MATHÉMATIQUES ET SCIENCES HUMAINES

J. L. PIEDNOIR

Modèle probabiliste et problème statistique

Mathématiques et sciences humaines, tome 28 (1969), p. 39-51

<http://www.numdam.org/item?id=MSH_1969__28__39_0>

M.S.H. (7e année, no 28, 1969, p. 39-51)

MODÈLE PROBABILISTE ET PROBLÈME STATISTIQUE

par

J. L. PIEDNOIR

Cet exposé s'adresse à des étudiants en sciences humaines ayant déjà suivi un cours élémentaire de mathématiques et de probabilité (cf. M. BARBUT, *Mathématiques des Sciences Humaines*, P.U.F., 1967). Son but est de fournir à leur niveau une vue synthétique du problème statistique. Il peut servir également de résumé de cours.

I. — QUELQUES SITUATIONS CONCRÈTES.

A) *Traitements agricoles.*

Un organisme d'étude agricole veut déterminer la valeur de deux nouveaux engrais. Il procède à 10 essais dans 10 parcelles différentes pour l'engrais A et à 8 essais dans 8 parcelles différentes pour l'engrais B.

On veut comparer les résultats, savoir si les engrais sont de bons engrais, savoir quel est le meilleur.

Mais pour décrire les résultats, le langage littéraire habituel est inadéquat, il manque de précision; pour décrire le phénomène, il faut parler un autre langage, c'est le langage mathématique. Pour décrire la réalité, nous nous servirons de concepts mathématiques empruntés à une théorie, cette opération s'appelle « faire un modèle descriptif ».

Le rendement de chaque parcelle sera un nombre réel, les résultats de l'engrais A pourront donc être représentés par une suite de 10 nombres réels, ceci représente pour les mathématiciens un espace vectoriel : l'espace $\mathbf{R}^{10}$, de même les résultats de l'engrais B seront représentés par un vecteur de l'espace $\mathbf{R}^8$; finalement le résultat de l'expérience peut se représenter par un élément de l'ensemble $\mathbf{R}^{10} \times \mathbf{R}^8$, produit cartésien du vectoriel $\mathbf{R}^{10}$ et du vectoriel $\mathbf{R}^8$.

Par exemple : $(x, y) = [(50, 49, 47{,}5, 44{,}3, 46, 54, 52, 53, 45, 58)$
$(48, 47, 44, 40, 46{,}5, 39, 41, 34)]$.

Se servir de la structure vectorielle de $\mathbf{R}^{10}$ et de $\mathbf{R}^8$ permet de calculer certains indicateurs de chacun de ces vecteurs : la moyenne, la variance :

$$\mathrm{m}\, x = 49{,}88 \qquad \mathrm{m}\, y = 42{,}44$$
$$\mathrm{Var}\, y = 17{,}369 \qquad \mathrm{Var}\, y = 20{,}21.$$

On peut également considérer $\mathbf{R}^{10}$ et $\mathbf{R}^8$ uniquement comme des espaces produits cartésiens et se servir de la relation d'ordre de R et en utilisant cette autre structure mathématique, calculer d'autres indicateurs : la médiane

$$\mu x \in [49{,}50] \qquad\qquad \mu y \in [40{,}44]$$

Le rang de chacun des nombres $y_1 \ldots y_8$ parmi les $10 + 8$ observations totales; ainsi :

$$R_1 = 12;\ R_2 = 10;\ R_3 = 5;\ R_4 = 3;\ R_5 = 9;\ R_6 = 2;\ R_7 = 4;\ R_8 = 1.$$

Divers indicateurs de description sont possibles chacun étant lié au modèle descriptif employé.

Une telle description est indispensable pour connaître la réalité, mais elle est insuffisante pour une analyse en profondeur du phénomène, utilisable en particulier dans un but prédictif; et on ne peut répondre à la question : l'engrais A est-il meilleur que l'engrais B ? Pour continuer l'analyse au-delà de la description, il faut utiliser un modèle probabiliste, c'est ce que nous ferons dans la deuxième partie.

B) *Contrôle de fabrication.*

Un acheteur reçoit de son fournisseur un lot de 100 000 cartouches. Il se pose la question : parmi ces 100 000 cartouches, quelle est la proportion de mauvaises cartouches ? Une idée lui vient : il en prélève 100 en des endroits notés d'avance et les tire; il s'aperçoit que 30 cartouches sont mauvaises.

Pour décrire la réalité, il se sert d'un nombre compris entre 0 et 1 indiquant la proportion de mauvaises cartouches, elle est de 0,30. Comme précédemment, on a utilisé un langage mathématique pour décrire la réalité, langage certes beaucoup moins élaboré que dans le cas A); mais comme dans ce cas, pour une analyse plus profonde, il faut utiliser un modèle probabiliste.

