REVUE DE STATISTIQUE APPLIQUÉE

J. ULMO

La décision statistique dans le cadre bayesien

Revue de statistique appliquée, tome 19, nº 3 (1971), p. 27-66

<http://www.numdam.org/item?id=RSA_1971__19_3_27_0>

LA DÉCISION STATISTIQUE DANS LE CADRE BAYÉSIEN

J. ULMO

Les méthodes statistiques dites classiques ou encore objectives se prêtent mal à la prise en compte explicite des informations ou des idées à priori que l'on peut avoir sur "l'état de la nature" c'est-à-dire sur les paramètres du modèle statistique considéré.

Dans les méthodes bayesiennes on introduit ces informations sous forme d'une distribution de probabilité sur l'espace des états de la nature.

Dans cet exposé on se propose de reprendre dans le cadre bayesien, les éléments de la théorie de la décision selon Wald et de montrer comment dans la recherche d'une "règle de décision de Bayes" on peut aussi choisir au préalable l'expérience à réaliser dans l'ensemble E des expériences possibles.

Par ailleurs, la décision bayesienne se prête particulièrement bien à l'expérimentation séquentielle.

Pour définir le modèle décisionnel bayesien, il a paru nécessaire de définir préalablement le modèle statistique décisionnel "classique" ; aussi le plan de cet exposé est le suivant :

I - Le modèle statistique décisionnel classique

II - Le modèle statistique de la décision bayesienne

III - Choix d'une distribution a priori pour le paramètre
Les conjuguées naturelles d'une famille de vraisemblances

IV - Application à un exemple. Estimation par "échantillonnage" de la moyenne μ d'une distribution normale de variance connue

V - Introduction à l'expérimentation séquentielle.

I - LE MODELE STATISTIQUE DECISIONNEL CLASSIQUE ([1], [2])

I.1 - Introduction

La décision statistique est une décision dans l'incertitude.

Elle est relative à une situation imparfaitement connue dans laquelle on a à prendre une décision dont les conséquences dépendent d'un paramètre inconnu θ ou "état" caractérisant la situation.

Le choix de la règle de conduite à adopter est supporté par des observations, résultats aléatoires d'une expérience.

C'est l'objet de la théorie de la décision statistique de formaliser mathématiquement le problème du choix des décisions adéquates dans un tel contexte aléatoire.

Toute formalisation mathématique reposant sur une certaine idéalisation de la réalité, il importe de préciser les concepts de base du modèle statistique.

Avant de définir ce modèle, nous considérerons deux exemples schématiques.

Exemple 1

Une machine fabrique automatiquement chaque jour un lot de pièces qu'un critère de conformité à l'usage auquel elles sont destinées permet de classer en bonnes ou mauvaises.

Le fabricant doit décider s'il livre le lot à son client (décision a_1) ou s'il le met au rebut (décision a_2), selon l'importance de la proportion p de pièces défectueuses du lot.

En effet, si le coût de mise au rebut d'une pièce est 1, le fabricant alloue une indemnité de 10 au client pour chaque pièce mauvaise livrée.

Le fabricant supporte ainsi un coût W(p, a) fonction de p et de la décision a prise, soit, si N est l'effectif du lot

$$W(p, a_1) = 10\ pN$$

$$W(p, a_2) = N$$

Si p était parfaitement connu, le choix de la décision à prendre résulterait de la comparaison des coûts résultants. Le fabricant déciderait :

$$a_1 \text{ (livraison du lot) si } p \leqslant 0,10$$

$$a_2 \text{ (rebut du lot) si } p > 0,10$$

Si N est grand, il est difficile d'examiner chaque pièce du lot et de déterminer ainsi la valeur de p.

Aussi le fabricant pratique-t-il un contrôle sur un échantillon d'effectif restreint, comportant n pièces prélevées au hasard dans le lot.

Le résultat du contrôle opéré sur l'échantillon peut être représenté par un ensemble de n variables x_i, $i = 1, 2, \ldots\ldots, n$, x_i prenant la valeur 1 si la $i^{\text{ème}}$ pièce est mauvaise et 0 si elle est bonne, c'est-à-dire par une variable vectorielle $\underline{X}$ de composantes X_i, $i = 1, \ldots\ldots, n$.

La démarche de la décision statistique consiste à faire choix avant d'effectuer le contrôle d'une règle de conduite ou "règle de décision" associant à chaque valeur possible pour $\underline{x}$ une décision a.

Dans l'exemple considéré ici, il paraît naturel de baser la règle de décision sur le nombre $k = \sum_{i=1}^{n} x_i$ des pièces défectueuses de l'échantillon.

Une règle de décision δ pourra être définie par :

$$\delta(k) \Longrightarrow a_1 \quad \text{si } k \leqslant k_o$$

$$\delta(k) \Longrightarrow a_2 \quad \text{si } k > k_o$$

Pour que δ soit complètement définie il reste à fixer la valeur critique k_o.

Pour choisir k_o, et même plus généralement pour choisir la règle de décision δ à adopter, on est amené à faire intervenir l'espérance mathématique des coûts entraînés par les décisions prises dans le cadre de la règle δ. Cette espérance qui dépend de p est " la fonction de risque attachée à la règle δ".

En effet, si n est petit devant N, les valeurs x_i associées aux pièces de l'échantillon peuvent être considérées comme des réalisations indépendantes d'une variable aléatoire de Bernoulli Y telle que $P_p\,[Y = 1] = p$, et k est une réalisation d'une variable aléatoire binomiale k d'ordre n et de paramètre p, telle que $P_p\,[K = k] = \binom{n}{k} p^k (1 - p)^{n-k}$

La décision prise $a = \delta(k)$ est donc une réalisation d'une variable aléatoire A telle que :

$$P_p\,(A = a_1) = P_p\,(k \leqslant k_o) = \sum_{k=o}^{k_o} \binom{n}{k} p^k (1 - p)^{n-k}$$

$$P_p\,(A = a_2) = P_p\,(k > k_o) = 1 - P_p\,(a_1)$$

et le coût associé W(A, p) est aussi une variable aléatoire.

On ne peut donc prendre comme critère de classement des règles de décision, le coût associé au résultat d'un contrôle particulier ; il paraît naturel de prendre comme critère la valeur moyenne de ce coût pour tous les contrôles possibles sur des lots de proportion de pièces défectueuses p fixée. Cette espérance mathématique de W(A, p) est la fonction de risque de la règle δ.

$$R(\delta, p) = 10p\ NP_p\,(a_1) + NP_p\,(a_2)$$

$$= 10p \sum_{k=o}^{k_o} \binom{n}{k} p^k (1 - p)^{n-k} + N \sum_{k=k_o+1}^{n} \binom{n}{k} p^k (1 - p)^{n-k}$$

Il n'existe pas de règle dont le risque soit minimum pour toute valeur de p comprise entre 0 et 1, c'est-à-dire pas de règle de décision optimale. Il est donc nécessaire de rechercher un autre critère de choix pour δ.

Remarque :

On peut montrer qu'à toute règle de décision basée sur $\underline{X}$, on peut associer une règle de décision au moins aussi bonne pour toute valeur de p, et basée sur la statistique k (c'est dû à ce que k est un résumé exhaustif de l'échantillon $\underline{X}$ pour p). Ceci justifie la recherche de la règle de décision à adopter parmi celles qui sont basées sur k.

La méthode conventionnelle consiste à limiter supérieurement le risque du fournisseur de rebuter un lot de qualité p_o qu'il aurait nettement avantage à livrer, c'est-à-dire à fixer un seuil α à $P_{p_o}(a_2)$ et à déterminer k_c de façon à ce que le "risque du client" de recevoir un lot de qualité $p_1 > p_o$ soit le plus petit possible. On montre que c'est avec une règle du type précédent[1] que ce résultat sera obtenu et que si $P_{p_1}(a_2)$ est minimum pour une valeur p_1, il l'est pour tout $p_1 > p_o$. La donnée de α et p_o détermine complètement la règle sans qu'il soit nécessaire de fixer la fonction de coût.

La méthode décisionnelle à laquelle est consacrée cette étude utilise la connaissance a priori que le fabricant a sur sa fabrication donc sur p pour choisir k_o et plus généralement δ.

Exemple 2 - Problème de commande ou d'estimation

Un marchand de journaux désire fixer le montant C_j de sa commande d'un quotidien déterminé pour chaque jour j de la semaine.

Si m_j est la vente moyenne de ce quotidien pour le jour j, la conséquence de son choix se traduit par une perte

$$W(C_j, m_j) = k|C_j - m_j| \text{ où k est une constante.}$$

(On admet que la perte par invendus est égale au manque à gagner par manque de marchandise).

Si le marchand connaissait la valeur de m_j, il prendrait naturellement $C_j = m_j$.

Afin d'avoir une information sur m_j, le marchand a noté les ventes x_i effectuées pendant n semaines, dont la distribution peut être considérée en première approximation comme une distribution de Laplace-Gauss de moyenne m_j, d'écart-type σ indépendant de m_j.

Quelle règle de décision δ (ou encore quel estimateur de m_j) :

$$(x_1, x_2, \ldots, x_n) \longrightarrow C_j$$

doit-il choisir ?

Si on admet que les ventes de n semaines sont des variables aléatoires indépendantes, on peut encore montrer que les meilleurs estimateurs peuvent être recherchés parmi ceux qui sont basés sur la moyenne $\overline{X}$ des observations X_i ($\overline{X}$ est un résumé exhaustif des observations pour m_j) et que dans la classe des estimateurs sans biais de m_j (c'est-à-dire tels que :

$$E[\delta(X_1, X_2, \ldots, X_n)] = m_j),$$

----

(1) Ou plus précisément avec une règle de décision "aléatoire" qui s'en déduit, en remettant pour $k = k_o$, la décision entre a_1 et a_2 au résultat d'un tirage au sort avec probabilité $\Pi(\alpha, p)$ de décider a_1.

$\overline{X}$ est un estimateur optimum indépendamment de la fonction de coût adoptée, à condition que celle-ci soit, pour toute valeur de m_j, une fonction convexe de C_j.

1.2 - Les données de base dans un problème de décision statistique

Dans un problème de décision statistique, on doit donc prendre une décision a dont les conséquences c dépendent non seulement de a, mais aussi de l'état de la nature θ, qui est inconnu du décideur.

Pour obtenir une information sur θ, on procède à une expérience ou épreuve aléatoire e, dont le résultat se traduit par une variable aléatoire X(e) dont la loi de probabilité dépend de θ.

La démarche statistique consiste à décider de la conduite à tenir, non pas après examen du résultat obtenu mais avant que ce résultat soit connu. On est ainsi conduit à définir une règle de décision δ associant à chaque valeur possible pour X une décision a. Un modèle de décision statistique comporte donc la définition des éléments suivants :

. L'espace Θ des états de la nature, ou des paramètres caractéristiques de ces états.

. L'espace mesurable $(\mathcal{X}, \mathcal{B})$(1) des résultats X de l'expérience e, ou des "observations".

. X est une variable aléatoire dont la loi de probabilité p_X^θ dépend de l'état de la nature θ.

. La famille des lois de probabilité p_X^θ de X fait partie de la définition du modèle.

. L'espace mesurable $(A, \mathcal{A})$ des décisions.

Une règle de décision δ est alors une application de $\mathcal{X}$ dans A. Pour que la variable aléatoire $\delta(X)$ puisse être définie, il faut que l'espace A et l'application δ soient mesurables, ce qui entraîne la nécessité de faire choix d'une tribu $\mathcal{A}$ de parties de A.

L'ensemble Δ des règles de décision possibles sera alors l'ensemble des applications mesurables δ de $(\mathcal{X}, \mathcal{B})$ dans $(A, \mathcal{A})$, telles que la fonction de risque $E_X[W(\delta(X), \theta)]$ existe pour tout $\theta \in \Theta$.

. Une fonction de coût (ou de perte) $W(a, \theta)$, fonction numérique définie sur l'espace produit $A \times \Theta$ qui permet de comparer entre elles les conséquences de deux ou plusieurs décisions pour chaque état de la nature (quand l'état de la nature est θ, a_1 est préférable à a_2 si $W(a_1, \theta) < W(a_2, \theta)$).

Dans l'exemple 1, l'état de la nature est représenté par la proportion p de pièces défectueuses du lot et Θ est l'intervalle $[0, 1]$, tandis que A est réduit à deux éléments

a_1 : livrer le lot

a_2 : le mettre au rebut.

----

(1) $\mathcal{B}$ est la "tribu" des événements ou parties de $\mathcal{X}$ auxquelles on peut associer une probabilité.

L'expérience e consiste à prélever au hasard dans le lot un échantillon de taille n fixée et le résultat de l'expérience est la variable aléatoire K, nombre de pièces défectueuses de l'échantillon dont la loi de probabilité dépend de p.

Dans l'exemple 2, l'état de la nature est représenté par m_j, vente moyenne du quotidien au jour j de la semaine. Θ est la droite réelle $R = (-\infty, +\infty)$, tandis que $A = \Theta$.

L'expérience e est le relevé des ventes pendant n semaines, et le résultat X est la variable aléatoire vectorielle à valeurs dans $R^n = [-\infty, +\infty]^n$, dont les composantes sont les ventes $(X_1, X_2, \ldots\ldots, X_n)$ des n semaines et dont la loi de probabilité $p_X^{m_j}$ est la loi "produit $(p_{X_i}^{m_j})$" de n lois de Laplace Gauss de moyenne m_j et d'écart-type σ.

Choix d'une règle de décision

Comme on l'a déjà indiqué dans l'exemple 1, au résultat aléatoire X, la règle de décision δ, associe la variable aléatoire $\delta(X)$ à valeurs dans A, et le critère de choix associe naturellement à la fonction de coût $W(a, \theta)$, la fonction de risque $R(\theta, \delta) = E[W(\delta(X), \theta)]$, valeur moyenne de la variable aléatoire $W[\delta(X), \theta]$ pour la loi p_X^{θ}(1). Comme il n'existe généralement pas de règle de décision optimale pour tout $\theta \in \Theta$, on a le choix entre deux types de méthodes :

La première consiste à restreindre l'ensemble Δ des règles de décision. C'est ce qui a été fait dans l'exemple 2 où on a imposé aux estimateurs de m_j d'être sans biais.

La deuxième consiste à utiliser un critère moins sévère que celui de la fonction de risque dans son ensemble, mais qui soit tel qu'une règle de décision optimale pour la fonction de risque soit également optimale pour ce critère. C'est ce qu'on fait avec le critère du minimax(2) et dans la décision bayesienne à laquelle est consacré cet exposé.

