Probabilités

Antonio Falcó

Dans une phase d'évaluation, un test est appliqué à un groupe d'individus "malades" et á un groupe d'individus "non malades". La présence de la maladie est établie à l'aide d'un test de référence (gold standard) dont le résultat est considéré comme sûr. Pour chaque individu, on s'intéresse donc aux caractères suivants:

$$\begin{aligned} M &\equiv \text{ avoir la ``maladie''}, \\ NM &\equiv \text{ ne pas avoir la ``maladie''}, \\ T &\equiv \text{ avoir un résultat positif au test}, \\ NT &\equiv \text{ avoir un résultat négatif au test}. \end{aligned}$$

Un peut modelé avec deux variables,

$$X, Y : \Omega \longrightarrow \{0, 1\}$$

où

$$X(\omega) := \begin{cases} 1 & \text{si} \quad \omega \text{ est dans } M \\ 0 & \text{si} \quad \omega \text{ est dans } NM \end{cases} \qquad Y(\omega) := \begin{cases} 1 & \text{si} \quad \omega \text{ est dans } T \\ 0 & \text{si} \quad \omega \text{ est dans } NT \end{cases}$$

Alors,

$$M = \{X = 1\}, \quad NM = \{X = 0\},$$

$$T = \{Y = 1\}, \quad NT = \{Y = 0\}$$

On détermine les fréquences absolues (comptages) des quatre résultats possibles:

	M	NM	
T	$n_{T\text{et}M}$	$n_{T\text{et}NM}$	n_T
NT	$n_{NT\text{et}M}$	$n_{NT\text{et}NM}$	n_{NT}
Total	n_M	n_{NM}	n

Définitions:

Sensibilité: $= \frac{n_{T\text{et}M}}{n_M}$ proportion de test positives chez les malades.

Spécificité: $= \frac{n_{NT\text{et}NM}}{n_{NM}}$ proportion de test negatives chez les no malades.

Exemple

Le test a été administré à 1000 personnes avec M et à 1000 personnes sans M:

	M	NM	
T	950	10	960
NT	50	990	1040
Total	1000	1000	2000

Donc:

$$Sensibilit\acute{e} = \frac{950}{1000} = 95\%, \; Sp\acute{e}cificit\acute{e} = \frac{990}{1000} = 99\%.$$

Remarque

La "précision" des valeurs obtenues ("estimations") dépend du nombre d'individus testés. Cet aspect n'est pas traité ici.

Exemple

On peut construire la table suivante (on a divisé toutes les colonnes par 1000):

	M	NM	
T	0.95	0.01	
NT	0.05	0.99	
Total	1	1	2

Alors la table pour *un test idéal* doit être:

	M	NM	
T	1	0	
NT	0	1	
Total	1	1	2

Problème

Supposons que la sensibilité et la spécificité d'un certain test soient:

$$\textit{Sensibilité} = 95\%, \ \textit{Spécificité} = 99\%.$$

Le médecin applique ce test à un patient ($\omega \in \Omega$) et obtient un résultat positif ($Y(\omega) = 1$). Information *a priori*.

Est qu'on peut calculer la probabilité que le patient soit réellement malade?

$$\Pr(X = 1 | Y = 1) = \Pr(M | T)?$$

On va déduire que

	Information *a priori* M	NM	
T	$\Pr(T\|M)$	$\Pr(T\|NM)$	
NT	$\Pr(NT\|M)$	$\Pr(NT\|NM)$	
Total	1	1	2

et en utilisant la formule de Bayes on peut calculer

$$\Pr(M|T) = \frac{\Pr(T|M)\Pr(M)}{\Pr(T|M)\Pr(M) + \Pr(T|NM)\Pr(NM)}$$

avec la connaissance de $\Pr(M)$ $(\Pr(NM) = 1 - \Pr(M))$.

Pour résoudre ce Probléme une information supplémentaire est nécessaire: la fréquence (relative) de la maladie M dans la population ou prévalence. Supposons que

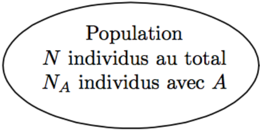
$$\text{prévalence} = \frac{1}{10000}$$

La prévalence de M dans la population est une probabilité a priori (avant connaissance du résultat du test) que le patient soit malade.