C) *Enquête sociologique.*

Un sociologue veut savoir s'il y a une relation entre le niveau d'instruction et le niveau de rémunération des personnes actives en France. Son organisme ne lui fournissant que des crédits limités, il ne peut interroger les 20 millions de personnes actives, il en interroge un nombre limité : n, et pour chacune note le revenu et le niveau d'instruction. Là aussi, plusieurs modèles mathématiques sont possibles pour la description. Pour les besoins de l'enquête, il se contente de celui-ci : l'I.N.S.E.E. a opéré une classification du niveau d'instruction en 5 classes (I, II, III, IV, V) et du niveau de revenu en 8 classes (1, 2, 3, 4, ... 8).

Appelons : $I = \{ I, II, III, IV, V \}$,
$J = \{ 1, 2, 3, 4, 5, 6, 7, 8 \}$,

à chaque élément de $I \times J$ il affecte un nombre compris entre 0 et 1, f_{ij} qui représente la proportion de personnes interrogées ayant le niveau d'instruction i et le niveau de revenu j. Pour une description de la réalité, il peut aussi calculer les nombres :

$f_i = \sum_j f_{ij}$ proportion des personnes ayant l'instruction i

$f_j = \sum_i f_{ij}$ proportion des personnes ayant la rémunération j

$g_{ij} = \dfrac{f_{ij}}{f_i}$ proportion des personnes ayant la rémunération j parmi celles qui ont l'instruction i

$g'_{ij} = \dfrac{f_{ij}}{f_j}$ proportion de personnes ayant l'instruction i parmi celles ayant le revenu j.

Mais tout cela ne permet guère d'aborder le problème posé, à savoir : y a-t-il quelque relation entre les phénomènes I et J ?

II. — LE MODÈLE PROBABILISTE.

Les questions que l'on se pose étant sans réponse, nous allons bâtir un modèle mathématique, c'est-à-dire remplacer la réalité concrète par des « idées pures », des concepts mathématiques, sur lesquels nous pourrons raisonner, tirer des conséquences que nous appliquerons ensuite à la situation réelle. Une telle démarche est souvent hasardeuse, le passage au modèle implique des hypothèses très fortes sur le réel, hypothèses que l'on est en général incapable de vérifier. Le problème se pose toujours de la validité du modèle, de sa robustesse, ce n'est qu'en confrontant la réalité aux résultats tirés du modèle que l'on peut conclure.

Il existe de nombreux modèles mathématiques et à une même réalité on peut appliquer plusieurs modèles.

Pour les cas qui nous intéressent, les phénomènes ne sont pas, à l'évidence, déterministes dans le premier cas : la variété d'engrais ne détermine pas un seul et même rendement; dans le deuxième cas, si l'on tirait 100 autres cartouches, on aurait un nombre différent de mauvaises. De même dans le troisième cas, interroger 100 autres personnes changerait les f_{ij} de la description. Il faut donc adopter un modèle probabiliste.

L'hypothèse fondamentale que l'on fait est la suivante : on considère chaque résultat observé comme la réalisation concrète d'un aléa (cf. cours de probabilité). Il faut donc construire un aléa, c'est-à-dire un espace et une loi de probabilité dessus.

A) *Construction de l'espace.*

L'espace de notre modèle sera l'ensemble de tous les cas possibles et imaginables (ensemble des possibilités).

Dans le cas des traitements agricoles, on peut se servir comme espace fondamental de l'espace qui a servi à la description c'est-à-dire $\mathbf{R}^{10} \times \mathbf{R}^{8} = \mathbf{R}^{18}$.

Dans le cas des cartouches, chaque cartouche peut être bonne (noté 1) ou mauvaise (noté 0); pour une cartouche, l'espace serait $\{ 0,1 \}$; pour 100 cartouches $\{ 0,1 \}^{100}$.

Dans le cas de l'enquête sociologique, pour un individu l'ensemble des cas possibles est l'ensemble $I \times J$ comme il y a n individus l'ensemble des cas possibles est donc : $(I \times J)^n$.

B) *Probabilisation.*

La loi de probabilité est en général impossible à déterminer a priori; mais à l'aide d'hypothèses complémentaires, on peut préciser à quel type elle appartient.

Cette réduction est d'ailleurs indispensable pour que l'on ait quelque chance de répondre aux questions posées.

Les aléas à construire se présentent le plus souvent sous forme d'aléas produits, ce qui traduit le fait que l'on fait plusieurs expériences; on verra plus loin la raison profonde de ce fait intuitif. On considère l'aléa produit comme un *aléa produit d'aléas indépendants*, ce qui implique que le résultat d'une expérience ne dépend que de cette expérience et non pas des autres expériences. Une telle hypo-

thèse, très importante, exige pour sa réalisation concrète, une expérimentation bien conduite. On considère de plus que tous les aléas correspondant à des expériences semblables sont les mêmes, c'est-à-dire qu'ils ont même distribution. Cette hypothèse suppose que les conditions expérimentales sont rigoureusement les mêmes, ce qui évidemment, a des conséquences sur la manière de faire les expériences.