1.3 - Le principe de la décision bayesienne

Le principe de la méthode consiste à ne pas attribuer le même poids à toutes les valeurs de θ dans l'utilisation du critère que constitue la fonction de risque, mais à privilégier des valeurs de θ qu'on considère comme plus probables au détriment d'autres, c'est-à-dire à définir une loi de probabilité p'_θ ou loi de probabilité a priori de θ, sur l'espace Θ des états de la nature, qu'il faut pour cela munir d'une tribu d'événements $\mathcal{C}$.

θ est alors une variable aléatoire, et il en est de même de $R(\theta, \delta)$. Il paraît alors naturel d'adopter comme critère de choix d'une règle de décision la valeur moyenne $R(\theta, \delta)$ de la fonction de risque $R(\theta, \delta)$ pour la loi p'_θ de θ.

Une règle de décision δ qui minimise $R(\theta, \delta)$ est dite "règle de Bayes" pour la loi a priori p'_θ.

----

(1) La fonction $W(a, \theta)$ doit être mesurable sur $(A, \mathcal{A})$ pour tout θ, pour que $W[\delta(X), \theta]$ soit une variable aléatoire.

(2) Le critère du minimax consiste à choisir une règle de décision δ qui minimise le maximum de la fonction de risque $R(\theta, \delta)$ quand θ parcourt Θ.

C'est ainsi que dans l'exemple 1, on peut considérer que la proportion p de pièces défectueuses du lot dont on a à décider l'acceptation ou le rejet est une réalisation de la variable aléatoire P, proportion de défectueux des lots susceptibles d'être fabriqués par la machine. dont la loi de probabilité est inconnue mais peut être évaluée à l'aide des résultats obtenus sur les lots fabriqués antérieurement par cette machine ou par d'autres semblables.

Dans d'autres cas, la nature aléatoire de θ est plus discutable, mais l'introduction d'une distribution de probabilité sur Θ permet de prendre en compte de façon explicite des informations ou des idées a priori (préalables à l'expérimentation envisagée) que l'on peut avoir sur l'état de la nature .

Les tenants des méthodes bayesiennes admettent en général la possibilité pour une personne donnée de définir une probabilité "subjective" ou "personnelle" sur $\mathcal{C}$ qui mesure le degré de confiance qu'a cette personne dans la réalisation des événements de $\mathcal{C}$

Une personne peut ainsi définir sa probabilité qu'il fasse beau demain , cette probabilité pouvant bien entendu différer d'une personne à l'autre.

Afin de prouver que leur affirmation n'était pas gratuite certains "subjectivistes" parmi lesquels nous citerons B. de Finetti, puis plus récemment L.J. Savage[1], ont effectivement construit une mesure de probabilité sur Θ en partant d'une relation de préférence sur les décisions. Savage admet en effet qu'il est possible à un individu d'ordonner totalement ses décisions. C'est ainsi que, dans le problème du choix de matériaux pour construire un ouvrage d'art (pont, barrage, etc...), on peut vouloir attribuer des probabilités de tenue en service à différents ensembles de matériaux envisagés. Si la personne qui est intérrogée à ce sujet préfère utiliser tel ensemble de matériaux plutôt que tel autre, on en déduit qu'elle attribue une plus grande probabilité de tenue en service au premier ensemble de matériaux qu'au deuxième. Des considérations de ce genre permettent de définir une probabilité "qualitative" sur $\mathcal{C}$, relation d'ordre total sur les ensembles de $\mathcal{C}$, à laquelle on peut associer une mesure de probabilité dite "compatible" avec cette probabilité qualitative.

Nous allons exposer en détail dans le cadre bayesien les éléments de statistique décisionnelle définis en I.2, en les élargissant pour montrer comment dans la recherche d'une règle de Bayes on peut aussi choisir au préalable l'expérience e à réaliser dans l'ensemble E des expériences possibles.

Dans l'exemple 1, on verra comment la prise en compte du coût du contrôle permet de déterminer non seulement la règle de décision à adopter mais aussi la taille optimale n de l'échantillon à prélever qui caractérise dans le cas présent l'expérience e à réaliser. E est ici l'ensemble des nombres entiers réels, positifs ou nuls, 0, 1, 2, ..., N.

En effet la décision statistique ne consiste pas seulement à choisir une règle de décision à la suite d'une expérimentation aléatoire particulière, mais également à choisir les expériences, et aussi les modèles interprétatifs, qui permettront de décider au mieux.

----

(1) The Foundations of Statistics - Wiley.

II - LE MODELE STATISTIQUE DE LA DECISION BAYESIENNE

II. 1. - Les éléments de base du modèle

II. 1.1. - Les espaces

Aux espaces $\mathcal{X}$ et A déjà définis en 1.2, il convient d'ajouter : l'espace E des expérimentations possibles.

Une expérimentation peut comporter plusieurs expériences élémentaires indépendantes du même type, comme c'est le cas dans l'exemple 1, où une expérience élémentaire consiste dans le prélèvement d'une pièce et l'examen de sa qualité : bonne ou mauvaise ; elle peut aussi plus généralement comporter des expériences élémentaires de type distinct, comme c'est le cas dans le diagnostic médical, ou une expérience peut par exemple être l'ensemble de plusieurs analyses de laboratoire et radiographies, ou plus généralement de plusieurs examens.

L'espace $\mathcal{X}$ de tous les résultats ou observations possibles, généralisation de l'espace des observations $\mathcal{X}$ défini en I.2.

A chaque expérience $e \in E$, correspond un espace probabilisable $(\mathcal{X}(e), \mathcal{B}(e))$ des observations ou résultats possibles pour cette expérience .

Il convient de remarquer qu'une même valeur x correspondant à deux expériences différentes e et e', correspond à la réalisation de deux événements distincts et doit, de ce fait, être représentée par deux points différents de $\mathcal{X}$.

C'est ainsi que dans l'exemple A à une même valeur x = k du nombre des pièces défectueuses observée sur deux échantillons de tailles différentes n et n' prélevés dans le lot, correspondent deux points différents (k, n) et (K , n') de $\mathcal{X}$; pour une même valeur de p, les probabilités de ces deux résultats sont différentes. [On rappelle que $p(k, n, p) = \binom{n}{p} p^k(1-p)^{n-k}$] . Il est donc indispensable de représenter les points de $\mathcal{X}$ par un couple de symboles. On utilisera par la suite l'une des deux notations x(e) ou (x, e).

$\mathcal{X}$ est la famille des espaces $\mathcal{X}(e)$, $e \in E$[1].

Définition 1 - Deux expériences e et e' sont "indépendantes" si leurs résultats X(e) et X(e') sont deux variables aléatoires indépendantes en probabilité.

Dans l'exemple A, le prélèvement au hasard de deux échantillons de tailles n et n' constitue deux expériences indépendantes.

L'expérience (ee') composée des deux expériences indépendantes e et e' , est l'expérience constituée par l'ensemble des expériences élémentaires constituant e et e' (Dans l'exemple 1 c'est le prélèvement d'un échantillon de taille n + n').

E doit être tel que $ee' \in E$ et la loi de composition τ, $e \,\tau\, e' \Longrightarrow (e, e')$ serait à étudier de façon plus précise[2].

----

(1) Du point de vue mathématique $\mathcal{X}$ est la "somme topologique" ou "l'atlas" des $\mathcal{X}(e)$, $e \in E$. Les espaces $\mathcal{X}(e)$ sont les "cartes" de cet atlas.

(2) Cette loi de composition est associative et pourvue d'un élément neutre (ne pas faire d'expérience), ce qui donne à E une structure de "monoïde" avec élément neutre.

Généralisation de l'espace Δ des règles de décision possibles :

Un élément δ de Δ est une règle de choix d'une décision a à la suite du résultat (x,e) d'une expérience e $\in$ E. On a a = δ(x, e) et Δ est l'espace des applications mesurables de la famille $\mathscr{X}$ des $\mathscr{X}$(e), e $\in$ E dans A. Un élément δ de Δ est une famille δ_e, e $\in$ E d'applications des $\mathscr{X}$(e), e $\in$ E dans A(1).

Dans l'exemple 1, une règle de décision δ est une famille δ_n, $n \in \{0,1,\ldots, N\}$ de règles de décision attachées à chaque taille n d'échantillon, les deux décisions à considérer étant :

$$\begin{cases} a_1 \text{ : accepter le lot} \\ a_2 \text{ : refuser le lot} \end{cases}$$

Avec la fonction de coût choisie (cf. 1.1) les règles de décision à envisager sont du type :

$$\text{si} \begin{cases} k(n) \leqslant k_0(n) \Longrightarrow a_1 \\ k(n) > k_0(n) \Longrightarrow a_2 \end{cases}$$

où k(n) désigne le nombre des pièces défectueuses observées sur le "n échantillon".

II. 1.2 - Les lois de probabilité

A la famille des lois de probabilité de X(e) qui est fonction de l'état de la nature θ et de l'expérience e, et qu'on notera $p_{x/\theta,e}$ pour indiquer que, quand θ et e, sont fixés, $p_{x/\theta,e}$ est une mesure de probabilité sur $\mathscr{X}$(e), il y a lieu d'ajouter la loi de probabilité a priori de θ, soit p'_θ.

Quand l'expérience e est fixée, la donnée de la loi a priori de θ, et de la famille de lois $p_{x/\theta,e}$ que l'on peut considérer comme les probabilités de X(e) conditionnées par Θ, définit une distribution de probabilité sur l'espace produit Θ x $\mathscr{X}$(e) soit $p_{\theta,x/e}$ dont p'_θ est la distribution marginale sur Θ.

De cette distribution produit on déduit, la loi marginale de X(e) soit $p_{x/e}$ et surtout la loi de θ conditionnée par le résultat x(e) soit $p''_{\theta/x,e}$ (2).

Définition de la loi de probabilité a posteriori de θ

C'est cette distribution de θ conditionnée par le résultat x(e), qu'on appelle "distribution a posteriori" de θ correspondant au résultat ou à l'observation x(e).

----

(1) On précisera en II. 2 une condition complémentaire à laquelle doivent satisfaire les éléments de Δ. Par ailleurs en statistique classique on est amené à considérer des règles de décision aléatoires ou mixtes. Dans la décision bayesienne c'est pratiquement inutile car on peut montrer sous des conditions très générales, qu'il existe toujours une règle simple au moins aussi bonne qu'une règle aléatoire ou mixte donnée.

(2) Dans toute la suite le ' sera réservé aux distributions à priori, tandis que le " sera réservé aux distributions à posteriori.

On se borne ici, au seul cas où, pour tout $e \in E$, les probabilités P'_θ et $P_{x/\theta, e}$ admettent respectivement des densités par rapport à des mesures naturelles de Θ et $\mathfrak{X}(e)$. La loi $P_{\theta, x/e}$ sur $\Theta \times \mathfrak{X}(e)$ admet alors une densité par rapport à la mesure produit et il en est de même des lois $P_{x/e}$ et $P''_{\theta/x, e}$. On désignera ces densités respectivement par $D'(\theta)$, $D(x/\theta, e)$ etc...

La densité a posteriori de θ pour le résultat x(e) est alors donnée par la formule de Bayes

$$D''(\theta/x, e) = \frac{D'(\theta), \ D(x/\theta, e)}{\int_\Theta D(x/\theta, e) \ D'(\theta) \ d\mu(\theta)} \tag{1}$$

en désignant par $\mu(\theta)$ la mesure adoptée sur Θ.

Remarques -

1/ $\int_\Theta D(x \theta, e). D'(\theta) \ d\mu(\theta)$ existe par hypothèse puisque c'est $D(x/e)$, densité de la loi marginale de X(e) au point x(e).

2/ Il résulte de (1) que $D''(\theta/x, e)$ n'est pas modifiée si on remplace $D(x/\theta, e)$ par $D^*(x/\theta, e) = \lambda(x, e) \ D(x/\theta, e)$ avec $\lambda(x, e) \neq 0$, $\forall (x, e) \in \mathfrak{X}$.

On peut en déduire que la densité a posteriori de θ, $D''(\theta/x, e)$ ne dépend pas du choix de la mesure de référence ν utilisée sur $\mathfrak{X}$ pour exprimer les densités $D(x/\theta, e)$ des lois $P_{x/\theta, e}$.

En effet, si on remplace la mesure ν par une mesure ν^* qui la domine et si :

$$\lambda(x, e) = \frac{d \ \nu(x, e)}{d \ \nu^*(x, e)}$$

est la densité de ν par rapport à ν^*, la densité de la loi $P_{x/\theta, e}$ par rapport à la mesure ν^* est $D^*(x/\theta, e) = \lambda(x, e) \ D(x/\theta, e)$.

On peut généraliser ce résultat au cas d'une mesure ν^* qui ne domine pas ν en introduisant une mesure $\nu^{**} = a + a^* \nu^*$ (a et a^* constantes positives), qui domine ν et ν^*.

Cette remarque sera utilisée ultérieurement lors de l'étude des statistiques exhaustives et de celle des conjuguées naturelles.

Exemple : Dans l'exemple 1 on peut prendre pour distribution a priori de p sur (0,1), une distribution béta du premier type de densité :

$$D'(p) = \text{cte} \ p^{\tilde{k}-1} (1-p)^{\tilde{n}-\tilde{k}-1}$$

où $\tilde{k}$ et $\tilde{n}$ ($\tilde{n} > \tilde{k} > 0$) sont deux constantes que l'on fixe de façon que la distribution ajustée rende compte au mieux des informations qu'on a sur p[1].

----

(1) On verra en III-1 et III-2 les raisons du choix d'une telle distribution à priori.

On peut prendre par exemple $\tilde{n} = 60$, $\tilde{k} = 3$, ce qui donne une moyenne $E'(p) = \tilde{k}/\tilde{n} = 0,05$, une valeur modale de l'ordre de 3 % et une probabilité négligeable d'avoir $p > 15$ %.

