Chapitre 2

Inférence bayesienne

Dans ce chapitre, nous allons voir de quelle manière certains aspects de la théorie des probabilités peuvent-être utiles dans d'autres domaines. Voyons un exemple de cela.

Exemple 2.0.1. Imaginons que nous disposions d'un test permettant de détecter si un individu est atteint d'une certaine maladie[1]. Puisque la mise en place de ces tests repose sur de nombreux facteurs (l'individu, le prélèvement médical, le traitement des données,...), il ne peut être fiable à 100%. En réfléchissant, nous constatons que quatre situations peuvent se produire :

- **Vrai positif** (noté $V+$) : le test est positif et l'individu est effectivement malade.
- **Faux positif** (noté $F+$) : le test est positif alors que l'individu n'est pas malade.
- **Vrai négatif** (noté $V-$) : le test est négatif et l'individu n'est pas malade.
- **Faux négatif** (noté $F-$) : le test est négatif alors que l'individu est malade.

Tout ceci se résume dans un **tableau de contingence** :

	Malade	Non malade
Test +	$V+$	$F+$
Test −	$F-$	$V-$

Ce tableau nous permettra de calculer toutes les caractéristiques probabilistes du test.

Remarque. Bien entendu, seuls les cas *faux négatif* et *faux positif* sont véritablement ennuyeux. Puisque c'est à ce moment là que le test est défaillant.

Voyons sur des exemples les quantités qu'il faudra être capable de déterminer. Pour cela nous allons travailler sur l'exemple de détection de *spams* sur une boite mail.

Exemple 2.0.2. Voici des données tirées de l'ouvrage *Enseignement scientifique - Belin*. Un test cherchant à détecter les spams parmi les mails reçu par un usager. Sur 4 326 mails il y avait 1 351 spam et 2 975 mails normaux. Le tableau de contingence associé est le suivant[2]

1. Comme nous le verrons, d'autres exemples sont envisageables : grossesses, spams,...
2. Celui-ci a été dressé par les gens qui voulaient tester leur algorithme permettant de détecter les spams.

	Spam	Non spam	Total
Test +	932	60	992
Test −	419	2915	3334
Total	1351	2975	4326

1. Tout d'abord, comme vu en classe de seconde, nous pouvons déterminer la **fréquence de faux positifs** (noté encore $F+$) et de **faux-négatifs** (noté à nouveau $F-$) :

$$F+ = \frac{60}{992} \approx 0.06 \quad \text{et} \quad F- = \frac{419}{3334} \approx 0.12.$$

Dans le premier cas, nous avons compté le nombre de mails normaux parmi les mails qui ont été considéré comme des spams (ceux avec un test positif) ; dans le second cas, nous avons dénombré les spams qui ont été considéré comme des mails normaux (ceux avec un test négatif).

2. Ensuite, nous pouvons déterminer la **Sensibilité** (notée S_e) du test : il s'agit de connaitre la **probabilité que le test soit positif lorsqu'il est censé l'être** (quand il s'agit effectivement d'un spam, d'un individu malade,...). Cela permet de mesure **l'efficacité du test**. Ici, avons

$$S_e = \frac{932}{1351} \approx 0,68.$$

3. Il faut également déterminer la **Spécificité** (notée S_p) du test : il s'agit de connaitre la **probabilité que le test soit négatif lorsqu'il est censé l'être**[3]. Ici, nous avons

$$S_p = \frac{2915}{2975} \approx 0,98.$$

Remarque. Observons que les fréquences calculées sont obtenues via les lignes tandis que la spécificité et la sensibilité sont obtenues via les colonnes du tableau de contingence.

Exercice à traiter : 1 et 2 (facultatif).

Poursuivons notre étude.

Exemple 2.0.3. La **prévalence** (note P_r) désigne la **proportion d'une population portant le caractère étudié** (être malade, être un spam,...). En conservant l'exemple des spams, nous voyons que

$$P_r = \frac{\text{nombres de spams}}{\text{nombres total de mails}} = \frac{1351}{4326} \approx 0.31.$$ [4]

Cette nouvelle quantité étant introduite, nous pouvons chercher à répondre à la question suivante :

comment estimer la probabilité d'être malade sachant que le test est positif.

