Probabilité, statistique, informatique et sciences : propositions et expériences pour l'enseignement transdisciplinaire

G. DI BIASE et A. MATURO - Chieti

Récemment en Italie on a introduit dans les programmes de mathématiques de plusieurs écoles des éléments de calcul des probabilités, de statistique et d'informatique. Notre contribution veut démontrer qu'il est possible de résoudre des problèmes très importants (par exemple dans le domaine de la médecine et de la biologie) à partir des concepts de base des trois disciplines susnommées.

Développer une mentalité probabiliste

L'objectif essentiel de cet article est de montrer les principes de base utiles pour préparer les élèves à une mentalité probabiliste par l'enseignement conjoint du calcul des probabilités et de la statistique avec le support de l'informatique.

En effet, l'informatique nous permet de résoudre des problèmes, même très complexes, de calcul des probabilités par simulation sur ordinateur de plusieurs expériences statistiques.

Les outils employés sont la génération des nombres au hasard et l'induction statistique.

En ce qui concerne le deuxième outil on épousera l'approche bayesienne, qui a son origine dans la conception subjective de la probabilité et qui, avec sa rigueur méthodologique et son lien naturel avec les problèmes inférenciaux, permet la fusion de la statistique et du calcul des probabilités.

Telle position est fondée sur l'emploi répété de la formule de Bayes, en simulant sur ordinateur plusieurs expériences statistiques. A l'aide des résultats obtenus, on peut mettre à jour, par une mesure quantitative, le *degré de confiance* qu'on avait *a priori* des événements. Ceux-ci expriment les causes qui ont produit les effets expérimentaux du problème examiné.

Par la position bayesienne l'ordinateur aussi joue un rôle éducatif différent. En effet, en employant cette méthodologie, l'ordinateur pousse l'utilisateur à une analyse critique des résultats des épreuves et en outre l'ordinateur devient sujet actif, au contraire de l'approche classique où il n'est employé que pour calculer des indices statistiques ou pour décrire les variables en jeu.

L'exposition du Shell-Center (ICME 7)

LA FORMULE DE BAYES DANS L'INDUCTION STATISTIQUE

Une grande partie de la statistique moderne est fondée sur la formule de Bayes. En effet il s'agit d'une façon différente de lire le théorème des probabilités composées (voir par exemple [14]). Ici on l'emploie pour résoudre des problèmes qui sont appelés : *probabilité des causes.* Il s'agit d'une catégorie de problèmes dont le but est de déterminer, à partir d'un événement connu, les probabilités de toutes les causes possibles qui pourraient le produire.

Bref, un effet s'étant réalisé, quelle est la probabilité pour qu'une certaine cause l'ait produit ?

Il est toujours possible, d'une manière ou d'une autre, d'assigner *un degré de confiance* aux événements antérieurs (causes ou hypothèses). Leurs mesures quantitatives sont appelées probabilités *a priori*, c'est-à-dire les probabilités qu'ils ont avant l'expérience, sur la base des informations possédées par la personne qui conduit l'expérience.

On se propose de déterminer leurs probabilités *a posteriori*, c'est-à-dire après l'observation des résultats de l'expérience.

Soit $H = \{H_1, H_2, \ldots H_m\}$ une partition de l'événement certain Ω, c'est-à-dire :

(i) $H_1 \cup H_2 \cup \ldots \cup H_m = \Omega$

(ii) $i \neq j \Rightarrow H_i \cap H_j = \emptyset$

Les (i) et (ii) signifient que les événements H_r, deux à deux, ne peuvent se réaliser simultanément (incompatibles) et que un et seulement un se réalisera sûrement.

On suppose que tout événement

$E \in P(\Omega)$ (l'ensemble de toutes les parties de Ω) représente une observation, c'est-à-dire l'issue d'une expérience statistique.

On veut évaluer la probabilité des événements H_r, r = 1, 2, ..., m, chacun est regardé comme une possible «explication» de l'observation E.