On obtient alors :

$$D''(p/k, n) = \frac{p^{k+\tilde{k}-1} (1-p)^{n+\tilde{n}-(k+\tilde{k})-1}}{\int_0^1 p^{k+\tilde{k}-1} (1-p)^{n+\tilde{n}-(k+\tilde{k})-1} dp}$$

soit

$$D''(p/k, n) = \frac{1}{\beta[k+\tilde{k}, n+\tilde{n}-(k+\tilde{k})]} p^{k+\tilde{k}-1} (1-p)^{n+\tilde{n}-(k+\tilde{k})-1}$$

c'est-à-dire la même distribution β du 1er type mais dont les paramètres sont maintenant $k + \tilde{k}$ et $n + \tilde{n} - (k + \tilde{k})$ au lieu de $\tilde{k}$ et $\tilde{n} - \tilde{k}$.

$$\left[B(a, b) = \frac{(a-1)!(b-1)!}{(a+b-1)!} \right]$$

La moyenne et la variance de la loi a priori de p sont respectivement égales à $E'(p) = \frac{\tilde{k}}{\tilde{n}}$ et $V'(p) = \frac{\tilde{k}(\tilde{n}-\tilde{k})}{\tilde{n}(\tilde{n}+1)}$, soit pour $\tilde{k} = 3$, $\tilde{n} = 60$, $V'(p) \simeq 8.10^{-4}$ et $\delta'(p) \simeq 0,028$, tandis que pour la loi à posteriori, on a :

$$E''(p/k, n) = \frac{k+\tilde{k}}{n+\tilde{n}} \qquad V''(p/k, n) = \frac{(k+\tilde{k})[n+\tilde{n}-(k+\tilde{k})]}{(n+\tilde{n})\ (n+\tilde{n}+1)}.$$

Quand n augmente indéfiniment $V''(p/k, n)$ tend vers 0, et ceci quelque soit k. Quant à la moyenne $E''(p/k, n)$ elle a même limite que $\frac{k}{n}$. Elle converge donc presque sûrement vers la proportion réelle ϖ de pièces défectueuses. La distribution a posteriori de p converge donc vers la variable certaine ϖ, quelles que soient les valeurs adoptées pour les paramètres $\tilde{k}$, $\tilde{n}$ de sa distribution a priori.

II. 1.3 - La fonction de coût

W définie en 1.2 doit être généralisée pour tenir compte du coût de l'expérimentation à choisir ; ce coût peut également dépendre du résultat x(e) de cette expérimentation. (Dans le diagnostic médical le coût d'une analyse dépend souvent de son résultat).

Exemple : Dans les problèmes de décision à partir d'un n-échantillon d'observations indépendantes comme ceux des exemples 1 et 2 dans lesquels on se propose également de déterminer la taille n de l'échantillon à prélever, il peut être utile d'ajouter au coût résultant du choix d'une décision une fois l'expérience réalisée, le coût de l'expérimentation proprement dite C(e, x) que l'on pourra prendre de la forme :

$$\begin{aligned} C(e, x) &= a + bn \quad \text{si } n > 0 \\ &= 0 \quad \text{si } n = 0 \end{aligned}$$

a représente le coût fixe de l'échantillonnage, tandis que b est le coût additionnel par unité.

Si on se propose d'estimer un paramètre réel θ, et si on adopte pour les erreurs d'estimation une fonction de coût quadratique $W_1(\theta, \hat{\theta}) = k(\theta-\hat{\theta})^2$, la fonction de coût résultante sera alors :

$$W(n,\hat{\theta},\theta) = k(\theta-\hat{\theta})^2 + a\ \delta(n) + bn$$

avec : $\delta(n) = 1$ si $n > 0$
$\delta(n) = 0$ si $n = 0$
W n'est nulle que si $n = 0$ et $\hat{\theta} = \theta$.

De façon générale la fonction de coût sera une fonction numérique dépendant de e, x(e), a, θ qu'on notera $W(e,x,a,\theta)$, telle que pour tout $a \in A$ et toute $\in E$, $E_{\theta, x/e}[W(e,x,a,\theta)]$ existe(1).

Quand on a $W(e, x, a, \theta) = W(a, \theta) + C(e, x)$, c'est-à-dire quand les coûts dus à la décision proprement dite et à l'expérimentation s'additionnent, on dit qu'on a une fonction de coût "additive".

Remarque -

Dans la décision bayésienne on introduit souvent une fonction d'utilité" plutôt qu'une fonction de coût. L'utilité varie dans le même sens que la préférence entre les différents points (e,x,a,θ) possibles, tandis que la fonction de coût établit un préordre opposé à la préférence et est généralement positive ou nulle. Une utilité associée au coût W sera donc $u = -W$ ou toute fonction de la forme $u' = Ku + K'$, K et K' constantes, $K > O$, c'est-à-dire de la forme $u' = -KW + K'$, $K > O$.

II. 2 - Processus de choix de l'expérience et de la règle de décision

Les étapes de l'expérimentation sont les suivantes :

On procède à une expérience $e \in E$

On obtient un résultat $x \in \mathfrak{X}(e)$ réalisation d'une variable aléatoire X (e) de loi Px/e.

On en déduit une décision (on dit aussi un acte) a par une règle de décision. $\delta \in \Delta$ qui pour e fixée est une application de $\mathfrak{X}(e)$ dans A : $a = \delta(x, e)$.

Parallèlement, on obtient une information sur θ donnée par la loi a posteriori $P''_{\theta/x,e}$.

Les conséquences de ce processus se traduisent par le coût :

$$W(e, x, a, \theta) = W(e, x, \delta(x, e), \theta)$$

qui comme θ dont il dépend n'est connu que par sa loi de probabilité.

Les problèmes qui se posent sont :

Déterminer l'expérience à faire : e^*
Déterminer la règle de décision δ^* à adopter.

----

(1) $E_{\theta\, x/e}$ signifie que l'espérance mathématique est à prendre pour la loi $P_{\theta,\, x/e}$

La méthode de Bayes consiste à choisir e^* et δ^* de façon à ce que l'espérance du coût pour la loi $P_{\theta, x/e}$ soit minimale (il revient au même de dire que l'espérance de l'utilité est maximale).

On doit donc avoir :

$$E_{\theta, x/e}\,[W(e^*, X, \delta^*(X, e)\theta] = \underset{\substack{\delta \in \Delta \\ e \in E}}{\text{Min}}\; E_{\theta, x/e}\,[W(e, X, \delta(X, e), \theta)]^{(1)} \quad (2)$$

$E_{\theta, x/e}$ signifie que l'espérance est à prendre pour la loi $P_{\theta, x/e}$.

<u>Remarque</u> -

L'équation (2) exige que la fonction $W[e, X, \delta(X, e)\ \theta]$ soit intégrable par rapport à la mesure $P_{\theta, x/e}$ pour tout $e \in E$, et tout $\delta \in \Delta$. L'espace Δ des décisions est constitué par les applications de $\mathfrak{X}$ dans A qui satisfont à cette condition.

La détermination de e^* et δ^* n'est pas simultanée et le processus de détermination de e^* et δ^* est inverse du processus expérimental en ce sens qu'on détermine d'abord la règle de décision optimale δ^*, si elle existe, puis ensuite l'expérimentation optimale e^*.

Utilisant le fait que $D_{\theta, x/e} = D_{x/e} \cdot D''_{\theta/x, e}$, on montre en effet (cf. annexe) que :

$$\underset{\substack{\delta \in \Delta \\ e \in E}}{\text{Min}}\; E_{\theta, x/e}\,[W(e, X, \delta(X, e), \theta] = \underset{e \in E}{\text{Min}}\; E_{x/e}\,[\underset{a \in A}{\text{Min}}\; E''_{\theta/x, e}[W(e, x, a, \theta)] \quad (3)$$

sous réserve que $E_{x/e} \left\{ \underset{a \in A}{\text{Min}}\; E''_{\theta/x, e}\,[W(e, x, a, \theta)] \right\}$ existe, $\forall\ e \in E$.

La notation E'' rappelle que la loi considérée pour θ est la loi a posteriori $P''_{\theta/x, e}$, et dans le deuxième membre de (3) figure a et non plus δ, parce que, quand x et e sont fixées, l'ensemble des valeurs $\delta(x, e)$, $\delta \in \Delta$ est l'ensemble A des $a \in A$.

Les deux étapes de la détermination de e^* et δ^* sont :

1/ <u>Choix de la règle de décision δ^*</u>. D'après le 2ème membre de (3) : $a^* = \delta^*(x, e)$ doit minimiser l'expérance du coût une fois l'expérience réalisée.

$\delta^*(x, e)$ sera donc construite par points, c'est-à-dire en déterminant ses valeurs pour tous les éléments (x, e) de $\mathfrak{X}$ de façon à satisfaire à :

$$E''_{\theta/x, e}\,[W[e, x, \delta^*(x, e), \theta] = \underset{a \in A}{\text{Min}}\; E''_{\theta/x, e}\,[W(e, x, a, \theta)] \quad (4)$$

On montre (cf. annexe) que δ^* existe et est bien une règle de Bayes.

----

(1) On rappelle qu'une fonction réelle admet un minimum quand elle atteint effectivement sa borne inférieure.

2/ Choix de l'expérience e^* - Ayant déterminé δ^*, on choisit l'expérience e^*, si elle existe, qui minimise l'espérance du coût par rapport à l'ensemble des résultats possibles pour cette expérience, pour la loi $P_{x/e}$. e^* est telle que l'on ait :

$$E_{x/e^*} \left\{ E''_{\theta/x,e^*} [W(e,X,\delta^*(X,e^*),\theta)] \right\} \tag{5}$$

$$= \operatorname*{Min}_{e \in E} E_{x/e} \left\{ E''_{\theta/x,e} [W(e,X,\delta^*(X,e),\theta] \right\}$$

Remarque -

Dans le cas d'une fonction de coût additive, le coût de l'expérimentation $C(e,x)$ n'intervient pas dans le choix de la règle de décision δ^* mais seulement dans celui de l'expérience e^* [cf (4)].

En conclusion, quand on a défini les espaces de base et fait choix d'une fonction de coût et d'une distribution a priori pour θ, le choix de l'expérience et la règle de décision sont complètement définis sous réserve qu'on sache procéder aux calculs nécessaires. On a donc intérêt à choisir une distribution a priori pour θ et une fonction de coût qui se prêtent bien aux calculs tout en étant une image suffisamment fidèle de ce que l'on sait.

Exemple. - Détermination de δ^* et n^* dans l'exemple 1, quand on prend pour loi a priori de p, la loi β du 1er type de densité

$$D'(p) = Cte\ p^{\tilde{k}-1} (1-p)^{\tilde{n}-\tilde{k}-1}$$

et une fonction de coût additive dont la composante dépendant de δ et θ est

$W_1(a_1, p) = 10\ p\ N$ avec a_1 : accepter le lot

$W_1(a_2, p) = N$ " a_2 : rebuter le lot

N est l'effectif du lot.

Détermination de δ^*

On voit sur l'équation (4) qu'il faut déterminer la densité a posteriori de p, sachant qu'on a obtenu pour $X = (X_1, X_2, \ldots, X_n)$ la réalisation $x = (x_1\ x_2, \ldots, x_n)$.

On a indiqué en I.1 que les X_i pouvaient être considérées comme des variables de Bernoulli indépendantes en sorte que

$$\begin{cases} P[X_i = 1] = p \\ P[X_i = 0] = 1 - p \end{cases}$$

ce que l'on peut résumer par :

$$P[X_i = x_i] = p^{x_i} (1 - p)^{1-x_i}\ ,\ x_i = 0,\ 1.$$

On en déduit :

$$D(x_1,\ x_2,\ \ldots\ldots\ x_n/p,\ n) = p^{\sum_{i=1}^{n} x_i} \quad (1-p)^{n-\sum_{i=1}^{n} x_i}$$

soit

$$D(x_1,\ x_2,\ \ldots\ldots\ x_n/p,\ n) = p^k\ (1-p)^{n-k} \text{ avec } k = \sum_{i=1}^{n} x_i,$$

La densité de X ne dépendant que de la variable $K = \sum_{i=1}^{n} X_i$, il en est de même de la densité a posteriori de p, et on a

$D''(p/x,n) = D''(p/k,n)$ où $D''(p/k,n)$ a la forme indiquée en II 1.2 (cf. équation 1)(1).

Puisque la loi a posteriori de p ne dépend que de la variable K, il en sera de même de la règle δ^* (cf. équation 4).

Dans la mise en oeuvre de l'équation (4) on peut alors remplacer x par k.

$$E''_{p/k,n}\ [W_1(a_1,\ p)] = 10\,N\,E''_{p/k,n}\ (p)$$

$$E''_{p/k,n}\ [W_1(a_2,\ p)] = N$$

On a vu en III 1.2 que $E''_{p/k,n}\ (p) = \dfrac{k+\tilde{k}}{n+\tilde{n}}$

en sorte que $E''_{p/k,n}\ [W_1(a_1,\ p)] = 10\,N\dfrac{k+\tilde{k}}{n+\tilde{n}}$

$\delta^*(k,n)$ doit être telle que $E''_{p/k,n}\ \{W_1[\delta^*(k,n),\ p]\}$ soit minimum.

On prendra donc $\delta^*(k,n) = a_1$ si $E''_{p/k,n}[W_1(a_1,p)] \leqslant E''_{p/k,n}[W_1(a_2,p)]$ et $\delta^*(k,n) = a_2$ dans le cas contraire.

Ceci conduit à $\delta^* \Rightarrow a_1$ si $\dfrac{k+\tilde{k}}{n+\tilde{n}} \leqslant \dfrac{1}{10}$ soit si $k \leqslant \dfrac{n+\tilde{n}}{10} - \tilde{k}$.

Pour fixer les idées on peut prendre $\tilde{k} = 3$, $\tilde{n} = 60$ ce qui donne pour valeur critique de k : $k_0(n) = \dfrac{n}{10} + 3$.

Remarques -

1/ Si $k = \dfrac{n+\tilde{n}}{10} - \tilde{k}$ on peut en fait choisir indifféremment a_1 ou a_2 ;

2/ $\dfrac{k+\tilde{k}}{n+\tilde{n}}$ est, on l'a vu, l'espérance à posteriori de p et la valeur $p = \dfrac{1}{10}$ est celle au-dessus de laquelle il est plus avantageux de rebuter le lot

----

(1) L'égalité ci-dessus est une propriété des statistiques exhaustives (cf. II, 3.2).

3/ $K \leqslant n$; par suite si $K_0(n) \geqslant n$ soit $n \leqslant \frac{n - 10\tilde{K}}{9}$, on aura toujours $\delta^* \Rightarrow a_1$. Avec $n = 60$, $\tilde{K} = 3$, cela se produira pour $n \leqslant 3$.

Détermination de n^* (1)

D'après (5), n^* doit minimiser

$$C(n) = E_{X/p} \left\{ E''_{p/x,n} [W(n, X, \delta^*(K,n), p] \right\}$$

$$= a\,\delta(n) + b\,n + E_{k/n} \left\{ E''_{p/k,n} \; W_1\,[\delta^*(K,n), p] \right\}$$

puisque W ne dépend de X que par l'intermédiaire de δ^* donc de K et $D''_{p/x,n} = D''_{p/k,n}$. (2)

On a $E''_{p/k,n}\,[W_1(\delta^*\; p)] = \frac{k + \tilde{k}}{n + \tilde{n}} 10N$ si $\delta(k,n) = a_1$ c'est-à-dire si $k \leqslant \inf \left[n, \frac{n + \tilde{n}}{10} - \tilde{k} \right]$.

et $E''_{p/k,n}\,[W_1(\delta^*, p] = N$ si $\delta^*(k,n) = a_2$, c'est-à-dire si $n \geqslant k > \frac{n + \tilde{n}}{10} - \tilde{k}$.