Pour encadrer ce problème, il conviendra d'utiliser les concepts fondamentaux et le formalisme du calcul des probabilités.

Considérons une population de taille N et soit N_A le nombre d'individus avec le caractère A.

Supposons le tirage au sort d'un individu.

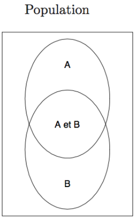


La probabilité de tirer un individu avec A est:

$$P(A) = \frac{N_A}{N}$$

On dit aussi que $P(A)$ est la probabilité de l'événement A. Dans notre définition elle coïncide avec la proportion d'individus avec la propiété A.

Supposons maintenant que les individus aient un deuxième caractère B, et indiquons par $N_{A\cap B}$ le nombre d'individus avec les deux caractères A et B simultanément.



La probabilité conjointe de A et B, l'évenement est noté par $A \cap B$, est:

$$P(A \cap B) = \frac{N_{A\cap B}}{N}.$$

Probabilité conditionelle

La probabilité conditionnelle de B sachant que A (est verifié) est:

$$P(B|A) = \frac{N_{A\cap B}}{N_A}.$$

Alors, on peut ecrire

$$P(B|A) = \frac{N_{A\cap B}}{N_A} = \frac{\frac{N_{A\cap B}}{N}}{\frac{N_A}{N}} = \frac{P(A\cap B)}{P(A)}.$$

On dit que deux événements A et B sont indépendants si:

$$P(B|A) = P(B).$$

Alors, si deux événements A et B sont indépendants:

$$P(A\cap B) = P(A)\cdot P(B).$$

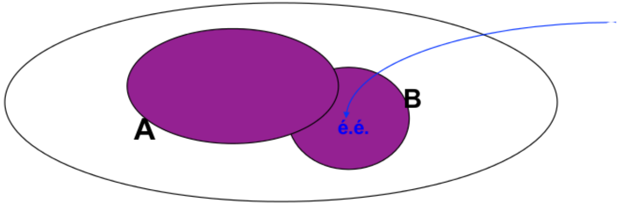
Un axiome est une propriété non démontré nécessaire pour construire une théorie.

Exemple : en Géométrie Euclidienne, "deux parallèles ne se coupent pas" est un axiome (la géométrie riemanienne n'utilise pas cet axiome).

Les 3 axiomes du calcul des probabilités

La probabilité est une function $\Pr : E \longrightarrow \mathbb{R}$ ou E est l'ensemble de tous les événements elementaires et qui a les trois propiétés suivantes.

1. La probabilité' d'un événement est un nombre compris entre 0 et 1
2. si A et B sont incompatibles ($A \cap B = \emptyset$),
$$\Pr(A \text{ ou } B) = \Pr(A \cup B) = \Pr(A) + \Pr(B)$$
3. $\Pr(E) = 1$.



s

Soit A un événement et NA l'événement contraire de A i.e.

$$A \cap NA = \emptyset \text{ et } A \cup NA = E.$$

Alors parmi 2. on a

$$\Pr(NA) = 1 - \Pr(A).$$

Si A et B sont des événements alors

$$\Pr(A \cup B) = \Pr(A) + \Pr(B) - \Pr(A \cap B).$$

L'idée est

$$A = (A \cap B) \cup (A \cap NB) \text{ et } (A \cap B) \cap (A \cap NB) = \emptyset,$$
$$B = (A \cap B) \cup (B \cap NA) \text{ et } (A \cap B) \cap (B \cap NA) = \emptyset,$$

$$A \cup B = (A \cap NB) \cup (B \cap NA) \cup (A \cap B),$$

et

$$(A \cap NB) \cap (B \cap NA) \cap (A \cap B) = \emptyset.$$

Alors

$$\Pr(A) = \Pr(A \cap B) + \underline{\Pr(A \cap NB)}$$
$$\Pr(B) = \Pr(A \cap B) + \underline{\Pr(B \cap NA)}$$
$$\Pr(A \cup B) = \underline{\Pr(A \cap NB)} + \underline{\Pr(B \cap NA)} + \Pr(A \cap B)$$

