L'ANALYSE STATISTIQUE BAYÉSIENNE

DOMINIQUE CELLIER (*)

...Toi, qui de l'univers en marche ne sait rien
Tu es bâti de vent : par suite tu n'es rien.
Ta vie est comme un pont jeté entre deux vides :
Tu n'as pas de limite, au milieu tu n'es rien...
Omar Khayyâm

- Introduction -

La difficulté de cet exposé réside dans le fait qu'il n'est pas évident de présenter en si peu de temps une théorie qui, d'une part, est un véritable champ de bataille de polémiques et qui, d'autre part, est peu développée en France tant au niveau de l'enseignement de la statistique qu'au niveau des applications pratiques.

Les raisons de cet état de fait seraient longues à exposer ici. Cependant on peut penser que "l'objectivité" des esprits dits "cartésiens" ne peut être que choquée, ignorante parfois ou méprisante vis à vis d'une théorie qui fait "apparemment" appel à la "subjectivité" : l'analyse bayésienne prend fondamentalement en compte les "a priori", l'apprentissage et l'expérience acquise de l'expérimentateur. Certains aspects élémentaires, intuitifs d'un point de vue mathématique font préférer souvent à l'analyse statistique bayésienne des théories plus complexes mathématiquement.

L'analyse statistique bayésienne est essentiellement un principe de dualité et une démarche cohérente d'inversion.

Le travail d'un statisticien consiste en dernière analyse à "*remonter des effets aux causes*". On observe les effets d'un phénomène et on cherche, sur la base de cette observation, à faire une déduction (une inférence) sur les causes qui provoquent ces effets :

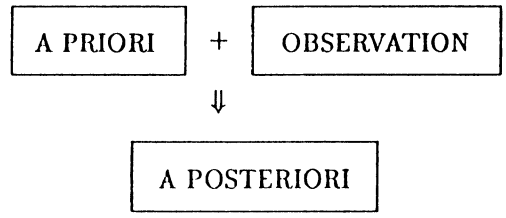
- estimer des paramètres inconnus,
- émettre, accepter ou rejeter des hypothèses,
- prédire des observations futures ...etc...

La statistique inférentielle est donc une démarche d'"*inversion*".

EFFETS $\hookrightarrow$? CAUSES

(*) Université de Rouen - Laboratoire Analyse et Modèles Stochastiques - U.R.A. C.N.R.S. 1378.

La démarche bayésienne est sans doute celle qui s'inscrit de façon la plus cohérente dans cette problématique d'inversion : elle met en œuvre effectivement cette dernière qui consiste à remonter des effets (les observations) aux causes (les paramètres).



L'a posteriori consiste en une réactualisation de la connaissance ou du degré d'ignorance du ou des paramètres.

- I - Généralités sur la démarche bayésienne -

...Quand vous avez éliminé l'impossible
Ce qui reste, même improbable,
Doit être la vérité...
Conan Doyle

I.1 - Statistique inférentielle et théorie des probabilités.

Quel rapport existe-t-il entre la Statistique qui repose sur l'observation de phénomènes concrets et la théorie des probabilités qui traite des propriétés de certaines structures modélisant des phénomènes où le hasard intervient ?

1 - Les données observées sont imprécises, entachées d'erreurs. Le modèle probabiliste permet de les présenter comme des variables aléatoires (l'aléa provenant de la déviation entre vraies valeurs et valeurs observées).

2 - On constate parfois que la répartition statistique d'une variable au sein d'une population est voisine de modèles mathématiques proposés par le calcul des probabilités.

3 - On est souvent amené à modéliser des situations très complexes (nombre important de paramètres, paramètres cachés...). Le modèle proposé est alors simplifié grâce au calcul des probabilités.

4 - Enfin, le lien le plus important réside dans le fait que les échantillons d'individus observés sont la plupart du temps tirés au hasard dans la population, ceci pour en assurer la "représentativité". Chaque individu a alors une certaine probabilité d'appartenir à l'échantillon. Les caractéristiques observées deviennent, grâce au "tirage au sort", des variables aléatoires dont le calcul des probabilités permet d'étudier les propriétés, le comportement etc...

I.2 - Modèle statistique et vraisemblance.

Compte tenu de ce qui précède, on comprend bien le rôle important joué par la modélisation en statistique inférentielle. Un modèle statistique consiste en l'observation d'une variable aléatoire X, de loi de probabilité P_θ. Le modèle peut alors s'écrire

$$\left(\mathcal{X}, (P_t)_{\theta\in\Theta}\right)$$

où

- $\mathcal{X}$ est l'espace des observations (les effets)
- P_θ est la loi de l'observation qui dépend d'un paramètre inconnu $\theta \in \Theta$ (représentant les causes).

L'espace des observations peut être discret. Par exemple X peut prendre ses valeurs dans $\{0, 1, 2, \cdots, n\}$ et avoir pour loi une loi Binomiale de paramètre (n, θ), $\theta \in [0, 1]$ inconnu. Dans ce cas on a

$$\forall k \in \{0, 1, 2, \cdots, n\} \qquad P_\theta(X = k) = C_n^k \theta^k (1-\theta)^{n-k} = l(\theta, k)$$

La fonction $l(\theta, \cdot)$ est appelée la *vraisemblance*.

L'espace des observations peut être continu. Par exemple X peut prendre ses valeurs dans $\mathbf{R}$ et avoir pour loi une loi de Laplace-Gauss $\mathcal{N}(m, \sigma^2)$, $\theta = (m, \sigma^2)$ inconnu. Dans ce cas on a

$$\forall [a.b] \subset \mathbf{R} \qquad P_\theta(X \in [a, b]) = \int_a^b \frac{1}{\sqrt{2\pi}\sigma} \cdot \exp\left(-\frac{1}{2\sigma^2}(x-m)^2\right) dx$$
$$= \int_a^b l(\theta, x) dx$$

La fonction $l(\theta, \cdot)$ est encore appelée la *vraisemblance*.

Une fois le modèle statistique construit, on cherche à établir, sur la base de l'observation et de sa vraisemblance, une inférence sur le paramètre θ inconnu.

I.3 - Le théorème de Bayes.