A partir du théorème des probabilités composées on a

$$p(E \cap H_r) = p(H_r) \cdot p(E/H_r) = p(E) \cdot p(H_r/E), \quad r = 1, 2, \ldots, m$$

et, en supposant p(E) > 0,

$$p(H_r/E) = p(H_r)\, p(E/H_r)/p(E)$$

Par ailleurs étant :

$$E = E \cap \Omega = E \cap \left(\bigcup_{r=1}^{m} H_r \right) = \bigcup_{r=1}^{m} (H_r \cap E)$$

la formule précédente peut être écrite :

$$p(H_r/E) = p(H_r) \frac{p(E/H_r)}{\sum_{r=1}^{m} p(H_r) \cdot p(E/H_r)}$$

$p(H_r)$ c'est la probabilité *a priori* de la cause, sans savoir si l'effet s'est produit ;

$p(E/H_r)$ c'est la probabilité de l'effet, sachant que la cause H_r a agi ;

$p(H_r/E)$ c'est la probabilité *a posteriori* de la cause, sachant que l'effet E s'est produit.

$\sum_{r=1}^{m} p(H_r) \cdot p(E/H_r)$ c'est une constante, indépendante de H_r.

L'égalité précédente permet, sur la base de l'observation E, d'évaluer les probabilités $p(H_r/E)$ de toutes les hypothèses H_r, que, avant la connaissance de les issues expérimentales, on avait considérées égales à $p(H_r)$.

Cette variation des probabilités, due à une augmentation de renseignement, est proportionnelle au produit $p(H_r) \cdot p(E/H_r)$ selon la constante $\left(\sum_{r=1}^{m} p(H_r) \cdot p(E/H_r) \right)^{-1}$

On appelle $p(E/H_r)$ *vraisemblable.*

Naturellement la formule de Bayes peut être répétée chaque fois que l'on connaît de nouveaux résultats expérimentaux, obtenant toujours des probabilités de plus en plus crédibles.

On simulera ces résultats par ordinateur (voir aussi [3]), en employant la génération des nombres au hasard, avec des techniques décrites dans un des paragraphes suivants.

UN EXEMPLE CLASSIQUE

Une urne renferme un nombre N (N grand) de boules blanches ou noires dont le nombre de celles de l'une des deux couleurs est inconnu ; on fait plusieurs

tirages successifs, par exemple m, en remettant la boule extraite après chaque tirage ; le résultat de ces tirages est l'extraction de k boules blanches et de m–k boules noires.

Quelle est la composition la plus probable de l'urne ?

C'est-à-dire que l'on se demandera, après l'observation d'un échantillon (avec remise) de taille m, quelle est la probabilité pour que la fraction de boules blanches dans l'urne soit $\vartheta_i = i/N$, i = 0, 1, 2, ..., N.

On peut faire sur la composition de l'urne N+1 hypothèses que nous désignerons par H_0, H_1, ..., H_N, l'hypothèse H_r étant celle où l'urne renferme r boules blanches et N–r noires.

Avant l'échantillonnage on peut faire une hypothèse plausible en supposant les probabilités a priori égales entre elles, c'est-à-dire en admettant que le N+1 compositions possibles sont représentées par N+1 urnes entre lesquelles on choisit au hasard ; ceci donne :

$p(H_i) = i/(N+1)$, pour tout i = 0, 1, 2, ..., N.

Ceci, naturellement, si on n'a pas de renseignements ultérieurs qui permettent d'assigner subjectivement d'autres probabilités a priori.

Soit E l'événement «dans l'échantillon de taille m il y a k boules blanches». La probabilité d'extraire k boules blanches avec m épreuves dans l'urne de composition H_1 est :

$$p(E/H_i) = \binom{m}{k} \cdot \vartheta_i^{k} \cdot (1-\vartheta_i)^{m-k}.$$

Nous pouvons utiliser le résultat de cette expérience pour mettre à jour les évaluations $p(H_i)$ par la formule de Bayes :

$$p(H_i/E) = p(H_i)\, p(E/H_i)/p(E)$$

ou

$$p(E) = \sum_{i=0}^{N} p(H_i) \cdot p(E/H_i).$$

On peut répéter plusieurs fois l'échantillonnage, même avec des tailles différentes, et donc l'application de la formule de Bayes, de manière que les résultats probabilistes soient de plus en plus «sûrs».

SIMULATION SUR ORDINATEUR

La simulation est réalisée par la génération des suites de nombres au hasard (voir par exemple [9] et [12]. Il s'agit de suites dont les éléments sont des nombres réels dans l'intervalle [0,1], appelés *pseudo-aléatoires*.