On en déduit :

$$E_K \; \{E''_p\,[W_1(\delta^*, p)]\} = N.\; \frac{10}{n + \tilde{n}} \sum_{k=0}^{[\frac{n+\tilde{n}}{10} - \tilde{k}]} (k + \tilde{k}) \; P\left[\frac{k}{n}\right] + N \sum_{k=[\frac{n+\tilde{n}}{10} - \tilde{k} + 1]}^{n} P[k/n]$$

ou $\left[\frac{n + \tilde{n}}{10} - \tilde{k}\right]$ désigne la partie entière de $\frac{n + \tilde{n}}{10} - \tilde{k}$ et P(k/n) est la probabilité marginale de la valeur $K = k$, dans un échantillon de taille n.

Par suite

$$P(k/n) = \int_{p=0}^{1} P(k/n, p)\, D'(p)$$

soit :

$$P(k/n) = \frac{\binom{n}{K}}{B(\tilde{k}, \tilde{n}-\tilde{k})} \int_0^1 p^{k+\tilde{k}-1} (1-p)^{n+\tilde{n}-(k+\tilde{k})-1} dp$$

$$= \frac{\binom{n}{K}}{B(\tilde{k}, \tilde{n}-\tilde{k})} B[k+\tilde{k},\; n+\tilde{n}-(k+\tilde{k})], \quad k = 0, 1 \ldots n$$

----

(1) On suppose que le coût d'échantillonnage est suffisamment élevé pour que n soit petit devant N et qu'on puisse considérer que les prélèvements de pièces sont non exhaustifs.

(2) On peut aussi remarquer que $E_{x/n}\,[E''_{p/x,n}(F)] = E_{x,p/n}(F)$ pour toute statistique F. Puisque $W_1\,[n, \delta^*(K,n), p]$ ne dépend de X que par l'intermédiaire de K, on a : $E_{x,p/n}\,[W_1(n, \delta^*(k,n), p] = E_{k,p/n}\,[W_1(n, \delta^*(k,n), p]$, l'espérance $E_{k,p/n}$ étant relative à la loi $P_{k,\theta/n}$ induite de $P_{x,\theta/n}$ par la statistique K.

Soit :

$$P(k/n) = \frac{(\tilde{n}-1)\,!}{(\tilde{k}-1)\,!\ (\tilde{n}-\tilde{k}-1)\,!} \times \frac{n\,!}{k\,!\ (n-k)\,!} \times \frac{(k+\tilde{k}-1)\,!\ [n+\tilde{n}(k-\tilde{k}-1)]\,!}{(n+\tilde{n}-1)\,!}$$

$$= f_{b\beta}\,(k/\tilde{k},\ \tilde{n},\ n) \quad 0 \leqslant k \leqslant n.$$

La distribution de K est dite "Béta-binomiale" de paramètres $\tilde{k}$, $\tilde{n}$, n.

On a donc

$$E_K \left\{ E''_p\,[W_1(\delta^*, p]\right\} = N.A(n,\tilde{n},\tilde{k})$$

$$= N.\ \frac{(\tilde{n}-1)\,!}{(\tilde{k}-1)\,!\ (\tilde{n}-\tilde{k}-1)\,!} \sum_{k=o}^{[\frac{n+\tilde{n}}{10}-\tilde{k}]} 10\binom{n}{K}\frac{(k+\tilde{k})\,!\ [n+\tilde{n}-(k+\tilde{k})-1]\,!}{(n+\tilde{n})\,!}$$

$$+ \sum_{k=[\frac{n+\tilde{n}}{10}-\tilde{k}+1]}^{n} \binom{n}{K}\frac{(k+\tilde{k}-1)\,!\ (n+\tilde{n}-k-\tilde{k}-1)\,!}{(n+\tilde{n}-1)\,!}$$

soit en posant $G_{b\beta}\,(a/\tilde{k},\tilde{n},n) = P\,[K \geqslant a]$ pour la loi béta binomiale de paramètres $\tilde{k},\tilde{n},n$:

$$A(n,\tilde{n},\tilde{k}) = 10\,\frac{\tilde{k}}{\tilde{n}} + G_{b\beta}\left\{\left[\frac{n+\tilde{n}}{10}-\tilde{k}+1\right]/\tilde{k},\tilde{n},n\right\}$$

$$- 10\,\frac{\tilde{k}}{\tilde{n}}\,G_{b\beta}\left\{\left[\frac{n+\tilde{n}}{10}-\tilde{k}+1\right]/\tilde{k}+1,\ \tilde{n}+1,n\right\}$$

La fonction $G_{b\beta}$ n'est pas tabulée et on ne peut déterminer n^* que par un calcul numérique qui nécessite la fixation de $\tilde{k},\tilde{n},b,N$ et de l'ordre de grandeur de a pour voir si on peut avoir $n^* = 0$.

Si n = 0, c'est-à-dire si on ne prélève pas l'échantillon, on doit remplacer la loi a posteriori de p, par sa loi a priori.

On a $E'(p) = \frac{\tilde{k}}{\tilde{n}}$ (cf. II. 1.2) et par suite $E'_p\ [W_1(a_1,p)] = 10N\,\frac{\tilde{k}}{\tilde{n}}$

$$E'_p\ [W_1(a_2,p)] = N.$$

On prendra donc toujours la décision a_1 (accepter le lot) si $\frac{\tilde{k}}{\tilde{n}} \leqslant \frac{1}{10}$, et la décision a_2 si $\frac{\tilde{k}}{\tilde{n}} > \frac{1}{10}$ ($\frac{\tilde{k}}{\tilde{n}}$ est la moyenne de la loi a priori de p).

Avec $\tilde{n} = 60$, $\tilde{k} = 3$, $\frac{\tilde{k}}{\tilde{n}} = 0{,}05$. On est donc amené à accepter le lot. L'espérance du coût est $10N\frac{\tilde{k}}{\tilde{n}}$, soit $\frac{N}{2}$ avec $\tilde{n} = 60$, $\tilde{k} = 3$.

Détermination de n^* quand $\tilde{n} = 60$, $\tilde{k} = 3$

$$A(n,60,3) = 0,5 + G_{b\beta}\left\{\left[\frac{n}{10}+4\right]/3,60,n\right\} - 0,5\, G_{b\beta}\left\{\left[\frac{n}{10}+4\right]/4,61,n\right\}$$

Comme $G_{b\beta}\{x/\tilde{k},\tilde{n},n\} = 0$ si $x > n$, on voit que pour $\frac{n}{10} + 4 \geqslant n + 1$ soit $n \leqslant 3$, $A(n,60,3) = 0,5$, de sorte que pour $n \leqslant 3$ le coût moyen $C(n) = a + bn + \frac{N}{2}$ est supérieur à $C(0) = \frac{N}{2}$.

On a remarqué que dans ce cas $k_0(n) > n$, de sorte qu'on est conduit à toujours accepter le lot. On a donc toujours avantage à accepter le lot sans échantillonnage si la taille de l'échantillon ne dépasse pas 3.

Le calcul de la fonction $A(n,60,3)$ pour les valeurs de n de 1 à 37(1), montre qu'elle décroît extrêmement lentement. Nous avons trouvé

$$A(37,60,3) = 0,4984052 = 0,5 - 0,0015948$$

L'utilisation de l'approximation de $G_{b\beta}(x/\tilde{k},\tilde{n},n)$ par $F_\beta\left(\frac{n}{n+\tilde{n}-1}/x,\tilde{k}\right)$ où F_β est la fonction de répartition de la distribution béta du 1er type de paramètres x et $\tilde{k}$ qui est, elle, tabulée, a montré que pour $n = 60$ et 100, A était encore très voisin de 0,5.

Cette approximation est très bonne si on a

$$x << n,\ \tilde{k} << \tilde{n},\ \text{et } x + \tilde{k} \ll \text{Max}\,(n, \tilde{n}),$$

ce qui est bien le cas ici où $x = \left[\frac{N}{10} + 4\right]$.

Il s'ensuit que pour les valeurs $b = 0,2$ et $0,05$ et $N = 100$ et $1\,000$ qu'on a essayées, (C(n) est minimum pour $n = 1$.

On est donc amené à prendre $n = 0$ puisque $C(0) \leqslant C(1)$, c'est-à-dire à accepter le lot sans aucun prélèvement, si le coût de l'échantillonnage n'est pas nul ou très faible. // Pour avoir $C(37) < 0,5$ avec $a = 0$, il faut que l'on ait $37\frac{b}{N} \leqslant 0,000159$, soit $b \leqslant 4N.10^{-5}$ c'est-à-dire $b \leqslant 0,04$ si $N = 1\,000$; par suite de la variation très lente de $A(n,60,3)$ avec n, le gain moyen par rapport à l'acceptation sans contrôle est pratiquement négaligeable ; avec $n = 37$, $b = 0$, $N = 1\,000$, il serait de 1,6 alors que le coût moyen $C(0) = N/2 = 500$ est la moitié du coût de fabrication du lot.

Ces résultats, curieux au premier abord, s'expliquent par le fait que la ditribution a priori adoptée pour p est très concentrée autour de sa moyenne 0,05. Par suite, la probabilité a priori d'avoir p inférieur à 1/10, c'est-à-dire de donner un lot acceptable, est de 0,93. Il faut alors un échantillon de très grande taille tiré d'une population où p serait supérieur à 1/10 pour que la probabilité a posteriori d'avoir $p < 1/10$ soit très sensiblement inférieure à 0,9. La valeur élevée trouvée pour le coût moyen est due au prix

----

(1) On s'est arrêté à $n = 37$, car pour $n > 37$ il y avait dépassement de la capacité de calcul en double précision avec le programme adopté.

très élevé d'une pièce défectueuse dans la livraison et à ce que la probabilité a priori d'avoir p petit, par exemple, $p < 0,02$, est faible. (elle vaut 0,10). Il en est donc de même de la probabilité a posteriori.

II. 3 - Les statistiques exhaustives dans la décision bayesienne

Dans l'exemple 1 on a établi l'égalité des densités à posteriori de p 0 relatives à (x,n) et à (k,n), ce qui a permis de conclure que puisque la fonction de coût dépendait de e, a, θ et non de x :

1/ Pour e fixé la règle de Bayes δ^* pouvait être cherchée comme règle de Bayes relative à (k, n),

2/ L'expérimentation optimale e^* pouvait également être cherchée comme expérimentation optimale relative à (k,n).

On remarquera que ces conclusions subsistent avec une fonction de coût additive, telle que le coût de l'expérience c(e,x) ne soit fonction de x que par l'intermédiaire de k soit $c(e,x) = C'(e,k)$.

Ces résultats sont dûs à ce que $K(\underline{x}, n) = \sum_{i=1}^{n} x_i$ est une statistique exhaustive de $\underline{x}(n) = (x_1, x_2, \ldots, x_n)$ pour p.

II. 3.1 - Définition classique d'une statistique exhaustive

On rappelle que, si T(x) est une statistique de X, c'est-à-dire une application mesurable définie sur $(\mathcal{X}, \mathcal{B})$ et à valeurs dans un espace mesurable $(\mathcal{T}, \mathcal{C})$, T(x) est exhaustive de X pour θ, si la probabilité de X conditionnée par T ne dépend pas de θ.

On traduit ainsi le fait que la statistique T(x) résume toute l'information sur θ contenue dans X, puisque quand on connaît la valeur t prise par T, la connaissance de la valeur x prise par X n'apporte aucune information supplémentaire sur θ.

C'est ainsi que dans l'exemple 1, quand on connaît le nombre k des pièces défectueuses de l'échantillon, la connaissance des numéros i = 1, 2, .., n de celles qui sont défectueuses n'apport aucune information sur p.

On vérifie aisément que tous les échantillons comportant k pièces défectueuses ont même probabilité quand k est fixé soit $1/\binom{n}{k}$.

II. 3.2 - Définition bayesienne d'une statistique exhaustive [6]

L'intervention explicite de l'expérimentation e, conduit à noter T(X,e) une statistique de (X,e), T étant pour tout $e \in E$ une application mesurable de $\mathcal{X}(e)$ dans $\mathcal{T}(e)$.

La définition intuitive d'une statistique exhaustive T de (X,e) pour θ, selon laquelle l'information sur θ fournie par (x,e) ou par T(x,e) est la même, conduit, puisque l'information sur θ est représentée par la loi a posteriori de θ relative à (x,e), à la :

Définition 2

T(x,e) est exhaustive de (x,e) pour θ si

(6) $$D''(\theta /t, e) = D''(\theta /x,e) \text{ avec } (t, e) = T(x, e)$$

pour tout (x, e) $\in \mathfrak{X}$ et toute densité a priori $D'(\theta)$ telle que

$$\int_{\Theta} D(x/\theta , e)\ D'(\theta)\ d\mu (\theta) > 0, \quad \forall (x, e) \in \mathfrak{X}$$

On a alors le Théorème 1 :

Les définitions bayesienne et classique d'une statistique exhaustive sont équivalentes.

Pour démontrer le théorème 1 et pour reconnaître qu'une statistique est exhaustive on utilise le résultat suivant.

II. 3.3. - Caractérisation des statistiques exhaustives au sens classique. Théorème de factorisation :

On démontre que si la famille P_{θ} des lois de probabilité de X admet pour tout θ une densité $l(x/\theta)$ par rapport à une mesure μ, une condition nécessaire et suffisante pour que T(x) soit une statistique exhaustive de x pour θ est qu'elle satisfasse au "théorème de factorisation" c'est-à-dire qu'il existe deux fonctions réelles $h(t/\theta)$ et $g(x)$, la deuxième étant positive ou nulle pour tout $x \in \mathfrak{X}$, telles que :

$$l(x/\theta) = h(t/\theta) \cdot g(x) \quad \begin{cases} \forall\ x \in \mathfrak{X} \\ \forall \theta \in \Theta \end{cases}$$

On peut alors montrer que $h(T/\theta)$ est la densité de T par rapport à une certaine mesure λ sur $\mathcal{T}$.

Théorème 1 - Les définitions bayesienne et classique d'une statistique exhaustive sont équivalentes

Démonstration :

1/ La définition 2/ entraine le théorème de factorisation.