i.e

$$\Pr(A) - \Pr(A \cap B) = \underline{\Pr(A \cap NB)}$$
$$\Pr(B) - \Pr(A \cap B) = \underline{\Pr(B \cap NA)}$$

$$\begin{aligned}\Pr(A \cup B) &= \Pr(A \cap NB) + \Pr(B \cap NA) + \Pr(A \cap B)\\ &= (\Pr(A) - \Pr(A \cap B)) + (\Pr(B) - \Pr(A \cap B)) + \Pr(A \cap B)\\ &= \Pr(A) + \Pr(B) - \Pr(A \cap B).\end{aligned}$$

Probabilité conditionelle

Si A et B sont des événements. La probabilité conditionnelle de B sachant que A (est verifié) est:

$$\Pr(B|A) = \frac{\Pr(A \cap B)}{\Pr(A)}.$$

On dit que deux événements A et B sont indépendants si:

$$\Pr(B|A) = \Pr(B).$$

Alors, si deux événements A et B sont indépendants:

$$\Pr(A \cap B) = \Pr(A) \cdot \Pr(B).$$

Formule de Bayes: Si A et B sont des événements. Alors

$$\Pr(B|A) = \frac{\Pr(A \cap B)}{\Pr(A)} = \frac{\Pr(A|B)\Pr(B)}{\Pr(A)}$$

En résumé:

$$\Pr(B|A) = \frac{\Pr(A|B)\Pr(B)}{\Pr(A)}.$$

Formule de la probabilité totale: Si A et B sont des événements. Alors

$$\begin{aligned}
\Pr(A) &= \Pr(A \cap E) = \Pr(A \cap (B \cup NB)) = \Pr((A \cap B) \cup (A \cap NB)) \\
&= \Pr(A \cap B) + \Pr(A \cap NB) \\
&= \Pr(A|B)\Pr(B) + \Pr(A|NB)\Pr(NB).
\end{aligned}$$

En résumé:

$$\Pr(A) = \Pr(A|B)\Pr(B) + \Pr(A|NB)\Pr(NB).$$

Formule de la probabilité totale: Si P_1, P_2 et P_3 sont des événements telles que

$$P_1 \cap P_2 = P_1 \cap P_3 = P_2 \cap P_3 = \emptyset$$

et

$$E = P_1 \cup P_2 \cup P_3,$$

alors pour tour événement A on a:

$$\Pr(A) = \Pr(A|P_1)\Pr(P_1) + \Pr(A|P_2)\Pr(P_2) + \Pr(A|P_3)\Pr(P_3).$$

Cette formule se généralise évidemment à une partition de E en plus de trois événements.

Solution du problème à l'aide de la formule de Bayes

On veut déterminer

$$\Pr(M|T) = \text{Probabilité que le patient soit malade sachant que le résultat du test est positif.}$$

On sait que:

$$\begin{aligned} \Pr(T|M) &= 95\% = \text{sensibilité}, \\ \Pr(NT|NM) &= 99\% = \text{spécificité}, \\ \Pr(M) &= 1/10000 = \text{prévalence}, \end{aligned}$$

et donc

$$\begin{aligned} \Pr(T|NM) &= 1\% \\ \Pr(NM) &= 9999/10000. \end{aligned}$$

Selon les formules de Bayes et de la probabilité totale:

$$\Pr(M|T) = \frac{\Pr(T|M)\Pr(M)}{\Pr(T)}$$
$$= \frac{\Pr(T|M)\Pr(M)}{\Pr(T|M)\Pr(M) + \Pr(T|NM)\Pr(NM)}$$

On obtient:

$$\Pr(M|T) = \frac{0.95 \times 0.0001}{0.95 \times 0.0001 + 0.01 \times 0.9999} = 0.0094115.$$

$\Pr(M|T)$ est la probabilité a posteriori (après connaissance du résultat du test) que le patient soit malade.

Solution intuitive

La prévalence de $1/10000$ nous permet d'affirmer que dans une population hypothétique de 1.000.000 d'individus, on peut s'attendre à 100 malades et 999.900 sains. Le test dépiste 95 cas positifs et 5 cas négatifs parmi les malades, car sa sensibilité est de 95%. Le test trouve aussi 9.999 résultats positifs et 989.901 résultats négatifs dans la partie saine de la population.

