Fête de la Science 2006

Mathématique du hasard

UNIVERSITÉ LOUIS PASTEUR STRASBOURG

Institut de Recherche Mathématique Avancée

La formule de Bayes : application aux tests de dépistage

La problématique biologique.

Dans de nombreux contextes, médicaux, sportifs ou biologiques par exemple, les praticiens sont amenés à réaliser des tests. Le test de dépistage du SIDA, du cancer du sein chez la femme, de la présence de substances dopantes en sont des exemples connus de tous. On appellera sain un individu ne présentant pas la caractéristique testée et malade un individu la présentant.

La chimie d'un individu est un ensemble de paramètres très complexes et les tests dont l'on se sert ne sont pas fiables à 100 %. C'est-à-dire qu'il est possible qu'il donne un résultat contraire à la réalité biologique : être sain alors que l'on est malade, être malade alors que l'on est sain. Comment savoir quel est le nombre de chances, la probabilité, d'être dans l'une de ces situations ?

Mesure expérimentale des performances d'un test

1. **La sensibilité :** la sensibilité d'un test est sa capacité à détecter les cas d'une maladie. On considère alors un groupe de personnes dont on est assuré qu'ils possèdent ce que l'on cherche à détecter. Deux situations peuvent alors se présenter.
   - Les individus pour lesquels le résultat du test est positif sont les **vrais positifs** (VP).
   - Les individus pour lesquels le résultat du test est négatif sont les **faux négatifs** (FN).

   La sensibilité du test est alors le rapport $\dfrac{VP}{VP+FN}$.

2. **La spécificité :** la spécificité d'un test est sa capacité à identifier correctement les individus qui ne sont pas atteints par la maladie. Pour mesurer la spécificité d'un test, on doit disposer d'un groupe de personnes dont on est sûr qu'aucune d'entre elles ne présente ce que l'on cherche à détecter à l'aide du test. À nouveau, deux situations peuvent se présenter.
   - Les individus pour lesquels le résultat du test est positif sont les **faux positifs** (FP).
   - Les individus pour lesquels le résultat du test est négatif sont les **vrais négatifs** (VN).

   La spécificité du test est alors le rapport $\dfrac{VN}{VN+FP}$.

   On a reproduit dans le tableau ci-contre les différents cas de figure possibles.

	Malade	Sain
Test positif	VP	FP
Test négatif	FN	VN

3. **Les valeurs prédictives :**
   Quelle **confiance** accorder au **résultat** du test ? On cherche alors à connaître la probabilité d'être malade si le test est positif et celle d'être sain si le test est négatif. Ce sont les **valeurs prédictives** du test.

   - On appelle valeur prédictive positive d'un test la probabilité d'être malade lorsque le résultat est positif.
   - On appelle valeur prédictive négative d'un test la probabilité d'être sain lorsque le résultat est négatif.

   Nous allons maintenant mettre le problème en équations.

Évènements

On considère une population d'individus Ω, c'est-à-dire un ensemble de personnes. On peut choisir à l'intérieur de ce groupe certaines personnes qui présentent des caractéristiques communes: les hommes, les femmes, ...

Dans le langage mathématique un tel regroupement est appelé un évènement. On note souvent les évènements par des lettres majuscules comme A. Introduisons S l'évènement être sain et M l'évènement être malade.

Probabilités conditionnelles

À tout évènement A on peut associer une probabilité, celle qu'un individu pris au hasard dans la population présente la caractéristique A. La définition est conforme à l'intuition : il s'agit du nombre de personnes qui présentent la caractéristique A divisé par le nombre total de personnes de la population. C'est un nombre qui est toujours compris entre 0 et 1. Ainsi :

$$\mathbb{P}[S] = \frac{\text{Nombre de sains}}{\text{Nombre total de personnes}} \qquad \mathbb{P}[M] = \frac{\text{Nombre de malades}}{\text{Nombre total de personnes}}.$$