On emploie le suffixe *pseudo* parce que, en réalité, ils sont obtenus de façon déterministe.

Le postfixe *aléatoire* est employé parce que, tout compte fait, il est très difficile de les prévoir.

Pour ce qui concerne les aspects proprement théoriques on renvoie par exemple à [1], [7], [8], [15].

Les suites de nombres *pseudo-aléatoires* sont obtenues en pratique par ordinateur à partir d'une formule de type récurrent :

(4.1) $Y_{i+1} = f(a_1, a_2, ..., a_h ; Y_i)$, $i \in N$

f est une fonction à valeurs dans $I_m = \{0, 1, 2, ..., m-1\}$, $m \in N$, et dépend des paramètres $\alpha_1, \alpha_2, ..., \alpha_h$, où Y_i, l'argument de la fonction, parcourt l'ensemble I_m.

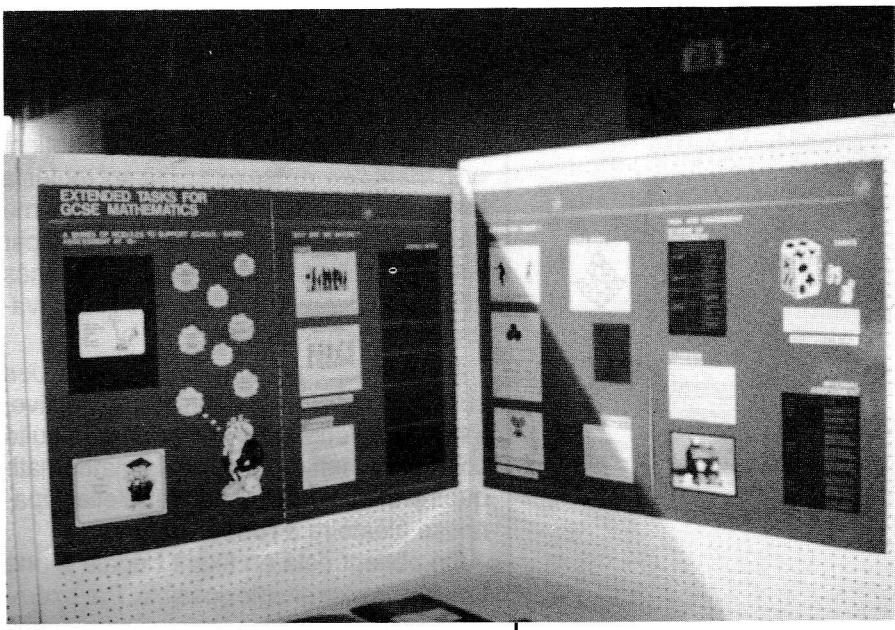
Y_1 étant fixé (valeur initiale), on obtient une suite $\{Y_i\}i \in N$ des éléments de l'ensemble I_m.

En posant $z_i = Y_i/m$ pour tout $i \in N$, on obtient une suite *pseudo-aléatoire* en [0,1].

Cette façon d'opérer est motivée par la définition subjective du nombre *pseudo-aléatoire* (voir par exemple [2] et [11]) qui affirme qu'une suite $\{z_i\}i \in N$ est appelée *pseudo-aléatoire* si :

(1) un utilisateur qui connaît le générateur peut engendrer la suite d'une façon univoque ;

(2) un utilisateur qui ne connaît pas le générateur, considère, de façon subjective, que la suite $\{z_i\}i \in N$ est une réalisation d'une suite des variables aléatoires discrètes, normalisées, de distribution uniforme.



Zoom sur le Shell-Center (ICME 7)

En effet si la fonction f et ses paramètres sont attribués de façon adéquate, la manière par laquelle on obtient Y_{i+1}, à partir de Y_1 apparaît comme le produit d'un «effet hasard» similaire à celui que l'on obtient en mêlant des cartes ou en faisant rouler la boule de la roulette. Effectivement, même dans ce cas-là, les parcours effectués par les cartes ou par la boule sont déterminés et le hasard dépend du fait que l'observateur ne les connaît pas.

