STATISTIQUE MATHEMATIQUE
PLANS D'EXPERIENCES

Daniel DUGUÉ
Professeur à la Sorbonne

I. — DEFINITION DE LA STATISTIQUE MATHEMATIQUE A L'INTERIEUR DU CALCUL DES PROBABILITES

Dans le calcul des probabilités, la notion de convergence se fractionne en deux notions différentes qui se réduisent à une seule dans le cas d'éléments certains :

1) la convergence en probabilité ou faible ;

2) la convergence presque certaine (ou presque sûre) ou forte, qui entraîne la précédente.

1° Une suite d'éléments aléatoires $X_1, X_2, ..., X_n$ tend en probabilité vers un élément X si :

$$\lim. \text{Prob}\,[|X_n - X| < \varepsilon] = 1,$$

quelque petit que soit ε.

2° Une suite d'éléments aléatoires $X_1, X_2, ..., X_n$ tend presque sûrement vers un élément X si :

$$\text{Prob}\,[\lim X_n = X] = 1.$$

A mon avis (cette opinion est contestée, et je citerai, parmi ceux qui ne la partagent pas, MM. Fréchet et Brard), la statistique mathématique est le domaine de la convergence en probabilité, le calcul des probabilités théoriques est celui de la convergence presque sûre.

Cela tient au fait que la statistique, science appliquée, ne peut s'intéresser à « une suite infinie de résultats », et que dire qu'une suite de variables aléatoires converge presque sûrement, c'est dire qu'il y a une probabilité unité pour qu'une suite infinie de résultats tende vers une limite.

Je rappelle, comme je l'ai dit au début de ce paragraphe, que si l'on a affaire à des quantités certaines, les deux manières de converger coïncident.

On peut présenter cette remarque en disant que le calcul des probabilités est une extension des mathématiques certaines au même titre que la géométrie non-euclidienne est une extension de la géométrie euclidienne (cette dernière s'obtenant à partir de la géométrie non-euclidienne par l'intervention de l'axiome d'Euclide : par tout point passe une seule parallèle à une droite).

Les mathématiques certaines sont le domaine du calcul des proba-

bilités dans lequel on ne considère que les éléments pour lesquels convergence en probabilité et convergence presque certaine coïncident.

En introduisant cette idée, je me permets d'insister d'une manière plus générale sur le caractère tout relatif d'un travail d'axiomatisation. Ce sont certes des recherches intéressantes, mais il ne faut pas perdre de vue que c'est une tâche qui ne sera jamais terminée. La rigueur définitive ne peut jamais être atteinte. Toute axiomatique sous-entend obligatoirement un certain nombre de conventions auxquelles on peut renoncer.

Un autre point qui montre le calcul des probabilités comme une extension des mathématiques certaines est le suivant, d'ailleurs lié au développement précédent.

Un ensemble $X_1, X_2, X_3, ..., X_n$ de quantités certaines est connu si l'on connaît chaque élément *isolément*. Il n'en est plus de même si $X_1, ..., X_n$ sont aléatoires. La connaissance de chacun des éléments $X_1, ..., X_n$, ou connaissance marginale, ne renseigne pas complètement sur l'ensemble des n éléments. C'est le phénomène de la corrélation qui est typiquement aléatoire.

Ces quelques idées théoriques étant amorcées, je voudrais décrire un problème statistique particulier, celui des *plans d'expériences* (*design of experiments*, en anglais).

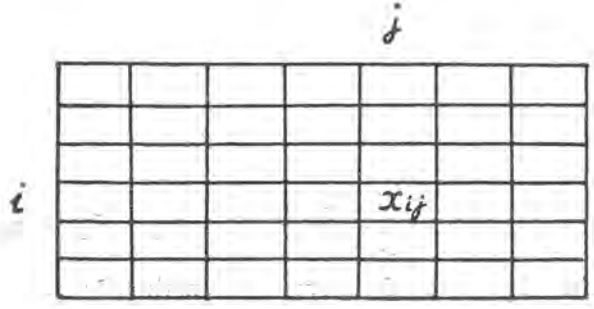
II. — PLANS D'EXPERIENCES

La statistique, je l'ai dit, ne s'intéresse qu'à un nombre fini de résultats. Elle va avoir, dans ce problème des plans d'expériences, à utiliser des notions d'algèbre finie et en particulier celle de corps de Galois et d'espaces construits sur ces corps de Galois.