On peut maintenant, dans chaque cas, préciser la forme des différentes distributions de probabilité.

Cas des traitements agricoles : nous avons « arrondi » à des nombres entiers, chaque résultat et nous pouvons poser :

P_i = Probabilité pour que le rendement d'une parcelle traitée avec l'engrais A soit de i (c'est-à-dire compris entre $i - 0{,}5$ et $i + 0{,}5$, de même on pose :

g_j = Probabilité pour que le rendement d'une parcelle traitée avec l'engrais B soit de j (compris entre $j - 0{,}5$ et $j + 0{,}5$), et du fait des hypothèses faites, la probabilité d'un vecteur de $R^{10} \times R^8$ dont le résultat se résume sous forme :

$$[(i_1, i_2, \ldots, i_{10}), (j_1, \ldots, j_8)]$$

est :

$$p_{i_1} \times p_{i_2} \times p_{i_{10}} \times \ldots \times p_{i_{10}} \times q_{j_1} \times \ldots \times q_{j_8},$$

les probabilités se multiplient du fait de l'indépendance, la même loi entraîne la présence de $10p$ et de $8q$.

Cas du contrôle. On suppose toutes les cartouches prélevées dans les mêmes conditions ; si on appelle p la proportion de mauvaises cartouches dans le lot, la probabilité d'une séquence de 100 cartouches 110 ... 1 où il y a k zéros est :

$$p^k (1 - p)^{100-k}.$$

Cas de l'enquête sociologique. Si on appelle p_{ij} la proportion dans la population totale des individus ayant l'instruction i et le revenu j la probabilité pour que l'individu k ait le profil $i_k j_k$, k variant de 1 à n est :

$$\prod_{k=1}^{n} p_{i_k j_k}$$

C'est une probabilisation explicite, à partir des hypothèses générales, elle est rarement faite, en effet on se contente en général d'aléa image de ceux-ci et c'est cet aléa image que l'on explicite. Pourquoi ? Parce qu'il faut non pas connaître l'aléa (Ω, p) mais répondre d'une manière précise aux questions posées dans la première partie. Mais les questions posées ne sont pas précises. Pour y répondre, il faut les formuler dans le modèle mathématique.

C) *Formulation dans le modèle mathématique des questions posées.*

Les questions posées se formulent dans le modèle mathématique sous forme de conditions sur la distribution de probabilité, puisque dans le modèle probabiliste, la spécificité du réel est traduit par la distribution de probabilité sur l'ensemble fondamental.

Dans le *cas des traitements agricoles*, dire que les engrais A et B ont même effet sur le rendement peut se traduire mathématiquement par :

$$\boxed{p_i = q_i \ \forall\, i}$$

Cas des cartouches. Il faut déterminer $\boxed{p}$.

Cas de l'enquête. Si le revenu d'un individu est indépendant de son instruction, on a en posant :

$$p_{i\cdot} = \sum_j p_{ij} \qquad \text{et} \qquad p_{\cdot j} = \sum_i p_{ij}$$

$$\boxed{p_{ij} = p_{i\cdot} \cdot p_{\cdot j}}$$

qui traduit une indépendance en probabilité.

D) *Réduction du problème.*

Une fois la question posée, l'aléa (Ω, p) construit initialement peut paraître trop compliqué; on a la même information en utilisant un aléa image (Ω', p') plus simple. *Dans le cas des cartouches,* soit X l'application : à une suite, on lui affecte le nombre de cartouches loupées ainsi l'aléa (Ω', p') est :

$$\Omega' = (0, 1, 2, \ldots, k, \ldots, 100)$$
$$p' = (\ldots\ldots, C_{100}^{k}\, p^k (1-p)^{100-k} \ldots)$$

Dans le cas de l'enquête, on s'intéresse uniquement au nombre de personnes ayant le niveau d'instruction i et le revenu j. Soit n_{ij} et l'ensemble Ω' est l'ensemble de tous les tableaux de nombres n_{ij} où :

$$i \in I \qquad \text{et} \qquad j \in J.$$

avec :

$$\sum_i \sum_j n_{ij} = n$$

III. — LE PROBLÈME STATISTIQUE.

A) *Présentation.*

La question posée étant maintenant précisée sous forme d'une condition sur la loi de probabilité, il faut maintenant voir si l'expérience faite est compatible ou non avec la loi ou le type de loi envisagé; ou bien encore, si la question est posée comme une interrogation sur la loi de probabilité, tirer de l'expérience une « mesure » de la probabilité.

Ces problèmes reviennent donc à choisir une loi ou un type de loi de probabilité parmi ceux qui sont a priori possibles. Mais comme on est dans un modèle probabiliste, il est impossible d'avoir de bonnes certitudes et on ne pourra pas faire mieux que de choisir la loi ou le type de loi qui a le plus de chance de cadrer avec l'expérience. Mais ce mot : « le plus de chance » est vague, la formalisation statistique va lui donner plusieurs sens précis suivant le type de questions posées, suivant le genre d'idées que l'on se fait sur le problème.

