{\lambda_3}} = 3.297$$

$$\frac{|\lambda_2|}{s_{\lambda_2}} = 6.951 \qquad \frac{|\lambda_4|}{s_{\lambda_4}} = 2.925$$

Les coefficients λ_2, λ_3, λ_4 sont hautement significatifs. Par contre λ_1 n'est pas significativement différent de 0, et on pourrait ne pas retenir le paramètre $\frac{A + \alpha_1}{P'}$ dans le calcul de la fonction discriminante, et envisager de recommencer les calculs en ne considérant que les 3 paramètres $\frac{\alpha_2}{P'}$; $\frac{\beta}{P'}$ et P. Cependant comme le coefficient de $\frac{A + \alpha_1}{P'}$ se révèle le plus significatif dans la discrimination entre calculs et cancers, il paraît préférable d'avoir les mêmes paramètres dans les deux fonctions discriminantes.

Variance de y_1 à l'intérieur des familles :

En désignant par $a_{11}\ a_{12}\ a_{13}\ a_{14}$, $a_{22}\ a_{23}\ a_{24}\ a_{33}$, a_{34}, a_{44} les termes de la matrice des sommes des carrés et produits centrés à l'intérieur des familles (Tableau II) on établit que la somme des carrés centrés de y à l'intérieur des familles a pour expression :

$$\sum_{ij} ({}_j y_1^{(i)} - \bar{y}_1^{(i)})^2 = \lambda_1 (\lambda_1\ a_{11} + \lambda_2\ a_{12} + \lambda_3\ a_{13} + \lambda_4\ a_{14})$$
$$+ \lambda_2 (\lambda_1\ a_{12} + \lambda_2\ a_{22} + \lambda_3\ a_{23} + \lambda_4\ a_{24})$$
$$+ \lambda_3 (\ldots) + \lambda_4 (\ldots)$$

Tous calculs faits on trouve :

$$\sum_{ij} ({}_j y_1^{(i)} - \bar{y}_1^{(i)})^2 = 10.748\ 165$$

d'où estimation var $y_1 = \frac{10.748\ 165}{195} = 0.055\ 118\ 795$

et $s_{y_1} = 0.23477.$

On peut obtenir la variance de y_1 à l'intérieur des familles d'une manière plus rapide, et qui permet dans une certaine mesure une vérification de l'exactitude des calculs.

Désignons par R le coefficient de corrélation entre t et y_1. La décomposition en carrés du tableau VI s'écrit :

$$\sum(t - \bar{t})^2 = R^2 \sum (t - \bar{t})^2 + (1 - R^2) \sum (t - \bar{t})^2$$

d'où :

$$R^2 = \frac{15.932\,291\,40}{48.964\,467\,00} = 0.325\,385$$

et

$$1 - R^2 = \frac{33.032\,175\,60}{48.964\,467\,00} = 0.674\,615$$

On peut décomposer la somme des carrés $\sum(y_1 - \bar{y}_1)^2$ autour de la moyenne générale en deux termes : l'un dû à la régression par rapport à t [$= R^2 \sum(y_1 - \bar{y}_1)^2$], et le résidu = somme des carrés à l'intérieur des familles [$=(1 - R^2) \sum(y_1 - \bar{y}_1)^2$].

La somme des carrés à l'intérieur des familles à donc pour expression :

$$15.932\,291\,40 \times 0.674\,615 = 10.748\,165.$$

On retrouve la valeur calculée directement plus haut.

Discrimination effectuée par y_1 - Les valeurs moyennes de y_1 au sein des deux familles sont :

$\bar{y}_1$ Hépatites = $-1.516\,550$

$\bar{y}_1$ Calculs-Cancers = $-1.841\,935$

La différence est égale à 0.325 385, soit 1.385 95 écarts-types.

On remarque que la différence des valeurs moyennes a la même valeur numérique que R^2 (0.325 385). Ce fait provient du choix des valeurs de t_1 et t_2.

2ème exemple - Plusieurs familles dépendant de 2 paramètres. Centres alignés.

Estimation du nombre de capes de cigares que l'on peut tirer d'une feuille de tabac, connaissant sa longueur et sa largeur.

La confection des cigares est mécanique, l'ouvrière présente la feuille de tabac sur un gabarit de découpe. La machine coupe donc une portion de feuille appelée cape qui a toujours la même forme, les mêmes dimensions, et enroule cette cape autour du cigare pour lui donner son aspect définitif.

Pour estimer le nombre de capes que l'on peut tirer d'un lot de tabac, on peut utiliser l'essai sur machine, ou l'essai au gabarit en Laboratoire. Ce dernier consiste à dénombrer à l'aide de morceaux de carton ayant la forme du gabarit de découpe le nombre de capes réalisable.