Dans la formule (4.1), pour obtenir des suites qui satisfont la définition subjective, il faut que l'utilisateur qui génère la suite connaisse bien ses propriétés mathématiques et statistiques, de façon qu'elle apparaisse aléatoire à l'observateur.

On obtient cela en fixant f égale à une fonction la plus simple possible. Par exemple, une des formules les plus employées est :

$$(4.2) \quad Y_{i+1} = (a.Y_i + b) \mod m, \; i \in N$$

où a et b sont des éléments de I_m.

Les propriétés mathématiques et statistiques de la suite $\{Y_i\}i \in N$ que l'on obtient du générateur (4.2) changent remarquablement quand m modifie les paramètres a et b, le module m et la valeur initiale Y1. Pour approfondir cette classe de générateurs on renvoie à [1], [7], [8], [12], [15].

On dispose de plusieurs méthodes pour juger le «degré de hasard» d'une suite de nombres *pseudo-aléatoires*. On en peut trouver quelques-unes qui sont fondées sur une méthodologie classique (voir par exemple [10] et [11]) et d'autres qui suivent la méthodologie bayesienne (voir [4], [5], [6]).

Mais revenons au problème de l'urne. Comment procède-t-on à la simulation du tirage des boules ?

On commence par générer un nombre entier dans l'intervalle [0,N] pour simuler le choix au hasard de l'une des N+1 urnes, ou plus précisément, de l'une des N+1 compositions possibles de l'urne. Ce choix reste inconnu à l'usager jusqu'à la fin des expériences. Il n'est connu que par l'ordinateur.

Ensuite, le tirage d'un échantillon est simulé de la façon suivante :

Soit H_i l'hypothèse vraie, à laquelle correspond la fraction de boules blanches $J_i = i/N$; après la génération d'un nombre pseudo-aléatoire, ayant distribution uniforme continue, dans l'intervalle [0,1] on va le comparer à J_i : s'il est moindre que J_i cela veut dire qu'une boule blanche est tirée, sinon c'est une noire.

On répète cette opération autant de fois que la taille de l'échantillon et l'échantillonnage autant de fois que l'on veut.

La composition de l'urne est inconnue jusqu'à la fin de toutes les épreuves et, jusqu'à ce moment-là, l'ordinateur donne simplement une liste des valeurs des probabilités des hypothèses, après tout échantillonnage, dans le but de connaître la plus probable.

Il est très important (conformément à la méthodologie bayesienne) de mettre en évidence que ce n'est pas sûr que la composition qui apparaît la plus probable soit la vraie !

Evidemment, en répétant l'échantillonnage plusieurs fois les conclusions probabilistes deviennent plus «sûres».

QUELQUES APPLICATIONS DANS LES DOMAINES DES SCIENCES

A) On présente le cas concret d'un problème de décision qu'un médecin dans l'exercice de sa profession doit résoudre quotidiennement.

Il faut formuler un diagnostic sur la base de quelques symptômes relevés sur un malade.

Puisqu'un diagnostic représente la synthèse de plusieurs facteurs incertains, il est traité à la façon d'un événement aléatoire, dont la valeur de vérité ne correspond pas aux situations limites : événement certain ou impossible.

Donc, l'intervention du calcul des probabilités peut être de quelque utilité, puisqu'il permet de quantifier le degré de confiance que l'on a d'un événement. En particulier l'événement dont on parle c'est l'événement subordonné

H_r/E = «un malade a contracté la maladie H_r, sachant qu'il a les symptômes E».

En utilisant les résultats décrits dans le deuxième paragraphe, pour formuler le diagnostic, il faudra que le médecin évalue la probabilité de H_r si E :

$$p(H_r/E) = p(H_r) \frac{p(E/H_r)}{\sum_{r=1}^{m} p(H_r) \cdot p(E/H_r)}$$

$$, r = 1, 2, \ldots, n$$

$H_1, H_2, \ldots, H_i$ étant toutes les maladies

possibles relatives aux symptômes relevés.

La solution de ce problème exige la connaissance des probabilités a priori des maladies $p(H_i)$, $i = 1, 2, \ldots, n$. En Italie il y a des statistiques officielles qui périodiquement mettent à jour ces données. En outre, on doit faire des recherches cliniques pour obtenir la matrice suivante :

	S_1 S_2 •••	S_j	••• S_m
H_1 H_2 • • •			
H_i		P_{ij}	
• • • H_n			

où p_{ij} représente la fréquence relative des malades qui présentent le symptôme S_j, ayant contracté la maladie H_i.