B) *La décision dichotomique.*

Il s'agit ici de voir si l'expérience est compatible avec un type de loi possible ou avec un autre. C'est le cas des rendements agricoles : $p_i = q_i \; \forall i$, ou bien il existe i tel que $p_i \neq q_i$.

C'est le cas de l'enquête : $p_{ij} = p_{i\cdot}\, p_{\cdot j} \; \forall i \; \forall j$, ou bien $\exists i \; \exists j$ tels que $p_{ij} \neq p_{i\cdot}\, p_j$.

Pour faire comprendre le problème, nous allons commencer par des cas très simples et purement abstraits.

Supposons une variété de terre, des informations antérieures à l'expérience font penser que deux cas peuvent se présenter : ou bien le rendement est compris entre 30 et 40, la probabilité d'un rendement extérieur à ces limites est nulle (loi de type I), ou bien le rendement est compris entre 40 et 60 et la probabilité d'un rendement extérieur à ces limites est nulle (loi de type II). On fait une expérience, on trouve un rendement égal à 43, tout esprit bien constitué pensera que l'expérience est compatible avec une distribution de probabilité de type II et incompatible avec une distribution de type I.

— Cas plus compliqué. Les rendements possibles pour une terre donnée peuvent obéir à deux lois de probabilité p, q entre lesquelles il s'agit de choisir (R le rendement) :

	$R < 30$	$30 \leqslant R < 35$	$35 \leqslant R < 40$	$40 \leqslant R < 45$	$R \leqslant 45$
p	0,80	0,10	0,07	0,02	0,01
q	0,01	0,02	0,07	0,10	0,80

Si on trouve un rendement de 29 tout esprit normalement constitué pensera que l'expérience a plus de chances d'être compatible avec l'hypothèse p qu'avec l'hypothèse q, conclusion opposée si on trouve un rendement de 51. Mais avec le seul critère du « flair » le même esprit sera bien ennuyé pour conclure s'il trouve un rendement de 37. Pour s'en sortir, il faut une formalisation plus rigoureuse.

Plusieurs formalisations ont été proposées, la plus « générale » est celle de Wald dite de la décision statistique. Mais celle qui est la plus employée est celle de Neyman et Pearson (qui peut d'ailleurs être considéré comme résultant d'une particularisation de la précédente).

Formalisation de Neyman-Pearson.

1) L'idée fondamentale de Neyman, est de dire que les deux types de loi ne sont pas de même nature.

Pour moi, expérimentateur, il y a un type de loi, disons maintenant hypothèse, qui est plus important que l'autre, c'est l'hypothèse nulle. *On choisit une hypothèse nulle*, elle est choisie pour des tas de raisons :

— opinion d'un pontife;

— confort intellectuel;

— elle est déduite d'une théorie actuellement en vigueur;

— son abandon, quand elle est vraie, me coûterait plus cher que le fait contraire : l'accepter quand elle est fausse.

2) On veut se garantir contre un rejet trop fréquent de l'hypothèse nulle *quand elle est vraie.* Pour cela on choisit un nombre $\alpha \in [\,0, 1\,[$ et la probabilité de rejeter l'hypothèse nulle, notée H_0, quand elle est vraie, devra être inférieure à α. Soit C l'événement : rejeter H_0.

$$\forall P \in H_0 \qquad P(C) \leqslant \alpha$$

On marque $\forall P \in H_0$ car H_0 peut être formée de lois différentes (exemple : les traitements agricoles).

Le choix du seuil de confiance α est en général délicat. Grossièrement, et pour autant que le problème est suffisamment régulier, on peut dire que si α est très faible, j'ai peu de chance de rejeter l'hypothèse nulle, donc je choisirai α d'autant plus petit que j'ai plus de confiance en l'hypothèse H_0, α sera dans cette interprétation une « mesure » de ma confiance en H_0.

Mais cette interprétation n'est pas à l'abri de toute critique : elle fait intervenir des notions subjectives : la confiance en l'hypothèse nulle, extérieure au modèle explicité, mal formalisées et que l'on veut traduire par un chiffre précis, ce qui est pour le moins une contradiction logique. Le choix de α et son interprétation ne peuvent recevoir de réponse satisfaisante que dans le cadre général de la décision statistique de A. Wald.

3) L'événement C rejeté, H_0 est donc l'ensemble des résultats expérimentaux parmi tous ceux qui sont possibles, que l'on juge trop peu vraisemblables avec l'hypothèse H_0; C est donc une partie de Ω ($C \subset \Omega$). C est appelée la *région critique du test.* Dans le cas abstrait donné plus haut si :

p est appelé l'hypothèse nulle
$\alpha = 0{,}1$

alors :

$C = (R \geqslant 35)$ est une région critique possible.