Soient n variables aléatoires $(n = pq)$, à deux indices, i prenant p valeurs et j en prenant q :

1) normales (obéissant à une loi de Laplace-Gauss) ;
2) indépendantes ;
3) ayant toutes le même écart-type σ.

Les p valeurs de l'indice i représentent l'influence d'un facteur, les q valeurs de l'indice j représentent l'influence d'un autre facteur ; répartissons les résultats dans les cases d'un tableau de p lignes et q colonnes. L'un de ces résultats sera x_{ij}, situé dans la i^e ligne et la j^e colonne.



Posons-nous le problème suivant : les $n = pq$ variables aléatoires ont-elles toutes la même valeur moyenne ; autrement dit, m_{ij}, valeur moyenne de la variable aléatoire X_{ij}, dont le résultat est x_{ij}, sera-t-elle indépendante de i et j ?

Pour arriver à résoudre « statistiquement » ce problème, la méthode va être de construire une ou plusieurs variables aléatoires, combinaisons linéaires des X_{ij}, et de calculer leur fonction de répartition dans l'hypothèse où $m_{ij} = m$, quels que soient i et j.

Supposons que l'une de ces variables aléatoires L ait une réalisation l. On rejettera l'hypothèse ($m_{ij} = m$ pour tous les i et j) si la probabilité de l'écart $|l - E(L)|$ est inférieure à ce qu'on appelle le *seuil de signification* (1/20 ou 1/100 sont les seuils les plus usuels et pour lesquels les tables sont dressées).

Dans le cas actuel où les réalisations sont x_{ij}, on utilise l'égalité suivante :

$$\sum_{ij}(x_{ij}-m)^2=\sum_{ij}(x_{ij}-x_{i.}-x_{.j}+x_{..})^2+p\sum_{j}(x_{.j}-x_{..})^2$$
$$+q\sum_{i}(x_{i.}-x_{..})^2+px(x_{..}-m)^2$$

où $x_{.i}$ est la moyenne dans la ligne i : $x_{i.}=\frac{1}{q}\sum_{j}x_{ij}$

$x_{j.}$ est la moyenne dans la colonne j : $x_{.j}=\frac{1}{p}\sum_{i}x_{ij}$

$x_{..}$ est la moyenne dans l'ensemble du tableau :

$$x_{..}=\frac{1}{pq}\sum_{i}\sum_{j}x_{ij}=\frac{1}{p}\sum_{i}x_{i.}=\frac{1}{q}\sum_{j}x_{.j}.$$

C'est ce qu'on appelle une *décomposition orthogonale* (car les termes rectangles disparaissent dans le développement) de la forme quadratique $\Sigma(x_{ij} - m)^2$.

La variable L définie tout à l'heure sera le quotient $\frac{q_c^2}{Q^2}$ ou $\frac{q_l^2}{Q^2}$, où :

$$Q^2=\sum_{ij}(x_{ij}-x_{i.}\quad x_{.j}+x_{..})^2\,;\ q_c^2=p\sum_{j}(x_{.j}-x_{..})^2\,:\ q_l^2=q\sum_{i}(x_{i.}-x_{..})^2.$$

Ce sont des variables dont la loi peut être aisément calculée et qu'on appelle loi de Behrens-Fisher.

Si $\frac{q_c^2}{Q^2}$ $(p-1)$ s'écarte significativement, au sens précisé, de sa valeur moyenne qui est l'unité, m_{ij} ne pourra pas être indépendante de j. De même, si $\frac{q_l^2}{Q^2}$ $(q-1)$ s'écarte significativement de l'unité, m_{ij} ne pourra être indépendante de i. Cela revient à rejeter le fait qu'un événement est dû au hasard s'il a une probabilité trop faible (« trop faible » étant fixé d'avance) de se réaliser. σ^2 est « estimé », dans le cas où m_{ij} ne dépend pas de i et j, soit par $\frac{q_c^2}{q-1}$, soit par $\frac{q_l^2}{p-1}$.

La méthode du plan d'expérience que je viens de décrire, et qui est connue sous le nom d'*analyse de variance*, se généralise aisément à plus de deux indices.