En effet, si $h(t/\theta, e)$ est la densité de (t,e) par rapport à une mesure ν sur $\mathcal{T}$ et $l(x/\theta, e)$ est la densité de (x, e) par rapport à une mesure μ sur $\mathfrak{X}$ on a :

$$D''(\theta/t, e) = \frac{h(t/\theta,e)\ D'(\theta)}{\int_{\Theta} h(t/\theta, e)\ D'(\theta)\ d\nu(\theta)} = \frac{h(t/\theta, e)\ D'(\theta)}{h(t, e)}$$

et

$$D''(\theta/x, e) = \frac{l(x/\theta, e)\ D'(\theta)}{\int_{\Theta} l(x/\theta, e)\ D'(\theta)\ d\nu(\theta)} = \frac{l(x/\theta, e)\ D'(\theta)}{l(x, e)}$$

en désignant par $h(t,e)$ et $l(x,e)$ respectivement les densités marginales de (t,e) et (x,e) par rapport aux mesures ν et θ qui ne dépendent pas de θ.

De $D''(\theta/t,e) = D''(\theta/x,e)$ on déduit $l(x/\theta,e) = h(t/\theta,e) \times \frac{h(t,e)}{l(x,e)}$ ce qui n'est autre que le théorème de factorisation puisque $\frac{h(t,e)}{l(x,e)}$ n'est fonction que de (x,e).

En effet, quand on introduit explicitement l'expérimentation e, le théorème de factorisation prend la forme :

$$l(x/\theta, e) = h\,[T(x,e)/\theta, e]\ .\ g(x,e) \tag{7}$$

où $g(x,e) \geqslant 0$ est indépendante de θ

et h est pour θ et e fixés une fonction réelle définie sur $T[\mathcal{X}(e)]$.

2/ Réciproquement le théorème de factorisation (7) entraine l'équation (6) de la définition 2

La démonstration repose sur le fait que la fonction $h[T/\theta,e]$ de l'équation (7) est la densité de probabilité de (T,e) par rapport à une certaine mesure λ_e sur $\mathcal{T}_e$.

Comme la densité à posteriori ne dépend pas du choix de la mesure de référence sur T (cf. remarque 2 de II. 1.2) on a :

$$D''(\theta/t,e) = \frac{h[t/\theta,e]}{\int_\Theta h[t/\theta e]\, D'(\theta)\, d\mu(\theta)} = \frac{h[t/\theta,e]\, g(x,e)}{\int_\Theta h(t/\theta,e)\, g(x,e)\, D'(\theta)\, d\mu(\theta)}$$

$$= \frac{l(x/\theta,e)}{\int_\Theta l(x/\theta,e)\, D'(\theta)\, d\mu(\theta)} = D''(\theta/x,e)$$

C.Q.F.D.

On peut alors énoncer le

<u>Théorème 2</u> - Si pour tout $e \in E$, il existe une statistique exhaustive T(x,e) de (x,e) pour θ, et si la fonction de coût ne dépend pas de x, ou est additive avec un coût d'expérimentation qui ne dépend de x que par l'intermédiaire de T, la recherche d'une règle de Bayes δ^* (définie sur $\mathcal{X}$) et d'une expérimentation optimale relatives à la variable (X,e) se ramène à la recherche d'une règle de Bayes δ^{**} (définie sur $T(\mathcal{X})$) et d'une expérimentation optimale relatives à la statistique exhaustive T(x,e).

On a $\delta^*(X,e) = \delta^{**}[T(x,e),e]$ et le coût moyen associé aux deux règles est le même pour tout $e \in E$.

III - CHOIX D'UNE LOI A PRIORI POUR θ. LES "CONJUGUEES NATURELLES" D'UNE FAMILLE DE VRAISEMBLANCES [(3), (6)].

III. 1 - <u>Définition et propriété fondamentale des conjuguées naturelles.</u>

Comme on vient de le voir dans l'exemple précédent, la mise en oeuvre pratique de la méthode bayesienne se heurte à des difficultés de calcul.

Pour qu'au moins la première étape de cette mise en oeuvre, c'est-à-dire le choix de δ^*, puisse être réalisée sans trop de peine, il importe que la densité a posteriori $D''(\theta/x,e)$ puisse être calculée aisément par la formule de Bayes(1) à partir de la densité a priori $D'(\theta)$ et de la fonction de vraisemblance de θ au point x(e), soit $l(x/\theta,e) = D(x/\theta,e)$.

On a donc recherché s'il était possible de trouver des densités a priori $D'(\theta)$ telles que la densité a posteriori $D''(\theta/x,e)$ se calcule facilement.

Définition 3 - Conjuguées naturelles d'une famille de fonctionns de vraisemblance $\theta \Longrightarrow l(x/\theta,e)$, $e \in E$, $x \in \mathcal{X}(e)$.

La famille des conjuguées naturelles est la famille des fonctions.

$$\theta \Longrightarrow K(\theta/x',e') = \frac{l(x'/\theta,e')}{\int_{\Theta} l(x'/\theta,e')\,d\mu(\theta)} \tag{8}$$

ou e' est un élément quelconque de E et x' un élément quelconque de $\mathcal{X}(e')$ tel que $0 < \int_{\Theta} l(x'/\theta,e')\,d\mu(\theta) < \infty$.

On définit ainsi une famille à deux groupes de paramètres $(x',e') \in \mathcal{X}$, $\mathcal{X} = \{\mathcal{X}(e)\}_{e\in E}$, de fonctions définies sur Θ et à valeurs dans $\mathbf{R}^+$. En fait la restriction $(x',e') \in \mathcal{X}$ est inutile et on est souvent amené à étendre le domaine des valeurs possibles pour x' et e' à des espaces E' et $\mathcal{X}'(e)$ plus grands que E et $\mathcal{X}(e)$. On a ainsi davantage de choix. C'est ainsi que dans un problème d'échantillonnage ou $E = N$, on peut prendre $E' = R^+$ ou même $E' = R$, c'est-à-dire admettre pour n' des valeurs non entières et même négatives, et si, par exemple x est entier, admettre pour x' des valeurs réelles quelconques. On en verra des exemples plus loin.

La fonction $\theta \Longrightarrow K(\theta/x',e')$ est la conjuguée naturelle de la fonction de vraisemblance $l(x'/\theta,e)$ de θ pour (x',e').

Remarque - Il résulte de la remarque 2 de II.1.2, que la conjuguée naturelle ne dépend pas du choix de la mesure sur $\mathcal{X}$ utilisée pour exprimer les densités $l(x/\theta,e)$ des lois $Px/\theta,e$.

Théorème 3 - Propriété fondamentale des conjuguées naturelles

Si la densité a priori de θ est une conjuguée naturelle de la vraisemblance de θ au point (x',e'), la densité a posteriori de θ après observation de (x,e) est conjuguée naturelle de la vraisemblance de θ au point $[(x,x')\ (e,e')]$.

(ee') est l'expérience composée des deux expériences indépendantes e et e' (cf. II.1.1) et (x,x') représente le résultat de l'expérience ee' constitué par l'ensemble des résultats x pour e et x' pour e'.

Démonstration - Par hypothèse

$$D'(\theta) = K(\theta/x',e') = \frac{l(x'/\theta,e')}{\int_{\Theta} l(x'/\theta,e')\,d\mu\,\theta}$$

La formule de Bayes (1) donne :

$$D''(\theta/x,e) = \frac{l(x/\theta,e)\,\dfrac{l(x'/\theta,e')}{\int_{\Theta} l(x'/\theta,e')\,d\mu(\theta)}}{\int_{\Theta} l(x/\theta,e).\,\dfrac{l(x'\ \theta,e')}{\int_{\Theta} l(x'\ \theta,e')\,d\mu(\theta)}\,d\mu(\theta)}$$

soit en simplifiant par $\int_{\theta} l(x'/\theta, e')\, d\mu(\theta)$:

$$D''(\theta/x,e) = \frac{l(x/\theta,e)\; l(x'/\theta,e')}{\int_{\theta} l(x/\theta,e)\; l(x'/\theta,e')\; d\mu(\theta)} \tag{9}$$

$$= \frac{l(x,x'/\theta,ee')}{\int_{\theta} l(x,x'/\theta,ee')\, d\mu(\theta)}$$

soit $\boxed{D''(\theta/x,e) = K[\theta/(xx'), (ee')]}$ C.Q.F.D.

Pratiquement, on choisira (x',e') de façon à ce que $K(\theta/x',e')$ représente au mieux la connaissance a priori que l'on a sur θ.

Exemple - Dans l'exemple 1, on a été amené à considérer la variable

$$(k, n) = \sum_{i+1}^{n} x_i$$

qui suit une loi binomiale de vraisemblance $l(k/p,n) = \binom{n}{p} p^k (1-p)^{n-k}$.

La conjuguée naturelle de la fonction $l(K/p,n)$ au point (k',n') est

$$K(p/k',n') = \frac{\binom{n'}{p} p^{k'} (1-p)^{n'-k'}}{\int_o^1 \binom{n'}{p} p^{k'} (1-p)^{n'-k'}\, dp} = \frac{p^{k'} (1-p)^{n'-k'}}{\int_o^1 p^{k'} (1-p)^{n'-k'}}$$

$K(p/k',n')$ n'est autre que la distribution béta du 1er type de paramètres $k'+1$, $n'-k'+1$. Pour $k' \in N$, $n' \in N^+$, $k' \leqslant n'$, elle est toujours définie, mais elle est également définie pour $-1 < k' \leqslant n'$, k' et n' n'étant plus forcément entiers, ce qui donne une plus grande souplesse pour l'ajustement d'une loi Béta aux connaissances a priori sur p.

C'est pourquoi dans l'exemple traité en II.1.2. on a choisi pour distribution a priori de p une distribution béta de 1er type de paramètres $\tilde{k}$ et $\tilde{n} - \tilde{k}$, c'est-à-dire conjuguée naturelle de la fonction de vraisemblance de p au point $(k',n') = (\tilde{k}-1, \tilde{n}-2)$, ce qui justifie les conditions $\tilde{n} > \tilde{k} > 0$ posées dans cet exemple. On a choisi $\tilde{k} = 3$, $\tilde{n} = 60$ pour avoir $E'(p) = 5\,\%$ (On a vue que $E'(p) = \tilde{k}/\tilde{n}$) et $P'(p > 15\,\%)$ négligeable.

On a obtenu, pour $D''(p/k,n)$ la même distribution β du 1er type de paramètres $k+\tilde{k}$ et $n+\tilde{n}$, c'est-à-dire la conjuguée naturelle de la fonction de vraisemblance de p au point $(k+k', n+n')$. La composition des expériences $e = n$, $e' = n' = \tilde{n} - 2$ n'est autre, ici, qu'une addition ; $ee' = e + e'$, de même que $xx' = x' + x' = k + k'$, puisque $\left\{\begin{array}{l} x = k \\ x' = k' \end{array}\right.$, comme il fallait s'y attendre

4

étant donné la signification concrète du problème (k est le nombre de pièces défectueuses observées sur un échantillon de n pièces et $k' = \tilde{k}^2 - 1$ peut être considéré comme le nombre de pièces défectueuses observées sur un échantillon de $n' = \tilde{n} - 2$ pièces).

Application du théorème 3 au cas particulier où l'expérience est un échantillonnage.

$E = \{e_n,\ n \in N\}$ où e_n est l'expérience : prélever un échantillon de taille n dans une même population.

$\mathscr{X}(e_n)$ est l'ensemble des résultats observables sur tous les échantillons de taille n.

et $\mathscr{X} = \{\mathscr{X}(e_n),\ n \in N\}$ est l'ensemble de tous les résultats observables sur l'ensemble de tous les échantillons.

On a $e_n \ , e_{n'} \ = e_{n+n'}$

et si $\quad x^{(n)} = (x_1,\ x_2 \ldots x_n) \in \mathscr{X}\ (e_n)$

et $\quad x'^{(n')} = (x'_1 \ldots x'_{n'}) \in \mathscr{X}\ (e_{n'})$

$$x^{(n)}\ .\ x'^{(n')} = (x_1,\ x_2, \ldots x_n,\ x'_1, \ldots,\ x'_{n'})$$

Si on prend pour densité a priori $D'(\theta)$ la conjuguée naturelle

$$k(\theta/x'^{(n')},\ n') = K(\theta/x'_1 \ldots x'_{n'}) \text{ de } l(x'^{(n')}/\theta, n') = l(x'_1 \ldots x'_{n'}/\theta)$$

la densité a posteriori de θ est

$$\begin{aligned} D''(\theta/x^{(n)},\ n) = D''(\theta/x_1 \ldots x_n) &= K(\theta/x^{(n)},\ x'^{(n')},\ n+n') \\ &= K(\theta/x_1 \ldots x_n,\ x'_1 \ldots x'_{n'}) \end{aligned}$$

Remarque :

Essai d'interprétation concrète des conjuguées naturelles.

Le rapprochement de l'équation (8) définissant $K(\theta/x', e')$ avec la formule de Bayes

$$D''(\theta/x, e) = \frac{D'(\theta)\ l(x/\theta, e)}{\int_\Theta l(x/\theta, e)\ D'(\theta)\ d\mu\ (\theta)} \qquad (1)$$

conduit à considérer $K(\theta/x', e')$ comme la densité a posteriori de θ correspondant à une densité à priori $D'(\theta)$ constante par rapport à la mesure μ, et au résultat (x', e').

On voit sur la formule (9) que $D''(\theta)$ correspond à la même densité à priori mais au résultat (xx', ee'). On peut donc dire que la donnée de $K(\theta/x', e')$ équivaut à la donnée du résultat x' d'une expérience fictive e' indépendante de e et au fait qu'on est ignorant a priori sur θ, si tant est que l'on puisse traduire son ignorance par une densité uniforme par rapport à la mesure μ sur Θ.

Ainsi, dans l'exemple 1, la donnée de D'(p) peut être considérée comme équivalente au tirage d'un échantillon fictif de taille $n' = \tilde{n} - 2$ sur lequel on aurait observé $k' = \tilde{k} - 1$ pièces défectueuses.

Cette interprétation présente plusieurs difficultés :

a) Si $\mu(\Theta)$ n'est pas fini, une densité uniforme par rapport à μ égale à λ n'est pas une densité de probabilité puisqu'elle attribue la probabilité $\lambda\,\mu(\Theta) = \infty$ à Θ. On rencontre cette difficulté dans les cas fréquents où $\Theta = R^k$ et μ est la mesure de Lebesgue sur R^k.