Il paraît bon de mettre à la disposition de l'Ingénieur et de la Maîtrise une règle simple, rapide, permettant de contrôler le travail effectué aussi bien sur machine qu'au Laboratoire. Comme l'essai en

Laboratoire donne également la longueur L et la largeur l des feuilles, *nous avons déterminé le nombre de capes que l'on peut tirer d'une feuille connaissant sa longueur et sa largeur.*

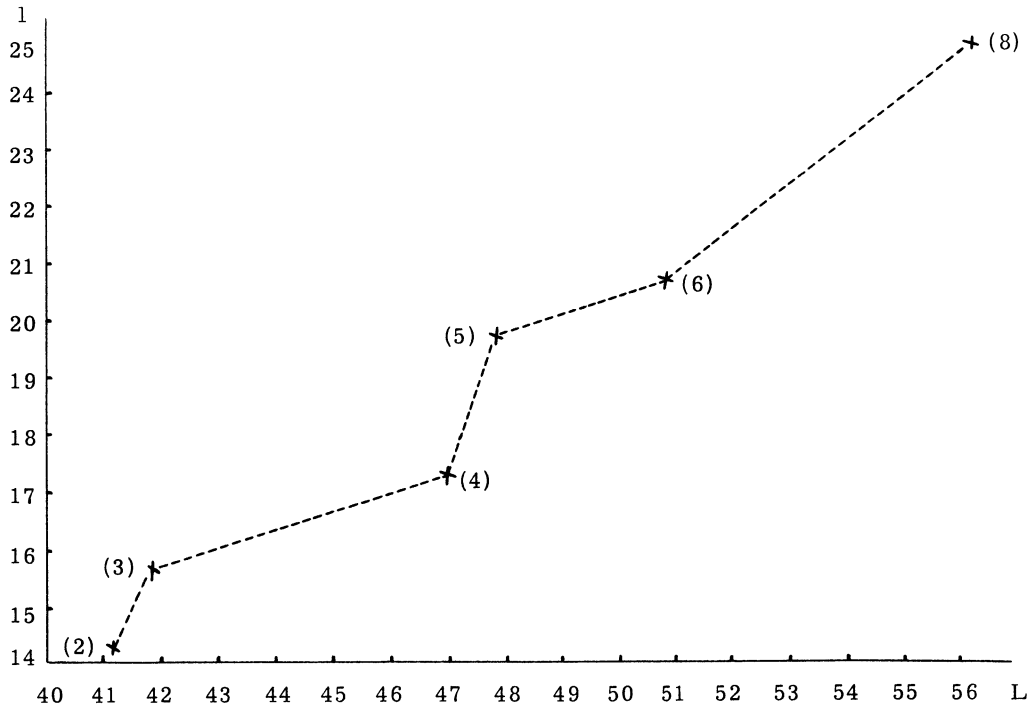
Les mesures ont porté sur 950 feuilles. Leur répartition entre les différentes familles ainsi que les valeurs moyennes observées sont indiquées dans le tableau I ci-après :

Tableau I

Valeurs moyennes

Familles	2 capes	3 capes	4 capes	5 capes	6 capes	8 capes	
Nbre de feuilles	200	100	200	135	200	115	950
$\bar{L}$ (cm)	41.135	41.740	46.890	47.830	50.855	56.191	
$\bar{l}$ (cm)	13.965	15.570	17.320	19.696	20.865	25.043	

L'examen des valeurs moyennes montre qu'on est en droit de considérer les centres comme alignés, ainsi qu'en fait foi le graphique ci-dessous :



Somme des carrés et produits centrés

	L^2	l^2	L l
2	2 223.355	398.755	247.945
3	823.240	284.510	- 79.180
4	1 431.580	437.520	314.080
5	1 087.082	316.547	-114.985
6	790.795	489.355	236.085
8	713.792	372.782	73.044
Total	7 069.844	2 299.469	676.989

La matrice des sommes des carrés et produits centrés est donc:

Tableau II

Matrice des sommes des carrés et produits centrés 944 degrés de liberté

	L	l
L	7.069,844	676,989
l	676,989	2 299,469

Désignons par c le nombre de capes que l'on peut tirer d'une feuille. Le caractère qui distingue les familles (c) est ici mesurable. Le paramètre c joue le rôle du paramètre t de la théorie générale.

$$\overline{c} = 4.521$$

$$\sum (c - \overline{c})^2 = 3\ 417.079$$

$$\sum L(c - \overline{c}) = 8\ 639.890$$

$$\sum l(c - \overline{c}) = 6\ 250.032$$

D'où le tableau III.