De manière générale, cette démarche d'inversion qui consiste à remonter des effets aux causes est décrite par le *Théorème de Bayes* : si A et E sont deux évènements tels que $P(E) \neq 0$, on a

$$P(A \mid E) = \frac{P(E \mid A) \cdot P(A)}{P(E \mid A) \cdot P(A) + P(E \mid \bar{A}) \cdot P(\bar{A})}$$

Ce théorème est une simple conséquence de la définition de la probabilité conditionnelle

$$P(A \mid E) = \frac{P(A \cap E)}{P(E)}$$

Ce théorème a constitué un saut conceptuel majeur dans l'histoire de la théorie des probabilités et de la statistique : c'est la première formule d'inversion des probabilités. En termes d'apprentissage, il décrit l'actualisation de la vraisemblance de A après que E ait été observé. En termes statistiques, il actualise l'information sur le paramètre inconnu θ (les causes) au vu de l'observation x (l'effet).

Ce théorème fondamental est à la base de la "*statistique bayésienne*" : il met sur un pied d'égalité causes et effets, tous deux pouvant être probabilisés.

- Exemple.

Dans une usine, deux machines M_1 et M_2 fabriquent des boulons de même type. M_1 sort en moyenne 0,3% de boulons défectueux et M_2 en sort 0,8% .

On mélange dans une caisse 250 boulons provenant de M_1 et 750 de M_2. On tire au hasard 1 boulon dans la caisse, on constate qu'il est défectueux. *Quelle est la probabilité qu'il ait été fabriqué par* M_1 *?*

Lorsqu'on tire un boulon au hasard, *a priori* la probabilité qu'il provienne

de M_1 est 0,25

de M_2 est 0,75 .

Si on observe qu'il est défectueux (évènement D), on calcule les probabilités conditionnelles $P(M_1 \mid D)$ et $P(M_2 \mid D)$ pour répondre à la question

$$P(M_1 \mid D) = \frac{P(D \mid M_1) \cdot P(M_1)}{P(D \mid M_1) \cdot P(M_1) + P(D \mid M_2) \cdot P(M_2)}$$
$$= \frac{0,003 \cdot 0,25}{0,003 \cdot 0,25 + 0,008 \cdot 0,75}$$
$$\simeq 0,11$$

Evidemment on a $P(M_2 \mid D) \simeq 0,89$.

$P(M_1 \mid D)$ et $P(M_2 \mid D)$ sont les probabilités *a posteriori* sachant que le boulon observé est défectueux.

- version continue du théorème de Bayes.

Supposons que l'observation X à valeurs réelles est de vraisemblance $l(\theta, \cdot)$, $\theta \in \mathbf{R}$ inconnu. Supposons que θ ait une loi Π de densité $\pi(\cdot)$, c'est-à-dire

$$\forall [a,b] \subset \mathbf{R} \qquad \Pi([a,b]) = \int_a^b \pi(t)dt$$

Alors, a posteriori, la loi conditionnelle $\Pi(\cdot \mid X = x)$ de θ sachant qu'on a observé $X = x$ a une densité $\pi(\cdot \mid x)$ donnée par

$$\pi(\theta \mid x) = \frac{l(\theta, x) \cdot \pi(\theta)}{\int_{-\infty}^{+\infty} l(t, x) \cdot \pi(t)dt}$$

Le numérateur de l'expression est la densité de la *loi conjointe* du couple (θ, X), le dénominateur est la densité, notée ρ, de la *loi prédictive* de X. Alors

$$\forall [a,b] \subset \mathbf{R} \qquad \Pi(\theta \in [a,b] \mid X = x) = \int_a^b \pi(t \mid x)dt$$

I.4 - Modèle statistique bayésien.

La problématique de l'analyse statistique bayésienne consiste à introduire une loi de probabilité Π sur l'espace des paramètres : idée révolutionnaire qui continue à diviser les statisticiens.

On passe alors de la notion de *paramètre inconnu* à la notion de *paramètre aléatoire*. Cette loi de probabilité sur l'espace des paramètres est appelée *loi a priori*.

La première question qui vient à l'esprit est évidemment : que représente en fait cette loi a priori ?

- Dans certains cas, le paramètre inconnu est réellement aléatoire ou peut être perçu comme tel.
- Mais dans la majorité des cas c'est impossible et c'est là que réside l'argument principal des adversaires de l'approche bayésienne.

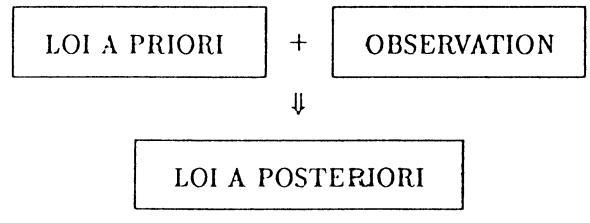
Prenons un exemple : La vitesse c de la lumière est en fait à jamais inconnue du fait de la limitation des appareils de mesure. Il est alors légitime de considérer c comme une variable aléatoire uniforme dans l'intervalle $[c_0 - \epsilon, c_0 + \epsilon]$ où

- ϵ représente la précision maximale actuelle des appareils de mesure
- et c_0 la mesure usuellement retenue.

L'importance de la loi a priori réside dans le fait qu'elle représente un moyen efficace de résumer l'information a priori disponible sur le paramètre inconnu ainsi que l'incertitude sur la valeur de cette information.

I.5 - Loi a priori, loi a posteriori.

L'essentiel des méthodes de l'analyse statistique bayésienne consiste, sur la base de la loi a priori et de l'observation effectuée, à déterminer la loi du paramètre conditionnellement à l'observation : la *loi a posteriori* qui actualise l'information sur le paramètre



Toute la statistique bayésienne repose sur cette loi a posteriori. Les difficultés essentielles proviennent, d'une part, de la détermination et du choix de la loi a priori et, d'autre part, du calcul explicite de la loi a posteriori.

Illustrons cela sur un exemple.

I.6 - Exemple

On observe une variable aléatoire réelle X dont la loi est une loi de Laplace-Gauss $\mathcal{N}(\theta, 1)$ où
- la moyenne $\theta \in \mathbf{R}$ est inconnue
- et la variance est connue et égale à 1.