Essayons maintenant de répondre à une question plus complexe : quelle est la probabilité que le test soit négatif sachant que l'on est sain. En posant cette question on a implicitement introduit deux évènements : S l'évènement être sain et N l'évènement avoir un test négatif. Pour pouvoir connaître la probabilité recherchée il faut bien entendu qu'il y ait des individus sains dans la population que l'on étudie et donc que $\mathbb{P}[S] > 0$. On dit que cette probabilité est la probabilité conditionnelle de N sachant S. Elle se calcule ainsi :

$$\mathbb{P}[N|S] = \frac{\text{Nombre de personnes saines et ayant un test négatif}}{\text{Nombre de personnes saines}} = \text{Spécificité}.$$

De même la probabilité conditionnelle d'avoir un test positif alors que l'on est malade est :

$$\mathbb{P}[P|M] = \frac{\text{Nombre de personnes malades et ayant un test positif}}{\text{Nombre de personnes malades}} = \text{Sensibilité}.$$

Les valeurs prédictives que l'on cherche à calculer sont $\mathbb{P}[M|P]$ et $\mathbb{P}[S|N]$.

La formule de Bayes

Cette formule va nous permettre de calculer les valeurs prédictives à l'aide des valeurs de $\mathbb{P}[M]$, $\mathbb{P}[S]$, $\mathbb{P}[P|M]$ et $\mathbb{P}[N|S]$:

$$\mathbb{P}[M|P] = \frac{\mathbb{P}[P|M] \times \mathbb{P}[M]}{\mathbb{P}[P|M] \times \mathbb{P}[M] + (1 - \mathbb{P}[N|S]) \times \mathbb{P}[S]}$$

$$\mathbb{P}[S|N] = \frac{\mathbb{P}[N|S] \times \mathbb{P}[S]}{\mathbb{P}[N|S] \times \mathbb{P}[S] + (1 - \mathbb{P}[P|M]) \times \mathbb{P}[M]}$$

Deux exemples

- **Le dépistage du SIDA** :

$\mathbb{P}[M] = 0,0001$, $\mathbb{P}[S] = 0,9999$, $\mathbb{P}[P|M] = 0,99$ et $\mathbb{P}[N|S] = 0,999$

$$\mathbb{P}[M|P] = \frac{0,99 \times 0,0001}{0,99 \times 0,0001 + (1 - 0,999) \times 0,9999} = \frac{10}{111} \simeq 0,090$$

$$\mathbb{P}[S|N] = \frac{0,999 \times 0,9999}{0,999 \times 0,9999 + (1 - 0,99) \times 0,0001} = \frac{9989001}{9989011} \simeq 0,999$$

Les valeurs prédictives obtenues permettent de conclure que :

- Avec une quasi-certitude, on est sain si le test est négatif.
- Que même si le test est positif, il y a **91 chances sur 100 que l'on soit sain !**

- **Le dépistage du cancer du sein chez la femme** :

$\mathbb{P}[M] = 0,20$, $\mathbb{P}[S] = 0,80$, $\mathbb{P}[P|M] = 0,98$ et $\mathbb{P}[N|S] = 0,90$

$$\mathbb{P}[M|P] = \frac{0,98 \times 0,20}{0,98 \times 0,20 + (1 - 0,90) \times 0,80} = \frac{49}{69} \simeq 0,710$$

$$\mathbb{P}[S|N] = \frac{0,90 \times 0,8}{0,90 \times 0,8 + (1 - 0,98) \times 0,2} = \frac{180}{181} \simeq 0,994$$

Les valeurs prédictives obtenues permettent de conclure que :

- Avec une quasi-certitude, on est sain si le test est négatif.
- Que même si le test est positif, il y a 28 chances sur 100 que l'on soit sain.

On comprend alors pourquoi on fait **systématiquement des analyses complémentaires** lorsqu'un test est positif. Le premier exemple est particulièremet frappant : le test utilisé a de très **bonnes performances** ($\mathbb{P}[\mathbf{S}] = \mathbf{0,9999}$ et $\mathbb{P}[\mathbf{P}|\mathbf{M}] = \mathbf{0,99}$) et pourtant cela **ne suffit pas** ! Le nombre de malades au sein de la population, 1 pour 10000, est **trop faible** .