3. Il parait naturel de requérir que le test ne détecte rien s'il n'y a rien.
4. Plus simplement : il s'agit de la fréquence du caractère « spam » parmi tous les mails reçus.

La probabilité que nous souhaitons trouver est appelée **valeur prédictive positive** (notée VPP) et celle-ci est donnée par

$$VPP = \frac{\text{proportion de } V+}{\text{proportion de tests positifs}} = \frac{932}{992} \approx 0.94.$$ [5]

En fait, il se trouve qu'il est possible d'exprimer cette quantité en fonction de P_r, S_e et S_p. [6]

Proposition 4 (Valeur prédictive positive)**.** *Avec les notations introduites plus haut, nous avons*

$$VPP = \frac{\textit{aire du rectangle hachuré en rouge}}{\textit{aire du rectangle hachuré en rouge} + \textit{aire du rectangle hachuré en bleu}}$$

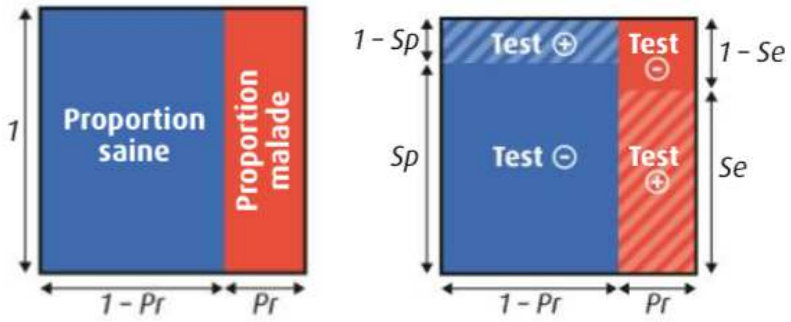

FIGURE 2.1: Représentation schématique de la population

Autrement dit :

$$VPP = \frac{P_r \times S_e}{P_r \times S_e + (1 - P_r) \times (1 - S_p)}.$$

Démonstration. La démonstration de ce résultat est assez simple et peu éclairer les lecteurs les plus curieux ayant suivi un cours de spécialité mathématiques en classe de 1ère.

Notons à nouveau M l'évènement « être malade » et T l'évènement « le test est positif ». La valeur prédictive positive correspond à $\mathbb{P}_T(M) = \frac{\mathbb{P}(M \cap T)}{\mathbb{P}(T)}$ la probabilité d'être malade sachant que le test est positif. Observons alors que

$$\mathbb{P}(M \cap T) = \mathbb{P}(M) \times \mathbb{P}_M(T) = P_r \times S_e.$$

La première égalité découle de la définition des probabilités conditionnelles. En outre,

$$\begin{aligned} \mathbb{P}(T) = \mathbb{P}(M \cap T) + \mathbb{P}(\overline{M} \cap T) &= P_r \times S_e + \mathbb{P}(\overline{M}) \times \mathbb{P}_{\overline{M}}(T) \\ &= P_r \times S_e + \big(1 - \mathbb{P}(M)\big) \times \mathbb{P}_{\overline{M}}(T) \\ &= P_r \times S_e + (1 - P_r) \times (1 - S_p). \end{aligned}$$

5. Si M désigne l'évènement « être malade » et T désigne l'évènement « le test est positif », la valeur prédictive positive n'est rien d'autre que $\mathbb{P}_T(M) = \frac{\mathbb{P}(M \cap T)}{\mathbb{P}(T)}$ la probabilité d'être malade sachant que le test est positif.

6. C'est à cet endroit qu'intervient Bayes et son théorème

Rappelons que $\overline{M}$ désigne l'évènement complémentaire de M.

□

Exercices à traiter : 3 et 4.