Par exemple on peut faire ces recherches dans le service d'un hôpital ou dans le cabinet d'un médecin.

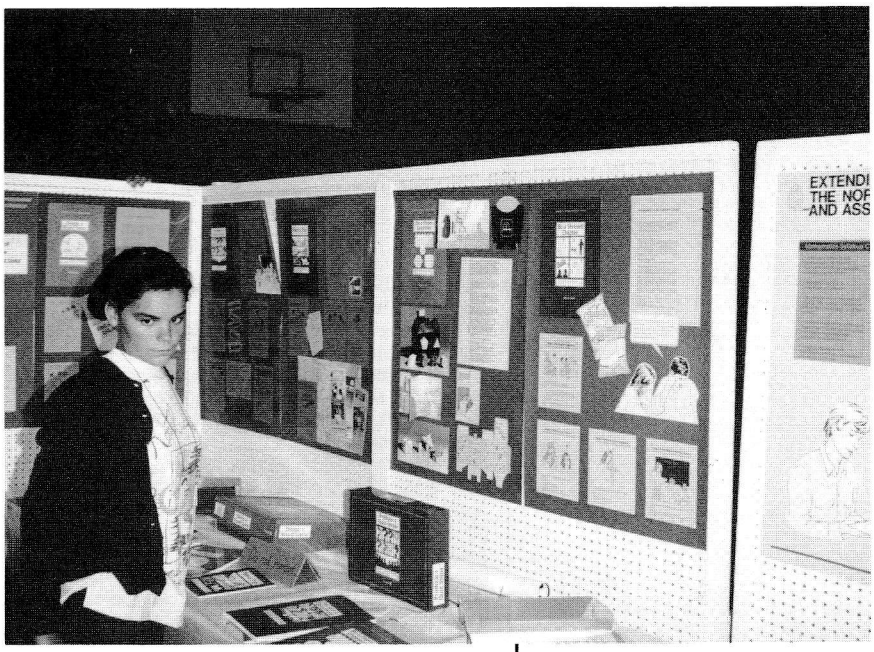
En supposant l'indépendance des symptômes, on peut calculer les $P(E/H_i) = p(S_1 \cap S_2 \cap \ldots \cap S_m/H_i)$, $i = 1, \ldots, n$, avec la formule multiplicative et, donc employer la formule de Bayes.

Le médecin, toutefois, pourrait décider d'approfondir la question et envoyer le malade faire des analyses cliniques.

La formule de Bayes pourrait être répétée en obtenant ainsi de nouvelles probabilités a posteriori. En effet, dès lors que l'on possède des informations nouvelles, les probabilités qui hier étaient a posteriori deviennent aujourd'hui les probabilités a priori.

Les jumeaux

B) Il y a deux espèces de couples de jumeaux : celle provenant d'un seul œuf fragmenté en deux (jumeaux univitellins ou vrais jumeaux), indiquée par VJ et celle



Re-zoom sur le Shell-Center (ICME 7)

venant de deux œufs simultanément émis (jumeaux bivitellins ou faux jumeaux), indiquée par FJ.

Les premiers, étant génétiquement identiques, ont le même sexe : donc on a MM ou FF avec probabilité 1/2.

Les deuxièmes se ressemblent comme deux frères et donc peuvent être de sexe différent : MM, FF, FM, ou MF avec probabilité 1/4.

Soit E l'événement «Désirée a mis au monde des jumeaux du même sexe», quelle est la probabilité qu'ils soient de vrais jumeaux ?