On cherche, sur la base d'une observation, à estimer la paramètre θ.
Supposons que l'on choisisse comme loi a priori pour θ une loi de Laplace-Gauss $\mathcal{N}(\mu, \tau^2)$.
Déterminons la loi a posteriori. En vertu de la version continue du théorème de Bayes on a

$$\pi(\theta \mid x) = \frac{\exp\left(-\frac{1}{2}(x-\theta)^2\right) \cdot \exp\left(-\frac{1}{2\tau^2}(\theta-\mu)^2\right)}{\int_{-\infty}^{+\infty} \exp\left(-\frac{1}{2}(x-\theta)^2\right) \cdot \exp\left(-\frac{1}{2\tau^2}(\theta-\mu)^2\right) d\theta}$$

Le calcul du dénominateur donne

$$\frac{1}{\sqrt{2\pi}} \cdot \frac{1}{\sqrt{1+\tau^2}} \cdot \exp\left(-\frac{1}{2(1+\tau^2)}(x-\mu)^2\right)$$

On reconnaît la loi de Laplace-Gauss $\mathcal{N}(\mu, 1+\tau^2)$: c'est la *loi prédictive* de X.
Il vient alors

$$\pi(\theta \mid x) = \frac{1}{\sqrt{2\pi}} \cdot \sqrt{\frac{1+\tau^2}{\tau^2}} \cdot \exp\left(-\frac{1+\tau^2}{2\tau^2}\left[\theta - \frac{\tau^2}{1+\tau^2}\left(x + \frac{\mu}{\tau^2}\right)\right]^2\right)$$

On reconnaît de nouveau une loi de Laplace-Gauss. La *loi a posteriori* est donc la loi

$$\mathcal{N}\left(\frac{\tau^2}{1+\tau^2}\left(x + \frac{\mu}{\tau^2}\right), \frac{\tau^2}{1+\tau^2}\right)$$

Il y a eu réactualisation des paramètres de la loi a priori :

- la moyenne a priori μ devient la moyenne a posteriori

$$\frac{\tau^2}{1+\tau^2}\left(x + \frac{\mu}{\tau^2}\right)$$

- la variance a priori τ^2 devient la variance a posteriori

$$\frac{\tau^2}{1+\tau^2}$$

- II - Estimation de la moyenne d'une loi de Laplace-Gauss. -

...Estimer ne coûte rien.
Estimer incorrectement coûte cher..
Vieux proverbe chinois

2.1 - Estimation.

Nous allons traiter un exemple d'analyse statistique bayésienne : le problème de l'estimation de la moyenne inconnue d'une loi de Laplace-Gauss. Reprenons pour cela le cadre de l'exemple précédent.

On observe une variable aléatoire réelle X de loi $\mathcal{N}(\theta, 1)$. La moyenne θ est inconnue et la variance est connue (ici égale à 1).

Le modèle statistique associé à une observation est donc

$$\left(\mathbf{R}, (\mathcal{N}(\theta, 1))_{\theta \in \mathbf{R}}\right)$$

La vraisemblance $l(\theta, \cdot)$ est égale à

$$\forall x \in \mathbf{R} \qquad l(\theta, x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(x-\theta)^2\right)$$

Le problème est donc d'estimer le paramètre θ inconnu sur la base d'une observation. Un *estimateur* ϕ de θ est une application de $\mathbf{R}$ dans $\mathbf{R}$. Pour tout $x \in \mathbf{R}$, $\phi(x)$ est une *estimation* de θ.

L'estimateur usuel, noté $\hat{\phi}$, est l'estimateur des moindres carrés et du maximum de vraisemblance dans ce cas. Il est défini par

$$\forall x \in \mathbf{R} \qquad \hat{\phi}(x) = x$$

Intuitivement, il est naturel, sur la base d'une seule observation, d'estimer la moyenne inconnue θ par cette observation elle-même.

Cependant, on peut imaginer d'autres estimateurs possibles de θ. Il est donc indispensable d'utiliser un estimateur "le meilleur possible" et pour cela, on doit disposer de critères pour juger de la performance d'un estimateur et pour pouvoir comparer les estimateurs entre eux.

2.2 - Coût quadratique, risque quadratique.

Estimer θ par $\phi(x)$ entraîne une erreur inévitable que l'on peut mesurer par $(\phi(x) - \theta)^2$: le coût quadratique (ou erreur quadratique) encouru si on estime θ par $\phi(x)$.

Pour un estimateur ϕ de θ, on définit alors le *risque quadratique* de ϕ, noté R_ϕ par

$$\forall \theta \in \mathbf{R} \qquad R_\phi(\theta) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} (\phi(x) - \theta)^2 \exp\left(-\frac{1}{2}(x-\theta)^2\right) dx$$

c'est-à-dire le coût moyen encouru si on utilise l'estimateur ϕ. Il s'agit d'une fonction de $\mathbf{R}$ dans $\bar{\mathbf{R}}_+$.

Par exemple, pour l'estimateur usuel $\hat{\phi}$ on a

$$\forall \theta \in \mathbf{R} \qquad R_{\hat{\phi}}(\theta) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} (x-\theta)^2 \exp\left(-\frac{1}{2}(x-\theta)^2\right) dx = 1 \ .$$

Le risque de l'estimateur usuel est donc constant en θ.

2.3 - Comparaison des estimateurs.

On peut utiliser le critère du risque quadratique pour comparer les estimateurs entre eux.
Si ϕ et ψ sont deux estimateurs de θ, on dira que ϕ est meilleur (préférable) que ψ, on note $\phi \prec \psi$, si

$$\forall \theta \in \mathbf{R} \qquad R_\phi(\theta) \leq R_\psi(\theta)$$

On remarque immédiatement que deux estimateurs ne sont pas toujours comparables.

Si un estimateur n'est pas améliorable, on dit qu'il est *admissible*. Le problème de l'admissibilité est très compliqué (existence, unicité, calcul...) et demeure aujourd'hui un secteur important de la recherche en statistique.

Appliquons maintenant l'analyse statistique bayésienne pour proposer d'autres estimateurs dans le cas particulier que nous étudions.