La probabilité de C, si p est vraie, est bien égale à 0,1.

Remarques.

a) La région critique dépend de l'hypothèse nulle choisie, elle dépend aussi du seuil choisi;

b) Pour une hypothèse nulle et un seuil donné, quelle est la meilleure région critique ? Pour résoudre ce problème, il faut donner un critère précis pour comparer deux régions critiques. (Cf. plus loin, la remarque *a*) du point 4).) Avec les critères actuellement employés, il n'admet de solutions que dans quelques cas particuliers.

Illustrons la remarque *a*). Dans l'exemple abstrait, si p est l'hypothèse nulle et $\alpha = 0{,}03$, la région critique est $C = (R \leqslant 40)$; si q est l'hypothèse nulle (nous faisons valser les hypothèses) et $\alpha = 0{,}03$, la région critique est $C = (R \leqslant 35)$. Attention, la décision finale : accepter ou rejeter l'hypothèse nulle dépend bien entendu, du choix de l'hypothèse nulle et du seuil. Dans l'exemple précédent, si on a un constat expérimental de 37, avec p comme hypothèse nulle et $\alpha = 0{,}03$, on accepte l'hypothèse nulle p; avec q comme hypothèse nulle et $\alpha = 0{,}003$, on accepte q l'hypothèse nulle.

4) Dans la problématique Neyman-pearsonnienne on a privilégié une hypothèse appelée hypothèse nulle et on se garantit contre un type d'erreur (erreur I) : « rejeter l'hypothèse nulle quand elle est vraie »; mais il y a un deuxième type d'erreur (erreur II), « accepter l'hypothèse nulle quand elle est fausse ». Et si on se garantit beaucoup contre l'erreur I en fixant α très petit, on aura peu de chance de rejeter l'hypothèse nulle (la région sera très petite) et on sera amené à l'accepter plus souvent même si elle est fausse et donc à augmenter le risque d'erreur II.

Il y a donc un équilibre à trouver, qui dépend du problème posé, de l'opinion de l'expérimentateur sur H_0. Pour mesurer le risque d'erreur II, on est amené à poser :

$$\beta(P) = P(c) \qquad \forall P \in H_1$$

H_1 contre hypothèse β (P) s'appelle *la puissance du test*, c'est la probabilité d'accepter H_1 *quand la loi* $P \in H_1$ *est vraie.*

β est donc une fonction, β varie quand P varie dans H_1;

β n'est qu'un nombre que si H_1 ne contient qu'une distribution possible. Ainsi, dans l'exemple précédent, p est l'hypothèse nulle : $\alpha = 0{,}03 \qquad C = (R \geqslant 40)$.

Probabilité de C avec la distribution q donne la puissance : $\beta = 0{,}90$, et l'erreur du type II est : $1 - \beta = 0{,}10$.

Remarques.

a) Pour une hypothèse nulle donnée et un seuil donné si on hésite entre deux régions critiques C et C_1, la meilleure est celle qui donne la puissance la plus grande, puisque toutes choses égales par ailleurs, l'erreur de type II est plus faible.

b) Si avec un seuil $\alpha = 0{,}10$ on trouve une puissance toujours supérieure à 0,99, le problème est mal posé, car finalement l'erreur de type II étant très petite, ce n'est pas H_0 qui est favorisée, mais H_1 et tout revient à avoir choisi en fait H_1 comme hypothèse nulle. Il y a une cohérence interne à cette formalisation qu'il faut respecter.

5) Résumons la démarche.

Problème. — A-t-on une distribution de type H_0 ou une distribution de type H_1 ?

1. On choisit une hypothèse nulle H_0.

2. On choisit un seuil α.

3. On détermine une région C dans l'ensemble des possibilités tel que $\forall\, P \in H_0 \qquad P(C) \leqslant \alpha$

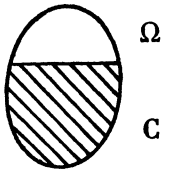

Fig. 1

4. Si $\omega \in C$, on rejette l'hypothèse nulle.

   Si $\omega \notin C$, on accepte l'hypothèse nulle.

5. On cherche la fonction puissance de C : $\beta(P) = P(C)$ si $P \in H_1$, à condition que cela ne soit mathématiquement pas trop compliqué.

Exemple : Les traitements agricoles.

Problème posé : $p_i = q_i \quad \forall\, i$ ou bien $\exists_i,\ p_i \neq q_i$.

a) On choisit comme hypothèse nulle que les engrais A et B ont des actions identiques, les distributions de type H_0 sont telles que : $p_i = q_i \quad \forall\, i$.

b) On choisit comme seuil $\alpha = 0{,}1$.

c) La détermination de la région critique est ici délicate, pour y arriver on code les situations. Coder, c'est affecter à chaque situation un nombre. Ici, ce sera un nombre entier, un code c'est donc une

application $U : \Omega \longrightarrow N$. Le code que nous adoptons ici est le suivant : sur les deux engrais qui ont une même action, les résultats sont très « mélangés », c'est-à-dire que parmi 10 + 8 observations totales ordonnées par ordre croissant, les observations y_j provenant de l'engrais B ne sont pas concentrées en un seul secteur.