On peut songer à limiter le support de la mesure uniforme à un ensemble fini de $\mathbf{R}^k$, par exemple à un intervalle fini de R soit Θ_0, mais il ne faut pas perdre de vue que tout ensemble de Θ qui a une probabilité à priori nulle, à une probabilité a posteriori également nulle et ceci quelles que soient les lois $P_{x/\theta,e}$. Autrement dit, ce qui est impossible à priori reste impossible à posteriori. Il convient donc de ne pas limiter à la légère le domaine a priori effectivement possible pour θ.

b) Au premier abord, il ne paraît pas très satisfaisant de traduire son ignorance par une densité a priori uniforme sur Θ. En effet, il semble que si on ignore tout de la valeur de θ, il en est de même pour $\frac{1}{\theta}$, θ^2 etc... si $\theta \in \mathbf{R}$, et plus généralement pour toute fonction de θ, alors qu'une densité constante et égale à λ pour θ par rapport à la mesure de Lebesgue de $\mathbf{R}$, implique par exemple pour $u = 1/\theta \in \mathbf{R}$ une densité par rapport à la même mesure égale à λ/u^2.

Le fait de pouvoir choisir la mesure par rapport à laquelle on exprime son ignorance, peut néanmoins justifier que l'ignorance sur θ ne soit pas équivalente à l'ignorance sur $1/\theta$. Dire qu'on est ignorant par rapport à une mesure μ, c'est dire qu'on attribue une même probabilité à des ensembles de même mesure pour μ.

III.2 - Recherche effective des conjuguées naturelles - Conjuguées naturelles des familles de vraisemblances admettant une statistique exhaustive.

Théorème 4 - Si T(x,e) est une statistique exhaustive, la conjuguée naturelle de la fonction de vraisemblance $l(x/\theta,e)$ de θ relative à (x,e) ne dépend que de $(t,e) = T(x,e)$ et c'est la conjuguée naturelle de la fonction de vraisemblance de θ relative à (t,e).

Démonstration - Le théorème de factorisation permet, en effet, d'écrire

$$K(\theta/x,e) = \frac{l(x/\theta,e)}{\int_\Theta l(x/\theta,e)\,d(\theta)} = \frac{h(t/\theta,e)}{\int_\Theta h(t/\theta,e)\,d\mu(\theta)}$$

où $h(t/\theta,e)$ est, on l'a remarqué en II.3.3, une fonction de vraisemblance de θ relative à (t,e).

On a donc $K(\theta/x,e) = K^*(\theta/t,e)$ où $K^*(\theta/t,e)$ est la conjuguée naturelle de la vraisemblance de θ relative à (t,e).

La densité a posteriori $D''(\theta/x,e) = D''(\theta/t,e)$ correspondant à une densité a priori conjuguée naturelle de la fonction de vraisemblance de θ relative à (x',e') ou à (t',e') avec t' = T(x',e') est donc $K^*(\theta/tt', ee')$. Elle ne dépend que de (t,e) et t',e') et non de (x,e) et (x',e').

Le théorème 4 joint au théorème 2, permet quand on adopte une densité à priori pour θ conjuguée naturelle de sa fonction de vraisemblance de remplacer la variate (X,e) par une statistique exhaustive, quand la fonction de coût ne dépend pas de X ou est additive avec un coût d'expérimentation ne dépendant de X que par l'intermédiaire de T.

Il justifie le choix de la loi a priori utilisée dans l'exemple 1, où K(p/k',n') est conjuguée naturelle de la vraisemblance de p relative à (x',n') = $(x'_1, x'_2, \ldots\ x'_n)$ aussi bien que de celle de p relative à (k', n') avec

$K' = \sum_{i=1}^{n} x'_i$.

La distribution Béta du 1er type de paramètres k' + 1, n' − k' + 1 est donc conjuguée naturelle de la fonction de vraisemblance de p relative à un n' échantillon d'une variable de Bernoulli, aussi bien que de la fonction de vraisemblance relative à une variable binomiale de taille n'.

Conjuguées naturelles des fonctions de vraisemblance de type exponentiel

Définition

Une famille exponentielle d'ordre s de lois de probabilité $P_{x/\theta,e}$ est caractérisée par une densité de la forme

$$l(x/\theta,e) = \Phi(x,e)\ e^{\sum_{j=1}^{s}\alpha_j(\theta)T_j(x,e)+\gamma(\theta,e)}$$

où les fonctions $\alpha_j(\theta)$ sont linéairement indépendantes.

Il résulte du théorème de factorisation (cf. II 3.3) que le vecteur T(x,e) de R^s dont les composantes sont $[T_1(x,e), T_2(x,e) \ldots T_s(x,e)]$ est une statistique exhaustive de (x,e) pour θ.

Remarque

Quand l'expérience e consiste dans le prélèvement d'un n échantillon de réalisations d'une variable aléatoire Y, on a e = n, et

$$\gamma(\theta,n) = n \cdot \gamma(\theta,1) = n\ \beta(\theta).$$

La conjuguée naturelle de $l(x/\theta,e)$ est :

$$K^*(\theta/t_1,\ t_2 \ldots t_s,\ e) = \frac{e^{\sum_{j=1}^{s}\alpha_j(\theta)t_j+\gamma(\theta,e)}}{\int_\Theta e^{\sum_{j=1}^{s}\alpha_j(\theta)t_j+\gamma(\theta,e)}\ d\mu(\theta)}$$

soit

$$K^*(\theta/t_1,\ t_2 \ldots t_s,\ e) = \varphi(t,e)\ e^{\sum_{j=1}^{s}\alpha_j(\theta)t_j+\gamma(\theta,e)}$$

avec $t_j = T_j(x,e)$

Elle n'existe que si $\frac{1}{\varphi(t,e)} = \int_\Theta e^{\sum_{j=1}^{s}\alpha_j(\theta)t_j+\gamma(\theta,e)}\ d\mu(\theta)$ est finie et non nulle.

Si

$$D'(\theta) = K(\theta/x', e') = K^*(\theta/t'_1, t'_2 .. t'_s, e')$$

on a vu que

$$D''(\theta/x, e) = D''(\theta/t, e) = K(\theta/xx', ee') = K^*(\theta/tt', ee')$$

X(e) et X(e') étant indépendantes en probabilité, la fonction de vraisemblance de θ relative au point (xx', ee') est :

$$l(x, x'/\theta, ee') = l(x/\theta, e)\ l(x'/\theta, e')$$

$$= \Phi(x,e).\ \Phi(x',e') \exp \left[\sum_{j=1}^{s} \alpha_j(\theta)\ [T_j(x,e) + T_j(x', e')] + \right.$$
$$\left. + \gamma(\theta, e) + \gamma(\theta, e') \right]$$

On a encore affaire à une vraisemblance de la même forme avec pour résumé exhaustif :

$$T(x, x', ee') = T(x, e) + T(x', e') \in \mathbf{R}^s,$$

(la somme dont il s'agit ici est une somme vectorielle), et pour paramètre caractéristique de l'expérience ee' :

$$\gamma(\theta, ee') = \gamma(\theta, e) + \gamma(\theta, e')$$

et

$$K(\theta/xx', ee') = K^*(\theta/tt', ee')$$

$$= \varphi(t+t', ee')\ e^{\sum_{j=1}^{s} \alpha_j(\theta)(t_j + t'_j) + \gamma(\theta, e) + \gamma(\theta, e')}$$

La composition de deux expériences indépendantes e et e' se traduit donc par l'addition vectorielle des résumés exhaustifs :

$$(tt', ee') \Longrightarrow T(xx', ee') = T(x, e) + T(x', e')$$

et par l'addition des paramètres caractéristiques de l'expérience

$$\gamma(\theta, ee') = \gamma(\theta, e) + \gamma(\theta, e')$$

Quand l'expérience est un échantillonnage, c'est-à-dire quand e = n, e' = n', $\gamma(\theta, ee') = (n + n')\ \beta(\theta)$, et par suite la composition de deux échantillonnages indépendants, se traduit par l'addition de leurs effectifs, ce qui est conforme à l'intuition (nn') $\Longrightarrow$ n + n'.

On a donc, dans le cas où

$$D'(\theta) = K(\theta/x', n') = K^*(\theta/t', n')$$

$$D''(\theta/\underline{x}, n) = D''(\theta/\underline{t}, n) = K^*(\theta/\underline{t+t'}, n+n')$$

C'est bien ce qu'on a obtenu dans l'exemple 1, car la famille des lois de Bernoulli de paramètre p est une famille exponentielle d'ordre un à laquelle est attachée le résumé exhaustif

$$T(X^{(n)}, n) = \sum_{i=1}^{n} X_i{}^{(1)}$$

III. 3 - Conjuguées naturelles des fonctions de vraisemblance usuelles correspondant à des expériences d'échantillonnage.

On montre que quand $X^{(1)} \in \mathbf{R}^{p}$ (2), $\Theta \subset \mathbf{R}^{k}$ et le domaine de variation effective de $X^{(1)}$ ne dépend pas de θ, une condition nécessaire et suffisante pour qu'il existe un résumé exhaustif d'ordre s de l'échantillon $X^{(n)}$, $n \geqslant s$, est que la fonction de vraisemblance de $X^{(1)}$ soit de type exponentiel d'ordre s.

a) Conjuguée naturelle de la vraisemblance associée à un échantillon d'une variable de Bernoulli de paramètre p

$$l(x/p, n) = p^{K}(1-p)^{n-K}, \quad x \in [\{0\}, \{1\}]^{n}$$

avec

$$K = \sum_{i=1}^{n} x_i \qquad O < p < 1$$

On a vu (cf. III.1, exemple 1) que la conjuguée naturelle de $l(x'/p, n')$ était

$$K^{*}(p/k', n') = \frac{p^{k'}(1-p)^{n'-k'}}{B(k'+1, n'-k'+1)}$$

$$k' \leqslant n'$$

c'est-à-dire une distribution β du 1er type, de paramètres $k'+1$, $n'-k'+1$, et que si $D'(p) = K^{*}(p/k', n')$, la densité à posteriori de p est

$$D''(p/k, n) = K^{*}(p/k'+k, n'+n) = K^{*}(p/k'', n'')$$

----

(1) Dans le cas général, le problème de la composition des statistiques exhaustives $T(x,e)$ et $T(x'e')$ correspondant à deux expériences e et e' n'est pas aussi simple.

Le théorème de factorisation montre que la vraisemblance $l(xx'/\theta, ee')$ admet également une statistique exhaustive $T(xx', ee')$, mais on peut se demander s'il est toujours possible de déterminer une fonction φ indépendante de θ telle que :

$$\varphi\,[T(x, e),\ T(x', e')] = T(xx', ee').$$

La réponse est affirmative quand la statistique $T(x,e)$ est minimale exhaustive $\forall\, e \in E$ puisque $T(xx', ee')$ qui est minimale exhaustive est alors fonction de toute autre statistique exhaustive de (xx', ee') et par suite de l'ensemble $T(x, e)$, $T(x', e')$ qui en est une.

(2) $x^{(1)}$ est la variable aléatoire correspondant à une expérience élémentaire, c'est-à-dire au prélèvement d'un échantillon de taille n = 1.

$$\text{avec} \begin{cases} k'' = k' + k \\ n'' = n' + n \end{cases}$$

La restriction k' et n' entiers et positifs est inutile.

Il suffit de prendre $-1 < k' \leqslant n'$ ce qui permet d'élargir la gamme des densités à priori possibles.

b) Conjuguée naturelle de la vraisemblance associée à un échantillon d'une variable binomiale de paramètre p inconnu, mais de taille N connue

$$l\,(x_1/p, 1) = \binom{N}{x_1} p^{x_1} (1-p)^{N-x_1}$$

et

$$l\,(x/p, n) = \binom{N}{x_1} \ldots \binom{N}{x_n} p^{\sum_{i=1}^{n} x_1} (1-p)^{nN-\sum_{i=1}^{n} x_i}$$

et on sait (cf. th. 4) que la conjuguée naturelle cherchée est celle de la fonction de vraisemblance de la statistique exhaustive $T(x,n) = \sum_{i=1}^{n} x_i$ qui est une variable binomiale de paramètre p, de taille n,N.

On a donc

$$K^*(p/t', n') = \frac{p^{t'} (1-p)^{n'N-t'}}{B(t'+1,\ n'N-t'+1)}$$

$$-1 < t' \leqslant n'N$$

c'est-à-dire encore une distribution β du 1er type mais de paramètres $t'+1$, $n'N - t'+1$.

Si $D'(p) = K^*(p/t', n')$, la densité à posteriori de p est

$$D''(p/t, n) = K^*(p/t+t',\ n+n') = \frac{p^{t+t'} (1-p)^{(n+n')N-(t+t')}}{B\,[t+t'+1,\ (n+n')\,N-(t+t')+1]}$$

avec

$$t = \sum_{i=1}^{n} x_i.$$

C'est une distribution β du 1er type de paramètres $t + t' + 1$,

$$(n + n')\ N - (t + t') + 1.$$

c) Conjuguée naturelle de la vraisemblance associée à un échantillon d'une variate de Poisson de paramètre m inconnu.

$$l\,(x/m, n) = e^{-mn}\, m^t \quad \text{avec} \quad t = \sum_{i=1}^{n} x_i$$

$$m > 0$$

La statistique exhaustive associée à la famille exponentielle est (T,n) avec $T = \sum_{i=1}^{n} x_i$. C'est une variable de Poisson de paramètre mn.

On a donc :

$$K^*(m/t',n') = \frac{e^{-mn'}(mn')^{t'}}{\Gamma(t'+1)}$$

$$\begin{cases} m > O \\ n' > O \\ t' > O \end{cases}$$

c'est-à-dire une distribution gamma d'ordre $t'+1$ et de paramètre $1/n'$. Les restrictions n' et t' entiers sont inutiles. Il suffit de prendre $n' > O$, $t' > -1$.

Si $D'(m) = K^*(m/t',n')$ alors $D''(m/t,n) = K^*(m, t+t', n+n')$

d) Cas d'un échantillon d'une variable gaussienne, d'écart type σ connu de moyenne μ inconnue.

$$l(x/\mu,n) \; \alpha^{(1)} \; e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2}$$

La statistique exhaustive intervenant dans la forme exponentielle est $(\sum_{i=1}^{n} X_i, n)$ et elle suit une loi de Gauss d'écart type $\sigma\sqrt{n}$, de moyenne $n\mu$.

On a donc

$$K^*(\mu/t',n') \; \alpha \; e^{-\frac{1}{2n'\sigma^2}(t'-n'\mu)^2} \qquad n' > O$$

Si $D'(\mu) = K^*(\mu/t',n')$, ce qui correspond à une distribution de Laplace Gauss de moyenne t'/n', d'écart type $\sigma/\sqrt{n'}$, alors $D''(\mu/t,n) = K^*(\mu/t+t',n+n')$ si $t = \sum_{i=1}^{n} x_i$.