2.4 - Loi a priori, loi a posteriori.

Comme dans l'exemple 1.6, nous choisissons comme loi a priori sur θ une loi de Laplace-Gauss $\mathcal{N}(\mu, \tau^2)$. Nous introduisons donc deux nouveaux paramètres

* μ : on pense a priori que θ est plutôt voisin de μ
* τ^2 : la variance, paramètre d'échelle mesurant la dispersion autour de μ, est ici un paramètre nuisible.

Comme dans l'exemple 1.6, on calcule la loi a posteriori $\Pi(\cdot \mid X = x)$

$$\Pi(\cdot \mid X = x) = \mathcal{N}\Big(\frac{\tau^2}{1+\tau^2}\big(x + \frac{\mu}{\tau^2}\big), \frac{\tau^2}{1+\tau^2}\Big)$$

Il y a réactualisation de notre information sur θ :

* la moyenne a posteriori est un barycentre de x et de μ
* la variance a posteriori vérifie la propriété suivante si on définit la précision comme l'inverse de la variance :

Précision a posteriori = Précision a priori + Précision de l'observation

2.5 - Risque bayésien, estimateur bayésien.

Soit ϕ un estimateur de θ. On peut calculer la moyenne de la fonction de risque R_ϕ de ϕ relativement à la loi a priori puisque θ est considéré comme aléatoire. On définit alors le *risque bayésien* de ϕ relativement à la loi a priori Π par

$$R_\Pi(\phi) = \frac{1}{\sqrt{2\pi}\tau} \int_{-\infty}^{+\infty} R_\phi(t) \cdot \exp\big(-\frac{1}{2\tau^2}(t-\mu)^2\big)\, dt$$

Par exemple, on a $R_\Pi(\hat{\phi}) = 1$.

Le risque bayésien permet de définir un nouveau critère de comparaison des estimateurs :

$$\phi \prec\prec \psi \qquad \Longleftrightarrow \qquad R_\Pi(\phi) \leq R_\Pi(\psi)$$

Deux remarques s'imposent.

1 - Comme R_Π est un nombre, deux estimateurs sont toujours comparables au sens du risque bayésien.

2 - Si $\phi \prec \psi$ alors $\phi \prec\prec \psi$. La réciproque est fausse.

Un estimateur ϕ^* est dit *bayésien* pour Π si, pour tout estimateur ϕ on a $\phi^* \prec\prec \phi$.

Plusieurs problèmes se posent alors : existence d'un estimateur bayésien, unicité, calcul.

2.6 - Coût moyen a posteriori.

Par interversion de l'ordre d'intégration, on montre que

$$\begin{aligned} R_\Pi(\phi) &= \int_{-\infty}^{+\infty} \Big(\int_{-\infty}^{+\infty} (\phi(x) - \theta)^2 l(\theta, x) dx\Big) \pi(\theta) d\theta \\ &= \int_{-\infty}^{+\infty} \Big(\int_{-\infty}^{+\infty} (\phi(x) - \theta)^2 \pi(\theta \mid x) d\theta\Big) \rho(x) dx \end{aligned}$$

L'intégrale $\int_{-\infty}^{+\infty} (\phi(x) - \theta)^2 \pi(\theta \mid x) d\theta$ représente le *coût moyen a posteriori*.

Le calcul précédent présente deux intérêts.

- Un intérêt pratique : si on minimise ce coût moyen a posteriori, alors on minimise le risque bayésien. Cela fournit donc une méthode de calcul de l'estimateur bayésien.
- Un intérêt sur le fond : on moyenne l'erreur quadratique sur toutes les valeurs possibles de θ ayant observé $X = x$. C'est plus naturel que de faire la moyenne sur toutes les valeurs possibles de X alors qu'on a observé une seule valeur de X.

2.7 - Existence et calcul de l'estimateur bayésien.

L'intégrale $\int_{-\infty}^{+\infty}(a-\theta)^2\pi(\theta \mid x)d\theta$ est minimisée pour

$$a = \int_{-\infty}^{+\infty} \theta\pi(\theta \mid x)d\theta$$

expression qui représente la moyenne de la loi a posteriori. Ainsi, *l'estimateur bayésien est la moyenne de la loi a posteriori.*

Dans notre exemple, l'estimateur bayésien ϕ_Π^* est donc défini par

$$\forall x \in \mathbf{R} \qquad \phi_\Pi^*(x) = \frac{\tau^2}{1+\tau^2}\left(x + \frac{\mu}{\tau^2}\right) = \mu + \left(1 - \frac{1}{1+\tau^2}\right)(x-\mu)$$

2.8 - Remarques.

On peut calculer le risque de l'estimateur bayésien ϕ_Π^* ainsi déterminé

$$\forall \theta \in \mathbf{R} \qquad R_{\phi_\Pi^*}(\theta) = \left(1 - \frac{1}{1+\tau^2}\right)^2 + \left(\frac{1}{1+\tau^2}\right)^2(\theta-\mu)^2$$

Quant au risque bayésien de ϕ_Π^*, il vérifie

$$R_\Pi(\phi_\Pi^*) = 1 - \frac{1}{1+\tau^2} < 1 = R_\Pi(\hat{\phi})$$

Pour conclure, on peut faire un certain nombre de remarques

1 - On a bien $\phi_\Pi^* \prec\prec \hat{\phi}$.

2 - Mais ϕ_Π^* et $\hat{\phi}$ ne sont pas comparables au sens du risque quadratique.

3 - ϕ_Π^* est meilleur que $\hat{\phi}$ (au sens du risque quadratique) au voisinage de μ.

4 - ϕ_Π^* est mauvais pour des grandes valeurs de θ puisque son risque quadratique converge vers l'infini lorsque θ tend vers l'infini.

5 - Si on fait tendre τ^2 vers l'infini alors ϕ_Π^* converge vers $\hat{\phi}$ et il y a aussi convergence du risque quadratique et du risque bayésien de ϕ_Π^* vers les risques respectifs de $\hat{\phi}$.

Le paramètre τ^2 apparaît bien comme un paramètre nuisible. On aimerait s'en débarrasser dans l'expression de l'estimateur bayésien. C'est une autre histoire qui sera abordée dans l'atelier correspondant.