Exemples : (I) *y xxxyyxxxxyyyyxxyyx*; par contre une observation (2) *yyyyyyyxyyxxxxxxxx* aurait tendance à prouver que l'engrais B est moins bon que l'engrais A. Pour mesurer ce mélange, on compte le nombre de transpositions U des y par rapport aux x

pour (1) $U = 45$,
pour (2) $U = 2$,

U peut prendre toutes les valeurs de 0 à 80.

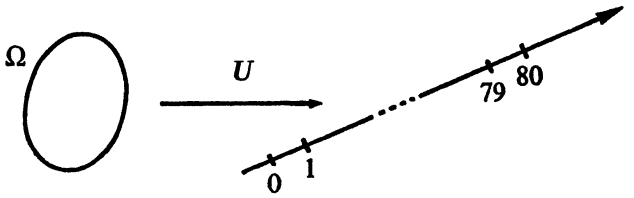

Fig. 2

Si U est trop petit ou trop grand, l'hypothèse nulle sera peu vraisemblable et donc rejetée, quand l'hypothèse nulle est vraie, l'aléa image de (Ω, P) par U : (Ω', P') est le suivant : $\Omega' = (0, 1, 2, \ldots, 80)$.

P' est une distribution que l'on trouve dans les tables et qui, fait remarquable, est la même quel que soit l'aléa P sur Ω pourvu qu'il fasse partie du type H_0 (les y et les x ont même distribution de probabilité) et c'est ce fait qui donne toute sa valeur à notre code.

On peut maintenant déterminer la région critique. La table dit que dans l'*hypothèse* H_0 la loi de probabilité marquée P_0 de U implique que :

$$P_0 [(21,79)] = 0,9$$

donc la région critique sera :

$$C = \{ \omega \in \Omega, \quad U(\omega) \leqslant 20 \text{ ou } \quad U(\omega) \geqslant 80 \}$$

d) La série expérimentale donnée ici est la suivante : $U(\omega) = 10$, d'où : au seuil $\alpha = 0,1$, on rejette l'hypothèse H_0.

e) Vu la complexité ici de l'hypothèse H_1, vrai fourre-tout, il est très difficile de calculer la puissance, ceci est possible pour quelques distributions de H_1 mais trop complexe pour figurer dans un enseignement élémentaire.

Remarque.

Pour tous les tests que nous ferons, la démarche sera la même, en particulier pour les problèmes complexes nous étudierons d'autres codes, le plus célèbre est celui dit du χ^2 (cf. Note de Barbut : une introduction au test du χ^2, *M.S.H.*, n° 14, pp. 23-30, 1966).

C) *Choix d'un paramètre.*

Dans l'exemple des cartouches, il ne s'agit pas comme précédemment, de choisir entre deux types de loi, mais à partir du fait expérimental constaté, choisir une loi parmi toutes celles qui sont possibles : ici choisir $p \in (0,1)$ le plus « vraisemblable ». C'est le problème de l'estimation. Résoudre le problème

de l'estimation c'est donc associer à tout résultat expérimental donné ω un nombre $\hat{p}(\omega)$ qui sera appelé l'estimateur de p. C'est donc faire une application : $\Omega \hat{p} \longrightarrow (0,1)$.

Le problème se pose de choisir, en un certain sens à préciser, la meilleure application possible. Pour le problème qui nous intéresse, les mathématiciens l'ont résolu, et d'abord avec le sens commun, ils ont démontré que pour un certain nombre de critères, la meilleure application $\hat{p}$ est la suivante :

$$\hat{p}(\omega) = \frac{\text{nombre de mauvaises cartouches dans la séquence } \omega}{\text{nombre total de cartouches tirées dans la dite séquence.}}$$

Dans l'essai ω effectué sur 100 cartouches, 30 sont mauvaises donc : $\hat{p}(\omega) = 0{,}30$ est une estimation de la proportion de mauvaises cartouches dans le lot global.

Mais un problème se pose : quelle confiance puis-je accorder à cette valeur ? Quelle est sa précision ? Précision probable bien entendue, car le modèle est un modèle probabiliste.