Ce schéma peut aussi être représenté dans un tableau de fréquences attendues:

	M	NM	
T	95	9.999	10.094
NT	5	989.901	989.906
Total	100	999.900	1.000.000

En conclusion, la proportion de malades parmi les cas positifs est de $95/10.094$, ce qui indique que les chances qu'un individu positif au test soit réellement malade sont seulement de $0.0094(\approx 1\%)$.

Terminologie

$$\begin{aligned} \Pr(T|M) &= \text{ sensibilité du test}, \\ \Pr(NT|NM) &= \text{spécificité du test}, \end{aligned}$$

$$\begin{aligned} \Pr(M|T) &= \text{valeur prédictive positive du test}, \\ \Pr(NM|NT) &= \text{valeur prédictive négative du test}, \end{aligned}$$

$$\begin{aligned} \Pr(T|NM) &= \text{ taux de faux positifs} = 1 - \text{spécificité}, \\ \Pr(NT|M) &= \text{ taux de faux négatifs } = 1 - \text{ sensibilité}. \end{aligned}$$

Attention pour certains auteurs:

$$\begin{aligned} \text{taux de faux positifs} &= \Pr(NM|T), \\ \text{taux de faux négatifs } &= \Pr(M|NT). \end{aligned}$$

Vérifiez donc toujours la définition utilisée !

Le rôle de la prévalence Il est souvent difficile de connaître $\Pr(M)$ avec précision. Il convient alors d'examiner le test pour différentes valeurs de $\Pr(M)$.

Par exemple, si $\Pr(T|M) = 0.95$ et $\Pr(NT|NM) = 0.99$, on obtient:

$\Pr(M)$	$\Pr(NM\|T)$	$\Pr(M\|NT)$
1/1.000.000	0.9999	0.00000
1/100.000	0.9991	0.00000
1/10.000	0.9906	0.00001
1/1.000	0.9132	0.00005
1/500	0.8401	0.00010
1/200	0.6769	0.00025
1/100	0.5103	0.00051

La taux $\Pr(M|NT)$ est faible: dans le pire des 10.000 échappent au test. Par contre le taux $\Pr(NM|T)$ est élevé ($> 50\%$) : sur 100 individus positifs plus de 50 sont sains. La décision de maintenir un tel test dépendra de l'importance de la maladie, des conséquences du test, des coûts des examens complémentaires et de l'éventuel traitement, des chances de succès du traitement, etc.

Évaluation basée sur un seul échantillon Dans certaines études d'évaluation, on ne considère pas deux groupes séparés (malades et non malades) de tailles fixées (n_M et n_{NM}): un seul échantillon de taille n est étudié; ses éléments sont classés dans les quatres cases du tableau:

	M	NM	
T	$n_{T \text{ et } M}$	$n_{T \text{ et } NM}$	n_T
NT	$n_{NT \text{ et } M}$	$n_{NT \text{ et } NM}$	n_{NT}
Total	n_M	n_{NM}	n

On obtient

$$\text{sensibilité du test} = \Pr(T|M) = \frac{\Pr(T \cap M)}{\Pr(M)} \approx \frac{n_{T \text{ et } M}}{n_M},$$

$$\text{spécificité du test} = \Pr(NT|NM) = \frac{\Pr(NT \cap NM)}{\Pr(NM)} \approx \frac{n_{NT \text{ et } NM}}{n_{NM}},$$

$$\text{valeur prédictive + du test} = \Pr(M|T) \approx \frac{n_{M \text{ et } T}}{n_T},$$

$$\text{valeur prédictive - du test} = \Pr(NM|NT) \approx \frac{n_{NM \text{ et } NT}}{n_{NT}},$$

On a de distinguer entre la probabilité théorique:

$$\Pr(T|M)$$

la probabilité empirique qu'on calcule à partir des observations:

$$P(T|M) = \frac{n_{T \text{ et } M}}{n_M}$$

Alors, dans la pratique:

$$\Pr(T|M) \approx P(T|M).$$