La loi a posteriori de μ est une loi de Gauss de moyenne $\frac{t+t'}{n+n'}$ d'écart d'écart type $\frac{\sigma}{\sqrt{n+n'}}$

Remarque :

$T^*(x,n) = (\overline{X},n)$ avec $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ est aussi une statistique exhaustive de l'échantillon, mais on n'a pas

$$T^*(xx', nn') = T^*(x, n) + T^*(x', n') = (\overline{X} + \overline{X}', n + n')$$

parce que (X, n) n'est pas la statistique exhaustive attachée à la forme exponentielle de la loi.

---

(1) α signifie "proportionnel à".

e) Cas d'un échantillon d'une variable suivant une loi exponentielle de paramètre θ

$$l(x/\theta, n) = e^{-\sum_{i=1}^{n} x_i/\theta} \cdot \frac{1}{\theta^n} \qquad \theta > O,\ x_i > O$$

$$i = 1,\ 2, \ldots,\ n.$$

La statistique exhaustive intervenant dans la forme exponentielle est (T, n) avec $T = \sum_{i=1}^{n} X_i$ et $\left(\frac{T}{\theta}\right)$ suit une loi γ_n.

On a donc

$$l(t/\theta, n) = e^{-t/\theta}\left(\frac{t}{\theta}\right)^{n-1} \cdot \frac{1}{\theta} \qquad t > O,\ \theta > O$$

et

$$K^*(\theta/t', n') \ \alpha\ e^{-t'/\theta}\left(\frac{t'}{\theta}\right)^{n'-1} \cdot \frac{1}{\theta}, \quad t' > O,\ \theta > O$$

soit

$$K^*(\theta/t', n') \ \alpha\ e^{-t'/\theta}\left(\frac{t'}{\theta}\right)^{n'-2} \qquad t' > O,\ \theta > O$$

le facteur t' peut, en effet, entrer dans le coefficient de proportionnalité puisqu'il ne joue plus que le rôle d'un paramètre.

Si $D'(\theta) = K^*(\theta/t', n')$, on a

$$D''(\theta/t, n) = K^*(\theta/t+t',\ n+n') \ \alpha\ e^{-\frac{t+t'}{\theta}}\left(\frac{t+t'}{\theta}\right)^{n+n'-2}$$

Il est plus simple de faire intervenir de paramètre $\omega = 1/\theta$ dont la densité par rapport à la mesure de Lebesgue sur R est $D^*(\omega) = D(\theta) \cdot \left|\frac{d\theta}{d\omega}\right| = \frac{D(\theta)}{\omega^2}$

soit

$$D(\omega) \ \alpha\ e^{-\omega t'} \frac{(\omega t')^{n'-2}}{\omega^2} \ \alpha\ e^{-\omega t'} (\omega t')^{n'}$$

On reconnaît pour $\omega = 1/\theta$ une loi $\gamma_{n'+1}$ de paramètre $1/t$, ce qui permet de préciser que $D^*(\omega) = \frac{e^{-\omega t'} (\omega t')^{n'} t'}{\Gamma(n'+1)}$

C'est une densité pour $t' > O$, $n' > -1$ quelconque et non nécessairement entier.

e') Cas d'un échantillon d'une variable suivant une loi γ_p de paramètre θ, p étant connu

$$l(x/\theta, n) \alpha\, e^{-\sum_{i=1}^{n} x_i/\theta} \left(\prod_{i=1}^{n} x_i\right)^{p-1} \cdot \frac{1}{\theta^{np}} \qquad \begin{array}{l}\theta > 0\\ x_i > 0\end{array}$$

La statistique exhaustive intervenant dans la forme exponentielle est (T, n) avec $T = \sum_{i=1}^{n} X_i$ et T/θ suit une loi γ_{np}. On est donc ramené au problème précédent en y remplaçant respectivement n et n' par np et n'p.

On a donc

$$K^*(\theta/t', n') \;\alpha\; e^{-\frac{t'}{\theta}} \left(\frac{t'}{\theta}\right)^{n'p-2}$$

et si

$$D'(\theta) = K^*(\theta/t', n')\ , \ D''(\theta) \alpha\, e^{-\frac{t+t'}{\theta}} \left(\frac{t+t'}{\theta}\right)^{(n+n')p-2}$$

f) Cas d'un échantillon d'une variable suivant une loi de Gauss de de moyenne μ et d'écart type σ inconnus

$$l(x/\mu, \sigma, n) \;\alpha \frac{1}{\sigma^n}\; e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2} = \frac{1}{\sigma^n} \cdot e^{-\frac{1}{2\sigma^2}(\sum x_i^2 - 2\mu\sum x_i + n\mu^2)}$$

La statistique exhaustive intervenant dans la forme exponentielle est $(T_1 = \sum_{i=1}^{n} X_i,\ T_2 = \sum_{i=1}^{n} X_i^2,\ n)$ et puisque $\frac{\sum(X_i - \overline{X})^2}{\sigma^2} = \frac{T_2 - \frac{T_1^2}{n}}{\sigma^2}$ suit, indépendamment de T_1, une loi de χ^2 à n-1 degrés de liberté, la densité de la statistique exhaustive est :

$$l(t_1, t_2/\mu, \sigma, n) \;\alpha\; \frac{1}{\sigma\sqrt{n}} \cdot \frac{1}{\sigma^2}\, e^{-\frac{1}{2\sigma^2 n}(t_1 - n\mu)^2}\; e^{-\frac{\left(t_2 - \frac{t_1^2}{n}\right)}{2\sigma^2}} \left(\frac{t_2 - \frac{t_1^2}{n}}{2\sigma^2}\right)^{(n-3)/2}$$

Il est plus simple de faire intervenir au lieu de σ, le paramètre $h = 1/\sigma^2$. Il vient alors :

$$l(t_1, t_2/\mu, h, n) \;\alpha\; h^{\frac{n}{2}} \times e^{-\frac{h}{2n}\left[(t_1 - n\mu)^2 + nt_2 - t_1^2\right]} \times \left(t_2 - \frac{t_1^2}{n}\right)^{(n-3)/2}$$

D'où on déduit

$$K^*(\mu, h/t_1', t_2', n') \;\alpha\; e^{-\frac{hn'}{2}\left(\mu - \frac{t_1'}{n'}\right)^2}\; e^{-\frac{h}{2}\left(t_2' - \frac{t_1'^2}{n'}\right)}\; h^{\frac{n'}{2}}$$

Cette distribution qui est définie pour μ et t_1' quelconques, $h > 0$, $n' > 0$, $t_2' > \frac{t_1'^2}{n'}$ est dite "normale Gamma" de paramètres $m' = \frac{t_1'}{n'}$, $v' = t_2' - \frac{t_1'^2}{n'}$, n' ; $v' > 0$, $n' > 0$ et la loi à posteriori de μ et h est une loi normale Gamma de paramètres m'', v'', n'' associés à $t_1'' = t_1' + t_1$, $t_2'' = t_2' + t_2$ $n'' = n' + n$.

La justification de cette dénomination provient de la remarque suivante :

En écrivant :

$$K^*(\mu, h/t'_1, t'_2, n') \propto h^{\frac{1}{2}} \underbrace{e^{-\frac{h}{2}n'(\mu-\frac{t'_1}{n'})^2}}_{A} \times \underbrace{e^{-\frac{h}{2}(t'_2-\frac{t'^2_1}{n'})} h^{\frac{n'-1}{2}}}_{B}$$

on voit apparaître en A la densité de la loi de μ conditionnée par h qui est une loi normale de moyenne t'_1/n et de variance $\frac{1}{n'h} = \frac{\sigma^2}{n'}$, c'est-à-dire la même que la loi de μ quand σ est connu.

On a alors en B la densité de la loi marginale de h et on voit que $\frac{h}{2}(t'_2 - \frac{t'^2_1}{n'})$ suit une loi γ à $\frac{n'-1}{2}$ degrés de liberté et par suite $h(t'_2 - \frac{t'^2_1}{n'})$ suit une loi de χ^2 à n'-1 degrés de liberté. Quant à la distribution marginale de μ on peut montrer que la variable.

$$\frac{(\mu - \frac{t'_1}{n'}) \sqrt{n'(n'-1)}}{\sqrt{t'_2 - \frac{t'^2_1}{n'}}} = \frac{(\mu' - \frac{t'_1}{n'}) \sqrt{n'(n'-1)}}{\sqrt{v'}}$$

suit une loi de Student à n' - 1 degrés de liberté.

On en déduit que la moyenne de la loi marginale de μ est t_1/n' tandis que sa variance, qui n'est finie que si n' > 3, est $V'(\mu) = \frac{v'}{n'(n'-3)}$ avec $v' = t'_2 - \frac{t_1^2}{n'}$ (la variance d'une variable de Student à ν degrés de liberté est $V(t) = \frac{\nu}{\nu - 2}$ et n'existe que si $\nu > 2$).

IV - APPLICATION A UN EXEMPLE - ESTIMATION PAR "ECHANTILLONNAGE" DE LA MOYENNE μ D'UNE DISTRIBUTION NORMALE DE VARIANCE CONNUE σ^2 (3).

On a $\Theta = \mathbf{R}$ et l'espace des décisions $A = \Theta$ puisqu'il s'agit d'un problème d'estimation.

$E = \mathbf{N}$ puisque e = n taille de l'échantillon à prélever et $\mathscr{X}(n) = \mathbf{R}^n$.

Une régle de décision δ est un estimateur $\hat{\mu}(x,n)$ de μ. Si on admet la possibilité de ne pas prélever d'échantillon, il convient de définir $\hat{\mu}(.,0)$.

On peut dire que $\hat{\mu}(x,n)$ est une application de $\{\mathscr{X}(n)\}_{n\in\mathbf{N}}$ dans $\mathbf{R}$ en convenant de définir arbitrairement $\mathscr{X}(0)$, par exemple par un ensemble réduit à un point et de prendre $\hat{\mu}(x,0) = \hat{\mu}(.,0)$.

Il s'agit de déterminer l'estimateur optimal $\hat{\mu}^*(x,n)$ s'il existe, et l'effectif optimal n^* de l'échantillon à prélever. On a adopté une fonction de coût additive par rapport au coût de l'expérimentation et aux erreurs d'estimation, et quadratique par rapport à ce dernir type d'erreurs, c'est-à-dire de la forme μ

$$W(n,\hat{\mu}\ \mu) = K(\mu - \mu)^2 + a\delta(n) + bn$$

avec

$$\delta(n) = 1 \text{ si } n > 0$$

$$\delta(n) = 0 \text{ si } n = 0$$

$$K > 0,\ a \geqslant 0,\ b > 0$$

On a vu, cf. II.2 qu'il convenait dans un premier stade, où n est supposé connu de rechercher un estimateur $\hat{\mu}^*(x,n)$ tel que l'espérance de $W(n,\hat{\mu},\mu)$ par rapport à la distribution à posteriori de μ soit minimum.

Comme le coût de l'échantillonnage est indépendant de μ, on est conduit à minimiser

$$E''_{\mu/x,n}\,[K(\hat{\mu}-\mu)^2]^2 = K\,E''_{\mu/x,n}(\hat{\mu}-\mu)^2$$

Puisque $E''_{\mu}\ (\hat{\mu}-\mu)^2$ est le moment d'ordre 2 de la loi à posteriori de μ par rapport à l'origine $\hat{\mu}(x,n)$, on sait qu'il sera minimum si $\hat{\mu}(x,n)$ est la moyenne de la loi a posteriori de μ.

On prendra donc $\hat{\mu}^*(x,n) = E''_{\mu/x,n}\ (\mu)$ de sorte que

$$E''_{\mu/x,n}\ (\hat{\mu}^* - \mu)^2 = V''_{\mu/x,n}\ (\mu)$$

variance de la distribution à posteriori de μ.

n doit être choisi (cf. II.2) de telle façon que

$$E_{x/n}[E''_{\mu/x,n}\{W(n,\mu^*,\mu)\}] \text{ soit minimum.}$$

$$E''_{\mu/x,n}\,[W(n,\mu^*,\mu)] = E''_{\mu/x,n}[K(\mu^* - \mu)^2] + a\delta(n) + bn$$

$$= KV''_{\mu/x,n}(\mu) + a\delta(n) + bn$$

Par suite, n^* doit être choisi de façon à minimiser

$$K E_{x,n}\ [V''(\mu)] + a\ \delta(n) + bn.$$

Pour déterminer n^* il faut expliciter $V''(\mu)$ et pour cela faire choix d'une loi à priori pour μ.

On prend pour densité à priori de μ une conjuguée naturelle des vraisemblances de μ relatives à un n' échantillon d'une variable gaussienne de moyenne inconnue et d'écart type σ connu, soit $D'(\mu) = N\left(m', \frac{\sigma^2}{n'}\right)$ densité d'une variate gaussienne de moyenne $m' = t'/n'$, de variance σ^2/n' (cf. III.3 d).

Pratiquement, on prendra pour m' la valeur à priori la plus probable pour μ et on choisira n' d'autant plus grand qu'on est a priori plus sûr que μ est proche de m' (On pourra utiliser par exemple le fait que l'intervalle $\left(m' - \frac{3\sigma}{\sqrt{n'}},\ m' + \frac{3\sigma}{\sqrt{n'}}\right)$ a une probabilité à priori égale à 99,8 % de contenir μ).

On sait qu'alors

$$D''(\mu/x,n) = N(m'', \frac{\sigma^2}{n+n'})$$

avec $n''m'' = n\bar{x} + n'm'$

$$(t'' = t + t' \text{ et } t = \sum x_i)$$

On a donc

$$V''(\mu) = \frac{\sigma^2}{n+n'}$$

et

$$E_{x/n}[V''(\mu)] = \frac{\sigma^2}{n+n'}$$

n^* doit donc minimiser $H(n) = k \frac{\sigma^2}{n+n'} + a\ \delta(n) + bn.$

où n', a, b, K et δ sont donnés.

L'étude des variations de $H(n)$ avec n conduit à choisir pour n^* les valeurs suivantes :

$$\begin{cases} n^* = 0 \text{ si } \sigma\sqrt{\frac{K}{b}} \leqslant n' \\ n^* = 0 \text{ si } \sigma\sqrt{\frac{K}{b}} > n' \text{ et } a > a_o \\ \text{avec :} \\ \qquad a_o = \frac{K\sigma^2}{n'} + b\,n' - 2\ \sigma\sqrt{bK} \\ n^* = \sigma\sqrt{\frac{K}{b}} - n' \text{ si } \sigma\sqrt{\frac{K}{b}} > n' \text{ et } a \leqslant a_o \end{cases}$$

Dans le dernier cas, on devra prendre en fait pour n^* l'un des deux entiers encadrant la valeur indiquée ci-dessus.

Rappelons que l'on doit prendre pour estimateur $\mu^*(x,n)$ la moyenne m'' de la loi a posteriori de μ.