L'erreur commise en remplaçant p par $\hat{p}(\omega)$ est $\hat{p}(\omega) - p$ l'application qui a ω fait correspondre $\hat{p}(\omega) - p$ détermine un aléa image dont l'ensemble fondamental est $(-1, +1)$. Pour déterminer une borne à l'erreur probable $\hat{p} - p$, on choisit un seuil α et on cherche sur l'intervalle $(-1, +1)$ un intervalle $(-l, +l)$ tel que la probabilité de cet intervalle soit $1 - \alpha$. La détermination rigoureuse de l est délicate mais les mathématiciens ont montré que si n est assez grand, disons $n > 20$ et la valeur $\hat{p}(\omega)$ trouvée pas trop petite, disons $n\hat{p} \geqslant 7$ l'aléa image de (Ω, p) par l'application $(\hat{p} - p)\, 2\sqrt{n}$ est à peu près un aléa normal dont il existe des tables. Ainsi dans le cas qui nous intéresse où $n = 100$, si on prend $\alpha = 0{,}05$, les tables donnent $l' = 1{,}96$, et donc :

$$\text{Prob}\,(-1{,}96 < (\hat{p} - p)\, 2\sqrt{n} < 1{,}96) = 1 - \alpha$$

et comme $2\sqrt{n} = 20$,

$$l = \frac{l'}{2\sqrt{n}} = \frac{l'}{20} = 0{,}098.$$

L'intervalle aléatoire $(\hat{p} - l,\ p + l)$ recouvre donc la vraie valeur p choisie avec une *probabilité* $1 - \alpha = 0{,}95$, c'est ce que l'on traduit en disant une fois calculé $\hat{p}(\omega)$ pour la séquence expérimentale ω donnée que : avec une confiance de 0,95, $0{,}20 < p < 0{,}40$.

Attention : ne pas confondre probabilité et confiance.

D) *Enrichissement du modèle, la statistique bayesienne.*

Nous avons jusqu'à présent considéré implicitement la probabilité comme une caractéristique du modèle, rendant compte de certains phénomènes réels et donc un peu comme une donnée du réel. Mais la notion de probabilité peut avoir un autre statut épistémologique que celui-là, elle peut traduire non pas la nature des choses, mais l'opinion qu'une personne a sur les choses. Il s'agit de systématiser, grâce à la probabilisation, une opinion a priori. Ce concept a déjà fait son apparition avec l'interprétation du seuil de confiance et dans la formalisation de Neyman et Pearson.

Pour faire comprendre ce dont il s'agit, nous raisonnons sur un exemple qui n'a de valeur que didactique, mais qui permet des calculs simples.

L'idée générale est la suivante :

a) Différents états de la nature sont possibles, chacun étant caractérisé par une certaine distribution de probabilité. *Exemple :* une urne contient des boules marquées 0 et des boules marquées 1. Des renseignements précédents, montrent que trois compositions notées I, II, III sont a priori possibles,

chaque composition étant caractérisée par la proportion de boules de chaque type, ce qui caractérise également la probabilité de tirer une telle boule. Le tout est résumé par le tableau suivant :

Tableau I

	0	1
I	0,3	0,7
II	0,5	0,5
III	0,7	0,3

b) J'ai une opinion a priori sur les différents états de la nature que je traduis en affectant une distribution de probabilité à ces différents états, ce qui donne pour l'exemple considéré, le fait suivant :

Un coup d'œil rapide sur l'urne, me donne une opinion a priori sur la composition possible que je caractérise par la distribution de probabilité suivante :

Tableau II

État I	État II	État III
0,2	0,4	0,4

c) Je fais une expérience, elle me donne un certain résultat; dans l'exemple j'ai tiré une boule marquée 0, comment vais-je modifier mon opinion pour tenir compte de cette information nouvelle ? et obtenir ainsi une opinion a posteriori intégrant le résultat de l'expérience. La formule du Révérend Thomas Bayes sur les probabilités conditionnelles sera l'axe de la formalisation probabiliste de cette situation.

Posons : $\Omega = (0, 1)$
$\Theta = (\text{I, II, III})$

Considérons l'espace $\Omega \times \Theta$, le tableau I peut être interprété comme étant les probabilités conditionnelles d'une distribution de probabilité sur $\Omega \times \Theta$:

$$P(\omega = i \mid \theta = j) \qquad i = 0, 1 \qquad j = \text{I, II, III.}$$

Le tableau II de mon opinion a priori sont les probabilités dites marginales sur Θ :

$$P(\theta = j) \qquad j = \text{I, II, III.}$$

Dans ce schéma une modification rationnelle de mon opinion pour obtenir mon opinion a posteriori intégrant l'information « il a été observé $\omega = 0$ » est interprétée comme la recherche des probabilités conditionnelles suivantes :

$$P(\theta. = j \mid \omega = 0.)$$

La formule de Bayes démontrée dans le cours de calcul des probabilités permet d'écrire :

$$P(\theta. = j \mid \omega = 0) = \frac{P(\omega = 0 \mid \theta. = j) \times P(\theta. = j)}{P(\omega = 0)}$$

avec :

$$P(\omega = 0) = \sum_{j=I}^{j=III} P(\omega = 0 \mid \theta\text{-} = j) \,.\, P(\theta = j)$$