Dans ce problème les hypothèses H_1 et H_2 sont les événements VJ et FJ ;
on a $VJ \cup FJ = \Omega$.
Soit p la probabilité p(VJ). On peut exprimer E par :

$$E = MM \cup FF = E \cap (H_1 \cup H_2)$$
$$= (VJ \cap (MM \cup FF) \cup (FJ \cap (MM \cup FF)).$$

Donc :

$$p(E) = P(VJ) \cdot p(MM \cup FF/VJ) + p(FJ) \cdot p(MM \cup FF/FJ).$$

La probabilité recherchée est :

$$p(VJ/MM \cup FF) = p(VJ) \cdot p(MM \cup FF/VJ) / p(VJ) \cdot p(MM \cup FF/VJ) + p(FJ) \cdot p(MM \cup FF/FJ).$$

Avec ces données on a :
$p(VJ) = p$; $p(MM \cup FF/VJ) = 1$;
$p(MM \cup FF/FJ) = 1/2$, donc

$$p(VJ/MM \cup FF) = p.1 / p.1 + (1-p).1/2 = 2p / 1 + p.$$

Puisque les spécialistes en génétique affirment que $p(VJ) = 1/3$, on peut considérer, par exemple, cette valeur comme la probabilité a priori. Cela donne :
$p(VJ/MM \cup FF) = 2/3 / (1 + 1/3) = 1/2$.

Si on précise l'information précédente en disant que Désirée a mis au monde deux garçons la probabilité cherchée ne change pas.

En effet, sachant que p(MM/VJ) = 1/2 et p(MM/FJ) = 1/4, on aura :

p(VJ/MM) = p(VJ) . p(MM/VJ)
/ p(VJ) . p(MM/VJ)
+ p(FJ) . p(MM/FJ)
= 1/3 . 1/2 / (1/3 . 1/2 + 2/3 . 1/4) = 1/2.

Bibliographie

[1] Ahrens J.H. - Dieter U. - Grube A., 1970. *Pseudo-random numbers. A new proposal for the choice of multiplicators.* Compunting 6°, 1970.

[2] Baldessari B., 1987. *Aspetti probabilistici della crittografia.* Atti del 1° Simposio Naz. su «Stato e Prospettive della Ricerca Crittografica in Italia», 1987, pp. 9-21.

[3] Colagrande V. e Di Biase G., 1990. *Simulazione su calcolatore di alcuni esempi di applicazione del teorema di Bayes.* Periodico di Matematiche, n. 4, 1990, pp. 53-60.

[4] Di Biase G. e Maturo A., 1990. Analisi Bayesiana di generatori pseudorandom. Atti del XIV Conv. Ann. AMASES, Pescara, 1990, pp. 293-313.

[5] Di Biase G. e Maturo A., 1990. *Su un'analisi bayesiana per la verifica di casualità di successioni pseudorandom : software applicativo ed interpretazioni dei risultati.* Atti del Conv. Naz. «Classificazione e Analisi dei dati, metodi, Software, Applicazioni», Pescara, 1990, pp. 115-129.

[6] Di Biase G. e Maturo A., 1991. *Confronto fra inferenza bayesiana e test classici per l'analisi di successioni di numeri pseudocasuali.* Ratio Math., n. 2, 1991, pp. 97-110.

[7] Maturo A., 1989. *Numeri pseudocasuali.* Libreria dell"Università, Pescara.

[8] Maturo A., e Cera N., 1991. *Analisi della bontà di alcuni generatori di numeri pseudocasuali per la cifratura dei messaggi e la simulazione.* Ratio Math., n. 2, 1991, pp. 83-96.

[9] Maturo A., e Piscione A., 1991. *Probabilità e Statistica con il calcolatore : problematiche di carattere logico ed operativo.* Metron, in corso di stampa.

[10] Knuth D.E., 1969. *The art of Computer programming.* 1969, vol. 2°, Addison Wesley, London.

[11] Rizzi A., 1989. *Verifiche di Pseudo-Casualità in crittografia.* Atti del 2° Simposio su citato, 1989, pp. 3-21.

[12] Rizzi A., 1977. *Generazione di distribuzioni statistiche mediante un eleboratore elettronico.* Ist. di Statistica e ricerca sociale «C. Gini», Roma.

[13] Rizzi B., 1968. *Funzioni quasi periodiche e funzioni pseudo-aleatorie.* Metron, 1968, vol. XXVII, n. 1-2, pp. 1-35.

[14] Scozzafava R., 1989. *La probabilità soggettiva e le sue applicazioni.* Editoriale Veschi, Masson, Milano.

[15] Smith C.S., 1971. *Multiplicative pseudo-random numbers generators with prime modulus.* Journal of the Ass. for Compunting Machinery, 18°, 1971.