Atelier :

ESTIMATION BAYÉSIENNE
- EFFET STEIN -

DOMINIQUE CELLIER (*)

- Le paradoxe de Stein(**) -

La meilleure estimation de la probabilité qu'un évènement se réalise est généralement identifiée à la moyenne arithmétique des résultats obtenus antérieurs. Le paradoxe de Stein définit les circonstances où existent des estimateurs meilleurs que cette moyenne.

Pour illustrer ce paradoxe, utilisons l'exemple développé par Efron et Morris dans leur article. On analyse le taux de réussite dans le renvoi de la balle avec la batte de 18 joueurs de baseball au cours de la saison 1970. Ces taux de réussite sont reproduits dans le tableau 1 page 2.

Pour le joueur numéro i, $1 \leq i \leq 18$, on note

θ_i : le taux de réussite du joueur pour l'année 1970,

y_i : le taux de réussite de ce même joueur à l'issu des 45 premiers essais de la même saison.

A l'issue des 45 premiers essais, on a donc observé $y = (y_1, y_2, \cdots, y_{18})$.

Si on nous avait demandé à ce moment précis d'estimer le taux de réussite $\theta = (\theta_1, \theta_2, \cdots, \theta_{18})$ sur l'ensemble de la saison, qu'aurions-nous prédit ? Probablement

$$\hat{\theta} = (y_1, y_2, \cdots, y_{18})$$

Car, traditionnellement, l'estimation fondée sur une observation est la valeur observée elle-même.

Le résultat de Stein est paradoxal en ce sens qu'il dément cette loi élémentaire de la théorie statistique : *si nous avons 3 joueurs ou plus, il existe une estimation "meilleure", c'est-à-dire avec plus de précision.*

Soit $\mu = \frac{1}{18} \sum_{i=1}^{18} y_i$ la moyenne des valeurs observées (ici $\mu = 0,265$). La phase essentielle de la méthode de Stein consiste à *rapprocher* chaque valeur observée y_i de μ de la façon suivante

$$\forall i \ , \ 1 \leq i \leq 18 \qquad\qquad \theta_i^* = \mu + c(y_i - \mu)$$

où c est une constante de rapprochement calculée à partir de l'observation (ici $c = 0,212$) .

On choisit alors l'estimateur $\theta^* = (\theta_1^*, \theta_2^*, \cdots, \theta_{18}^*)$ de θ défini par

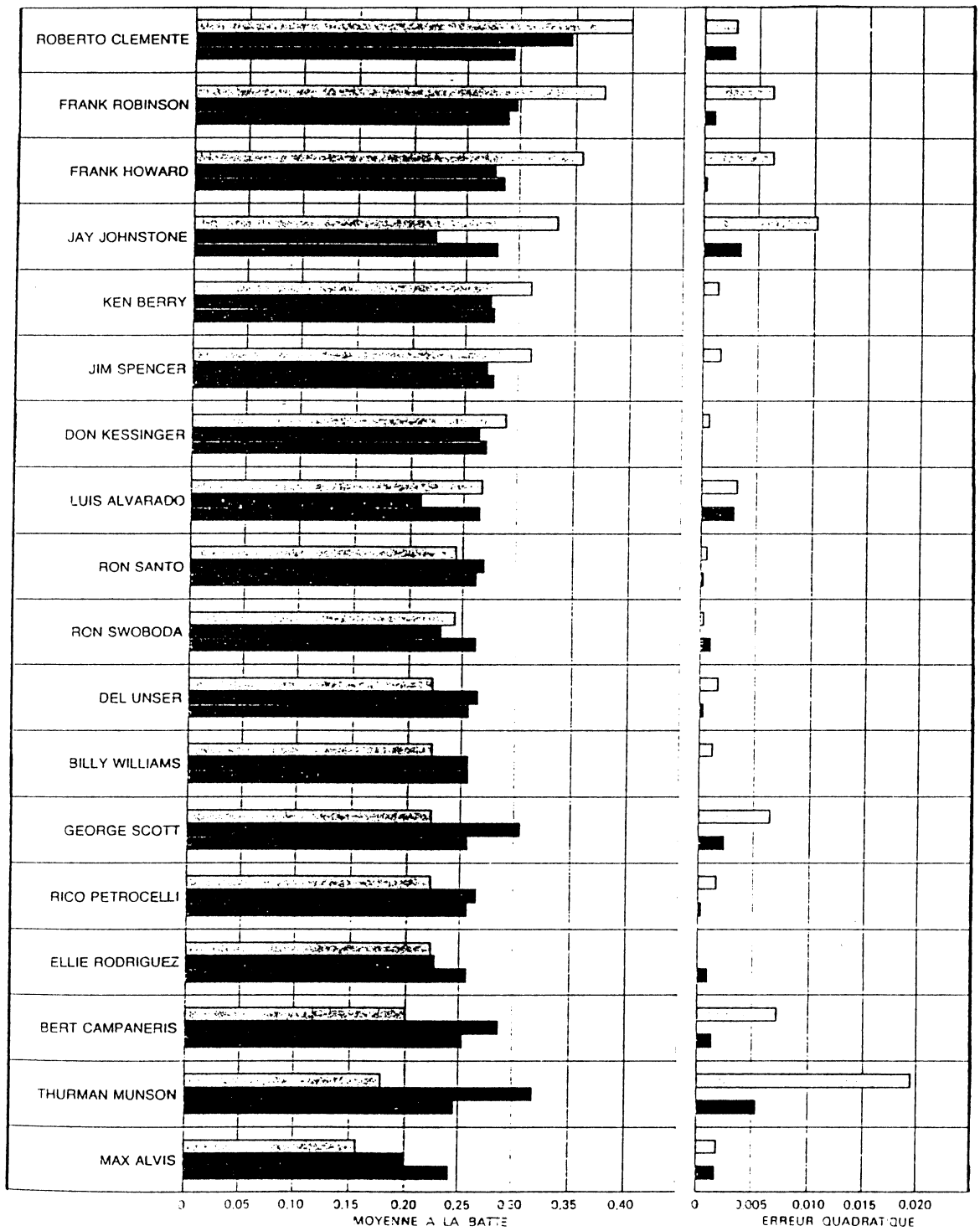
$$\forall i \ , \ 1 \leq i \leq 18 \qquad\qquad \theta_i^* = 0,265 + 0,212 \cdot (y_i - 0,265)$$

(voir tableau 2 page 3).