Et le problème est entièrement formalisé. Effectuons les calculs :

$$\begin{aligned} P(\omega = 0) &= 0{,}3 \times 0{,}2 + 0{,}5 \times 0{,}4 + 0{,}7 \times 0{,}4 \\ &= 0{,}06 + 0{,}20 + 0{,}28 \\ &= 0{,}54 \end{aligned}$$

$$P(\theta. = I \mid \omega = 0) = \frac{0{,}06}{0{,}54} = 0{,}11$$

$$P(\theta. = II \mid \omega = 0) = \frac{0{,}20}{0{,}54} = 0{,}37$$

$$P(\theta. = III \mid \omega = 0) = \frac{0{,}28}{0{,}54} = 0{,}52$$

Et mon opinion après information est donc résumée par le tableau :

Tableau III

État I	État II	État III
0,11	0,37	0,52

d) Au vu de cette opinion a posteriori c'est à moi d'en tirer les conséquences pour une action ultérieure.

Remarque.

L'opinion a posteriori est naturellement dépendante de l'opinion a priori et aussi au vu des mêmes expériences deux individus différents tireraient des conséquences différentes. Mais on démontre que plus les expériences faites sont nombreuses et donc apportent une information riche (ici il n'y a qu'une expérience !) plus les opinions a posteriori se ressemblent et ceci quelles que soient les opinions a priori prises au départ pourvu que celles-ci ne contiennent pas d'état probabilisé avec la probabilité 0. Les opinions a posteriori se ressemblent car elles tendent toutes vers la distribution suivante : 1 à l'état qui traduit la nature réelle de l'urne et zéro ailleurs quand le nombre d'expériences tend vers l'infini par exemple si l'état de l'urne est l'état II après un nombre infini de tirages mon opinion a posteriori sera la suivante :

Tableau IV

État I	État II	État III
0	1	0

quelque soit l'opinion a priori pourvu que celle-ci ne donne pas la probabilité 0 à l'un des états.

Ce théorème est quand même réconfortant pour l'esprit, même s'il est impossible de faire un nombre infini d'expériences.

E) *Statistique et grands nombres.*

Le théorème précédent montre qu'il est intéressant de voir l'évolution des procédures statistiques quand le nombre d'expériences tend vers l'infini.

Nous avons vu lors de la modélisation probabiliste que les aléas que l'on considère sont des aléas produits d'aléas indépendants et de même loi.

Ainsi, dans le cas des *traitements agricoles*, si on fait n_1 essais avec l'engrais A et n_2 essais avec l'engrais B, l'espace fondamental est $R^{n_1} \times R^{n_2}$ (cf. paragraphe I-A).) Pour décider entre l'hypothèse H_0 : les deux engrais ont même effet et l'hypothèse H_1 les deux engrais ont des effets différents, on se sert d'une région $C_{n_1 n_2} \subset R^{n_1} \times R^{n_2}$. Si l'expérience faite $\omega_0 \in C_{n_1 n_2}$, on accepte H_0, sinon on rejette H_0.

On démontre le théorème suivant : il existe une suite de régions $C_{n_1 n_2}$ qui sont telles que :

$$\begin{array}{lll} \forall\, P \in H_0 & P\,(C_{n_1 n_2}) & \longrightarrow 0 \\ \forall\, P \in H_1 & P\,(C_{n_1 n_2}) & \longrightarrow 1 \\ & \text{quand } n_1 & \longrightarrow \infty \\ & \text{et } n_2 & \longrightarrow \infty \end{array}$$

Théorème qui signifie donc qu'avec un nombre infini d'expériences, les erreurs de type I et de type II peuvent être nulles toutes les deux et donc que l'on peut décider sans ambiguïté entre H_0 et H_1. Dans les cas des cartouches, on démontre en calcul des probabilités, qu'avec n cartouches tirées :

$$\forall \in \text{Prob}\,[(\hat{p}_n - p) > \in] \underset{n \to \infty}{\to} 0$$

donc l'erreur devient nulle si on fait un nombre infini d'expériences.

CONCLUSION.

Tous les problèmes statistiques que l'on se pose sont résumés par la phrase : Trouver la « bonne » loi de probabilité ou le « bon » type de loi parmi celles ou ceux qui sont possibles.

Les procédures que nous avons données sont convergentes, c'est-à-dire qu'avec un nombre infini d'expériences, les résultats expérimentaux étant connu sans erreur, on arriverait à une réponse certaine. On peut dire aussi : plus le nombre d'expériences est grand plus la discrimination entre les types de lois possibles est facile. Mais on ne peut justifier la statistique par des considérations asymptotiques uniquement. Le but de la statistique, comme nous l'avons vu, est de proposer des procédures rationnelles pour répondre au mieux, suivant des critères donnés à l'avance, aux problèmes posés quand justement, une connaissance complète est impossible. Les théorèmes limites que nous avons suggérés disent simplement qu'avec un nombre infini d'expériences la précision, au sens des critères précédents, est parfaite.