Quand $n' = 0$, la loi a posteriori de μ est la même que sa loi a priori, par suite $\mu^*(0) = m'$, moyenne de la loi a priori de μ.

Quand $n^* > 0$, $\mu^*(x,n) = m'' = \frac{n\bar{x} + n'm'}{n+n'}$

Commentaires des résultats trouvés

Si a coût fixe de l'échantillonnage dépasse une certaine valeur ou (et) si n' est déjà grand, c'est-à-dire si on a déjà des idées assez précises sur μ, on peut avoir intérêt à ne pas faire d'expérience.

Par ailleurs, lorsque $n^* > 0$, n croît avec σ (on doit faire d'autant plus d'observations qu'on sait a priori moins de choses sur μ) et avec K/b (on doit faire d'autant plus d'observations que le coût d'une erreur d'estimation est grand par rapport au coût d'une observation).

Enfin, la part de la connaissance a priori sur μ représentée par n' m' dans l'estimateur $\hat{\mu} = \frac{n'\, m' + nx}{n + n'}$ est d'autant plus faible que n est plus grand. Elle tend vers 0 quand n augmente indéfiniment puisqu'alors $\hat{\mu}$ tend vers $\bar{x}$ qui converge presque sûrement vers μ. $\hat{\mu}$ est donc un estimateur convergent de μ. // Il en est de même pour l'influence de n' dans $V''(\mu) = \frac{\sigma^2}{n+n'}$ qui tend vers 0 quand $n \longrightarrow \infty$.

La distribution à posteriori de μ converge donc vers la vraie valeur de μ quand l'effectif de l'échantillon augmente indéfiniment et ceci quelles que soient les valeurs adoptées pour les paramètres de la distribution a priori de μ.

Convergence de la distribution à posteriori de θ vers sa vraie valeur dans une expérience d'échantillonnage.

On peut se demander si ce résultat, déjà trouvé en II.1.2, exemple 1, est général, c'est-à-dire si dans une expérience d'échantillonnage la distribution à posteriori de θ converge vers la vraie valeur de θ, quand $n \longrightarrow \infty$, quelle que soit la loi à priori adoptée pour θ.

On peut démontrer, sous des conditions assez générales, qu'il en est bien ainsi, la convergence étant une convergence presque sûre, sous réserve que la densité à priori de cette vraie valeur ne soit pas nulle, c'est-à-dire qu'on n'ait pas déclaré cette valeur impossible à priori.

Extension du problème précédent au cas où la variance σ^2 n'est pas connue.

Si σ n'est pas connu on adoptera pour distribution à priori pour le couple (μ, h) avec $h = 1/\sigma^2$ une distribution normale gamma définie par t'_1, t'_2, n' (cf. III.2.3.f) où $m' = t'_1/n'$ est la moyenne de la loi marginale à priori de μ et $\frac{v'}{n'(n'-3)}$ avec $v' = t'_2 - \frac{t'^2_1}{n'}$ est la variance de cette loi.

Puisque la fonction de coût ne dépend pas de h, tout ce qui précède reste valable en considérant la loi marginale à posteriori de μ. On aura donc la même expression pour l'estimateur $\mu^*(x,n)$ de μ, qui ne dépend que de m', moyenne de la loi marginale à priori de μ, n', $\bar{x}$ et n.

Par contre, la détermination de n^* devient plus compliquée. La variance de la loi marginale à posteriori de μ est

$$V''(\mu/x,n) = \frac{v''}{n''(n''-3)} \text{ au lieu de } \frac{\sigma^2}{n'} \text{ (cf. III.3.f)}$$

Elle n'est finie que si $n'' > 3$ et dépend de x par l'intermédiaire de

$$v'' = \left(t''_2 - \frac{t''^2_1}{n''}\right) \text{ avec } \begin{array}{l} t''_1 = t'_1 + t_1 = t'_1 + \sum x_i \\ t''_2 = t'_2 + t_2 = t'_2 + \sum x_i^2 \end{array}$$

de sorte qu'on a

$$E_{x/n}\,[V''(\mu)] = \frac{E(v''/n)}{n''(n''-3)}$$

On peut montrer que $E_{x/n}\,[V''(\mu)] = \frac{v'}{(n'+n)(n'-3)}$ n'est finie que si $n' > 3$, de sorte que ce n'est que si $n' > 3$ que n^* est fini.

Dans le cas où $n' > 3$, l'étude qui a été faite dans le cas σ^2 connu, subsiste en remplaçant σ^2 par $\frac{v'}{n'-3}$. Les résultats peuvent être mis sous forme indépendante du fait qu'on connaît ou non σ^2 en les exprimant en fonction du paramètre $\lambda = \frac{\sigma^2}{n'}$ au lieu de σ^2, qui représente dans tous les cas la variance de la loi à priori de μ.

Que σ^2 soit connu ou non, la solution du problème ne dépend donc finalement que des paramètres de la loi à priori de μ, soit n', sa moyenne et sa variance, mais il faut noter que cette loi n'est pas la même si σ^2 est connu ou s'il ne l'est pas.

V - INTRODUCTION A L'EXPERIMENTATION SEQUENTIELLE

Il est formellement très aisé d'introduire un processus séquentiel dans le cadre bayesien. Il suffit dans l'espace A des décisions d'adjoindre aux décisions terminales, la décision a_0 qui consiste à recommencer un nouveau cycle : expérimentation, examen du résultat, décision, en prenant pour loi a priori de θ pour le nouveau cycle, la loi a posteriori que l'on a obtenue à la fin du cycle précédent, soit au re cycle : $D'_r(\theta) = D''_{r-1}(\theta)$.

Le problème le plus délicat est celui de la définition de la fonction de coût correspondant à a_0, à la fin du $r^{\text{ème}}$ cycle. C'est de sa définition pour $r = 1, 2...$ que dépend essentiellement la règle d'arrêt du processus.

Si les résultats X(e) des expériences successives sont des variables indépendantes et de même loi $P_{x/\theta}$, on démontre que tout se passe comme si on avait affaire à une suite de décisions non séquentielles indépendantes.

Ce résultat suggère une méthode de définition d'une règle d'arrêt.

On a vu que le choix de la décision $\delta^*(x_r, e_r) = a_r$ était fait de façon à ce que l'espérance $W_r^*(e_r, x_r, a_r)$ de la perte pour la loi a posteriori de θ, soit $D''_r(\theta/x_r, e_r)$ soit minimum.

Il s'agit donc de définir $W_r^*(e_r, x_r, a_r)$ et de la comparer à

$$\inf_{a \in A_r - \{a_0\}} [W_r^*(e_r, x_r, a)] = W_r^*(e_r, x_r)$$

qui correspond à la meilleure décision terminale après le $r^{\text{ème}}$ cycle.

A_r est l'ensemble des décisions possibles après le $r^{\text{ème}}$ cycle. Si on fait une nouvelle expérience e_{r+1}, elle est telle que

$$E_{x_{r+1}/e_{r+1}} \left\{ W^*_{r+1} [e_{r+1}, X_{r+1}, \delta_{r+1}(e_{r+1}, X_{r+1})] \right\} \text{soit} \qquad (10)$$

minimum (cf. II.2).

Pour déterminer e_{r+1} on se bornera aux règles de décision δ_{r+1} qui conduisent à une décision terminale, c'est-à-dire à des applications de $\{\mathcal{X}(e_{r+1})\}$ $e_{r+1} \in E_{r+1}$ dans $A_{r+1} - \{a_o\}$.

Soit W^{**}_{r+1} la valeur minimale de l'expression (10).

On posera $W^*_r(e_r, x_r, a_o) = \hat{W}^{**}_{r+1}$, valeur minimale de la perte moyenne pour la loi du couple (θ, X_{r+1}), quand on prend une décision terminale au r + 1ème cycle.

Par suite, si $W^*_r(e_r, x_r) \leqslant W^*_r(e_r, x_r, a_o)$ on arrête l'expérimentation après le rème cycle, tandis que dans le cas contraire on procède à une nouvelle expérience.

La règle d'arrêt à la fin du rème cycle est donc basée sur la comparaison de la valeur minimale de la perte moyenne associée au résultat (x_r, e_r) pour toutes les décisions terminales à la fin de ce cycle, à la valeur minimale de la perte moyenne au cycle suivant pour toutes les règles de décision terminales à la fin de ce nouveau cycle (le r+1ème).

Il semble que ce résultat devrait pouvoir être généralisé au cas où les résultats des expériences successives sont des variates, indépendantes mais n'ont pas toutes la même loi(1).

C'est certainement un des avantages de la méthode bayésienne que de permettre d'introduire très naturellement l'expérimentation séquentielle, tout au moins d'un point de vue formel.

BIBLIOGRAPHIE

Ouvrages de base

[1] T.S. FERGUSON - Mathematical statistics. Academic Press, New-York (1967).

[2] H. CHERNOFF and L.E. MOSES - Elementary decision theory-Wiley and Sons (1959).

[3] H. RAIFFA and R. SCHLAIFER - Applied statistical decision theory. Harvard University Press, Boston (1961). Ouvrage fondamental pour la mise en oeuvre de la décision bayesienne.

---

(1) Si la densité a priori de θ, $D'_1(\theta)$ est conjuguée naturelle de sa vraisemblance relative a (X_1, e_1), la densité à posteriori $D''_1(\theta) = D'_2(\theta)$ risque fort dans ce cas de ne pas être conjuguée naturelle de la vraisemblance relative à (X_2, e_2) et on aura des difficultés pratiques à calculer $D''_2(\theta)$. Les conjuguées naturelles ne présentent donc d'intérêt dans l'expérimentation séquentielle que si les vraisemblances relatives aux résultats des expériences successives (X_1, e_1), (X_2, e_2)... (X_r, e_r)... ont la même forme, ou du moins les mêmes conjuguées naturelles.

Articles de cette Revue

[4] G. d'HERBEMONT - Aspects de la théorie statistique des décisions (1963), Vol. XI N° 3 p. 41/82.

[5] D. GROUCKHO - Application de la formule de Bayes au calcul d'un volant de rechanges pour des pièces de faible consommation (1964) Vol XII, N° 4.

[6] G. HANSEL et D. GROUCHKO - Prévision séquentielle par la méthode de Bayes (1965) Vol. XIII N° 3, p. 67/81.

[7] Ph. L'HARDY - Application des méthodes bayesiennes à l'estimation d'élasticités de consommation (1966) Vol. XIV N° 4.

[8] G. MAAREK et A. SEGOND - L'estimation séquentielle dans les modèles linéaires (1969) Vol XVII N° 2, p. 21/45.

ANNEXE

DEMONSTRATION DE LA FORMULE (3) de II.2 et justification de la méthode proposée

$$\underset{\substack{\delta \in \Delta \\ e \in E}}{\text{Min}} \; E_{\theta,\, x/e} \{W[e, X, \delta(X, e)\theta]\} = \underset{e \in E}{\text{Min}} \; E_{x/e} \left\{ \underset{a \in A}{\text{Min}} \; E''_{\theta/x,\, e} [W(e, x, a, \theta]\right\} \qquad (3)$$

Montrons que pour chaque $e \in E$, on a :

$$\underset{\delta \in \Delta}{\text{Min}} \; E_{\theta,\, x/e} \{W[e, X, \delta(X, e), \theta]\} = E_{x/e} \left\{ \underset{a \in A}{\text{Min}} \; E''_{\theta/x,\, e} [W(e, x, a, \theta]\right\} \qquad (3A)$$

sous réserve que le deuxième membre de (3A) existe effectivement.

1°) On a

$$\underset{\delta \in \Delta}{\text{Inf}} \; E_{\theta,\, x/e}\{W[e, X, \delta(X, e), \theta]\} > E_{x/e} \left\{ \underset{a \in A}{\text{Min}} \; E''_{\theta/x,\, e} [W(e, x, a, \theta)]\right\}$$

$$\forall \; e \in E.$$

En effet, on a pour toute règle $\delta \in \Delta$

$$E''_{\theta/x,\, e} \{W[e, x, \delta(x, e), \theta]\} \geqslant \underset{a \quad A}{\text{Min}} \; E''_{\theta/x,\, e} [W(e, x, a, \theta)]$$

puisque $\delta(x, e) \in A$

On en déduit :

$$\underbrace{E_{x/e} \; E''_{\theta/x,\, e}}_{E_{\theta,\, x/e}} \{W[e, X, \delta(X, e), \theta]\} \geqslant E_{x/e} \left\{ \underset{a \in A}{\text{Min}} \; E''_{\theta/x,\, e} [W(e, x, a, \theta)]\right\}$$

et par suite :

$$\operatorname*{Inf}_{\delta \in \Delta} E_{\theta, x/e}\{W[e, X, \delta(X,e), \theta]\} \geqslant E_{x/e}\left\{\operatorname*{Min}_{a \in A} E''_{\theta/x, e}\ [W(e,x,a,\theta)]\right\}$$

C.Q.F.D.

2°) <u>Il existe $\delta^*(x,e) \in \Delta$, telle que $\forall\ e \in E$</u>

$$: \ E_{\theta, x/e}\{W[e,x,\delta^*(x,e), \theta]\} = E_{x/e}\left\{\operatorname*{Min}_{a \in A} E''_{\theta/x, e}\ [W(e,x,a,\theta)]\right\} \qquad (4\,A)$$

Soit $\delta^*(x,e)$ l'application de $\mathfrak{X}$ dans A définie par :

$$E''_{\theta/x, e}\{W[e,x,\delta^*(x,e), \theta]\} = \operatorname*{Min}_{a \in A} E''_{\theta/x, e}[W(e,x,a,\theta)] \text{ pour tout } (x,e) \in \mathfrak{X} \quad (4)$$

$\delta^*(x,e)$ existe puisque le 2ème membre de (4) existe. On peut la construire par points.

Montrons que $\delta^* \in \Delta$.

De (4), on déduit

$$: \ \underbrace{E_{x/e}\{E_{\theta/x, e}}_{E_{\theta, x/e}}\{W[e,x,\delta^*(x,e), \theta]\} = E_{x/e}\left\{\operatorname*{Min}_{a \in A} E''_{\theta/x, e}\ [W(e,x,a,\theta)]\right\} \qquad (5\,A)$$

Puisque le 2ème membre de cette égalité existe, il en est de même du premier membre ce qui prouve que $\delta^* \in \Delta$.

3°) De 1°) et 2°) on déduit que $E_{\theta, x/e}[W(e,x,\delta(X,e),\theta]$ passe effectivement par un minimum pour tout $e \in E$ quand δ parcourt Δ et que ce minimum est atteint pour $\delta = \delta^*$ définie par (4) ce qui démontre (3) et justifie la méthode proposée en II.2.