Tableau III

Matrice des termes de régression 1 degré de liberté

	L	l
L	21 845.470	15 802.856
l	15 802.856	11 431.666

En ajoutant les matrices des Tableaux II et III, nous obtenons la matrice du Tableau IV, appelée Matrice des sommes des carrés et produits centrés autour de la moyenne générale, mais il convient de ne pas oublier que l'on suppose avoir ramené les centres des différentes familles rigoureusement sur la droite estimée des centres.

Tableau IV

Matrice des sommes des carrés et produits centrés autour de la moyenne générale
945 degrés de liberté

	L	1
L	28 915.3	16 479.8
1	16 479.8	13 731.1

Désignons par Δ la valeur du déterminant formé par les termes de cette matrice.

$$\Delta = 125\ 455\ 067.79$$

$$\frac{I}{\Delta} = 10^{-9}\ 7.970\ 985$$

La matrice inverse est :

Tableau V

Matrice inverse 10^{-6}

	L	1
L	109.4504	-131.3602
1	-131.3602	230.4834

d'où

$\lambda_1 = 0.124\ 6342$ $\quad \lambda_2 = 0.305\ 5905$

proportionnels à 1 et 2.451

Test des coefficients λ :

Décomposons la somme des carrés $\sum(c - \bar{c})^2 = 3\,417.079$ qui dépend de 945 degrés de liberté (et non pas 949) en deux parties, l'une correspondant à la régression par rapport à L et 1 (2 degrés de liberté), l'autre étant le résidu dépendant de 943 degrés de liberté.

Le terme correspondant à la régression s'obtient par :

$$\lambda_1 \sum L(c - \bar{c}) + \lambda_2 \sum 1(c - \bar{c})$$

Le tableau d'analyse de la variance est donc :

Tableau VI

Analyse de la variance

	Sommes des carrés	Degrés de liberté	Carré moyen
Régression	2 986. 776	2	
Résidu	430. 303	943	0. 456 3128
Total	3 417. 079	945	

La variance de λ_1 s'obtient en multipliant le carré moyen du résidu par le terme correspondant à L^2 dans la matrice inverse, soit dans le cas présent :

$$10^{-6} \quad 109.4504$$

d'où variance $\lambda_1 = 10^{-4} \quad 0.664\,159 \qquad s_{\lambda_1} = 0.00815$

variance $\lambda_2 = 10^{-4} \quad 2.041\,991 \qquad s_{\lambda_2} = 0.01429$

Les valeurs λ_1 et λ_2 observées sont donc très significatives.

Calculons maintenant les coefficients λ_1 λ_2 en ne tenant compte que des familles à 2, 4, 6 et 8 capes. Il est plus logique de ne tenir compte que de ces familles, puisque en voulant déduire le nombre de capes de la seule connaissance de la longueur et de la largeur, il est normal de supposer la feuille symétrique. D'autre part, l'alignement des centres de ces 4 familles est meilleur que celui de tous les centres et les produits centrés sont positifs pour ces familles alors qu'ils sont négatifs pour les familles à 3 et 5 capes.

Tableau des valeurs moyennes :

	2 capes	4 capes	6 capes	8 capes
n	200	200	200	115
$\bar{L}$	41. 135	46. 890	50. 855	56. 191
$\bar{l}$	13. 965	17. 320	20. 865	25. 043

Somme des carrés et produits centrés :

	L^2	l^2	L l
2	2 223. 355	398. 755	247. 945
4	1 431. 580	437. 520	314. 080
6	790. 795	489. 355	236. 085
8	713. 792	372. 782	73. 044
Total	5 159. 522	1 698. 412	871. 154

D'où la matrice :

	L	l
L	5 159. 522	871. 154
l	871. 154	1 698. 412

La matrice inverse est proportionnelle à :

	L	l
L	1 698. 412	- 871. 154
l	- 871. 154	5 159. 522

(Il suffit d'intervertir les termes de la diagonale principale et de changer les signes des termes relatifs à L,l).

Comme $\sum L\,(c - \overline{c}) = 7\ 708.76$

$\sum l\,(c - \overline{c}) = 5\ 716.92$

Il vient :

$\lambda_1 = 8\ 112\ 332 \qquad \lambda_2 = 22\ 781\ 057$

proportionnels à 1 et 2. 808

Equation des droites discriminantes.