Pour le cas de trois indices i, j, k prenant respectivement p, q, r valeurs, on a :

$$\sum_{ijk}(x_{ijk}-m)^2=\sum_{ijk}(x_{ijk}-x_{ij.}-x_{i.k}-x_{.jk}+x_{i..}+x_{.j.}+x_{..k}-x_{...})^2$$
$$+r\sum_{ij}(x_{ij.}-x_{i..}-x_{.j.}+x_{...})^2+q\sum_{ik}(x_{i.k}-x_{i..}-x_{..k}+x_{...})^2$$
$$+p\sum_{jk}(x_{.jk}-x_{.j.}-x_{..k}+x_{...})^2+pq\sum_{k}(x_{..k}-x_{...})^2$$
$$+qr\sum_{i}(x_{i..}-x_{...})^2+pr\sum_{j}(x_{.j.}-x_{...})^2+pqr(x_{...}-m^2),$$

$x_{ij.}$, par exemple, sera la moyenne des résultats dont les deux premiers indices sont i et j.

Chacune des sommes de carrés divisées par ce qu'on appelle le nombre de degrés de liberté, — pour $r\sum_{ij}(x_{ij.}-x_{i..}-x_{.j.}+x_{...})^2$ ce sera $(p-1)(q-1)$ —, fournit une estimation de σ^2 dans le cas où m_{ijk} est indépendant de i, j, k.

Ces estimations sont indépendantes au sens du calcul des probabilités et leur quotient ne doit pas s'écarter significativement de l'unité. Les tables de Behrens-Fisher-Snedecor permettent encore de résoudre la question de la signification.

Dans le cas de trois indices, il faut donc pqr résultats, et si $p=q=r$, p^3 résultats pour appliquer cette méthode.

Nous allons étudier un procédé qui permet de n'en utiliser que p^2. Cette économie est un des buts du plan d'expérience, l'autre étant la simplification des calculs qui deviennent rapidement d'une complication monstrueuse à mesure qu'augmente le nombre des indices.

Supposons, par exemple, que p soit égal à 5 et examinons la figure suivante, que l'on appelle un *carré latin :*

A B C D E
C D E A B
E A B C D
B C D E A
D E A B C

Dans chaque ligne figure une fois et une seule chaque lettre. Il en est de même dans chaque colonne. Supposons que chaque case contienne un résultat aléatoire gaussien ; tous ces résultats étant indépendants les uns des autres et ayant même écart-type. Nous voulons encore tester l'hypothèse que la moyenne est la même pour toutes les variables.

Nous aurons ici une décomposition quadratique de la forme suivante :

$$\sum_{ij}(x_{ij}-m)^2=\sum_{ij}(x_{ij}-x_{i.}-x_{.j}-x_t+2x_{..})^2+5\sum_{i}(x_{i.}-x_{..})^2$$
$$+5\sum_{j}(x_{.j}-x_{..})^2+5\sum_{t}(x_t-x_{..})^2+25(x_{..}-m)^2,$$

$x_{i.}$ et $x_{.j}$ étant encore les moyennes des résultats dans les colonnes i et j et x_t la moyenne des résultats dans les cases portant la même lettre que la case considérée.

On aura encore, si la moyenne est la même pour toutes les cases (c'est-à-dire ne dépend ni de la ligne, ni de la colonne, ni de la lettre), des variables indépendantes dont les quotients obéiront aux lois de Behrens-Fisher-Snedecor, avec les degrés de liberté appropriés.

On dit ici que les lignes, colonnes et lettres sont orthogonales. Les termes rectangles disparaissent, car il y a une lettre donnée et une seule dans chaque ligne et dans chaque colonne.

De cette façon, avec 5^2 résultats, on obtient la même précision d'analyse de la variance qu'avec 5^3 dans le modèle factoriel à trois indices, et les calculs sont beaucoup plus simples.

On peut pousser plus loin l'orthogonalité en prenant deux carrés latins orthogonaux. Prenons un carré de 5 cases de côté, et considérons les deux carrés suivants :

A B C D E	α β γ δ ε
C D E A B	δ ε α β γ
E A B C D	β γ δ ε α
B C D E A	ε α β γ δ
D E A B C	γ δ ε α β

Nous pouvons constater que si on les superpose et si l'on considère les couples formés par la lettre latine et la lettre grecque qui se trouvent dans une case, tous les couples sont représentés :

Aα	Bβ	Cγ	Dδ	Eε
Cδ	Dε	Eα	Aβ	Bγ
Eβ	Aγ	Bδ	Cε	Dα
Bε	Cα	Dβ	Eγ	Aδ
Dγ	Eδ	Aε	Bα	Cβ

A est ainsi associé à α, β, γ, δ, ε, une fois et une seule, et de même pour toutes les lettres latines.