(*) Université de Rouen - Laboratoire Analyse et Modèles Stochastiques - U.R.A. C.N.R.S. 1378.

(**) D'après Bradley Efron et Carl Morris : Le paradoxe de Stein - Pour la Science, 1979, n° 1, p28-37.

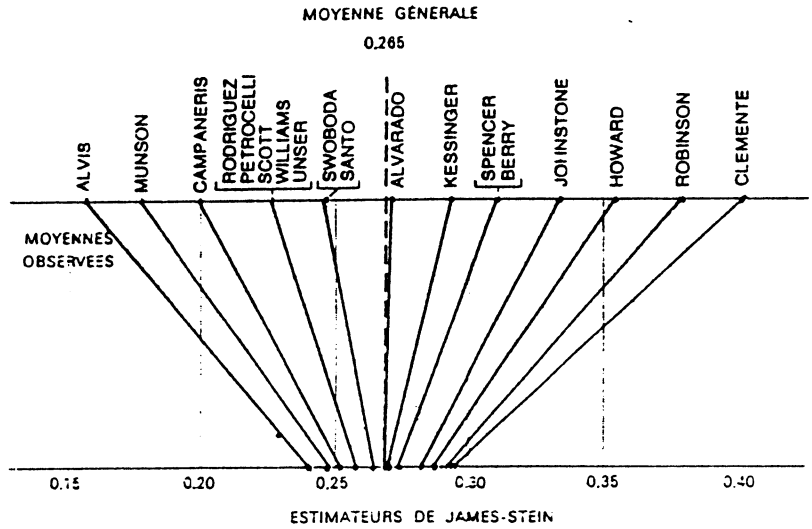
TABLEAU 1(*)



1. UNE ESTIMATION DES CAPACITÉS A LA BATTE de 18 joueurs de baseball est obtenue de façon plus précise par la méthode de James-Stein qu'en prenant les moyennes individuelles. Les moyennes employées comme estimateurs ont été calculées après que chaque joueur ait réalisé 45 essais pendant la saison 1970. La capacité réelle d'un joueur à la batte est une quantité inobservable, mais elle est approchée de près par la moyenne de ses performances sur une longue période. Ici, la capacité réelle est représentée par la moyenne à la batte obtenue pendant le reste de la saison de 1970. Pour l'estimation de la capacité à la batte de 16 joueurs sur 18, la moyenne arithmétique individuelle constitue une estimation moins bonne que l'estimation de James-Stein. L'ensemble des estimateurs de James-Stein a une erreur quadratique totale associée inférieure.

(*) Bradley Efron et Carl Morris : Le paradoxe de Stein - Pour la Science, 1979, n° 1, p29.

TABLEAU 2(*)



2. LES ESTIMATEURS DE JAMES STEIN pour les 18 joueurs de baseball ont été calculés en « rapprochant » les moyennes individuelles à la batte d'une « moyenne des moyennes individuelles » (moyenne globale). Dans ce cas, la moyenne globale vaut 0,265 et chacune des moyennes voit diminuer d'environ 80 % sa distance à cette valeur. Ainsi, le théorème sur lequel est basé la méthode de Stein affirme que les capacités réelles à la batte sont plus étroitement regroupées que les moyennes préliminaires n'auraient d'abord semblé le suggérer.

Nous disposons donc de deux estimateurs $\hat{\theta}$ et θ^* de θ. Lequel est le meilleur ?

Pour le joueur numéro i, les erreurs de prévision dans l'utilisation de ces estimateurs sont $(\hat{\theta}_i - \theta_i)$ et $(\theta_i^* - \theta_i)$.

On compare alors les deux estimateurs $\hat{\theta}$ et θ^* à l'aide de l'erreur quadratique globale

$$e(\hat{\theta}) = \sum_{i=1}^{18} (\hat{\theta}_i - \theta_i)^2 = 0,077$$

$$e(\theta^*) = \sum_{i=1}^{18} (\theta_i^* - \theta_i)^2 = 0,022$$

Ainsi l'*estimateur de Stein* θ^* est 3,5 fois plus précis au sens de l'erreur quadratique globale, il est par ailleurs meilleur pour 16 des 18 joueurs (voir tableau 1 page 2).

C'est un véritable défi au bon sens : pourquoi la réussite ou l'insuccès d'un joueur devrait influencer notre estimation d'un autre joueur ?

Plus choquant encore !

Supposons qu'on choisisse un échantillon de 45 voitures à Paris. Notons y_{19} la proportion de voitures étrangères dans l'échantillon et θ_{19} la proportion de voitures étrangères à Paris.

On peut injecter cette observation dans l'exemple précédent. L'observation est maintenant y', la moyenne globale devient μ' et la constante c c'.

On obtient un nouvel estimateur de $\theta' = (\theta_1, \theta_2, \cdots, \theta_{18}, \theta_{19})$

$$\theta' = \mu' + c'(y' - \mu')$$

qui est encore meilleur que l'estimateur usuel au sens de l'erreur quadratique globale.

Pour comprendre ce paradoxe appliquons l'analyse bayésienne dans le cas général suivant.

(*) Bradley Efron et Carl Morris : Le paradoxe de Stein - Pour la Science, 1979, n° 1, p30.

D.CELLIER : Estimation bayésienne - effet Stein

- Estimation bayésienne empirique -

On observe n variables aléatoires réelles indépendantes $(X_1, X_2, \cdots, X_n)$. On suppose que pour tout i, $1 \leq i \leq n$, la loi de X_i est une loi de Laplace-Gauss $\mathcal{N}(\theta_i, 1)$. Le paramètre $\theta = (\theta_1, \theta_2, \cdots, \theta_n) \in \mathbf{R}^n$ est inconnu et on désire l'estimer.