Pour ce 2ème cas que nous venons de traiter, il est tout à fait logique de prendre pour droites discriminantes celles qui passent par les milieux des segments reliant 2 centres consécutifs, bien que l'on n'ait pas observé le même nombre de feuilles à 2, 4, 6 et 8 capes, dans le lot étudié, et que pour les calculs on n'ait pas le même nombre de feuilles dans chaque famille. La règle que nous établissons doit être valable pour tout lot, et l'on doit donner la même importance à toutes les familles.

On est ainsi conduit à tracer les droites (graphique I).

L + 2. 808 l = 87. 987 séparant les feuilles à 2 et à 4 capes

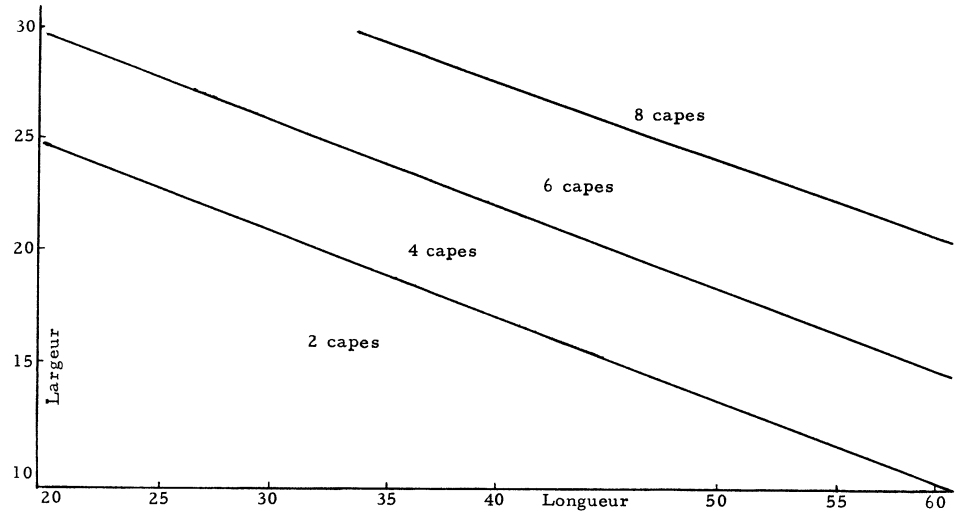
L + 2. 801 l = 102. 485 séparant les feuilles à 4 et à 6 capes

L + 2. 088 l = 117. 978 séparant les feuilles à 6 et à 8 capes

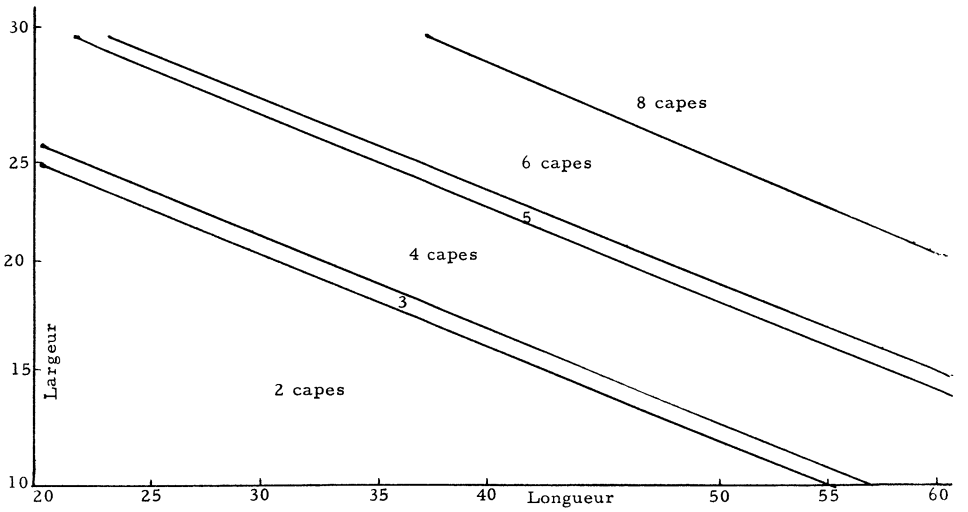
Par contre, pour le 1er cas, où l'on tient compte des feuilles à nombre impair de capes, il y a lieu de donner plus d'importance aux familles à 2, 4, 6 et 8 capes. Nous avons choisi, compte tenu de la proportion de feuilles à 3 et 5 capes, de tracer des bandes de largeur dixième des bandes correspondant à 2, 4, 6 capes. Pour cela, nous avons tracé sur le graphique II les droites

$$L + 2.451\ l = K$$

passant par les centres, et tracé les bandes graphiquement.



GRAPHIQUE I



GRAPHIQUE II