Cette disposition permet la décomposition de la forme quadratique suivante :

$$\sum(x_{ij}-m)^2=\sum_{ij}(x_{ij}-x_{i.}-x_{.j}-x_t-x_\tau+3x_{..})^2$$
$$+5\sum(x_{i.}-x_{..})^2+5\sum(x_{.j}-x_{..})^2+5\sum(x_t-x_{..})^2$$
$$+5\sum(x_\tau-x_{..})^2$$

à cause de l'orthogonalité des lignes et des colonnes, des lettres latines et des lettres grecques.

Ici, 5^2 résultats suffisent pour une analyse qui en aurait nécessité 5^4 avec le modèle factoriel ordinaire, puisqu'on a affaire à quatre indices (lignes, colonnes, lettres latines et lettres grecques).

Dans le cas d'un carré de 5 cases de côté, on peut construire quatre carrés latins orthogonaux deux à deux au sens que nous avons précisé :

A α a 0	B β b 1	C γ c 2	D δ d 3	E ε e 4
C δ b 4	D ε c 0	E α d 1	A β e 2	B γ a 3
E β c 3	A γ d 4	B δ e 0	C ε a 1	D α b 2
B ε d 2	C α e 3	D β a 4	E γ b 0	A δ c 1
D γ e 1	E δ a 2	A ε b 3	B α c 4	C β d 0

Dans ce cas, la forme quadratique $\Sigma(x_{ij} - m)^2$ peut se décomposer de la façon suivante :

$$\sum (x_{ij} - x_{i.} - x_{.j} - x_{\tau} - x_{\tau} - x_{l} - x_{n} + 5x_{..})^2$$

$$+ 5 \sum (x_{i.} - x_{..})^2 + 5 \sum (x_{.j} - x_{..})^2 + 5 \sum (x_{\tau} - x_{..})^2 + 5 \sum (x_{l} - x_{..})^2$$

$$+ 5 \sum (x_{\tau} - x_{..})^2 + 5 \sum (x_{n} - x_{..})^2 + 25(x_{..} - m)^2$$

x_{τ} étant la moyenne des résultats dans les cases contenant la même lettre majuscule que la case considérée et des définitions analogues pour les autres indices. Ici, le plan d'expérience permet, avec 5^2 résultats, une analyse de variance qui en aurait nécessité 5^6 avec le plan factoriel ordinaire.

On voit aisément qu'on ne peut avoir plus de $p - 1$ carrés latins, de p cases de côté, orthogonaux deux à deux. Peut-on les avoir tous ? On conçoit que c'est là une question importante pour le plan d'expérience. On dit qu'on a alors orthogonalisation complète.

La solution de ce problème n'est pas obtenue à l'heure actuelle. On sait que pour les nombres p tels qu'il existe un corps de Galois ayant p éléments (c'est-à-dire tels que $p = \omega^n$, ω étant un nombre premier), il y a possibilité d'orthogonalisation complète. Mais on ne sait pas si la réciproque est exacte. Cela pose un problème d'algèbre finie très difficile.

On sait que pour tous les nombres p qui ne sont pas de la forme $2(2n + 1)$, il existe au moins deux carrés latins orthogonaux ayant p cases de côté. On sait qu'il n'y a pas deux carrés latins orthogonaux de 6 cases de côté. C'est le vieux problème des 36 officiers posé par Euler au milieu du XVIII^e siècle et résolu par la négative par Tarry, en 1900. On ne sait rien pour le cas $p = 10$.

Je vais maintenant présenter un autre modèle que l'ensemble orthogonal de carrés latins, qui permet à la fois une simplification des calculs et une économie des résultats. Il s'agit de ce que Yates appelle le *bloc incomplet équilibré*.

Il s'agit dans ce problème de placer v lettres dans b lignes (blocs) de k cases chacune :

1) dans chaque ligne, chaque lettre figure 0 ou 1 fois ;

2) chaque lettre est répétée le même nombre de fois r dans tout le modèle (donc $bk = rv$) ;

3) chaque combinaison de deux lettres figure le même nombre de fois λ dans tout le modèle. Cette condition entraîne que l'on doit avoir :

$$r(k-1) = \lambda(v-1).$$

La solution générale de ce problème reste encore à trouver. On a des solutions particulières qui s'obtiennent à partir des géométries projectives ou euclidiennes construites sur des corps de Galois, également à partir de modules, mais la solution générale n'est pas encore atteinte.

Donnons un exemple :

A B D H	B C F M	C H J K	G H I M
A C E I	B E G K	D E J M	
A F G J	B I J L	D F I K	
A K L M	C D G L	E F H L	

Ici : $b = 13$; $v = 13$; $r = 4$; $k = 4$; $\lambda = 1$.