Un estimateur de θ est donc une application Φ de $\mathbf{R}^n$ dans $\mathbf{R}^n$

$$\forall (x_1, x_2, \cdots, x_n) \in \mathbf{R}^n \quad \Phi(x_1, x_2, \cdots, x_n) = \big(\phi_1(x_1, x_2, \cdots, x_n), \phi_2(x_1, x_2, \cdots, x_n), \cdots, \phi_n(x_1, x_2, \cdots, x_n)\big)$$

où $\phi_i(x_1, x_2, \cdots, x_n)$ estime θ_i.

Par exemple, l'estimateur usuel est $\hat{\Phi}$ défini par

$$\forall (x_1, x_2, \cdots, x_n) \in \mathbf{R}^n \qquad \Phi(x_1, x_2, \cdots, x_n) = (x_1, x_2, \cdots, x_n)$$

Les variables observées étant indépendantes, la vraisemblance $L(\theta, \cdot)$ du modèle statistique est définie par

$$\forall (x_1, x_2, \cdots, x_n) \in \mathbf{R}^n \quad L(\theta, (x_1, x_2, \cdots, x_n)) = \frac{1}{(2\pi)^{\frac{n}{2}}} \cdot \exp\Big(-\frac{1}{2}\sum_{i=1}^{n}(x_i - \theta_i)^2\Big)$$

Si on utilise le critère du risque quadratique, le risque R_Φ de tout estimateur Φ de θ est défini par

$$\forall \theta \in \mathbf{R}^n \qquad R_\Phi(\theta) = \sum_{i=1}^{n} \int_{\mathbf{R}^n} \big(\phi_i(x_1, x_2, \cdots, x_n) - \theta_i\big)^2 L(\theta, (x_1, x_2, \cdots, x_n)) dx_1 dx_2 \cdots dx_n$$

On vérifie facilement que le risque de l'estimateur usuel $\hat{\Phi}$ est constant et égal à n.

Estimation bayésienne.

Choisissons comme loi a priori pour θ_i une loi de Laplace-Gauss $\mathcal{N}(\mu_i, \tau^2)$ et supposons que les θ_i sont indépendants. Un calcul simple montre que la loi a posteriori de θ_i est une loi de Laplace-Gauss

$$\Pi(\cdot \mid (x_1, x_2, \cdots, x_n)) = \mathcal{N}\Big(\frac{\tau^2}{1+\tau^2}(x_i + \frac{\mu_i}{\tau^2}), \frac{\tau^2}{1+\tau^2}\Big)$$

L'estimateur bayésien $\Phi^* = (\phi_1^*, \phi_2^*, \cdots, \phi_n^*)$ de θ est la moyenne de la loi a posteriori et est donc défini par

$$\forall i \in \{1, 2, \cdots, n\} \quad \forall (x_1, x_2, \cdots, x_n) \in \mathbf{R}^n \quad \phi_i^*(x_1 . x_2, \cdots, x_n) = \frac{\tau^2}{1+\tau^2} \cdot (x_i + \frac{\mu_i}{\tau^2}) = \mu_i + \big(1 - \frac{1}{1+\tau^2}\big)(x_i - \mu_i)$$

On montre que le risque quadratique de cet estimateur est défini par

$$\forall \theta \in \mathbf{R}^n \qquad R_{\Phi^*}(\theta) = n\big(1 - \frac{1}{1+\tau^2}\big)^2 + \frac{1}{(1+\tau^2)^2}\sum_{i=1}^{n}(\theta_i - \mu_i)^2$$

Nous pouvons faire alors les mêmes remarques que dans l'exposé précédent concernant la comparaison de $\hat{\Phi}$ et Φ^*

1 - Φ^* est meilleur que $\hat{\Phi}$ au sens du risque bayésien ($\Phi^* \prec\prec \hat{\Phi}$).
2 - Ces deux estimateurs ne sont pas comparables au sens du risque quadratique.
3 - Φ^* est meilleur que $\hat{\Phi}$ (au sens du risque quadratique) au voisinage de $\mu = (\mu_1, \mu_2, \cdots, \mu_n)$.
4 - Φ^* est mauvais pour des grandes valeurs de $\|\theta\|$ puisque son risque quadratique converge vers l'infini lorsque $\|\theta\|$ tend vers l'infini.
5 - Si on fait tendre τ^2 vers l'infini, alors Φ^* converge vers $\hat{\Phi}$ et il y a aussi convergence du risque quadratique et du risque bayésien de Φ^* vers ceux de $\hat{\Phi}$.

Le paramètre τ^2 apparaît encore comme un paramètre nuisible dont on aimerait se débarrasser. Mais comment ?

Estimation bayésienne empirique.

Si on utilise la loi prédictive de l'observation, on montre que relativement à cette loi, la variable aléatoire

$$\frac{n-2}{\sum_{i=1}^{n}(x_i-\mu_i)^2}$$

est de moyenne $1/(1+\tau^2)$, c'est-à-dire que cette variable aléatoire est un *estimateur sans biais* de la quantité $1/(1+\tau^2)$.

On peut donc remplacer cette dernière dans l'expression de l'estimateur bayésien Φ^* par une estimation sans biais. On obtient ainsi un *estimateur bayésien empirique* Φ^s défini par

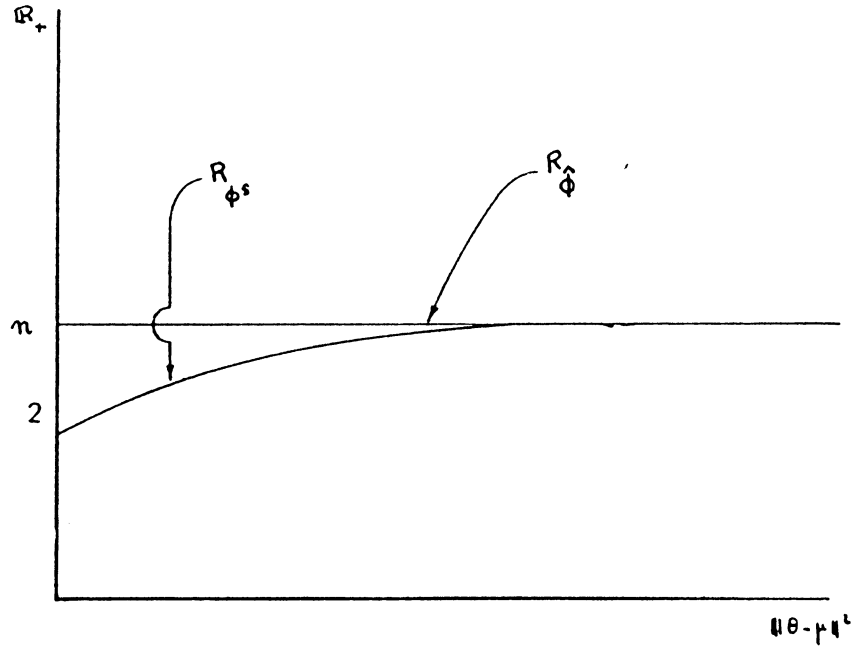
$$\forall i \in \{1,2,\cdots,n\} \quad \forall (x_1,x_2,\cdots,x_n) \in \mathbb{R}^n \quad \phi_i^s(x_1,x_2,\cdots,x_n) = \mu_i + \left(1 - \frac{n-2}{\sum_{i=1}^{n}(x_i-\mu_i)^2}\right)(x_i-\mu_i)$$

Cet estimateur porte le nom d'*estimateur de Stein.*

On reconnaît évidemment la forme de l'estimateur construit dans l'exemple introductif d'Efron et Morris dans lequel

$$\mu = \frac{1}{18}\sum_{j=1}^{18} y_i = 0,265 \qquad \text{et} \qquad c = 1 - \frac{16}{\sum_{j=1}^{18}(y_j-\mu)^2} = 0,212$$

Pour terminer, un calcul plus compliqué permet de vérifier que cet estimateur de Stein a un risque quadratique strictement inférieur à celui de l'estimateur usuel dès que $n \geq 3$.



Nous renvoyons aux deux tableaux 3 et 4 de simulations donnés en annexe pour analyer et juger des performances réciproques des différents types d'estimateurs introduits et de l'intérêt des estimateurs de Stein.

TABLEAU 3

Lois simulées	N(0,1)	N(5,1)	N(10,1)	N(15,1)	N(20,1)
	0.11	5.74	9.22	14.93	18.82
	-0.58	5.72	9.97	14.71	18.63
	1.00	6.98	10.26	14.21	19.11
	-1.62	5.08	9.41	15.41	20.56
	-0.92	4.70	10.68	14.33	22.54
	-0.07	4.13	9.67	15.03	18.91
	0.63	5.48	7.74	15.25	20.86
	0.43	4.59	8.96	14.60	20.16
	-0.08	4.97	10.98	15.31	20.75
	-1.61	4.83	10.16	16.41	20.02
	0.36	4.88	11.30	14.14	21.44
	-0.05	4.41	9.84	14.55	18.63
	-0.53	6.15	8.84	15.55	20.58
	-1.47	4.04	11.52	14.33	20.36
	0.04	5.40	12.08	15.44	20.61
	1.32	6.36	10.02	15.00	19.75
	-0.56	6.20	10.29	14.58	20.81
	0.01	5.60	10.55	13.34	20.25
	0.73	3.79	10.98	13.49	19.88
	-0.28	3.31	11.59	14.66	21.09
Estimateur usuel	-0.16	5.12	10.20	14.76	20.19
Erreur quadratique	0.025	0.014	0.041	0.056	0.035
Loi a priori N(0,1)					
Estimateur bayésien	-0.08	2.56	5.10	7.38	10.09
Erreur quadratique	0.006	5.959	23.994	58.039	98.126
Loi a priori N(0,10)					
Estimateur bayésien	-0.16	5.07	10.10	14.62	19.99
Erreur quadratique	0.024	0.004	0.010	0.147	0.000
Loi a priori N(0,5)					
Estimateur Bayésien	-0.15	4.92	9.81	14.20	19.41
Erreur quadratique	0.023	0.006	0.036	0.647	0.346

D.CELLIER : Estimation bayésienne - effet Stein

TABLEAU 4

Lois simulées	N(0,1)	N(10,1)	N(20,1)	N(30,1)	N(40,1)	N(50,1)	N(60,1)	N(70,1)	N(80,1)	N(90,1)
Valeur du paramètre	0	10	20	30	40	50	60	70	80	90
Valeurs observées	-0.26	10.51	19.21	30.25	39.00	50.86	59.01	72.09	81.15	89.54
ESTIMATEUR USUEL	-0.26	10.51	19.21	30.25	39.00	50.86	59.01	72.09	81.15	89.54
Erreurs quadratiques	0.07	0.26	0.63	0.06	1.00	0.73	0.98	4.35	1.33	0.21
Erreur quadratique globale	9.62									
ESTIMATEUR DE STEIN										
Moyenne des observations	45.13									
(Observ. - Moyen.)**2	2060.25	1199.01	672.33	221.68	37.63	32.73	192.48	726.35	1297.30	1971.87
Rétrécissement	0.00095									
Estimateur de Stein	-0.21	10.54	19.23	30.26	39.01	50.85	58.99	72.06	81.12	89.50
Erreurs quadratiques	0.05	0.29	0.59	0.07	0.99	0.72	1.01	4.24	1.25	0.25
Erreur quadratique globale	9.46									
Performance	+	-	+	-	+	+	-	+	+	-
ESTIMATEUR DE BAYES										
Loi a priori N(45.13;1)										
Estimateur bayésien	22.44	27.82	32.17	37.69	42.07	47.99	52.07	58.61	63.14	67.34
Erreur quadratique	503.53	317.59	148.11	59.14	4.27	4.02	62.86	129.73	284.14	513.59
Erreur quadratique globale	2027.00									
Performance	-	-	-	-	-	-	-	-	-	-
ESTIMATEUR DE BAYES										
Loi a priori N(45.13;10)										
Estimateur bayésien	0.19	10.85	19.46	30.39	39.06	50.80	58.87	71.82	80.80	89.10
Erreur quadratique	0.04	0.72	0.29	0.15	0.88	0.64	1.28	3.31	0.63	0.81
Erreur quadratique globale	8.75									
Performance	+	-	+	-	+	+	-	+	+	-